



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0010489
(43) 공개일자 2020년01월30일

- (51) 국제특허분류(Int. Cl.)
G16B 20/00 (2019.01) G06K 9/62 (2006.01)
G06N 3/04 (2006.01) G06N 3/08 (2006.01)
G16B 30/00 (2019.01) G16B 40/20 (2019.01)
G16B 50/00 (2019.01)
- (52) CPC특허분류
G16B 20/00 (2019.02)
G06K 9/6267 (2013.01)
- (21) 출원번호 10-2019-7038078
- (22) 출원일자(국제) 2018년10월15일
심사청구일자 2019년12월23일
- (85) 번역문제출일자 2019년12월23일
- (86) 국제출원번호 PCT/US2018/055919
- (87) 국제공개번호 WO 2019/079200
국제공개일자 2019년04월25일
- (30) 우선권주장
62/573,125 2017년10월16일 미국(US)
(뒷면에 계속)

- (71) 출원인
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
- (72) 발명자
자가나탄 키쇼르
미국 캘리포니아주 92122 샌디에이고 5200 일루미나 웨이
파 카이-하우
미국 캘리포니아주 92122 샌디에이고 5200 일루미나 웨이
(뒷면에 계속)
- (74) 대리인
특허법인아주김장리

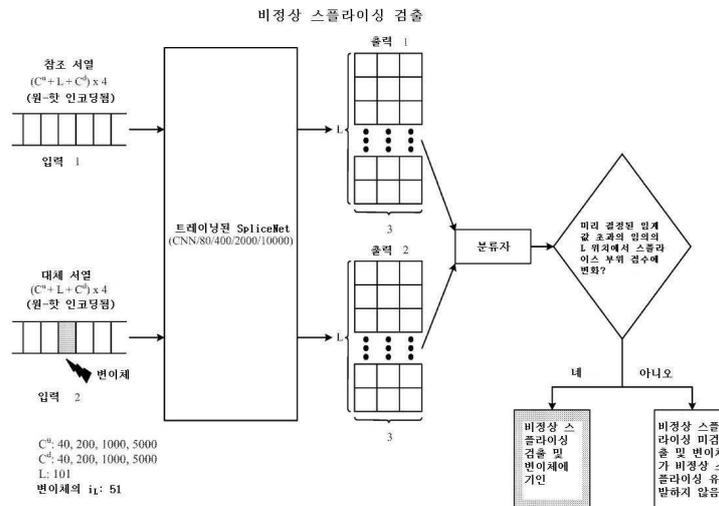
전체 청구항 수 : 총 38 항

(54) 발명의 명칭 심층 학습 기반 비정상 스플라이싱 검출

(57) 요약

개시된 기술은 변이체 분류를 위한 컨볼루션 신경망-기반 분류자의 구성에 관한 것이다. 구체적으로, 이 기술은, 컨볼루션 신경망-기반 분류자의 출력을 대응하는 실측 자료 표지와 점진적으로 매칭시키는 역전파-기반 그라디언트 업데이트 기술을 사용하여 트레이닝 데이터에 대한 컨볼루션 신경망-기반 분류자를 트레이닝하는 것에 관한 것이다. 컨볼루션 신경망-기반 분류자는 잔여 블록의 그룹을 포함하며, 잔여 블록의 각 그룹은, 잔여 블록의 컨볼루션 필터의 수, 잔여 블록의 컨볼루션 윈도우 크기 및 잔여 블록의 아트로스 컨볼루션 레이트에 의해 파라미터화되고, 컨볼루션 윈도우의 크기는 잔여 블록의 그룹들 간에 달라지며, 아트로스 컨볼루션 레이트는 잔여 블록의 그룹들 간에 달라진다. 트레이닝 데이터는, 양성 변이체 및 병원성 변이체로부터 생성되는 번역된 서열 쌍들의 양성 트레이닝 예와 병원성 트레이닝 예를 포함한다.

대표도 - 도34



(52) CPC특허분류

G06N 3/0472 (2013.01)

G06N 3/0481 (2013.01)

G06N 3/08 (2013.01)

G16B 30/00 (2019.02)

G16B 40/20 (2019.02)

G16B 50/00 (2019.02)

(72) 발명자

키리아조플루 파나기오토폴루 소피아

미국 캘리포니아주 92122 샌디에이고 5200 일루미
나 웨이

맥레 제레미 프란시스

미국 캘리포니아주 92122 샌디에이고 5200 일루미
나 웨이

(30) 우선권주장

62/573,131 2017년10월16일 미국(US)

62/573,135 2017년10월16일 미국(US)

62/726,158 2018년08월31일 미국(US)

명세서

청구범위

청구항 1

비정상 스플라이싱 검출기로서,

병렬로 작동하고 메모리에 연결된 다수의 프로세서;

상기 다수의 프로세서 상에서 실행되는 트레이닝된 아트러스 컨볼루션 신경망(atrous convolutional neural network: ACNN)으로서,

입력 서열(input sequence)에서 표적 뉴클레오타이드를 분류하고, 상기 표적 뉴클레오타이드의 각각이 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 할당하되, 상기 입력 서열은 적어도 801개의 뉴클레오타이드를 포함하고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되는, 상기 트레이닝된 ACNN; 및

상기 다수의 프로세서 중 적어도 하나 상에서 실행되는 분류자(classifier)로서,

상기 ACNN을 통해 참조 서열 및 변이체 서열을 처리하여 상기 참조 서열에서의 그리고 상기 변이체 서열에서의 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 생성하되, 상기 참조 서열 및 상기 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 갖고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되고, 그리고

상기 참조 서열에서의 상기 표적 뉴클레오타이드와 상기 변이체 서열에서의 상기 표적 뉴클레오타이드의 상기 스플라이스 부위 점수의 차이로부터, 상기 변이체 서열을 생성한 변이체가 비정상 스플라이싱을 유발하고 따라서 병원성인지의 여부를 결정하는, 상기 분류자를 포함하는, 비정상 스플라이싱 검출기.

청구항 2

제1항에 있어서, 상기 스플라이스 부위 점수의 상기 차이가 상기 참조 서열에서의 상기 표적 뉴클레오타이드와 상기 변이체 서열에서의 표적 뉴클레오타이드 간의 위치에 따라 결정되는, 비정상 스플라이싱 검출기.

청구항 3

제1항 또는 제2항에 있어서, 적어도 표적 뉴클레오타이드 위치에 대해서, 상기 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값을 초과할 경우, 상기 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 4

제1항 내지 제3항 중 어느 한 항에 있어서, 적어도 표적 뉴클레오타이드 위치에 대해서, 상기 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값 미만인 경우, 상기 변이체를 비정상 스플라이싱을 유발하지 않고 따라서 양성인 것으로 분류하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 5

제1항 내지 제4항 중 어느 한 항에 있어서, 상기 임계값은,

복수의 후보 임계값에 대해서,

양성 공통 변이체에 의해 생성된 제1 세트의 참조 및 변이체 서열쌍을 처리하여 제1 세트의 비정상 스플라이싱 검출을 초래하고;

병원성 희귀 변이체에 의해 생성된 제2 세트의 참조 및 변이체 서열쌍을 처리하여 제2 세트의 비정상 스플라이싱 검출을 초래하고; 그리고

분류자에 의해 사용하기 위해, 상기 제2 세트에서 비정상 스플라이싱 검출의 계수치를 최대화하고 상기 제1 세트에서 비정상 스플라이싱 검출의 계수치를 최소화하는 적어도 하나의 임계값을 선택함으로써, 결정되는, 비정상 스플라이싱 검출기.

청구항 6

제1항 내지 제5항 중 어느 한 항에 있어서, 상기 참조 서열 및 상기 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 갖고 그리고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 5000개의 뉴클레오타이드가 축적되는, 비정상 스플라이싱 검출기.

청구항 7

제1항 내지 제6항 중 어느 한 항에 있어서, 상기 참조 서열에서의 상기 표적 뉴클레오타이드의 상기 스플라이스 부위 점수는 상기 ACNN의 제1 출력에서 인코딩되고 그리고 상기 변이체 서열에서의 상기 표적 뉴클레오타이드의 상기 스플라이스 부위 점수는 상기 ACNN의 제2 출력에서 인코딩되는, 비정상 스플라이싱 검출기.

청구항 8

제1항 내지 제7항 중 어느 한 항에 있어서, 상기 제1 출력은 제1 101×3 매트릭스로서 코딩되고 그리고 상기 제2 출력은 제2 101×3 매트릭스로서 인코딩되는, 비정상 스플라이싱 검출기.

청구항 9

제1항 내지 제8항 중 어느 한 항에 있어서, 상기 제1 101×3 매트릭스의 각 행은 상기 참조 서열에서의 상기 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타내는, 비정상 스플라이싱 검출기.

청구항 10

제1항 내지 제9항 중 어느 한 항에 있어서, 상기 제2 101×3 매트릭스의 각 행은 상기 변이체 서열에서의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타내는, 비정상 스플라이싱 검출기.

청구항 11

제1항 내지 제10항 중 어느 한 항에 있어서, 상기 제1 101×3 매트릭스 및 상기 제2 101×3 매트릭스의 각 행에 서의 스플라이스 부위 점수는 지수적으로 정규화되고 단일로 합산되는, 비정상 스플라이싱 검출기.

청구항 12

제1항 내지 제11항 중 어느 한 항에 있어서, 상기 분류자는 상기 제1 101×3 행렬과 상기 제2 101×3 행렬의 행-대-행 비교를 수행하고, 행 단위로, 스플라이스 부위 점수의 분포의 변화를 결정하는, 비정상 스플라이싱 검출기.

청구항 13

제1항 내지 제12항 중 어느 한 항에 있어서, 상기 행-대-행 비교의 적어도 하나의 경우에 대해서, 분포의 변화가 미리 결정된 임계값을 초과할 경우, 상기 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 14

제1항 내지 제13항 중 어느 한 항에 있어서, 상기 참조 서열 및 상기 변이체 서열은 원-핫 인코딩되는(one-hot encoded), 비정상 스플라이싱 검출기.

청구항 15

제1항 내지 제14항 중 어느 한 항에 있어서,

유전 장애를 갖는 개체의 코호트로부터 샘플링된 특정 유전자에 대해서,

상기 트레이닝된 ACNN을 적용하여 비정상 스플라이싱을 유발하는 상기 특정 유전자의 후보 변이체를 식별하고;

상기 후보 변이체의 관찰된 트라이뉴클레오타이드 돌연변이율을 합산하고 그 합에 전달 계수치 및 상기 코호트의 크기를 곱한 것에 기초하여 상기 특정 유전자에 대한 돌연변이의 베이스라인 수를 결정하고;

상기 트레이닝된 ACNN을 적용하여 비정상 스플라이싱을 유발하는 상기 특정 유전자에서 드 노보 변이체를 식별하고; 그리고

상기 돌연변이의 베이스라인 수를 상기 드 노보 변이체의 계수치와 비교하고, 비교의 출력에 기초하여, 상기 특정 유전자가 상기 유전 장애에 연관되어 있음과 상기 드 노보 변이체가 병원성임을 결정함으로써

비정상 스플라이싱을 유발하는 것으로 결정된 변이체의 병원성을 결정하는 유전자당 농축 분석(per-gene enrichment analysis)을 구현하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 16

제1항 내지 제15항 중 어느 한 항에 있어서, p-값을 출력으로서 생성하는 통계 테스트를 사용하여 상기 비교를 수행하도록 더 구성된, 비정상 스플라이싱 검출기.

청구항 17

제1항 내지 제16항 중 어느 한 항에 있어서, 상기 돌연변이의 베이스라인 수를 상기 드 노보 변이체의 계수치와 비교하고, 비교의 출력에 기초하여, 상기 특정 유전자가 상기 유전 장애와 연관되지 않음과 상기 드 노보 변이체가 양성임을 결정하도록 더 구성된, 비정상 스플라이싱 검출기.

청구항 18

제1항 내지 제17항 중 어느 한 항에 있어서, 상기 유전 장애는 자폐 스펙트럼 장애(autism spectrum disorder: ASD)인, 비정상 스플라이싱 검출기.

청구항 19

제1항 내지 제18항 중 어느 한 항에 있어서, 상기 유전 장애는 발달 지연 장애(developmental delay disorder: DDD)인, 비정상 스플라이싱 검출기.

청구항 20

제1항 내지 제19항 중 어느 한 항에 있어서, 상기 후보 변이체의 적어도 일부는 단백질-절단 변이체인, 비정상 스플라이싱 검출기.

청구항 21

제1항 내지 제20항 중 어느 한 항에 있어서, 상기 후보 변이체의 적어도 일부는 미스센스 변이체인, 비정상 스플라이싱 검출기.

청구항 22

제1항 내지 제21항 중 어느 한 항에 있어서,

상기 트레이닝된 ACNN을 적용하여 건강한 개체의 코호트로부터 샘플링된 복수의 유전자 내에서 비정상 스플라이싱을 유발하는 드 노보 변이체들의 제1 세트를 식별하고;

상기 트레이닝된 ACNN을 적용하여 유전 장애가 있는 개체의 코호트로부터 샘플링된 상기 복수의 유전자 내에서 비정상 스플라이싱을 유발하는 드 노보 변이체들의 제2 세트를 식별하고; 그리고

상기 제1 세트 및 상기 제2 세트의 각 계수치를 비교하고, 비교의 출력에 기초하여, 상기 드 노보 변이체들의 제2 세트가 유전 장애가 있는 개체의 코호트에 농축되어 있고 이에 따라 병원성임을 결정함으로써,

비정상 스플라이싱을 유발하는 것으로 결정된 변이체의 병원성을 결정하는 게놈 전체 농축 분석(genome-wide enrichment analysis)을 구현하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 23

제1항 내지 제22항 중 어느 한 항에 있어서, p-값을 출력으로서 생성하는 통계 테스트를 사용하여 상기 비교를 수행하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 24

제1항 내지 제23항 중 어느 한 항에 있어서, 상기 비교는 각각의 코호트 크기에 의해 추가로 파라미터화되는, 비정상 스플라이싱 검출기.

청구항 25

제1항 내지 제24항 중 어느 한 항에 있어서, 상기 제1 세트 및 상기 제2 세트의 각각의 계수치를 비교하고, 비교의 출력에 기초하여, 상기 드 노보 변이체들의 제2 세트가 유전 장애를 가진 개체의 코호트에 농축되어 있지 않으며 이에 따라 양성임을 결정하도록 더 구성되는, 비정상 스플라이싱 검출기.

청구항 26

제1항 내지 제25항 중 어느 한 항에 있어서, 상기 유전 장애는 ASD인, 비정상 스플라이싱 검출기.

청구항 27

제1항 내지 제26항 중 어느 한 항에 있어서, 상기 유전 장애는 DDD인, 비정상 스플라이싱 검출기.

청구항 28

비정상적 스플라이싱을 유발하는 계놈 변이체를 검출하는 방법으로서,

표적 하위서열(target sub-sequence)에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 차등 스플라이싱 패턴을 검출하도록 트레이닝된 아트러스 컨볼루션 신경망(약칭 ACNN)을 통해 참조 서열을 처리하는 단계;

상기 처리에 기초하여, 참조 표적 하위서열에서의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 참조 표적 하위서열에서 제1 차등 스플라이싱 패턴을 검출하는 단계;

상기 ACNN을 통해 변이체 서열을 처리하는 단계로서, 상기 변이체 서열과 상기 참조 서열은 변이체 표적 하위서열에 위치한 하나 이상의 변이체 뉴클레오타이드만큼 상이한, 상기 변이체 서열을 처리하는 단계;

상기 처리에 기초하여, 상기 변이체 표적 하위서열의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 변이체 표적 하위서열에서 제2 차등 스플라이싱 패턴을 검출하는 단계;

뉴클레오타이드별로, 상기 참조 표적 하위서열과 상기 변이체 표적 하위서열의 스플라이스 부위 분류를 비교함으로써 상기 제1 차등 스플라이싱 패턴과 상기 제2 차등 스플라이싱 패턴 사이의 차이를 결정하는 단계; 및

상기 차이가 미리 결정된 임계값을 초과할 때, 상기 변이체를 비정상 스플라이싱을 유발하며 따라서 병원성인 것으로 분류하고 그 분류를 메모리에 저장하는 단계를 포함하는, 계놈 변이체를 검출하는 방법.

청구항 29

제28항에 있어서, 차등 스플라이싱 패턴은 표적 하위서열에서 스플라이싱 이벤트 발생의 위치 분포를 식별하는, 계놈 변이체를 검출하는 방법.

청구항 30

제28항 또는 제29항에 있어서, 상기 스플라이싱 이벤트는 크립틱 스플라이싱(cryptic splicing), 엑손 스킵핑(exon skipping), 상호 배타적인 엑손(mutually exclusive exon), 대체 공여체 부위, 대체 수용체 부위 및 인트론 보유(intron retention) 중 적어도 하나를 포함하는, 계놈 변이체를 검출하는 방법.

청구항 31

제28항 내지 제30항 중 어느 한 항에 있어서, 상기 참조 표적 하위서열과 상기 변이체 표적 하위서열은 뉴클레오타이드 위치에 대해서 정렬되고, 적어도 하나의 변이체 뉴클레오타이드만큼 상이한, 게놈 변이체를 검출하는 방법.

청구항 32

제28항 내지 제31항 중 어느 한 항에 있어서, 상기 참조 표적 하위서열 및 상기 변이체 표적 하위서열은 각각 적어도 40개의 뉴클레오타이드를 갖고 그리고 각각 각 측면에 적어도 40개의 뉴클레오타이드가 축적되는, 게놈 변이체를 검출하는 방법.

청구항 33

제28항 내지 제32항 중 어느 한 항에 있어서, 상기 참조 표적 하위서열 및 상기 변이체 표적 하위서열은 각각 적어도 101개의 뉴클레오타이드를 갖고 그리고 각각 각 측면에 적어도 5000개의 뉴클레오타이드가 축적되는, 게놈 변이체를 검출하는 방법.

청구항 34

제28항 내지 제33항 중 어느 한 항에 있어서, 상기 변이체 표적 하위서열은 2개의 변이체를 포함하는, 게놈 변이체를 검출하는 방법.

청구항 35

비정상 스플라이싱을 검출하는 방법으로서,

트레이닝된 아트러스 컨볼루션 신경망(약칭 ACNN)을 사용하여 입력 서열에서의 표적 뉴클레오타이드를 분류하고, 상기 표적 뉴클레오타이드의 각각이 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 플라이스 부위 점수를 할당하는 단계로서, 상기 입력 서열은 적어도 801개의 뉴클레오타이드를 포함하고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되는, 상기 할당하는 단계;

상기 ACNN을 통해 참조 서열 및 변이체 서열을 처리하여, 상기 참조 서열 및 상기 변이체 서열에서의 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 생성하는 단계로서, 상기 참조 서열 및 상기 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 갖고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되는, 상기 스플라이스 부위 점수를 생성하는 단계; 및

상기 참조 서열 및 상기 변이체 서열에서의 상기 표적 뉴클레오타이드의 상기 스플라이스 부위 점수의 차이로부터, 상기 변이체 서열을 생성한 변이체가 비정상 스플라이싱을 유발하고 따라서 병원성인지 여부를 결정하는 단계를 포함하는, 게놈 변이체를 검출하는 방법.

청구항 36

비정상 스플라이싱을 검출하는 컴퓨터 프로그램 명령어가 부여된 비일시적 컴퓨터 판독 가능 저장 매체로서, 상기 명령어는, 프로세서 상에서 실행될 경우,

트레이닝된 아트러스 컨볼루션 신경망(약칭 ACNN)을 사용하여 입력 서열에서의 표적 뉴클레오타이드를 분류하고, 상기 표적 뉴클레오타이드의 각각이 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 플라이스 부위 점수를 할당하는 단계로서, 상기 입력 서열은 적어도 801개의 뉴클레오타이드를 포함하고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되는, 상기 할당하는 단계;

상기 ACNN을 통해 참조 서열 및 변이체 서열을 처리하여, 상기 참조 서열 및 상기 변이체 서열에서의 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 생성하는 단계로서, 상기 참조 서열 및 상기 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 갖고 각각의 표적 뉴클레오타이드는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되는, 상기 스플라이스 부위 점수를 생성하는 단계; 및

상기 참조 서열 및 상기 변이체 서열에서의 상기 표적 뉴클레오타이드의 상기 스플라이스 부위 점수의 차이로부

터, 상기 변이체 서열을 생성한 변이체가 비정상 스플라이싱을 유발하고 따라서 병원성인지 여부를 결정하는 단계

를 포함하는 방법을 구현하는, 비일시적 컴퓨터 판독 가능 저장 매체.

청구항 37

비정상 스플라이싱을 검출하는 컴퓨터 프로그램 명령어가 부여된 비일시적 컴퓨터 판독 가능 저장 매체로서, 상기 명령어는, 프로세서 상에서 실행될 경우,

표적 하위서열에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 차등 스플라이싱 패턴을 검출하도록 트레이닝된 아트리스 컨볼루션 신경망(약칭 ACNN)을 통해 참조 서열을 처리하는 단계;

상기 처리에 기초하여, 참조 표적 하위서열에서의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 참조 표적 하위서열에서 제1 차등 스플라이싱 패턴을 검출하는 단계;

상기 ACNN을 통해 변이체 서열을 처리하는 단계로서, 상기 변이체 서열과 상기 참조 서열은 변이체 표적 하위서열에 위치한 하나 이상의 변이체 뉴클레오타이드만큼 상이한, 상기 변이체 서열을 처리하는 단계;

상기 처리에 기초하여, 상기 변이체 표적 하위서열의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 변이체 표적 하위서열에서 제2 차등 스플라이싱 패턴을 검출하는 단계;

뉴클레오타이드별로, 상기 참조 표적 하위서열과 상기 변이체 표적 하위서열의 스플라이스 부위 분류를 비교함으로써 상기 제1 차등 스플라이싱 패턴과 상기 제2 차등 스플라이싱 패턴 사이의 차이를 결정하는 단계; 및

상기 차이가 미리 결정된 임계값을 초과할 때, 상기 변이체를 비정상 스플라이싱을 유발하며 따라서 병원성인 것으로 분류하고 그 분류를 메모리에 저장하는 단계

를 포함하는 방법을 구현하는, 비일시적 컴퓨터 판독 가능 저장 매체.

청구항 38

메모리에 결합된 하나 이상의 프로세서를 포함하는 시스템으로서, 상기 메모리에는 비정상 스플라이싱을 검출하는 컴퓨터 명령어가 로딩되고, 상기 명령어는, 프로세서 상에서 실행될 경우,

입력 서열의 표적 하위서열에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 표적 하위서열에서 차등 스플라이싱 패턴을 검출하도록 트레이닝된 아트리스 컨볼루션 신경망(약칭 ACNN)을 통해 참조 서열을 처리하는 단계;

상기 처리에 기초하여, 참조 표적 하위서열에서의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 참조 표적 하위서열에서 제1 차등 스플라이싱 패턴을 검출하는 단계;

상기 ACNN을 통해 변이체 서열을 처리하는 단계로서, 상기 변이체 서열과 상기 참조 서열은 변이체 표적 하위서열에 위치한 하나 이상의 변이체 뉴클레오타이드만큼 상이한, 상기 변이체 서열을 처리하는 단계;

상기 처리에 기초하여, 상기 변이체 표적 하위서열의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 상기 변이체 표적 하위서열에서 제2 차등 스플라이싱 패턴을 검출하는 단계;

뉴클레오타이드별로, 상기 참조 표적 하위서열과 상기 변이체 표적 하위서열의 스플라이스 부위 분류를 비교함으로써 상기 제1 차등 스플라이싱 패턴과 상기 제2 차등 스플라이싱 패턴 사이의 차이를 결정하는 단계; 및

상기 차이가 미리 결정된 임계값을 초과할 때, 상기 변이체를 비정상 스플라이싱을 유발하며 따라서 병원성인 것으로 분류하고 그 분류를 메모리에 저장하는 단계

를 포함하는 동작을 구현하는, 시스템.

발명의 설명

기술 분야

부록

[0001]

[0002]

부록에는, 본 발명자들이 작성한 논문에 열거된 잠재적으로 관련된 참고문헌들의 목록이 포함되어 있다. 그 논문의 주제는, 본 출원이 우선권/이익을 주장하는 미국 가특허 출원에서 다루어진다. 이들 참고문헌은 요청 시 대리인에 의해 제공될 수 있거나 글로벌 도시에(Global Dossier)를 통해 액세스될 수 있다.

[0003]

우선권 출원

[0004]

본 출원은, 미국 가특허 출원 제62/573,125호(대리인 정리번호 ILLM 1001-1/IP-1610-PRV)(출원일: 2017년 10월 16일, 발명의 명칭: "Deep Learning-Based Splice Site Classification", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae); 미국 가특허 출원 제62/573,131호(대리인 정리번호 ILLM 1001-2/IP-1614-PRV)(출원일: 2017년 10월 16일, 발명의 명칭: "Deep Learning-Based Aberrant Splicing Detection", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae); 미국 가특허 출원 제62/573,135호(대리인 정리번호 ILLM 1001-3/IP1615-PRV)(출원일: 2017년 10월 16일, 발명의 명칭: "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae); 및 미국 가특허 출원 제62/726,158호(대리인 정리번호 ILLM 1001-10/IP-1749-PRV)(출원일: 2018년 8월 31일, 발명의 명칭: "Predicting Splicing from Primary Sequence with Deep Learning", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae)의 우선권 또는 이점을 주장한다. 상기 가특허 출원은 모든 면에서 본 명세서에 참고로 인용된다.

[0005]

원용 문헌

[0006]

이하의 것들은, 본 명세서에 그 전체가 기재된 것처럼 모든 면에서 참고로 인용되는 것이다:

[0007]

2018년 10월 15일자로 출원된 PCT 특허출원번호 PCT/US18/55915(대리인 정리번호 ILLM 1001-7/IP-1610-PCT)(발명의 명칭: "Deep Learning-Based Splice Site Classification", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae), 후속하여 PCT 공보 WO____로서 공개됨.

[0008]

2018년 10월 15일자로 동시에 출원된 PCT 특허출원번호 PCT/US18/____(대리인 정리번호 ILLM 1001-9/IP-1615-PCT)(발명의 명칭: "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae), 후속하여 PCT 공보 WO____로서 공개됨.

[0009]

동시에 출원된 미국 정규출원(대리인 정리번호 ILLM 1001-4/IP-1610-US)(발명의 명칭: "Deep Learning-Based Splice Site Classification", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae).

[0010]

동시에 출원된 미국 정규출원(대리인 정리번호 ILLM 1001-5/IP-1614-US)(발명의 명칭: "Deep Learning-Based Aberrant Splicing Detection", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae).

[0011]

동시에 출원된 미국 정규출원(대리인 정리번호 ILLM 1001-6/IP-1615-US)(발명의 명칭: "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)", 발명자: Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou 및 Jeremy Francis McRae).

[0012]

문헌 1 - S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO," arXiv:1609.03499, 2016;

문헌 2 - S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J.

Raiman, S. Sengupta and M. Shoeybi, "DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH,"

arXiv:1702.07825, 2017;

[0013]

- [0014] 문헌 3 - F. Yu and V. Koltun, "MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS," arXiv:1511.07122, 2016;
- [0015] 문헌 4 - K. He, X. Zhang, S. Ren, and J. Sun, "DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION," arXiv:1512.03385, 2015;
- [0016] 문헌 5 - R.K. Srivastava, K. Greff, and J. Schmidhuber, "HIGHWAY NETWORKS," arXiv: 1505.00387, 2015;
- [0017] 문헌 6 - G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "DENSELY CONNECTED CONVOLUTIONAL NETWORKS," arXiv:1608.06993, 2017;
- [0018] 문헌 7 - C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "GOING DEEPER WITH CONVOLUTIONS," arXiv: 1409.4842, 2014;
- [0019] 문헌 8 - S. Ioffe and C. Szegedy, "BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT," arXiv: 1502.03167, 2015;
- [0020] 문헌 9 - J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE," arXiv:1704.03669, 2017;
- [0021] 문헌 10 - L. C. Piqueras, "AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION," Tampere University of Technology, 2016;
- [0022] 문헌 11 - J. Wu, "Introduction to Convolutional Neural Networks," Nanjing University, 2017;
- [0023] 문헌 12 - I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "CONVOLUTIONAL NETWORKS", Deep Learning, MIT Press, 2016; 및
- [0024] 문헌 13 - J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS," arXiv:1512.07108, 2017.
- [0025] 문헌 1은, 동일한 컨볼루션 윈도우 크기를 갖는 컨볼루션 필터와 함께 잔여 블록의 그룹, 일괄 정규화층(batch normalization layer), 정류 선형 유닛(rectified linear unit: ReLU) 층, 차원 변경층(dimensionality altering layer), 지수적으로 성장하는 아트러스 컨볼루션 레이트(atrous convolution rate)를 갖는 아트러스 컨볼루션층, 스킵 연결, 및 입력 서열(input sequence)을 수용하고 입력 서열의 엔트리를 점수 매기는 출력 서열을 생성하도록 소프트맥스 분류층(softmax classification layer)을 이용하는 심층 컨볼루션 신경망 아키텍처를 기술한다. 개시된 기술은 문헌 1에 기술된 신경망 구성요소 및 파라미터를 사용한다. 일 구현예에서, 개시된 기술은 문헌 1에 기술된 신경망 구성요소의 파라미터를 수정한다. 예를 들어, 문헌 1과는 달리, 개시된 기술에서의 아트러스 컨볼루션 레이트는 낮은 잔여 블록 그룹으로부터 높은 잔여 블록 그룹으로 비지수적으로 진행된다. 다른 일례로, 문헌 1과는 달리, 개시된 기술에서의 컨볼루션 윈도우 크기는 잔여 블록의 그룹 간에 가변된다.
- [0026] 문헌 2는 문헌 1에 기술된 심층 컨볼루션 신경망 아키텍처의 세부사항을 기술한다.
- [0027] 문헌 3은 개시된 기술에 의해 사용되는 아트러스 컨볼루션을 기술한다. 본 명세서에서 사용되는 바와 같이, 아트러스 컨볼루션은 "팽창 컨볼루션"(dilated convolutions)이라고도 한다. 아트러스/팽창 컨볼루션은 트레이닝 가능한 파라미터가 거의 없는 큰 수용장을 허용한다. 아트러스/팽창 컨볼루션은, 아트러스 컨볼루션 레이트 또는 팽창 인자라고도 하는 소정의 단차로 입력값들을 스킵함으로써 커널이 자신의 길이보다 큰 면적에 걸쳐 적용되는 컨볼루션이다. 아트러스/팽창 컨볼루션은, 컨볼루션 동산이 수행될 때 넓은 간격으로 이웃하는 입력 엔트리(예를 들어, 뉴클레오타이드, 아미노산)가 고려되도록 컨볼루션 필터/커널의 요소들 사이에 간격을 추가한다. 이는 입력에 장거리 컨텍스트 종속성을 통합할 수 있게 한다. 아트러스 컨볼루션은, 인접한 뉴클레오타이드가 처리될 때 재사용을 위해 부분 컨볼루션 계산을 보존한다.
- [0028] 문헌 4는 개시된 기술에 의해 사용되는 잔여 블록 및 잔여 연결을 기술한다.
- [0029] 문헌 5는 개시된 기술에 의해 사용되는 스킵 연결을 기술한다. 본 명세서에서 사용되는 바와 같이, 스킵 연결은 "고속도로 네트워크"라고도 한다.
- [0030] 문헌 6은 개시된 기술에 의해 사용되는 조밀하게 연결된 컨볼루션 망 아키텍처를 기술한다.

[0031] 문헌 7은 개시된 기술에 의해 사용되는 차원 변경 컨볼루션층 및 모듈 기반 처리 파이프라인을 기술한다. 차원 변경 컨볼루션의 일례는 1×1 컨볼루션이다.

[0032] 문헌 8은 개시된 기술에 의해 사용되는 일괄 정규화층을 기술한다.

[0033] 문헌 9도 개시된 기술에 의해 사용되는 아트러스/팽창 컨볼루션을 기술한다.

[0034] 문헌 10은, 컨볼루션 신경망, 심층 컨볼루션 신경망, 및 아트러스/팽창 컨볼루션을 갖는 심층 컨볼루션 신경망을 포함하여, 개시된 기술에 의해 사용될 수 있는 심층 신경망의 다양한 아키텍처를 기술한다.

[0035] 문헌 11은, 서브샘플링층(예를 들어, 풀링(pooling)) 및 완전히 연결된 층을 갖는 컨볼루션 신경망을 트레이닝하기 위한 알고리즘을 포함하여, 개시된 기술에 의해 사용될 수 있는 컨볼루션 신경망의 세부사항을 기술한다.

[0036] 문헌 12는 개시된 기술에 의해 사용될 수 있는 다양한 컨볼루션 동작의 세부사항을 기술한다.

[0037] 문헌 13은 개시된 기술에 의해 사용될 수 있는 컨볼루션 신경망의 다양한 아키텍처를 기술한다.

[0038] 출원 시 전자적으로 함께 제출된 참고 표의 원용

[0039] ASCII 텍스트 포맷으로 되어 있는 이하의 표 파일들은 본 출원과 함께 제출되며 참고로 원용되는 것이다. 파일들의 명칭, 작성일 및 크기는 아래와 같다:

[0040] table_S4_mutation_rates.txt 2018년 8월 31일 2,452 KB

[0041] table_S5_gene_enrichment.txt 2018년 8월 31일 362 KB

[0042] table_S6_validation.txt 2018년 8월 31일 362 KB

[0043] 개시된 기술은, 불확실성이 있는 추론을 위한 시스템(예를 들어, 퍼지 논리 시스템), 적응형 시스템, 기계 학습 시스템, 및 인공 신경망을 포함하여, 인공 지능형 컴퓨터 및 디지털 데이터 처리 시스템 및 대응하는 데이터 처리 방법 및 지능 애플리케이션을 위한 제품(즉, 지식 기반 시스템, 추론 시스템, 및 지식 획득 시스템)에 관한 것이다. 특히, 개시된 기술은, 심층 컨볼루션 신경망을 트레이닝하기 위한 심층 학습 기반 기술의 사용에 관한 것이다.

배경 기술

[0044] 이 부문에서 개시되는 주제는, 단지 이 부문에서의 언급의 결과로서 종래 기술인 것으로 가정되어서는 안 된다. 유사하게, 이 부문에서 언급되거나 배경으로서 제공된 주제에 연관된 문제는 종래 기술에서 이전에 인식된 것으로 가정되어서는 안 된다. 이 부문의 주제는 단지 다른 방안을 나타내는 것이며, 이러한 방안은 그 자체가 청구된 기술의 구현에 또한 대응할 수 있는 것이다.

기계 학습

[0046] 기계 학습에서, 입력 변수는 출력 변수를 예측하는 데 사용된다. 입력 변수는, 종종 피처(feature)라고 하며, $X = (X_1, X_2, \dots, X_k)$ 로 표현되며, 여기서 각 $X_i, i \in 1, \dots, k$ 가 피처이다. 출력 변수는, 종종 응답 또는 종속 변수라고 칭하며, Y 로 표현된다. Y 와 대응 X 간의 관계는 다음과 같이 일반적으로 형태로 표현될 수 있다.

[0047]
$$Y = f(X) + \epsilon$$

[0048] 위 수학적식에서, f 는 피처 (X_1, X_2, \dots, X_k) 의 함수이고, ϵ 는 랜덤 에러 항이다. 에러 항은 X 와는 독립적이며 제로인 평균값을 갖는다.

[0049] 실제로, 피처 X 는, Y 를 갖지 않고서 또는 X 와 Y 간의 정확한 관계를 몰라도 이용 가능하다. 에러 항은 제로인 평균값을 가지므로, 목적은 f 를 추정하는 것이다.

[0050]
$$\hat{Y} = \hat{f}(X)$$

[0051] 위 수학적식에서, \hat{f} 는 ϵ 의 추정이고, 이는, 종종 블랙 박스라고 간주되며, f 의 출력과 입력 간의 관계만이 알려져 있지만 왜 그렇게 기능하는지에 대한 답은 없음을 의미한다.

[0052] 함수 f 는 학습을 이용하여 발견된다. 감독 학습과 비감독 학습은 이 작업을 위한 기계 학습에 사용되는 두 가지 방법이다. 감독 학습에서는, 지표 데이터가 트레이닝에 사용된다. 입력과 대응 출력(=표지)을 표시함으로써, 함수 f 는 출력에 근접하도록 최적화된다. 비감독 학습에서는, 목적이 지표 없는 데이터로부터 숨겨진 구조를 찾는 것이다. 이 알고리즘은 입력 데이터의 정확도를 측정하지 않으므로, 감독 학습과 구별된다.

[0053] **신경망**

[0054] 단층 퍼셉트론(single layer perceptron: SLP)은 신경망의 가장 단순한 모델이다. 단층 퍼셉트론은 도 1에 도시된 바와 같이 하나의 입력층과 하나의 활성화 함수를 포함한다. 입력들은 가중 그래프를 통과한다. 함수 f 는 입력들의 합을 인수로 사용하고 이를 임계값 θ 와 비교한다.

[0055] 도 2는 다수의 층을 갖는 완전히 연결된 신경망의 일 구현예를 도시한다. 신경망은, 서로 메시지를 교환하는 상호 연결된 인공 뉴런(예를 들어, a_1, a_2, a_3)의 시스템이다. 예시된 신경망은, 3개의 입력, 숨겨진 층에서의 2개의 뉴런, 및 출력층에서의 2개의 뉴런을 갖는다. 숨겨진 층은 활성화 함수 $f(\bullet)$ 를 갖고, 출력층은 활성화 함수 $g(\bullet)$ 를 갖는다. 연결에는 트레이닝 프로세스 동안 조정되는 숫자 가중치(예를 들어, $w_{11}, w_{21}, w_{12}, w_{31}, w_{22}, w_{32}, v_{11}, v_{22}$)가 있으므로, 인식할 이미지를 공급할 때 올바르게 트레이닝된 네트워크가 올바르게 응답한다. 입력층은 원시 입력을 처리하고, 숨겨진 층은, 입력층과 숨겨진 층 간의 연결의 가중치에 기초하여 입력층으로부터의 출력을 처리한다. 출력층은, 숨겨진 층으로부터 출력을 가져 와서 숨겨진 층과 출력층 간의 연결의 가중치에 기초하여 처리한다. 망은 피쳐 검출 뉴런의 다수의 층을 포함한다. 각 층은, 이전 층으로부터의 입력들의 상이한 조합에 응답하는 많은 뉴런을 갖는다. 이들 층은, 제1 층이 입력 화상 데이터에서 프리미티브 패턴들의 세트를 검출하고 제2 층이 패턴들 중 패턴을 검출하고 제3 층이 그러한 패턴들 중 패턴을 검출하도록 구성된다.

[0056] 유전체학에서의 심층 학습의 적용에 대한 조사는 이하의 간행물에서 찾을 수 있다:

[0057] • T. Ching et al., Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org:142760, 2017;

[0058] • Angermueller C, P rnamaa T, Parts L, Stegle O. Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878;

[0059] • Park Y, Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825-826. (doi:10.1038/nbt.3313);

[0060] • Min, S., Lee, B. & Yoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016);

[0061] • Leung MK, Delong A, Alipanahi B et al. Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016; 및

[0062] • Libbrecht MW, Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32.

도면의 간단한 설명

[0063] 도면에서, 유사한 참조 문자는 일반적으로 상이한 도면 전체에 걸쳐 유사한 부분을 지칭한다. 또한, 도면은, 반드시 축척대로 도시된 것은 아니며, 대신 개시된 기술의 원리를 설명하도록 일반적으로 강조된 것이다. 이하의 설명에서는, 개시된 기술의 다양한 구현예를 이하의 도면을 참조하여 설명한다.

도 1은 단층 퍼셉트론(SLP)을 도시한다.

도 2는 다수의 층을 갖는 피드포워드 신경망(feed-forward neural network)의 일 구현예를 도시한다.

도 3은 컨볼루션 신경망의 동작의 일 구현예를 도시한다.

도 4는 개시된 기술의 일 구현예에 따라 컨볼루션 신경망을 트레이닝하는 블록도를 도시한다.

- 도 5는 개시된 기술의 일 구현예에 따라 ReLU 비선형 층의 일 구현예를 도시한다.
- 도 6은 팽창 컨볼루션을 도시한다.
- 도 7은 개시된 기술의 일 구현예에 따라 서브샘플링층(평균/최대 풀링)의 일 구현예이다.
- 도 8은 컨볼루션층의 2-층 컨볼루션의 일 구현예를 도시한다.
- 도 9는 피쳐 맵 추가를 통해 하류측으로 이전 정보를 재주입하는 잔여 연결을 도시한다.
- 도 10은 잔여 블록과 스킵 연결의 일 구현예를 도시한다.
- 도 11은 적층된 팽창 컨볼루션의 일 구현예를 도시한다.
- 도 12는 일괄 정규화 순방향 패스(batch normalization forward pass)를 도시한다.
- 도 13은 테스트 시간에서의 일괄 정규화 변환을 도시한다.
- 도 14는 일괄 정규화 역방향 패스를 도시한다.
- 도 15는 컨볼루션 또는 밀집 연결층(densely connected layer)과 함께 일괄 정규화층을 사용하는 것을 도시한다.
- 도 16은 1D 컨볼루션의 일 구현예를 도시한다.
- 도 17은 글로벌 평균 풀링(global average pooling: GAP)의 동작 방식을 도시한다.
- 도 18은 개시된 기술을 구현하는 데 사용될 수 있는 트레이닝 서버들과 생성 서버들을 구비한 연산 환경의 일 구현예를 도시한다.
- 도 19는, 본 명세서에서 "SpliceNet"이라고 하는, 아트러스 컨볼루션 신경망(atrous convolutional neural network: ACNN)의 아키텍처의 일 구현예를 도시한다.
- 도 20은 ACNN과 컨볼루션 신경망(convolutional neural network: CNN)에 의해 사용될 수 있는 잔여 블록의 일 구현예를 도시한다.
- 도 21은, 본 명세서에서 "SpliceNet80"이라고 하는, ACNN의 아키텍처의 다른 일 구현예를 도시한다.
- 도 22는, 본 명세서에서 "SpliceNet400"이라고 하는, ACNN의 아키텍처의 또 다른 일 구현예를 도시한다.
- 도 23은, 본 명세서에서 "SpliceNet2000"이라고 하는, ACNN의 아키텍처의 또 다른 일 구현예를 도시한다.
- 도 24는, 본 명세서에서 "SpliceNet10000"이라고 하는, ACNN의 아키텍처의 또 다른 일 구현예를 도시한다.
- 도 25, 도 26 및 도 27은 ACNN 및 CNN에 의해 처리되는 다양한 유형의 입력들을 도시한다.
- 도 28은 ACNN이 적어도 8억개의 비-스플라이싱 부위(non-splicing site)에서 트레이닝될 수 있고, CNN은 적어도 1백만개의 비-스플라이싱 부위에서 트레이닝될 수 있음을 도시한다.
- 도 29는 원-핫 인코더(one-hot encoder)를 도시한다.
- 도 30은 ACNN의 트레이닝을 도시한다.
- 도 31은 CNN을 도시한다.
- 도 32는 ACNN 및 CNN의 트레이닝, 검증 및 테스트를 도시한다.
- 도 33은 참조 서열 및 대체 서열을 도시한다.
- 도 34는 비정상 스플라이싱 검출을 도시한다.
- 도 35는 스플라이스 부위 분류(splice site classification)를 위한 SpliceNet10000의 프로세싱 피라미드를 도시한다.
- 도 36은 비정상 스플라이싱 검출을 위한 SpliceNet10000의 프로세싱 피라미드를 도시한다.
- 도 37A, 도 37B, 도 37C, 도 37D, 도 37E, 도 37F, 도 37G 및 도 37H는 심층 학습에 의해 1차 서열(primary sequence)로부터 스플라이싱을 예측하는 것의 일 구현예를 도시한다.

도 38A, 도 38B, 도 38C, 도 38D, 도 38E, 도 38F 및 도 38G는 RNA-seq 데이터에서 희귀 크립틱 스플라이스 돌연변이의 검증의 일 구현예를 도시한다.

도 39A, 도 39B 및 도 39C는 조직특이적 대체 스플라이싱을 빈번하게 생성하는 크립틱 스플라이스 변이체들의 일 구현예를 도시한다.

도 40A, 도 40B, 도 40C, 도 40D 및 도 40E는 인간 개체군에서 크게 유해한 예측된 크립틱 스플라이스 변이체들의 일 구현예를 도시한다.

도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F는 희귀 유전질환을 가진 환자들에서 드 노보 크립틱 스플라이스 돌연변이(de novo cryptic splice mutation)의 일 구현예를 도시한다.

도 42A 및 도 42B는 lincRNAs 에 대한 다양한 스플라이싱 예측 알고리즘들의 평가를 도시한다.

도 43A 및 도 43B는 TACTAAC 분기점 및 GAAGAA 엑손-스플라이스 인헨서 모티프의 위치종속 효과를 도시한다.

도 44A 및 도 44B는 스플라이싱에 대한 뉴클레오솜 포지셔닝의 효과를 도시한다.

도 45는 복합 효과를 가지는 스플라이스-과피 변이체에 대한 효과 크기 계산의 예시를 도시한다.

도 46A, 도 46B 및 도 46C는 싱글톤(singleton)과 공통 변이체(common variant)에 대한 SpliceNet-10k 모델의 평가를 도시한다.

도 47A 및 도 47B는 변이체의 위치에 의해 분할된, 스플라이스 부위가 생성하는 변이체의 유효성확인율과 효과 크기를 도시한다.

도 48A, 도 48B, 도 48C 및 도 48D는 트레이닝 및 테스트 염색체에 대한 SpliceNet-10k 모델의 평가를 도시한다.

도 49A, 도 49B 및 도 49C는 동의, 인트론 또는 비번역 영역 부위들만으로부터, 희귀 유전질환을 가지는 환자에서 드 노보 크립틱 스플라이스 돌연변이를 도시한다.

도 50A 및 도 50B는 ASD에서 크립틱 스플라이스 드 노보 돌연변이를 도시하고 또한 크립틱 스플라이스 드 노보 돌연변이를 병원성 DNMs의 일 부분으로 도시한다.

도 51a, 도 51b, 도 51c, 도 51d, 도 51e, 도 51f, 도 51g, 도 51h, 도 51i 및 도 51j는 ASD 환자들에게서 예측된 크립틱 스플라이스 드 노보 돌연변이의 RNA-seq 유효성확인을 도시한다.

도 52A 및 도 52B는 표준 전사체에 대해서만 트레이닝된 모델의 RNA-seq에 대한 유효성확인율과 민감도를 도시한다.

도 53A, 도 53B 및 도 53C는 SpliceNet-10k 성능을 개선시키는 앙상블 모델링을 도시한다.

도 54A 및 도 54B는 가변 엑손 밀도 영역에서 SpliceNet-10k의 평가를 도시한다.

도 55는 효과 크기 계산과 조직특이적 스플라이싱을 입증하는 데 사용되는 GTEx 샘플들의 일 구현예를 도시하는 표 S1이다.

도 56은 다른 알고리즘들의 유효성확인율과 민감도를 평가하는 데 사용되는 컷오프의 일 구현예를 도시하는 표 S2이다.

도 57은 유전자당 농축 분석(per-gene enrichment analysis)의 일 구현예를 도시한다.

도 58은 게놈 전체 농축 분석(genome-wide enrichment analysis)의 일 구현예를 도시한다.

도 59는 개시된 기술을 구현하는 데 사용될 수 있는 컴퓨터 시스템의 단순화된 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0064] 이하의 설명은, 통상의 기술자가 개시된 기술을 제조 및 사용할 수 있도록 제시된 것이며, 특정 응용분야 및 그 요건과 관련하여 제공된 것이다. 개시된 구현예에 대한 다양한 변형은 통상의 기술자에게 명백할 것이며, 본 명세서에서 정의된 일반적인 원리는 개시된 기술의 사상 및 범위를 벗어나지 않고 다른 구현예와 응용분야에 적용될 수 있다. 따라서, 개시된 기술은, 도시된 구현예들로 제한되도록 의도된 것이 아니라, 본 명세서에 개시된

원리 및 특징과 일치하는 가장 넓은 범위를 따른 것이다.

[0065] **도입부**

[0066] **컨볼루션 신경망**

[0067] 컨볼루션 신경망은 특수한 유형의 신경망이다. 조밀하게 연결된 층과 컨볼루션층 간의 근본적인 차이점은 다음과 같은데, 즉, 조밀하게 연결된 층은 입력 피쳐 공간에서 글로벌 패턴을 학습하는 반면, 컨볼루션층은 로컬 패턴을 학습하며, 이 경우, 입력의 작은 2D 윈도우에서 발견되는 패턴을 학습한다. 이러한 핵심 특성은 컨볼루션 신경망에 두 개의 흥미로운 특성을 제공하는데, 즉, (1) 컨볼루션층이 학습하는 패턴은 불변 번역이고, (2) 패턴의 공간 계층을 학습할 수 있다는 점이다.

[0068] 첫 번째와 관련하여, 컨볼루션층은, 화상의 우측 하단 코너에서 소정의 패턴을 학습한 후, 임의의 위치, 예를 들어, 좌측 상단 코너에서 그 패턴을 인식할 수 있다. 조밀하게 연결된 망은, 새로운 패턴이 새로운 위치에 나타나면 그 새로운 패턴을 학습해야 한다. 따라서, 이는, 일반화 능력이 있는 표현을 학습하도록 더 적은 트레이닝 샘플을 필요로 하기 때문에, 컨볼루션 신경망 데이터를 효율적으로 되게 한다.

[0069] 두 번째와 관련하여, 제1 컨볼루션층은 예지와 같은 작은 국소 패턴을 학습할 수 있고, 제2 컨볼루션층은 제1 컨볼루션층의 피쳐로 이루어진 큰 패턴 등을 학습한다. 이를 통해 컨볼루션 신경망이 점점 더 복잡해지고 추상적인 시각적 개념을 효율적으로 학습할 수 있다.

[0070] 컨볼루션 신경망은, 다른 많은 층에 배치된 인공 뉴런 층들을 그 층들을 증속시키는 활성화 함수와 상호 연결함으로써 고도의 비선형 맵핑을 학습한다. 이것은, 하나 이상의 서브샘플링층과 비선형 층이 산재된 하나 이상의 컨볼루션층을 포함하며, 이들 층에는 통상적으로 하나 이상의 완전히 연결된 층이 뒤따른다. 컨볼루션 신경망의 각 요소는 이전 층의 피쳐들의 세트로부터 입력을 수신한다. 컨볼루션 신경망은, 동일한 피쳐 맵의 뉴런이 동일한 가중치를 가질 수 있기 때문에 동시에 학습한다. 이러한 국소 공유 가중치는, 다차원 입력 데이터가 컨볼루션 신경망에 진입할 때 컨볼루션 신경망이 피쳐 추출 및 회귀 또는 분류 프로세스에서 데이터 재구성의 복잡성을 피하도록 신경망의 복잡성을 감소시킨다.

[0071] 컨볼루션은, 2개의 공간 축(높이 및 폭)과 깊이 축(채널 축이라고도 함)을 갖는 피쳐 맵이라고 하는 3D 텐서에서 동작한다. RGB 이미지의 경우, 이미지가 3개의 색상인 적색, 녹색, 청색 채널을 갖기 때문에, 깊이 축의 치수가 3이다. 흑백 사진의 경우, 깊이는 1(회색 수준)이다. 컨볼루션 동작은, 자신의 입력 피쳐 맵으로부터 패치를 추출하고 이러한 패치 모두에 동일한 변환을 적용하여, 출력 피쳐 맵을 생성한다. 이러한 출력 피쳐 맵은, 여전히 3D 텐서이며, 폭과 높이를 갖는다. 그 출력 깊이는 임의적일 수 있는데, 그 이유는 출력 깊이가 층의 파라미터이고, 해당 깊이 축의 상이한 채널들이 더 이상 RGB 입력에서와 같이 특정 색상을 나타내지 않고 오히려 필터를 나타내기 때문이다. 필터는 입력 데이터의 특정 양태를 인코딩하며, 예를 들어, 높이 수준에서, 단일 필터는 "입력에 얼굴이 존재함"이라는 개념을 인코딩할 수 있다.

[0072] 예를 들어, 제1 컨볼루션층은, 크기(28, 28, 1)의 피쳐 맵을 취하고 크기(26, 26, 32)의 피쳐 맵을 출력하며, 자신의 입력에 대해 32개의 필터를 연산한다. 이러한 32개의 출력 채널의 각각은 26×26 그리드의 값을 포함하며, 이것은 입력에 대한 필터의 응답 맵이며, 입력의 상이한 위치에서의 해당 필터 패턴의 응답을 나타낸다. 이것이 피쳐 맵이라는 용어의 의미이며, 깊이 축의 모든 치수는 피쳐(또는 필터)이며, 2D 텐서 출력(:, :, n)은 입력에 대한 이러한 필터의 응답의 2D 공간 맵이다.

[0073] 컨볼루션은, 두 개의 주요 파라미터에 의해 정의되는데, 즉, (1) 입력으로부터 추출된 패치의 크기 - 이들은 통상적으로 1×1, 3×3, 또는 5×5이고, (2) 출력 피쳐 맵의 깊이 - 필터의 수는 컨볼루션에 의해 연산된다. 종종, 이들 컨볼루션은, 깊이 32에서 시작하여, 깊이 64로 계속되며, 깊이 128 또는 256으로 종료된다.

[0074] 컨볼루션은, 3D 입력 피쳐 맵 위로 3×3 또는 5×5 크기의 이들 윈도우를 슬라이딩하고, 모든 위치에서 정지하고, 주변 피쳐의 3D 패치(형상(window_height, window_width, input_depth))를 추출함으로써 동작한다. 이어서, 이러한 각 3D 패치는, (컨볼루션 커널이라고 하는 동일한 학습 가중치 행렬을 갖는 텐서 곱을 통해) 형상(output_depth)의 1D 벡터로 변환된다. 이어서, 이러한 벡터는 모두 형상(높이, 폭, output_depth)의 3D 출력 맵으로 공간적으로 재조립된다. 출력 피쳐 맵의 모든 공간 위치는 입력 피쳐 맵의 동일한 위치에 대응한다 (예를 들어, 출력의 우측 하단 코너는 입력의 우측 하단 코너에 대한 정보를 포함한다). 예를 들어, 3×3 윈도우의 경우, 벡터 출력([i, j, :])은 3D 패치 입력([i-1: i+1, j-1: J+1, :])으로부터 온 것이다. 전체 프로세스는 도 3에 상세히 설명되어 있다.

[0075] 컨볼루션 신경망은, 트레이닝 동안 많은 그라디언트 업데이트 반복에 걸쳐 학습되는 컨볼루션 필터(가중치 행렬)와 입력값 간의 컨볼루션 동작을 수행하는 컨볼루션층을 포함한다. (m, n)을 필터 크기라고 하고 W를 가중치 행렬이라고 설정하면, 컨볼루션 층은, 내적 를 계산함으로써 입력 X와 W의 컨볼루션을 수행하며, 여기서, x는 X의 인스턴스이고, b는 편향이다. 컨볼루션 필터가 입력을 가로질러 슬라이딩하는 단차 크기를 보폭이라고 하며, 필터 면적(m×n)을 수용장(receptive field)이라고 한다. 동일한 컨볼루션 필터가 입력의 상이한 위치에 걸쳐 적용되며, 이는 학습되는 가중치의 수가 감소시킨다. 이것은, 또한, 위치 불변 학습을 가능하게 하며, 즉, 중요한 패턴이 입력에 존재하는 경우, 컨볼루션 필터는 시퀀스의 위치에 관계없이 그 패턴을 학습한다.

[0076] **컨볼루션 신경망의 트레이닝**

[0077] 도 4는 개시된 기술의 일 구현예에 따라 컨볼루션 신경망을 트레이닝하는 블록도를 도시한다. 컨볼루션 신경망은, 입력 데이터가 특정 출력 추정값으로 이어지도록 조정되거나 트레이닝된다. 컨볼루션 신경망은, 출력 추정값이 실측 자료(ground truth)에 점진적으로 일치하거나 근접할 때까지 출력 추정값과 실측 자료 간의 비교에 기초하여 역전파(backpropagation)를 이용하여 조정된다.

[0078] 컨볼루션 신경망은, 실측 자료와 실제 출력 간의 차이에 기초하는 뉴런들 간의 가중치를 조정함으로써 트레이닝된다. 이것은 수학적으로 다음과 같이 설명된다:

[0079]
$$\Delta w_i = x_i \delta$$

[0080] 여기서 $\delta = (\text{실측 자료}) - (\text{실제 출력})$

[0081] 일 구현예에서, 트레이닝 규칙은 다음과 같이 정의된다:

[0082]
$$w_{nm} \leftarrow w_{nm} + \alpha(t_m - \varphi_m) a_n$$

[0083] 위 수식에서, 화살표는 값의 업데이트를 나타내고, t_m 은 뉴런 M의 목표 값이고, φ_m 은 뉴런 m의 연산된 현재 출력이고, a_n 은 입력 n이고, α 는 학습률이다.

[0084] 트레이닝의 중간 단계는, 컨볼루션층을 사용하여 입력 데이터로부터 피쳐 벡터를 생성하는 단계를 포함한다. 출력에서 시작하여 각 층의 가중치에 대한 그라디언트를 계산한다. 이것을 역방향 패스 또는 후진이라고 한다. 네거티브 그라디언트와 이전 가중치의 조합을 사용하여 망의 가중치를 업데이트한다.

[0085] 일 구현예에서, 컨볼루션 신경망은, 그라디언트 하강에 의해 에러의 역전파를 수행하는 (ADAM과 같은) 확률적 그라디언트 업데이트 알고리즘을 사용한다. 시그모이드 함수 기반 역전파 알고리즘의 일례가 아래에 설명되어 있다:

[0086]
$$\varphi = f(h) = \frac{1}{1 + e^{-h}}$$

[0087] 위 시그모이드 함수에서, h는 뉴런에 연산된 가중 합이다. 시그모이드 함수는 이하의 도함수를 갖는다:

[0088]
$$\frac{\partial \varphi}{\partial h} = \varphi(1 - \varphi)$$

[0089] 알고리즘은, 망의 모든 뉴런의 활성화를 연산하여, 순방향 패스를 위한 출력을 생성하는 것을 포함한다. 숨겨진 층의 뉴런 m의 활성화는 다음과 같이 기술된다:

[0090]
$$\varphi_m = \frac{1}{1 + e^{-h_m}}$$

$$h_m = \sum_{n=1}^N a_n w_{nm}$$

[0091] 이것은 아래와 같이 기술되는 활성화를 얻도록 모든 숨겨진 층에 대하여 행해진다:

$$\varphi_k = \frac{1}{1 + e^{-h_k}}$$

[0092]

$$h_k = \sum_{m=1}^M \varphi_m v_{mk}$$

[0093] 이어서, 층당 에러와 보정된 가중치를 계산한다. 출력에서의 에러는 다음과 같이 연산된다:

[0094]

$$\delta_{ok} = (t_k - \varphi_k) \varphi_k (1 - \varphi_k)$$

[0095] 숨겨진 층의 가중치는 다음과 같이 계산된다:

[0096]

$$\delta_{hm} = \varphi_m (1 - \varphi_m) \sum_{k=1}^K v_{mk} \delta_{ok}$$

[0097] 출력층의 가중치는 다음과 같이 업데이트된다:

[0098]

$$v_{mk} \leftarrow v_{mk} + \alpha \delta_{ok} \varphi_m$$

[0099] 숨겨진 층의 가중치는 다음과 같이 학습률 α 를 사용하여 업데이트된다:

[0100]

$$v_{nm} \leftarrow v_{nm} + \alpha \delta_{hm} \varphi_n$$

[0101] 일 구현예에서, 컨볼루션 신경망은, 그라디언트 하강 최적화를 이용하여 모든 층에 걸쳐 에러를 연산한다. 이러한 최적화에 있어서, 입력 피쳐 벡터 x 와 예측 출력 \hat{y} 에 대하여, 손실 함수는, 표적이 y 인 경우, \hat{y} 를 예측하는 비용에 대하여 l 로서, 즉, $l(\hat{y}, y)$ 로서 정의된다. 예측 출력 \hat{y} 은, 함수 f 를 사용하여 입력 피쳐 벡터 x 로부터 변환된다. 함수 f 는 컨볼루션 신경망의 가중치에 의해 파라미터화되며, 즉, $\hat{y} = f_w(x)$ 이다. 손실 함수는, $l(\hat{y}, y) = l(f_w(x), y)$ 또는 $Q(z, w) = l(f_w(x), y)$ 로서 기술되며, 여기서, z 는 입력 및 출력 데이터 쌍 (x, y) 이다. 그라디언트 하강 최적화는 이하의 식에 따라 가중치를 업데이트함으로써 수행된다:

[0102]

$$v_{t+1} = \mu v_t - \alpha \frac{1}{n} \sum_{i=1}^n \nabla_{w_i} Q(z_i, w_i)$$

$$w_{t+1} = w_t + v_{t+1}$$

[0103] 위 수식에서, α 는 학습률이다. 또한, 손실은 데이터 쌍들의 세트에 대한 평균으로서 연산된다. 연산은, 선형 수렴시 학습률 α 가 충분히 작을 때 종료된다. 다른 구현예에서, 그라디언트는, 연산 효율을 주입하도록 네스테로브(Nesterov)의 가속 그라디언트 및 적응형 그라디언트에 공급되는 선택된 데이터 쌍만을 사용하여 계산된다.

[0104] 일 구현예에서, 컨볼루션 신경망은, 확률적 그라디언트 하강(stochastic gradient descent: SGD)을 이용하여 비용 함수를 계산한다. SGD는, 다음과 같이 기술되는 그라디언트를 하나의 랜덤화된 데이터 쌍 z_i 로부터만 연산함으로써, 손실 함수의 가중치에 대하여 그라디언트를 근사화한다:

[0105]

$$v_{t+1} = \mu v_t - \alpha \nabla_w Q(z_i, w_i)$$

$$w_{t+1} = w_t + v_{t+1}$$

[0106] 위 수학적식에서, α 는 학습률이고, μ 는 모멘텀이고, t 는 업데이트 전의 현재 가중 상태이다. SGD의 수렴 속도는, 학습률 α 가 빠르고 느린 경우 모두에 대하여 충분히 감소될 때 대략 $O(1/t)$ 이다. 다른 구현예에서, 컨볼루션 신경망은 유클리드 손실 및 소프트맥스 손실 등의 상이한 손실 함수를 사용한다. 추가 구현예에서는, 컨볼루션 신경망이 아담(Adam) 확률적 최적화기를 사용한다.

[0107] **컨볼루션층**

[0108] 컨볼루션 신경망의 컨볼루션층은 피쳐 추출기로서 기능한다. 컨볼루션층은, 입력 데이터를 학습하고 계층적 피쳐로 분해시킬 수 있는 적응형 피쳐 추출기로서 기능한다. 일 구현예에서, 컨볼루션층은, 2개의 이미지를 입력으로서 취하고 제3 이미지를 출력으로서 생성한다. 이러한 구현예에서, 컨볼루션은 2차원(2D)인 2개의 이미지로 동작하며, 이때 하나의 이미지는 입력 이미지이고 나머지 이미지는 "커널"이라고 하며 입력 이미지에 대한 필터로서 적용되어, 출력 이미지를 생성한다. 따라서, 길이 n 의 입력 벡터와 길이 m 의 커널 g 에 대해, f 와 g 의 컨볼루션 $f * g$ 는 다음과 같이 정의된다:

[0109]
$$(f * g)(i) = \sum_{j=1}^m g(j) \cdot f(i - j + m/2)$$

[0110] 컨볼루션 동작은 입력 이미지 위로 커널을 슬라이딩하는 것을 포함한다. 커널의 각 위치에 대해, 커널과 입력 이미지의 중첩 값들이 승산되고 결과가 가산된다. 곱들의 합은, 커널이 중심에 있는 입력 이미지의 지점에서의 출력 이미지의 값이다. 많은 커널로부터의 상이한 출력 결과를 피쳐 맵이라고 한다.

[0111] 일단 컨볼루션층이 트레이닝되면, 이러한 컨볼루션층은 새로운 추론 데이터에 대한 인식 작업을 수행하는 데 적용된다. 컨볼루션층은, 트레이닝 데이터로부터 학습하므로, 명시적 피쳐 추출을 피하고 트레이닝 데이터로부터 은연중에 학습한다. 컨볼루션층은, 트레이닝 프로세스의 일부로서 결정 및 업데이트되는 컨볼루션 필터 커널 가중치를 사용한다. 컨볼루션층은, 상위 층들에서 결합되는 입력의 상이한 피쳐들을 추출한다. 컨볼루션 신경망은 다양한 수의 컨볼루션층을 사용하며, 각 컨볼루션층은 커널 크기, 보폭, 패딩, 피쳐 맵의 수, 및 가중치 등의 상이한 컨볼빙 파라미터를 갖는다.

[0112] **비선형 층**

[0113] 도 5는, 개시된 기술의 일 구현예에 따른 비선형 층의 일 구현예를 도시한다. 비선형 층은, 상이한 비선형 트리거 기능을 사용하여 각 숨겨진 층에서의 가능한 피쳐의 명확한 식별을 시그널링한다. 비선형 층은, 정류 선형 유닛(ReLU), 쌍곡 탄젠트, 쌍곡 탄젠트의 절대, 시그모이드 및 연속 트리거(비선형) 함수를 포함하여 다양한 특정 기능을 사용하여 비선형 트리거링을 구현한다. 일 구현예에서, ReLU 활성화는 함수 $y = \max(x, 0)$ 를 구현하고 층의 입력 및 출력 크기를 동일하게 유지한다. ReLU 사용의 장점은, 컨볼루션 신경망이 여러 번 더욱 빠르게 트레이닝된다는 점이다. ReLU는, 입력값이 제로보다 크면 입력에 대해 선형이고 그렇지 않으면 제로인 비연속 비포화 활성화 함수이다. 수학적으로 ReLU 활성화 함수는 다음과 같이 기술된다:

[0114]
$$\varphi(h) = \max(h, 0)$$

$$\varphi(h) = \begin{cases} h & \text{if } h > 0 \\ 0 & \text{if } h \leq 0 \end{cases}$$

[0115] 다른 구현예에서, 컨볼루션 신경망은, 다음과 같이 기술되는 연속 비포화 함수인 파워 유닛 활성화 함수를 사용한다:

[0116]
$$\varphi(h) = (a + bh)^c$$

[0117] 위 수학적식에서, a , b , c 는 각각 시프트, 스케일, 및 파워를 제어하는 파라미터이다. 파워 활성화 함수는, c 가 홀수이면 x 와 y -비대칭 활성화를 생성할 수 있고 c 가 짝수이면 y -대칭 활성화를 생성할 수 있다. 일부 구현예에서, 유닛은 비정류 선형 활성화를 생성한다.

[0118] 또 다른 구현예에서, 컨볼루션 신경망은, 이하의 로직 함수에 의해 기술되는 연속 포화 함수인 시그모이드 유닛 활성화 함수를 사용한다:

$$\varphi(h) = \frac{1}{1 + e^{-\beta h}}$$

[0119]

[0120] 위 수학적식에서, $\beta = 1$ 이다. 시그모이드 유닛 활성화 함수는, 네거티브 활성화를 생성하지 않으며, y축에 대해서만 비대칭이다.

[0121] **팽창 컨볼루션**

[0122] 도 6은 팽창 컨볼루션을 도시한다. 팽창 컨볼루션은, 때때로 아트러스 컨볼루션이라고 하며, 글자 그대로 홀(hole)을 갖는 것을 의미한다. 프랑스 이름은 알고리즘 아 트러스에서 유래되었으며, 이것은 빠른 이분구간(dyadic) 웨이브렛 변환을 연산한다. 이러한 유형의 컨볼루션층에서, 필터의 수용장에 대응하는 입력은 이웃 지점이 아니다. 이것은 도 6에 도시되어 있다. 입력들 사이의 거리는 팽창 인자에 의존한다.

[0123] **서브샘플링층**

[0124] 도 7은 개시된 기술의 일 구현예에 따른 서브샘플링층의 일 구현예이다. 서브샘플링층은, 컨볼루션층에 의해 추출된 피처의 해상도를 감소시켜 추출된 피처 또는 피처 맵을 노이즈 및 왜곡에 대해 강력하게 만든다. 일 구현예에서, 서브샘플링층은 평균 풀링 및 최대 풀링인 두 가지 유형의 풀링 동작을 사용한다. 풀링 동작은 입력을 중복되지 않는 2차원 공간으로 나눈다. 평균 풀링의 경우, 영역에 있는 4개 값의 평균이 계산된다. 최대 풀링의 경우, 4개 값 중 최대값이 선택된다.

[0125] 일 구현예에서, 서브샘플링층은, 그 출력을 최대 풀링의 입력들 중 하나의 입력에만 맵핑하고 그 출력을 평균 풀링의 입력들의 평균에 맵핑함으로써 이전 층들의 뉴런들의 세트에 대한 풀링 동작을 포함한다. 최대 풀링에 있어서, 풀링 뉴런의 출력은 다음에 기술된 바와 같이 입력 내에 있는 최대값이다:

$$\varphi_o = \max(\varphi_1, \varphi_2, \dots, \varphi_N)$$

[0126]

[0127] 위 수학적식에서, N은 뉴런 세트 내의 요소들의 총 수이다.

[0128] 평균 풀링에 있어서, 풀링 뉴런의 출력은, 이하에서 기술되는 바와 같이 입력 뉴런 세트와 함께 상주하는 입력 값들의 평균값이다:

$$\varphi_o = \frac{1}{N} \sum_{n=1}^N \varphi_n$$

[0129]

[0130] 위 수학적식에서, N은 입력 뉴런 세트 내의 요소들의 총 수이다.

[0131] 도 7에서, 입력의 크기는 4×4이다. 2×2 서브샘플링에 대하여, 4×4 이미지가 2×2 크기의 4개의 비중복 행렬로 분할된다. 평균 풀링에 대하여, 4개 값의 평균은 완전한 정수 출력이다. 최대 풀링에 대하여, 2×2 행렬의 4개 값의 최대값은 완전한 정수 출력이다.

[0132] **컨볼루션 예**

[0133] 도 8은 컨볼루션층의 2-층 컨볼루션의 일 구현예를 도시한다. 도 8에서, 크기가 2048인 입력값이 컨볼루션된다. 컨볼루션 1에서, 입력은, 크기가 3×3인 16개의 커널의 2개의 채널로 구성된 컨볼루션층에 의해 컨볼루션된다. 이어서, 생성되는 16개의 피처 맵은, ReLU1에서 ReLU 활성화 기능에 의해 정류된 후 3×3 크기의 커널이 있는 16개의 채널 풀링층을 사용하여 평균 풀링에 의해 풀 1에서 풀링된다. 이어서, 컨볼루션 2에서, 풀 1의 출력은, 크기가 3×3인 30개의 커널의 16개 채널로 구성된 다른 컨볼루션층에 의해 컨볼루션된다. 이어서, 또 다른 ReLU2 및 커널 크기가 2×2인 풀 2의 평균 풀링이 이어진다. 컨볼루션층은, 다양한 수의 보폭과 패딩, 예를 들어, 0, 1, 2, 3을 사용한다. 일 구현예에 따르면, 생성되는 피처 벡터는 오백십이(512) 치수이다.

[0134] 다른 구현예에서, 컨볼루션 신경망은, 상이한 수의 컨볼루션층, 서브샘플링층, 비선형 층, 및 완전히 연결된 층을 사용한다. 다른 일 구현예에서, 컨볼루션 신경망은, 층당 적은 층 및 많은 뉴런을 갖는 얇은 망이며, 예를 들어, 층당 100개 내지 200개의 뉴런을 갖는 한 개, 두 개, 또는 세 개의 완전히 연결된 층을 갖는다. 또 다른 일 구현예에서, 컨볼루션 신경망은, 층당 많은 층 및 적은 뉴런을 갖는 심층 망이며, 예를 들어, 3삼십(30)개 내지 오십(50)개의 뉴런을 갖는 다섯(5)개, 여섯(6)개 또는 여덟(8)개의 완전히 연결된 층을 갖는다.

[0135] 순방향 패스

[0136] 피쳐 맵의 f개의 컨볼루션 코어들의 수에 대한 제k 피쳐 스냅 및 제l 컨볼루션층의 행 x, 열 y의 뉴런의 출력은 이하의 식에 의해 결정된다:

$$O_{x,y}^{(l,k)} = \tanh\left(\sum_{t=0}^{f-1} \sum_{r=0}^{k_h} \sum_{c=0}^{k_w} W_{(r,c)}^{(k,t)} O_{(x+r,x+c)}^{(l-1,t)} + Bias^{(l,k)}\right)$$

[0137] 제k 피쳐 맵 및 제l 서브샘플층의 행 x, 열 y의 뉴런의 출력은 이하의 식에 의해 결정된다:

$$O_{x,y}^{(l,k)} = \tanh\left(W^{(k)} \sum_{r=0}^{S_h} \sum_{c=0}^{S_w} O_{(x \times S_h + r, y \times S_w + c)}^{(l-1,k)} + Bias^{(l,k)}\right)$$

[0140] 제l 출력층의 제i 뉴런의 출력은 이하의 식에 의해 결정된다:

$$O_{(l,i)} = \tanh\left(\sum_{j=0}^H O_{(l-1,j)} W_{(i,j)}^l + Bias^{(l,i)}\right)$$

[0142] 역전파

[0143] 출력층의 제k 뉴런의 출력 편차는 이하의 식에 의해 결정된다:

$$d(O_k^o) = y_k - t_k$$

[0145] 출력층의 제k 뉴런의 입력 편차는 이하의 식에 의해 결정된다:

$$d(I_k^o) = (y_k - t_k) \varphi'(v_k) = \varphi'(v_k) d(O_k^o)$$

[0147] 출력층의 제k 뉴런의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

$$\Delta W_{k,x}^o = d(I_k^o) y_{k,x}$$

$$\Delta Bias_k^o = d(I_k^o)$$

[0149] 숨겨진 층의 제k 뉴런의 출력 편향은 이하의 식에 의해 결정된다:

$$d(O_k^H) = \sum_{i=0}^{i-84} d(I_i^o) W_{i,k}$$

[0151] 숨겨진 층의 제k 뉴런의 입력 편향은 이하의 식에 의해 결정된다:

$$d(I_k^H) = \varphi'(v_k) d(O_k^H)$$

[0153] 숨겨진 층의 k개 뉴런으로부터의 입력을 수신하는 이전 층의 제m 피쳐 맵의 행 x, 열 y의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

$$\Delta W_{m,x,y}^{H,k} = d(I_k^H) y_{x,y}^m$$

$$\Delta Bias_k^H = d(I_k^H)$$

[0155] 서브샘플층 S의 제m 피쳐 맵의 행 x, 열 y의 출력 편향은 이하의 식에 의해 결정된다:

$$d(O_{x,y}^{S,m}) = \sum_k^{170} d(I_{m,x,y}^H) W_{m,x,y}^{H,k}$$

[0157] 서브샘플층 S의 제m 피쳐 맵의 행 x, 열 y의 입력 편향은 이하의 식에 의해 결정된다:

$$d(I_{x,y}^{S,m}) = \varphi'(v_k) d(O_{x,y}^{S,m})$$

[0159] 서브샘플층 S와 컨볼루션층 C의 제m 피쳐 맵의 행 x, 열 y의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

$$\Delta W^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m}) O_{x,y}^{C,m}$$

$$\Delta Bias^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m})$$

[0161] 컨볼루션층 C의 제k 피쳐 맵의 행 x, 열 y의 출력 편향은 이하의 식에 의해 결정된다:

$$d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k}) W^k$$

[0163] 컨볼루션층 C의 제k 피쳐 맵의 행 x, 열 y의 입력 편향은 이하의 식에 의해 결정된다:

$$d(I_{x,y}^{C,k}) = \varphi'(v_k) d(O_{x,y}^{C,k})$$

[0165] 제l 컨볼루션층 C의 제k 피쳐 맵의 제m 컨볼루션 코어의 행 r, 열 c의 가중치 및 편향 편차는 다음과 같다:

$$\Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{l-1,m}$$

$$\Delta Bias^{C,k} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k})$$

[0167] **잔여 연결**

[0168] 도 9는 피쳐 맵 추가를 통해 이전 정보를 하향 재주입하는 잔여 연결을 도시한다. 잔여 연결은, 과거 출력 텐서를 이후 출력 텐서에 추가함으로써 이전 표현을 데이터의 다운스트림 흐름으로 재주입하는 것을 포함하며, 이는 데이터 처리 흐름을 따른 정보 손실을 방지하는 데 도움이 된다. 잔여 연결은, 임의의 대규모 심층 학습 모델을 피로하는 두 가지 일반적인 문제점인, 그라디언트 소실 및 표현적 병목 현상에 대처한다. 일반적으로, 10개를 초과하는 층을 갖는 모델에 잔여 연결을 추가하는 것이 유익할 수 있다. 전술한 바와 같이, 잔여 연결은, 이전 층의 출력을 이후 층에 대한 입력으로서 이용 가능하게 하여 순차적 망에서의 지름길을 효과적으로 생성하는 것을 포함한다. 이후 활성화에 연결, 즉, 연쇄화(concatenate)되기보다는, 이전 출력이 이후 활성화와 합산되며, 이는 양측 활성화의 크기가 같다고 가정 한 것이다. 이들의 크기가 다른 경우, 이전 활성화를 목표 형상으로 재구성하는 선형 변환을 사용할 수 있다.

[0169] **잔여 학습 및 스킵 연결**

[0170] 도 10은 잔여 블록 및 스킵 연결의 일 구현예를 도시한다. 잔여 학습의 주요 아이디어는, 잔여 맵핑이 원래 맵보다 학습되는 것이 훨씬 쉽다는 것이다. 잔여 망은, 트레이닝 정확도의 저하를 완화하도록 많은 잔여 유닛을 적층한다. 잔여 블록은, 심층 신경망에서 소실되는 그라디언트를 방지하도록 특수 추가 스킵 연결을 이용한다. 잔여 블록의 시작시, 데이터 흐름은 두 개의 스트림으로 분리되며, 제1 스트림은 블록의 변경되지 않은 입력을

반송하고, 제2 스트림은 가중치와 비선형성을 적용한다. 블록의 끝에서, 두 개의 스트림은 요소별 합을 사용하여 병합된다. 이러한 구성의 주요 장점은, 그라디언트가 망을 통해 더욱 쉽게 흐를 수 있게 한다는 점이다.

[0171] 잔여 망으로부터의 이점을 통해, 심층 컨볼루션 신경망(CNN)을 쉽게 트레이닝할 수 있고, 이미지 분류 및 오브젝트 검출에 대한 정확도가 개선되었다. 컨볼루션 순방향 망은, 제1 층의 출력을 제(1+1) 층에 입력으로서 연결하며, 이는 이하의 층 친이 $x_t = H_t(x_{t-1})$ 를 발생시킨다. 잔여 블록은 식별 함수 $x_t = H_t(x_{t-1}) + x_{t-1}$ 로 비선형 변환을 우회하는 스킵 연결을 추가한다. 잔여 블록의 장점은, 그라디언트가 식별 함수를 통해 이후 층으로부터 이전 층으로 직접 흐를 수 있다는 점이다. 그러나, 식별 함수와 H_t의 출력은 합산에 의해 결합되며, 이는 네트워크의 정보 흐름을 방해할 수 있다.

[0172] **웨이브넷**

[0173] 웨이브넷은 원시 오디오 파형을 생성하기 위한 심층 신경망이다. 웨이브넷은, 저렴한 비용으로 비교적 큰 '시아'를 취할 수 있으므로 다른 컨볼루션 망과 구별된다. 또한, 국부적으로 그리고 전세계적으로 신호의 컨디셔닝을 추가할 수 있으며, 이는 웨이브넷을 여러 음성이 있는 텍스트 투 스피치(text to speech: TTS) 엔진으로서 사용할 수 있게 하며, 즉, TTS는 국부적 컨디셔닝 및 특정 음성에 글로벌 컨디셔닝을 제공한다.

[0174] 웨이브넷의 주요 빌딩 블록은 인과적 팽창 컨볼루션이다. 인과적 팽창 컨볼루션에 대한 확장으로서, 웨이브넷은, 도 11에 도시된 바와 같이 이들 컨볼루션의 스택을 허용한다. 이 도면에서 팽창 컨볼루션이 있는 동일한 수용장을 취득하려면, 다른 팽창 층이 필요하다. 스택은, 팽창 컨볼루션층의 출력을 단일 출력에 연결하는 팽창 컨볼루션의 반복이다. 이는, 웨이브넷이 비교적 작은 연산 비용으로 하나의 출력 노드의 큰 '시각적' 필드를 얻을 수 있게 한다. 비교를 위해, 512개 입력의 시각적 필드를 얻기 위해서는, 완전 컨볼루션 망(FCN)이 511개의 층을 필요로 한다. 팽창 컨볼루션 망의 경우에는, 8개의 층이 필요하다. 적층된 팽창 컨볼루션에는, 2개의 스택이 있는 7개의 층 또는 4개의 스택이 있는 6개의 층만이 필요하다. 동일한 시각적 필드를 커버하는 데 필요한 연산 능력의 차이에 대한 아이디어를 얻기 위해, 이하의 표는, 층당 하나의 필터와 2의 필터 폭을 가정할 때 망에 필요한 가중치의 수를 나타낸다. 또한, 네트워크가 8비트의 이진 인코딩을 사용하고 있다고 가정한다.

네트워크 유형	스택 수	채널당 가중치 수	가중치 총 수
FCN	1	$2.6 \cdot 10^5$	$2.6 \cdot 10^6$
WN	1	1022	8176
WN	2	1022	8176
WN	4	508	4064

[0175] .

[0176] 웨이브넷은, 잔여 연결이 이루어지기 전에 스킵 연결을 추가하며, 이는 이하의 잔여 블록을 모두 우회한다. 이러한 스킵 연결의 각각은, 일련의 활성화 함수 및 컨볼루션을 통과하기 전에 합산된다. 직관적으로, 이것은 각 층에서 추출된 정보의 합이다.

[0177] **일괄 정규화**

[0178] 일괄 정규화는, 데이터 표준화를 네트워크 아키텍처의 필수 부분으로서 만듦으로써 심층 네트워크 트레이닝을 가속화하는 방법이다. 일괄 정규화는 트레이닝 동안 시간이 지남에 따라 평균 및 분산이 변하더라도 데이터를 적응적으로 정규화할 수 있다. 이것은, 트레이닝 중에 보여지는 데이터의 일괄별 평균 및 분산의 지수 이동 평균을 내부적으로 유지함으로써 가능하다. 일괄 정규화의 주요 효과는, 잔여 연결과 매우 유사하게 그라디언트 전파에 도움이 되어 심층 망을 허용한다는 점이다. 일부 고 심층 망은, 여러 일괄 정규화층을 포함하는 경우에만 트레이닝될 수 있다.

[0179] 일괄 정규화는, 완전히 연결된 층 또는 컨볼루션층과 같이 모델 아키텍처에 삽입될 수 있는 또 다른 층이라고 할 수 있다. 일괄정규화층은 통상적으로 컨볼루션 또는 조밀하게 연결된 층 뒤에 사용된다. 이것은, 컨볼루션층 또는 조밀하게 연결된 층 전에도 사용될 수 있다. 양측 구현에는 개시된 기술에 의해 사용될 수 있으며 도 15에 도시되어 있다. 일괄정규화층은 축 인수를 사용하며, 이러한 축 인수는 정규화되어야 하는 피쳐 축을 특정한다. 이 인수의 기본값은 입력 텐서의 마지막 축인 -1이다. 이것은, data_format이 "channels_last"로 설정된 조밀층, Conv1D 층, RNN 층, 및 Conv2D 층을 사용할 때의 올바른 값이다. 그러나, Conv2D 층이

"channels_first"로 설정된 data_format을 갖는 틸새 사용의 경우에, 피쳐 축은 축 1이고, 일괄정규화에서의 축 인수는 1로 설정될 수 있다.

[0180] 일괄 정규화는, 입력을 피드포워드하고 역전파를 통해 파라미터 및 자신의 고유 입력에 대한 그라디언트를 연산하기 위한 정의를 제공한다. 실제로, 일괄 정규화층은, 컨볼루션층 또는 완전히 연결된 층 뒤에 삽입되지만, 출력이 활성화 함수에 공급되기 전에 삽입된다. 컨볼루션층의 경우, 컨볼루션 속성을 준수하기 위해 상이한 위치에 있는 동일한 피쳐 맵의 상이한 요소- 즉, 활성화 -가 동일한 방식으로 정규화된다. 따라서, 미니-일괄의 모든 활성화는, 활성화마다 정규화되기보다는 모든 위치에서 정규화된다.

[0181] 내부 공변량 시프트는, 왜 심층 아키텍처의 트레이닝 속도가 심각하게 느렸는지에 대한 주요 이유이다. 이는, 심층망이 각 층에서 새로운 학습할 필요가 있을 뿐만 아니라 해당 분포의 변화도 고려해야 한다는 사실에서 비롯된 것이다.

[0182] 공변량 시프트는, 일반적으로 심층 학습 영역에서 알려진 문제이며, 실세계 문제에서 자주 발생한다. 일반적인 공변량 시프트 문제는 트레이닝 및 테스트 세트의 분포 차이이며, 이는 준최적화된 일반화 성능으로 이어질 수 있다. 이 문제는 일반적으로 표준화 또는 미백 전처리 단계에서 다루어진다. 그러나, 특히 미백 연산은, 연산적으로 비싸고, 따라서 특히 공변량 시프트가 상이한 층에 걸쳐 발생하는 경우 온라인 설정에서 비실용적이다.

[0183] 내부 공변량 시프트는, 트레이닝 동안 망 파라미터의 변화로 인해 망 활성화의 분포가 층에 걸쳐 변화하는 현상이다. 이상적으로는 각 층이, 동일한 분포를 갖지만 기능적 관계는 동일하게 유지되는 공간으로 변환되어야 한다. 모든 층과 단계에서 데이터를 상관해제하고 피백하기 위한 공분산 행렬의 고가의 계산을 피하기 위해, 각 미니-일괄에 걸쳐 각 층의 각 입력 피쳐 분포를 평균 0과 표준 편차 1을 갖도록 정규화한다.

[0184] **순방향 패스**

[0185] 순방향 패스 동안, 미니-일괄 평균 및 분산이 계산된다. 이러한 미니-일괄 통계를 사용하면, 데이터는 평균을 빼고 표준 편차로 나눔으로써 정규화된다. 마지막으로, 학습된 스케일 및 시프트 파라미터를 사용하여 데이터를 스케일링하고 시프트한다. 일괄 정규화 순방향 패스 f_{BN} 은 도 12에 도시되어 있다.

[0186] 도 12에서, 각각 μ_{β} 는 일괄 평균이고 σ_{β}^2 는 일괄 분산이다. 학습된 스케일 및 시프트 파라미터는 각각 γ 및 β 로 표시된다. 명확성을 위해, 일괄 정규화 절차는 활성화마다 본 명세서에 기재되고 상응하는 지수를 생략한다.

[0187] 정규화는 미분가능한 변환이므로, 오류는, 이러한 학습된 파라미터로 전파되므로, 식별 변환을 학습함으로써 망의 표현력을 복원할 수 있다. 반대로, 해당 일괄 통계와 동일한 스케일 및 시프트 파라미터를 학습하면, 일괄 정규화 변환이 최적의 수행 작업인 경우 네트워크에 영향을 미치지 않는다. 테스트 시간 때, 일괄 평균 및 분산은, 입력이 미니-일괄의 다른 샘플에 의존하지 않으므로, 각 모집단 통계량에 의해 대체된다. 다른 방법은, 트레이닝 중에 일괄 통계의 이동 평균을 계속 유지하고 이를 사용하여 테스트 시간에 망 출력을 연산하는 것이다.

테스트 시간에, 일괄 정규화 변환은 도 13에 도시된 바와 같이 표현될 수 있다. 도 13에서, μ_D 와 σ_D^2 는 각각 일괄 통계라기보다는 모집단 평균과 분산을 나타낸다.

[0188] **역방향 패스**

[0189] 정규화는 미분가능한 동작이므로, 역방향 통과는 도 14에 도시된 바와 같이 연산될 수 있다.

[0190] **1D 컨볼루션**

[0191] 1D 컨볼루션은, 도 16에 도시한 바와 같이, 서열로부터 로컬 1D 패치 또는 서브시퀀스를 추출한다. 1D 컨볼루션은 입력 서열의 시간 패치로부터 각 출력 타임 스텝을 취득한다. 1D 컨볼루션층은 서열의 국부 패턴을 인식한다. 모든 패치에 대하여 동일한 입력 변환이 수행되므로, 입력 서열의 특정 위치에서 학습된 패턴을 나중에 다른 위치에서 인식할 수 있으므로, 1D 컨볼루션층 변환이 시간 변환에 대해 변하지 않게 한다. 예를 들어, 크기가 5인 컨볼루션 윈도우를 사용하는 염기의 1D 컨볼루션층 처리 서열은, 길이가 5 이하인 염기 또는 염기 서열을 학습할 수 있어야 하며, 입력 서열의 모든 컨텍스트에서 염기 모티프를 인식할 수 있어야 한다. 따라서, 염기 수준의 1D 컨볼루션은 염기 형태에 대해 학습할 수 있다.

[0192] **글로벌 평균 풀링**

- [0193] 도 17은 글로벌 평균 풀링(GAP) 작업의 동작 방식을 도시한다. 점수 매김(scoring)을 위해 마지막 층의 피치들의 공간 평균을 취함으로써 분류를 위해 완전히 연결된(FC) 층들을 대체하는 데 글로벌 평균 풀링을 사용할 수 있다. 이것은 트레이닝 부하를 감소시키고 과적합 문제를 우회한다. 글로벌 평균 풀링은, 모델 이전에 구조를 적용하며, 사전 정의된 가중치를 이용하는 선형 변환과 같다. 글로벌 평균 풀링은 파라미터의 수를 감소시키고 완전히 연결된 층을 제거한다. 완전히 연결된 층들은 통상적으로 가장 파라미터가 많고 연결이 많은 층들이며, 글로벌 평균 풀링은 비슷한 결과를 얻기 위해 훨씬 저렴한 비용의 접근 방식을 제공한다. 글로벌 평균 풀링의 주요 아이디어는, 각 마지막 층 피치 맵으로부터 평균값을 점수 매김을 위한 신뢰인자로서 생성하여 소프트맥스 층에 직접 공급하는 것이다.
- [0194] 글로벌 평균 풀링은 3가지 이점을 갖는데, 즉, (1) 글로벌 평균 풀링층에 추가 파라미터가 없으므로, 글로벌 평균 풀링층에서 과적합을 피하고, (2) 글로벌 평균 풀링의 출력은 전체 피치 맵의 평균이므로, 글로벌 평균 풀링이 공간 변환에 더 강력하고, (3) 전체 맵의 모든 파라미터에서 일반적으로 50% 넘게 차지하는 완전히 연결된 층들의 파라미터의 수가 많기 때문에, 이들을 글로벌 평균 풀링층으로 대체하면, 모델의 크기를 상당히 감소시킬 수 있고, 이는 글로벌 평균 풀링을 모델 압축에 있어서 매우 유용하게 한다.
- [0195] 글로벌 평균 풀링은, 마지막 층에서 더 강한 피치가 더 높은 평균값을 가질 것으로 예상되므로, 의미가 있다. 일부 구현예에서, 글로벌 평균 풀링은 분류 점수에 대한 프록시로서 사용될 수 있다. 글로벌 평균 풀링의 영향을 받는 피치 맵은, 신뢰 맵으로서 해석될 수 있으며, 피치 맵과 카테고리 간의 대응을 강제할 수 있다. 글로벌 평균 풀링은, 마지막 층 피치가 직접 분류에 대한 충분한 추상에 있는 경우 특히 효과적일 수 있지만, 멀티레벨 피치 기능을 부분 모델과 같은 그룹으로 결합해야 하는 경우 글로벌 평균 풀링만으로는 충분하지 않으며, 이러한 결합은 글로벌 평균 풀링 후 단순한 완전히 연결된 층 또는 다른 분류자(classifier)를 추가함으로써 가장 잘 수행된다.
- [0196] **용어**
- [0197] 특허, 특허출원, 기사, 서적, 논문, 및 웹페이지를 포함하지만 이에 제한되지 않는 본 명세서에 인용된 모든 문헌 및 유사 자료의 전문은, 이러한 문헌 및 유사 자료의 형식에 관계없이, 본 명세서에 참고로 인용된다. 통합된 문헌과 유사 자료 중 하나 이상이 정의 용어, 용어 사용, 설명된 기술 등을 포함하지만 이에 제한되지 않는 본 발명과 상이하거나 상반되는 경우에는, 본 발명이 우선한다.
- [0198] 본 명세서에서 사용되는 바와 같이, 하기 용어들은 지시된 의미를 갖는다.
- [0199] 염기는 뉴클레오타이드 염기 또는 뉴클레오타이드, A(아데닌), C(사이토신), T(티민), 또는 G(구아닌)를 가리킨다.
- [0200] 본 출원은 "단백질" 및 "번역된 서열"이라는 용어를 호환 가능하게 사용한다.
- [0201] 본 출원은 "코돈" 및 "염기 트리플렛"이라는 용어를 호환 가능하게 사용한다.
- [0202] 본 출원은 "아미노산" 및 "번역된 유닛"이라는 용어를 호환 가능하게 사용한다.
- [0203] 본 출원은 "변이체 병원성 분류자", "변이체 분류를 위한 컨볼루션 신경망-기반 분류자", "변이체 분류를 위한 심층 컨볼루션 신경망 기반 분류자"라는 어구를 호환 가능하게 사용한다.
- [0204] "염색체"라는 용어는, DNA 및 단백질 성분(특히 히스톤)을 포함하는 염색질 가닥으로부터 유도된 살아있는 세포의 유전-보유 유전자 운반체를 지칭한다. 종래의 국제적으로 인정되는 개별 인간 게놈 염색체 넘버링 시스템이 본 명세서에서 사용된다.
- [0205] "부위"라는 용어는, 참조 게놈 상의 고유한 위치(예를 들어, 염색체 ID, 염색체 위치, 및 배향)를 지칭한다. 일부 구현예에서, 부위는 잔기, 서열 태그, 또는 서열 상의 세그먼트의 위치일 수 있다. "좌위"(locus)라는 용어는 참조 염색체 상의 핵산 서열 또는 다형성의 특정 위치를 지칭하는 데 사용될 수 있다.
- [0206] 본 명세서에서 "샘플"이라는 용어는, 통상적으로 핵산을 함유하는 생물학적 유체, 세포, 조직, 기관, 또는 유기체, 혹은 서열분석될 그리고/또는 상처리(phase)될 적어도 하나의 핵산 서열을 함유하는 핵산들의 혼합물로부터 유도된 샘플을 지칭한다. 이러한 샘플은, 객담/경구 액, 양수, 혈액, 혈액 분획물, 미세 침 생검 샘플(예를 들어, 외과적 생검, 미세 침 생검 등), 소변, 복막액, 흉막액, 조직 외식편, 기관 배양물, 및 다른 임의의 조직 또는 세포 제제, 또는 이들의 분획물이나 유도체 또는 이들로부터 분리된 분획물이나 유도체를 포함하지만 이에 제한되지는 않는다. 샘플은 종종 인간 대상(예를 들어, 환자)으로부터 채취되지만, 샘플은, 개, 고양이, 말, 염

소, 양, 소, 돼지 등을 포함하지만 이들로 제한되지 않는 염색체를 갖는 임의의 유기체로부터 채취될 수 있다. 샘플은, 생물학적 공급원으로부터 취득되었을 때 그대로 또는 샘플의 특성을 변경하도록 전처리에 이어서 사용될 수 있다. 예를 들어, 이러한 전처리는, 혈액으로부터 혈장을 제조하고 점성 유체 등을 희석하는 것을 포함할 수 있다. 전처리 방법은, 또한, 여과, 침전, 희석, 증류, 혼합, 원심분리, 동결, 동결건조, 농축, 증폭, 핵산 단편화, 간섭 성분의 비활성화, 시약의 첨가, 용해 등을 포함할 수 있지만, 이들로 제한되지는 않는다.

[0207] "서열"이라는 용어는 서로 연결된 뉴클레오타이드의 가닥을 포함하거나 나타낸다. 뉴클레오타이드는 DNA 또는 RNA에 기초할 수 있다. 하나의 서열은 다수의 하위서열(sub-sequence)을 포함할 수 있음을 이해해야 한다. 예를 들어, (예를 들어, PCR 앰플리콘의) 단일 서열은 350개의 뉴클레오타이드를 가질 수 있다. 샘플 리드(read)는 이들 350개 뉴클레오타이드 내에 다수의 하위서열을 포함할 수 있다. 예를 들어, 샘플 리드는, 예를 들어, 20개 내지 50개의 뉴클레오타이드를 갖는 제1 및 제2 플랭킹 서열(flanking subsequence)을 포함할 수 있다. 제1 및 제2 플랭킹 하위서열은, 상응하는 하위서열(예를 들어, 40개 내지 100개의 뉴클레오타이드)를 갖는 반복 세그먼트의 어느 일측에 위치할 수 있다. 플랭킹 하위서열의 각각은 프라이머 하위서열(예를 들어, 10개 내지 30개의 뉴클레오타이드) 또는 프라이머 하위서열의 일부를 포함할 수 있다. 용이한 판독을 위해, "서열"이라는 용어는, "서열"로 지칭될 것이나, 두 개의 서열이 반드시 공통 가닥 상에서 서로 분리될 필요는 없음을 이해할 수 있다. 본 명세서에 기재된 다양한 서열을 구별하기 위해, 서열에는 상이한 표지(예를 들어, 목표 서열, 프라이머 서열, 측면 서열, 참조 서열 등)가 제공될 수 있다. "대립유전자"와 같은 다른 용어에는 유사한 대상들을 구별하도록 다른 표지가 부여될 수 있다.

[0208] "페어드-엔드 서열분석"(paired-end sequencing)이라는 용어는 목표 분획물의 양측 말단을 서열분석하는 서열분석 방법을 지칭한다. 페어드 엔드 서열분석은, 유전자 융합 및 신규한 전사뿐만 아니라 게놈 재배열 및 반복 세그먼트의 검출을 용이하게 할 수 있다. 페어드-엔드 서열분석 방법은, PCT 공보 W007010252, PCT 출원 일련번호 PCTGB2007/003798, 및 미국 특허출원 공개공보 US 2009/0088327에 기재되어 있으며, 이들 각각은 본 명세서에 참고로 인용된다. 일례로, 일련의 동작을 다음과 같이 수행할 수 있는데, 즉, (a) 핵산들의 클러스터를 생성하고; (b) 핵산들을 선형화하고; (c) 제1 서열분석 프라이머를 혼성화하고 상기한 바와 같이 확장, 스캐닝 및 디블로킹(deblocking)의 반복 사이클을 수행하고, (d) 상보적 사본을 합성함으로써 유동 세포면 상의 목표 핵산을 "반전"시키고, (e) 재합성된 가닥을 선형화하고, (f) 제2 서열분석 프라이머를 혼성화하고 상기한 바와 같이 확장, 스캐닝 및 디블로킹의 반복 사이클을 수행한다. 단일 사이클의 브리지 증폭에 대해 전술한 바와 같은 시약을 전달하여 반전 작업을 수행할 수 있다.

[0209] "참조 게놈" 또는 "참조 서열"이라는 용어는, 대상으로부터 확인된 서열을 참조하는 데 사용될 수 있는, 부분적 인지 완전한지에 상관 없이 임의의 유기체의 임의의 특정한 알려진 게놈 서열을 지칭한다. 예를 들어, 인간 대상 및 다른 많은 유기체에 사용되는 참조 게놈은 ncbi.nlm.nih.gov의 국립 생명공학 정보 센터에서 찾을 수 있다. "게놈"은, 핵산 서열로 발현된 유기체 또는 바이러스의 완전한 유전자 정보를 지칭한다. 게놈에는 유전자와 DNA의 비암호화 서열이 모두 포함된다. 참조 서열은 이러한 서열에 정렬된 리드보다 클 수 있다. 예를 들어, 참조 서열은, 적어도 약 100배 이상, 또는 적어도 약 1000배 이상, 또는 적어도 약 10,000배 이상, 또는 적어도 약 10^5 배 이상, 또는 적어도 약 10^6 배 이상, 또는 적어도 약 10^7 배 이상일 수 있다. 일례로, 참조 게놈 서열은 전장 인간 게놈의 서열이다. 다른 일례에서, 참조 게놈 서열은 염색체 13과 같은 특정 인간 염색체로 제한된다. 일부 구현예에서, 참조 염색체는 인간 게놈 버전 hg19로부터의 염색체 서열이다. 참조 게놈이라는 용어는 이러한 서열을 커버하도록 의도되었지만, 이러한 서열은 염색체 기준 서열이라고 칭할 수 있다. 참조 서열의 다른 예는, 임의의 종의 염색체, (가닥과 같은) 부염색체 영역 등뿐만 아니라 다른 종의 게놈도 포함한다. 다양한 구현예에서, 참조 게놈은 컨센서스 서열 또는 다수의 개체로부터 유도된 다른 조합이다. 그러나, 소정의 응용분야에서, 참조 서열은 특정 개체로부터 취해질 수 있다.

[0210] "리드"라는 용어는, 뉴클레오타이드 샘플 또는 참조의 분획물을 기술하는 서열 데이터의 수집을 지칭한다. "리드"라는 용어는 샘플 리드 및/또는 참조 리드를 지칭할 수 있다. 통상적으로, 반드시 그런 것은 아니지만, 리드는 샘플 또는 참조에서의 연속 염기쌍의 짧은 서열을 나타낸다. 리드는 샘플 또는 참조 분획물의 (ATCG로 된) 염기쌍 서열에 의해 상징적으로 표현될 수 있다. 리드는, 리드가 참조 서열과 일치하는지 또는 다른 기준을 충족하는지를 결정하도록 메모리 장치에 저장될 수 있고 적절하게 처리될 수 있다. 리드는, 서열분석 장치로부터 직접 또는 샘플에 관한 저장된 서열 정보로부터 간접적으로 취득될 수 있다. 일부 경우에, 리드는, 더 큰 서열 또는 영역을 확인하도록 사용될 수 있는, 예를 들어, 염색체 또는 게놈 영역 또는 유전자에 정렬되고 특정하게 할당될 수 있는 충분한 길이(예를 들어, 적어도 약 25bp)의 DNA 서열이다.

- [0211] 차세대 서열분석 방법은, 예를 들어, 합성 기술(일루미나(Illumina))에 의한 서열분석, 파이로시퀀싱(454), 이온 반도체 기술(이온 토렌트(Ion Torrent) 서열분석), 단일-분자 실시간 서열분석(퍼시픽 바이오사이언시스사(Pacific Biosciences)), 및 결찰(SOLiD 서열분석)에 의한 시퀀싱을 포함한다. 서열분석 방법에 따라, 각 리드의 길이는 약 30bp 내지 10,000bp를 초과하도록 가변될 수 있다. 예를 들어, SOLiD 시퀀서를 이용한 일루미나 서열분석 방법은 약 50bp의 핵산 리드를 생성한다. 다른 예에서, 이온 토렌트 서열분석은 최대 400bp의 핵산 리드를 생성하고, 454 파이로시퀀싱은 약 700bp의 핵산 리드를 생성한다. 또 다른 예에서, 단일-분자 실시간 서열분석 방법은 10,000bp 내지 15,000bp의 리드를 생성할 수 있다. 따라서, 소정의 구현예에서, 핵산 서열 리드의 길이는 30bp 내지 100bp, 50bp 내지 200bp, 또는 50np 내지 400bp의 길이를 갖는다.
- [0212] "샘플 리드", "샘플 서열" 또는 "샘플 분획물"이라는 용어는 샘플로부터의 관심 게놈 서열에 대한 서열 데이터를 지칭한다. 예를 들어, 샘플 리드는, 순방향 및 역방향 프라이머 서열을 갖는 PCR 앰플리콘으로부터의 서열 데이터를 포함한다. 서열 데이터는 임의의 선택 서열 방법으로부터 취득될 수 있다. 샘플 리드는, 예를 들어, 합성에 의한 서열분석(SBS) 반응, 결찰에 의한 서열분석 반응, 또는 다른 임의의 적합한 서열분석 방법으로부터 발생하는 것일 수 있으며, 이를 위해 이미지의 요소의 길이 및/또는 동일성을 결정하는 것이 필요하다. 샘플 리드는, 다수의 샘플 리드로부터 유도된 컨센서스(예를 들어, 평균 또는 가중) 서열일 수 있다. 소정의 구현예에서, 참조 서열을 제공하는 것은, PCR 앰플리콘의 프라이머 서열에 기초하여 관심 좌위를 식별하는 것을 포함한다.
- [0213] "원시 분획물"이라는 용어는, 샘플 리드 또는 샘플 분획물 내의 관심있는 지정된 위치 또는 이차 위치와 적어도 부분적으로 중복되는 관심 게놈 서열의 일부에 대한 서열 데이터를 지칭한다. 원시 분획물의 비제한적인 예로는, 이중 스티치 분획물, 단일 스티치 분획물, 이중 언스티치 분획물, 및 단일 언스티치 분획물을 포함한다. "원시"라는 용어는, 원시 분획물이 샘플 리드의 잠재적 변이체에 대응하고 이러한 잠재적 변이체를 인증 또는 확인하는 변이체를 나타내는지의 여부에 관계없이, 원시 분획물이 샘플 리드에서 서열 데이터와 일부 관계가 있는 서열 데이터를 포함한다는 것을 나타내는 데 사용된다. "원시 분획물"이라는 용어는, 분획물이 반드시 샘플 리드에서 변이체 호출을 유효성 확인하는 지지 변이체를 포함한다는 것을 나타내지는 않는다. 예를 들어, 제1 변이체를 나타내기 위해 변이체 호출 애플리케이션에 의해 샘플 리드가 결정될 때, 변이체 호출 애플리케이션은, 하나 이상의 원시 분획물이 다른 경우엔 샘플 리드의 변이체가 주어지는 경우 발생할 것으로 예상될 수 있는 대응 유형의 "지지" 변이체를 갖지 않는다고 결정할 수 있다.
- [0214] "맵핑", "정렬된", "정렬", 또는 "정렬하는"이라는 용어는, 리드 또는 태그를 참조 서열과 비교하여 참조 서열이 리드 서열을 포함하는지를 결정하는 프로세스를 지칭한다. 참조 서열이 리드를 포함하는 경우, 리드는, 참조 서열에 맵핑될 수 있고, 또는 특정 구현예에서 참조 서열의 특정 위치에 맵핑될 수 있다. 일부 경우에, 정렬은, 리드가 특정 참조 서열의 구성원인지 여부(즉, 리드가 참조 서열에 존재하는지 또는 부재하는지)를 단순히 알려준다. 예를 들어, 인간 염색체 13에 대한 참조 서열에 대한 리드의 정렬은, 염색체 13에 대한 참조 서열에 리드가 존재하는지의 여부를 알려줄 것이다. 이 정보를 제공하는 도구를 세트 멤버십 테스터라고 한다. 일부 경우에, 정렬은, 리드 태그가 맵핑되는 참조 서열의 위치를 추가로 나타낸다. 예를 들어, 참조 서열이 전체 인간 게놈 서열인 경우, 정렬은, 리드가 염색체 13에 존재함을 나타내고, 리드가 특정 가닥 및/또는 염색체 13의 부위에 있음을 추가로 나타낼 수 있다.
- [0215] "인델"(indel)이라는 용어는, 유기체의 DNA에서의 염기의 삽입 및/또는 삭제를 지칭한다. 마이크로-인델은, 1개 내지 50개 뉴클레오타이드의 순 변화를 초래하는 인델을 나타낸다. 게놈의 코딩 영역에서, 인델의 길이가 3의 배수가 아닌 한, 이것은 프레임시프트 돌연변이를 생성할 것이다. 인델은 점 돌연변이와 대조될 수 있다. 인델은 뉴클레오타이드를 삽입하고 서열로부터 삭제하는 반면, 점 돌연변이는 DNA의 전체 수를 변경하지 않고 뉴클레오타이드들 중 하나를 대체하는 치환 형태이다. 인델은, 또한, 인접한 뉴클레오타이드에서의 치환으로서 정의될 수 있는 탠덤 염기 돌연변이(TBM)와 대조될 수 있다 (주로 2개의 인접한 뉴클레오타이드에서의 치환에 해당하지만, 3개의 인접한 뉴클레오타이드에서의 치환이 관찰되었다).
- [0216] "변이체"라는 용어는, 핵산 참조와는 다른 핵산 서열을 지칭한다. 통상적인 핵산 서열 변이체는, 단일 뉴클레오타이드 다형성(SNP), 짧은 삭제 및 삽입 다형성(Indel), 카피 수 변이(CNV), 마이크로위성 마커, 또는 짧은 탠덤 반복 및 구조적 변이를 제한 없이 포함한다. 체세포 변이체 호출은, DNA 샘플에서 낮은 빈도로 존재하는 변이체를 식별하기 위한 노력이다. 체세포 변이체 호출은 암 치료의 맥락에서 중요하다. 암은, DNA에 돌연변이가 축적되어 발생하는 것이다. 종양으로부터의 DNA 샘플은, 일반적으로 일부 정상 세포, (돌연변이가 적은) 암 진행의 초기 단계의 일부 세포, 및 (돌연변이가 많은) 일부 후기 단계 세포를 포함하여 이종성이다. 이러한 이종성 때문에, (예를 들어, FFPE 샘플로부터) 종양을 시퀀싱할 때, 체세포 돌연변이는 종종 낮은 빈도로 나타난다.

예를 들어, SNV는 주어진 염기를 커버하는 리드의 10%에서만 보일 수 있다. 변이체 분류자에 의해 체세포 또는 생식세포로서 분류되는 변이체도, 본 명세서에서 "테스트 중인 변이체"라고 지칭된다.

- [0217] "노이즈"라는 용어는, 서열분석 프로세스 및/또는 변이체 호출 애플리케이션에서의 하나 이상의 에러로 인한 잘못된 변이체 호출을 지칭한다.
- [0218] "변이체 빈도"라는 용어는, 모집단의 특정 좌위에서의 대립유전자(유전자의 변이체)의 상대 빈도를 분획율 또는 백분율로서 표현한 것을 나타낸다. 예를 들어, 분획율 또는 백분율은 해당 대립유전자를 보유하는 모집단에서의 모든 염색체의 분획률일 수 있다. 예를 들어, 샘플 변이체 빈도는, 개인으로부터 관심 게놈 서열에 대하여 취득된 샘플 및/또는 리드의 수에 상응하는 "모집단"에 대한 관심 게놈 서열을 따른 특정 좌위/위치에서의 대립유전자/변이체의 상대 빈도를 나타낸다. 다른 일례로, 베이스라인 변이체 빈도는, 하나 이상의 베이스라인 게놈 서열을 따른 특정 좌위/위치에서의 대립유전자/변이체의 상대 빈도를 나타내며, 여기서 "모집단"은, 정상적인 개인들의 모집단으로부터 하나 이상의 베이스라인 게놈 서열에 대하여 취득된 샘플 및/또는 리드의 수에 상응한다.
- [0219] 용어 "변이체 대립유전자 빈도"(VAF)는, 변이체를 목표 위치에서의 전체 커버리지로 나눈 값과 일치하는 것으로 관찰된 서열분석된 리드의 백분율을 지칭한다. VAF는 변이체를 전달하는 서열분석된 리드의 비율을 측정하는 것이다.
- [0220] "위치", "지정된 위치" 및 "좌위"라는 용어는, 뉴클레오타이드들의 서열 내에서의 하나 이상의 뉴클레오타이드의 위치 또는 좌표를 지칭한다. "위치", "지정된 위치" 및 "좌위"라는 용어들은, 또한, 뉴클레오타이드들의 서열에서의 하나 이상의 염기 쌍의 위치 또는 좌표를 지칭한다.
- [0221] "일배체형"이라는 용어는 함께 유전되는 염색체 상의 인접 부위에 있는 대립유전자들의 조합을 지칭한다. 일배체형은, 좌위의 주어진 세트가 발생하였다면, 이러한 세트 간에 발생한 재조합 이벤트들의 수에 따라 하나의 좌위, 여러 개의 좌위, 또는 전체 염색체일 수 있다.
- [0222] 본 명세서에서 "임계값"이라는 용어는, 샘플, 핵산, 또는 그 일부(예를 들어, 리드)를 특성화하도록 쿼리로 사용되는 숫자 또는 비슷자 값을 지칭한다. 임계값은 경험적 분석에 기초하여 가변될 수 있다. 임계값은, 이러한 값을 발생시키는 소스가 특정 방식으로 분류되어야 하는지의 여부를 결정하도록 측정된 값 또는 계산된 값과 비교될 수 있다. 임계값은 경험적으로 또는 분석적으로 식별될 수 있다. 임계값의 선택은, 사용자가 분류를 원하는 신뢰 수준에 의존한다. 임계값은, 특정 목적을 위해(예를 들어, 감도 및 선택성의 균형을 맞추기 위해) 선택될 수 있다. 본 명세서에서 사용되는 바와 같이, "임계값"이라는 용어는, 분석 과정이 변경될 수 있는 지점 및/또는 동작이 트리거될 수 있는 지점을 나타낸다. 임계값은 미리 정해진 수일 필요가 없다. 대신, 임계값은, 예를 들어, 복수의 인자에 기초한 함수일 수 있다. 임계값은 상황에 적응적일 수 있다. 또한, 임계값은 상한값, 하한값, 또는 한계값들 사이의 범위를 나타낼 수 있다.
- [0223] 일부 구현예에서는, 서열분석 데이터에 기초한 메트릭 또는 점수가 임계값과 비교될 수 있다. 본 명세서에서 사용되는 바와 같이, "메트릭" 또는 "점수"라는 용어는, 서열분석 데이터로부터 결정된 값 또는 결과를 포함할 수 있다. 임계값과 마찬가지로, 메트릭 또는 점수는 상황에 따라 적응적일 수 있다. 예를 들어, 메트릭 또는 점수는 정규화된 값일 수 있다. 점수 또는 메트릭의 예로서, 하나 이상의 구현예는 데이터를 분석할 때 계수치 점수를 사용할 수 있다. 계수치 점수는 샘플 리드의 수에 기초할 수 있다. 샘플 리드는, 샘플 리드가 하나 이상의 공통 특성 또는 품질을 갖도록 하나 이상의 필터링 단계를 겪을 수 있다. 예를 들어, 계수치 점수를 결정하기 위해 사용되는 각각의 샘플 리드는 참조 서열과 정렬되었을 수 있고 또는 잠재적 대립유전자로서 할당될 수 있다. 공통 특성을 갖는 샘플 리드의 수는 리드 카운*를 결정하기 위해 계수될 수 있다. 계수치 점수는 리드 계수치에 기초할 수 있다. 일부 구현예에서, 계수치 점수는 리드 계수치와 동일한 값일 수 있다. 다른 구현예에서, 계수치 점수는 리드 계수치 및 다른 정보에 기초할 수 있다. 예를 들어, 계수치 점수는, 유전 좌위의 특정 대립유전자에 대한 리드 수 및 유전 좌위에 대한 총 리드 수에 기초할 수 있다. 일부 구현예에서, 계수치 점수는 유전 좌위에 대한 리드 계수치 및 이전에 취득된 데이터에 기초할 수 있다. 일부 구현예에서, 계수치 점수는 미리 결정된 값들 간에 정규화된 점수일 수 있다. 계수치 점수는, 또한, 샘플의 다른 좌위로부터의 리드 계수치의 함수 또는 관심 샘플과 동시에 실행된 다른 샘플로부터의 리드 카운*의 함수일 수 있다. 예를 들어, 계수치 점수는, 특정 대립유전자의 리드 계수치 및 샘플 내의 다른 좌위의 리드 계수치 및/또는 다른 샘플로부터의 리드 계수치의 함수일 수 있다. 일례로, 다른 좌위로부터의 리드 계수치 및/또는 다른 샘플로부터의 리드 계수치는 특정 대립유전자에 대한 계수치 점수를 정규화하는 데 사용될 수 있다.

- [0224] "커버리지" 또는 "프래그먼트 커버리지"라는 용어는, 서열의 동일한 프래그먼트에 대한 다수의 샘플 리드의 계수치 또는 다른 측정값을 지칭한다. 리드 계수치는 대응하는 프래그먼트를 커버하는 리드 수의 계수치를 나타낼 수 있다. 대안으로, 커버리지는, 이력 지식, 샘플의 지식, 좌위의 지식 등에 기초하는 지정된 계수에 리드 계수치를 곱함으로써 결정될 수 있다.
- [0225] "리드 깊이"(통상적으로 " \times "가 후속하는 수)라는 용어는 목표 위치에서 중복되는 정렬을 갖는 서열분석된 리드의 수를 지칭한다. 이는 종종 간격들의 세트(예를 들어, 엑손, 유전자 또는 패널)에 걸쳐 컷오프를 초과하는 평균 또는 백분율로서 표현된다. 예를 들어, 임상 보고서에 따르면, 패널 평균 커버리지가 $1,105 \times$ 이고 목표 염기의 98%가 $>100 \times$ 를 커버한다고 말할 수 있다.
- [0226] "염기 호출 품질 점수" 또는 "Q 점수"라는 용어는, 단일 서열분석된 염기가 정확한 확률에 반비례하여 0 내지 20 범위의 PHRED-스케일 확률을 지칭한다. 예를 들어, Q가 20인 T 염기 호출은, 신뢰도 P-값이 0.01인 경우 올바른 것으로 간주될 수 있다. $Q < 20$ 인 모든 염기 호출은 품질이 낮은 것으로 간주되어야 하며, 변이체를 지지하는 서열분석된 리드의 상당 부분이 품질이 낮은 것으로 식별된 임의의 변이체는 잠재적 위양성으로 간주되어야 한다.
- [0227] "변이체 리드" 또는 "변이체 리드 수"라는 용어는 변이체의 존재를 지지하는 서열분석된 리드의 수를 지칭한다.
- [0228] **서열분석 프로세스**
- [0229] 본 명세서에 설명된 구현예들은, 서열 변이를 식별하기 위해 핵산 서열을 분석하는 데 적용될 수 있다. 구현예들은, 유전자 위치/좌위의 잠재적 변이체/대립유전자를 분석하고 유전 좌위의 유전자형을 결정하거나 다시 말하면 좌위를 위한 유전자형 호출을 제공하는 데 사용될 수 있다. 예를 들어, 핵산 서열은 미국 특허출원 공개번호 2016/0085910 및 미국 특허출원 공개번호 2013/0296175에 기술된 방법 및 시스템에 따라 분석될 수 있으며, 이들 문헌의 완전한 주제 전문은 본 명세서에서 인용된다.
- [0230] 일 구현예에서, 서열분석 프로세스는 DNA와 같은 핵산을 포함하거나 포함하는 것으로 의심되는 샘플을 수신하는 단계를 포함한다. 샘플은, 동물(예를 들어, 인간), 식물, 박테리아 또는 진균과 같이 공지된 또는 알려지지 않은 공급원으로부터 유래될 수 있다. 샘플은 공급원으로부터 직접 취해질 수 있다. 예를 들어, 혈액 또는 타액은 개인으로부터 직접 취해질 수 있다. 대안으로, 샘플은 공급원으로부터 직접 취득되지 않을 수 있다. 이어서, 하나 이상의 프로세서는 서열분석을 위해 샘플을 준비하도록 시스템에 지시한다. 준비는 외부 물질을 제거 및/또는 소정의 물질(예를 들어, DNA)을 격리하는 것을 포함할 수 있다. 생물학적 샘플은 특정 분석에 대한 피처를 포함하도록 준비될 수 있다. 예를 들어, 생물학적 샘플은 합성에 의한 서열분석(SBS)을 위해 준비될 수 있다. 소정의 구현예에서, 준비는 게놈의 소정의 영역의 증폭을 포함할 수 있다. 예를 들어, 준비는 STR 및/또는 SNP를 포함하는 것으로 알려진 미리 결정된 유전 좌위를 증폭시키는 것을 포함할 수 있다. 유전 좌위는 미리 결정된 프라이머 서열을 사용하여 증폭될 수 있다.
- [0231] 다음에, 하나 이상의 프로세서는 시스템이 샘플을 서열분석하도록 지시한다. 서열분석은 공지된 다양한 서열분석 프로토콜을 통해 수행될 수 있다. 특정 구현예에서, 서열분석은 SBS를 포함한다. SBS에서, 복수의 형광-표지된 뉴클레오타이드는, 광학 기관의 표면(예를 들어, 유동 세포의 채널을 적어도 부분적으로 정의하는 표면)에 존재하는 증폭된 DNA의 복수의 클러스터(수백만의 클러스터일 수 있음)를 서열분석하는 데 사용된다. 유동 세포는, 유동 세포가 적절한 유동 세포 홀더 내에 배치되는 서열분석을 위한 핵산 샘플을 포함할 수 있다.
- [0232] 핵산은, 핵산이 알려지지 않은 표적 서열에 인접한 공지된 프라이머 서열을 포함하도록 준비될 수 있다. 제1 SBS 서열분석 사이클을 개시하기 위해, 하나 이상의 상이하게 표지된 뉴클레오타이드, 및 DNA 폴리머라제 등이 유체 흐름 서브시스템에 의해 유동 세포 내로/유동 세포를 통해 흐를 수 있다. 단일 유형의 뉴클레오타이드가 한 번에 추가될 수 있거나, 서열분석 절차에 사용되는 뉴클레오타이드는 가역적 종결 특성을 갖도록 특별히 설계될 수 있으며, 따라서 서열분석 반응의 각 사이클이 여러 유형의 표지된 뉴클레오타이드(예를 들어, A, C, T, G)가 존재하는 가운데 동시에 일어날 수 있게 한다. 뉴클레오타이드는 형광단과 같은 검출가능한 표지 모이어티를 포함할 수 있다. 4개의 뉴클레오타이드가 함께 혼합되는 경우, 폴리머라제는 혼입할 정확한 염기를 선택할 수 있고, 각 서열은 단일 염기에 의해 확장된다. 비혼합 뉴클레오타이드는 유동 세포를 통해 세척액을 흐르게 함으로써 세척될 수 있다. 하나 이상의 레이저가 핵산을 자극하고 형광을 유발할 수 있다. 핵산으로부터 방출되는 형광은 혼입된 염기의 형광단에 기초하고, 상이한 형광단은 상이한 파장의 방출 광을 방출할 수 있다. 디블로킹 시약을 유동 세포에 첨가하여 확장 및 검출된 DNA 가닥으로부터 가역적 종결자 그룹을 제거할 수 있다. 이어서, 디블로킹 시약은 유동 세포를 통해 세척 용액을 흐르게 함으로써 세척될 수 있다. 이어서, 유동 세포는,

상기 기재된 바와 같이 표지된 뉴클레오타이드의 도입으로 시작하여 서열분석의 추가 사이클에 대하여 준비된다. 서열분석 실행을 완료하기 위해 유체 및 검출 동작을 여러 번 반복할 수 있다. 서열분석 방법의 예는, 예를 들어, 문헌[Bentley et al., Nature 456:53-59 (2008)]; 국제출원공개번호 WO 04/018497; 미국 특허번호 7,057,026; 국제출원공개번호 WO 91/06678; 국제출원공개번호 WO 07/123744; 미국 특허번호 7,329,492; 미국 특허번호 7,211,414; 미국 특허번호 7,315,019; 미국 특허번호 7,405,281; 및 미국 특허출원 공개번호 2008/0108082에 개시되어 있으며, 이들 문헌의 각각은 본 명세서에 참고로 인용된다.

[0233] 일부 구현예에서, 핵산은, 표면에 부착될 수 있고 서열분석 전에 또는 서열분석 동안 증폭될 수 있다. 예를 들어, 증폭은, 브리지 증폭을 이용하여 수행되어 표면 상에 핵산 클러스터를 형성할 수 있다. 유용한 브리지 증폭 방법은, 예를 들어, 미국 특허번호 5,641,658; 미국 특허출원 공개번호 2002/0055100; 미국 특허번호 제 7,115,400호; 미국 특허출원 공개번호 2004/0096853; 미국 특허출원 공개번호 2004/0002090; 미국 특허출원 공개번호 2007/0128624; 및 미국 특허출원 공개번호 2008/0009420에 개시되어 있으며, 이들 문헌 각각의 전문은 본 명세서에 참고로 인용된다. 표면 상의 핵산을 증폭시키는 또 다른 유용한 방법은, 예를 들어, Lizardi 등의 Nat. Genet. 19:225-232 (1998) 및 미국 특허출원 공개번호 2007/0099208 A1에 개시된 바와 같은 롤링 서클 증폭(RCA)이며, 이들 문헌 각각은 본 명세서에 참고로 인용된다.

[0234] SBS 프로토콜의 일례는, 예를 들어, 국제공개번호 WO 04/018497, 미국 특허출원 공개번호 2007/0166705A1, 및 미국 특허번호 제7,057,026호에 기재된 바와 같이, 제거가능한 3' 블럭을 갖는 변형된 뉴클레오타이드를 이용하여, 이들 문헌 각각은 본 명세서에 참고로 인용된다. 예를 들어, SBS 시약의 반복 사이클은, 예를 들어, 브리지 증폭 프로토콜의 결과로 목표 핵산이 부착된 유동 세포로 전달될 수 있다. 핵산 클러스터는 선형화 용액을 사용하여 단일 가닥 형태로 전환될 수 있다. 선형화 용액은, 예를 들어, 각 클러스터의 하나의 가닥을 절단할 수 있는 제한 엔도뉴클레아제를 함유할 수 있다. 다른 절단 방법이, 특히, 화학적 절단(예를 들어, 과옥소산염에 의한 디올 연결의 절단), 엔도뉴클레아제에 의한 절단에 의한 염기성 부위의 절단(예를 들어, 미국 매사추세츠 인스위치에 소재하는 NEB사에 의해 공급되는 바와 같은 'USER', 부품 번호 M5505S), 열이나 알칼리에 대한 노출, 테옥시리보뉴클레오타이드로 달리 구성된 증폭 산물로 혼입된 리보뉴클레오타이드의 절단, 광화학적 절단, 또는 펩타이드 링커의 절단을 포함하여, 효소 또는 니킹 효소를 제한하기 위한 대체 방법으로서 사용될 수 있다. 선형화 동작 후에, 서열분석 프라이머를 서열분석될 목표 핵산에 혼성하기 위한 조건 하에서 서열분석 프라이머를 유동 세포로 전달할 수 있다.

[0235] 이어서, 유동 세포를, 단일 뉴클레오타이드 첨가에 의해 각각의 목표 핵산에 혼성화된 프라이머를 확장시키는 조건 하에서 제거가능한 3' 블럭 및 형광 표지를 갖는 변형된 뉴클레오타이드를 갖는 SBS 확장 시약과 접촉시킬 수 있다. 일단 변형된 뉴클레오타이드가 서열분석되는 템플릿의 영역에 상보적인 성장하는 폴리뉴클레오타이드 쇄에 혼합되었다면, 추가 서열 확장을 지시하기 위해 이용 가능한 유리 3'-OH기가 없기 때문에, 단일 뉴클레오타이드만이 각 프라이머에 첨가되고, 따라서, 중합효소가 추가의 뉴클레오타이드를 첨가할 수 없다. SBS 확장 시약은, 제거될 수 있고 방사선으로 여기 상태에서 샘플을 보호하는 성분을 포함하는 스캔 시약으로 교체될 수 있다. 스캔 시약을 위한 예시적인 성분은 미국 특허출원 공개번호 2008/0280773 A1 및 미국 특허출원번호 13/018,255에 기재되어 있으며, 이들 문헌 각각은 본 명세서에 참고로 인용된다. 이어서, 확장된 핵산은 스캔 시약의 존재 하에서 형광 검출될 수 있다. 일단 형광이 검출되었다면, 사용된 블로킹 기에 적합한 디블로킹 시약을 사용하여 3' 블럭을 제거할 수 있다. 각 블로킹 기에 유용한 예시적인 디블로킹 시약(deblock reagent)은 WO004018497, US 2007/0166705 A1, 및 미국 특허번호 7,057,026에 기재되어 있으며, 이들 문헌 각각은 본 명세서에 참고로 인용된다. 디블로킹 시약을 세척하여, 목표 핵산을, 이제 추가의 뉴클레오타이드의 첨가를 위한 성분인 3'-OH기를 갖는 확장된 프라이머에 혼성되게 한다. 따라서, 하나 이상의 동작 사이클에서의 선택적 세척에 의해 확장 시약, 스캔 시약, 및 디블로킹 시약을 첨가하는 주기는, 원하는 서열이 취득될 때까지 반복될 수 있다. 상기 사이클은, 각각의 변형된 뉴클레오타이드 각각이 특정 염기에 상응하는 것으로 공지된 상이한 표지로 부착될 때 사이클당 단일 확장 시약 전달 동작을 사용하여 수행될 수 있다. 상이한 표지는, 각각의 혼입 동작 동안 첨가되는 뉴클레오타이드의 구별을 용이하게 한다. 대안으로, 각 사이클은, 확장 시약 전달의 개별 동작 및 후속하는 시약 전달 및 검출의 개별 동작을 포함할 수 있으며, 이 경우, 2개 이상의 뉴클레오타이드가 동일한 표지를 가질 수 있고 공지된 전달 순서에 기초하여 구별될 수 있다.

[0236] 서열분석 동작을 특정 SBS 프로토콜과 관련하여 전술하였지만, 임의의 다양한 다른 분자 분석 중 임의의 것을 서열분석하기 위한 다른 프로토콜이 필요에 따라 수행될 수 있음을 이해할 것이다.

[0237] 이어서, 시스템의 하나 이상의 프로세서는 후속 분석을 위해 서열분석 데이터를 수신한다. 서열분석 데이터는 .BAM 파일과 같이 다양한 방식으로 포맷화될 수 있다. 서열분석 데이터는 예를 들어 다수의 샘플 리드를 포함할

수 있다. 서열분석 데이터는 뉴클레오타이드의 상응하는 샘플 서열을 갖는 복수의 샘플 리드를 포함할 수 있다. 하나의 샘플 리드만이 설명되고 있지만, 서열분석 데이터는 예를 들어 수백, 수천, 수십만 또는 수백만 개의 샘플 리드를 포함할 수 있음을 이해해야 한다. 상이한 샘플 리드는 상이한 수의 뉴클레오타이드를 가질 수 있다. 예를 들어, 샘플 리드는 10개의 뉴클레오타이드 내지 약 500개의 뉴클레오타이드 이상의 범위에 있을 수 있다. 샘플 리드들은 공급원(들)의 전체 계능에 걸쳐 이어질 수 있다. 일례로, 샘플 리드값은, STR이 의심되거나 SNP가 의심되는 그러한 유전 좌위와 같은 미리 정해진 유전 좌위에 관한 것이다.

[0238] 각각의 샘플 리드는, 샘플 서열, 샘플 분획물 또는 표적 서열이라고 칭할 수 있는 뉴클레오타이드들의 서열을 포함할 수 있다. 샘플 서열은, 예를 들어, 프라이머 서열, 측면 서열, 및 표적 서열을 포함할 수 있다. 샘플 서열 내의 뉴클레오타이드의 수는 30, 40, 50, 60, 70, 80, 90, 100 이상을 포함할 수 있다. 일부 구현예에서, 하나 이상의 샘플 리드(또는 샘플 서열)는, 적어도 150개의 뉴클레오타이드, 200개의 뉴클레오타이드, 300개의 뉴클레오타이드, 400개의 뉴클레오타이드, 500개의 뉴클레오타이드 이상을 포함한다. 일부 구현예에서, 샘플 리드는 1000개를 초과하는 뉴클레오타이드, 2000개 이상의 뉴클레오타이드를 포함할 수 있다. 샘플 리드(또는 샘플 서열)는 한쪽 또는 양쪽 말단에 프라이머 서열을 포함할 수 있다.

[0239] 다음에, 하나 이상의 프로세서는 서열분석 데이터를 분석하여 잠재적 변이체 호출(들) 및 샘플 변이체 호출(들)의 샘플 변이체 빈도를 취득한다. 상기 동작은, 또한, 변이체 호출 애플리케이션 또는 변이체 호출자라고 칭할 수 있다. 따라서, 변이체 호출자는 변이체를 식별 또는 검출하고, 변이체 분류자는 검출된 변이체를 체세포 또는 생식세포로서 분류한다. 대안의 변이체 호출자는 본 발명의 구현예에 따라 이용될 수 있고, 여기서 상이한 변이체 호출자들은, 관심 샘플의 피처 등에 기초하여 수행되는 서열분석 동작의 유형에 기초하여 사용될 수 있다. 변이체 호출 애플리케이션의 비제한적인 일례는, <https://github.com/Illumina/Pisces>에 호스팅되고 Dunn, Tamsen & Berry, Gwenn & Emig-Agius, Dorothea & Jiang, Yu & Iyer, Anita & Udar, Nitin & Stromberg, Michael. (2017). Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller. 595-595. 10.1145/3107411.3108203 기사에 개시된 일루미나사(Illumina Inc.)(캘리포니아주 샌디에고 소재)에 의한 Pisces™이 있으며, 이 문헌의 완전한 주제 전문은 명백하게 본 명세서에 참고로 인용된다.

[0240] 이러한 변이체 호출 애플리케이션은 다음과 같이 4개의 순차적으로 실행되는 모듈을 포함할 수 있다:

[0241] (1) 파이스즈 리드 스티치(Pisces Read Stitcher): BAM(동일한 분자의 리드 1과 리드 2)의 페어드 리드들을 컨센서스 리드로 스티칭함으로써 노이즈를 감소시킨다. 출력은 스티칭된 BAM이다.

[0242] (2) 파이스즈 변이체 호출자(Pisces Variant Caller): 작은 SNV, 삽입, 및 삭제를 호출한다. 파이스즈는, 리드 경계, 기본 필터링 알고리즘, 및 간단한 푸아송 기반 변이체 신뢰도 점수매김 알고리즘에 의해 분해된 변이체들을 병합하는 변이체 허탈 알고리즘을 포함한다. 출력은 VCF이다.

[0243] (3) 파이스즈 변이체 품질 재조정기(Pisces Variant Quality Recalibrator: VQR): 변이체 호출이 열적 손상 또는 FFPE 탈아민에 연관된 패턴을 압도적으로 추종하는 경우, VQR 단계는 의심되는 변이체 호출의 변이체 Q 점수를 다운그레이드한다. 출력은 조정된 VCF이다.

[0244] (4) 파이스즈 변이체 위상기(Pisces Variant Phase)(Scylla): 리드-백 그리디(read-backed greedy) 클러스터링 방법을 사용하여 작은 변이체를 클론 하위모집단의 복잡한 대립유전자들로 조립한다. 이를 통해 하향 틀에 의한 기능적 결과를 더욱 정확하게 결정할 수 있다. 출력은 조정된 VCF이다.

[0245] 부가적으로 또는 대안적으로, 동작은, <https://github.com/Illumina/strelka>에 호스팅되고 T Saunders, Christopher & Wong, Wendy & Swamy, Sajani & Becq, Jennifer & J Murray, Lisa & Cheetham, Keira. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England). 28. 1811-7. 10.1093/bioinformatics/bts271 기사에 개시된 Illumina Inc.에 의한 변이체 호출 애플리케이션 Strelka™을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 인용된다. 게다가, 부가적으로 또는 대안적으로, 동작은, <https://github.com/Illumina/strelka>에 호스팅되고 Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Beyter, D., Krusche, P., and Saunders, C.T. (2017). Strelka2: Fast and accurate variant calling for clinical sequencing applications 기사에 개시된 일루미나사에 의한 변이체 호출 애플리케이션 Strelka2™을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 인용된다. 게다가, 부가적으로 또는 대안적으로, 동작은, <https://github.com/Illumina/Nirvana/wiki>에 호스팅되고 Stromberg, Michael & Roy, Rajat & Lajugie, Julien & Jiang, Yu & Li, Haochen & Margulies,

Elliott. (2017). Nirvana: Clinical Grade Variant Annotator. 596-596. 10.1145/3107411.3108204 기사에 개시된 일루미나사에 의한 변이체 주석/호출 툴 NirvanaTM을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 인용된다.

- [0246] 이러한 변이체 주석/호출 툴은, 아래와 같이 Nirvana에 개시된 알고리즘 기술 등의 상이한 알고리즘 기술을 적용할 수 있다:
- [0247] a. 간격 어레이를 사용하여 중복되는 모든 전사를 식별: 기능적 주석의 경우, 변이체와 중복되는 모든 전사를 식별할 수 있고 간격 트리를 사용할 수 있다. 그러나, 일련의 간격은 정적일 수 있으므로, 이를 간격 어레이에 추가로 최적화할 수 있었다. 간격 트리는 $O(\min(n, k \lg n))$ 시간으로 모든 중복되는 전사를 리턴하며, 여기서, n 은 트리의 간격의 수이고, k 는 중복되는 간격의 수이다. 실제로, k 는 대부분의 변이체에 대한 n 에 비해 실제로 작기 때문에, 간격 트리의 유효 런타임은 $O(k \lg n)$ 이다. 제1 중복 간격만 찾으려 하고 이어서 남아 있는 $(k-1)$ 개를 통해 열거 처리하도록 모든 간격이 정렬된 어레이로 저장되는 간격 어레이를 생성함으로써 $O(\lg n + k)$ 으로 개선하였다.
- [0248] b. CNV/SV (Yu): 카피 수 변이 및 구조 변이체에 대한 주석을 제공할 수 있다. 작은 변이체의 주석과 유사하게, sv 및 또한 이전에 보고된 구조 변이체와 중복되는 전사체는 온라인 데이터베이스에서 주석 표시될 수 있다. 작은 변이체와는 달리, 너무 많은 전사체가 큰 sv와 중복되므로 모든 중복되는 전사체에 주석을 달 필요는 없다. 대신, 부분 중첩 유전자에 속하는 모든 중복되는 전사체에 주석을 달 수 있다. 구체적으로, 이들 전사체에 대해, 영향을 받은 인트론, 엑손, 및 구조 변이체에 의해 야기된 결과가 보고될 수 있다. 모든 중복 전사체를 출력할 수 있는 옵션을 사용할 수 있지만, 유전자 심볼, 전사체와 정규적으로 중복되는지 또는 부분적으로 중복되는지의 플래그 등의 이러한 전사체에 대한 기본 정보를 보고할 수 있다. 각각의 SV/CNV에 대해, 이들 변이체 및 해당 빈도가 다른 모집단에서 연구되었는지를 아는 것도 중요하다. 따라서, 1000개의 게놈, DGV, 및 ClinGen 과 같이 외부 데이터베이스에서 중복되는 sv를 보고하였다. 어떤 sv가 중복되는지를 결정하도록 임의의 컷오프를 사용하는 것을 피하기 위해, 대신에 모든 중복되는 전사체를 사용할 수 있고 상호 중복을 계산할 수 있으며, 즉, 중복되는 길이를 이들 두 개의 sv의 길이의 최소값으로 나눌 수 있다.
- [0249] c. 보충 주석 보고: 보충 주석은 소형 및 구조 변이체(SV)의 두 가지 유형이 있다. SV는, 간격으로서 모델링될 수 있으며, 전술한 간격 어레이를 사용하여 중복되는 SV를 식별할 수 있다. 소형 변이체는 점으로서 모델링되며 위치 및 (선택적으로) 대립유전자에 의해 일치된다. 이처럼, 이들은 이진-검색-유사 알고리즘을 사용하여 검색된다. 보충 주석 데이터베이스는 상당히 클 수 있으므로, 검색체 위치를 보충 주석이 상주하는 파일 위치에 맵핑하기 위해 훨씬 작은 인덱스가 생성된다. 인덱스는, 위치를 사용하여 이진 검색될 수 있는 (검색체 위치와 파일 위치로 구성된) 객체의 정렬된 어레이이다. 인덱스 크기를 작게 유지하기 위해, (최대 특정 개수의) 다수의 위치가, 제1 위치에 대한 값과 후속 위치에 대한 델타만을 저장하는 하나의 객체로 압축된다. 이진 검색을 사용하므로, 런타임은 $O(\lg n)$ 이며, 여기서 n 은 데이터베이스의 항목 수이다.
- [0250] d. VEP 캐시 파일
- [0251] e. 전사 데이터베이스: 전사 캐시(캐시) 및 보충 데이터베이스(SAdb) 파일은 전사 및 보충 주석과 같은 데이터 객체의 직렬화된 덤프이다. 본 발명자들은 Ensembl VEP 캐시를 캐시를 위한 본 발명자들의 데이터 소스로서 사용한다. 캐시를 생성하기 위해, 모든 전사체가 간격 어레이에 삽입되고, 어레이의 최종 상태가 캐시 파일에 저장된다. 따라서, 주석 표시 중에는, 미리 연산된 간격 어레이를 로딩하고 이에 대한 검색을 수행하면 된다. (전술한 바와 같이) 캐시가 메모리에 로딩되고 검색이 매우 빠르므로, Nirvana에서 중복되는 전사체를 찾는 것이 매우 빠르다(총 런타임의 1% 미만으로 프로파일링되었는가?).
- [0252] f. 보충 데이터베이스: SAdb용 데이터 공급원들은 보충 자료에서 열거되어 있다. 소형 변이체에 대한 SAdb는, (참조명과 위치에 의해 식별되는) 데이터베이스의 각 객체가 모든 관련된 보충 주석을 보유하도록 모든 데이터 공급원의 k-way 병합에 의해 생성된다. 데이터 소스 파일을 구문 분석하는 동안 발생하는 문제는 Nirvana의 홈페이지에 자세히 설명되어 있다. 메모리 사용을 제한하기 위해, SA 인덱스만이 메모리에 로딩된다. 이 인덱스에 의해, 보충 주석에 대한 파일 위치를 빠르게 찾을 수 있다. 그러나, 데이터를 디스크에서 가져와야 하므로, 보충 주석 추가는 Nirvana의 최대 병목 현상(전체 런타임의 ~30%로 프로파일링됨)으로서 식별되었다.
- [0253] g. 결과 및 서열 온톨로지: Nirvana의 기능 주석(제공된 경우)은 서열 온톨로지(SO)(<http://www.sequenceontology.org/>) 지침을 따른다. 경우에 따라, 현재 SO에서 문제를 식별하고 SO 팀과 협력하여 주석 상태를 개선할 수 있는 기회가 있었다.

[0254] 이러한 변이체 주식 틀은 전처리를 포함할 수 있다. 예를 들어, Nirvana에는, ExAC, EVS, 1000 게놈 프로젝트, dbSNP, ClinVar, Cosmic, DGV, 및 ClinGen과 같은 외부 데이터 공급원의 많은 주식이 포함되었다. 이러한 데이터베이스를 최대한 활용하려면, 데이터베이스로부터 정보를 삭제해야 한다. 상이한 데이터 공급원으로부터 발생하는 상이한 충돌을 처리하기 위해 상이한 전략을 구현하였다. 예를 들어, 동일한 위치와 대체 대립유전자에 대해 다수의 dbSNP 엔트리가 있는 경우, 모든 ID를 쉼표로 구분된 ID 목록에 입력하고, 동일한 대립유전자에 대해 상이한 CAF 값을 가진 다수의 엔트리가 있는 경우, 제1 CAF 값을 사용한다. ExAC 엔트리와 EVS 엔트리가 충돌하는 경우, 샘플 계수치의 수를 고려하고, 샘플 계수치가 높은 엔트리를 사용한다. 1000개의 게놈 프로젝트에서, 충돌 대립유전자의 대립유전자 빈도를 제거하였다. 또 다른 문제는 부정확한 정보이다. 주로 1000개의 게놈 프로젝트로부터 대립유전자 빈도 정보를 추출했지만, GRCh38의 경우, 정보 필드에 보고된 대립유전자 빈도가 유전자형을 사용할 수 없는 샘플을 배제하지 않아서, 모든 샘플에 대하여 사용할 수 없는 변이체의 빈도가 감소된다는 점에 주목하였다. 본 발명자들의 주식의 정확성을 보장하기 위해, 본 발명자들은 모든 개별 수준 유전자형을 사용하여 실제 대립유전자 빈도를 컴퓨팅한다. 본 발명자들이 알고 있는 바와 같이, 동일한 변이체는 상이한 정렬에 기초하여 상이한 표현을 가질 수 있다. 이미 식별된 변이체에 대한 정보를 정확하게 보고할 수 있으려면, 다른 자원들로부터의 변이체를 전처리하여 일관성 있는 표현을 유지해야 한다. 모든 외부 데이터 공급원에 대해, 대립유전자를 트리밍하여 참조 대립유전자와 대체 대립유전자 모두에서 중복된 뉴클레오타이드를 제거하였다. ClinVar의 경우, 모든 변이체에 대해 5-프라임 정렬을 수행한 xml 파일을 직접 구문 분석하였으며, 이는 종종 vcf 파일에서 사용된다. 다른 데이터베이스에는 정보의 동일한 세트가 포함될 수 있다. 불필요한 중복을 피하기 위해, 일부 중복된 정보를 제거하였다. 예를 들어, 1000개의 게놈에서의 DGV의 변이체가 더욱 자세한 정보와 함께 이미 보고되었으므로, 데이터 공급원을 1000개의 게놈 프로젝트로서 갖는 이러한 변이체를 제거하였다.

[0255] 적어도 일부 구현예에 따르면, 변이체 호출 애플리케이션은 저 빈도 변이체, 생식세포 호출 등에 대한 호출을 제공한다. 비제한적인 예로서, 변이체 호출 애플리케이션은 종양 전용 샘플 및/또는 종양-정상 쌍을 이룬 샘플에서 실행될 수 있다. 변이체 호출 애플리케이션은, 단일 뉴클레오타이드 변이(SNV), 다중 뉴클레오타이드 변이(MNV), 인델 등을 검색할 수 있다. 변이체 호출 애플리케이션은, 변이체를 식별하면서 서열분석 또는 샘플 준비 오류로 인한 불일치를 필터링한다. 각각의 변이체에 대해, 변이체 호출자는, 참조 서열, 변이체의 위치 및 잠재적 변이체 서열(들)(예를 들어, A에서 C SNV로, 또는 AG에서 A 삭제)을 식별한다. 변이체 호출 애플리케이션은, 샘플 서열(또는 샘플 분획물), 참조 서열/분획물, 및 변이체 호출을 변이체가 존재함을 나타내는 표시로서 식별한다. 변이체 호출 애플리케이션은, 원시 분획물을 식별할 수 있고, 원시 분획물의 지정, 잠재적 변이체 호출을 검증하는 원시 분획물의 수, 지지 변이체가 발생한 원시 분획물 내의 위치, 및 기타 관련 정보를 출력할 수 있다. 원시 분획물의 비제한적인 예로는, 이중 스티치 분획물, 단일 스티치 분획물, 이중 언스티치 분획물, 및 단순한 언스티치 분획물을 포함한다.

[0256] 변이체 호출 애플리케이션은, .VCF 또는 .GVCF 파일과 같은 다양한 형식으로 호출을 출력할 수 있다. 단지 예로서, 변이체 호출 애플리케이션은 (예를 들어, MiSeq174; 시퀀서 기기 상에 구현될 때) MiSeqReporter 파이프라인에 포함될 수 있다. 선택적으로, 이 애플리케이션은 다양한 워크플로우로 구현될 수 있다. 분석은, 원하는 정보를 취득하도록 지정된 방식으로 샘플 리드를 분석하는 단일 프로토콜 또는 프로토콜들의 조합을 포함할 수 있다.

[0257] 이어서, 하나 이상의 프로세서는 잠재적 변이체 호출과 관련하여 유효성확인 동작을 수행한다. 유효성확인 동작은 이하에 설명되는 바와 같이 품질 점수 및/또는 계층적 테스트의 층에 기초할 수 있다. 유효성확인 동작이 잠재적 변이체 호출을 인증하거나 검증하면, 유효성확인 동작은 (변이체 호출 애플리케이션으로부터) 변이체 호출 정보를 샘플 보고서 생성기에 전달한다. 대안으로, 유효성확인 동작이 잠재적 변이체 호출을 무효화 또는 실격화하는 경우, 유효성확인 동작은, 대응하는 표시(예를 들어, 음성 표시기, 무 호출 표시기, 무효 호출 표시기)를 샘플 보고서 생성기에 전달한다. 유효성확인 동작은, 또한, 변이체 호출이 정확하거나 무효 호출 지정이 정확하다는 신뢰도와 관련된 신뢰도 점수를 전달할 수 있다.

[0258] 다음에, 하나 이상의 프로세서는 샘플 보고서를 생성하고 저장한다. 샘플 보고서는, 예를 들어, 샘플에 대한 복수의 유전 좌위에 관한 정보를 포함할 수 있다. 예를 들어, 미리 결정된 유전 좌위의 세트의 각각의 유전 좌위에 대해, 샘플 보고서는, 유전자형 호출을 제공하는 것, 유전자형 호출을 할 수 없음을 나타내는 것, 유전자형 호출의 확실성에 대한 신뢰 점수를 제공하는 것, 또는 하나 이상의 유전 좌위에 관한 분석법의 잠재적 문제를 나타내는 것 중 적어도 하나일 수 있다. 샘플 보고서는, 또한, 샘플을 제공한 개인의 성별을 나타낼 수 있고 및/또는 샘플이 다수의 공급원을 포함함을 나타낼 수 있다. 본 명세서에서 사용되는 바와 같이, "보고"는, 유전

좌위의 디지털 데이터(예를 들어, 데이터 파일) 또는 유전 좌위의 미리 결정된 세트 및/또는 유전 좌위 또는 유전 좌위의 세트의 인쇄된 보고서를 나타낼 수 있다. 따라서, 생성 또는 제공은, 데이터 파일의 생성 및/또는 샘플 보고서의 인쇄, 또는 샘플 보고서의 표시를 포함할 수 있다.

[0259] 샘플 보고서는, 변이체 호출이 결정되었지만 유효성확인되지 않았음을 나타낼 수 있다. 변이체 호출이 무효한 것으로 결정되면, 샘플 보고서는 변이체 호출을 유효성확인하지 않는 결정의 근거에 관한 추가 정보를 나타낼 수 있다. 예를 들어, 보고서의 추가 정보는, 원시 분획물의 설명 및 원시 분획물이 변이체 호출을 지지하거나 반박하는 정도(예를 들어, 계수치)를 포함할 수 있다. 추가적으로 또는 대안으로, 보고서의 추가 정보는 본 명세서에서 설명되는 구현예에 따라 취득된 품질 점수를 포함할 수 있다.

[0260] **변이체 호출 애플리케이션**

[0261] 본 명세서에 개시된 구현예들은 잠재적 변이체 호출을 식별하기 위해 서열분석 데이터를 분석하는 것을 포함한다. 변이체 호출은 이전에 수행된 서열분석 동작을 위해 저장된 데이터에 대해 수행될 수 있다. 추가적으로 또는 대안으로, 변이체 호출은 서열분석 동작이 수행되는 동안 실시간으로 수행될 수 있다. 각각의 샘플 리드 값은 상응하는 유전 좌위에 할당된다. 샘플 리드는, 샘플 리드의 뉴클레오타이드의 서열, 즉, 샘플 리드 내의 뉴클레오타이드의 서열(예를 들어, A, C, G, T)에 기초하여 대응하는 유전 좌위에 할당될 수 있다. 이 분석에 기초하여, 샘플 리드는, 특정 유전 좌위의 가능한 변이체/대립유전자를 포함하는 것으로서 지정될 수 있다. 샘플 리드는, 유전 좌위의 가능한 변이체/대립유전자를 포함하는 것으로서 지정된 다른 샘플 리드와 함께 수집(또는 집계 또는 비닝)될 수 있다. 할당 동작은, 또한, 샘플 리드가 특정 유전자 위치/좌위에 연관될 수 있는 것으로서 식별되는 호출 동작이라고 칭할 수 있다. 샘플 리드는, 샘플 리드를 다른 샘플 리드로부터 구별하는 뉴클레오타이드의 하나 이상의 식별 서열(예를 들어, 프라이머 서열)을 위치시키기 위해 분석될 수 있다. 보다 구체적으로, 식별 서열(들)은 다른 샘플 리드로부터의 샘플 리드를 특정 유전 좌위에 연관된 것으로서 식별할 수 있다.

[0262] 할당 동작은, 식별 서열의 일련의 n개의 뉴클레오타이드를 분석하여 식별 서열의 일련의 n개의 뉴클레오타이드가 하나 이상의 선택 서열과 효과적으로 일치하는지를 결정하는 것을 포함할 수 있다. 특정 구현예에서, 할당 동작은, 샘플 서열의 제1 n개의 뉴클레오타이드를 분석하여 샘플 서열의 제1 n개의 뉴클레오타이드가 하나 이상의 선택 서열과 효과적으로 일치하는지를 결정하는 것을 포함할 수 있다. 수 n은, 다양한 값을 가질 수 있으며, 프로토콜로 프로그래밍될 수 있거나 사용자에게 의해 입력될 수 있다. 예를 들어, 수 n은 데이터베이스 내에서 가장 짧은 선택 서열의 뉴클레오타이드의 수로서 정의될 수 있다. 수 n은 미리 결정된 수일 수 있다. 미리 결정된 수는, 예를 들어, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 또는 30개의 뉴클레오타이드일 수 있다. 그러나, 다른 구현예에서는 더 적거나 더 많은 뉴클레오타이드가 사용될 수 있다. 수 n은, 또한, 시스템의 사용자와 같은 개인에 의해 선택될 수 있다. 수 n은 하나 이상의 조건에 기초할 수 있다. 예를 들어, 수 n은 데이터베이스 내에서 가장 짧은 프라이머 서열의 뉴클레오타이드의 수 또는 지정된 수 중 작은 수로서 정의될 수 있다. 일부 구현예에서, 15개 미만의 임의의 프라이머 서열이 예외로 지정될 수 있도록, 15와 같은 n에 대한 최소값이 사용될 수 있다.

[0263] 일부 경우에, 식별 서열의 일련의 n개의 뉴클레오타이드는 선택 서열의 뉴클레오타이드와 정확하게 일치하지 않을 수 있다. 그럼에도 불구하고, 식별 시퀀스가 선택 시퀀스와 거의 동일한 경우 식별 시퀀스가 선택 시퀀스와 효과적으로 일치될 수 있다. 예를 들어, 식별 서열의 일련의 n개의 뉴클레오타이드(예를 들어, 제1 n개의 뉴클레오타이드)가 불일치의 지정된 수(예를 들어, 3) 이하 및/또는 시프트의 지정된 수(예를 들어, 2)를 갖는 선택 서열과 일치하는 경우, 유전 좌위에 대하여 샘플 리드가 호출될 수 있다. 각각의 불일치 또는 시프트가 샘플 리드와 프라이머 서열 간의 차로서 계수될 수 있도록 규칙이 확립될 수 있다. 차의 수가 지정된 수보다 작으면, 상응하는 유전 좌위(즉, 상응하는 유전 좌위에 할당됨)에 대해 샘플 리드가 호출될 수 있다. 일부 구현예에서, 샘플 리드의 식별 서열과 유전 로커에 연관된 선택 서열 간의 차의 수에 기초하여 일치 점수가 결정될 수 있다. 일치 점수가 지정된 일치 임계값을 통과하면, 선택 서열에 대응하는 유전 좌위가 샘플 리드를 위한 잠재적 좌위로서 지정될 수 있다. 일부 구현예에서는, 샘플 리드가 유전 좌위에 대해 호출되는지를 결정하기 위해 후속 분석이 수행될 수 있다.

[0264] 샘플 리드가 데이터베이스에서의 선택 서열들 중 하나와 효과적으로 일치하는 경우(즉, 전술한 바와 같이 정확히 일치하거나 거의 일치하는 경우), 샘플 리드는 선택 서열과 상관되는 유전 좌위에 할당되거나 지정된다. 이것은 유전 좌위 호출 또는 잠정 좌위 호출이라고 칭할 수 있으며, 여기서 샘플 리드는 선택 서열과 상관되는 유전 좌위에 대하여 호출된다. 그러나, 전술한 바와 같이, 샘플 리드는 하나보다 많은 유전 좌위에 대하여 호출될

수 있다. 이러한 구현예에서, 잠재적 유전 좌위들 중 하나에 대해서만 샘플 리드를 호출하거나 할당하도록 추가 분석이 수행될 수 있다. 일부 구현예에서, 참조 서열들의 데이터베이스와 비교되는 샘플 리드는 페어드-엔드 서열분석으로부터의 제1 리드이다. 페어드-엔드 서열분석을 수행할 때, 샘플 리드와 상관되는 제2 리드(원시 분획물을 나타냄)가 취득된다. 할당 후, 할당된 리드로 수행되는 후속 분석은, 할당된 리드를 위해 호출된 유전 좌위의 유형에 기초할 수 있다.

[0265] 다음에, 잠재적 변이체 호출을 식별하도록 샘플 리드가 분석된다. 무엇보다도, 분석 결과는, 잠재적 변이체 호출, 샘플 변이체 빈도, 참조 서열, 및 변이체가 발생한 게놈 서열 내의 위치를 식별한다. 예를 들어, 유전 좌위가 SNP를 포함하는 것으로 알려진 경우, 유전 좌위를 호출된 할당된 리드는 할당된 리드의 SNP를 식별하도록 분석을 거칠 수 있다. 유전 좌위가 다형성 반복 DNA 요소를 포함하는 것으로 알려진 경우, 할당된 리드는 샘플 리드 내의 다형성 반복 DNA 요소를 식별하거나 특성화하도록 분석될 수 있다. 일부 구현예에서, 할당된 리드가 STR 좌위 및 SNP 좌위와 효과적으로 일치하면, 경고 또는 플래그가 샘플 리드에 할당될 수 있다. 샘플 리드는 STR 유전 좌위와 SNP 좌위 모두로서 지정될 수 있다. 분석은, 할당된 리드의 서열 및/또는 길이를 결정하기 위해 정렬 프로토콜에 따라 할당된 리드를 정렬하는 것을 포함할 수 있다. 정렬 프로토콜은, 2013년 3월 15일에 출원된 국제 특허출원번호 PCT/US2013/030867(공개번호 WO 2014/142831)에 기술된 방법을 포함할 수 있으며, 이 문헌의 전문은 본 명세서에 참고로 인용된다.

[0266] 이어서, 하나 이상의 프로세서는, 원시 분획물을 분석하여 원시 분획물 내의 해당 위치에 지지 변이체가 존재하는지를 결정한다. 다양한 종류의 원시 분획물이 식별될 수 있다. 예를 들어, 변이체 호출자는, 초기 변이체 호출자를 유효성확인하는 변이체를 나타내는 원시 분획물의 유형을 식별할 수 있다. 예를 들어, 원시 분획물의 유형은, 이중 스티치 분획물, 단일 스티치 분획물, 이중 언스티치 분획물, 또는 단일 언스티치 분획물 나타낼 수 있다. 선택적으로, 전술한 예 대신 또는 추가로 다른 원시 분획물을 식별할 수 있다. 각 원시 분획물의 유형을 식별하는 것과 관련하여, 변이체 호출자는, 또한, 지지 변이체를 나타낸 원시 분획물 수의 계수치뿐만 아니라 지지 변이체가 발생한 원시 분획물 내의 위치도 식별한다. 예를 들어, 변이체 호출자는, 특정 위치 X에서 지지 변이체를 갖는 이중 스티치 분획물을 나타내도록 10개의 원시 분획물이 식별되었다는 표시를 출력할 수 있다. 변이체 호출자는, 또한, 특정 위치 Y에서 지지 변이체를 갖는 단일 언스티치 분획물을 나타내도록 원시 분획물의 5개 리스가 식별되었음을 출력할 수 있다. 변이체 호출자는, 또한, 참조 서열에 대응한 많은 원시 분획물을 출력할 수 있으므로, 다른 경우엔 관심 게놈 서열에서 잠재적 변이체 호출을 유효성확인하는 증거를 제공하는 지지 변이체를 포함하지 않았다.

[0267] 이어서, 지지 변이체가 발생한 위치뿐만 아니라 지지 변이체를 포함하는 원시 분획물의 계수치를 유지한다. 부가적으로 또는 대안적으로, (샘플 리드 또는 샘플 분획물의 잠재적 변이체 호출의 위치에 관한) 관심 위치에서 지지 변이체를 포함하지 않은 원시 분획물의 계수치를 유지할 수 있다. 부가적으로 또는 대안적으로, 참조 서열에 대응하고 잠재적 변이체 호출을 인증 또는 확인하지 않는 원시 분획물의 카운*를 유지할 수 있다. 결정된 정보는, 잠재적 변이체 호출을 지지하는 원시 분획물의 계수치와 유형, 원시 분획물의 지지 분산의 위치, 잠재적 변이체 호출을 지지하지 않는 원시 분획물의 수 등을 포함하여 변이체 호출 유효성확인 애플리케이션으로 출력된다.

[0268] 잠재적 변이체 호출이 식별되면, 프로세서는 잠재적 변이체 호출, 변이체 서열, 변이체 위치, 및 이에 연관된 참조 서열을 나타내는 표시를 출력한다. 변이체 호출은, 에러로 인해 호출 프로세스가 거짓 변이체를 식별할 수 있으므로 "잠재적" 변이체를 나타내도록 지정된다. 본 발명의 구현예에 따라, 잠재적 변이체 호출을 분석하여 거짓 변이체 또는 위양성을 감소 및 제거한다. 부가적으로 또는 대안적으로, 이 프로세서는, 샘플 리드에 연관된 하나 이상의 원시 분획물을 분석하고 원시 분획물에 연관된 해당 변이체 호출을 출력한다.

[0269] **유전체학에서의 심층 학습**

[0270] 유전자 변이는 많은 질환을 설명하는 데 도움이 될 수 있다. 모든 인간에게는 고유한 유전자 코드가 있으며, 개인 그룹 내에는 많은 유전자 변이체가 있다. 해로운 유전자 변이체의 대부분은 자연적인 선택에 의해 게놈으로부터 고갈되었다. 어떠한 유전자 변이체가 병원성이거나 해로운 가능성이 있는지를 식별하는 것이 중요하다. 이는, 연구자들이 병원성 유전자 변이체에 집중하고 많은 질환의 진단 및 치료 속도를 가속화하는 데 도움을 줄 수 있다.

[0271] 변이체의 특성 및 기능적 효과(예를 들어, 병원성)를 모델링하는 것은 유전체학 분야에서 중요하지만 어려운 과제이다. 기능적 게놈 서열분석 기술의 급속한 발전에도 불구하고, 변이체의 기능적 결과의 해석은, 세포 유형-특이적 전사 조절 시스템의 복잡성으로 인해 여전히 큰 도전으로 남아 있다.

- [0272] 병원성 분류자와 관련하여, 심층 신경망은, 다중 비선형 및 복잡한 변환 층을 사용하여 고레벨 피처를 연속적으로 모델링하는 유형의 인공 신경망이다. 심층 신경망은, 파라미터를 조정하기 위해 관찰된 출력과 예측 출력 간의 차이를 전달하는 역전파를 통해 피드백을 제공한다. 심층 신경망은, 큰 트레이닝 데이터세트의 가용성, 병렬 및 분산 연산의 힘, 및 정교한 트레이닝 알고리즘으로 진화했다. 심층 신경망은, 컴퓨터 비전, 음성 인식, 및 자연어 처리와 같은 다양한 영역에서 주요 발전을 촉진하였다.
- [0273] 컨볼루션 신경망(CNN) 및 순환 신경망(RNN)은 심층 신경망의 구성요소들이다. 컨볼루션 신경망은, 특히 컨볼루션층, 비선형 층, 및 풀링층을 포함하는 아키텍처로 이미지를 인식하는 데 성공하였다. 순환 신경망은, 퍼셉트론(perceptron), 장기 단기 메모리 유닛, 및 게이트 순환 유닛과 같은 빌딩 블록들 간의 주기적 연결을 통해 입력 데이터의 서열 정보를 이용하도록 설계되었다. 또한, 심층 시공간 신경망, 다차원 순환 신경망, 및 컨볼루션 자동 인코더와 같이 제한된 컨텍스트에 대해 다른 많은 응용 심층 신경망이 제안되어 왔다.
- [0274] 심층 신경망을 트레이닝하는 목적은 각 층의 가중치 파라미터를 최적화하는 것이며, 이는 간단한 피처들을 복잡한 피처로 점진적으로 결합하여 가장 적합한 계층적 표현이 데이터로부터 학습될 수 있도록 한다. 최적화 프로세스의 단일 사이클은 다음과 같이 구성된다. 먼저, 트레이닝 데이터세트가 주어지면, 순방향 패스는 각 층의 출력을 순차적으로 연산하고 기능 신호를 망을 통해 순방향으로 전파한다. 최종 출력층에서, 객관적인 손실 함수는 추론된 출력과 주어진 지표 간의 오류를 측정한다. 트레이닝 에러를 최소화하기 위해, 역방향 패스는 체인 규칙을 사용하여 에러 신호를 역전파하고 신경망 전체에 걸쳐 모든 가중치에 대한 그라디언트를 연산한다. 마지막으로, 가중치 파라미터는, 확률적 그라디언트 하강에 기반한 최적화 알고리즘을 사용하여 업데이트된다. 일괄 그라디언트 하강은 각각의 전체 데이터세트에 대한 파라미터 업데이트를 수행하는 반면, 확률적 그라디언트 하강은 데이터 예들의 작은 세트 각각에 대한 업데이트를 수행함으로써 확률적 근사치를 제공한다. 여러 최적화 알고리즘은 확률적 그라디언트 하강으로부터 비롯된다. 예를 들어, Adagrad 및 Adam 트레이닝 알고리즘은, 확률적 그라디언트 하강을 수행하면서 각 파라미터에 대한 그라디언트의 업데이트 빈도 및 모멘트를 기반으로 학습률을 각각 적응적으로 수정한다.
- [0275] 심층 신경망을 트레이닝하는 데 있어서 또 다른 핵심 요소는 규제화(regularization)인데, 이는 과적합을 피하고 이에 따라 우수한 일반화 성능을 달성하도록 의도된 전략을 가리킨다. 예를 들어, 가중치 감소는, 가중치 파라미터가 더 작은 절대값으로 수렴하도록 표적 손실 함수에 페널티 항을 추가한다. 드롭아웃은, 트레이닝 중에 신경망으로부터 숨겨진 유닛을 랜덤하게 제거하며 가능한 서브네트워크들의 앙상블로서 간주될 수 있다. 드롭아웃 기능을 향상시키기 위해, rnnDrop이라는 순환 신경망에 대하여 드롭아웃의 변형 및 새로운 활성화 함수 maxout이 제안되었다. 또한, 일괄 정규화는, 각 평균 및 분산을 파라미터로서 학습하고 미니-일괄 내의 각 활성화에 대한 스칼라 피처의 정규화를 통해 새로운 규제화 방법을 제공한다.
- [0276] 서열분석된 데이터가 다차원적이고 고차원적이라는 점을 감안할 때, 심층 신경망은, 광범위한 적용가능성과 향상된 예측 능력으로 인해 생물정보학 연구에 큰 가능성을 갖고 있다. 컨볼루션 신경망은, 모티프 발견, 병원성 변이체 식별, 및 유전자 발현 추론 등의 유전체학에서의 서열-기반 문제를 해결하도록 구성되었다. 컨볼루션 신경망은 DNA를 연구하는 데 특히 유용한 가중치 공유 전략을 사용하는데, 이는 중요한 생물학적 기능을 갖는 것으로 추정되는 DNA에서의 짧고 순환되는 국부 패턴인 서열 모티프를 포착할 수 있기 때문이다. 컨볼루션 신경망의 특징은 컨볼루션 필터를 사용하는 것이다. 정교하게 설계되고 수동으로 제작된 피처를 기반으로 하는 기존의 분류 방법과 달리, 컨볼루션 필터는 원시 입력 데이터를 지식의 정보 표현에 맵핑하는 프로세스와 유사한 피처의 적응적 학습을 수행한다. 이러한 의미에서, 컨볼루션 필터는, 이러한 필터들의 세트가 입력의 관련 패턴을 인식할 수 있고 트레이닝 과정 중에 스스로 업데이트할 수 있으므로, 일련의 모티프 스캐너 역할로서 기능한다. 순환 신경망은, 단백질 또는 DNA 서열과 같은 다양한 길이의 서열 데이터의 장거리 의존성을 포착할 수 있다.
- [0277] 따라서, 변이체의 병원성을 예측하기 위한 강력한 연산 모델은 기본 과학 및 변환 연구 모두에 대하여 많은 이점을 가질 수 있다.
- [0278] 현재, 희귀질환 환자의 25 내지 30%만이 단백질-코딩 서열의 검사로부터 분자 진단을 받는데, 이는 나머지 진단율이 비코딩인 게놈의 99%에 있을 수 있음을 시사한다. 본 명세서에서 본 발명자들은, 임의의 전구체-mRNA(pre-mRNA)의 전사체 서열로부터 스플라이스 접합부를 정확하게 예측하여, 비코딩 변이체의 스플라이스 변경 효과의 정확한 예측을 가능하게 하는 신규한 심층 학습망을 설명한다. 예측된 스플라이스 변경 결과를 갖는 동의 및 인트론 돌이변이는 RNA-seq에서 높은 속도로 유효성확인되고 인간 개체군에서 크게 해롭다. 예측된 스플라이스 변경 결과를 갖는 드 노보 돌이변이는 건강한 대조군에 비해 자폐증 및 지적장애 환자에 상당히 농축되고 이들 환자 28명 중 21명에서 RNA-seq 데이터에 대해 유효성확인된다. 본 발명자들은 희귀 유전 장애를 가진 환자에서

병원성 돌연변이의 9-11%가 이러한 이전에 정당하게 평가되지 않은 클래스의 질병 변이에 의해 초래된 것으로 추정한다.

[0279] 엑솜 서열분석은 희귀 유전 장애를 가진 환자와 가족의 임상 진단을 변화시켜왔으며, 첫 번째 테스트로 채택될 경우, 장기간의 진단 방랑의 시간과 비용을 크게 저감시킨다(Monroe et al., 2016; Stark et al., 2016; Tan et al., 2017). 그러나, 엑솜 서열분석의 진단율은 희귀 유전질환 코호트에서 25-30%로서, 대다수의 환자들을 엑솜 및 마이크로어레이 복합 테스트 후에도 진단 없이 방치한다(Lee et al., 2014; Trujillano et al., 2017; Yang et al., 2014). 비코딩 영역은 유전자 조절에서 중대한 역할을 수행하고, 인간 복합 질환의 비편향적 게놈 전체 연합 연구에서 발견된 원인 질병 좌위의 90%를 설명하며(Ernst et al., 2011; Farh et al., 2015; Maurano et al., 2012), 이는 침투성 비코딩 변이체가 희귀 유전질환에서 원인 돌연변이의 중대한 책임을 또한 설명할 수 있을 것임을 시사한다. 사실, 필수적인 GT 및 AG 스플라이스 다이뉴클레오타이드의 외측에 놓여짐에도 불구하고 mRNA 스플라이싱의 정규 패턴을 파괴하며 크립틱 스플라이스 변이체로도 종종 불리우는 침투성 비코딩 변이체는 희귀 유전질환에서 중요한 역할을 하는 것으로 오랫동안 인식되어 왔다(Cooper et al., 2009; Padgett, 2012; Scotti and Swanson, 2016; Wang and Cooper, 2007). 그러나, 크립틱 스플라이스 돌연변이는, 스플라이싱 코드에 대한 불완전한 이해와 그 결과로, 필수적인 GT 및 AG 다이뉴클레오타이드 외측의 스플라이스 변경 변이체를 정확히 식별하는 것의 어려움 때문에, 임상 실무에서 종종 간과된다(Wang and Burge, 2008).

[0280] 최근, RNA-seq가 멘델 유전질환에서 스플라이싱 이상을 검출하기 위한 촉망받는 분석방법으로 등장했으나(Cummings et al., 2017; Kremer et al., 2017), 여태껏 임상 환경에서 그 유용성은 관련 세포 유형이 알려져 있고 생체검사를 이용 가능한 소수의 케이스들로 한정될 뿐이다. 잠재적인 스플라이스 변경 변이체의 고능률 선별 분석(Soemedi et al., 2017)은 스플라이싱 변이의 특성화를 확장시켰으나, 유전질환에서 무작위의 드 노보 돌연변이를 평가하기에는 덜 실용적인데, 이는 스플라이스 변경 돌연변이가 발생하는 게놈 공간이 지극히 넓기 때문이다. 임의의 전구체-mRNA 서열로부터 스플라이싱을 일반적으로 예측하면 비코딩 변이체의 스플라이스 변경 결과들의 정밀한 예측이 잠재적으로 가능해져서, 유전질환을 가진 환자들의 진단을 현저하게 향상시킬 것이다. 오늘날까지, 스플라이소좀의 특이성에 접근하는, 원시 서열로부터의 스플라이싱에 대한 일반적인 예측 모델은, 핵심 스플라이싱 모티프의 서열 특성 모델링(Yeo and Burge, 2004), 엑손 스플라이스 인핸서 및 침투자의 특성화(Fairbrother et al., 2002; Wang et al., 2004), 및 카세트 엑손 인클루전(cassette exon inclusion)의 예측(Barash et al., 2010; Jha et al., 2017; Xiong et al., 2015)와 같은 특정 분야의 진전에도 불구하고, 여전히 파악하기 어렵다.

[0281] 긴 전구체-mRNA를 성숙한 전사체로 스플라이싱하는 것은 그 정밀도와 스플라이스 변경 돌연변이의 임상적 가혹함에서 볼 때 놀랍지만, 세포 기계가 그 특이성을 결정짓게 되는 메커니즘은 불완전하게 이해된 상태이다. 본 명세서에서 본 발명자들은 스플라이소좀의 정확성에 접근하는 심층 학습망을 인 실리코(in silico)로 트레이닝 시켜, 전구체-mRNA 서열로부터 엑손-인트론 경계를 95% 정확도로 식별하고, 기능성 크립틱 스플라이스 돌연변이를 RNA-seq 상에서 80%를 넘는 유효성확인율로 예측한다. 스플라이싱을 변경할 것으로 예측된 비코딩 변이체는 인간 개체군에서 매우 유해하여, 새롭게 생성된 크립틱 스플라이스 돌연변이의 80%가, 타 클래스의 단백질 절단 변이의 충격과 유사한, 부정적 선택을 겪게 된다. 자폐증 및 지적장애 환자에서 드 노보 크립틱 스플라이스 돌연변이는 단백질 절단 돌연변이에 의해 반복적으로 돌연변이가 발생하는 동일한 유전자를 적중시켜, 추가적인 후보 질환 유전자들을 발견할 수 있도록 한다. 본 발명자들은 희귀 유전 장애를 가진 환자들에서 침투성 원인 돌연변이의 최대 24%가 이러한 이전에 정당하게 평가되지 않은 클래스의 질병 변이 때문인 것으로 추정하며, 이에 따라 임상 서열분석 분야에서 비코딩인 게놈의 99%의 해석을 향상시킬 필요성이 강조된다.

[0282] 임상 엑솜 서열분석은 희귀 유전 장애를 가진 환자 및 가족에 대한 진단을 혁신해왔고, 첫 번째 테스트로 채택될 경우, 장기간의 진단 방랑의 시간과 비용을 크게 저감시킨다. 그러나, 엑솜 서열분석의 진단율은 희귀 질환 환자 및 그 부모의 다수의 큰 코호트에서 25-30%로 보고되었으며, 대다수의 환자들을 엑솜 및 마이크로어레이 복합 테스트 후에도 진단 없이 방치한다. 비코딩 게놈은 유전자 조절에서 매우 활동적이고, 비코딩 변이체는 일반적인 질환에 대한 ~90%의 GWAS 적중률을 설명하므로, 이는 비코딩 게놈에서 희귀 변이체가 희귀 유전 장애 같은 침투성 질환과 중앙학에서 원인 돌연변이의 중요한 일부분을 또한 설명할 수 있음을 시사한다. 그러나, 비코딩 게놈에서 변이체 해석의 어려움은, 큰 구조의 변이체의 외측에서, 현재 비코딩 게놈이 임상 관리에 제일 큰 영향을 주는 희귀 침투성 변이체에 관하여 추가적인 진단상 이익을 거의 주지 않는다는 것을 의미한다.

[0283] 필수적인 GT 및 AG 스플라이스 다이뉴클레오타이드의 외측에서 스플라이스 변경 돌연변이의 역할은 희귀 질환에서 오랫동안 인정되어 왔다. 사실, 이들 크립틱 스플라이스 변이체는 글리코젠 축적 질환 XI(폼페병)과 적혈구 생성프로토포르피린증과 같은 일부 희귀 유전 장애에 있어 가장 흔한 돌연변이이다. 인트론의 5' 및 3' 말단에

서의 연장된 스플라이스 모티프는 매우 퇴화해 있고 동일하게 우수한 모티프가 게놈에서 빈번하게 발생하므로, 기존 방법들로는 비실용적인 크립틱 스플라이싱(cryptic splicing)을 어느 비코딩 변이체가 유발할 수 있을지 정확한 예측을 형성한다.

[0284] 스플라이소좀이 어떻게 특이성을 획득하는지 더 잘 이해하기 위하여, 본 발명자들은 전구체-mRNA의 전사체에서 각각의 뉴클레오타이드에 대해, 그것이 스플라이스 수용체이든, 스플라이스 공여체이든, 또는 둘 다 아니든 간에, 전사체 서열만을 그 입력으로 사용하여, 예측하도록 심층 학습 신경망을 트레이닝시켰다(도 37A). 짝수 염색체 상의 정준 전사체를 트레이닝 세트로 사용하고 홀수 염색체 상의 전사체를 테스트용으로 사용하여(패럴로그는 제외됨), 심층 학습망은 엑손-인트론 경계를 95% 정확도로 호출하고, CFTR과 같이 100KB를 초과하는 전사체들조차도 종종 뉴클레오타이드 정밀도로 완벽하게 재건된다(도 37B).

[0285] 다음으로 본 발명자들은 이러한 놀라운 정밀도로 엑손-인트론 경계를 인식하기 위하여 망에 의해 사용되는 특이성 결정자를 이해하고자 하였다. 통계적 피쳐 또는 인간 공학 피쳐에 작용하는 이전 분류자들과 대조적으로, 심층 학습은 위계적 방식으로 서열로부터 피쳐를 직접 학습하므로, 장거리의 서열 컨텍스트로부터 추가적인 특이성이 알려질 수 있도록 한다. 사실, 본 발명자들은 망의 정확도가 망에 입력으로 제공되는 예측중인 뉴클레오타이드를 플랭킹하는 서열 컨텍스트의 길이에 크게 의존함을 확인하며(표 1), 본 발명자들이 서열 40-nt 만을 사용하는 심층 학습 모델을 트레이닝시킬 경우, 성능은 기존 통계적 방법을 조금 넘어설 뿐이다. 이는 개별 9 내지 23nt 스플라이싱 모티프를 인식함에 있어서 심층 학습이 기존 통계적 방법 이상으로 더하는 것이 거의 없음을 나타내지만, 더 넓은 서열 컨텍스트는 기능성 스플라이스 부위를 동일하게 강력한 모티프를 가진 비기능성 스플라이스 부위로부터 구별하기 위한 핵심이다. 어디에서 서열이 교란되는지 엑손에 대해 예측하도록 망에 부탁하는 것은, 인 비보(in vivo)의 엑손 스킵핑(exon skipping) 이벤트들에서 빈번히 관찰되는 바와 같이, 공여체 모티프의 파괴가 전형적으로 또한 수용체 신호가 사라지게 함을 나타내며(도 37C), 허용가능한 거리에서 강력한 수용체와 공여체 모티프 사이의 페어링을 요구하는 것만으로 상당한 정도의 특이성이 주어짐을 나타낸다.

[0286] 엑손 길이의 실험적 교란이 엑손 인클루전이나 엑손 스킵핑이나에 강력한 영향을 가짐을 다량의 증거가 보여주었음에도 불구하고, 그것은 심층 학습망의 정확도가 컨텍스트의 1000-nt 너머로 계속 증가하는지 이유를 설명하지 않는다. 국부적인 스플라이스 모티프에 의해 작동하는 특이성과 장거리 특이성 결정자들 사이를 더 잘 구별하기 위하여, 본 발명자들은 컨텍스트 100-nt 만을 입력으로 받아들이는 국부적인 망을 트레이닝시켰다. 국부적인 망을 사용하여 알고 있는 접합부들을 기록함으로써, 본 발명자들은 엑손과 인트론 양쪽 모두가 모티프 길이가 최소가 되는 최적의 길이(엑손의 경우 ~115nt, 인트론의 경우 ~1000nt)를 가진다는 것을 알았다(도 37D). 이러한 관계는 10000-nt 심층 학습망에는 존재하지 않으며(도 37E), 이는 인트론과 엑손 길이의 가변성이 광범위 컨텍스트 심층 학습망 내에 이미 완전히 감안되어 있음을 나타낸다. 특히, 인트론과 엑손의 경계는 광범위 컨텍스트 심층 학습 모델에 결코 주어진 바 없으므로, 이는 광범위 컨텍스트 심층 학습 모델이 서열만으로부터 엑손과 인트론의 위치를 추론함으로써 이러한 거리를 도출해낼 수 있었음을 나타낸다.

[0287] 핵사머 공간의 체계적 탐색은 심층 학습망이 모티프, 특히 위치 -34에서 -14까지에서 분기점 모티프 TACTAAC, 엑손 말단 부근의 잘 특성화된 엑손 스플라이스 인핸서 GAAGAA, 그리고 일반적으로 폴리피리미딘 트랙트의 일부지만 엑손 스플라이스 침묵자라도 작용하는 것으로 보이는 poly-U 모티프를 엑손-인트론 정의에 활용함을 또한 나타냈다(도 21, 도 22, 도 23 및 도 24).

[0288] 본 발명자들은, 참조 전사체 서열과 변이체를 수용하는 대체 전사체 서열 양쪽에서의 엑손-적어도 하나의 변이체 뉴클레오타이드만큼 상이 경계를 예측하고 엑손-인트론 경계에서의 모든 변화를 탐색함으로써 심층 학습망을 스플라이스 변경 기능을 위한 유전적 변이체의 평가로 확장한다. 60,706명의 인간들로부터 모은 엑솜 데이터를 최근 이용 가능하게 되면서, 본 발명자들은 스플라이스 기능을 변경할 것으로 예측된 변이체들에 대하여 대립유전자 빈도 스펙트럼에서 그들의 분포를 조사함으로써 부정적 선택의 효과를 평가할 수 있게 되었다. 본 발명자들은 예측된 크립틱 스플라이스 변이체가, 고 대립유전자 빈도에서 예상 계수치 대비 그들의 상대적인 고갈에 의해 입증되는 바와 같이, 강력하게 부정적 선택 하에 있음을 확인하며(도 38A), 그들의 고갈의 규모는 AG 또는 GT 스플라이스 파괴 변이체 및 정지 게인 변이체에 비견할 만하다. 부정적 선택의 영향은 프레임내 변화를 유발하는 변이체보다 프레임시프트를 유발할 수 있는 크립틱 스플라이스 변이체를 고려할 경우 더욱 크다(도 38B). 단백질 절단 변이의 타 클래스에 비하여 프레임을 시프트시키는 크립틱 스플라이스 변이체의 고갈에 기하면, 본 발명자들은 자신있게 예측된 크립틱 스플라이스 돌연변이의 88%가 기능성인 것으로 추정한다.

[0289] 전체 게놈 데이터를 엑솜 데이터로서 사용할 수 있는 것은 아니지만, 심층적 인트론 지역에서 자연적 선택의 영향을 검출하는 능력을 제한하며, 또한 엑손 영역으로부터 멀리 떨어진 크립틱 스플라이스 돌연변이의 관찰된 계

수치 대 예상 계수치를 계산할 수 있었다. 전반적으로, 본 발명자들은 엑손-인트론 경계로부터 >50nt의 거리에서 크립틱 스플라이스 돌연변이에서의 60% 고갈을 관찰한다(도 38C). 감쇠된 신호는, 엑소좀과 비교하여 더 작은 표본 크기와 전체 게놈 데이터의 조합일 수 있으며, 심층적 인트로 변이체의 영향을 예측하기가 더 어려울 수 있다.

[0290] 본 발명자들은 크립틱 스플라이스 변이체의 관측된 개수 대 예상된 개수를 사용하여 선택 하에 있는 크립틱 스플라이스 변이체의 개수와 이 개수가 단백질 절단 변이의 타 클래스와 어떻게 비교되는지를 추정할 수도 있다. 크립틱 스플라이스 변이체는 스플라이스 기능을 부분적으로만 파괴할 수 있으므로, 본 발명자들은 더욱 완화된 임계값에서 크립틱 스플라이스 변이체의 관측된 개수 대 예상된 개수를 또한 평가하였고, ExAC 데이터세트에 있는 희귀 AG 또는 GT 스플라이스 파괴 변이체에 비해 약 3배 많은 해로운 희귀 크립틱 스플라이스 변이체가 있는 것으로 추정한다(도 38D). 각 개인은 약 ~20의 희귀 크립틱 스플라이스 돌연변이를 지니며, 이는 단백질 절단 변이체의 개수와 대략 동일하지만(도 38E), 이 변이체들 모두가 스플라이스 기능을 완전히 파괴하는 것은 아니다.

[0291] 148명의 개인의 전체 게놈 서열분석과 다수의 조직 부위로부터의 RNA-seq를 둘 다 포함하는 GTEx 데이터가 최근 공개됨으로써, 본 발명자들은 RNA-seq 데이터에서 직접 희귀 크립틱 스플라이스 변이체의 효과를 찾아볼 수 있게 되었다. 희귀 질환 서열분석에서 직면했던 시나리오를 근사화하기 위하여, 본 발명자들은 희귀 변이체(GTEx 코호트에서의 싱글톤, 그리고 1000 게놈에 대해 1% 미만의 대립유전자 빈도)만을 고려하였고, 이들을 변이체를 가진 개인에 특유한 스플라이싱 이벤트와 페어링했다. 유전자 발현과 조직 발현에서의 차이점과 스플라이스 이상의 복잡성이 심층 학습 예측의 민감도와 특이성을 평가하는 것을 어렵게 만들긴 하지만, 엄격한 특이성 임계값에서 희귀 크립틱 스플라이스 돌연변이의 90% 이상이 RNA-seq 상에서 유효성확인됨을 확인했다(도 39A). RNA-seq에 존재하는 다수의 비정상 스플라이싱 이벤트는 심층 학습 분류자에 따르면 적당한 효과를 가지는 것으로 예측된 변이체들과 연계되는 것으로 보이며, 이는 그들이 스플라이스 기능에 부분적으로만 영향을 미치는 것임을 시사한다. 이러한 보다 민감한 임계값에서는, 새로운 접합부의 약 75%가 스플라이싱 기능에 비정상을 유발하는 것으로 예측된다(도 38B).

[0292] 심층 학습망이, 개체군 서열분석 데이터에 매우 해롭고 RNA-seq 상에서 높은 속도로 유효성확인되는 크립틱 스플라이스 변이체의 예측에 성공한 것은, 그 방법이 희귀 질환 서열분석 연구에서 추가적인 진단을 식별하는 데 사용될 수 있음을 시사한다. 이 가설을 테스트하기 위하여, 본 발명자들은 자폐증 및 신경발달장애를 위한 엑솜 서열분석 연구에서 드 노보 변이체를 조사하였고, 크립틱 스플라이스 돌연변이가 영향을 받는 개인들 대 그들의 건강한 형제자매들에 상당히 농축되어 있음을 입증한다(도 40A). 더욱이, 크립틱 스플라이스 돌연변이의 농축은 단백질 절단 변이체에 대한 농축보다 약간 낮으며, 이는 우리의 예측된 크립틱 스플라이스 변이체의 약 90%가 기능성임을 나타낸다. 이들 값에 근거하여, 질병유발 단백질 절단 변이체의 약 ~20%는 엑손과 엑손에 바로 인접한 뉴클레오타이드에서의 크립틱 스플라이스 돌연변이에 기인하는 것일 수 있다(도 40B). 전체 인트론 서열을 조사할 수 있는 전체 게놈 연구에 대해 이 수치를 확장하여 추정함으로써, 본 발명자들은 희귀 유전 장애에서 원인 돌연변이의 24%는 크립틱 스플라이스 돌연변이 때문이라고 추정한다.

[0293] 각 개별 유전자에 대해 드 노보 크립틱 스플라이스 돌연변이를 호출할 확률을 추정함으로써, 본 발명자들은 확률과 비교하여 후보 질환 유전자에서 크립틱 스플라이스 돌연변이의 농축을 추정할 수 있게 된다. 드 노보 크립틱 스플라이스 돌연변이는 미스센스 변이가 아닌 단백질 절단 변이에 의해 이전에 적중된 유전자 내에 강력하게 농축되었고(도 40C), 이는 그들이 다른 작용 모드보다는 일배체 불충분(haploinsufficiency)을 통해 대개 질병을 유발함을 나타낸다. 예측된 크립틱 스플라이스 돌연변이를 단백질 절단 변이체의 목록에 추가함으로써, 본 발명자들은, 단백질 절단 변이 하나만을 사용하는 경우와 비교하여, 자폐증에서 3가지 추가적인 질병 유전자 지적장애에서 11가지 추가적인 질병 유전자를 식별할 수 있게 된다(도 40D).

[0294] 질병 가능성이 있는 조직(이 경우 두뇌)을 구할 수 없었던 환자에서 크립틱 스플라이스 돌연변이를 유효성확인할 가능성을 평가하기 위하여, 본 발명자들은 사이먼의 심플렉스 컬렉션(Simon's Simplex Collection)에서 입수된, 예측된 드 노보 크립틱 스플라이스 돌연변이를 가지는 37명의 개인에 대하여 심층 RNA-seq를 수행하였고, 그 개인에게는 존재하였지만 실험에서의 다른 모든 개인들과 GTEx 코호트로부터의 149명의 개인들에게는 부재하였던 비정상 스플라이싱 이벤트를 탐색하였다. 본 발명자들은 37명의 환자 중 NN이 RNA-seq 상에서, 예측된 크립틱 스플라이스 변이체에 의해 설명되는, 독특한 비정상 스플라이싱을 나타냈음을 확인한다(도 40E).

[0295] 요약하면, 본 발명자들은 희귀 유전 장애에서 원인 질환 돌연변이를 식별하는 데 유용할 정도로 충분한 정밀도로 크립틱 스플라이스 변이체를 정확히 예측하는 심층 학습 모델을 입증한다. 본 발명자들은 크립틱 스플라이싱

에 의해 유발되는 희귀 질환 진단의 상당한 부분을 단백질 코딩 영역만을 고려함으로써 현재 놓치고 있다고 추정하며, 비코딩 게놈에서 침투성 희귀 변이의 효과를 해석하기 위한 방법을 개발할 필요성을 강조한다.

[0296]

결과

[0297]

심층 학습을 이용한 1차 서열로부터의 스플라이싱의 정확한 예측

[0298]

본 발명자들은, 전구체-mRNA 전사체의 게놈 서열만을 입력으로 사용하여, mRNA 전사체에서의 각 위치가 스플라이스 공여체인지, 스플라이스 수용체인지, 또는 둘 다 아닌지 예상하는 심층 잔여 신경망(He et al., 2016a)을 구성하였다(도 37A와 도 21, 도 22, 도 23 및 도 24). 스플라이스 공여체와 스플라이스 수용체가 수만개의 뉴클레오타이드에 의해 분리될 수 있기 때문에, 본 발명자들은 매우 큰 게놈 거리에 걸쳐 놓여있는 서열 결정자를 인지할 수 있는 32개의 팽창 컨볼루션층으로 구성된 새로운 망 아키텍처를 사용했다(Yu and Koltun, 2016). 엑손-인트론 경계에 인접한 짧은 뉴클레오타이드 윈도우만을 고려했거나(Yeo and Burge, 2004), 인간 공학 피처에 의존했거나(Xing et al., 2015), 또는 발현 또는 스플라이스 인자 바인딩과 같은 실험 데이터에 의존했던(Jha et al., 2017) 이전 방법들과 대조적으로, 본 발명의 신경망은, 전구체-mRNA 전사체에서 각 위치의 스플라이스 기능을 예측하기 위해 플랭킹 컨텍스트 서열의 10,000개의 뉴클레오타이드를 평가함으로써, 1차 서열로부터 직접 스플라이싱 결정자를 학습한다.

[0299]

본 발명자들은 신경망의 파라미터들을 트레이닝하기 위하여 인간 염색체의 서브세트에 대해 GENCODE 주석이 달린 전구체-mRNA 전사체 서열(Harrow et al., 2012)을 사용하고, 신경망의 예측을 테스트하기 위하여 페달로그를 제외한 나머지 염색체에 대해 전사체를 사용했다. 테스트 데이터세트에 있는 전구체-mRNA 전사체에 대해, 신경망은 스플라이스 접합부를 95% top-k 정확도로 예측하며, 이는 예측된 부위 개수가 테스트 데이터세트에 존재하는 스플라이스 부위의 실제 개수와 동일한 경우의 임계값에서 정확히 예측된 스플라이스 부위의 분율이다(Boyd et al., 2012; Yeo and Burge, 2004). CFTR과 같이 100 kb를 초과하는 유전자들조차도 종종 뉴클레오타이드 정밀도로 완벽하게 재건된다(도 37B). 신경망이 엑손 서열 편향에 단지 의존하고 있는 것이 아님을 확인하기 위하여, 본 발명자들은 긴 비코딩 RNAs에 대해 또한 신경망을 테스트했다. 본 발명자들의 정확도를 낮출 것으로 예상되는 비코딩 전사체 주해의 불완전함에도 불구하고, 신경망은 lincRNAs 내의 알려진 스플라이스 접합부를 84% top-k 정확도로 예측하며(도 42A 및 도 42B), 이는 신경망이 단백질-코딩 선택적 압력이 없는 임의의 서열들에 대한 스플라이소좀의 거동을 근사화할 수 있음을 나타낸다.

[0300]

테스트 데이터세트에서 GENCODE 주석이 달린 각각의 엑손(각 유전자의 최초 및 최후 엑손은 제외)에 대하여, 본 발명자들은, Gene and Tissue Expression 도해서(GTEx)로부터의 RNA-seq 데이터에 기하여(The GTEx Consortium et al., 2015)(도 37C), 신경망의 예측 점수가 엑손 인클루전 대 엑손 스킵핑을 지원하는 리드의 분율과 상관관계가 있는지 여부를 또한 조사했다. GTEx 조직을 가로질러 구조적으로 스플라이스 인되거나 또는 스플라이스 아웃되었던 엑손들은 각각 1 또는 0에 가까운 예측 점수가 나온 반면, 상(샘플들 평균이 엑손 인클루전 10 내지 90% 사이인) 상당한 정도의 대체 스플라이스를 겪은 엑손들은 중간 점수(피어슨 상관관계 = 0.78, P ≈ 0)를 향하는 경향이 있었다.

[0301]

다음으로 본 발명자들은 놀라운 정확도를 획득하기 위하여 신경망에 의해 활용되는 서열 결정자를 이해하고자 하였다. 본 발명자들은 주석이 달린 엑손들 주위의 각 뉴클레오타이드의 체계적인 인 실리코 치환을 수행하여, 인접한 스플라이스 부위들에서 신경망의 예측 점수에 대한 영향을 측정하였다(도 37E). 스플라이스 공여체 모티프의 서열의 파괴는 빈번히, 엑손 스킵핑 이벤트에서 인 비보로 관찰되는 바와 같이, 신경망이 상류 스플라이스 수용체 부위 또한 분실될 것으로 예측하도록 함을 발견하였으며, 이는 최적 거리에 설정된 페어링된 상류 수용체 모티프와 하류 공여체 모티프 사이에 엑손 정의에 의해 상당한 정도의 특이성이 주어지는 것을 나타낸다(Berget, 1995). 스플라이싱 시그널에 기여하는 추가적 모티프는 SR-단백질족과 그 분기점의 잘 특성화된 바인딩 모티프를 포함한다(도 43A 및 도 43B)(Fairbrother et al., 2002; Reed and Maniatis, 1988). 이러한 모티프의 효과는 엑손 내의 그들의 위치에 크게 의존하는데, 이는 그들의 역할이, 경합하는 수용체 및 공여체 부위들 사이를 구별함으로써 인트론-엑손 경계의 정밀한 포지셔닝을 구체화하는 것을 포함함을 시사한다.

[0302]

가변 입력 서열 컨텍스트를 가진 신경망을 트레이닝시키는 것은 스플라이스 예측의 정확도에 현저하게 영향을 미치며(도 37E), 이는 스플라이스 부위로부터 최대 10,000 nt 떨어진 장거리 서열 결정자가 기능성 스플라이스 접합부를 최적으로 근접한 모티프를 가진 다수의 비기능성 부위로부터 식별하는 데 필수적이다. 장거리 및 단거리의 특이성 결정자들을 조사하기 위하여, 본 발명자들은, 서열 컨텍스트 80 nt에 대해 트레이닝된 모델(SpliceNet-80nt) 대 컨텍스트 10,000 nt에 대해 트레이닝된 완전체 모델(SpliceNet-10k)에 의해, 주석이 달린 접합부들에 부여된 점수를 비교했다. 서열 컨텍스트 80 nt에 대해 트레이닝된 망은 일반적인 길이(엑손의 경우

150 nt, 인트론의 경우 ~1000 nt)의 엑손 또는 인트론에 인접한 접합부에 더 낮은 점수를 부여하며(도 37F), 이는 비정상적으로 길거나 짧은 엑손과 인트론의 스플라이스 부위와 비교할 때 이러한 부위가 더 약한 스플라이스 모티프를 가지는 경향이 있다는 이전의 관찰과 일치하는 것이다(Amit et al., 2012; Gelfman et al., 2012; Li et al., 2015). 대조적으로, 서열 컨텍스트의 10,000 nt 에 대해 트레이닝된 망은 더 약한 스플라이스 모티프에도 불구하고 평균적인 길이의 인트론과 엑손에 대한 선호를 나타내는데, 이는 그것이 엑손 또는 인트론 길이에 의해 수여된 장거리 특이성을 설명할 수 있기 때문이다. 긴 연속된 인트론에서 더 약한 모티프의 스킵핑은 엑손 포징(pausing)의 부재시에 실험적으로 관찰된 더 빠른 RNA 중합요소 II 연장과 일관되며, 이는 준최적 모티프를 인식하기 위해 스플라이소좀에 더 적은 시간을 허용할 수 있다(Close et al., 2012; Jonkers et al., 2014; Veloso et al., 2014). 본 발명자들의 발견은 평균적인 스플라이스 접합부는 중대한 특이성을 수여하는 호의적인 장거리 서열 결정자를 보유함을 시사하며, 이는 대부분의 스플라이스 모티프에서 용인되는 고도의 서열 퇴화를 설명한다.

[0303] 스플라이싱은 공동 전사적으로(co-transcriptionally) 발생하기 때문에(Cramer et al., 1997; Tilgner et al., 2012), 염색질 상태와 공동 전사 스플라이싱 사이의 상호작용은 또한 엑손 정의를 가이드하고(Luco et al., 2011), 염색질 상태가 1차 서열로부터 예측가능한 정도까지 망에 의해 활용될 수 있는 잠재력을 가질 수 있다. 특히, 뉴클레오솜 포지셔닝의 전체 게놈 연구는 뉴클레오솜 점유가 엑손에서 더 높음을 보여주었다(Andersson et al., 2009; Schwartz et al., 2009; Spies et al., 2009; Tilgner et al., 2009). 스플라이스 부위 예측을 위해 망이 뉴클레오솜 포지셔닝의 서열 결정자를 사용하는지 여부를 테스트하기 위하여, 본 발명자들은 게놈 전체에 걸쳐 150 nt(대략 평균적인 엑손의 크기)만큼 분리된 한 쌍의 최적 수용체 및 공여체 모티프를 왔다갔다했고 그 한 쌍의 모티프가 그 좌위에서 결과적으로 엑손 인클루전이 될 것인지 여부를 예측할 것을 망에게 부탁했다(도 37G). 본 발명자들은 엑손 인클루전에 호의적일 것으로 예측된 위치들은, 유전자간 영역(스피어맨 상관관계 = 0.36, $P < 0$)에서조차도, 높은 뉴클레오솜 점유와 상관관계가 있음을 알았으며, 이러한 결과는 GC 함량 조절 후에도 지속된다(도 44A). 이러한 결과들은 망이 1차 서열로부터 뉴클레오솜 포지셔닝을 예측하도록 은연 중에 학습하였고 그것을 엑손 정의에서 특이성 결정자로 활용함을 시사한다. 평균적인 길이의 엑손 및 인트론과 유사하게, 뉴클레오솜 너머에 위치된 엑손은 더 약한 국부적 스플라이스 모티프를 가지며(도 44B), 이는 보상인자 존재 시에 퇴화된 모티프에 대한 용인이 더 큰 것과 일관된다(Spies et al., 2009).

[0304] 다수의 연구가 엑손과 뉴클레오솜 점유 사이의 상관관계를 발표하긴 했지만, 엑손 정의에서 뉴클레오솜 포지셔닝에 대한 인과적 역할은 견고하게 확립되지 않았다. RNA-seq와, Genotype-Tissue Expression(GTEx) 코호트(The GTEx Consortium et al., 2015)로부터의 전체 게놈 서열분석 양쪽 모두를 가진 149명 개인으로부터의 데이터를 사용하여, 본 발명자들은, 단일 개인에게 개인적이며, 개인적인 스플라이스 부위를 생성하는 유전적 돌연변이에 대응하는 새로운 엑손을 식별하였다. 이들 개인적인 엑손 생성 이벤트는 K562 및 GM12878 세포들(치환 테스트에 의해 $P = 0.006$, 도 37H)에서의 현존 뉴클레오솜 포지셔닝과 크게 연관되어 있었는데, 이들 세포주들은 상응하는 개인적인 유전적 돌연변이가 결핍되어 있을 가능성이 큼에도 불구하고 그러했다. 본 발명자들의 결과는, 만약 결과적으로 생겨난 새로운 엑손이 현존 뉴클레오솜 점유의 일 영역을 뒤덮을 것이라면, 유전적 변이체가 새로운 엑손의 생성을 촉발할 가능성이 더 크다는 점을 나타내며, 이는 엑손 정의 촉진시에 뉴클레오솜 포지셔닝에 대한 인과적 역할을 지지한다.

[0305] **RNA-seq 데이터 중 예측된 크립틱 스플라이스 돌연변이의 검증**

[0306] 본 발명자들은 심층 학습망을, 참조 전구체-mRNA 전사체 서열과 변이체를 포함하는 대체 전사체 서열 모두에 대한 엑손-인트론 경계를 예측하고, 점수 간의 차이(Δ Score)를 취함으로써, 스플라이스 변경 기능에 대한 유전적 변이체의 평가에 확장했다(도 38A). 중요한 것은, 이 망은 참조 전사체 서열 및 스플라이스 접합부 주석에 대해 서만 트레이닝되었고, 트레이닝 중에 변이체 데이터를 볼 수 없었는데, 이는 변이체 효과의 예측을 스플라이싱의 서열 결정자를 정확하게 모델링하는 망의 능력에 대한 도전적인 테스트로 만들었다.

[0307] 본 발명자들은 전체 게놈 서열분석과 다수의 조직으로부터의 RNA-seq 모두를 갖는 149명의 개인으로 구성된 GTEx 코호트(GTEx Consortium et al., 2015)에서 RNA-seq 데이터의 크립틱 스플라이스 돌연변이의 영향을 탐색했다. 희귀 질환의 서열분석에서 직면한 시나리오를 근사화하기 위하여, 본 발명자들은 먼저 희귀한 개인적 돌연변이(GTEx 코호트 내의 1명에만 존재)에 초점을 맞췄다. 본 발명자들은 신경망에 의해 기능적인 결과를 가져올 것으로 예측되는 개인적인 돌연변이가 개인적인 신규 스플라이스 접합부와 개인적인 엑손 스킵핑 이벤트에서 스킵된 엑손의 경계에서 강력하게 농축됨을 발견하였고(도 38B), 이는 이러한 예측들의 큰 일부가 기능적임을 시사한다.

- [0308] 정상 및 비정상 스플라이스 동형의 상대적인 생성에 대한 스플라이스 부위 생성 변이체의 영향을 정량화하기 위해, 본 발명자들은 신규한 스플라이스 이벤트를 지원하는 리드의 개수를 부위를 커버하는 리드의 총 개수의 비율로서 측정하였다(도 38C)(Cummings et al., 2017). 스플라이스 부위 파괴 변이체에 대하여, 많은 엑손이 엑손 스킵핑의 낮은 베이스라인 비율을 가지며, 변이체의 효과는 엑손 스킵핑 리드의 비율을 증가시키는 것을 관찰했다. 따라서 파괴된 접합부에서 스플라이싱되는 리드 비율의 감소와 엑손을 스킵한 리드 비율의 증가 양쪽 모두를 계산하고, 두 효과 중 큰 쪽을 취했다(도 45 및 STAR 방법).
- [0309] 확실하게 예측된 크립틱 스플라이스 변이체($\Delta\text{Score} \geq 0.5$)는 필수 GT 또는 AG 스플라이스 파괴 속도의 3/4으로 RNA-seq에서 유효성확인된다(도 38D). 크립틱 스플라이스 변이체의 유효성확인율과 효과 크기는 모두 ΔScore 를 면밀히 추적하며(도 38D 및 도 38E), 모델의 예측 점수가 변이체의 스플라이스 변경 잠재력에 대한 훌륭한 대용품을 보여준다. 유효성확인된 변이체, 특히 낮은 점수($\Delta\text{Score} < 0.5$)를 갖는 변이체는 종종 불완전하게 침투성이며, 결과적으로 RNA-seq 데이터에서 비정상 및 정상 전사체 모두의 혼합물을 생성함으로써 대체 스플라이싱한다(도 38E). 유효성확인율 및 효과 크기에 대한 본 발명의 추정치는 보수적이며, 설명되지 않은 스플라이스 동형 변화와 언센스 매개 붕괴로 인해 실제 값을 과소평가하기 쉬운데, 언센스 매개 붕괴는 우선적으로 이상 스플라이싱된 전사체를 퇴화시키며 이는 그들이 빈번히 미성숙한 정지 코돈을 도입하기 때문이다(도 38C 및 도 45). 이는 필수 GT 및 AG 스플라이스 다이뉴클레오타이드를 파괴하는 변이체의 평균 효과 크기가 완전 침투성 이형접합 변이체에 대해 예상되는 50% 미만인 것에 의해 입증된다.
- [0310] mRNA 전사체의 관찰된 사본의 적어도 3/10에서 비정상 스플라이스 동형을 생성하는 크립틱 스플라이스 변이체에 대해, 많은 변이체가 엑손 근처에 있을 때 71%, 변이체가 심층 인트론 서열에 있을 때 41%의 민감도를 갖는다($\Delta\text{Score} \geq 0.5$, 도 38F). 이러한 발견은 심층 인트론 변이체가 예측하기가 더 어렵다는 것을 나타내며, 이는 아마도 심층 인트론 영역이 엑손 근처에 존재하도록 선택된 특이성 결정자가 더 적게 포함하기 때문일 수 있다.
- [0311] 기존의 방법에 대한 본 발명의 망 성능을 벤치마킹하기 위하여, 본 발명자들은 희귀 유전 질환 진단에 대해 문헌에 언급되었던 세 가지 인기 분류자, GeneSplicer(Pertea et al., 2001), MaxEntScan (Yeo and Burge, 2004), 그리고 NNSplice(Reese et al., 1997)를 선택하여, 다양한 임계값에서 RNA-seq 유효성확인율 및 민감도를 도표화하였다(도 38G). 해당 분야의 다른 사람들의 경험과 마찬가지로(Cummings et al., 2017), 아마도 스플라이싱에 영향을 미칠 수 있는 매우 많은 수의 비코딩 변이체가 게놈 전체에 있음을 감안할 때 기존의 분류자는 특이성이 불충분한 것으로 나타났으며, 추정컨대 이는 그들이 국부적인 모티프에 집중하고 장거리 특이성 결정자를 설명하지 않기 때문이다.
- [0312] 기존 방법과 비교할 때 성능의 큰 차이를 감안하여, 본 발명자들은 RNA-seq 데이터의 결과가 과대적합에 의하여 혼동될 수 있는 가능성을 배제하기 위해 추가 통계를 수행했다. 먼저, 개인적인 변이체와 GTEX 코호트에서 둘 이상의 개인에 존재하는 변이체에 대한 유효성확인 및 민감도 분석을 개별적으로 반복했다(도 46A, 도 46B 및 도 46C). 스플라이스 기계나 심층 학습 모델 모두 대립유전자 빈도 정보에 접근할 수 없기 때문에, 대립유전자 빈도 스펙트럼 전체에 걸쳐서 망이 유사한 성능을 갖는지 확인하는 것은 중요한 통계이다. 동일한 ΔScore 임계값에서, 개인 및 공통 크립틱 스플라이스 변이체는 RNA-seq ($P > 0.05$, Fisher Exact 테스트)에서 유효성확인율에 유의미한 차이가 없음을 발견하였으며, 이는 망의 예측이 대립유전자 빈도에 대해 강건하다는 것을 나타낸다.
- [0313] 두 번째로, 새로운 스플라이스 접합부를 생성할 수 있는 서로 다른 유형의 크립틱 스플라이스 변이체들에 걸쳐 모델의 예측을 유효성확인하기 위하여, 본 발명자들은 신규한 GT 또는 AG 다이뉴클레오타이드를 생성하는 변이체, 확장된 수용체 또는 공여체 모티프에 영향을 미치는 변이체, 그리고 더 원위 영역에서 발생하는 변이체를 별도로 평가했다. 본 발명자들은 크립틱 스플라이스 변이체가 이들 세 그룹 간에 거의 균등하게 분포되어 있으며 동일한 ΔScore 임계값에서 그룹 간의 유효성확인율 또는 효과 크기에 유의미한 차이가 없음을 확인했다(균일성 테스트에서 $P > 0.3$ χ^2 및 Mann Whitney U 테스트에서 $P > 0.3$, 도 47A 및 도 47B).
- [0314] 세 번째로, 트레이닝에 사용된 염색체 상의 변이체와 나머지 염색체 상의 변이체에 대해 각각 RNA-seq 유효성확인 및 민감도 분석을 수행했다(도 48A 및 48b). 비록 망은 참조 게놈 서열 및 스플라이스 주석에 대해서만 트레이닝되었고, 트레이닝 중에 변이체 데이터에 노출되지 않았지만, 본 발명자들은 망이 트레이닝 염색체에서 참조 서열을 보았다는 사실로부터 발생하는 변이체 예측에서의 편향 가능성을 배제하고 싶었다. 본 발명자들은 트레이닝 및 테스트 염색체로부터의 변이체에 대해 망이 동일하게 잘 작동한다는 것을 알았으며, 유효성확인율이나 민감도($P > 0.05$, Fisher Exact 테스트)에 유의미한 차이가 없었는데, 이는 망의 변이체 예측이 트레이닝 서열의 과대적합에 의해 설명될 것 같지 않음을 나타낸다.

[0315] 크립틱 스플라이스 변이체를 예측하는 것은, 다른 스플라이스 예측 알고리즘뿐만 아니라 본 발명자들의 모델의 결과에 의해 보여지는 바와 같이, 주석이 달린 스플라이스 접합부를 예측하는 것보다 어려운 문제이다(도 37E 및 도 38G 비교). 중요한 이유는 두 가지 유형의 분석 사이의 엑손 인클루전 비율의 기본 분포의 차이 때문이다. 대부분의 GENCODE 주석이 달린 엑손은 강한 특이성 결정자를 가지며, 결과적으로 1에 가까운 구조적 스플라이싱 및 예측 점수를 갖는다(도 37C). 대조적으로, 대부분의 크립틱 스플라이스 변이체는 부분적으로만 침투성이고(도 38D 및 도 38E), 예측 점수가 낮거나 중간이며, 정상 및 비정상 전사체의 혼합물의 생성에 의해 대체 스플라이싱으로 빈번히 인도한다. 이로 인해 크립틱 스플라이스 변이체의 영향을 예측하는 후자의 문제는 주석이 달린 스플라이스 부위를 식별하는 것보다 본질적으로 어려운 문제가 된다. 넌센스 매개 붕괴와 같은 추가 인자, 설명되지 않은 동형 변화, RNA-seq 분석의 한계는 RNA-seq 유효성확인율을 낮추는 데 추가로 기여한다(도 38C 및 도 45).

[0316] **조직 특이적 대체 스플라이싱은 약한 크립틱 스플라이스 변이체에서 자주 발생한다**

[0317] 대체 스플라이싱은 상이한 조직 및 발달 단계에서 전사체의 다양성을 증가시키는 역할을 하는 주요 유전자 조절 모드이며, 조절 장애는 질병 과정과 관련이 있다(Blencowe, 2006; Irimia et al., 2014; Keren et al., 2010) Licatalosi and Darnell, 2006; Wang et al., 2008). 예기치 않게, 본 발명자들은 크립틱 스플라이스 돌연변이에 의해 생성된 신규 스플라이스 접합부의 상대적인 사용이 조직에 따라 상당히 달라질 수 있음을 발견했다(도 39A). 또한, 스플라이싱에서 조직 특이적 차이를 야기하는 변이체는 다수의 개인에 걸쳐 재현 가능하며(도 39B), 이는 확률론적 영향보다는 조직 특이적 생물학이 이러한 차이의 기초가 될 수 있음을 나타낸다. 본 발명자들은 약하고 중간 예측 점수(Δ Score 0.35 - 0.8)를 가진 크립틱 스플라이스 변이체의 35%가 조직 전체에 걸쳐 생성된 정상 및 비정상 전사체의 분율에서 유의미한 차이를 나타냄을 발견했다(χ^2 테스트에 대해 Bonferroni-보정된 $P < 0.01$, 도 39C). 이는 조직 특이적 효과를 생성할 가능성이 훨씬 더 작은($P = 0.015$) 높은 예측 점수(Δ Score > 0.8)를 갖는 변이체와 대조를 이룬다. 본 발명자들의 발견은 대체 스플라이싱된 엑손이, 각각 1 또는 0에 가까운 점수를 갖는 구조적으로 스플라이싱 인되거나 스플라이싱 아웃된 엑손과 비교하여, 중간 예측 점수를 갖는 경향이 있다는 이전의 관찰과 일치한다(도 37C).

[0318] 이러한 결과는 염색질 컨택스트 및 RNA-결합 단백질의 결합과 같은 조직 특이적 인자가 선호도가 가까운 두 스플라이스 접합부 사이에서 경쟁을 일으킬 수 있는 모델을 지지한다(Gelfman et al., 2013; Luco et al., 2010; Shukla et al., 2011; Ule et al., 2003). 강한 크립틱 스플라이스 변이체는 후성 유전적 컨택스트에 상관없이 스플라이싱을 정상으로부터 비정상 동형으로 완전히 이동시킬 가능성이 있는 반면, 약한 변이체는 스플라이스 접합부 선택을 결정 경계에 더 가깝게 하여, 상이한 조직 유형 및 세포 컨택스트에서 대체 접합부 사용을 유발한다. 이것은 신규한 대체 스플라이싱 다양성을 생성하는 데 있어서 크립틱 스플라이스 돌연변이에 의해 수행된 예기치 않은 역할을 강조하는데, 이는 그 후에 자연적 선택이 유용한 조직 특이적 대체 스플라이싱을 생성하는 돌연변이를 보존할 기회를 가질 것이기 때문이다.

[0319] **예측된 크립틱 스플라이스 변이체는 인간 개체군에서 강력하게 유해하다**

[0320] 예측된 크립틱 스플라이스 변이체는 RNA-seq에서 높은 속도로 유효성확인되지만, 많은 경우에 효과가 완전 침투성이 아니고 정상 및 비정상 스플라이스 동형의 혼합물이 생성되어, 이들 크립틱 스플라이스 변경 변이체의 일부가 기능적으로 중요하지 않을 수 있는 가능성이 발생할 가능성이 높아진다. 예측된 크립틱 스플라이스 변이체에 대한 자연 선택의 특징적 흔적을 탐색하기 위하여, 본 발명자들은 Exome Aggregation Consortium(ExAC) 데이터베이스(Lek et al., 2016)로부터의 60,706개의 인간 엑솜에 존재하는 각 변이체의 점수를 매기고 엑손-인트론 경계를 변경할 것으로 예측되는 변이체를 식별했다.

[0321] 예측된 스플라이스 변경 변이체에 작용하는 부정적 선택의 정도를 측정하기 위해, 본 발명자들은 일반적인 대립 유전자 빈도에서 발견되는 예측된 스플라이스 변경 변이체의 수(인간 인구에서 $\geq 0.1\%$)를 세고 ExAC의 싱글톤 대립유전자 빈도에서의 예측된 스플라이스 변경 변이체의 수(즉, 60,706명 중 1명)와 비교했다. 최근 인간 인구 규모의 기하급수적 팽창으로 인해 싱글톤 변이체는 정제 선택에 의해 최소한으로 여과된 최근에 생성된 돌연변이를 나타낸다(Tennessen et al., 2012). 대조적으로, 공통 변이체는 정제 선택의 체를 통과한 중성 돌연변이의 서브 세트를 나타낸다. 따라서, 싱글톤 변이체에 대한 공통 대립유전자 빈도 스펙트럼에서 예측된 스플라이스 변경 변이체의 고갈은, 유해하고 따라서 기능성인, 예측된 스플라이스 변경 변이체의 분율의 추정치를 제공한다. 단백질-코딩 서열에 대한 혼란스러운 영향을 피하기 위해, 본 발명자들은 필수 GT 또는 AG 다이뉴클레오타이드 외부에 있는 동의 변이체 및 인트론 변이체로 분석을 제한하였으며, 역시 스플라이스 변경 효과가 있을 것으로 예상되는 미스센스 돌연변이는 제외하였다.

- [0322] 공통 대립유전자 빈도에서, 예상과 비교한 상대적 고갈에 의해 입증되는 바와 같이, 자신있게 예측된 크립틱 스플라이스 변이체($\Delta\text{Score} \geq 0.8$)는 강한 부정적 선택을 받고 있다(도 40A). 대부분의 변이체가 RNA-seq 데이터에서 완전 침투성에 근접할 것으로 예상되는 이 임계값에서(도 38D), 예측된 동의 및 인트론 크립틱 스플라이스 돌연변이는 일반적인 대립유전자 빈도에서 78% 고갈되는데, 이는 프레임시프트, 정지 게인 및 필수 GT 또는 AG 스플라이스 파괴 변이체의 82% 고갈에 비견될만하다(도 40B). 프레임내 변경을 일으키는 것보다 프레임시프트를 유발하는 크립틱 스플라이스 변이체를 고려할 때 부정적 선택의 영향이 더 크다(도 40C). 프레임시프트 결과를 갖는 크립틱 스플라이스 변이체의 고갈은 다른 클래스의 단백질 절단 변이와 거의 동일하며, 이는 인트론 부근 영역(알려진 엑손-인트론 경계로부터 ≤ 50 nt)에서 자신있게 예측된 크립틱 스플라이스 돌연변이의 대다수가 기능성이며 인간 개체군에 매우 해로운 영향을 미친다는 것을 나타낸다.
- [0323] 이 분석을 알려진 엑손-인트론 경계에서 > 50 nt 인 침묵 인트론 영역으로 확장하기 위해, 본 발명자들은 Genome Aggregation Database(gnomAD) 코호트(Lek et al., 2016)에서 나온 15,496명의 인간으로부터 집계된 전체 게놈 서열분석 데이터를 사용하여 공통 대립유전자 빈도에서 크립틱 스플라이스 돌연변이의 관찰 계수치 및 예상 계수치를 계산했다. 전반적으로, 본 발명자들은 엑손-인트론 경계로부터 > 50 nt 이상의 거리에서 공통 크립틱 스플라이스 돌연변이($\Delta\text{Score} \geq 0.8$)의 56% 고갈을 관찰하며(도 40D), 이는 RNA-seq 데이터에서 관찰했듯이 침묵 인트론 변이체의 영향을 예측하는데 있어서의 더 큰 어려움과 일관된다.
- [0324] 다음으로 본 발명자들은, gnomAD 코호트에서 개인당 희귀한 크립틱 스플라이스 돌연변이의 수를 측정함으로써, 다른 유형의 단백질 코딩 변이에 비해, 크립틱 스플라이스 돌연변이가 침투성 유전 질환에 기여할 가능성을 추정하려고 노력했다. 부정적 선택 하에 있는 예측된 크립틱 스플라이스 돌연변이의 분율에 기초하여(도 40A), 평균적인 인간은 ~ 11 개의 희귀 단백질 절단 변이체와 비교하여 ~ 5 개의 희귀 기능성 크립틱 스플라이스 돌연변이(대립유전자 빈도 $< 0.1\%$)를 보유한다(도 40E). 크립틱 스플라이스 변이체는 필수 GT 또는 AG 스플라이스 파괴 변이체보다 약 2:1 더 많다. 이러한 크립틱 스플라이스 변이체의 상당 부분은, 그들이 프레임내 변경을 생성하기 때문이거나 또는 스플라이싱을 이상 동형으로 완전히 시프트시키지 않기 때문에, 유전자 기능을 완전히 파괴하지 않을 수 있음을 주의한다.
- [0325] **드 노보 크립틱 스플라이스 돌연변이는 희귀 유전 장애의 주요 원인이다**
- [0326] 자폐 스펙트럼 장애 및 심각한 지적장애를 가진 환자에 대한 대규모 서열분석 연구는 신경 발달 경로에서 유전자를 파괴하는 드 노보 단백질-코딩 돌연변이(미스센스, 넌센스, 프레임시프트 및 필수 스플라이스 다이뉴클레오타이드)의 중심적인 역할을 입증했다(Fitzgerald et al., 2015; Iossifov et al., 2014; McRae et al., 2017; Neale et al., 2012; De Rubeis et al., 2014; Sanders et al., 2012). 변경된 스플라이싱을 통해 작용하는 비코딩 돌연변이의 임상적 영향을 평가하기 위해, 본 발명자들은 신경망을 적용하여 Deciphering Developmental Disorders 코호트(DDD)의 지적장애를 가진 4,293명(McRae et al., 2017), Simons Simplex Collection(De Rubeis et al., 2014; Sanders et al., 2012; Turner et al., 2016)과 Autism Sequencing Consortium의 자폐 스펙트럼 장애(ASD) 환자 3,953명, 및 영향을 받지 않은 Simons Simplex Collection의 형제 자매 대조군 2,073명에서 드 노보 돌연변이의 영향을 예측했다. 여러 연구에 걸친 드 노보 변이체 확정의 차이를 통제하기 위해, 본 발명자들은 개개인당 동의 돌연변이의 수가 코호트들에 걸쳐 동일하도록 드 노보 변이체의 예상되는 수를 정규화하였다.
- [0327] 스플라이싱을 파괴할 것으로 예측된 드 노보 돌연변이는 건강한 대조군($\Delta\text{Score} \geq 0.1$, 도 41A, 도 43A 및 도 43B)과 비교하여 지적장애가 1.51배($P = 0.000416$) 그리고 자폐 스펙트럼 장애가 1.30배($P = 0.0203$) 농축된다. 스플라이스 파괴 돌연변이는 또한, 이중 단백질-코딩 및 스플라이싱 효과를 갖는 돌연변이에 의해서만 농축이 설명될 가능성을 제외하면, 동의 및 인트론 돌연변이만을 고려할 때, 대조군과 비교하는 경우에 현저하게 농축된다(도 49A, 도 49B 및 도 49C). 각 연구에서 서열분석 범위 또는 변이체 확정이 걸려진 영역에서 돌연변이의 예상 분율을 조정한 후, 영향을 받는 사람 대 영향을 받지 않은 사람의 과도한 드 노보 돌연변이에 기초하여, 크립틱 스플라이스 돌연변이는 자폐 스펙트럼 장애에서 병원성 돌연변이의 약 11%, 지적장애에서 병원성 돌연변이의 9%를 구성하는 것으로 추정된다(도 41B). 영향을 받은 개체에서 대부분의 드 노보 예측된 크립틱 스플라이스 돌연변이는 $\Delta\text{Score} < 0.5$ 를 가졌으며(도 41C, 도 50A 및 도 50B), GTEx RNA-seq 데이터셋에서 유사한 점수를 갖는 변이체에 기초하여 정상 및 비정상 전사체의 혼합물을 생성할 것으로 예상된다.
- [0328] 확률과 비교하여 후보 질환 유전자에서 크립틱 스플라이스 돌연변이의 농축을 추정하기 위해, 본 발명자들은 돌연변이율을 조정하기 위해 트라이뉴클레오타이드 컨텍스트를 사용하여 각각의 개별 유전자에 대한 드 노보 크립틱 스플라이스 돌연변이를 호출할 확률을 계산 하였다(Samocha et al., 2014)(표 S4). 신규 유전자 발견에서 크

립틱 스플라이스 돌연변이 및 단백질-코딩 돌연변이 둘 다를 조합하면, 단백질 코딩 돌연변이만을 고려할 때 발견 임계치 (FDR < 0.01) 미만일 수도 있었던(Kosmicki et al., 2017; Sanders et al., 2015), 지적장애와 관련된 5개의 추가 후보 유전자 및 자폐 스펙트럼 장애와 관련된 2개의 추가 유전자가 생성된다(도 41D 및 도 45).

[0329] 자폐증 환자에서 드 노보 크립틱 스플라이스 돌연변이의 실험적 유효성확인

[0330] 본 발명자들은 Simons Simplex Collection의 36명으로부터 말초 혈액 유래 림프 모세포 세포주(LCL)를 얻었으며, 이는 적어도 최소 수준의 LCL 발현을 가진 유전자에 예측된 드 노보 크립틱 스플라이스 돌연변이를 가지고 있었다(De Rubeis et al., 2014; Sanders et al., 2012). 각 개인은 직계 가족 내에서 유일한 자폐증 사례를 나타냈다. 대부분의 회귀 유전질환의 경우와 마찬가지로, 관련 있는 조직 및 세포 유형(집작하건대 발달 중인 뇌)에는 접근할 수 없었다. 따라서, 본 발명자들은 LCL에서 이들 전사체 다수의 약한 발현을 보상하기 위해 고수준의 mRNA 서열분석(샘플 당 ~ 3억 5천만×150 bp 단일 리드, GTEx 커버리지의 대략 10 배)를 수행하였다. 단순히 최고 예측이 아닌 예측된 크립틱 스플라이스 변이체를 대표하는 세트를 유효성확인하는 것임을 확실히 하기 위하여, 본 발명자들은 비교적 허용되는 임계값 (스플라이스 손실 변이체의 경우 ΔScore > 0.1, 스플라이스 이득 변이체의 경우 ΔScore > 0.5; STAR 방법)을 적용했고, 이러한 기준을 충족시키는 모든 드 노보 변이체에 대해 실험적 유효성확인을 수행했다.

[0331] 관심 유전자에서 RNA-seq 커버리지가 불충분한 8명의 개인을 배제한 후, 본 발명자들은 28명의 환자 중 21명에서 예측된 드 노보 크립틱 스플라이스 돌연변이와 관련된 독특하고 비정상적인 스플라이싱 이벤트를 확인하였다(도 41E 및 도 51a, 도 51b, 도 51c, 도 51d, 도 51e, 도 51f, 도 51g, 도 51h, 도 51i 및 도 51j). 이러한 비정상적인 스플라이싱 이벤트는 심층 LCL RNA-seq를 얻은 다른 35명의 개인과 GTEx 코호트의 149명의 개인에게는 없었다. 21개의 확인된 드 노보 크립틱 스플라이스 돌연변이 중에서, 9건의 신규 접합부 생성, 8건의 엑손 스킵핑 및 4건의 인트론 보유(intron retention), 그리고 더욱 복잡한 스플라이싱 이상이 관찰되었다(도 41F, 도 46A, 도 46B 및 도 46C). 7개의 사례는 전사체의 적절한 발현에도 불구하고 LCL에서 비정상적인 스플라이싱을 나타내지 않았다. 이들의 서브세트가 위양성 예측을 나타낼 수 있지만, 일부 크립틱 스플라이스 돌연변이는 이러한 실험 조건 하에서 LCL에서 관찰될 수 없는 조직 특이적 대체 스플라이싱을 야기할 수 있다.

[0332] RNA-seq 분석의 한계에도 불구하고 자폐 스펙트럼 장애 환자에서 예측된 크립틱 스플라이스 돌연변이의 높은 유효성확인율(75%)은 대부분의 예측이 기능적임을 나타낸다. 그러나 대조군과 비교한 경우의 드 노보 크립틱 스플라이스 변이체의 농축(DDD에서 1.5배 및 ASD에서 1.3배, 도 41A)은 드 노보 단백질 절단 변이체에 대해 관찰된 효과 크기(DDD에서 2.5배, ASD에서 1.7배)의 38%에 불과하다(Iossifov et al., 2014; McRae et al., 2017; De Rubeis et al., 2014). 이를 통해 기능적 크립틱 스플라이스 돌연변이가 고전적인 형태의 단백질 절단 돌연변이(정지-게인, 프레임시프트 및 필수 스플라이스 다이뉴클레오타이드)에 대한 임상적 침투도의 약 50%를 갖는 것을 정량화할 수 있는데, 이는 그들 중 다수는 정상적인 전사체의 생성을 부분적으로만 방해하기 때문이다. 실제로, FECH의 c.315-48T>C (Gouya et al., 2002) 및 GAA의 c.-32-13T>G (Boerkoel et al., 1995)와 같은 멘델 유전질환에서 가장 잘 특성화된 크립틱 스플라이스 돌연변이의 일부는, 온화한 표현형 또는 후기 발병과 관련된 저형질 대립유전자이다. 임상 침투도의 추정치는 비교적 허용되는 임계값(ΔScore ≥ 0.1)을 충족하는 모든 드 노보 변이체에 대해 계산되며, 예측 점수가 더 강한 변이체는 상응하는 높은 침투도를 가질 것으로 예상될 것이다.

[0333] ASD 및 DDD 코호트 전체에 걸쳐 대조군과 비교하는 경우에서 드 노보 돌연변이의 과잉에 근거하여, 드 노보 단백질 절단 변이체에 의해 설명될 수 있는 909건의 사례와 비교하여 250건의 사례는 드 노보 크립틱 스플라이스 돌연변이에 의해 설명될 수 있다(도 41B). 이는, 일단 크립틱 스플라이스 돌연변이의 감소된 침투도가 고려되면, 전체 모집단에서 1인당 회귀 단백질 절단 변이체(~ 11)와 비교하여 회귀 크립틱 스플라이스 돌연변이(~ 5)의 평균 개수에 대한 본 발명자들의 초기 추정치와 일치한다(도 38A). 게놈 전반에 걸친 크립틱 스플라이스 돌연변이의 광범위한 분포는 신경 발달 장애에서 크립틱 스플라이스 돌연변이에 의해 설명된 사례의 비율(9 내지 11%, 도 41B)이 일차 질병 기전이 기능성 단백질의 손실인 다른 회귀 유전 장애에 일반화될 가능성이 있음을 시사한다. 스플라이스 변경 돌연변이의 해석을 용이하게 하기 위해, 본 발명자들은 게놈 전체에서 모든 가능한 단일 뉴클레오타이드 치환에 대한 ΔScore 예측을 미리 계산하고, 이를 과학계에 자료로서 제공한다. 본 발명자들은 이 자료가 이전에 제대로 평가되지 않은 유전자 변이의 근원에 대한 이해를 증진시킬 것이라고 믿는다.

[0334] 구체적인 구현예

[0335] 트레이닝된 아트러스 컨볼루션 신경망을 사용하여 게놈 서열(예를 들어, 뉴클레오타이드 서열 또는 아미노산 서

열)에서 스플라이스 부위를 검출하기 위한 시스템, 방법, 및 제조 물품을 설명한다. 구현의 하나 이상의 피처를 기본 구현과 결합할 수 있다. 상호 배타적이지 않은 구현들은 결합 가능하도록 교시된다. 구현의 하나 이상의 피처는 다른 구현과 결합될 수 있다. 본 개시 내용은 사용자에게 이러한 옵션을 주기적으로 상기시킨다. 이러한 옵션을 반복하는 설명이 일부 구현에서 누락되더라도 이는 이전 부문에서 설명한 조합들을 제한하는 것으로 간주되어서는 안 되며, 이러한 설명은 이하의 각 구현예에 참조로 통합되는 것이다.

- [0336] 본 부문에서는 모듈(들)과 단계(들)라는 용어를 호환 가능하게 사용한다.
- [0337] 개시된 기술의 시스템 구현예는 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는 계층 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위를 식별하는 스플라이스 부위 검출기를 트레이닝시키기 위한 컴퓨터 명령어가 로딩된다.
- [0338] 도 30에 도시된 바와 같이, 시스템은 공여체 스플라이스 부위의 적어도 50000개의 트레이닝 예, 수용체 스플라이스 부위의 적어도 50000개의 트레이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대한 아트리스 컨볼루션 신경망(ACNN)을 트레이닝시킨다. 각각의 트레이닝 예는 각 측면에 적어도 20개의 뉴클레오타이드가 플랭킹, 즉, 측접(flank)된 적어도 하나의 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열이다.
- [0339] ACNN은 트레이닝 가능한 파라미터가 거의 없는 큰 수용장을 허용한다. 아트리스/팽창 컨볼루션은, 아트리스 컨볼루션 레이트 또는 팽창 인자라고도 하는 소정의 단차로 입력값들을 스킵함으로써 커널이 자신의 길이보다 큰 면적에 걸쳐 적용되는 컨볼루션이다. 아트리스/팽창 컨볼루션은, 컨볼루션 동작이 수행될 때 넓은 간격으로 있는 이웃하는 입력 엔트리들(예를 들어, 뉴클레오타이드, 아미노산)이 고려되도록 컨볼루션 필터/커널의 요소들 사이에 간격을 추가한다. 이는 입력에 장거리 컨텍스트 종속성을 통합할 수 있게 한다. 아트리스 컨볼루션은, 인접한 뉴클레오타이드들이 처리될 때 재사용을 위해 부분 컨볼루션 계산을 보존한다.
- [0340] 도 30에 도시된 바와 같이, ACNN을 사용한 트레이닝 예를 평가하기 위해, 시스템은 ACNN에 대한 입력으로서, 적어도 40개의 상류 컨텍스트(upstream context) 뉴클레오타이드 및 적어도 40개의 하류 컨텍스트(downstream context) 뉴클레오타이드가 추가적으로 측접된 표적 뉴클레오타이드 서열을 제공한다.
- [0341] 도 30에 도시된 바와 같이, 평가에 기초하여, ACNN은 표적 뉴클레오타이드 서열의 각 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대해서 트리플렛 점수를 출력으로서 생성한다.
- [0342] 개시된 이러한 시스템 구현예 및 다른 시스템들은 다음 피처들 중 하나 이상을 선택적으로 포함한다. 시스템은, 또한, 개시된 방법과 관련하여 설명된 피처를 포함할 수 있다. 간결성을 위해, 시스템 피처들의 대체 조합은 개별적으로 열거되지 않는다. 시스템, 방법, 및 제조 물품에 적용되는 피처는 기본 피처들의 각각의 범용 클래스 세트에 대하여 반복되지 않는다. 독자는, 이 부문에서 식별되는 피처를 다른 범용 클래스의 기본 피처와 쉽게 결합할 수 있는 방법을 이해할 것이다.
- [0343] 도 25, 도 26 및 도 27에 도시된 바와 같이, 입력은 각 측면에 2500개의 뉴클레오타이드가 측접된 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 이러한 구현에서, 표적 뉴클레오타이드 서열에는 5000개의 상류 컨텍스트 뉴클레오타이드 및 5000개의 하류 컨텍스트 뉴클레오타이드가 추가로 측접한다.
- [0344] 입력은 각 측면에 100개의 뉴클레오타이드가 측접된 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 이러한 구현예에서, 표적 뉴클레오타이드 서열에는 200개의 상류 컨텍스트 뉴클레오타이드 및 200개의 하류 컨텍스트 뉴클레오타이드가 추가로 측접한다.
- [0345] 입력은 각 측면에 500개의 뉴클레오타이드가 측접된 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 이러한 구현예에서, 표적 뉴클레오타이드 서열에는 1000개의 상류 컨텍스트 뉴클레오타이드 및 1000개의 하류 컨텍스트 뉴클레오타이드가 추가로 측접한다.
- [0346] 도 28에 도시된 바와 같이, 시스템은 공여체 스플라이스 부위의 150000개 트레이닝 예, 수용체 스플라이스 부위의 150000개 트레이닝 예 및 비-스플라이싱 부위의 800000000개 트레이닝 예에 대해 ACNN을 트레이닝할 수 있다.
- [0347] 도 19에 도시된 바와 같이, ACNN은 최저에서 최고로 순차적으로 배열된 잔여 블록의 그룹을 포함할 수 있다. 각각의 잔여 블록 그룹은 잔여 블록 내의 컨볼루션 필터의 수, 잔여 블록의 컨볼루션 윈도우 크기, 및 잔여 블록의 아트리스 컨볼루션 레이트에 의해 파라미터화된다.

- [0348] 도 21, 도 22, 도 23 및 도 24에 도시된 바와 같이, ACNN에서, 아트러스 컨볼루션 레이트는 낮은 잔여 블록 그룹에서 높은 잔여 블록 그룹으로 비-기하급수적으로 진행된다.
- [0349] 도 21, 도 22, 도 23 및 도 24에 도시된 바와 같이, ACNN에서, 컨볼루션 윈도우 크기는 잔여 블록의 그룹들 사이에서 달라진다.
- [0350] ACNN은 40개의 상류 컨텍스트 뉴클레오타이드 및 40개의 하류 컨텍스트 뉴클레오타이드가 추가로 축적된 표적 뉴클레오타이드 서열을 포함하는 입력을 평가하도록 구성될 수 있다. 이러한 구현예에서, ACNN은 4개의 잔여 블록의 그룹 하나와 적어도 하나의 스킵 연결을 포함한다. 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 1가지 아트러스 컨볼루션 레이트를 가진다. ACNN의 이러한 구현예는 본 명세서에서 "SpliceNet80"으로 지칭되고 도 21에 도시된다.
- [0351] ACNN은 200개의 상류 컨텍스트 뉴클레오타이드 및 200개의 하류 컨텍스트 뉴클레오타이드가 추가로 축적된 표적 뉴클레오타이드 서열을 포함하는 입력을 평가하도록 구성될 수 있다. 이러한 구현예에서, ACNN은 4개의 잔여 블록의 그룹 적어도 둘과 적어도 둘의 스킵 연결을 포함한다. 제1 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 1가지 아트러스 컨볼루션 레이트를 가진다. 제2 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 4가지 아트러스 컨볼루션 레이트를 가진다. ACNN의 이러한 구현예는 본 명세서에서 "SpliceNet400"으로 지칭되고 도 22에 도시된다.
- [0352] ACNN은 1000개의 상류 컨텍스트 뉴클레오타이드 및 1000개의 하류 컨텍스트 뉴클레오타이드가 추가로 축적된 표적 뉴클레오타이드 서열을 포함하는 입력을 평가하도록 구성될 수 있다. 이러한 구현예에서, ACNN은 4개의 잔여 블록의 그룹 적어도 셋과 적어도 셋의 스킵 연결을 포함한다. 제1 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 1가지 아트러스 컨볼루션 레이트를 가진다. 제2 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 4가지 아트러스 컨볼루션 레이트를 가진다. 제3 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 21가지 컨볼루션 창 크기 및 19가지 아트러스 컨볼루션 레이트를 가진다. ACNN의 이러한 구현예는 본 명세서에서 "SpliceNet2000"으로 지칭되고 도 23에 도시된다.
- [0353] ACNN은 5000개의 상류 컨텍스트 뉴클레오타이드 및 5000개의 하류 컨텍스트 뉴클레오타이드가 추가로 축적된 표적 뉴클레오타이드 서열을 포함하는 입력을 평가하도록 구성될 수 있다. 이러한 구현예에서, ACNN은 4개의 잔여 블록의 그룹 적어도 넷과 적어도 넷의 스킵 연결을 포함한다. 제1 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 1가지 아트러스 컨볼루션 레이트를 가진다. 제2 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 11가지 컨볼루션 창 크기 및 4가지 아트러스 컨볼루션 레이트를 가진다. 제3 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 21가지 컨볼루션 창 크기 및 19가지 아트러스 컨볼루션 레이트를 가진다. 제4 그룹의 각 잔여 블록은 32가지 컨볼루션 필터, 41가지 컨볼루션 창 크기 및 25가지 아트러스 컨볼루션 레이트를 가진다. ACNN의 이러한 구현예는 본 명세서에서 "SpliceNet10000"으로 지칭되고 도 24에 도시된다.
- [0354] 표적 뉴클레오타이드 서열에서 각각의 뉴클레오타이드에 대한 트리플렛 점수는 단일로 합산하기 위해 지수적으로 정규화될 수 있다. 이러한 구현예에서, 시스템은 각각의 트리플렛 점수에서 가장 높은 점수에 기초하여 표적 뉴클레오타이드 내의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류한다.
- [0355] 도 35에 도시된 바와 같이, ACNN 입력의 차원(dimensionality)은 $(C^u + L + C^d) \times 4$ 로 정의될 수 있으며, 여기서 C^u 는 다수의 상류 컨텍스트 뉴클레오타이드, C^d 는 다수의 하류 컨텍스트 뉴클레오타이드, 그리고 L은 표적 뉴클레오타이드 서열 내의 다수의 뉴클레오타이드이다. 일 구현예에서, 입력의 차원은 $(5000 + 5000 + 5000) \times 4$ 이다.
- [0356] 도 35에 도시된 바와 같이, ACNN의 출력의 차원은 $L \times 3$ 으로 정의될 수 있다. 일 구현예에서, 출력의 차원은 5000×3 이다.
- [0357] 도 35에 도시된 바와 같이, 각 잔여 블록 그룹은 선행 입력을 처리함으로써 중간 출력을 생성할 수 있다. 중간 출력의 차원은 $(I - \{(W-1) * D\} * A) \times N$ 으로 정의할 수 있고, 여기서 I는 이전 입력의 차원, W는 잔여 블록의 컨볼루션 창 크기, D는 잔여 블록의 아트러스 컨볼루션 레이트, A는 그룹 내의 아트러스 컨볼루션층의 수, 및 N은 잔여 블록의 컨볼루션 필터의 수이다.
- [0358] 도 32에 도시된 바와 같이, ACNN은 에포크(epoch) 동안 트레이닝 예를 일괄 평가한다. 트레이닝 예는 배치(batch)로 무작위로 샘플링된다. 각 배치에는 미리 정해진 배치 크기가 있다. ACNN은 복수의 에포크(예를 들어,

1 내지 10) 동안 트레이닝 예의 평가를 반복한다.

- [0359] 입력은 2개의 인접한 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 2개의 인접한 표적 뉴클레오타이드는 아데닌(약칭 A) 및 구아닌(약칭 G)일 수 있다. 2개의 인접한 표적 뉴클레오타이드는 구아닌(약칭 G) 및 우라실(약칭 U) 일 수 있다.
- [0360] 시스템은 트레이닝 예를 드물게 인코딩하고 입력으로서 원-핫 인코딩을 제공하는 원-핫 인코더(도 29에 도시됨)를 포함한다.
- [0361] ACNN은 잔여 블록의 수, 스킵 연결(skip connection)의 수 및 잔여 연결(residual connection)의 수에 의해 파라미터화될 수 있다.
- [0362] ACNN은 선행 입력의 공간 및 피쳐 차원을 재구성하는 차원 변경 컨볼루션층을 포함할 수 있다.
- [0363] 도 20에 도시된 바와 같이, 각각의 잔여 블록은 적어도 하나의 일괄 정규화층, 적어도 하나의 정류된 선형 유닛(약칭 ReLU) 층, 적어도 하나의 아트러스 컨볼루션층 및 적어도 하나의 잔여 연결을 포함할 수 있다. 이러한 구현예에서, 각각의 잔여 블록은 2개의 일괄 정규화층, 2개의 ReLU 비선형성 층, 2개의 아트러스 컨볼루션층, 및 하나의 잔여 연결을 포함한다.
- [0364] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0365] 개시된 기술의 다른 시스템 구현예는 병렬로 동작하고 메모리에 연결된 다수의 프로세서 상에서 실행되는 트레이닝된 스플라이스 부위 예측기를 포함한다. 이 시스템은 공여체 스플라이스 부위의 적어도 50000개의 트레이닝 예, 수용체 스플라이스 부위의 적어도 50000개의 트레이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대하여, 다수의 프로세서 상에서 실행되는 아트러스 컨볼루션 신경망(약칭 ACNN)을 트레이닝시킨다. 트레이닝에 사용된 각각의 트레이닝 예는 각 측면에 적어도 400개의 뉴클레오타이드가 축적된 표적 뉴클레오타이드를 포함하는 뉴클레오타이드 서열이다.
- [0366] 시스템은 다수의 프로세서 중 적어도 하나에서 작동하고 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 공급하는 ACNN의 입력 단계를 포함한다. 각각의 표적 뉴클레오타이드에는 각 측면에 적어도 400개의 뉴클레오타이드가 축적되어 있다. 다른 구현예에서, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되고 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 공급하는 ACNN의 입력 모듈을 포함한다.
- [0367] 상기 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 ACNN의 출력 단계를 포함하고, ACNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환한다. 다른 구현예에서, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 ACNN의 출력 모듈을 포함하고, ACNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환한다.
- [0368] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피쳐는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피쳐가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0369] ACNN은 공여체 스플라이스 부위의 150000개 트레이닝 예, 수용체 스플라이스 부위의 150000개 트레이닝 예 및 비-스플라이싱 부위의 800000000개 트레이닝 예에 대해 트레이닝될 수 있다. 시스템의 다른 구현예에서, ACNN은 최저에서 최고로 순차적으로 배열된 잔여 블록의 그룹을 포함한다. 시스템의 또 다른 구현예에서, 각각의 잔여 블록 그룹은 잔여 블록 내의 컨볼루션 필터의 수, 잔여 블록의 컨볼루션 윈도우 크기, 및 잔여 블록의 아트러스 컨볼루션 레이트에 의해 파라미터화된다.
- [0370] ACNN은 최저에서 최고로 순차적으로 배열된 잔여 블록의 그룹을 포함할 수 있다. 각각의 잔여 블록 그룹은 잔여 블록 내의 컨볼루션 필터의 수, 잔여 블록의 컨볼루션 윈도우 크기, 및 잔여 블록의 아트러스 컨볼루션 레이트에 의해 파라미터화된다.
- [0371] ACNN에서, 아트러스 컨볼루션 레이트는 낮은 잔여 블록 그룹에서 높은 잔여 블록 그룹으로 비-기하급수적으로 진행된다. 또한, ACNN에서, 컨볼루션 윈도우 크기는 잔여 블록의 그룹들 사이에서 달라진다.

- [0372] ACNN은 도 18에 도시된 바와 같이 하나 이상의 트레이닝 서버에서 트레이닝될 수 있다.
- [0373] 트레이닝된 ACNN은, 도 18에 도시된 바와 같이, 요청하는 클라이언트로부터 입력 서열을 수신하는 하나 이상의 생성 서버에 배치될 수 있다. 이러한 구현예에서, 생성 서버는, 도 18에 도시된 바와 같이, 클라이언트에게 전송되는 출력을 생성하기 위해 ACNN의 입력 및 출력 단계를 통해 입력 서열을 처리한다. 다른 구현예에서, 생성 서버는, 도 18에 도시된 바와 같이, 클라이언트에게 전송되는 출력을 생성하기 위해 ACNN의 입력 및 출력 모듈을 통해 입력 서열을 처리한다.
- [0374] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비밀시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0375] 개시된 기술의 방법 구현은 계층 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위를 식별하는 스플라이스 부위 검출기를 트레이닝시키는 것을 포함한다.
- [0376] 상기 방법은 각 측면에 적어도 400개의 뉴클레오타이드가 축적된 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 아트라스 컨볼루션 신경망(약칭 ACNN)에 공급하는 단계를 포함한다.
- [0377] ACNN은 공여체 스플라이스 부위의 50000개 이상의 트레이닝 예, 수용체 스플라이스 부위의 50000개 이상의 트레이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대해 트레이닝을 받는다. 트레이닝에 사용된 각각의 트레이닝 예는 각 측면에 400개 이상의 뉴클레오타이드가 축적된 표적 뉴클레오타이드를 포함하는 뉴클레오타이드 서열이다.
- [0378] 상기 방법은 ACNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환하는 단계를 더욱 포함한다.
- [0379] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0380] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비밀시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 메모리 및 메모리에 저장된 명령을 실행하여 전술한 방법을 수행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.
- [0381] 트레이닝된 아트라스 컨볼루션 신경망을 사용하여 계층 서열(예를 들어, 뉴클레오타이드 서열)에서 비정상 스플라이싱을 검출하기 위한 시스템, 방법, 및 제조 물품을 설명한다. 구현의 하나 이상의 피처를 기본 구현과 결합할 수 있다. 상호 배타적이지 않은 구현들은 결합 가능하도록 교시된다. 구현의 하나 이상의 피처는 다른 구현과 결합될 수 있다. 본 개시 내용은 사용자에게 이러한 옵션을 주기적으로 상기시킨다. 이러한 옵션을 반복하는 설명이 일부 구현에서 누락되더라도 이는 이전 부문에서 설명한 조합들을 제한하는 것으로 간주되어서는 안 되며, 이러한 설명은 이하의 각 구현에 참조로 통합되는 것이다.
- [0382] 개시된 기술의 시스템 구현은 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는, 병렬로 작동하고 메모리에 연결된 다수의 프로세서 상에서 실행되는 비정상 스플라이싱 검출기를 구현하도록 컴퓨터 명령이 로딩된다.
- [0383] 도 34에 도시된 바와 같이, 시스템은 다수의 프로세서 상에서 실행되는 트레이닝된 아트라스 컨볼루션 신경망(약칭 ACNN)을 포함한다. ACNN은 트레이닝 가능한 파라미터가 거의 없는 큰 수용장을 허용하는 아트라스/팽창 컨볼루션을 사용하는 컨볼루션 신경망이다. 아트라스/팽창 컨볼루션은, 아트라스 컨볼루션 레이트 또는 팽창 인자라고도 하는 소정의 단차로 입력값들을 스킵함으로써 커널이 자신의 길이보다 큰 면적에 걸쳐 적용되는 컨볼루션이다. 아트라스/팽창 컨볼루션은, 컨볼루션 동작이 수행될 때 넓은 간격으로 이웃하는 입력 엔트리들(예를 들어, 뉴클레오타이드, 아미노산)이 고려되도록 컨볼루션 필터/커널의 요소들 사이에 간격을 추가한다. 이는 입력에 장거리 컨텍스트 종속성을 통합할 수 있게 한다. 아트라스 컨볼루션은, 인접한 뉴클레오타이드들이 처리될 때 재사용을 위해 부분 컨볼루션 계산을 보존한다.
- [0384] 도 34에 도시된 바와 같이, ACNN은 입력 서열에서 표적 뉴클레오타이드를 분류하고, 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 할당한다. 입력 서열은 801개 이상의 뉴클레오타이드를 포함하고 각각의 표적 뉴클레오타이드에는 각 측면에 400개 이상의 뉴클레오타이드가 축적되어 있다.

- [0385] 도 34에 도시된 바와 같이, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 분류자를 포함하는데, 이 분류자는 ACNN을 통해 참조 서열 및 변이체 서열을 처리하여 참조 서열에서의 그리고 변이체 서열에서의 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 생성한다. 참조 서열 및 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 가지며, 각각의 표적 뉴클레오타이드에는 각 측면에 400개 이상의 뉴클레오타이드가 축적되어 있다. 도 33은 참조 서열과 대체/변이체 서열을 도시한다.
- [0386] 도 34에 도시된 바와 같이, 시스템은 참조 서열 및 변이체 서열에서 표적 뉴클레오타이드의 스플라이스 부위 점수의 차이로부터, 변이체 서열을 생성한 변이체가 비정상 스플라이싱을 유발하고 따라서 병원성인지 여부를 결정한다.
- [0387] 개시된 이러한 시스템 구현에 및 다른 시스템들은 다음 피쳐들 중 하나 이상을 선택적으로 포함한다. 시스템은 또한 개시된 방법과 관련하여 설명된 피쳐를 포함할 수 있다. 간결성을 위해, 시스템 피쳐들의 대체 조합은 개별적으로 열거되지 않는다. 시스템, 방법, 및 제조 물품에 적용되는 피쳐는 기본 피쳐들의 각각의 법정 클래스 세트에 대하여 반복되지 않는다. 독자는, 이 부문에서 식별되는 피쳐를 다른 법정 클래스의 기본 피쳐와 쉽게 결합할 수 있는 방법을 이해할 것이다.
- [0388] 도 34에 도시된 바와 같이, 스플라이스 부위 점수의 차이는 참조 서열 및 변이체 서열에서 표적 뉴클레오타이드 사이의 위치에 따라 결정될 수 있다.
- [0389] 도 34에 도시된 바와 같이, 적어도 하나의 표적 뉴클레오타이드 위치에 대해서, 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값을 초과할 때, ACNN은 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류한다.
- [0390] 도 17에 도시된 바와 같이, 적어도 하나의 표적 뉴클레오타이드 위치에 대해서, 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값 미만일 때, ACNN은 변이체를 비정상 스플라이싱을 유발하지 않고 따라서 양성인 것으로 분류한다.
- [0391] 임계값은 복수의 후보 임계값으로부터 결정될 수 있다. 여기에는 양성 공통 변이체에 의해 생성된 제1 세트의 참조 및 변이체 서열쌍을 처리하여 제1 세트의 비정상 스플라이싱 검출을 초래하고, 병원성 회귀 변이체에 의해 생성된 제2 세트의 참조 및 변이체 서열쌍을 처리하여 제2 세트의 비정상 스플라이싱 검출을 초래하며, 분류자에 의해 사용하기 위해, 제2 세트에서 비정상 스플라이싱 검출의 계수치를 최대화하고 제1 세트에서 비정상 스플라이싱 검출의 계수치를 최소화하는 적어도 하나의 임계값을 선택하는 것이 포함된다.
- [0392] 일 구현예에서, ACNN은 자폐 스펙트럼 장애(autism spectrum disorder: ASD)를 유발하는 변이체를 식별한다. 다른 구현예에서, ACNN은 발달 지연 장애(developmental delay disorder: DDD)를 유발하는 변이체를 식별한다.
- [0393] 도 36에 도시된 바와 같이, 참조 서열 및 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 가질 수 있고, 각각의 표적 뉴클레오타이드에는 각 측면에 적어도 5000개의 뉴클레오타이드가 축적될 수 있다.
- [0394] 도 36에 도시된 바와 같이, 참조 서열에서의 표적 뉴클레오타이드의 스플라이스 부위 점수는 ACNN의 제1 출력에서 인코딩될 수 있고 변이체 서열에서의 표적 뉴클레오타이드의 스플라이스 부위 점수는 ACNN의 제2 출력에서 인코딩될 수 있다. 일 구현예에서, 제1 출력은 제1 101×3 매트릭스로서 인코딩되고 제2 출력은 제2 101×3 매트릭스로서 인코딩된다.
- [0395] 도 36에 도시된 바와 같이, 이러한 구현예에서, 제1 101×3 매트릭스의 각 행은 참조 서열에서의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타낸다.
- [0396] 도 36에 도시된 바와 같이, 또한 이러한 구현예에서, 제2 101×3 매트릭스의 각 행은 변이체 서열의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타낸다.
- [0397] 도 36에 도시된 바와 같이, 일부 구현예들에서, 제1 101×3 매트릭스 및 제2 101×3 매트릭스의 각 행에서의 스플라이스 부위 점수는 단일로 합산하기 위해 지수적으로 정규화될 수 있다.
- [0398] 도 36에 도시된 바와 같이, 분류자는 제1 101×3 행렬과 제2 101×3 행렬의 행-대-행 비교를 수행하고, 행 단위로, 스플라이스 부위 점수의 분포의 변화를 결정할 수 있다. 행-대-행 비교의 적어도 하나의 경우에 있어서, 분

포의 변화가 미리 결정된 임계값을 초과할 때, ACNN은 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류한다.

- [0399] 시스템은 참조 서열 및 변이체 서열을 최소화해 인코딩하는 원-핫 인코더(도 29에 도시됨)를 포함한다.
- [0400] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0401] 다른 구현에는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현에는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0402] 개시된 기술의 방법 구현은 비정상 스플라이싱을 유발하는 게놈 변이체를 검출하는 것을 포함한다.
- [0403] 상기 방법은 표적 하위서열(sub-sequence)에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 입력 서열의 표적 하위서열에서 차등 스플라이싱 패턴을 검출하도록 트레이닝된 아트러스 컨볼루션 신경망(약칭 ACNN)을 통해 참조 서열을 처리하는 단계를 포함한다.
- [0404] 상기 방법은, 처리에 기초하여, 참조 표적 하위서열에서의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 참조 표적 하위서열에서 제1 차등 스플라이싱 패턴을 검출하는 단계를 포함한다.
- [0405] 상기 방법은 ACNN을 통해 변이체 서열을 처리하는 단계를 포함한다. 변이체 서열과 참조 서열은 변이체 표적 하위서열에 위치한 적어도 하나의 변이체 뉴클레오타이드만큼 상이하다.
- [0406] 상기 방법은, 처리에 기초하여, 변이체 표적 하위서열의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 변이체 표적 하위서열에서 제2 차등 스플라이싱 패턴을 검출하는 단계를 포함한다.
- [0407] 상기 방법은 뉴클레오타이드 단위 기초로 참조 표적 하위서열과 변이체 표적 하위서열의 스플라이스 부위 분류를 비교함으로써 제1 차등 스플라이싱 패턴과 제2 차등 스플라이싱 패턴 사이의 차이를 결정하는 단계를 포함한다.
- [0408] 그 차이가 미리 결정된 임계값을 초과할 때, 상기 방법은 변이체를 비정상 스플라이싱을 유발하며 따라서 병원성인 것으로 분류하고 그 분류를 메모리에 저장하는 단계를 포함한다.
- [0409] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0410] 차등 스플라이싱 패턴은 표적 하위서열에서 스플라이싱 이벤트 발생의 위치 분포를 식별할 수 있다. 스플라이싱 이벤트의 예는 크립틱 스플라이싱, 엑손 스킵핑, 상호 배타적인 엑손(mutually exclusive exon), 대체 공여체 부위, 대체 수용체 부위 및 인트론 보유 중 적어도 하나를 포함한다.
- [0411] 참조 표적 하위서열 및 변이체 표적 하위서열은 뉴클레오타이드 위치에 대해 정렬될 수 있고, 적어도 하나의 변이체 뉴클레오타이드만큼 상이할 수 있다.
- [0412] 참조 표적 하위서열 및 변이체 표적 하위서열은 각각 적어도 40개의 뉴클레오타이드를 가질 수 있고 각각 각 측면에 적어도 40개의 뉴클레오타이드가 축적될 수 있다.
- [0413] 참조 표적 하위서열 및 변이체 표적 하위서열은 각각 적어도 101개의 뉴클레오타이드를 가질 수 있고 각각 각 측면에 적어도 5000개의 뉴클레오타이드가 축적될 수 있다.
- [0414] 변이체 목표 하위서열은 두 개의 변이체를 포함할 수 있다.
- [0415] 다른 구현에는, 전술한 방법을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현에는 메모리 및 메모리에 저장된 명령을 실행하여 전술한 방법을 수행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

- [0416] 트레이닝된 컨볼루션 신경망을 사용하여 계층 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위와 비정상 스플라이싱을 검출하기 위한 시스템, 방법, 및 제조 물품을 설명한다. 구현의 하나 이상의 피처를 기본 구현과 결합할 수 있다. 상호 배타적이지 않은 구현들은 결합 가능하도록 교시된다. 구현의 하나 이상의 피처는 다른 구현과 결합될 수 있다. 본 개시 내용은 사용자에게 이러한 옵션을 주기적으로 상기시킨다. 이러한 옵션을 반복하는 설명이 일부 구현에서 누락되더라도 이는 이전 부문에서 설명한 조합들을 제한하는 것으로 간주되어서는 안 되며, 이러한 설명은 이하의 각 구현에 참조로 통합되는 것이다.
- [0417] 개시된 기술의 시스템 구현은 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는 계층 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위를 식별하는 스플라이스 부위 검출기를 트레이닝시키기 위한 컴퓨터 명령이 로딩된다.
- [0418] 시스템은 공여체 스플라이스 부위의 적어도 50000개의 트레이닝 예, 수용체 스플라이스 부위의 적어도 50000개의 트레이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대해 컨볼루션 신경망(약칭 CNN)을 트레이닝시킨다. 각 트레이닝 예는 각 측면에 적어도 20개의 뉴클레오타이드가 축적된 하나 이상의 표적 뉴클레오타이드를 포함하는 표적 뉴클레오타이드 서열이다.
- [0419] CNN을 사용하여 트레이닝 예를 평가하기 위해, 시스템은, CNN에 입력으로서, 적어도 40개의 상류 컨텍스트 뉴클레오타이드 및 적어도 40개의 하류 컨텍스트 뉴클레오타이드가 추가적으로 축적된 표적 뉴클레오타이드 서열을 제공한다.
- [0420] 평가에 기초하여, 이어 CNN은 표적 뉴클레오타이드 서열의 각 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대해서 트리플렛 점수를 출력으로서 생성한다.
- [0421] 개시된 이러한 시스템 구현에 및 다른 시스템들은 다음 피처들 중 하나 이상을 선택적으로 포함한다. 시스템은 또한 개시된 방법과 관련하여 설명된 피처를 포함할 수 있다. 간결성을 위해, 시스템 피처들의 대체 조합은 개별적으로 열거되지 않는다. 시스템, 방법, 및 제조 물품에 적용되는 피처는 기본 피처들의 각각의 범용 클래스 세트에 대하여 반복되지 않는다. 독자는, 이 부문에서 식별되는 피처를 다른 범용 클래스의 기본 피처와 쉽게 결합할 수 있는 방법을 이해할 것이다.
- [0422] 입력은 각 측면에 100개의 뉴클레오타이드가 축적된 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 이러한 구현에서, 표적 뉴클레오타이드 서열에는 200개의 상류 컨텍스트 뉴클레오타이드 및 200개의 하류 컨텍스트 뉴클레오타이드가 추가로 축적한다.
- [0423] 도 28에 도시된 바와 같이, 시스템은 공여체 스플라이스 부위의 150000개 이상의 트레이닝 예, 수용체 스플라이스 부위의 150000개 이상의 트레이닝 예 및 비-스플라이싱 부위의 1000000개 이상의 트레이닝 예에 대해 CNN을 트레이닝시킬 수 있다.
- [0424] 도 31에 도시된 바와 같이, CNN은 컨볼루션층의 수, 컨볼루션 필터의 수, 및 서브샘플링층(예를 들어, 최대 풀링 및 평균 풀링)의 수에 의해 파라미터화된다.
- [0425] 도 31에 도시된 바와 같이, CNN은 하나 이상의 완전 연결층(more fully-connected layer)과 최종 분류층(terminal classification layer)을 포함할 수 있다.
- [0426] CNN은 선행 입력의 공간 및 피처 차원을 재구성하는 차원 변경 컨볼루션층을 포함할 수 있다.
- [0427] 표적 뉴클레오타이드 서열에서 각각의 뉴클레오타이드에 대한 트리플렛 점수는 단일로 합산하기 위해 지수적으로 정규화될 수 있다. 이러한 구현예에서, 시스템은 각각의 트리플렛 점수에서 가장 높은 점수에 기초하여 표적 뉴클레오타이드 내의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류한다.
- [0428] 도 32에 도시된 바와 같이, CNN은 에포크(epoch) 동안 트레이닝 예를 일괄 평가한다. 트레이닝 예는 배위로 무작위로 샘플링된다. 각 배취에는 미리 정해진 배취 크기가 있다. CNN은 복수의 에포크(예를 들어, 1-10) 동안 트레이닝 예의 평가를 반복한다.
- [0429] 입력은 2개의 인접한 표적 뉴클레오타이드를 갖는 표적 뉴클레오타이드 서열을 포함할 수 있다. 2개의 인접한 표적 뉴클레오타이드는 아데닌(약칭 A) 및 구아닌(약칭 G)일 수 있다. 2개의 인접한 표적 뉴클레오타이드는 구아닌(약칭 G) 및 우라실(약칭 U) 일 수 있다.
- [0430] 시스템은 트레이닝 예를 드물게 인코딩하고 입력으로서 원-핫 인코딩을 제공하는 원-핫 인코더(도 32에 도시

됨)를 포함한다.

- [0431] CNN은 잔여 블록의 수, 스킵 연결의 수 및 잔여 연결의 수에 의해 파라미터화될 수 있다.
- [0432] 각각의 잔여 블록은 적어도 1개의 일괄 정규화층, 적어도 1개의 정류된 선형 유닛(약칭 ReLU) 층, 적어도 1개의 차원 변경층 및 적어도 1개의 잔여 연결을 포함할 수 있다. 각각의 잔여 블록은 2개의 일괄 정규화층, 2개의 ReLU 비선형성 층, 2개의 차원 변경층, 및 1개의 잔여 연결을 포함한다.
- [0433] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0434] 다른 구현에는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비밀시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현에는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0435] 개시된 기술의 다른 시스템 구현에는 병렬로 동작하고 메모리에 연결된 다수의 프로세서 상에서 실행되는 트레이닝된 스플라이스 부위 예측기를 포함한다. 이 시스템은 공여체 스플라이스 부위의 적어도 50000개의 트레이닝 예, 수용체 스플라이스 부위의 적어도 50000개의 트레이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대하여, 다수의 프로세서 상에서 실행되는 컨볼루션 신경망(약칭 CNN)을 트레이닝시킨다. 트레이닝에 사용된 각각의 트레이닝 예는 각 측면에 적어도 400개의 뉴클레오타이드가 측정된 표적 뉴클레오타이드를 포함하는 뉴클레오타이드 서열이다.
- [0436] 시스템은 다수의 프로세서 중 적어도 하나에서 작동하고 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 공급하는 CNN의 입력 단계를 포함한다. 각각의 표적 뉴클레오타이드에는 각 측면에 적어도 400개의 뉴클레오타이드가 측정되어 있다. 다른 구현예에서, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되고 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 공급하는 CNN의 입력 모듈을 포함한다.
- [0437] 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 CNN의 출력 단계를 포함하고, CNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환한다. 다른 구현예에서, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 CNN의 출력 모듈을 포함하고, CNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환한다.
- [0438] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0439] CNN은 공여체 스플라이스 부위의 150000개 트레이닝 예, 수용체 스플라이스 부위의 150000개 트레이닝 예 및 비-스플라이싱 부위의 800000000개 트레이닝 예에 대해 트레이닝될 수 있다.
- [0440] CNN은 하나 이상의 트레이닝 서버에서 트레이닝될 수 있다.
- [0441] 트레이닝된 CNN은 요청하는 클라이언트로부터 입력 서열을 수신하는 하나 이상의 생성 서버에 배치될 수 있다. 이러한 구현예에서, 생성 서버는 클라이언트에게 전송되는 출력을 생성하기 위해 CNN의 입력 및 출력 단계를 통해 입력 서열을 처리한다. 다른 구현예에서, 생성 서버는 클라이언트에게 전송되는 출력을 생성하기 위해 CNN의 입력 및 출력 모듈을 통해 입력 서열을 처리한다.
- [0442] 다른 구현에는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비밀시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현에는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0443] 개시된 기술의 방법 구현은 게놈 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위를 식별하는 스플라이스 부위 검출기를 트레이닝시키는 것을 포함한다. 상기 방법은 각각 각 측면에 적어도 400개의 뉴클레오타이드가 측정된 표적 뉴클레오타이드의 평가를 위해 적어도 801개의 뉴클레오타이드의 입력 서열을 컨볼루션 신경망(약칭 CNN)에 공급하는 것을 포함한다.
- [0444] CNN은 공여체 스플라이스 부위의 50000개 이상의 트레이닝 예, 수용체 스플라이스 부위의 50000개 이상의 트레

이닝 예 및 비-스플라이싱 부위의 적어도 100000개의 트레이닝 예에 대해 트레이닝을 받는다. 트레이닝에 사용된 각각의 트레이닝 예는 각 측면에 400개 이상의 뉴클레오타이드가 측정된 표적 뉴클레오타이드를 포함하는 뉴클레오타이드 서열이다.

- [0445] 상기 방법은 CNN에 의한 분석을 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 분류 점수로 변환하는 단계를 더욱 포함한다.
- [0446] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0447] 다른 구현에는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현에는 메모리 및 메모리에 저장된 명령을 실행하여 전술한 방법을 수행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.
- [0448] 개시된 기술의 또 다른 시스템 구현은 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는, 병렬로 작동하고 메모리에 연결된 다수의 프로세서 상에서 실행되는 비정상 스플라이싱 검출기를 구현하도록 컴퓨터 명령이 로딩된다.
- [0449] 시스템은 다수의 프로세서 상에서 실행되는 트레이닝된 컨볼루션 신경망(약칭 CNN)을 포함한다.
- [0450] 도 34에 도시된 바와 같이, CNN은 입력 서열에서 표적 뉴클레오타이드를 분류하고, 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 할당한다. 입력 서열은 801개 이상의 뉴클레오타이드를 포함하고 각각의 표적 뉴클레오타이드에는 각 측면에 400개 이상의 뉴클레오타이드가 측정되어 있다.
- [0451] 도 34에 도시된 바와 같이, 시스템은 다수의 프로세서 중 적어도 하나에서 실행되는 분류자를 또한 포함하는데, 이 분류자는 CNN을 통해 참조 서열 및 변이체 서열을 처리하여 참조 서열 및 변이체 서열에서 각각의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 생성한다. 참조 서열 및 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 가지며, 각각의 표적 뉴클레오타이드에는 각 측면에 400개 이상의 뉴클레오타이드가 측정되어 있다.
- [0452] 도 34에 도시된 바와 같이, 이 시스템은 참조 서열 및 변이체 서열에서 표적 뉴클레오타이드의 스플라이스 부위 점수의 차이로부터, 변이체 서열을 생성한 변이체가 비정상 스플라이싱을 유발하고 따라서 병원성인지 여부를 결정한다.
- [0453] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0454] 스플라이스 부위 점수의 차이는 참조 서열 및 변이체 서열에서 표적 뉴클레오타이드 사이의 위치에 따라 결정될 수 있다.
- [0455] 적어도 하나의 표적 뉴클레오타이드 위치에 대해서, 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값을 초과할 때, CNN은 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류한다.
- [0456] 적어도 하나의 표적 뉴클레오타이드 위치에 대해서, 스플라이스 부위 점수의 전체 최대 차이가 미리 결정된 임계값 미만일 때, CNN은 변이체를 비정상 스플라이싱을 유발하지 않고 따라서 양성인 것으로 분류한다.
- [0457] 임계값은 복수의 후보 임계값으로부터 결정될 수 있다. 여기에는 양성 공통 변이체에 의해 생성된 제1 세트의 참조 및 변이체 서열쌍을 처리하여 제1 세트의 비정상 스플라이싱 검출을 생성하고, 병원성 회귀 변이체에 의해 생성된 제2 세트의 참조 및 변이체 서열쌍을 처리하여 제2 세트의 비정상 스플라이싱 검출을 생성하며, 분류자에 의해 사용하기 위해, 제2 세트에서 비정상 스플라이싱 검출의 계수치를 최대화하고 제1 세트에서 비정상 스플라이싱 검출의 계수치를 최소화하는 적어도 하나의 임계값을 선택하는 것이 포함된다.
- [0458] 일 구현예에서, CNN은 자폐 스펙트럼 장애(약칭 ASD)를 유발하는 변이체를 식별한다. 다른 구현예에서, CNN은 발달 지연 장애(약칭 DDD)를 유발하는 변이체를 식별한다.
- [0459] 참조 서열 및 변이체 서열은 각각 적어도 101개의 표적 뉴클레오타이드를 가질 수 있고, 각 표적 뉴클레오타이드

드에는 각 측면에 적어도 1000개의 뉴클레오타이드가 측정될 수 있다.

- [0460] 참조 서열에서 표적 뉴클레오타이드의 스플라이스 부위 점수는 CNN의 제1 출력에서 인코딩될 수 있고 변이체 서열에서 표적 뉴클레오타이드의 스플라이스 부위 점수는 CNN의 제2 출력에서 인코딩될 수 있다. 일 구현예에서, 제1 출력은 제1 101×3 매트릭스로 인코딩되고 제2 출력은 제2 101×3 매트릭스로 인코딩된다.
- [0461] 이러한 구현예에서, 제1 101×3 매트릭스의 각 행은 참조 서열의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타낸다.
- [0462] 또한 이러한 구현예에서, 제2 101×3 매트릭스의 각 행은 변이체 서열의 표적 뉴클레오타이드가 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위일 가능성에 대한 스플라이스 부위 점수를 독자적으로 나타낸다.
- [0463] 일부 구현예들에서, 제1 101×3 매트릭스 및 제2 101×3 매트릭스의 각 행에서의 스플라이스 부위 점수는 단일로 합산하기 위해 지수적으로 정규화될 수 있다.
- [0464] 분류자는 제1 101×3 행렬과 제2 101×3 행렬의 행-대-행 비교를 수행하고, 행 단위로 스플라이스 부위 점수의 분포 변화를 결정할 수 있다. 행-대-행 비교의 적어도 하나의 경우에 있어서, 분포의 변화가 미리 결정된 임계값을 초과할 때, CNN은 변이체를 비정상 스플라이싱을 유발하고 따라서 병원성인 것으로 분류한다.
- [0465] 시스템은 참조 서열 및 변이체 서열을 최소하게 인코딩하는 원-핫 인코더(도 29에 도시됨)를 포함한다.
- [0466] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.
- [0467] 개시된 기술의 방법 구현은 비정상 스플라이싱을 유발하는 계층 변이체를 검출하는 것을 포함한다.
- [0468] 상기 방법은 표적 하위서열에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 입력 서열의 표적 하위서열에서 차등 스플라이싱 패턴을 검출하도록 트레이닝된 아트리스 컨볼루션 신경망(약칭 CNN)을 통해 참조 서열을 처리하는 단계를 포함한다.
- [0469] 상기 방법은, 처리에 기초하여, 참조 표적 하위서열에서의 각각의 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 참조 표적 하위서열에서 제1 차등 스플라이싱 패턴을 검출하는 단계를 포함한다.
- [0470] 상기 방법은 CNN을 통해 변이체 서열을 처리하는 단계를 포함한다. 변이체 서열 및 참조 서열은 변이체 표적 하위서열에 위치한 적어도 하나의 변이체 뉴클레오타이드만큼 상이하다.
- [0471] 상기 방법은, 처리에 기초하여, 변이체 표적 하위서열의 각 뉴클레오타이드를 공여체 스플라이스 부위, 수용체 스플라이스 부위 또는 비-스플라이싱 부위로 분류함으로써 변이체 표적 하위서열에서 제2 차등 스플라이싱 패턴을 검출하는 단계를 포함한다.
- [0472] 상기 방법은 뉴클레오타이드 단위 기초로 참조 표적 하위서열과 변이체 표적 하위서열의 스플라이스 부위 분류를 비교함으로써 제1 차등 스플라이싱 패턴과 제2 차등 스플라이싱 패턴 사이의 차이를 결정하는 단계를 포함한다.
- [0473] 그 차이가 미리 결정된 임계값을 초과할 때, 상기 방법은 변이체를 비정상 스플라이싱을 유발하며 따라서 병원성인 것으로 분류하고 그 분류를 메모리에 저장하는 단계를 포함한다.
- [0474] 다른 시스템 및 방법 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피쳐는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피쳐가 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.
- [0475] 차등 스플라이싱 패턴은 표적 하위서열에서 스플라이싱 이벤트 발생의 위치 분포를 식별할 수 있다. 스플라이싱 이벤트의 예는 크립틱 스플라이싱, 엑손 스킵핑, 상호 배타적인 엑손들, 대체 공여체 부위, 대체 수용체 부위 및 인트론 보유 중 적어도 하나를 포함한다.
- [0476] 참조 표적 하위서열 및 변이체 표적 하위서열은 뉴클레오타이드 위치에 대해 정렬될 수 있고, 적어도 하나의 변

이체 뉴클레오타이드만큼 상이할 수 있다.

[0477] 참조 표적 하위서열 및 변이체 표적 하위서열은 각각 40개 이상의 뉴클레오타이드를 가질 수 있고 각각 각 측면에 40개 이상의 뉴클레오타이드가 축적될 수 있다.

[0478] 참조 표적 하위서열 및 변이체 표적 하위서열은 각각 적어도 101개의 뉴클레오타이드를 가질 수 있고 각각 각 측면에 1000개 이상의 뉴클레오타이드가 축적될 수 있다.

[0479] 변이체 목표 하위서열은 두 개의 변이체를 포함할 수 있다.

[0480] 다른 구현예는, 전술한 방법을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 메모리 및 메모리에 저장된 명령을 실행하여 전술한 방법을 수행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0481] 전술한 설명은 개시된 기술의 제조 및 이용을 가능하게 하기 위해 제시된다. 개시된 구현들에 대한 다양한 변경들이 명백할 것이며, 본 명세서에서 정의된 일반적인 원리들은 개시된 기술의 사상 및 범위를 벗어나지 않고 다른 구현들 및 활용에 적용될 수 있다. 따라서, 개시된 기술은 도시된 구현으로 제한되도록 의도된 것이 아니라, 본 명세서에 개시된 원리 및 특징과 일치하는 가장 넓은 범위에 따라야한다. 개시된 기술의 범위는 첨부된 청구 범위에 의해 정의된다.

[0482] **유전자당 농축 분석**

[0483] 도 57은 유전자당 농축 분석의 일 구현예를 도시한다. 일 구현예에서, 비정상 스플라이싱 검출기는 비정상적인 스플라이싱을 유발하는 것으로 결정된 변이체의 병원성을 결정하는 유전자당 농축 분석을 구현하도록 추가로 구성된다. 유전 장애를 갖는 개체의 코호트로부터 샘플링된 특정 유전자에 대해서, 유전자당 농축 분석은, 트레이닝된 ACNN을 적용하여 비정상 스플라이싱을 유발하는 특정 유전자의 후보 변이체를 식별하는 것, 후보 변이체의 관찰된 트라이뉴클레오타이드 돌연변이율을 합산하고 그 합에 전달 계수치(transmission count) 및 코호트의 크기를 곱한 것에 기초하여 특정 유전자에 대한 돌연변이의 베이스라인 수를 결정하고, 트레이닝된 ACNN을 적용하여 비정상 스플라이싱을 유발하는 특정 유전자에서 드 노보 변이체를 식별하는 것, 그리고 돌연변이의 베이스라인 수를 드 노보 변이체의 계수치와 비교하는 것을 포함한다. 비교의 출력에 기초하여, 유전자당 농축 분석은, 특정 유전자가 유전 장애에 연관되어 있음과 드 노보 변이체가 병원성을 결정한다. 일부 구현예에서, 유전 장애는 자폐 스펙트럼 장애(약칭 ASD)이다. 다른 구현예에서, 유전 장애는 발달 지연 장애(약칭 DDD)이다.

[0484] 도 57에 도시된 예에서, 특정 유전자의 5개의 후보 변이체는 비정상 스플라이싱 검출기에 의해 비정상 스플라이싱을 유발하는 것으로 분류되었다. 이들 5개의 후보 변이체는 각각 10^{-8} , 10^{-2} , 10^{-1} , 10^5 및 10^1 의 트라이뉴클레오타이드 돌연변이율을 관찰하였다. 특정 유전자에 대한 돌연변이의 베이스라인 수는, 5개의 후보 변이체의 각각의 관찰된 트라이뉴클레오타이드 돌연변이율을 합산하고 그 합을 코호트(1000)의 크기 및 전달/염색체 계수치(2)와 곱한 것에 기초하여 10^{-5} 인 것으로 결정된다. 이것은 이어서 드 노보 변이체 계수치(3)와 비교된다.

[0485] 일부 구현예에서, 비정상 스플라이싱 검출기는, 또한, p-값을 출력으로서 생성하는 통계 테스트를 사용하여 비교를 수행하도록 구성된다.

[0486] 다른 구현예에서, 비정상 스플라이싱 검출기는, 또한, 돌연변이의 베이스라인 수를 드 노보 변이체의 계수치와 비교하고 비교의 출력에 기초하여 특정 유전자가 유전 장애와 연관되지 않음과 드 노보 변이체가 양성임을 결정하도록 구성된다.

[0487] 일 구현예에서, 후보 변이체의 적어도 일부는 단백질 절단 변이체이다.

[0488] 다른 일 구현예에서, 후보 변이체의 적어도 일부는 미스센스 변이체이다.

[0489] **계놈 전체 농축 분석**

[0490] 도 58은 계놈 전체 농축 분석의 일 구현예를 도시한다. 다른 일 구현예에서, 비정상 스플라이싱 검출기는, 또한, 비정상 스플라이싱을 유발하는 것으로 결정된 변이체의 병원성을 결정하는 계놈 전체 농축 분석을 구현하도록 구성된다. 계놈 전체 농축 분석은, 트레이닝된 ACNN을 적용하여 건강한 개체의 코호트로부터 샘플링된 복수의 유전자 내에서 비정상 스플라이싱을 유발하는 드 노보 변이체들의 제1 세트를 식별하는 것, 트레이닝된 ACNN을 적용하여 유전 장애가 있는 개체의 코호트로부터 샘플링된 복수의 유전자 내에서 비정상 스플라이싱을 유발하는 드 노보 변이체들의 제2 세트를 식별하는 것, 및 제1 및 제2 세트의 각 계수치를 비교하고, 비교의 출

력에 기초하여 드 노보 변이체들의 제2 세트가 유전 장애가 있는 개체의 코호트에 농축되어 있고 이에 따라 병 원성임을 결정하는 것을 포함한다. 일부 구현예에서, 유전 장애는 자폐 스펙트럼 장애(약칭 ASD)이다. 다른 구현예에서, 유전 장애는 발달 지연 장애(약칭 DDD)이다.

[0491] 일부 구현예에서, 비정상 스플라이싱 검출기는, 또한, p-값을 출력으로서 생성하는 통계 테스트를 사용하여 비교를 수행하도록 구성된다. 일 구현예에서, 비교는 각각의 코호트 크기에 의해 추가로 파라미터화된다.

[0492] 일부 구현예에서, 비정상 스플라이싱 검출기는, 또한, 제1 및 제2 세트의 각각의 계수치를 비교하고, 비교의 출력에 기초하여 드 노보 변이체들의 제2 세트가 유전 장애를 가진 개체의 코호트에 농축되어 있지 않으며 이에 따라 양성임을 결정하도록 구성된다.

[0493] 도 58에 도시된 예에서, 건강한 코호트에서의 돌연변이율(0.001) 및 영향을 받는 코호트에서의 돌연변이율(0.004)은 개체별 돌연변이 양(4)과 함께 도시되어 있다.

[0494] **논의**

[0495] 심각한 유전 장애가 있는 환자에서 엑손 서열분석의 제한된 진단율에도 불구하고, 임상 서열분석은 최소한 코딩 돌연변이에 중점을 두었으며, 해석의 어려움으로 인해 비코딩 게놈의 변화는 대개 무시했다. 여기서 본 발명자들은 1차 뉴클레오타이드 서열에서 스플라이싱을 정확하게 예측함으로써, 엑손 및 인트론의 정상적인 패턴을 방해하여 그 결과 생성된 단백질에 심각한 결과를 초래하는 비코딩 돌연변이를 식별하는 심층 학습망을 소개한다. 본 발명자들은 예측된 크립틱 스플라이스 돌연변이가 RNA-seq에 의해 높은 속도로 유효성확인되고, 인간 개체군에서 강력하게 유해하며, 희귀 유전질환의 주요 원인임을 보여준다.

[0496] 스플라이소좀의 인 실리코 모델로 심층 학습망을 사용함으로써, 본 발명자들은 스플라이소좀이 인 비보로 뛰어난 정밀도를 달성할 수 있도록 하는 특이성 결정자를 재구성할 수 있었다. 본 발명자들은 스플라이싱의 기전에 대한 지난 40년간의 연구에서 이루어진 많은 발견을 재확인하고, 스플라이소좀이 다수의 단거리 및 장거리 특이성 결정자들을 그 결정에서 통합한다는 것을 보여준다. 특히, 본 발명자들은 대부분의 스플라이스 모티프의 인 지된 퇴행성이 엑손/인트론 길이 및 뉴클레오솜 포지셔닝과 같은 장거리 결정자의 존재에 의해 설명되며, 이는 보상하는 것 이상으로 모티프 레벨에서 추가적인 특이성을 불필요하게 하는 것을 발견한다. 본 발명자들의 연구 결과는 단순히 블랙박스 분류자 역할을 하기보다는 생물학적 통찰력을 제공하기 위한 심층 학습 모델의 유망함을 입증한다.

[0497] 심층 학습은 생물학에서 비교적 새로운 기술이며, 잠재적인 타협이 없는 것은 아니다. 심층 학습 모델은 서열에서 피처를 자동으로 추출하는 방법을 학습함으로써, 인간 전문가가 잘 설명하지 않은 새로운 서열 결정자를 활용할 수 있지만, 모델이 스플라이소좀의 실제 반응을 반영하지 않는 피처를 통합할 위험이 또한 있다. 이러한 무관계한 피처는 주석이 달린 엑손-인트론 경계를 예측하는 외관상 정확도를 증가시킬 수 있지만, 유전적 변이에 의해 유도된 임의의 서열 변화의 스플라이스 변경 효과를 예측하는 정확도는 감소시킬 것이다. 변이체의 정확한 예측은 모델이 진정한 생물학으로 일반화될 수 있다는 가장 강력한 증거를 제공하기 때문에, 본 발명자들은 RNA-seq, 인간 개체군에서의 자연 선택 및 사례 대 대조군 코호트에서의 드 노보 변이체라고 하는 세 가지 완전 직교하는 방법을 사용하여 예측된 스플라이스 변경 변이체의 유효성확인을 제공한다. 이것이 무관계한 피처를 모델에 포함시키는 것을 완전히 배제하지는 못하지만, 이렇게 만들어진 모델은 진정한 스플라이싱의 생물학에 충분히 충실하여 유전병 환자의 크립틱 스플라이스 돌연변이를 식별하는 것과 같은 실제 응용에 중요한 가치가 있을 것으로 보인다.

[0498] 다른 클래스의 단백질 절단 돌연변이와 비교하여, 크립틱 스플라이스 돌연변이의 특히 흥미로운 측면은 불완전하게 침투성인 스플라이스 변경 변이체로 인한 대체 스플라이싱의 광범위한 현상인데, 이는 대체 스플라이스 부위에 비해 표준 스플라이스 부위를 약화시키는 경향이 있어서, RNA-seq 데이터에서 비정상 및 정상 전사체의 혼합물을 생성시키게 된다. 이들 변이체가 종종 조직-특이적 대체 스플라이싱을 유도한다는 관찰은 신규한 대체 스플라이싱 다양성을 생성하는데 있어서 크립틱 스플라이스 돌연변이에 의해 수행되는 예기치 않은 역할을 강조한다. 잠재적인 미래 방향은 관련 조직의 RNA-seq으로부터의 스플라이스 접합부 주석에 대해 심층 학습 모델을 트레이닝시켜 대체 스플라이싱의 조직별 모델을 얻는 것이다. RNA-seq 데이터로부터 직접 유도된 주석에 대해 망을 트레이닝시키는 것은 또한 GENCODE 주석의 빈틈을 메우는 데 도움이 되고, 이는 변이체 예측에 대한 모델의 성능을 향상시킨다(도 52A 및 도 52B).

[0499] 비코딩 게놈에서의 돌연변이가 어떻게 인간의 질병으로 이어지는지에 대한 본 발명자들의 이해는 여전히 완전함과 거리가 멀다. 유년기 신경 발달 장애에서 침투성 드 노보 크립틱 스플라이스 돌연변이의 발견은 비코딩 게놈

의 개선된 해석이 심각한 유전 장애를 가진 환자에게 직접 혜택을 줄 수 있음을 보여준다. 크립틱 스플라이스 돌연변이는 또한 암에서 주요한 역할을 하며(Jung et al., 2015; Sanz et al., 2010; Supek et al., 2014), 스플라이스 인자의 재발성 체세포 돌연변이는 스플라이싱 특이성에 광범위한 변화를 일으키는 것으로 나타났다(Graubert et al., 2012; Shirai et al., 2015; Yoshida et al., 2011). 상이한 조직 및 세포 컨텍스트에서의, 특히 스플라이소좀의 단백질에 직접적으로 영향을 미치는 돌연변이의 경우에서의 스플라이싱 조절을 이해하기 위해 아직 많은 연구가 이루어져야 한다. 서열-특이적 방식으로 스플라이싱 결함을 잠재적으로 표적화할 수 있는 올리고뉴클레오타이드 요법에서의 최근의 발전에 비추어(Finkel et al., 2017), 이 놀라운 과정을 지배하는 조절 기전에 대한 더 큰 이해는 치료적 개입을 위한 새로운 후보자에게 길을 열어줄 수 있다.

- [0500] 도 37A, 37b, 37c, 37d, 37e, 37f, 37g 및 37h는 심층 학습을 통해 1차 서열로부터 스플라이싱을 예측하는 일 구현예를 도시한다.
- [0501] 도 37A와 관련하여, 전구체-mRNA 전사체의 각 위치에 대해, SpliceNet-10k는 플랭킹 서열의 10,000개의 뉴클레오타이드를 입력으로서 사용하여 그 위치가 스플라이스 수용체인지 공여체인지 또는 둘 다 아닌지를 예측한다.
- [0502] 도 37B와 관련하여, MaxEntScan(상단) 및 SpliceNet-10k(하단)를 사용하여 점수가 매겨진 CFTR 유전자에 대한 전체 전구체-mRNA 전사체가 예측된 수용체(적색 화살표) 및 공여체(녹색 화살표) 부위 및 엑손의 실제 위치(검정 박스)와 함께 도시되어 있다. 각 방법에 대해 예측된 부위 수를 실제 부위의 총 수와 같게 하는 임계값을 적용했다.
- [0503] 도 37C와 관련하여, 각 엑손에 대해, RNA-seq에서 엑손의 인클루전 레이트를 측정하고 상이한 인클루전 레이트에서 엑손에 대한 SpliceNet-10k 점수 분포를 보여준다. 엑손의 수용체 및 공여체 점수의 최대값이 표시된다.
- [0504] 도 37D와 관련하여, *U2SURF* 유전자에서 엑손 9 주위의 각 뉴클레오타이드를 인 실리코로 돌연변이시키는 것의 영향. 각 뉴클레오타이드의 수직 크기는 뉴클레오타이드가 돌연변이될 때 수용체 부위(검정 화살표)의 예측된 강도의 감소를 나타낸다(Δ Score).
- [0505] 도 37E와 관련하여, 입력 서열 컨텍스트의 크기가 망의 정확도에 미치는 영향. Top-k 정확도는 예측된 부위의 수가 실제 부위 수와 동일한 임계값에서 올바르게 예측된 스플라이스 부위의 비율이다. PR-AUC는 정밀-재호출 곡선화 면적이다. 또한 스플라이스 부위 검출을 위한 3가지 다른 알고리즘에 대한 top-k 정확도와 PR-AUC를 보여준다.
- [0506] 도 37F와 관련하여, SpliceNet-80nt(로컬 모티프 점수) 및 SpliceNet-10k에 의해 예측된 엑손/인트론 길이와 인접 스플라이스 부위의 강도 사이의 관계. 엑손 길이(노란색) 및 인트론 길이(분홍색)의 게놈 전체 분포가 배경에 표시된다. x 축은 로그 스케일이다.
- [0507] 도 37G와 관련하여, 150 nt 간격으로 배치된 한 쌍의 스플라이스 수용체 및 공여체 모티프가 HMGR 유전자를 따라 이동된다. SpliceNet-10k에 의해 예측된 바와 같이, 각각의 위치에서, K562 뉴클레오솜 신호 및 그 위치에서 상기 한 쌍이 엑손을 형성할 가능성이 도시되어 있다.
- [0508] 도 37H와 관련하여, GTEx 코호트에서 신규 엑손을 생성하기 위해 SpliceNet-10k 모델에 의해 예측되는 개인적 돌연변이 근처의 평균 K562 및 GM12878 뉴클레오솜 신호. 치환 테스트에 의한 p-값이 표시된다.
- [0509] 도 38A, 도 38B, 도 38C, 도 38D, 도 38E, 도 38F 및 도 38G는 RNA-seq 데이터에서 희귀 크립틱 스플라이스 돌연변이의 유효성확인의 일 구현예를 도시한다.
- [0510] 도 38A와 관련하여, 돌연변이의 스플라이스-변경 영향을 평가하기 위해, SpliceNet-10k는, 심근 병증과 관련된 MYBPC3 인트론에서 병원성 크립틱 스플라이스 변이체인 rs397515893에 대해 나타낸 바와 같이, 돌연변이가 있는 유전자와 돌연변이가 없는 유전자의 전구체-mRNA 서열의 각 위치에서 수용체 및 공여체 점수를 예측한다. 돌연변이에 대한 Δ Score 값은 변이체로부터 50 nt 내에서 스플라이스 예측 점수의 가장 큰 변화이다.
- [0511] 도 38B와 관련하여, SpliceNet-10k 모델을 사용하여 개인적 유전자 변이체 (GTEx 코호트에서 149명의 개체 중 하나에서만 관찰됨)를 기록하였다. 개인적 엑손 스킵핑 접합부(상단) 또는 개인적 수용체 및 공여체 부위(하단) 부근에서 스플라이싱을 변경하거나(Δ Score > 0.2, 청색) 또는 스플라이싱에 영향을 미치지 않을(Δ Score < 0.01, 적색) 것으로 예상되는 개인적 변이체의 농축이 표시된다. y 축은 개인적 스플라이스 이벤트와 근처의 개인적 유전자 변이체가 동일한 개체에서 동시에 발생하는 횟수를 치환을 통해 얻은 예상 수와 비교하여 나타낸다.

- [0512] 도 38C와 관련하여, 불완전한 침투성을 갖는 신규한 공여체 부위를 생성하는 *PYGB*의 이형접합 등의 변이체의 예이다. 변이체를 갖는 개체와 대조군 개체에 대해 RNA-seq 커버리지, 접합부 리드 계수치 및 접합부 위치(청색 및 회색 화살표)가 도시되어 있다. 효과 크기는 변이체가 있는 개체와 변이체가 없는 개체 간의 신규 접합부(AC) 사용의 차이로 계산된다. 아래의 누적 막대그래프에는 주석이 달린 또는 새로운 접합부(각각 "스플라이스 없음" 및 "신규 접합부")를 사용한 참조 또는 대체 대립유전자가 있는 리드 횟수가 표시된다. 총 참조 리드 수는 총 대체 리드 수($P = 0.018$, 이항 테스트)와 크게 달랐으며, 이는 새로운 접합부에서 스플라이싱하는 전사체의 60%가 RNA-seq 데이터에서 누락된 것으로 보이는데, 이는 아마도 넨센스 매개 붕괴(nonsense-mediated decay: NMD)로 인한 것이다.
- [0513] 도 38D와 관련하여, GTEx RNA-seq 데이터에 대해 유효성확인된 SpliceNet-10k 모델에 의해 예측된 크립틱 스플라이스 돌연변이의 분율. 필수 수용체 또는 공여체 다이뉴클레오타이드(파선)의 파괴의 유효성확인율은 커버리지 및 넨센스 매개 붕괴로 인해 100% 미만이다.
- [0514] 도 38E와 관련하여, 유효성확인된 크립틱 스플라이스 예측을 위한 효과 크기의 분포. 파선(50%)은 완전 침투성 이형접합 변이체의 예상 효과 크기에 해당한다. 필수 수용체 또는 공여체 다이뉴클레오타이드 파괴의 측정된 효과 크기는 넨센스 매개 붕괴 또는 설명되지 않은 동형 변화로 인해 50% 미만이다.
- [0515] 도 38F와 관련하여, 상이한 Δ Score 컷오프에서 GTEx 코호트에서 스플라이스-변경 개인적 변이체를 검출할 때 SpliceNet-10k의 감도. 변이체는 심층 인트론 변이체(엑손으로부터 > 50 nt)와 엑손 근처의 변이체(엑손과 겹치거나 또는 엑손-인트론 경계로부터 ≤ 50 nt)로 나뉜다.
- [0516] 도 38G와 관련하여, SpliceNet-10k의 유효성확인을 및 민감도와 상이한 신뢰도 컷오프에서 스플라이스 부위 예측을 위한 세 가지 다른 방법. SpliceNet-10k 곡선의 세 점은 0.2, 0.5 및 0.8의 Δ Score 컷오프에서 SpliceNet-10k의 성능을 나타낸다. 다른 세 가지 알고리즘의 경우, 곡선의 세 점은 0.2, 0.5 및 0.8의 Δ Score 컷오프에서 SpliceNet-10k와 동일한 수의 크립틱 스플라이스 변이체를 예측하는 임계값에서의 성능을 나타낸다.
- [0517] 도 39A, 도 39B 및 도 39C는 조직-특이적 대체 스플라이싱을 빈번하게 생성하는 크립틱 스플라이스 변이체의 일 구현예를 도시한다.
- [0518] 도 39A와 관련하여, 신규 공여체 부위를 생성하는 *CDC25B*에서의 이형접합 엑손 변이체의 예. 이 변이체는 GTEx 코호트에서 단일 개체에 대해 개인적이며, 섬유 아세포(Fisher Exact test에 의하면 $P = 0.006$)에 비해 근육에서 신규 스플라이스 동형의 더 큰 부분을 선호하는 조직-특이적 대체 스플라이싱을 나타낸다. RNA-seq 커버리지, 접합부 리드 계수치 및 접합부 위치(청색 및 회색 화살표)는 근육 및 섬유 아세포 모두에서 변이체를 가진 개체와 대조군 개체에 대해 표시된다.
- [0519] 도 39B와 관련하여, 변이체를 보유한 GTEx 코호트에서 3명의 개체 모두에 걸쳐 일관된 조직-특이적 효과를 나타내는 *FAM229B*에서 이형접합 엑손 수용체-생성 변이체의 예. 동맥 및 폐에 대한 RNA-seq가 변이체를 갖는 3명의 개체와 대조군 개체에 대해 제시되어 있다.
- [0520] 도 39C와 관련하여, 발현 조직에 걸쳐 신규한 접합부의 현저히 불균일한 사용과 관련되는, GTEx 코호트에서의 스플라이스 부위-생성 변이체의 분율은, 균질성에 대한 카이-제곱 테스트에 의해 평가된다. 낮거나 내지는 중간 Δ Score 값을 갖는 유효성확인된 크립틱 스플라이스 변이체는 조직-특이적 대체 스플라이싱($P = 0.015$, Fisher Exact test)을 일으킬 가능성이 더 컸다.
- [0521] 도 40A, 도 40B, 도 40C, 도 40D 및 도 40E는 인간 개체군에서 강하게 유해한 예측된 크립틱 스플라이스 변이체의 일 구현예를 도시한다.
- [0522] 도 40A와 관련하여, 확실하게 예측된 스플라이스 변경 효과 (Δ Score ≥ 0.8)를 갖는 동의 및 인트론 변이체(알려진 엑손-인트론 경계로부터 ≤ 50 nt, 필수 GT 또는 AG 다이뉴클레오타이드 제외)는 60,706명의 개체에서 한 번만 관찰되는 희귀 변이체에 비해 인간 개체군에서는 일반적인 대립유전자 빈도($\geq 0.1\%$)로 강하게 고갈된다. 4.58의 승산비(카이 제곱 검정에 의하면 $P < 10^{-127}$)는 최근에 발생하는 예측된 크립틱 스플라이스 변이체의 78%가 자연 선택에 의해 제거되기에 충분히 해롭다는 것을 나타낸다.
- [0523] 도 40B에 도시된 바와 같이, 단백질-절단 변이체와 ExAC 데이터 세트에서 유해한 예측된 동의 및 인트론 크립틱 스플라이스 변이체의 분율은 (A)에서와 같이 계산된다.
- [0524] 도 40C와 관련하여, 변이체가 프레임시프트를 야기할 것으로 예상되는지 여부에 기초하여 분할된, ExAC 데이터

세트에서 유해한($\Delta\text{Score} \geq 0.8$) 동의 및 인트론 크립틱 스플라이스 이득 변이체의 분율.

- [0525] 도 40D와 관련하여, 단백질 절단 변이체와 gnomAD 데이터 세트에서 유해한 예측된 심층 인트론(알려진 엑손-인트론 경계로부터 > 50 nt) 크립틱 스플라이스 변이체의 분율.
- [0526] 도 40E와 관련하여, 개별 인간 게놈 당 회귀(대립유전자 빈도 < 0.1%) 단백질 절단 변이체 및 회귀 기능성 크립틱 스플라이스 변이체의 평균 수. 기능적일 것으로 예상되는 크립틱 스플라이스 돌연변이의 수는 유해한 예측의 비율에 기초하여 추정된다. 총 예측 수는 더 높다.
- [0527] 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F는 회귀 유전질환을 갖는 환자에서 드 노보 크립틱 스플라이스 돌연변이의 일 구현예를 보여준다.
- [0528] 도 41A와 관련하여, Deciphering Developmental Disorders(DDD) 코호트 환자, Simons Simplex Collection 및 Autism Sequencing Consortium의 자폐 스펙트럼 장애(ASD) 환자 및 건강한 대조군에 대해 1인당 예측된 크립틱 스플라이스 드 노보 돌연변이. 건강한 대조군보다 높은 DDD 및 ASD 코호트에서의 농축이 보여지고, 코호트 사이의 변이체 확정을 조정한다. 오차 막대는 95% 신뢰 구간을 나타낸다.
- [0529] 도 41B와 관련하여, 건강한 대조군과 비교한 각 카테고리의 농축에 기초하여, DDD 및 ASD 코호트에 대한 기능적 카테고리에 의한 병원성 드 노보 돌연변이의 추정된 비율.
- [0530] 도 41C와 관련하여, 상이한 ΔScore 임계값에서 건강한 대조군과 비교한, DDD 및 ASD 코호트에서의 크립틱 스플라이스 드 노보 돌연변이의 농축 및 과잉.
- [0531] 도 41D에서, 예측된 크립틱 스플라이스 돌연변이가 단백질-코딩 돌연변이와 함께 농축 분석에 포함되었을 때, DDD 및 ASD 코호트에서 드 노보 돌연변이에 대해 농축된 신규 후보 질병 유전자의 목록(FDR < 0.01). 여러 개인에게 존재하는 표현형이 표시된다.
- [0532] 도 41E와 관련하여, RNA-seq에서 유효성확인되며 인트론 보유, 엑손 스킵핑 및 엑손 연장을 각각 초래하는 자폐증 환자에서의 예측된 드 노보 크립틱 스플라이스 돌연변이의 3가지 예. 각각의 예에서, 영향을 받는 개체에 대한 RNA-seq 커버리지 및 접합부 계수치는 상단에 표시되고, 돌연변이가 없는 대조군 개체는 하단에 표시된다. 유전자의 전사와 관련하여 센스 가닥에 서열이 제시되어 있다. 청색 화살표와 회색 화살표는 변이체를 갖는 개체와 대조군 개체 각각에서 접합부의 위치를 구분한다.
- [0533] 도 41F와 관련하여, RNA-seq에 의한 실험적 유효성확인을 위해 선택된 36개의 예측된 크립틱 스플라이스 부위에 대한 유효성확인 상태.
- [0534] **실험적 모델 대상체 세부사항**
- [0535] 36명의 자폐증 환자에 대한 대상체 세부 사항은 Iossifov et al., Nature 2014(표 S1)에 의해 이전에 공개되었으며, 본 논문의 표 S4의 1열의 익명화된 식별자를 사용하여 상호 참조될 수 있다.
- [0536] **방법 세부사항**
- [0537] **I. 스플라이스 예측을 위한 심층 학습**
- [0538] **SpliceNet 아키텍처**
- [0539] 본 발명자들은 컴퓨터 연산에 의해 전구체-mRNA 뉴클레오타이드 서열로부터 스플라이싱을 예측하기 위해 여러 초심층 컨볼루션 신경망 기반 모델을 트레이닝했다. 본 발명자들은 4개의 아키텍처, 즉, SpliceNet-80nt, SpliceNet-400nt, SpliceNet-2k 및 SpliceNet-10k의 아키텍처를 설계했고, 이들 아키텍처는 관심 위치의 각 측면에 각각 40, 200, 1,000 및 5,000개의 뉴클레오타이드를 입력으로 사용하며, 위치가 스플라이스 수용체와 기증자가 될 확률을 출력한다. 보다 정확하게는, 모델로의 입력은 원-핫 인코딩된 뉴클레오타이드의 서열이고, 이때 A, C, G 및 T (또는 동등하게 U)는 각각 [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0] 및 [0, 0, 0, 1]으로 인코딩되고, 모델의 결과는 합산되어 1이 되는 3개의 점수로 구성되는데, 이들은 관심 위치가 스플라이스 수용체이거나, 스플라이스 공여체이거나, 및 둘 다 아닐 가능성에 해당한다.
- [0540] SpliceNet 아키텍처의 기본 단위는 잔여 블록이며(He et al., 2016b), 이는 일괄 정규화 계층(Ioffe and Szegedy, 2015), 정류 선형 유닛(ReLU) 및 특정 방식으로 구성된 컨볼루션 유닛(도 21, 도 22, 도 23 및 도 24)으로 구성된다. 잔여 블록은 일반적으로 심층 신경망을 설계할 때 사용된다. 잔여 블록을 개발하기 전에는 폭발/소실 그라디언트 문제로 인해 여러 개의 컨볼루션 유닛이 차례대로 적층되어 구성된 심층 신경망은 트레이

닝시키기가 매우 어려웠고(Glorot and Bengio, 2010), 그러한 신경망의 깊이를 증가시키는 것은 종종 더 큰 트레이닝 오차를 초래했다(He et al., 2016a). 종합적인 계산 실험 세트를 통해, 여러 개의 잔여 블록으로 구성된 아키텍처가 차례로 이러한 문제를 극복하는 것으로 나타났다(He et al., 2016a).

[0541] 완전한 SpliceNet 아키텍처는 도 21, 도 22, 도 23 및 도 24에 제공된다. 아키텍처는 입력층을 두 번째 층에 연결하는 K개의 적층된 잔여 블록과, 두 번째 층을 출력층에 연결하는 소프트맥스 활성화를 갖는 컨볼루션 유닛을 포함하는 장치로 구성된다. 잔여 블록은 i 번째 잔여 블록의 출력이 i+1 번째 잔여 블록의 입력에 연결되도록 적층된다. 또한, 모든 제4 잔여 블록의 출력은 두 번째 층의 입력에 추가된다. 이러한 "스킵 연결"은 일반적으로 심층 신경망에서 트레이닝 중 수렴 속도를 높이기 위해 사용된다(Oord et al., 2016).

[0542] 각 잔여 블록에는 3개의 하이퍼 파라미터 N, W 및 D가 있고, 여기서 N은 컨볼루션 커널 수를 나타내고, W는 윈도우 크기를 나타내며 D는 각 컨볼루션 커널의 팽창률(Yu 및 Koltun, 2016)을 나타낸다. 윈도우 크기 W와 팽창률 D의 컨볼루션 커널은 (W-1)D 인접 위치에 걸쳐있는 피처를 추출하므로, 하이퍼 파라미터 N, W 및 D를 갖는 잔여 블록은 2(W-1)D 인접 위치에 걸쳐있는 피처를 추출한다. 따라서 SpliceNet 아키텍처의 전체 인접 스펠은

$$S = \sum_{i=1}^K 2(W_i - 1)D_i$$

으로 주어지며, 여기서 N_i , W_i 및 D_i 는 i 번째 잔여 블록의 하이퍼 파라미터이다. SpliceNet-80nt, SpliceNet-400nt, SpliceNet-2k 및 SpliceNet-10k 아키텍처의 경우 S가 각각 80, 400, 2,000 및 10,000이 되도록 각 잔여 블록의 잔여 블록 수와 하이퍼 파라미터가 선택되었다.

[0543] SpliceNet 아키텍처에는 컨볼루션 유닛 외에 정규화 및 비선형 활성화 유닛만 있다. 결과적으로, 모델은 가변 시퀀스 길이를 갖는 시퀀스-대-시퀀스 모드에서 사용될 수 있다(Oord et al., 2016). 예를 들어 SpliceNet-10k 모델 입력(S = 10,000)은 길이가 S/2 + 1 + S/2 인 원-핫 인코딩된 뉴클레오타이드 시퀀스이고 출력은 1×3 행렬이며, 이는 입력에서 1개의 중심 위치의 3개의 점수, 즉, 첫 번째 및 마지막 S/2 뉴클레오타이드를 배제한 후 남아있는 위치에 상응하는 것이다. 이 피처를 사용하면 트레이닝 및 테스트 중에 엄청난 양의 계산 비용을 절약할 수 있다. 이것은 서로 가까운 위치에 대한 대부분의 계산이 공통적이기 때문이며, 공유 계산은 시퀀스-대-시퀀스 모드에서 사용될 때 모델에 의해 한 번만 수행하면 된다.

[0544] 본 발명자들의 모델은 잔여 블록의 아키텍처를 채택하였으며, 이는 이미지 분류에서의 성공 덕분에 널리 채택되어 오고 있다. 잔여 블록은, 초기 층으로부터의 정보가 잔여 블록을 스킵할 수 있게 하는 스킵 연결들이 산재된 컨볼루션 반복 단위를 포함한다. 각각의 잔여 블록에서, 입력층은 먼저 일괄 정규화되고, 이어서 정류 선형 유닛(ReLU)을 사용하는 활성화 층이 뒤따른다. 이어서, 활성화는 1D 컨볼루션층을 통과한다. 1D 컨볼루션층으로부터의 이러한 중간 출력은, 다시 일괄 정규화되고 ReLU가 활성화되고, 이어서 다른 1D 컨볼루션층이 이어진다. 제2 1D 컨볼루션의 종료시, 그 출력을 초기 입력과 함께 잔여 블록으로 합산하며, 이는 초기 입력 정보가 잔여 블록을 우회할 수 있게 함으로써 스킵 연결로서 기능한다. 저자에 의해 심층 잔여 학습망이라고 불리는 이러한 아키텍처에서, 입력은 초기 상태로 유지되고, 잔여 연결은 모델로부터의 비선형 활성화 없이 유지되어, 더욱 심층인 망의 효과적인 트레이닝이 가능하다.

[0545] 잔여 블록에 이어서, 소프트맥스층은, 각각의 아미노산에 대한 3개 상태의 확률을 연산하고, 그 중에서 가장 큰 소프트맥스 확률이 아미노산의 상태를 결정한다. 이 모델은, ADAM 옵티 마이저를 사용하여 전체 단백질 서열에 대해 누적된 카테고리 교차 엔트로피 손실 함수로 트레이닝된다.

[0546] 아트러스/팽창된 컨볼루션은 트레이닝 가능한 파라미터가 거의 없는 큰 수용장을 허용한다. 아트러스/팽창된 컨볼루션은, 입력값을 아트러스 컨볼루션을 또는 팽창 인자라고도 하는 소정의 단차로 스킵함으로써 길이보다 넓은 면적에 걸쳐 커널이 적용되는 컨볼루션이다. 아트러스/팽창 컨볼루션은, 컨볼루션 필터/커널의 요소들 사이에 간격을 추가하여, 컨볼루션 연산이 수행될 때 더욱 큰 간격으로 이웃하는 입력 엔트리들(예를 들어, 뉴클레오타이드, 아미노산)이 고려되도록 한다. 이를 통해 입력에 장거리 컨텍스트 종속성을 통합할 수 있다. 아트러스 컨볼루션은 인접한 뉴클레오타이드들이 처리될 때 재사용을 위한 부분 컨볼루션 계산을 보존한다.

[0547] 예시된 예는 1D 컨볼루션을 사용한다. 다른 구현예에서, 모델은, 2D 컨볼루션, 3D 컨볼루션, 팽창 또는 아트러스 컨볼루션, 전치된 컨볼루션, 분리가 가능한 컨볼루션, 및 깊이별 분리가 가능한 컨볼루션 등의 상이한 유형의 컨볼루션을 사용할 수 있다. 일부 층은, 또한, 시그모이드 또는 쌍곡 탄젠트와 같은 포화 비선형성에 비해 확률적 그라디언트 하강의 수렴을 크게 가속하는 ReLU 활성화 함수를 사용한다. 개시된 기술에 의해 사용될 수 있는 활성화 함수의 다른 예는 파라메트릭 ReLU, 누설 ReLU, 및 지수 선형 유닛(ELU)을 포함한다.

[0548] 일부 층은, 또한, 일괄 정규화를 사용한다(Ioffe and Szegedy 2015). 일괄 정규화와 관련하여, 컨볼루션 신경망(CNN)의 각 층의 분포는 트레이닝 중에 변경되며 층마다 가변된다. 이는 최적화 알고리즘의 수렴 속도를 감소시킨다. 일괄 정규화는 이러한 문제를 극복하는 기술이다. x 를 사용한 일괄 정규화층의 입력과 z 를 사용한 출력을 이용하여, 일괄 정규화는 x 에 대한 이하의 변환을 적용한다.

$$z = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \beta$$

[0549] 일괄 정규화는, μ 와 σ 를 사용하여 입력 x 에 평균-분산 정규화를 적용하고, 이를 γ 와 β 를 사용하여 선형으로 스케일링하고 시프트한다. 정규화 파라미터 μ 와 σ 는, 지수 이동 평균이라고 하는 방법을 사용하여 트레이닝 세트에 걸쳐 현재 층에 대하여 연산된다. 즉, 이들은 트레이닝 가능한 파라미터가 아니다. 대조적으로, γ 와 β 는 트레이닝 가능한 파라미터이다. 트레이닝 중에 계산된 μ 와 σ 의 값은 추론 동안 순방향 패스에 사용된다.

[0551] **모델 트레이닝 및 테스트**

[0552] UCSC 테이블 브라우저에서 GENCODE (Harrow et al., 2012) V24lift37 유전자 주석 테이블을 다운로드하고 20,287 개의 단백질 코딩 유전자 주석을 추출하여 여러 동형을 사용할 수 있을 때 주요 전사체를 선택했다. 본 발명자들은 스플라이스 접합부가 없는 유전자를 제거하고 나머지를 트레이닝 및 테스트 세트 유전자로 다음과 같이 나누었다: 염색체 2, 4, 6, 8, 10-22, X 및 Y에 속하는 유전자를 모델 트레이닝에 사용했다(13,384개 유전자, 130,796개 공여체-수용체 쌍). 본 발명자들은 트레이닝 유전자의 10%를 무작위로 선택하여 트레이닝 중 조기 정지 지점을 결정하는 데 사용했으며 나머지는 모델 트레이닝에 사용되었다. 모델을 테스트하기 위해, 본 발명자들은 패럴로그가 없는 염색체 1, 3, 5, 7 및 9의 유전자를 사용했다(1,652개의 유전자, 14,289개의 공여체-수용체 쌍). 이를 위해 <http://grch37.ensembl.org/biomart/martview> 에서 인간 유전자 패럴로그 목록을 참조했다.

[0553] 다음 절차를 사용하여 1 = 5,000 크기의 청크로 시퀀스-대-시퀀스 모드에서 모델을 트레이닝하고 테스트했다. 각각의 유전자에 대해, 표준 전사 시작 및 종료 부위 사이의 mRNA 전사체 서열을 hg19/GRCh37 어셈블리로부터 추출하였다. 입력 mRNA 전사체 서열은 다음과 같이 원-핫 인코딩되었다: A, C, G, T/U는 각각 [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1]에 매핑됨. 원-핫 인코딩된 뉴클레오타이드 서열은 길이가 5,000의 배수가 될 때까지 제로-패딩된 후, 길이 S/2의 플랭킹 서열로 시작 및 끝에서 추가로 제로-패딩되고, 여기서 S는 SpliceNet-80nt, SpliceNet-400nt, SpliceNet-2k 및 SpliceNet-10k 모델의 경우 각각 80, 400, 2,000 및 10,000이다. 이어서 패딩된 뉴클레오타이드 서열은 i 번째 블록이 $5,000(i - 1) - S/2 + 1$ 에서 $5,000i + S/2$ 까지의 뉴클레오타이드 위치로 구성되는 방식으로 길이 $S/2 + 5,000 + S/2$ 의 블록으로 분할되었다. 유사하게, 스플라이스 출력 라벨 서열은 다음과 같이 원-핫 인코딩되었다: 스플라이스 부위 아님, 스플라이스 수용체(해당 엑손의 첫 번째 뉴클레오타이드) 및 스플라이스 공여체(해당 엑손의 마지막 뉴클레오타이드)는 각각 [1, 0, 0], [0, 1, 0] 및 [0, 0, 1]에 매핑되었다. 원-핫 인코딩된 스플라이스 출력 라벨 시퀀스는 길이가 5,000의 배수가 될 때까지 제로-패딩된 다음, i 번째 블록이 $5,000(i - 1) + 1$ 부터 $5,000i$ 까지의 위치로 구성되는 방식으로 길이가 5,000인 블록으로 분할된다. 원-핫 인코딩된 뉴클레오타이드 서열 및 상응하는 원-핫 인코딩 라벨 서열은 각각 모델의 입력 및 모델의 표적 출력으로 사용되었다.

[0554] 이 모델은 2개의 NVIDIA GeForce GTX 1080 Ti GPU에서 배치 크기가 12로 10 에포크 동안 트레이닝되었다. 트레이닝 동안 Adam 옵티마이저(Kingma and Ba, 2015)를 사용하여 목표 출력과 예측 출력 간의 범주형 교차 엔트로피 손실을 최소화했다. 옵티마이저의 학습 속도는 처음 6 에포크 동안 0.001로 설정되었으며 이후의 모든 에포크에서 2의 약수로 감소되었다. 각 아키텍처에 대해 트레이닝 절차를 5회 반복하고 5개의 트레이닝된 모델을 얻었다(도 53A 및 도 53B). 테스트하는 동안 5개의 트레이닝된 모델을 사용하여 각 입력을 평가하고 그 출력들의 평균을 예측 출력으로 사용했다. 본 발명자들은 이러한 모델을 도 37A 및 다른 관련 도면의 분석에 사용하였다.

[0555] 스플라이스 변경 변이체의 식별을 포함하는 도 38A-g, 도 39A-c, 도 40A-e 및 도 41A-f의 분석을 위해, 본 발명자들은 염색체 2, 4, 6, 8, 10-22, X, Y의 GTEx 코호트에서 일반적으로 관찰되는 새로운 스플라이스 접합부를 포함하도록 GENCODE 주석의 트레이닝 세트를 보강했다(67,012 스플라이스 공여체 및 62,911 스플라이스 수용체). 이로써 트레이닝 세트에서 스플라이스 접합부 주석 수가 ~ 50% 증가했다. 결합된 데이터 세트에서 망을 트레이닝시키면, 특히 심층적 인트론 스플라이스 변경 변이체를 예측함에 있어, GENCODE 주석만으로 트레이닝된 망과 비교하여 RNA-seq 데이터에서 스플라이스 변경 변이체 검출의 감도가 향상되었으며(도 52A 및 도 52B), 본

발명자들은 이 망을 변이체 평가와 관련된 분석에 사용했다(도 38A-g, 도 39A-c, 도 40A-e, 도 41A-f 및 관련 도면들). GTEx RNA-seq 데이터 세트에 트레이닝과 평가 사이의 중복이 포함되지 않도록 하기 위해, 트레이닝 데이터 세트에서 5명 이상의 개인에게 존재하는 집합부만 포함시켰고, 4개 이하만 존재하는 변이체에 대해서는 망 성능만 평가했다. 신규 스플라이스 집합부 식별에 대한 자세한 내용은 방법의 GTEx 분석 부문에서 "스플라이스 집합부 검출"에 설명되어 있다.

[0556] **Top-k 정확도**

[0557] 정확하게 분류된 위치의 백분율과 같은 정확도 메트릭은 대부분의 위치가 스플라이스 부위가 아니기 때문에 대개 비효율적이다. 대신 이러한 설정에 효과적인 두 가지 메트릭, 즉, top-k 정확도와 정밀 재호출 곡선하 면적을 사용하여 모델을 평가했다. 특정 클래스에 대한 top-k 정확도는 다음과 같이 정의된다: 테스트 세트에 클래스에 속하는 k 개의 위치가 있다고 가정한다. 정확히 k 개의 테스트 세트 위치가 클래스에 속하는 것으로 예측되도록 임계값을 선택한다. 진정으로 클래스에 속하는 이러한 k 개의 예측된 위치의 비율은 top-k 정확도로 보고된다. 실제로, 이는 정밀도와 재호출이 동일한 값을 갖도록 임계값을 선택할 때의 정밀도와 같다.

[0558] **lincRNA에 대한 모델 평가**

[0559] 본 발명자들은 GENCODE V24lift37 주석에 기초한 모든 lincRNA 전사체의 목록을 얻었다. 단백질 코딩 유전자와는 달리, lincRNA는 GENCODE 주석에서 주요한 전사체가 할당되지 않았다. 유효성확인 세트에서 잉여를 최소화하기 위해, 본 발명자들은 lincRNA 유전자당 가장 긴 총 엑손 서열을 갖는 전사체를 확인하고 이를 유전자에 대한 표준 전사체로 지칭하였다. lincRNA 주석은 단백질 코딩 유전자에 대한 주석보다 신뢰성이 낮을 것으로 예상되며, 이러한 잘못된 주석은 top-k 정확도 추정치에 영향을 미치므로 GTEx 데이터를 사용하여 잠재적 주석 문제가 있는 lincRNA를 제거했다(이 데이터에 대한 자세한 내용은 아래의 "GTEx 데이터 세트 분석" 부문 참조). 각 lincRNA에 대해, 본 발명자들은 모든 GTEx 샘플에서 lincRNA의 길이에 걸쳐 맵핑된 모든 분할된 리드를 계수치했다(자세한 내용은 아래의 "스플라이스 집합부 검출" 참조). 이것은 주석이 달린 또는 신규의 집합부를 사용하는 lincRNA의 총 집합부-스패닝 리드의 추정치이다. 또한 표준 전사체의 집합부에 걸친 리드 수를 계수치했다. 모든 GTEx 샘플에 걸친 집합부-스패닝 리드의 95% 이상이 표준 전사체에 해당하는 lincRNA만을 고려하였다. 또한 표준 전사체의 모든 집합부는 GTEx 코호트에서 한 번 이상 관찰되어야 했다(길이 < 10 nt의 인트론에 걸친 집합부 제외). top-k 정확도를 계산하기 위해, 본 발명자들은 상기 필터를 통과한 lincRNA의 표준 전사체의 집합부만을 고려했다(781개의 전사체, 1047개의 집합부).

[0560] **전구체-mRNA 서열로부터 스플라이스 집합부 식별하기**

[0561] 도 37B에서, 본 발명자들은 유전자 서열로부터 유전자의 표준 엑손 경계를 식별하는 것과 관련하여 MaxEntScan 및 SpliceNet-10k의 성능을 비교한다. 본 발명자들의 테스트 세트에 있으며 26개의 표준 스플라이스 수용체와 기증자가 있는 CFTR 유전자를 사례 연구로 사용했고, MaxEntScan 및 SpliceNet-10k를 사용하여 표준 전사 개시 부위(chr7:117,120,017)에서 표준 전사 종결 부위(chr7:117,308,719)까지 188,703개의 위치 각각에 대한 수용체 및 공여체 점수를 획득하였다. 해당 점수가 top-k 정확도를 평가하는 동안 선택한 임계값보다 큰 경우, 위치는 스플라이스 수용체 또는 공여체로 분류된다. MaxEntScan은 49개의 스플라이스 수용체와 22개의 스플라이스 공여체를 예측했으며, 그중 9개와 5개는 각각 실제 스플라이스 수용체와 공여체이다. 더 나은 시각화를 위해 MaxEntScan의 사전 로그 점수를 표시한다(최대 2,500으로 제한). SpliceNet-10k는 26개의 스플라이스 수용체와 26개의 스플라이스 공여체를 모두 정확하게 예측했다. 도 42B에서, LINC00467 유전자를 사용하여 분석을 반복하였다.

[0562] **GENCODE 주석이 달린 스플라이스 집합부에서의 엑손 인클루전의 추정**

[0563] 본 발명자들은 GTEx RNA-seq 데이터로부터의 모든 GENCODE 주석이 달린 엑손의 인클루전을 계산하였다(도 37C). 각 유전자의 첫 번째와 마지막 엑손을 제외한 각 엑손에 대해 인클루전을 다음과 같이 계산했다.

$$\frac{(L + R)/2}{S + (L + R)/2}$$

[0564] 여기서 L은 모든 GTEx 샘플 걸쳐 이전 표준 엑손에서 검토 중인 엑손까지의 집합부의 총 리드 계수치이고, R은 고려 중인 엑손에서 다음 표준 엑손까지의 집합부의 총 리드 계수치이며, S는 이전 표준 엑손에서 다음 표준 엑손까지의 스킵핑 집합부의 총 리드 계수치이다.

[0566] **스플라이스 부위 인식을 위한 다양한 뉴클레오타이드의 중요성**

[0567] 도 37D에서, 스플라이스 수용체로서 위치의 분류를 향해 SpliceNet-10k에 의해 중요하게 고려되는 뉴클레오타이드를 식별한다. 이를 위해 테스트 세트에있는 U2SURP 유전자의 chr3:142,740,192에서 스플라이스 수용체를 고려했다. 스플라이스 수용체에 대한 뉴클레오타이드의 "중요도 점수"는 다음과 같이 정의된다: s_{ref} 는 고려중인 스플라이스 수용체의 수용체 점수를 나타낸다. 고려되는 뉴클레오타이드를 A, C, G 및 T로 대체함으로써 수용체 점수가 재계산된다. 이 점수를 각각 s_A , s_C , s_G 및 s_T 로 표시한다. 뉴클레오타이드의 중요도 점수는 다음과 같이 추정된다.

[0568]
$$s_{ref} = \frac{s_A + s_C + s_G + s_T}{4}$$

[0569] 이 절차는 종종 인 실리콘 돌연변이 유발로 지칭된다(Zhou and Troyanskaya, 2015). chr3:142,740,137에서 chr3:142,740,263까지 127개의 뉴클레오타이드를 플로팅하되, 각 뉴클레오타이드의 높이가 chr3:142,740,192에서 스플라이스 수용체에 대한 중요도 점수가 되도록 한다. 플로팅 기능은 DeepLIFT(Shrikumar et al., 2017) 소프트웨어에서 변용되었다.

[0570] **TACTAAC 및 GAAGAA 모티프가 스플라이싱에 미치는 영향**

[0571] 수용체 강도에 대한 분기점 시퀀스 위치의 영향을 연구하기 위해, 먼저 SpliceNet-10k를 사용하여 14,289 개의 테스트 세트 스플라이스 수용체의 수용체 점수를 얻었다. y_{ref} 가 이 점수를 포함하는 벡터를 나타낸다. 0에서 100 사이의 각 i 값에 대해 다음을 수행했다: 각각의 테스트 세트 스플라이스 수용체에 대해, 본 발명자들은 스플라이스 수용체 앞의 i 에서 $i-6$ 위치로부터의 뉴클레오타이드를 TACTAAC로 교체하고 SpliceNet-10k를 사용하여 수용체 점수를 재계산하였다. 이 점수를 포함하는 벡터는 $y_{alt,i}$ 로 표시된다. 다음의 수량을 i 의 함수로서 도 43A에서 플로팅한다.

[0572]
$$mean(y_{alt,i} - y_{ref})$$

[0573] 도 43B에 대하여, SR-단백질 모티프 GAAGAA를 사용하여 동일한 절차를 반복하였다. 이 경우, 스플라이스 수용체 이후에 존재할 때 모티프의 영향과 공여체 강도에 대한 영향을 연구했다. GAAGAA와 TACTAAC는 k -mer 공간의 포괄적인 탐색을 기반으로 수용체와 공여체 강도에 가장 큰 영향을 미치는 모티프였다.

[0574] **스플라이싱에서 엑손과 인트론 길이의 역할**

[0575] 스플라이싱에 대한 엑손 길이의 효과를 연구하기 위해, 첫 번째 또는 마지막 엑손 인 테스트 세트 엑손을 필터링했다. 이 필터링 단계는 14,289 개의 엑손에서 1,652개를 제거했다. 나머지 12,637 개의 엑손을 길이가 증가하는 순서대로 정렬했다. 각각에 대해 SpliceNet-80nt를 사용하여 스플라이스 수용체 부위의 수용체 점수와 스플라이스 공여체 부위의 공여체 점수를 평균하여 스플라이싱 점수를 계산했다. 본 발명자들은 스플라이싱 점수를 도 37F에 엑손 길이의 함수로서 플로팅한다. 플로팅하기 전에 다음과 같은 스무딩 절차를 적용했다: x 는 엑손의 길이를 포함하는 벡터를 나타내고, y 는 대응하는 스플라이싱 점수를 포함하는 벡터를 나타내도록 한다. 크기가 2,500 인 평균화 윈도우를 사용하여 x 와 y 를 모두 스무딩하였다.

[0576] SpliceNet-10k를 사용하여 스플라이싱 점수를 계산함으로써 이 분석을 반복 하였다. 백그라운드에서 이 분석에 고려된 12,637개의 엑손의 길이의 히스토그램을 보여준다. 인트론 길이가 스플라이싱에 미치는 영향을 연구하기 위해 유사한 분석을 적용했는데, 주요 차이점은 첫 번째 엑손과 마지막 엑손을 배제할 필요가 없다는 것이다.

[0577] **스플라이싱에서 뉴클레오솜의 역할**

[0578] UCSC 게놈 브라우저로부터 K562 세포주에 대한 뉴클레오솜 데이터를 다운로드 하였다. 본 발명자들은 SpliceNet-10k 점수에 대한 뉴클레오솜 포지셔닝의 영향을 입증하기 위해 일화적인 예로서 HMGR 유전자를 테스트 세트에 사용했다. 유전자에서 각 위치 p 에 대해, 본 발명자들은 다음과 같이 "심어진 스플라이싱 점수"를 계산했다.

[0579] • 위치 $p + 74$ 에서 $p + 81$ 까지의 8개 뉴클레오타이드를 공여체 모티프 AGGTAAGG로 교체했다.

[0580] • 위치 $p-78$ 에서 $p-75$ 까지의 4개의 뉴클레오타이드가 수용체 모티프 TAGG로 대체되었다.

- [0581]  위치 p-98에서 p-79까지의 20개 뉴클레오타이드를 폴리 피리 미딘 트랙 CCTCCTTTTTCCTGCCTC로 대체했다.
- [0582]  위치 p-105에서 p-99까지의 7개 뉴클레오타이드가 분 지점 서열 CACTAAC로 대체되었다.
- [0583]  SpliceNet-10k에 의해 예측된 p-75의 수용자 점수 및 p + 75의 기증자 점수의 평균은 심어진 스플라이싱 점수로 사용된다.
- [0584] chr5:74,652,154에서 chr5:74,657,153까지 5,000개 위치에 대한 K562 뉴클레오솜 신호 및 심어진 스플라이싱 점수는 도 37G에 도시되어있다.
- [0585] 이들 두 트랙 사이의 게놈 전체 Spearman 상관관계를 계산하기 위해, 모든 표준 유전자로부터 적어도 100,000 nt 떨어진 백만개의 유전자 간 위치를 무작위로 선택하였다. 이 위치들 각각에 대해, 본 발명자들은 평균 K562 뉴클레오솜 신호뿐만 아니라 그 심어진 스플라이싱 점수를 계산하였다(윈도우 크기 50이 평균화에 사용됨). 백만개의 위치에 걸친 이들 두 값 사이의 상관관계가 도 37G에 도시되어 있다. 빈의 크기가 0.02 인 GC 함량(심어진 수용체와 공여체 모티프 사이의 뉴클레오타이드를 사용하여 추정)을 기준으로 이러한 위치를 하위분류했다. 도 44A의 각 빈에 대한 게놈 전체 Spearman 상관관계를 보여준다.
- [0586] 14,289개의 테스트 세트 스플라이스 수용체 각각에 대해, 각각의 측면에서 50개 뉴클레오타이드 내의 뉴클레오솜 데이터를 추출하고 엑손 측면의 평균 신호를 인트론 측면의 평균 신호로 나눈 뉴클레오솜 농축을 계산하였다. 본 발명자들은 그들의 뉴클레오솜 농축의 순서대로 스플라이스 수용체를 분류하고 SpliceNet-80nt를 사용하여 그들의 수용체 점수를 계산했다. 수용체 점수는 도 44B에서 뉴클레오솜 농축의 함수로서 플로팅된다. 플로팅하기 전에, 도 37F에 사용된 스무딩 절차가 적용되었다. SpliceNet-10k를 이용하여 이 분석을 반복했고 또한 14,289개의 테스트 세트 스플라이스 공여체에 대해서도 이 분석을 반복했다.
- [0587] **신규 엑손에서 뉴클레오솜 신호의 농축**
- [0588] 도 37H의 경우, 본 발명자들은 예측된 신규 엑손 주위의 뉴클레오솜 신호를 보고 싶었다. 우리가 매우 자신있는 신규 엑손을 보고 있음을 보장하기 위해, 예측된 획득된 접합부가 변이체를 가진 개인에게 전적으로 개인적인 싱글톤 변이체(단일 GTEX 개인에 존재하는 변이체)만 선택했다. 또한 주변 엑손에서 혼란스러운 효과를 제거하기 위해, 주석이 달린 엑손에서 최소 750 nt 떨어진 인트론 변이체만을 조사했다. UCSC 브라우저에서 GM12878 및 K562 세포주에 대한 뉴클레오솜 신호를 다운로드하고 예측된 각각의 신규 수용체 또는 공여체 부위로부터 750 nt 내의 뉴클레오솜 신호를 추출하였다. 본 발명자들은 두 세포주 사이의 뉴클레오솜 신호를 평균화하고 음성 가닥에서 유전자가 겹치는 변이체에 대한 신호 백터를 뒤집었다. 본 발명자들은 수용체 부위로부터의 신호를 오른쪽으로 70 nt만큼 이동시켰고, 공여체 부위로부터의 신호를 왼쪽으로 70 nt만큼 이동시켰다. 이동 후, 수용체 및 공여체 부위 둘 다에 대한 뉴클레오솜 신호는 길이 140 nt의 이상적인 엑손의 중간에 중심을 두었으며, 이는 GENCODE v19 주석에서 엑손의 중간 길이이다. 마지막으로 모든 시프트된 신호의 평균을 구하고 각 위치를 중심으로 11 nt 윈도우 내의 평균을 계산하여 결과 신호를 스무딩했다.
- [0589] 연관성을 테스트하기 위해, 주석이 달린 엑손으로부터 750 nt 이상 떨어져 있고 모델에 의해 스플라이싱에 영향을 미치지 않는 것으로 예측된(Δ Score < 0.01) 무작위 싱글톤 SNV를 선택하였다. 이러한 SNV의 무작위 샘플 1000개를 생성하였으며, 각각의 샘플은 도 37H에 사용된 스플라이스-사이트 이득 부위의 세트만큼 많은 SNV를 갖는다(128개 부위). 각 무작위 샘플에 대해 위에서 설명한대로 스무딩된 평균 신호를 계산했다. 무작위 SNV가 새로운 엑손을 생성할 것으로 예측되지 않았기 때문에, 본 발명자들은 각각의 SNV로부터의 뉴클레오솜 신호를 SNV 자체의 중심에 놓고 무작위로 왼쪽으로 70 nt 또는 오른쪽으로 70 nt 이동시켰다. 이어서, 도 37H의 중간 염기에서의 뉴클레오솜 신호를 해당 염기의 1000회 시뮬레이션에서 얻은 신호와 비교하였다. 경험적 p-값은 스플라이스-사이트 이득 변이체에 대해 관찰된 것보다 크거나 같은 중간 값을 갖는 시뮬레이션된 세트의 비율로 계산되었다.
- [0590] **엑손 밀도의 차이에 대한 망의 견고성**
- [0591] 망의 예측의 일반화가능성을 조사하기 위해, 다양한 엑손 밀도 영역에서 SpliceNet-10k를 평가하였다. 먼저 10,000개 뉴클레오타이드 윈도우(각면에 5,000개 뉴클레오타이드)에 존재하는 표준 엑손의 수에 따라 테스트 세트 위치를 5가지 범주로 분리했다(도 54). 엑손 계수치가 각 위치에 대한 전체 값이 되도록 하기 위해, 본 발명자들은 윈도우에 존재하는 엑손 시작의 수를 대리인으로 사용했다. 각 범주에 대해 top-k 정확도와 정밀 재호출 곡선 하 면적을 계산했다. 위치의 개수와 k의 값은 범주에 따라 다르다(아래 표 참조).

엑손 계수치	위치 수	스플라이스 수용체 수	스플라이스 공여체 수
1 엑손	15,870,045	1,712	1,878
2 엑손	10,030,710	2,294	2,209
3 엑손	6,927,885	2,351	2,273
4 엑손	4,621,341	2,095	2,042
≥ 5 엑손	7,247,582	5,679	5,582

[0592]

[0593]

양상블의 5가지 모델 각각에 대한 망의 견고성

[0594]

다수의 모델을 트레이닝시키고 그 예측의 평균을 출력으로서 사용하는 것은 양상블 학습으로 불리는 더 나은 예측 성능을 얻기 위한 기계학습의 일반적인 전략이다. 도 53A에서, 본 발명자들은 양상블을 구성하기 위해, 트레이닝한 5가지 SpliceNet-10k 모델의 top-k 정확도와 정밀 재호출 곡선하 면적을 보여준다. 결과는 트레이닝 과정의 안정성을 분명히 보여준다.

[0595]

또한, 예측들 간의 Pearson 상관관계를 계산하였다. 계몽에서 대부분의 위치는 스플라이스 부위가 아니기 때문에 대부분 모델의 예측들 간의 상관관계는 1에 가까워 분석이 무의미하다. 이 문제를 극복하기 위해, 본 발명자들은 테스트 세트에서 적어도 하나의 모델에 의해 수용자 또는 공여체 점수가 0.01 이상으로 지정된 위치만을 고려했다. 이 기준은 53,272개의 위치(대략 동일한 수의 스플라이스 부위 및 비-스플라이스 부위)에 의해 충족되었다. 결과는 도 53B에 요약되어 있다. 모델들의 예측들 간의 매우 높은 Pearson 상관관계는 그들의 견고성을 더 잘 보여준다.

[0596]

본 발명자들은 도 53C에서 성능에 대한 양상블을 구성하기 위해 사용된 모델의 수의 효과를 보여준다. 결과는 모델 수가 증가함에 따라 성능이 향상됨을 보여주지만, 수익은 감소한다.

[0597]

II. GTEx RNA-seq 데이터 세트에 대한 분석

[0598]

단일 뉴클레오타이드 변이체의 ΔScore

[0599]

단일 뉴클레오타이드 변이체로 인한 스플라이싱 변화를 다음과 같이 정량화했다. 먼저 참조 뉴클레오타이드를 사용하여 변이체 주변의 101개 위치(양쪽에서 50개 위치)에 대한 수용체 및 공여체 점수를 계산했다. 이 점수가 각각 벡터 a_{ref} 및 d_{ref} 로 표시되어 있다고 가정한다. 그런 다음 대체 뉴클레오타이드를 사용하여 수용체와 공여체 점수를 다시 계산했다. 이 점수를 각각 벡터 a_{alt} 및 d_{alt} 로 표시한다.

[0600]

다음과 같은 4가지 수량을 평가했다:

$$\Delta \text{Score (수용체 이득)} = \max(a_{alt} - a_{ref})$$

$$\Delta \text{Score (수용체 손실)} = \max(a_{ref} - a_{alt})$$

$$\Delta \text{Score (공여체 이득)} = \max(d_{alt} - d_{ref})$$

$$\Delta \text{Score (공여체 손실)} = \max(d_{ref} - d_{alt})$$

[0601]

이들 4개의 점수 중 최대 점수는 변이체의 ΔScore로 지칭된다.

[0602]

변이체의 품질 관리 및 필터링의 기준

[0603]

본 발명자들은 dbGaP에서 GTEx VCF 및 RNA-seq 데이터를 다운로드했다(study accession phs000424.v6.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1).

[0604]

본 발명자들은 GTEx 코호트에서 최대 4명의 개체에 나타난 상염색체 SNV에 대한 SpliceNet의 성능을 평가하였다. 특히 변이체가 하나 이상의 개체 A에서 다음 기준을 만족하는 경우 그 변이체는 고려되었다.

[0605]

1. 변이체가 필터링되지 않았다(VCF의 FILTER 필드가 PASS였음).

[0606]

2. 개체 A의 VCF의 INFO 필드에 변이체가 MULTI_ALLELIC으로 표시되지 않았고 VCF는 ALT 필드에 단일 대립 유전

자를 포함했다.

- [0608] 3. 개체 A는 변이체에 대해 이형접합성이었다.
- [0609] 4. $\text{alt_depth}/(\text{alt_depth} + \text{ref_depth})$ 의 비율이 0.25와 0.75 사이였으며, 여기서 alt_depth 와 ref_depth 는 각각 개체 A에서 대체 및 참조 대립 유전자를 지지하는 리드의 횟수이다.
- [0610] 5. 전체 깊이 $\text{alt_depth} + \text{ref_depth}$ 가 개체 A의 VCF에서 20에서 300 사이였다.
- [0611] 6. 변이체는 유전자 본체 영역과 겹쳤다. 유전자 본체는 GENCODE (V24lift37)로부터의 표준 전사체의 전사 시작과 끝 사이의 영역으로 정의되었다.

[0612] 적어도 하나의 개인에서 이들 기준을 만족시키는 변이체에 대하여, 변이체가 (비록 상기 기준을 만족시키지 않았다 하더라도) 변이체를 갖는 것으로 나타난 모든 개체를 고려하였다. 본 발명자들은 단일 개인에 나타나는 변이체를 싱글톤이라고 하고 2-4명의 개인에 나타나는 변이체를 공통이라고 지칭한다. 트레이닝 데이터 세트와의 중복을 방지하기 위해 5명 이상의 개인에게 나타나는 변이체는 평가하지 않았다.

[0613] **RNA-seq 리드 정렬**

[0614] 본 발명자들은 OLego(Wu et al., 2013)를 사용하여 hg19 참조에 대해 GTEx 샘플의 리드를 맵핑하여 질의 리드와 기준 사이에서 최대 4의 편집 거리를 허용하였다(파라미터-M 4). OLego는 완전히 드 노보로 작동할 수 있으며 유전자 주석이 필요하지 않은 점에 유의한다. OLego는 분할 리드의 끝에 스플라이싱 모티프가 있는지 확인하기 때문에 스플라이스 부위를 각각 방해하거나 생성하는 SNV 주변의 기준에 대해 또는 그에 대한 정렬을 편향시킬 수 있다. 이러한 편향성을 제거하기 위해 PASS 필터를 사용하여 개인의 모든 SNV를 hg19 참조에 삽입하여 각 GTEx 개인에 대한 대체 참조 서열을 추가로 생성했다. 본 발명자들은 OLego를 동일한 파라미터와 함께 사용하여 각 개인의 모든 샘플을 해당 개인의 대체 참조 시퀀스에 매핑했다. 그런 다음 각 샘플에 대해 각 리드 쌍에 가장 적합한 정렬을 선택하여 두 개의 정렬 세트(hg19 참조 및 개인의 대체 참조와 비교)를 결합했다. 리드 쌍 P에 가장 적합한 정렬을 선택하기 위해 다음 절차를 사용했다.

- [0615] 1. 두 정렬 모두에서 P의 두 리드가 모두 매핑 해제된 경우 hg19 또는 P의 대체 정렬을 임의로 선택했다.
- [0616] 2. P가 한 정렬 세트에서 다른 정렬보다 매핑되지 않은 끝이 더 많으면(예컨대, P의 양쪽 끝이 대체 참조에 대해 매핑되었지만 한 끝만 hg19에 대해 매핑된 경우) P의 두 끝이 있는 정렬을 선택했다.
- [0617] 3. P의 양쪽 끝이 두 정렬 집합 모두에 매핑된 경우, 전체 불일치가 가장 적은 정렬을 선택하거나 불일치 수가 동일한 경우 임의의 정렬을 선택했다.

[0618] **정렬된 RNA-seq 데이터에서 스플라이스 접합부의 검출**

[0619] 본 발명자들은 각 샘플에서 스플라이스 접합부를 검출하고 계수치하기 위해 리프커터 패키지(Li et al., 2018)의 유틸리티인 leafcutter_cluster를 사용하였다. 접합부를 지원하기 위해 단일 분할 리드가 필요했으며 최대 인트론 길이는 500Kb (파라미터 -m 1 -l 500000)로 가정했다. 심층 학습 모델의 트레이닝을 위한 높은 신뢰도의 접합부 세트를 얻기 위해, 모든 샘플에서 모든 리프커터 접합부의 합집합을 컴파일한 후 다음 기준 중 하나를 충족한 접합부는 고려대상에서 제거했다.

- [0620] 1. 접합부의 어느 한쪽 끝이 ENCODE 블랙리스트 영역(UCSC 게놈 브라우저에서 hg19에 포함된 테이블 wgEncodeDacMapabilityConsensusExcludable) 또는 단순 반복(UCSC 게놈 브라우저에서 hg19의 Simple Repeats 트랙)과 겹쳤다.
- [0621] 2. 접합부의 양끝이 비표준 엑손(GENCODE 버전 V24lift37의 표준 전사체에 근거)에 있었다.
- [0622] 3. 접합부의 두 끝이 서로 다른 유전자에 있거나 한쪽 끝이 비유전자 영역에 있었다.
- [0623] 4. 양쪽 끝 모두 필수 GT/AG 다이뉴클레오타이드가 없었다.
- [0624] 5개 이상의 개체에 존재하는 접합부는 변이체 예측에 대한 분석을 위해 GENCODE 주석이 붙은 스플라이스 접합부의 목록을 확대하는 데 사용되었다(도 38A-g, 도 39A-c, 도 40A-e 및 도 41A-f). 모델 학습에 사용된 스플라이스 접합부 목록이 포함된 파일에 대한 링크는 핵심 자원 표에 제공된다.

[0625] 비록 트레이닝 데이터 세트를 증대시키기 위해 리프커터에 의해 검출된 접합부를 사용했지만, 이완된 파라미터의 사용에도 불구하고, 리프커터는 RNA-seq 데이터에서 우수한 지지를 갖는 많은 접합부를 필터링하고 있음을

알았다. 이것은 인공적으로 본 발명자들의 검증 속도를 낮추었다. 따라서, GTEx RNA-seq 유효성확인 분석(도 38A-g 및 도 39A-c)을 위해, RNA-seq 리드 데이터로부터 직접 접합부 및 접합부 계수치 세트를 재계산하였다. MAPQ가 10 이상이고 접합부의 양쪽에 정렬된 nt가 5개 이상인 모든 비중복 분할매핑된 리드를 계산했다. 리드는 2개 이상의 엑손에 걸쳐있을 수 있으며, 이 경우 리드는 양쪽에서 적어도 5 nt의 맵핑된 서열을 갖는 각각의 접합부로 계수치되었다.

[0626] **개인적 접합부의 정의**

[0627] 접합부는 다음 기준 중 하나 이상을 만족시키는 경우, 개인 A에서 개인적인 접합부로 간주되었다:

[0628] 1. 접합부는 A로부터 적어도 하나의 샘플에서 적어도 3개의 리드를 가졌으며, 다른 개체에서는 관찰되지 않았다.

[0629] 2. 다음 두 기준을 모두 만족하는 조직이 두 개 이상 있었다.

[0630] a. 조직의 개체 A로부터의 샘플에서 접합부의 평균 리드 계수치가 10 이상이었다.

[0631] b. 개체 A는 그 조직의 다른 개체보다 평균적으로 두 배 이상의 정규화된 리드를 가졌다. 여기서, 샘플에서 접합부의 정규화된 리드 계수치는 상응하는 유전자에 대한 모든 접합부에 걸친 총 리드의 수에 의해 정규화된 접합부의 리드 수로 정의되었다.

[0632] 다른 개체(A 아님)로부터 5개 미만의 샘플을 갖는 조직은 테스트에서 무시되었다.

[0633] **개인적 접합부 주변의 싱글톤 SNV 농축**

[0634] GENCODE 주석에 기초하여 개인적 접합부가 정확히 한쪽 끝에 주석을 달았다면, 본 발명자들은 이를 수용체 또는 공여체 획득의 후보로 간주하고 주석이 없는 끝에서 150 nt 내의 동일 개체에 개인적이었던 싱글톤 SNV(단일 GTEx 개체에 나타나는 SNV)를 탐색했다. 개인적 접합부에 양쪽 끝이 주석 처리된 경우, GENCODE 주석을 기반으로 동일한 유전자의 하나 이상 3개 이하의 엑손을 건너뛴 경우에 개인 엑손 스킵핑 이벤트의 후보로 간주했다. 그런 다음 건너뛴 각 엑손의 끝에서 150 nt 내의 싱글톤 SNV를 탐색했다. GENCODE 엑손 주석에서 부재하는 양쪽 끝이 있는 개인적인 접합부는 무시되었는데, 이들 중 상당 부분은 정렬 오류였기 때문이다.

[0635] 신규한 개인적 수용체 또는 공여체 주위의 싱글톤 SNV의 농축을 계산하기 위해(도 38B, 하단), 본 발명자들은 개인적 접합부에 대한 각 위치에서 싱글톤 SNV의 수를 집계하였다. 겹치는 유전자가 음성 가닥에 있다면, 상대 위치가 뒤집어졌다. 본 발명자들은 SNV를 두 개의 그룹으로 나누었다: 개인적 접합부를 가진 개체의 개인적 SNV와 다른 개체의 개인적 SNV. 결과 신호를 스무딩하기 위해 각 위치를 중심으로 7 nt 윈도우에서 계수치를 평균화했다. 그런 다음 첫 번째 그룹(같은 개체에서 개인적인)의 스무딩한 계수치에 대한 두 번째 그룹(다른 개체에서 개인적인)의 스무딩한 계수치의 비율을 계산했다. 신규한 개인적 엑손 건너뛰기(도 38B, 상단)의 경우, 유사한 절차에 따라 건너뛴 엑손의 끝 주변에서 싱글톤 SNV의 계수치를 집계했다.

[0636] **GTEx RNA-seq 데이터에서 모델 예측의 유효성확인**

[0637] 개인적 변이체(GTEx 코호트에 한 개인에게 나타남) 또는 공통 변이체(GTEx 코호트에 2-4명의 개인에 나타남)에 대해, 참조 및 대체 대립 유전자 둘 다에 대한 심층 학습 모델의 예측을 얻었고 Δ Score를 계산했다. 본 발명자들은 또한 모델이 비정상(신규 또는 파괴된) 접합부를 예측한 위치를 얻었다. 그런 다음 예측된 위치에 변이체가 있는 개체의 스플라이싱 이상을 지원하는 증거가 RNA-seq 데이터에 있는지 확인하려 시도했다. 많은 경우, 모델은 같은 변이체에 대해 여러 효과를 예측할 수 있다(예컨대, 주석이 달린 스플라이스 공여체를 방해하는 변이체는 또한 도 45에서와 같이 준최적 공여체의 사용을 증가시킬 수 있다. 이 경우, 모델은 주석이 달린 스플라이스 부위에서 공여체 손실과 준최적 부위에서 공여체 이득을 모두 예측할 수 있다. 그러나 유효성확인 목적으로 각 변이체에 대해 가장 높은 예측 Δ Score를 가진 효과만 고려했다. 따라서, 각 변이체에 대해, 예측된 스플라이스 부위 생성 및 스플라이스 부위 중단 효과를 개별적으로 고려했다. 모델이 트레이닝된 신규 접합부에서 모델을 평가하는 것을 피하기 위해, 5명 미만의 개인에게 나타나는 접합부는 모델 트레이닝 중에 제외되었다.

[0638] **개인적 스플라이스 접합부에 기초한 예측된 크립틱 스플라이스 돌연변이의 유효성확인**

[0639] 신규한 접합부 형성을 초래할 것으로 예측되는 각각의 개인적 변이체에 대해, 본 발명자들은 새로 생성된 비정상 스플라이스 접합부의 위치를 예측하기 위해 망을 사용하였고, 이러한 신규 접합부가 오직 SNV를 가진 개인에게만 나타나고 다른 GTEx 개인에게는 나타나지 않는지 확인하기 위해 RNA-seq 데이터를 조사하였다. 유사하게, 엑손 X의 스플라이스 부위에 영향을 미치는 스플라이스-사이트 손실을 야기할 것으로 예상되는 변이체에 대해,

본 발명자들은 이전의 표준 엑손(GENCODE 주석에 기하여 X 상류측의 것)에서 다음의 표준 엑손(X 하류측의 것)에 이르기까지 변이체를 가진 개인에게만 나타나고 GTEX의 다른 개인에게는 나타나지 않은 새로운 엑손 스킵핑 이벤트를 찾았다. 모델에 의해 손실될 것으로 예상되는 스플라이스 부위가 GENCODE에서 주석이 달리지 않았거나 변이체가 없는 GTEX 개인에서는 관찰되지 않은 경우 예측 손실을 배제했다. 또한 GENCODE에서 이득이 예상되는 스플라이스 부위에 이미 주석이 달린 경우 예측된 이득도 제외했다. 이 분석을 공통 변이체 (2-4명의 개체로 존재)로 확장하기 위해, 변이체가 있는 개체의 적어도 절반에 존재하고 변이체가 없는 모든 개체에는 없는 새로운 접합부를 유효성확인했다.

[0640] 예측된 비정상 스플라이스 이벤트가 변이체를 가진 개인에게 개인적이라는 요건을 사용하여, 예측된 높은 점수 ($\Delta\text{Score} \geq 0.5$) 수용체 및 공여체 이득의 40%를 유효성확인할 수 있었지만, 예측된 높은 점수 손실의 3.4% 및 필수 GT 또는 AG 중단의 5.6%(치환을 기준으로 $< 0.2\%$ 의 잘못된 유효성확인율에서 - "잘못된 유효성확인율 추정" 부분 참조)만 유효성확인할 수 있었다. 이득과 손실의 유효성확인율에 차이가 있는 이유는 두 가지이다. 첫째, 이득과는 달리, 엑손 스킵핑 이벤트는 변이체를 가진 개인에게 전적으로 개인적인 경우가 거의 없는데, 이는 엑손은 종종 낮은 베이스라인 수준에서 건너뛰기 때문이며, 이는 충분히 깊은 RNA-seq로 관찰할 수 있다. 둘째로, 스플라이스-사이트 손실은 엑손 스킵 증가 외에 인트론 보유 증가 또는 대체 준최적 스플라이스 부위의 사용 증가와 같은 다른 효과를 가질 수 있다. 이러한 이유로, 본 발명자들은 모델의 예측을 검증하기 위해 개인적인 신규 접합부에 전적으로 의존하지 않았으며, 변이체를 가진 개인에게 영향을 줄 것으로 예측된 접합부 사용의 증가 또는 감소에 대한 정량적 증거를 기반으로 변이체의 유효성을 확인했다.

[0641] 정량적 기준을 통한 예측된 크립틱 스플라이스 돌연변이의 유효성확인

[0642] 샘플 s에서 접합부 j에 대해 정규화된 접합부 계수치 c_{js} 를 얻었다:

[0643]
$$c_{js} = \text{asinh} \left(\frac{r_{js}}{\sum_g r_{gs}} \right) \quad (1)$$

[0644] 여기서 r_{js} 는 샘플 s의 접합부 j에 대한 원시 접합부 계수치이며 분모의 합계는 주석이 달린 수용체와 j와 동일한 유전자의 공여체 사이의 다른 모든 접합부에 대해 취해진다(GENCODE v19의 주석 사용). asinh 변환은 $\text{asinh}(x) = \ln(x + \sqrt{x^2 + 1})$ 로 정의된다. 이는 RNA-seq 데이터를 변환하는 데 자주 사용되는 로그 변환과 유사하지만(Lonsdale et al., 2013), 0에서 정의되어 있으므로, 많은 접합부, 특히 신규 접합부들이 낮거나 0인 계수치를 가지기 때문에, 값을 상당히 왜곡했을 수 있는 유사 계수치가 필요하지 않다. asinh 변환은 큰 값의 경우 로그 변환처럼 작동하지만 작은 값의 경우 선형에 가깝다. 이러한 이유로, 0에 가까운 값을 다수 갖는 데이터 세트(RNA-seq 또는 ChIP-seq 데이터 세트와 같은)에서 종종 사용되어 소수의 큰 값이 신호를 지배하지 못하게 한다(Azad et al., 2016; Herring et al., 2018; Hoffman et al., 2012; Kasowski et al., 2013; SEQC/MAQC-III Consortium, 2014). 후술하는 바와 같이, "유효성확인에 대한 고려 기준"부분에서, 식 (1)의 분모가 200 미만인 샘플은 모든 유효성확인 분석에 대해 제외되어 수치 문제를 피한다.

[0645] 한 세트의 개인들 I에 나타나는 SNV에 의해 야기될 것으로 예상되는 각각의 획득 또는 손실된 접합부 j에 대해, 본 발명자들은 각각의 조직 t에서 다음 z-점수를 개별적으로 계산하였다:

[0646]
$$z_{jt} = \frac{\text{mean}_{s \in A_t}(c_{js}) - \text{mean}_{s' \in U_t}(c_{js'})}{\text{std}_{s' \in U_t}(c_{js'})} \quad (2)$$

[0647] 여기서 A_t 는 조직 t의 I에 있는 개인의 샘플 세트이고 U_t 는 조직 t에 있는 다른 모든 개인의 샘플 세트이다. 동일한 개인 및 조직에 대해 GTEX 데이터 세트에 여러 샘플이 있을 수 있음에 유의한다. 이전과 마찬가지로 c_{js} 는 샘플 s에서 접합부 j의 계수치이다. 예측된 손실에 대해, 추정 영향을 받는 엑손을 건너뛰는 접합부 k에 대해 유사한 z-점수를 계산했다.

[0648]
$$z_{kt} = \frac{\text{mean}_{s' \in U_t}(c_{ks'}) - \text{mean}_{s \in A_t}(c_{ks})}{\text{std}_{s' \in U_t}(c_{ks'})} \quad (3)$$

[0649] 스킵핑을 초래한 손실은 접합부 손실의 상대적 감소와 스킵핑의 상대적 증가를 초래함에 유의한다. 이것은 z_{jt} 와

z_{kt} 의 분자의 차이의 역전을 정당화하므로, 이 두 점수는 실제 스플라이스 부위 손실에 대해 음의 경향이 있다.

[0650] 마지막으로, 본 발명자들은 모든 고려된 조직에 걸친 중간 z -점수를 계산하였다. 손실에 대해, 식 (2)와 (3)에서 각각의 z -점수의 중앙값을 별도로 계산했다. 다음 중 하나라도 해당되는 경우 수용체 또는 공여체 손실 예측이 유효성확인된 것으로 간주된다.

[0651] 1. 접합부의 상대 손실을 정당화하는 식 (2)의 z -점수의 중앙값은 치환된 데이터(-1.46)의 해당 값의 5 번째 백분위 수와 식 (3)의 z -점수의 중앙값보다 작으므로, 스킵핑의 상대적인 변화를 정당화하는 것은 양성이 아니었다(제로, 음성, 또는 누락, 이는 스킵핑 접합부가 어느 개인에서도 관찰되지 않은 경우에 해당). 다시 말해, 영향을 받는 접합부의 사용 감소에 대한 강력한 증거가 있었고 영향을 받는 개인의 스킵핑의 감소를 시사하는 증거는 없었다.

[0652] 2. 식 (3)의 z -점수의 중앙값은 치환된 데이터(-0.74)에서 해당 값의 5 번째 백분위 수보다 작았고 식 (3)의 z -점수의 중앙값은 양수가 아니었다.

[0653] 3. 식 (2)에서 z -점수의 중앙값은 치환된 데이터(-2.54)에서 해당 값의 1 번째 백분위 수보다 작았다.

[0654] 4. 식 (3)으로부터 z -점수의 중앙값은 치환된 데이터(-4.08)에서 상응하는 값의 1 번째 백분위 수보다 작았다.

[0655] 5. 영향 받은 엑손을 건너뛰는 접합부는 변이체를 가진 개인의 적어도 절반에서 관찰되고 다른 개인에서는 관찰되지 않았다(상기 "개인적 스플라이스 접합부에 기초한 예측된 크립틱 스플라이스 돌연변이의 유효성확인"부분에 기술된 바와 같이).

[0656] 상기 컷오프를 얻는 데 사용된 치환에 대한 설명은 "잘못된 유효성확인을 추정"부분에 제공된다.

[0657] 경험적으로, "개인적 스플라이스 접합부에 기초한 예측된 크립틱 스플라이스 돌연변이의 유효성확인" 부분에서 설명된 바와 같이, 손실은 이득보다 더 혼합된 효과를 초래하는 경향이 있기 때문에, 이득과 비교하여 손실에 대해 보다 엄격한 유효성확인 기준을 적용할 필요가 있음을 관찰하였다. 개인적 SNV 근처에서 새로운 접합부를 관찰하는 일은 우연히 일어날 가능성이 거의 없기 때문에 접합부에 대한 사소한 증거조차도 유효성확인에 충분하다. 대조적으로, 대부분의 예측 손실은 기존 접합부의 약화를 초래했으며, 이러한 약화는 이득에 의해 야기된 온-오프 변화보다 검출하기가 더 어렵고 RNA-seq 데이터의 노이즈에 더욱 기인한 것으로 보인다.

[0658] **유효성확인 분석의 인클루전 기준**

[0659] 계수치가 낮거나 커버리지가 형편없는 상태에서 z -점수를 계산하는 것을 피하기 위해, 유효성확인 분석을 위해 변이체를 필터링하기 위하여 다음 기준을 사용했다:

$$\sum_g r_{gs} > 200$$

[0660] 1. 샘플은 유전자를 발현한 경우에만 상기 z -점수 계산을 위해 고려되었다($\sum_g r_{gs} > 200$).

[0661] 2. 변이체가 없는 개인에서 각각 손실 또는 "참조" 접합부의 평균 계수치가 10 미만인 경우, 조직은 손실 또는 이득 z -점수 계산에 대해 고려되지 않았다. "참조" 접합부는, GENCODE 주석에 기초할 때, 새로운 접합부를 얻기 이전에 사용된 표준 접합부이다(자세한 내용은 효과 크기 계산 부분 참조). 직관에 따르면 대조군 개체에 발현되지 않은 접합부에 영향을 미치는 스플라이스 손실 변이체의 유효성을 확인해서는 안 된다. 유사하게, 대조군 개체가 영향을 받는 부위에 걸친 전사체를 충분히 발현시키지 못한 경우, 스플라이스-게인 변이체의 검증을 시도하지 않아야 한다.

[0662] 3. 스플라이스-사이트 손실이 예상되는 경우, 변이체가 없는 개인의 샘플은 10 계수치 이상의 손실된 접합부가 있는 경우에만 고려된다. 예측된 수용체 또는 공여체 이득의 경우, 대조군 개인으로부터의 샘플은 "참조" 접합부의 계수치가 10회 이상인 경우에만 고려되었다. 직관에 따르면 영향을 받는 접합부의 큰 평균 발현이 있는 조직(즉, 통과 기준 2)에서도 서로 다른 샘플이 서열분석 깊이가 크게 다를 수 있으므로, 충분한 발현을 가진 대조군 샘플만 포함해야 한다.

[0663] 4. 변이체를 갖는 개체로부터 상기 기준을 통과하는 적어도 하나의 샘플이 있는 경우뿐만 아니라, 적어도 2개의 별개의 대조군 개체로부터 상기 기준을 통과하는 적어도 5개의 샘플이 존재하는 경우에만 조직을 고려하였다.

[0664] 상기 고려 기준을 만족시키는 조직이 존재하지 않는 변이체는 확정할 수 없는 것으로 간주되고, 유효성확인율을 계산할 때 배제되었다. 스플라이스 이득 변이체의 경우, 기존 GENCODE 주석이 달린 스플라이스 부위에서 발생하

는 변이체는 필터링했다. 마찬가지로 스플라이스 손실 변이체의 경우, 기존 GENCODE 주석이 달린 스플라이스 부위의 점수를 낮추는 변이체만 고려했다. 전반적으로, 55% 및 44%의 높은 점수 ($\Delta\text{Score} \geq 0.5$)의 예측 가능한 이득과 손실은 각각 확정가능한 것으로 간주되었고 유효성확인 분석에 사용되었다.

[0665] **잘못된 유효성확인을 추정**

[0666] 상기 절차가 합리적인 실제 유효성확인을 갖도록 하기 위해, 먼저 1-4 GTEx 개체에서 나타나는 SNV를 보고 필수 GT/AG 다이뉴클레오타이드를 파괴하였다. 본 발명자들은 그러한 돌연변이가 거의 확실히 스플라이싱에 영향을 미치므로 그들의 유효성확인율은 100%에 가까워져야 한다고 주장했다. 이러한 파괴 중에서, 위에서 언급한 기준에 따라 39%는 확정가능했으며, 확정가능한 것들 중 유효성확인율은 81%였다. 잘못된 유효성확인을 추정하기 위해 SNV 데이터의 개체 라벨을 치환했다. k명의 GTEx 개인에 나타난 각 SNV에 대해, k명의 GTEx 개인의 무작위 하위 집합을 선택하고 SNV를 그들에게 할당했다. 본 발명자들은 10개의 무작위 데이터 세트를 생성하고 그에 대한 유효성확인 프로세스를 반복했다. 치환된 데이터 세트의 유효성확인율은 이득의 경우 1.7-2.1%, 손실의 경우 4.3-6.9%였으며, 각각 중앙값은 1.8%와 5.7%이다. 손실에 대한 잘못된 유효성확인율이 높고 본질적 파괴에 대한 유효성확인율이 상대적으로 낮은 것은 "개인적 스플라이스 접합부에 기초하여 예측된 크립틱 스플라이스 돌연변이의 유효성확인" 부문에서 강조된 대로 스플라이스 부위 손실의 유효성확인이 어렵기 때문이다.

[0667] **RNA-seq 데이터에서 크립틱 스플라이스 변이체의 효과 크기 계산**

[0668] 본 발명자들은 변이체의 "효과 크기"를 변이체로 인해 스플라이싱 패턴을 변화시킨 영향을 받는 유전자의 전사체의 분율(예를 들어, 신규 수용체 또는 공여체로 전환된 분율)로서 정의하였다. 예측된 스플라이스-이득 변이체에 대한 참조 예로서, 도 38C의 변이체를 고려한다. 예측된 획득 공여체 A에 대해, 본 발명자들은 가장 가까운 주석이 달린 수용체 C에 대한 접합부(AC)를 먼저 식별했다. 그리고 나서 본 발명자들은 "참조" 접합부(BC)를 식별했는데, 여기서 $B \neq A$ 는 A에 가장 가까운 주석 달린 공여체이다. 그런 다음 각 샘플 s에서, 참조 접합부(B)와 비교하여 신규 접합부(AC)의 상대적인 사용량을 계산했다:

[0669]
$$u_{(AB)s} = \frac{r_{(AC)s}}{r_{(AC)s} + r_{(BC)s}} \quad (4)$$

[0670] 여기서 $r_{(AC)s}$ 는 샘플 s에서 접합부(AC)의 원시 리드 계수치이다. 각 조직에 대해, 변이체 있는 개인과 다른 모든 개인 사이의 접합부(AC) 사용의 변화를 계산했다.

[0671]
$$mean_{s \in A_t} u_{(AC)s} - mean_{s' \in U_t} u_{(AC)s'} \quad (5)$$

[0672] 여기서 A_t 는 조직 t에서 변이체를 갖는 개체의 샘플 세트이고 U_t 는 조직 t의 다른 개체의 샘플 세트이다. 최종 효과 크기는 모든 고려된 조직에 걸친 상기 차이의 중앙값으로 계산되었다. 획득된 수용체의 경우 또는 스플라이스 부위 생성 변이체가 인트론인 경우에는 계산이 비슷했다. 효과 크기 계산의 단순화된 버전(변이체 유무에 관계없이 개인으로부터 단일 샘플을 가정)이 도 38C에 도시되어 있다.

[0673] 예측된 손실에 대해, 먼저 영향을 받는 엑손을 건너뛴 전사체의 분율을 계산 하였다. 계산은 도 45에 도시되어 있다. 공여체 C의 예측된 손실을 위해, 다음 하류측 주석 달린 엑손에 대한 접합부(CE)와 상류측 엑손에서 추정적으로 영향을 받는 엑손에 이르는 접합부(AB)를 확인하였다. 본 발명자들은 영향을 받는 엑손을 건너뛴 전사체의 분율을 다음과 같이 정량화했다:

[0674]
$$k_{(AE)s} = \frac{r_{(AE)s}}{r_{(AE)s} + mean(r_{(AB)s} + r_{(CE)s})} \quad (6)$$

[0675] 이득과 관련하여, 다음으로 변이체를 갖는 개체의 샘플과 변이체가 없는 개체의 샘플 사이에서 스킵된 분율의 변화를 계산하였다:

[0676]
$$mean_{s \in A_t} k_{(AE)s} - mean_{s' \in U_t} k_{(AE)s'} \quad (7)$$

[0677] 상기 계산된 스킵된 전사체의 분율은 수용체 또는 공여체 손실의 효과를 완전히 포착하지 못하는데, 그러한 파괴는 증가된 수준의 인트론 보유 또는 준최적 스플라이스 부위의 사용을 또한 야기할 수 있기 때문이다. 이러한 영향 중 일부를 설명하기 위해 동일한 수용체 E를 사용하는 다른 접합부의 사용과 비교하여 손실된 접합부(CE)

의 사용량도 계산했다:

$$l_{(CE)s} = \frac{r_{(CE)s}}{\sum r_{(E)s}} \quad (8)$$

여기서 $\sum r_{(E)s}$ 는 임의의 (주석 달린 또는 신규의) 수용자에서 공여체 E까지의 모든 접합부들의 합이다. 여기에는, 도 45의 예에 도시된 바와 같이, 영향을 받는 접합부(CE), 스킵핑 접합부(AE), 그리고 C의 손실을 보상한 다른 준최적 공여체로부터의 잠재적 접합부가 포함된다. 그런 다음 영향을 받는 접합부의 상대적인 사용량 변화를 계산했다:

$$mean_{s' \in U_t} l_{(CE)s'} - mean_{s \in A_t} l_{(CE)s} \quad (9)$$

변이체가 있는 개인의 획득된 또는 스킵핑 접합부의 사용량 증가를 측정하는 (5) 및 (7)과 달리, (9)에서는 손실된 접합부의 사용량 감소를 측정하고자 한다는 점에 유의하고, 따라서 차이의 두 부분을 되돌린다. 각 조직에 대해, 효과 크기는 (7) 및 (9) 중 최대로 계산되었다. 이득에 관해서는, 변이체의 최종 효과 크기는 조직 전체의 중앙값 효과 크기였다.

효과 크기 분석을 위한 인클루전 기준

변이체는 이전 부문에서 기술된 기준에 기초하여 유효성확인된 것으로 간주되는 경우에만 효과 크기 계산을 위해 고려되었다. 매우 적은 수의 비정상 전사체의 비율을 계산하지 않기 위해 비정상 및 참조 접합부의 계수치들 이상인 샘플 만 고려했다. 대부분의 크립틱 스플라이스 변이체는 인트론에 있으므로, 변이체와 중복되는 참조 및 대체 리드 수를 계산하여 효과 크기를 직접 계산할 수 없었다. 따라서 손실의 효과 크기는 정상 스플라이스 접합부의 상대적 사용 감소로부터 간접적으로 계산된다. 신규한 접합부 이득의 효과 크기에 대해, 비정상 전사체가 넌센스 매개 붕괴에 의해 영향을 받아, 관찰된 효과 크기를 약화시킬 수 있다. 이러한 측정의 한계에도 불구하고, 본 발명자들은 이득과 손실 이벤트 모두에서 낮은 점수의 크립틱 스플라이스 변이체에 대해 더 작은 효과 크기를 향한 일관된 경향을 관찰한다.

완전 침투 이형접합성 개인적 SNV의 예상 효과 크기

변이체를 갖는 개체의 변이체 1배체형으로부터의 모든 전사체가 신규 접합부로 전환되도록 초래하는 완전 침투성 스플라이스 부위 생성 변이체에 대해, 신규한 접합부가 대조군 개체에서 발생하지 않는다고 가정하면, 예상되는 효과 크기는 식 (5)에 의해 0.5가 된다.

유사하게, 이형접합성 SNV가 신규 엑손 스킵핑 이벤트를 야기하고, 영향 받은 1배체형의 모든 전사체가 스킵핑 접합부로 전환된 경우, 식 (7)에서의 예상 효과 크기는 0.5이다. 변이체가 있는 개인의 모든 전사체가 다른 접합부(건너뛰는 접합부 또는 다른 보상 접합부 중 하나)로 전환된 경우, 식 (8)의 비율은 변이체가 있는 개인의 샘플에서 0.5, 다른 개인의 샘플에서 1이 되므로, 식 (9)에서의 차이는 0.5이다. 이는 변이체가 없는 개인에서 스킵핑이나 수용자 E로의 다른 접합부가 없는 것으로 가정한다. 또한 스플라이스 부위 파괴가 인트론 보유를 촉발하지 않는다고 가정한다. 실제로, 적어도 낮은 수준의 인트론 보유는 종종 스플라이스 부위 파괴와 관련이 있다. 또한, 스플라이스-변경 변이체가 없는 경우에도 엑손 스킵핑은 널리 퍼져 있다. 이는 필수 GT/AG 다이뉴클레오타이드를 방해하는 변이체에 대해서조차도, 측정된 효과 크기가 0.5 미만인 이유를 설명한다.

완전 침투성 이형접합 변이체에 대해 0.5의 효과 크기의 기대는 또한 변이체가 넌센스 매개 붕괴(NMD)를 유발하지 않았다고 가정한다. NMD가있는 경우 식 (4), (6) 및 (8)의 분자와 분모가 모두 떨어지므로 관측된 효과 크기가 줄어든다.

넌센스 매개 붕괴(NMD)를 통해 열화된 전사체의 비율

도 38C의 경우 변이체가 엑손이었으므로, 변이체에 걸쳐 있고 참조 또는 대체 대립 유전자 (각각 "Ref(no splicing)" 및 "Alt(no splicing)")를 갖는 리드 횟수를 계수치할 수 있었다. 또한 새로운 스플라이스 부위에서 스플라이싱하고 추정컨대 대체 대립 유전자 ("Alt(novel junction)")를 가지고 있는 리드 횟수도 계수치할 수 있었다. 도 38C의 예와 우리가 살펴본 많은 다른 경우에서, 본 발명자들은 대체 대립 유전자 ("Alt(no splicing)" 및 "Alt(novel junction)"의 합)를 가진 1배체형으로부터 오는 총 리드 횟수가 참조 대립 유전자를 사용한 리드 횟수("Ref(no splicing)")보다 적었음을 관찰한다. 본 발명자들은 리드 맵핑 동안의 참조 편향을

제거했다고 믿으므로, 참조 및 대체 1배체형 모두에 매핑하고, 리드 횟수가 각 대립 유전자의 전사체 수에 비례한다고 가정함으로써, 참조 대립 유전자가 변이체 좌위에서의 리드의 절반을 설명할 것으로 예상하였다. 본 발명자들은 "누락된" 대체 대립유전자 리드가 신규 접합부에서 스플라이싱했던 대체 대립유전자 1배체형으로부터의 전사체에 대응하며 년센스 매개 붕괴(NMD)를 통해 열화되었다고 가정한다. 본 발명자들은 이 그룹을 "Alt(NMD)"라고 불렀다.

[0690] 관찰된 참조 리드 수와 대체 리드 수 사이의 차이가 유의미한지 여부를 결정하기 위해, 본 발명자들은 성공 확률 0.5를 갖는 이항 분포 하에서 Alt(no splicing) + Alt(novel junction)(또는 더 적은) 리드를 관찰할 확률과 Alt(no splicing) + Alt(novel junction) + Ref(no splicing)의 총 시행 횟수를 계산하였다. 이것은 잠재적으로 열화된 전사체를 계수하지 않음으로써 총 "시도" 수를 과소평가하기 때문에 보수적인 p-값이다. 도 38C의 NMD 전사체의 비율은 신규 접합부(Alt(NMD) + Alt(novel junction))에서 스플라이싱하는 총 리드 수에 대한 "Alt(NMD)" 리드의 수로 계산되었다.

[0691] **크립틱 스플라이스 접합부 검출시 망의 민감도**

[0692] SpliceNet 모델의 민감도를 평가하기 위해(도 38F), 본 발명자들은 영향을 받는 스플라이스 부위로부터 20 nt 이내에 있고(즉, 신규 또는 파괴된 수용체 또는 공여체) 주석이 달린 엑손의 필수 GT/AG 다이뉴클레오타이드와 겹치지 않으며 추정 효과 크기는 0.3 이상인("효과 크기 계산"부문 참조) SNV를 사용하였다. 모든 민감도 플롯에서, SNV는 그들이 주석이 달린 엑손과 겹치거나 주석이 달린 엑손의 경계의 50 nt 이내에 있는 경우 "엑손 부근"으로 정의되었다. 다른 모든 SNV는 "심층 인트론의"로 간주되었다. 강력하게 지원되는 크립틱 스플라이스 부위의 이 진실 데이터 세트를 사용하여 다양한 ΔScore 임계값에서 모델을 평가하고 해당 컷오프에서 모델이 예측한 진실 데이터 세트에서 크립틱 스플라이스 부위의 비율을 보고했다.

[0693] **기존 스플라이싱 예측 모델과의 비교**

[0694] 본 발명자들은 다양한 메트릭과 관련하여 SpliceNet-10k, MaxEntScan(Yeo and Burge, 2004), GeneSplicer(Pertea et al., 2001) 및 NNSplice(Reese et al., 1997)의 일대일 비교를 수행하였다. MaxEntScan 및 GeneSplicer 소프트웨어는 각각 <http://genes.mit.edu/burgelab/maxent/download/> 및 <http://www.cs.jhu.edu/~genomics/GeneSplicer/> 에서 다운로드했다. NNSplice는 다운로드 가능한 소프트웨어로 제공되지 않으므로 http://www.fruitfly.org/data/seq_tools/datasets/Human/GENIE_96/splicesets/ 에서 트레이닝 및 테스트 세트를 다운로드했으며, (Reese et al., 1997)에 설명된 최고 성능의 아키텍처로 모델을 트레이닝했다. 온전성 검사로서, 본 발명자들은 (Reese et al., 1997)에 보고된 테스트 세트 메트릭을 재현했다.

[0695] 이들 알고리즘의 top-k 정확도 및 정밀 재호출 곡선 하의 면적을 평가하기 위해, 각각의 알고리즘으로 테스트 세트 유전자 및 lincRNA의 모든 위치에 대해 점수를 매겼다(도 37D).

[0696] MaxEntScan 및 GeneSplicer 출력은 로그 승산비에 대응하는 반면, NNSplice 및 SpliceNet-10k 출력은 확률에 대응한다. 본 발명자들은 MaxEntScan과 GeneSplicer에게 최고의 성공 기회를 제공하기 위해, 기본 출력과, 출력을 확률에 대응하도록 먼저 변환하는 변환된 출력을 사용하여 ΔScore를 계산했다. 보다 정확하게는 MaxEntScan의 기본 출력은,

$$x = \log_2 \frac{p(\text{스플라이스 부위})}{p(\text{비-스플라이스 부위})}$$

[0698] 에 대응하는데,

$$\frac{2^x}{2^x + 1}$$

[0699] 이는, 변환 후, 원하는 수량에 해당한다. RETURN_TRUE_PROB 플래그를 한번은 0으로 설정하고 한번은 1로 설정하여 GeneSplicer 소프트웨어를 두 번 컴파일했다. RNA-seq 데이터에 대해 최상의 유효성확인율을 제공하는 출력전략을 선택했다(MaxEntScan: 변환된 출력, GeneSplicer: 기본 출력).

[0700] 다양한 알고리즘의 유효성확인율 및 민감도를 비교하기 위해(도 38G), 모든 알고리즘이 게놈 전체에서 동일한 수의 이득 및 손실을 예측하는 컷오프를 발견하였다. 즉, SpliceNet-10k ΔScore 값의 각 컷오프에 대해 각 경쟁 알고리즘이 SpliceNet-10k와 동일한 수의 이득 예측 및 동일한 수의 손실 예측을 하는 컷오프를 발견했다. 선택된 컷오프는 표 S2에 나와 있다.

[0701] **싱글톤 대 공통 변이체에 대한 변이체 예측의 비교**

[0702] 본 발명자들은 2-4 GTE_x 개체에서 나타나는 싱글톤 SNV와 SNV에 대해 개별적으로 유효성확인 및 민감도 분석("민감도 분석" 및 "모델 예측의 유효성확인" 부문에서 설명됨)을 수행하였다(도 46A, 도 46B 및 도 46C). 싱글톤과 공통 변이체간에 유효성확인율이 유의미하게 다른지 여부를 테스트하기 위해 각 ΔScore 그룹 (0.2 - 0.35, 0.35 - 0.5, 0.5 - 0.8, 0.8 - 1)에서 유효성확인율과 각 예측된 효과(수용자 또는 공여체 이득 또는 손실)에 대한 유효성확인율을 비교하여 Fisher Exact 테스트를 수행했다. 16개의 테스트를 설명하기 위해 Bonferroni 보정 후 모든 P-값이 0.05보다 컸다. 본 발명자들은 싱글톤 또는 공통 변이체를 검출하는 민감도를 유사하게 비교했다. 본 발명자들은 Fisher Exact 테스트를 사용하여 두 변이체 그룹 간에 유효성확인율이 크게 다른지 여부를 테스트했다. 심층 인트론 변이체와 엑손 부근의 변이체를 별개로 고려하여 두 가지 테스트에 대해 Bonferroni 보정을 수행했다. 0.05 컷오프를 사용해서는 P-값 중 어느 것도 유의미하지 않았다. 따라서, 싱글톤 및 공통 GTE_x 변이체를 조합하고 이들을 도 48A, 도 48B, 도 48C, 도 48d, 도 48e, 도 48f 및 도 48g 및 도 39A, 도 39B 및 도 39C에 제시된 분석을 위해 함께 고려하였다.

[0703] **트레이닝 대 테스트 염색체에 대한 변이체 예측의 비교**

[0704] 본 발명자들은 트레이닝 동안 사용된 염색체상의 변이체 및 나머지 염색체의 변이체 사이에서 RNA-seq에 대한 유효성확인율 및 SpliceNet-10k의 민감도를 비교하였다(도 48A 및 도 48B). Bonferroni 보정 후 모든 P-값이 0.05보다 컸다. 또한, 아래의 "유해 변이체의 분율" 부문에 기술된 바와 같이, 트레이닝 및 테스트 염색체의 변이체에 대해 유해한 변이체의 분율을 별도로 계산하였다(도 48C). 각 ΔScore 그룹 및 각 유형의 변이체에 대해 Fisher Exact 테스트를 사용하여 트레이닝 및 테스트 염색체 간의 공통 및 희귀 변이체의 수를 비교했다. 12번의 테스트에 대한 Bonferroni 보정 후 모든 P-값은 0.05보다 컸다. 마지막으로, 본 발명자들은 "코호트 당 드 노보 돌연변이의 농축" 부문에 기술된 바와 같이 트레이닝 및 테스트 염색체(도 48d)에 대한 크립틱 스플라이스 드 노보 변이체의 수를 계산하였다.

[0705] **다양한 유형의 크립틱 스플라이스 변이체에 대한 변이체 예측의 비교**

[0706] 본 발명자들은 예측된 스플라이스 부위 생성 변이체를 3개의 그룹으로 나누었다: 신규 GT 또는 AG 스플라이스 다이뉴클레오타이드를 생성하는 변이체, 나머지 스플라이싱 모티프와 겹치는 변이체(엑손-인트론 경계 주위의 위치는 엑손으로 최대 3 nt, 인트론으로 최대 8 nt), 및 스플라이스 모티프 외부의 변이체(도 47A 및 도 47B). 각 ΔScore 그룹 (0.2 - 0.35, 0.35 - 0.5, 0.5 - 0.8, 0.8 - 1)에 대해 χ^2 테스트를 수행하여 세 가지 유형의 스플라이스 부위 생성 변이체에서 유효성확인율이 균일하다는 가설을 테스트했다. 모든 테스트는 다중 가설 교정 이전에도 P-값 > 0.3을 산출했다. 세 가지 유형의 변이체 간의 효과 크기 분포를 비교하기 위해 Mann-Whitney U 테스트를 사용하여 각 ΔScore 그룹에 대해 세 가지 변이체 유형을 모두 비교했다(총 4×3 = 12 테스트). 12번의 테스트에 대한 Bonferroni 보정 후 모든 P-값은 > 0.3 이었다.

[0707] **조직-특이적 스플라이스-이득 변이체의 검출**

[0708] 도 39C의 경우, 본 발명자들은 신규 접합부의 사용률이 영향을 받는 유전자를 발현하는 조직에 걸쳐 균일한지를 테스트하고 싶었다. 본 발명자들은 새로운 개인적 스플라이스 부위를 생성한 SNV, 즉, 이 변이체를 가진 개인의 절반 이상에서만 나타났고 다른 개인에게는 나타나지 않은 스플라이스 접합부를 얻는 결과가 된 SNV에 중점을 두었다. 이러한 새로운 접합부 j 각각에 대해, 본 발명자들은 각 조직 t에서 조직 내 변이체를 가진 개인의 모

든 샘플에 대한 접합부의 총 개수를 계산했다: $\sum_{s \in A_t} n_{js}$. 여기서 A_t 는 조직 t의 변이체를 가진 개인의 샘플 세트이다. 마찬가지로, 동일한 표본에 대해 유전자의 모든 주석이 달린 접합부의 총 수를 계산했는데 $\sum_{s \in A_t} \sum_g n_{gs}$, 여기서 g는 유전자의 주석이 달린 접합부를 나타낸다. 유전자의 배경 계수치에 대해 정규화된, 조직 t에서 신규 접합부의 상대적인 사용은 다음과 같이 측정될 수 있다:

$$m_t = \frac{\sum_{s \in A_t} n_{js}}{\sum_{s \in A_t} (n_{js} + \sum_g n_{gs})}$$

[0709]

[0710] 본 발명자들은 또한 조직 전체에 걸친 집합부의 평균 사용을 계산하였다:

$$m = \frac{\sum_t \sum_{s \in A_t} r_{js}}{\sum_t \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})}$$

[0711]

[0712] 집합부의 상대적인 사용이 조직 전체에서 균일하고 m과 같다는 가설을 테스트하고자 했다. 본 발명자들은 관찰

된 조직 계수치 $\sum_{s \in A_t} r_{js}$ 를 균일한 속도를 가정한 예상 계수치 $m \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})$ 와 비교하여

χ^2 테스트를 수행했다. Bonferroni- 보정된 χ^2 p-값이 10^{-2} 미만인 경우 스플라이스-사이트 생성 변이체는 조직-특이적인 것으로 간주되었다. 테스트의 자유도는 $T - 1$ 이며, 여기서 T는 고려되는 조직의 수이다. 유효성확인 부문에 기술된 고려 기준을 만족하는 조직만이 테스트에 사용되었다. 또한, 균일성 테스트에 전력이 부족해서 계수치가 적었던 사례를 피하기 위해, 적어도 3개의 고려된 조직, 평균적으로 조직 당 하나 이상의 비정상 리드(즉, $m > 1$), 및 고려된 모든 조직 걸쳐 총계로 적어도 15 이상의 비정상 리드를 갖는(즉,

$$\sum_t \sum_{s \in A_t} r_{js} > 15$$

) 균일성 변이체에 대해서만 테스트하였다. 이 클래스의 변이체는 일반적으로 효과 크기가 낮고 집합부 계수치가 적으므로 Δ Score가 0.35 미만인 모든 변이체는 무시했다. 본 발명자들은 이 클래스에서 조직-특이적 변이체의 비율이 매우 낮다는 것을 관찰했지만, 이것은 전력 문제 때문이라고 생각한다.

[0713] **III. ExAC 및 gnomAD 데이터 세트에 대한 분석**

[0714] **변이체 필터링**

[0715] 본 발명자들은 ExAC 브라우저(Lek et al., 2016)로부터 Sites VCF 릴리스 0.3 파일(60,706 엑숨) 및 gnomAD 브라우저로부터 Sites VCF 릴리스 2.0.1 파일(15,496 전체 게놈)을 다운로드 하였다. SpliceNet-10k를 평가하기 위해 그것들로부터 필터링된 변이체 목록을 작성했다. 특히 다음 기준을 만족하는 변이체가 고려되었다.

- [0716] ● FILTER 필드는 PASS이다.
- [0717] ● 변이체는 단일 뉴클레오타이드 변이체이고 단지 하나의 대체 뉴클레오타이드가있었다.
- [0718] ● AN 필드(호출된 표현형에서 총 대립유전자 수)의 값은 10,000 이상이었다.
- [0719] ● 이 변이체는 표준 GENCODE 전사체의 전사 시작과 끝 부위 사이에 있었다.
- [0720] 총 7,615,051 및 73,099,995 변이체가 각각 ExAC 및 gnomAD 데이터 세트에서 이들 필터를 통과하였다.

[0721] **유해 변이체의 분류**

[0722] 이 분석을 위해, 코호트에서 싱글톤 또는 공통(대립 유전자 빈도 (AF) $\geq 0.1\%$)인 ExAC 및 gnomAD 필터링된 목록의 변이체만 고려했다. 본 발명자들은 GENCODE 표준 주석에 따른 게놈 위치에 기하여 이러한 변이체들을 하위 분류했다.

- [0723] ● 엑소닉: 이 그룹은 동의 ExAC 변이체(676,594 싱글톤 및 66,524 공통)으로 구성된다. 이 그룹에서 변이체의 가장 유해한 부분이 스플라이싱 변화로 인한 것임을 보장하기 위해 여기에서 미스센스 변이체는 고려되지 않았다.
- [0724] ● 인트로닉 부근: 이 그룹은 표준 엑손 경계에서 3 ~ 50nt 사이의 인트로닉 ExAC 변이체로 구성된다. 더 정확하게는, 수용체 이득/손실 및 공여체 이득/손실 변이체의 분석을 위해, 스플라이스 수용체 및 공여체로부터 각각 3-50 nt 인 변이체만 고려되었다(수용체 이득/손실에 대해 575,636 싱글톤 및 48,362 공통, 공여체 이득/손실에 대해 567,774 싱글톤 및 50,614 공통).
- [0725] ● 심층적 인트로닉: 이 그룹은 표준 엑손 경계에서 50 nt 이상 떨어진 인트로닉 gnomAD 변이체로 구성된다 (34,150,431 싱글톤 및 8,215,361 공통).

[0726] 각 변이체에 대해 SpliceNet-10k를 사용하여 4가지 스플라이스 유형에 대한 ΔScore를 계산했다. 그런 다음 각 스플라이스 유형에 대해, 본 발명자들은 2개의 행이 예측된 스플라이스 변경 변이체(스플라이스 유형에 대한 적절한 범위의 ΔScore) 대 예측된 스플라이스 미변경 변이체(모든 스플라이스 유형에 대해 ΔScore < 0.1)에 대응하는 2 × 2 카이-제곱 우연성 표를 구성했으며, 두 열은 싱글톤과 공통 변이체에 해당한다. 스플라이스 이득 변이체의 경우 기존 GENCODE 주석이 달린 스플라이스 부위에서 발생하는 변이체들을 필터링했다. 마찬가지로 스플라이스 손실 변이체의 경우 기존 GENCODE 주석이 달린 스플라이스 부위의 점수를 낮추는 변이체만 고려했다. 승산비(odds ratio)를 계산하고 유해한 변이체의 분율을 다음과 같이 추정했다:

$$\left(1 - \frac{1}{\text{승산비}}\right) \times 100\%$$

[0727] ExAC 및 gnomAD 필터링된 목록에서 단백질 절단 변이체는 다음과 같이 식별되었다:
 [0728] ExAC 및 gnomAD 필터링된 목록에서 단백질 절단 변이체는 다음과 같이 식별되었다:

[0729] • **넌센스:** VEP(McLaren et al., 2016) 결과는 'stop_gained'(ExAC에서 44,046 싱글톤 및 722 공통, gnomAD에서 20,660 싱글톤 및 970 공통)이었다.

[0730] • **프레임시프트:** VEP 결과는 'frameshift_variant'이다. 변이체 필터링 동안 단일 뉴클레오타이드 변이체 기준을 완화하여 이 그룹(ExAC에서 48,265개의 싱글톤 및 896 공통, gnomAD에서 30,342개의 싱글톤 및 1,472 공통)을 생성하였다.

[0731] • **필수 수용체/공여체 손실:** 이 변이체는 표준 인트론의 첫 번째 또는 마지막 두 위치에 있었다(ExAC에서 29,240 싱글톤 및 481 공통, gnomAD에서 12,387 싱글톤 및 746 공통).

[0732] 단백질-절단 변이체에 대한 2x2 카이-제곱 우발성 표는 ExAC 및 gnomAD 필터링된 리스트를 위해 구성되었고, 유해 변이체의 분율을 추정하는데 사용되었다. 여기서, 2개의 행은 단백질 절단 대 동의 변이체에 대응하고, 2개의 열은 이전과 같이 싱글톤 대 공통 변이체에 대응하였다.

[0733] ExAC(엑소닉 및 니어 인트로닉) 및 gnomAD(심층 인트로닉) 변이체에 대한 결과가 도 40B 및 도 40D에 도시되어 있다.

[0734] **프레임시프트 대 프레임내 스플라이스 이득**

[0735] 이 분석을 위해, 본 발명자들은 엑소닉(동의 만) 또는 니어 인트로닉이고 코호트에서 싱글톤 또는 공통(AF ≥ 0.1%)인 ExAC 변이체에 주의를 집중시켰다. 수용체 이득 변이체를 프레임내 또는 프레임시프트로 분류하기 위해, 표준 스플라이스 수용체와 새로 생성된 스플라이스 수용체 사이의 거리를 측정하고 그것이 3의 배수인지 아닌지를 점검했다. 표준 스플라이스 공여체와 새로 생성된 스플라이스 공여체 간의 거리를 측정하여 공여체 게인 변이체를 유사하게 분류했다.

[0736] 유해한 프레임내 스플라이스 이득 변이체의 분율은, 2개의 행이 예측된 프레임내 스플라이스 이득 변이체(수용체 또는 공여체 이득에 대해 ΔScore ≥ 0.8) 대 예측된 스플라이스 미변경 변이체(모든 스플라이스 유형에 대해 ΔScore < 0.1 미만)에 대응하는 2 × 2 카이-제곱 우연성 표에서 추정되었고, 두 열은 싱글톤 대 공통 변이체에 해당하였다. 이 절차는 우발성 테이블의 첫 번째 행을 예측된 프레임시프트 스플라이스 이득 변이체로 대체함으로써 프레임시프트 스플라이스 이득 변이체에 대해 반복되었다.

[0737] 도 40C에 도시된 p-값을 계산하기 위해, 예측된 스플라이스 이득 변이체만을 사용하여 2 × 2 카이-제곱 우발성 표를 구성하였다. 여기서, 2개의 행은 프레임내 대 프레임시프트 스플라이스 이득 변이체에 대응하고, 2개의 열은 이전과 같이 싱글톤 대 공통 변이체에 대응했다.

[0738] **개인당 크립틱 스플라이스 변이체의 수**

[0739] 개인당 희귀 기능성 크립틱 스플라이스 변이체의 수를 추정하기 위해(도 40E), 먼저 각 대립 유전자에 각 gnomAD 변이체를 대립 유전자 빈도와 동일한 확률로 포함시켜 100개의 gnomAD 개체를 시뮬레이션하였다. 다시 말해, 각 변이체는 각 개인에 대해 독립적으로 2배로 샘플링되어 이배수체를 모방한다. 본 발명자들은 각각 0.2, 0.2 및 0.5 이상의 ΔScore를 가진 1 인당 희귀(AF < 0.1%) 엑소닉(동의 만), 니어 인트로닉 및 심층적 인트로닉 변이체의 수를 계수치했다. 이들은 비교적 허용되는 ΔScore 임계값으로, 민감도를 최적화하는 동시에 예측된 변이체의 40% 이상이 유해하다는 것을 보장한다. 이 컷오프에서, 본 발명자들은 1인당 평균 7.92 동의/

니어 인트로닉 및 3.03 심층적 인트로닉 회귀 크립틱 스플라이스 변이체를 얻었다. 이러한 변이체가 모두 기능적인 것은 아니기 때문에, 이러한 컷오프에서 유해한 변이체의 비율을 계수치에 곱했다.

[0740] **IV. DDD 및 ASD 데이터 세트에 대한 분석**

[0741] **크립틱 스플라이싱 드 노보 돌연변이**

[0742] 본 발명자들은 공개된 드 노보 돌연변이(DNM)를 수득하였다. 여기에는 자폐 스펙트럼 장애가 있는 3953명의 프로밴드(Dong et al., 2014; Iossifov et al., 2014; De Rubeis et al., 2014), Deciphering Developmental Disorders 코호트(McRae et al., 2017)의 4293명의 프로밴드 및 2073명 건강한 대조군(Iossifov et al., 2014)이 포함되었다. 품질이 낮은 DNM은 분석에서 제외되었다(ASD 및 건강한 대조군: 신뢰도 = lowConf, DDD: PP(DNM) < 0.00781, (McRae et al., 2017)). DNM을 망으로 평가하였고, 컨텍스트에 따라 크립틱 스플라이스 돌연변이를 분류하기 위해 ΔScore(상기 방법 참조)를 사용하였다. synonymous_variant, splice_region_variant, intron_variant, 5_prime_UTR_variant, 3_prime_UTR_variant 또는 missense_variant의 VEP 결과로 주석이 달린 돌연변이만 고려했다. 본 발명자들은 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F 및 도 50A 및 도 50B에 대해 ΔScore > 0.1 인 부위와 도 49A, 도 49B 및 도 49C에 대해 ΔScore > 0.2 인 부위를 사용하였다.

[0743] 도 20, 도 21, 도 22, 도 23 및 도 24는 SpliceNet-80nt, SpliceNet-400nt, SpliceNet-2k 및 SpliceNet-10k 아키텍처에 대한 자세한 설명을 보여준다. 4개의 아키텍처는 관심 위치의 각 측면에서 각각 길이 40, 200, 1,000 및 5,000의 플랭킹 뉴클레오타이드 서열을 입력으로서 사용하고, 위치가 스플라이스 수용체일 확률, 스플라이스 공여체일 확률 및 둘 다 아닐 확률을 출력한다. 아키텍처는 주로 컨볼루션층 Conv(N, W, D)로 구성되며, 여기서 N, W 및 D는 각각 층의 각 컨볼루션 커널의 컨볼루션 커널 수, 윈도우 크기 및 팽창률이다.

[0744] 도 42A 및 도 42B는 lincRNA에 대한 다양한 스플라이싱 예측 알고리즘의 평가를 도시한다. 도 42A는 lincRNA상에서 평가될 때 다양한 스플라이싱 예측 알고리즘의 top-k 정확도와 정밀도-재호출 곡선 하의 면적을 보여준다. 도 42B는 MaxEntScan 및 SpliceNet-10k를 사용하여 접수가 매겨진 LINC00467 유전자에 대한 전체 전구체-mRNA 전사체를 예측된 수용체(적색 화살표) 및 공여체(녹색 화살표) 부위 및 엑손의 실제 위치와 함께 보여준다.

[0745] 도 43A 및 도 43B는 TACTAAC 분기점 및 GAAGAA 엑소닉-스플라이스 인헨서 모티프의 위치의존적 효과를 예시한다. 도 43A와 관련하여, 최적의 분기점 서열 TACTAAC를 14,289개의 테스트 세트 스플라이스 수용체 각각로부터 다양한 거리로 도입하였고, SpliceNet-10k를 사용하여 수용체 점수를 계산하였다. 예측된 수용체 점수의 평균 변화는 스플라이스 수용체로부터의 거리의 함수로 표시된다. 스플라이스 수용체로부터의 거리가 20 내지 45 nt 일 때 예측 점수가 증가하고; 20nt 미만의 거리에서, TACTAAC는 폴리퍼리미딘 트랙트를 방해하여 예측된 수용체 점수가 매우 낮다.

[0746] 도 43B에서, SR-단백질 육량체 모티프 GAAGAA는 14,289개의 테스트 세트 스플라이스 수용체 및 공여체 각각으로부터 다양한 거리에서 유사하게 도입되었다. 예측된 SpliceNet-10k 수용체 및 공여체 점수의 평균 변화는 각각 스플라이스 수용체 및 공여체로부터의 거리의 함수로 표시된다. 모티프가 엑소닉 측에 있고 스플라이스 부위로부터 ~ 50nt 미만일 때 예측 점수가 증가한다. 엑손까지 더 먼 거리에서, GAAGAA 모티프는 고려중인 스플라이스 수용체 또는 공여체 부위의 사용을 선호하지 않는 경향이 있는데, 이는 아마도 더 근위적인 수용체 또는 공여체 모티프를 우선적으로 지지하기 때문이다. GAAGAA가 인트론에 매우 가까운 위치에 있을 때 매우 낮은 수용체 및 공여체 점수는 연장된 수용체 또는 공여체 스플라이스 모티프의 붕괴로 인한 것이다.

[0747] 도 44A 및 도 44B는 스플라이싱에 대한 뉴클레오솜 포지셔닝의 효과를 도시한다. 도 44A와 관련하여, 1백만개의 무작위로 선택된 유전자간 위치에서, 150 nt 이격된 강한 수용체 및 공여체 모티프가 도입되었고 엑손 인클루전 가능성은 SpliceNet-10k를 사용하여 계산되었다. SpliceNet-10k 예측과 뉴클레오솜 포지셔닝 사이의 상관관계가 GC 구성과 무관하게 발생함을 보여주기 위해, 위치는 GC 함량(도입된 스플라이스 부위 사이의 150개 뉴클레오타이드를 사용하여 계산)과 SpliceNet-10k 예측 사이의 Spearman 상관관계에 따라 비닝되었고, 뉴클레오솜 신호는 각 빈에 대해 플로팅된다.

[0748] 도 44B와 관련하여, 테스트 세트로부터의 스플라이스 수용체 및 공여체 부위를 SpliceNet-80nt(국부적 모티프 점수로 지칭함) 및 SpliceNet-10k를 사용하여 점수를 매기고, 점수를 뉴클레오솜 농축의 함수로서 플로팅하였다. 뉴클레오솜 농축은 스플라이스 부위의 엑소닉 측에서 평균 50 nt에 걸쳐 평균화된 뉴클레오솜 신호를 스플라이스 부위의 인트로닉 측에서 평균 50 nt에 걸쳐 평균화된 뉴클레오솜 신호로 나눈 값으로 계산된다. 모티프 강도의 대용물인 SpliceNet-80nt 점수는 뉴클레오솜 농축과 음의 상관관계를 맺고 있는 반면, SpliceNet-10k 점수는 뉴클레오솜 농축과 양의 상관관계를 맺고 있다. 이것은 뉴클레오솜 위치가 약한 국부적

스플라이스 모티프를 보상할 수 있는 장거리 특이성 결정자임을 시사한다.

- [0749] 도 45는 복잡한 효과를 갖는 스플라이스 중단 변이체에 대한 효과 크기를 계산하는 예를 도시한다. 인트로닉 변이체 chr9:386429 A>G는 정상 공여체 부위(C)를 교란시키고 이전에 억제된 인트로닉 하류 공여체(D)를 활성화시킨다. 변이체를 갖는 개체 및 대조군 개체로부터의 전혈에서의 RNA-seq 커버리지 및 접합부 리드 계수치가 도시되어 있다. 변이체를 갖는 개체 및 대조군 개체의 공여체 부위는 각각 청색 및 회색 화살표로 표시된다. 굵은 적색 글자는 접합부 끝점에 해당한다. 가시성을 위해 엑손 길이는 인트론 길이에 비해 4배 과장되었다. 효과 크기를 추정하기 위해, 동일한 공여체 E를 갖는 다른 모든 접합부에 비해 엑손 스킵핑 접합부(AE)의 사용 증가와 중단 접합부(CE)의 사용 감소를 모두 계산한다. 최종 효과 크기는 두 값 중 최대값(0.39)이다. 돌연변이된 샘플에는 증가된 양의 인트론 보유가 존재한다. 이러한 가변 효과는 엑손 스킵핑 이벤트에서 일반적이며 수용체 또는 공여체 부위 손실을 유발할 것으로 예측되는 희귀 변이체를 유효성확인하는 복잡성을 증가시킨다.
- [0750] 도 46A, 도 46B 및 도 46C는 싱글톤 및 공통 변이체에 대한 SpliceNet-10k 모델의 평가를 보여준다. 도 46A와 관련하여, GTEx RNA-seq 데이터에 대해 유효성확인된 SpliceNet-10k에 의해 예측된 크립틱 스플라이스 돌연변이 분율. 모델은 GTEx 코호트에서 최대 4명의 개인에게 나타나는 모든 변이체에 대해 평가되었다. 스플라이스 변경 효과가 예측된 변이체는 RNA-seq 데이터에 대해 유효성확인되었다. 단일 GTEx 개인(왼쪽)에 나타나는 변이체와 2-4명의 GTEx 개인(오른쪽)에 나타나는 변이체에 대해서는 유효성확인율이 별도로 표시된다. 예측은 Δ Score로 그룹화된다. 각 Δ Score 그룹에서 4가지 변이체 클래스(수용자 또는 공여체의 이득 또는 손실) 각각에 대해 싱글톤과 공통 변이체 간에 유효성확인율을 비교했다. 차이는 크지 않다($P > 0.05$, 16회 테스트에 대해 Bonferroni 보정을 사용한 Fisher Exact 테스트).
- [0751] 도 46B와 관련하여, 상이한 Δ Score 컷오프에서 GTEx 코호트에서 스플라이스 변경 변이체를 검출할 때 SpliceNet-10k의 민감도. 모델의 민감도는 싱글톤(왼쪽)과 공통(오른쪽) 변이체에 대해 별도로 표시된다. 0.2의 Δ Score 컷오프에서 싱글톤과 공통 변이체 사이의 민감도 차이는 엑손 또는 심층적 인트론 변이체 근처의 변이체에서 유의미하지 않다($P > 0.05$, 2개의 테스트에 대한 Bonferroni 보정을 사용한 Fisher Exact 테스트).
- [0752] 도 46C와 관련하여, 유효성확인된 싱글톤 및 공통 변이체에 대한 Δ Score 값의 분포. P-값은 싱글톤 및 공통 변이체 점수를 비교하는 Mann-Whitney U 테스트에 대한 것이다. 공통 변이체는 큰 효과로 스플라이스 파괴 돌연변이를 걸러내는 자연적 선택으로 인해 Δ Score 값이 상당히 약하다.
- [0753] 도 47A 및 도 47A는 스플라이스 부위 생성 변이체의 유효성확인율 및 효과 크기를 변이체의 위치에 의해 분할하여 도시한 것이다. 예측된 스플라이스 부위-생성 변이체는 변이체가 새로운 필수 GT 또는 AG 스플라이스 다이뉴클레오타이드를 생성 하였는지, 스플라이스 모티프의 나머지 부분과 겹쳤는지(엑손-인트론 경계 주위의 모든 위치는 엑손에서 3nt, 인트론에서 8nt까지, 필수 다이뉴클레오타이드는 제외), 또는 스플라이스 모티프 외부에 있었는지 여부에 기초하여 그룹화되었다.
- [0754] 도 47A와 관련하여, 스플라이스 부위 생성 변이체의 세 가지 범주 각각에 대한 유효성확인율. 각 카테고리의 총 변이체 수는 막대 위에 표시된다. 각각의 Δ Score 그룹 내에서, 세 그룹의 변이체 사이의 유효성확인율의 차이는 유의미하지 않다($P > 0.3$, χ^2 균일성 테스트).
- [0755] 도 47B와 관련하여, 스플라이스 부위 생성 변이체의 세 가지 범주 각각에 대한 효과 크기의 분포. 각각의 Δ Score 그룹 내에서, 세 그룹의 변이체 사이의 효과 크기의 차이는 유의미하지 않다($P > 0.3$, χ^2 Bonferroni 보정을 사용한 Mann-Whitney U 테스트).
- [0756] 도 48A, 도 48B, 도 49C 및 도 48D는 트레이닝 및 테스트 염색체에서의 SpliceNet-10k 모델의 평가를 도시한다. 도 48A와 관련하여, GTEx RNA-seq 데이터에 대해 유효성확인된 SpliceNet-10k 모델에 의해 예측된 크립틱 스플라이스 돌연변이의 분율. 트레이닝 동안 사용된 염색체(chr1, chr3, chr5, chr7 및 chr9를 제외한 모든 염색체; 왼쪽) 및 나머지 염색체(오른쪽)에 대한 유효성확인율은 별도로 표시된다. 예측은 Δ Score로 그룹화된다. 각 Δ Score 그룹에서 4가지 클래스의 변이체(수용체 또는 공여체의 이득 또는 손실) 각각에 대한 트레이닝 및 테스트 염색체 사이의 유효성확인율을 비교하였다. 이것은 트레이닝 및 테스트 염색체 사이의 예측된 Δ Score 값의 분포의 잠재적인 차이를 설명한다. 유효성확인율의 차이는 크지 않다($P > 0.05$, 16개 테스트에 대해 Bonferroni 보정을 사용한 Fisher Exact 테스트).
- [0757] 도 48B와 관련하여, 상이한 Δ Score 컷오프에서 GTEx 코호트에서 스플라이스 변경 변이체를 검출할 때 SpliceNet-10k의 민감도. 모델 민감도는 트레이닝에 사용된 염색체(왼쪽)와 나머지 염색체(오른쪽)의 변이체에

대해 별도로 표시된다. 본 발명자들은 Fisher Exact 테스트를 사용하여 트레이닝 및 테스트 염색체 사이의 Δ Score 컷오프 0.2에서 모델의 민감도를 비교했다. 엑손 근처의 변이체 또는 심층적 인트로닉 변이체에서는 차이가 유의미하지 않다(두 테스트에 대한 Bonferroni 보정 후 $P > 0.05$).

- [0758] 도 48C와 관련하여, ExAC 데이터 세트에서 유해한 예측된 동의 및 인트로닉 크립틱 스플라이스 변이체의 비율로, 트레이닝에 사용되는 염색체의 변이체(왼쪽)와 나머지 염색체(오른쪽)에 대해 별도로 계산된다. 비율과 P-값은 도 4a와 같이 계산된다. 각 Δ Score 그룹에서 4가지 종류의 변이체(수용자 또는 공여체의 이득 또는 손실) 각각에 대한 트레이닝 및 테스트 염색체 사이의 공통 및 희귀 변이체의 수를 비교했다. 차이는 크지 않다($P > 0.05$, 12개 테스트에 대해 Bonferroni 보정을 사용한 Fisher Exact 테스트).
- [0759] 도 48d에 대하여 DDD, ASD 및 대조군 코호트에 대해 1인당 예측된 크립틱 스플라이스 드 노보 돌연변이(DNM)로서, 트레이닝(왼쪽)에 사용된 염색체의 변이체 및 나머지 염색체(오른쪽)에 대해 별도로 표시됨. 오차 막대에는 95% 신뢰구간(CI)이 표시된다. 1인당 크립틱 스플라이스 드 노보 변이체의 수는 테스트 세트가 트레이닝 세트 크기의 약 절반이기 때문에 테스트 세트에서 더 적다. 샘플 크기가 작기 때문에 숫자가 잡음처럼 보인다.
- [0760] 도 49A, 도 49B 및 도 49C는 동의, 인트로닉 또는 번역되지 않은 영역 부위에서만 희귀 유전질환을 가진 환자에서 드 노보 크립틱 스플라이스 돌연변이를 나타낸다. 도 49A와 관련하여, Deciphering Developmental Disorders 코호트(DDD)의 환자에 대해 1명당 0.2 초과의 크립틱 스플라이스 Δ Score를 갖는 예측된 크립틱 스플라이스 드 노보 돌연변이(DNM), Simons Simplex Collection 및 Autism Sequencing Consortium의 자폐 스펙트럼 장애(ASD) 환자, 그리고 건강한 대조군. 건강한 대조군 위의 DDD 및 ASD 코호트에서의 농축이 보여지고, 코호트 사이의 변이체 확정을 조정한다. 오차 막대는 95% 신뢰 구간을 나타낸다.
- [0761] 도 49B와 관련하여, 건강한 대조군과 비교하여 각 카테고리의 농축에 기초한, DDD 및 ASD 코호트에 대한 기능적 카테고리에 의한 병원성 DNM의 추정된 비율. 크립틱 스플라이스 비율은 미스센스의 부족과 더 심층적인 인트로닉 부위에 맞게 조정된다.
- [0762] 도 49C와 관련하여, 상이한 Δ Score 임계치에서의 건강한 대조군과 비교한 DDD 및 ASD 코호트에서의 크립틱 스플라이스 DNM의 농축 및 과잉. 크립틱 스플라이스 과잉은 미스센스의 부족과 더 심층적인 인트로닉 부위에 맞게 조정된다.
- [0763] 도 50A 및 도 50B는 ASD 및 병원성 DNM의 비율로서 크립틱 스플라이스 드 노보 돌연변이를 나타낸다. 도 50A와 관련하여, 크립틱 스플라이스 부위를 예측하기 위해 상이한 Δ Score 임계치에서 ASD 내에서의 크립틱 스플라이스 DNM의 농축 및 과잉.
- [0764] 도 50B와 관련하여, 크립틱 스플라이스 부위를 예측하기 위해 상이한 Δ Score 임계값을 사용하여, 모든 클래스의 병원성 DNM(단백질 코딩 돌연변이 포함)의 일부로서 크립틱 스플라이스 부위에 기인한 병원성 DNM의 비율. 더 허용적인 Δ Score 임계값은 승산비가 낮아지는 타협으로 백그라운드 예상 이상으로 식별된 크립틱 스플라이스 부위의 수를 증가시킨다.
- [0765] 도 51a, 도 51b, 도 51c, 도 51d, 도 51e, 도 51f, 도 51g, 도 51h, 도 51i 및 도 51j는 ASD 환자에서 예측된 크립틱 스플라이스 드 노보 돌연변이의 RNA-seq 유효성확인을 도시한다. RNA-seq에 의한 실험적 유효성확인을 위해 선택된 36개의 예측된 크립틱 스플라이스 부위로부터의 RNA 발현의 커버리지 및 스플라이스 접합부 계수치. 각각의 샘플에 대해, 영향을 받은 개체에 대한 RNA-seq 커버리지 및 접합부 계수치가 상단에 표시되고, 돌연변이가 없는 대조군 개체가 하단에 표시된다. 플롯은 유효성확인 상태 및 스플라이스 이상 유형별로 그룹화된다.
- [0766] 도 52A 및 도 52B는 표준 전사체에 대해서만 트레이닝된 모델의 RNA-seq에 대한 유효성확인을 및 민감도를 도시한다. 도 52A와 관련하여, 본 발명자들은 표준 GENCODE 전사체의 접합부만 사용하여 SpliceNet-10k 모델을 트레이닝하고 이 모델의 성능을 GTEx 코호트에서 5명 이상이 나타나는 표준 접합부와 스플라이스 접합부 모두에서 트레이닝된 모델과 비교했다. 각 Δ Score 그룹에서 4가지 클래스의 변이체(수용자 또는 공여체의 이득 또는 손실) 각각에 대해 두 모델의 검증 속도를 비교했다. 두 모델 간의 유효성확인을 차이는 크지 않다($P > 0.05$, 16 회 테스트에 대해 Bonferroni 보정을 사용한 Fisher Exact 테스트).
- [0767] 도 52B와 관련하여, 상이한 Δ Score 컷오프에서 GTEx 코호트에서 스플라이스-변경 변이체를 검출할 때 표준 접합부에 대해 트레이닝된 모델의 민감도. 심층적 인트로닉 영역에서 이 모델의 민감도는 도 2의 모델의 민감도보다 낮다($P < 0.001$, Bonferroni 보정을 사용한 Fisher Exact 테스트). 엑손 근처의 민감도는 크게 다르지

않다.

- [0768] 도 53A, 도 53B 및 도 53C는 양상블 모델링이 SpliceNet-10k 성능을 향상시키는 것을 도시한다. 도 53A와 관련하여, 5개의 개별 SpliceNet-10k 모델의 top-k 정확도 및 정밀-재호출 곡선하 면적이 도시되어 있다. 모델은 동일한 아키텍처를 가지며 동일한 데이터 세트를 사용하여 학습되었다. 그러나 파라미터 초기화, 데이터 서플링 등과 같은 트레이닝 과정과 관련된 다양한 임의의 측면으로 인해 서로 다르다.
- [0769] 도 53B와 관련하여, 5개의 개별 SpliceNet-10k 모델의 예측들은 높은 상관관계가 있다. 이 연구에서는, 적어도 하나의 모델에 의해 수용자 또는 공여체 점수가 0.01 이상으로 지정된 테스트 세트의 위치만을 고려했다. 서브플롯(i, j)은 모델 #j의 예측에 대해 모델 #i의 예측을 플로팅하여 구성된다(해당 피어슨 상관관계가 서브플롯 위에 표시됨).
- [0770] 도 53C와 관련하여, SpliceNet-10k 양상블을 구성하는데 사용된 모델의 수가 1에서 5로 증가함에 따라 성능이 향상된다.
- [0771] 도 54A 및 도 54B는 다양한 엑손 밀도의 영역에서 SpliceNet-10k의 평가를 보여준다. 도 54A와 관련하여, 테스트 세트 위치는 10,000개 뉴클레오타이드 윈도우에 존재하는 표준 엑손의 수에 따라 5개의 빈으로 범주화되었다. 각 빈에 대해 SpliceNet-10k에 대한 top-k 정확도와 정밀-재호출 곡선하 면적을 계산했다.
- [0772] 도 54B와 관련하여, 본 발명자들은 비교로서 MaxEntScan으로 분석을 반복했다. 양의 테스트 사례의 수가 음의 테스트 사례의 수에 비해 증가하기 때문에 top-k 정확도 및 Precision-Recall AUC로 측정된 두 모델의 성능은 더 높은 엑손 밀도에서 향상된다.
- [0773] **코호트당 드 노보 돌연변이의 농축**
- [0774] 후보 크립틱 스플라이스 DNM을 3개의 코호트 각각에서 계수치하였다. DDD 코호트는 엑손으로부터 8 nt 초과하여 떨어진 인트로닉 DNM을 보고하지 않았기 때문에, DDD와 ASD 코호트 사이의 동등한 비교를 가능하게 하기 위해 농축 분석의 목적으로 엑손으로부터 8 nt 초과인 영역이 모든 코호트에서 제외되었다(도 41A). 또한, 이중 크립틱 스플라이싱 및 단백질-코딩 기능 결과를 갖는 돌연변이를 배제하여 영향을 받는 코호트 내에서 단백질-코딩 효과를 갖는 돌연변이가 풍부한 것이 농축의 원인이 아님을 입증하는 별도의 분석을 수행하였다(도 49A, 도 49B 및 도 49C). 건강한 대조군 코호트를 베이스라인으로 사용하여 코호트 사이의 개체 당 동의 DNM의 비율을 정규화함으로써 코호트 사이의 DNM의 상이한 확정에 대해 계수치를 조정하였다. 본 발명자들은 두 개의 포아송 비율을 비교하기 위해 E-테스트를 사용하여 코호트 당 크립틱 스플라이스 DNM의 비율을 비교했다(Krishnamoorthy and Thomson, 2004).
- [0775] 기대 이상의 농축에 대한 플로팅된 비율(도 41C)은 엑손으로부터 8 nt 초과인 DNM의 결여에 대해 엑손으로부터 9 내지 50 nt 사이에서 발생할 것으로 예상되는 모든 크립틱 스플라이스 DNM의 비율로 트라이뉴클레오타이드 서열 컨텍스트 모델을 사용하여 상향 스케일링함으로써 조정되었다(하기, 유전자당 드 노보 돌연변이의 농축 참조). 자동 전용 진단 비율 및 과잉 크립틱 부위(도 49B 및 도 49C)는 또한 미스센스 부위 대 동의 부위에서 발생할 것으로 예상되는 크립틱 스플라이스 부위의 비율에 따라 크립틱 계수치를 스케일링함으로써 미스센스 부위 부족에 대해 조정되었다. 농축에 대한 ΔScore 임계값의 영향은 컷오프 범위에 걸쳐 DDD 코호트 내에서 크립틱 스플라이스 DNM의 농축을 계산함으로써 평가되었다. 관찰된 이들 각각에 대해, 과잉 크립틱 스플라이스 DNM과 함께 예상 승산비가 계산되었다.
- [0776] **병원성 DNM의 비율**
- [0777] 베이스라인 돌연변이율과 비교하여 과량의 DNM은 코호트 내에서 병원성 수율로 간주될 수 있다. 건강한 대조군 코호트의 배경에 대하여, ASD 및 DDD 코호트 내의 기능적 유형에 의한 DNM의 초과를 추정하였다(도 41B). DNM 계수치는 상기한 바와 같이 개체 당 동의 DNM의 비율로 정규화되었다. DDD 크립틱 스플라이스 계수치는 전술한 바와 같이 인트로닉으로부터 9 내지 50 nt 떨어진 DNM의 부족에 대해 조정되었다. ASD 및 DDD 코호트 둘 다에 대해, 본 발명자들은 엑손으로부터 50 nt 초과하여 떨어진 침묵적 인트로닉 변이체의 확정 결여에 대해, 음성 선택 분석으로부터 니어 인트로닉(<50 nt) 대 침묵적 인트로닉(> 50 nt) 크립틱 스플라이스 변이체의 비율을 사용하여 조정했다(도 38G).
- [0778] **유전자당 드 노보 돌연변이의 농축**
- [0779] 본 발명자들은 트라이뉴클레오타이드 서열 컨텍스트 모델을 사용하여 계놈의 모든 변이체에 대해 널 돌연변이율을 결정하였다(Samocha et al., 2014). 엑손 내에서 및 인트로닉에 최대 8 nt에서 가능한 모든 단일 뉴클레오타이드

드 치환에 대한 Δ Score를 예측하기 위해 망을 사용하였다. 널 돌연변이율 모델을 기반으로, 본 발명자들은 유전자당 예상된 드 노보 크립틱 스플라이스 돌연변이 수를 얻었다(컷오프로서 Δ Score > 0.2 사용).

[0780] DDD 연구(McRae et al., 2017)에 따르면, 유전자는 단백질 절단(PTV) DNM만을 고려하는 모델과 모든 단백질 변경 DNM(PTV, 미스센스 및 프레임내 인델)을 고려하는 두 가지 모델에서 확률과 비교하여 DNM의 농축에 대해 평가되었다. 각 유전자에 대해 가장 중요한 모델을 선택하고 다중 가설 검정을 위해 P-값을 조정했다. 이 테스트는 우리가 크립틱 스플라이스 DNM 또는 크립틱 스플라이스 비율(원래 DDD 연구에서 사용된 기본 테스트)을 고려하지 않은 곳에서 한 번 실행되었으며, 본 발명자들은 크립틱 스플라이스 DNM과 그 돌연변이율을 또한 계수치한 곳에서 실행되었다. 본 발명자들은 크립틱 스플라이스 DNM을 포함할 때 FDR 조정 P-값 < 0.01 인 유전자로 식별된 추가 후보 유전자를 보고하지만 크립틱 스플라이스 DNM을 포함하지 않을 때에는 FDR 조정 P-값 > 0.01 (기본 테스트)인 유전자로 식별된 추가 후보 유전자를 보고한다. ASD 코호트에 대해서도 유사하게 농축 테스트를 수행하였다.

[0781] **예측된 크립틱 스플라이스 부위의 유효성확인**

[0782] 본 발명자들은 림프 모세포 세포주에서 적어도 RPKM > 1 RNA-seq 발현을 갖는 Simons Simplex Collection의 영향을 받는 프로벤드로부터 높은 신뢰도를 갖는 드 노보 모델을 선택하였다. 스플라이스 손실 변이체의 경우 Δ Score 임계값 > 0.1을, 그리고 스플라이스 이득 변이체의 경우 Δ Score 임계값 > 0.5를 기반으로 유효성확인을 위해 드 노보 크립틱 스플라이스 변이체를 선택했다. 세포주를 미리 조달해야했기 때문에, 이 임계값은 논문의 다른 곳에서 채택한 임계값(도 38G 및 도 41A, 도 41B, 도 41C 및 도 41D)과 비교하여 방법의 초기 반복을 반영한다. 망에는 모델 트레이닝을 위한 GTEx 신규 스플라이스 접합부가 포함되지 않았다.

[0783] 이들 프로벤드에 대해 SSC로부터 림프 모세포 세포주를 수득하였다. 세포를 배양 배지(RPMI 1640, 2mM L-글루타민, 15% 소 태아 혈청)에서 최대 세포 밀도 1×10^6 개 세포/ml로 배양하였다. 세포가 최대 밀도에 도달했을 때, 세포를 4배 또는 5배 피펫팅하고 200,000 내지 500,000 생존 세포/ml의 밀도로 과중함으로써 세포를 분리함으로써 계대시켰다. 세포를 37°C, 5% CO₂ 조건 하에서 10일 동안 성장시켰다. 대략 5×10^5 개의 세포를 분리하고 4°C에서 5분 동안 300 × g에서 회전시켰다. 제조사의 프로토콜에 따라 RNeasy® Plus 마이크로 키트(QIAGEN)를 사용하여 RNA를 추출하였다. RNA 품질은 Agilent RNA 6000 Nano Kit(애질런트 테크놀로지스사(Agilent Technologies))를 사용하여 평가되었고 Bioanalyzer 2100(애질런트 테크놀로지스사)에서 실행되었다. RNA-seq 라이브러리는 Ribo-Zero Gold Set A(일루미나사)가 있는 TruSeq® Stranded Total RNA Library Prep Kit에 의해 생성되었다. 라이브러리는 270-388 백만 리드(중양값 358 백만 리드)의 커버리지에서 150-nt 단일 리드 서열분석을 사용하여 UCSF(Center for Advanced Technology)의 HiSeq 4000 기기에서 서열분석되었다.

[0784] 각 환자에 대한 서열분석 리드는 상응하는 대체 대립유전자로 환자의 드 노보 변이체(Iossifov et al., 2014)를 대체함으로써 hg19로부터 생성된 참조에 대해 OLego(Wu et al., 2013)와 정렬되었다. 서열분석 커버리지, 스플라이스 접합부 사용 및 전사 위치는 MISO의 사시미 플롯으로 플로팅되었다(Katz et al., 2010). 모델 예측 부분의 유효성확인에서 위에서 설명한 대로 예측된 크립틱 스플라이스 부위를 평가했다. 13개의 신규 스플라이스 부위(9개의 신규 접합부, 4개의 엑손 스킵핑)는 크립틱 스플라이스 부위를 함유하는 샘플에서만 관찰되었고 149개의 GTEx 샘플 중 임의의 것 또는 다른 35개의 서열분석된 샘플에서는 관찰되지 않았기 때문에 확인되었다. 4개의 추가 엑손 스킵핑 이벤트의 경우, GTEx에서 종종 낮은 수준의 엑손 스킵핑이 관찰되었다. 이 경우에, 본 발명자들은 스킵핑 접합부를 사용한 리드의 비율을 계산하고 이 비율이 다른 샘플과 비교하여 샘플을 포함하는 크립틱 스플라이스 부위에서 가장 높음을 확인했다. 다른 샘플에서는 없었거나 현저히 낮은 두드러진 인트론 보유에 기초하여 4가지 추가 사례가 유효성확인되었다. 대조군 샘플에서 적당한 인트론 보유는 DDX11 및 WDR4에서 이벤트를 해결하지 못하게 하였다. 변이체가 서열분석 리드에 존재하지 않았기 때문에 두 이벤트(CSAD 및 GSA P)가 유효성확인 실패로 분류되었다.

[0785] **데이터 및 소프트웨어 입수가능성**

[0786] 트레이닝 및 테스트 데이터, 참조 게놈에서의 모든 단일 뉴클레오타이드 치환에 대한 예측 점수, RNA-seq 유효성확인 결과, RNA-seq 접합부 및 소스 코드는 공개적으로 다음에서 호스팅된다:

[0787] <https://basespace.illumina.com/s/5u6Th0blecrh>

[0788] 36개의 림프 모세포 세포주에 대한 RNA-seq 데이터는 EMBL-EBI (www.ebi.ac.uk/arrayexpress)의 ArrayExpress 데이터베이스에 수탁 번호 E-MTAB-xxxx로 기탁되고 있다.

[0789] 예측 점수 및 소스 코드는 공개 소스 수정된 Apache License v2.0 하에서 공개적으로 공개되며 학술 및 비상업적 소프트웨어 응용 프로그램에 무료로 사용할 수 있다. 해당 분야에서 우려되는 원 형성 문제를 줄이기 위해, 저자는 이 방법의 예측 점수를 다른 분류자의 구성 요소로 통합하지 말고 이해 당사자가 제공된 소스 코드와 데이터를 사용하여 직접 자신의 심층 학습 모델을 트레이닝하고 향상시키도록 요청한다.

[0790] **핵심 자원 표**

피험자 또는 공급원	출처	식별자
Deposited Data		
RNA-seq data and variant calls for the GTEx cohort		dbGAP accession:
De-novo mutations for autism patients and healthy controls	(Iossifov et al., 2014)	N/A
De-novo mutations from the Deciphering Developmental Disorders cohort	(McRae et al., 2017)	N/A
Splice junctions from GENCODE principal transcripts used to train the canonical SpliceNet model	본 연구	https://basespace.illumina.com/s/
Splice junctions from GTEx used to augment the training dataset	본 연구	https://basespace.illumina.com/s/
Splice junctions from GENCODE principal transcripts used to test the model, with paralogs excluded	본 연구	https://basespace.illumina.com/s/
Splice junctions of lincRNAs used to test the model	본 연구	https://basespace.illumina.com/s/
Predictions of canonical model	본 연구	https://basespace.illumina.com/s/
Predictions of GTEx-supplemented model	본 연구	https://basespace.illumina.com/s/
All GTEx junctions in all GTEx v6.p1 samples	본 연구	https://basespace.illumina.com/s/
List of validated GTEx private variants with Δ Score > 0.1	본 연구	https://basespace.illumina.com/s/
Aligned BAM files for RNA-seq in 36 autism patients	본 연구	ArrayExpress accession: E-MTAB-xxxxx
소프트웨어 및 알고리즘		
SpliceNet source code	본 연구	https://basespace.illumina.com/s/

[0791]

[0792] **보충 표 제목**

[0793] 표 S1은 효과 크기 계산 및 조직-특이적 스플라이싱 효과를 입증하기 위해 사용된 GTEx 샘플을 보여준다. 도 38A, 38b, 38c, 38d, 38e, 38f 및 38g. 도 39A, 도 39B 및 도 45에 관련된다.

[0794] 표 S2는 모든 알고리즘이 게놈 전체에서 동일한 수의 이득 및 손실을 예측하는 SpliceNet-10k, GeneSplicer, MaxEntScan 및 NNSplice에 대한 일치된 신뢰도 컷오프를 보여준다. 도 38G와 관련된다.

[0795] 표 S3은 각 코호트에서 예측된 크립틱 스플라이스 DNMs의 수를 나타낸다. 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F와 관련되며, 다음과 같이 생성된다.

코호트	프로밴드 (n)	프로밴드 당 동일 드 노보	엑손 + 인트론 최대 8 nt		인트론 > 엑손에서 8 nt	
			비조정	동의로 정규화	비조정	동의로 정규화
DDD	4293	0.28744	347	298.7	14	12.1
ASD	3953	0.24462	236	238.7	64	64.7
대조군	2073	0.24747	98	98	20	20

[0796]

[0797]

[0798]

[0799]

[0800]

[0801]

[0802]

[0803]

[0804]

[0805]

[0806]

[0807]

[0808]

표 S4는 각각의 돌연변이 카테고리에 대한 유전자당 예상되는 드 노보 돌연변이율 보여준다. 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F와 관련된다.

표 S5는 DDD 및 ASD에서 유전자 농축에 대한 p-값을 예시한다. 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F와 관련된다.

표 S6은 자폐증 환자에서 36개의 예측된 크립틱 스플라이스 DNM에 대한 유효성확인 결과를 도시한다. 도 41A, 도 41B, 도 41C, 도 41D, 도 41E 및 도 41F와 관련된다.

컴퓨터 시스템

도 59는 개시된 기술을 구현하는 데 사용될 수 있는 컴퓨터 시스템의 단순화된 블록도이다. 컴퓨터 시스템은 통상적으로 버스 서브시스템을 통해 다수의 주변 장치와 통신하는 적어도 하나의 프로세서를 포함한다. 이러한 주변 장치는, 예를 들어, 메모리 장치와 파일 저장 서브시스템을 포함하는 저장 서브시스템, 사용자 인터페이스 입력 장치, 사용자 인터페이스 출력 장치, 및 네트워크 인터페이스 서브시스템을 포함할 수 있다. 입력 및 출력 장치는 컴퓨터 시스템과의 사용자 상호작용을 허용한다. 네트워크 인터페이스 서브시스템은, 다른 컴퓨터 시스템의 해당 인터페이스 장치에 대한 인터페이스를 포함하여 외부 네트워크에 대한 인터페이스를 제공한다.

일 구현예에서, ACNN 및 CNN과 같은 신경망들은 저장 서브시스템 및 사용자 인터페이스 입력 장치에 통신가능하게 연결된다.

사용자 인터페이스 입력 장치는, 키보드; 마우스, 트랙볼, 터치패드 또는 그래픽 태블릿과 같은 포인팅 장치; 스캐너; 디스플레이에 통합된 터치 스크린; 음성 인식 시스템 및 마이크와 같은 오디오 입력 장치; 및 다른 유형의 입력 장치를 포함할 수 있다. 일반적으로, "입력 장치"라는 용어의 사용은, 정보를 컴퓨터 시스템에 입력하도록 모든 가능한 유형의 장치 및 방법을 포함하고자 하는 것이다.

사용자 인터페이스 출력 장치는, 디스플레이 서브시스템, 프린터, 팩스기, 또는 오디오 출력 장치와 같은 비시각적 디스플레이를 포함할 수 있다. 디스플레이 서브시스템은, 음극선관(CRT), 액정 디스플레이(LCD)와 같은 평판 장치, 투영 장치, 또는 시각적 이미지를 생성하기 위한 다른 일부 메커니즘을 포함할 수 있다. 디스플레이 서브시스템은, 또한, 오디오 출력 장치와 같은 비시각적 디스플레이를 제공할 수 있다. 일반적으로, "출력 장치"라는 용어의 사용은, 컴퓨터 시스템으로부터 사용자 또는 다른 기계 또는 컴퓨터 시스템으로 정보를 출력하기 위한 모든 가능한 유형의 장치 및 방법을 포함하고자 하는 것이다.

저장 서브시스템은, 본 명세서에서 설명된 모듈과 방법 중 일부 또는 전부의 기능을 제공하는 프로그래밍 및 데이터 구성을 저장한다. 이러한 소프트웨어 모듈은 일반적으로 프로세서 단독으로 또는 다른 프로세서와 함께 실행된다.

저장 서브시스템에 사용되는 메모리는, 프로그램 실행 동안 명령어와 데이터를 저장하기 위한 메인 랜덤 액세스 메모리(RAM) 및 고정 명령어가 저장된 판독 전용 메모리(ROM)를 포함하는 다수의 메모리를 포함할 수 있다. 파일 저장 서브시스템은, 프로그램 및 데이터 파일을 위한 영구 저장소를 제공할 수 있으며, 하드 디스크 드라이브, 연관된 탈착가능 매체를 갖는 플로피 디스크 드라이브, CD-ROM 드라이브, 광 드라이브, 또는 탈착가능 매체 카트리지를 포함할 수 있다. 소정의 구현예의 기능을 구현하는 모듈들은, 파일 저장 서브시스템에 의해 저장 서브시스템에 또는 프로세서가 액세스할 수 있는 다른 기계에 저장될 수 있다.

버스 서브시스템은, 컴퓨터 시스템의 다양한 구성요소와 서브시스템들이 서로 의도된 바와 같이 통신하게 하는 메커니즘을 제공한다. 버스 서브시스템이 단일 버스로 개략적으로 표시되어 있지만, 버스 서브시스템의 대체 구현예에서는 다수의 버스를 사용할 수 있다.

컴퓨터 시스템 자체는, 개인용 컴퓨터, 휴대용 컴퓨터, 워크스테이션, 컴퓨터 단말, 네트워크 컴퓨터, 텔레비전, 메인프레임, 서버 팜, 느슨하게 네트워크화된 컴퓨터들의 광범위하게 분산된 세트, 또는 다른 임의의

데이터 처리 시스템이나 사용자 장치를 포함하는 다양한 유형일 수 있다. 컴퓨터 및 네트워크의 특성이 계속 변화함으로써 인해, 도 59에 도시된 컴퓨터 시스템의 설명은, 개시된 기술을 예시하기 위한 목적의 특정한 일례를 의도한 것일 뿐이다. 도 59에 도시된 컴퓨터 시스템보다 많거나 적은 구성요소를 갖는 컴퓨터 시스템의 다른 많은 구성이 가능하다.

[0809] 심층 학습 프로세서는, GPU 또는 FPGA 일 수 있으며, 구글 클라우드 플랫폼, 자일링스, 및 시라스케일과 같은 심층 학습 클라우드 플랫폼에 의해 호스팅될 수 있다. 심층 학습 프로세서의 예로는, Google의 텐서 처리 유닛(TPU), GX4 Rackmount Series, GX8 Rackmount Series와 같은 랙마운트 솔루션, NVIDIA DGX-1, Microsoft의 Stratix V FPGA, Graphcore의 Intelligent Processor Unit(IPU), Qualcomm의 Zeroth platform with Snapdragon processors, NVIDIA의 Volta, NVIDIA의 DRIVE PX, NVIDIA의 JETSON TX1/TX2 MODULE, Intel의 Nirvana, Movidius VPU, Fujitsu DPI, ARM의 DynamicIQ, IBM TrueNorth 등이 있다.

[0810] 전술한 설명은 개시된 기술의 제조 및 이용을 가능하게 하기 위해 제시된다. 개시된 구현들에 대한 다양한 변경들이 명백할 것이며, 본 명세서에서 정의된 일반적인 원리들은 개시된 기술의 사상 및 범위를 벗어나지 않고 다른 구현들 및 활용에 적용될 수 있다. 따라서, 개시된 기술은 도시된 구현으로 제한되도록 의도된 것이 아니라, 본 명세서에 개시된 원리 및 특징과 일치하는 가장 넓은 범위에 따라야한다. 개시된 기술의 범위는 첨부된 청구 범위에 의해 정의된다.

[0811] **부록**

[0812] 이하는, 발명자들이 작성한 논문에 열거된 잠재적으로 관련된 참고문헌들의 목록이 포함되어 있다. 그 논문의 주제는, 본 출원이 우선권/이익을 주장하는 미국 가특허 출원에서 다루어진다. 이들 참고문헌은 요청 시 대리인에 의해 제공될 수 있거나 글로벌 도시에를 통해 액세스될 수 있다. 논문은 가장 먼저 언급된 참고문헌이다.

1. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Rep.* *1*, 543–556.
2. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* *19*, 1732–1741.
3. Azad, A., Rajwa, B., and Pothén, A. (2016). flowVS: channel-specific variance stabilization in flow cytometry. *BMC Bioinformatics* *17*, 291.
4. Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. *Nature* *465*, 53–59.
5. Berget, S.M. (1995). Exon recognition in vertebrate splicing. *J. Biol. Chem.* *270*, 2411–2414.
6. Blencowe, B.J. (2006). Alternative Splicing: New Insights from Global Analyses. *Cell* *126*, 37–47.
7. Boerkoel, C.F., Exelbert, R., Nicastrì, C., Nichols, R.C., Miller, F.W., Plotz, P.H., and Raben, N. (1995). Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II. *Am. J. Hum. Genet.* *56*, 887–897.
8. Boyd, S., Cortes, C., Mohri, M., and Radovanovic, A. (2012). Accuracy at the Top. In *Advances in Neural Processing Systems*, pp. 953–961.
9. Close, P., East, P., Dirac-Svejstrup, A.B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Söding, J., Skehel, M., and Svejstrup, J.Q. (2012). DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. *Nature* *484*, 386–389.
10. Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and Disease. *Cell* *136*, 777–793.
11. Cramer, P., Pesce, C.G., Baralle, F.E., and Komblitt, a R. (1997). Functional association between promoter structure and transcript alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* *94*, 11456–11460.
12. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O’Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* *9*, eaal5209.
13. Dong, S., Walker, M.F., Carriero, N.J., DiCola, M., Willsey, A.J., Ye, A.Y., Waqar, Z., Gonzalez, L.E., Overton, J.D., Frahm, S., et al. (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* *9*, 16–23.
14. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* *473*, 43–49.
15. Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* (80-). *297*, 1007–1013.
16. Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shores, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* *518*, 337–343.

[0813]

17. Finkel, R.S., Mercuri, E., Darras, B.T., Connolly, A.M., Kuntz, N.L., Kirschner, J., Chiriboga, C.A., Saito, K., Servais, L., Tizzano, E., et al. (2017). Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *N. Engl. J. Med.* *377*, 1723–1732.
18. Fitzgerald, T.W., Gerety, S.S., Jones, W.D., van Kogelenberg, M., King, D.A., McRae, J., Morley, K.I., Parthiban, V., Al-Turki, S., Ambridge, K., et al. (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature* *519*, 223–228.
19. Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res.* *22*, 35–50.
20. Gelfman, S., Cohen, N., Yearim, A., and Ast, G. (2013). DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res.* *23*, 789–799.
21. Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proc. 13th Int. Conf. Artif. Intell. Stat.* *9*, 249–256.
22. Gouya, L., Puy, H., Robreau, A.-M., Bourgeois, M., Lamoril, J., Silva, V. Da, Grandchamp, B., and Deybach, J.-C. (2002). The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH. *Nat. Genet.* *30*, 27–28.
23. Graubert, T.A., Shen, D., Ding, L., Okeyo-Owuor, T., Lunn, C.L., Shao, J., Krysiak, K., Harris, C.C., Koboldt, D.C., Larson, D.E., et al. (2012). Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat. Genet.* *44*, 53–57.
24. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* *22*, 1760–1774.
25. He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
26. He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645.
27. Herring, C.A., Banerjee, A., McKinley, E.T., Simmons, A.J., Ping, J., Roland, J.T., Franklin, J.L., Liu, Q., Gerdes, M.J., Coffey, R.J., et al. (2018). Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Syst.* *6*, 37–51.e9.
28. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* *9*, 473–476.
29. Ioffe, S., and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proc. Mach. Learn. Res.* *37*, 448–456.
30. Iossifov, I., O’Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* *515*, 216–221.
31. Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O’Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* *159*, 1511–1523.
32. Jha, A., Gazzara, M.R., and Barash, Y. (2017). Integrative deep models for alternative splicing. 274–282.
33. Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* *3*, e02407.

[0814]

34. Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* *47*, 1242–1248.
35. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. *Science* (80-.). *342*.
36. Katz, Y., Wang, E.T., Airoidi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* *7*, 1009–1015.
37. Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. *Nat. Rev. Genet.* *11*, 345–355.
38. Kingma, D.P., and Ba, J.L. (2015). Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent.*
39. Kosmicki, J.A., Samocha, K.E., Howrigan, D.P., Sanders, S.J., Slowikowski, K., Lek, M., Karczewski, K.J., Cutler, D.J., Devlin, B., Roeder, K., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* *49*, 504–510.
40. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* *8*, 15824.
41. Krishnamoorthy, K., and Thomson, J. (2004). A more powerful test for comparing two Poisson means. *J. Stat. Plan. Inference* *119*, 23–35.
42. Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* *312*, 1880–1887.
43. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* *536*, 285–291.
44. Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* *25*, 1–13.
45. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* *50*, 151–158.
46. Licatalosi, D.D., and Darnell, R.B. (2006). Splicing Regulation in Neurologic Disease. *Neuron* *52*, 93–101.
47. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* *45*, 580–585.
48. Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science* *327*, 996–1000.
49. Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in Alternative Pre-mRNA Splicing. *Cell* *144*, 16–26.
50. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* (80-.). *337*, 1190–1195.
51. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* *17*, 122.
52. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S.,

[0815]

- Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature* *542*, 433–438.
53. Monroe, G.R., Frederix, G.W., Savelberg, S.M.C., de Vries, T.I., Duran, K.J., van der Smagt, J.J., Terhal, P.A., van Hasselt, P.M., Kroes, H.Y., Verhoeven-Duif, N.M., et al. (2016). Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet. Med.* *18*, 949–956.
54. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* *485*, 242–245.
55. Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. ArXiv:1609.03499 [Cs.SD].
56. Padgett, R.A. (2012). New connections between splicing and human disease. *Trends Genet.* *28*, 147–154.
57. Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* *29*, 1185–1190.
58. Reed, R., and Maniatis, T. (1988). The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev.* *2*, 1268–1276.
59. Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. *J. Comput. Biol.* *4*, 311–323.
60. Robinson, E.B., Samocha, K.E., Kosmicki, J.A., McGrath, L., Neale, B.M., Perlis, R.H., and Daly, M.J. (2014). Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc. Natl. Acad. Sci. U. S. A.* *111*, 15161–15165.
61. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* *515*, 209–215.
62. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo variation in human disease. *Nat. Genet.* *46*, 944–950.
63. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* *485*, 237–241.
64. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* *87*, 1215–1233.
65. Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardenosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin. Cancer Res.* *16*, 1957–1967.
66. Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* *16*, 990–995.
67. Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* *17*, 19–32.
68. SEQC/MAQC-III Consortium, S. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* *32*, 903–914.

[0816]

69. Shirai, C.L., Ley, J.N., White, B.S., Kim, S., Tibbitts, J., Shao, J., Ndonwi, M., Wadugu, B., Duncavage, E.J., Okeyo-Owuor, T., et al. (2015). Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell* 27, 631–643.
70. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proc. Mach. Learn. Res.* 70, 3145–3153.
71. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479, 74–79.
72. Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* 49, 848–855.
73. Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* 36, 245–254.
74. Stark, Z., Tan, T.Y., Chong, B., Brett, G.R., Yap, P., Walsh, M., Yeung, A., Peters, H., Mordaunt, D., Cowie, S., et al. (2016). A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet. Med.* 18, 1090–1096.
75. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156, 1324–1335.
76. Tan, T.Y., Dillon, O.J., Stark, Z., Schofield, D., Alam, K., Shrestha, R., Chong, B., Phelan, D., Brett, G.R., Creed, E., et al. (2017). Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. *JAMA Pediatr.* 171, 855–862.
77. Tennessen, J.A., Biggam, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* (80-.). 337, 64–69.
78. The GTEx Consortium, Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* (80-.). 348, 648–660.
79. Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* 16, 996–1001.
80. Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakraborty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625.
81. Trujillano, D., Bertoli-Avella, A.M., Kumar Kandaswamy, K., Weiss, M.E., Köster, J., Marais, A., Paknia, O., Schröder, R., Garcia-Aznar, J.M., Werber, M., et al. (2017). Clinical exome sequencing: Results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* 25, 176–182.
82. Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., et al. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* 98, 58–74.
83. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science* (80-.). 302, 1212–1215.
84. Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic

[0817]

modifications. *Genome Res.* 24, 896–905.

85. Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8, 749–761.

86. Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14, 802–813.

87. Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476.

88. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* 119, 831–845.

89. Wu, J., Anczukow, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLEgo: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* 41, 5149–5163.

90. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* (80-.), 347, 1254806.

91. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA* 312, 1870.

92. Yeo, G., and Burge, C.B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* 11, 377–394.

93. Yoshida, K., Sanada, M., Shiraiishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478, 64–69.

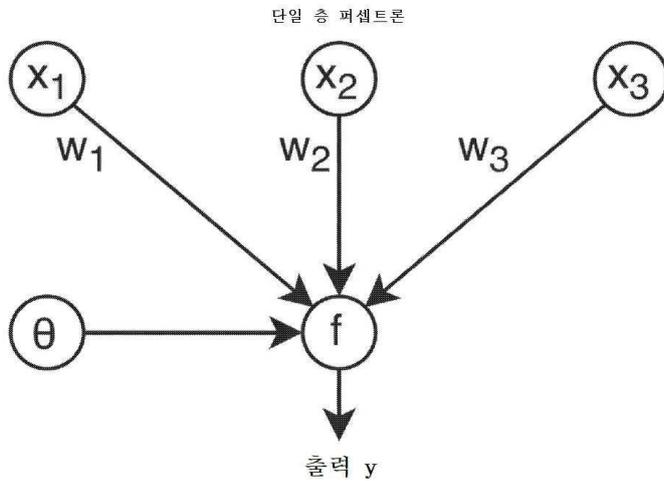
94. Yu, F., and Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. [ArXiv:1511.07122](https://arxiv.org/abs/1511.07122) [Cs.CV].

95. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931–934.

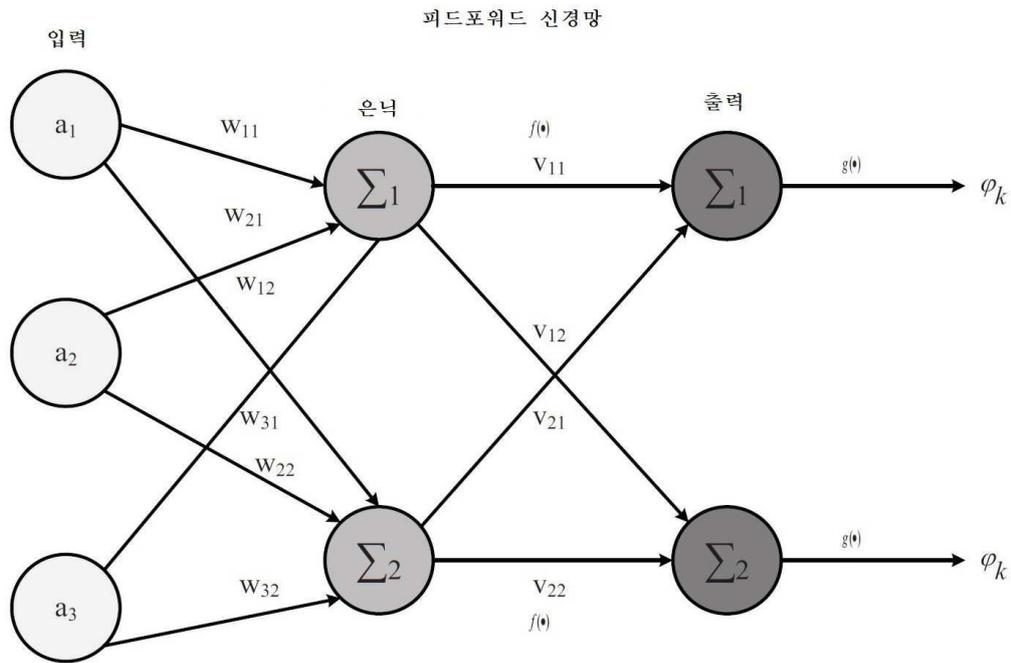
[0818]

도면

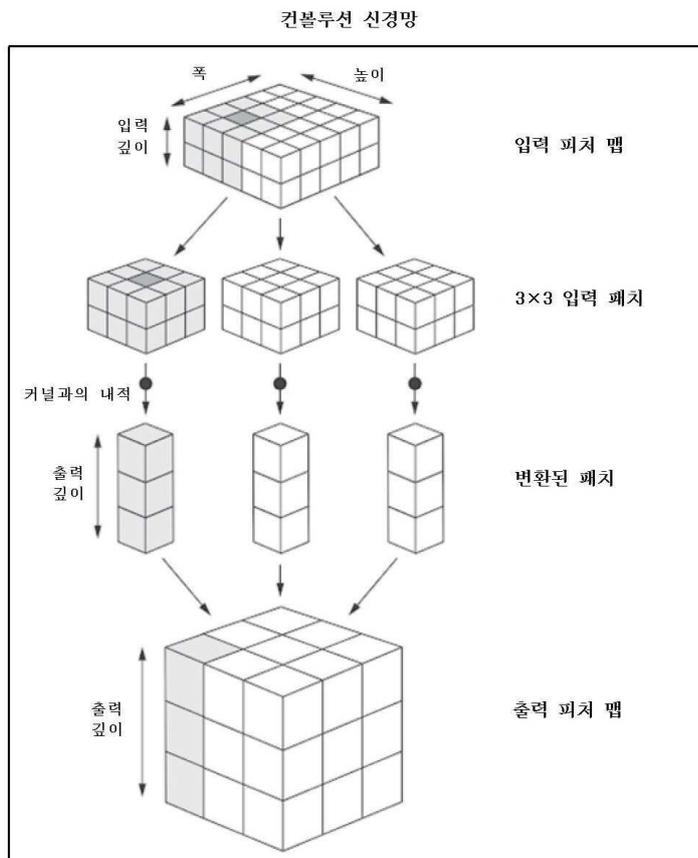
도면1



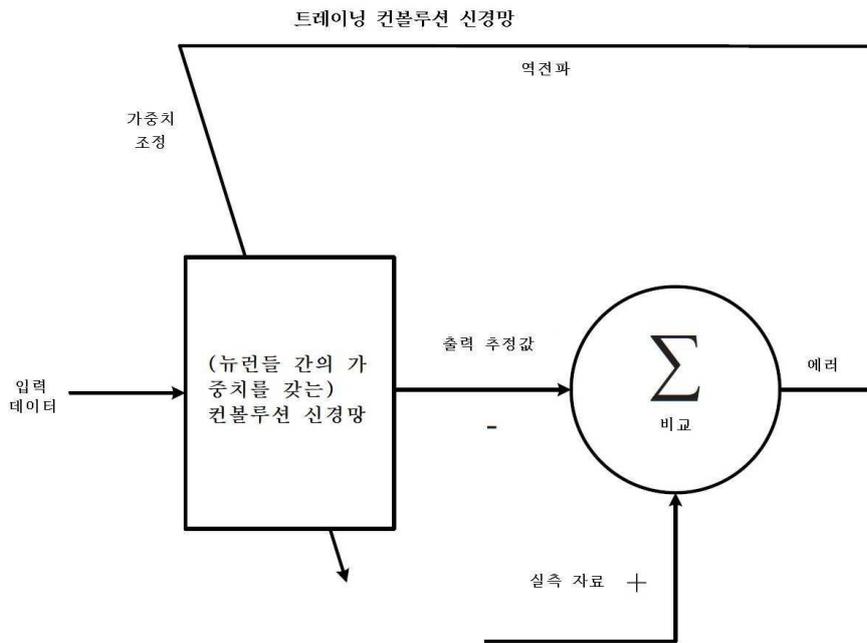
도면2



도면3

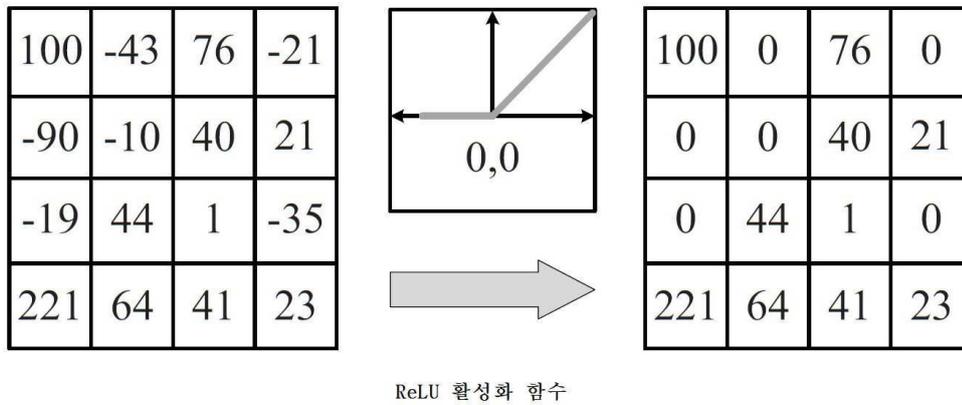


도면4

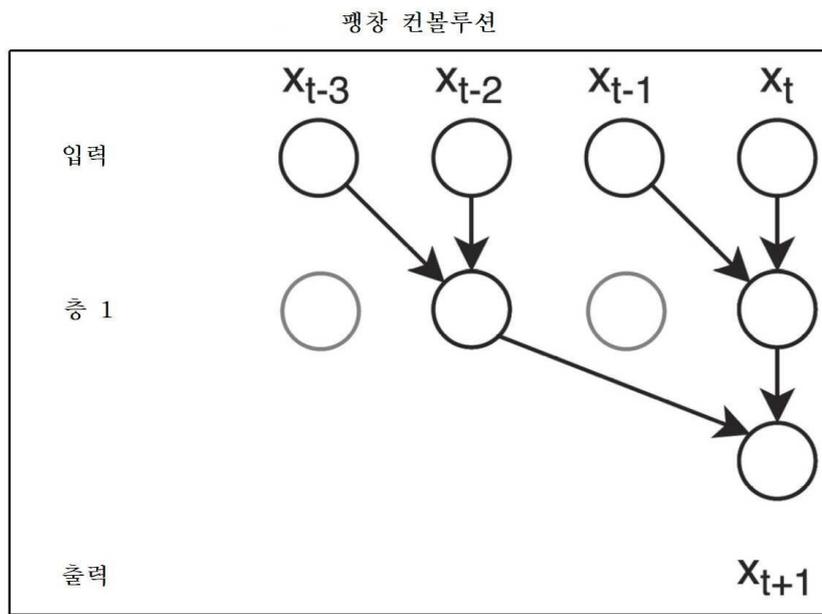


도면5

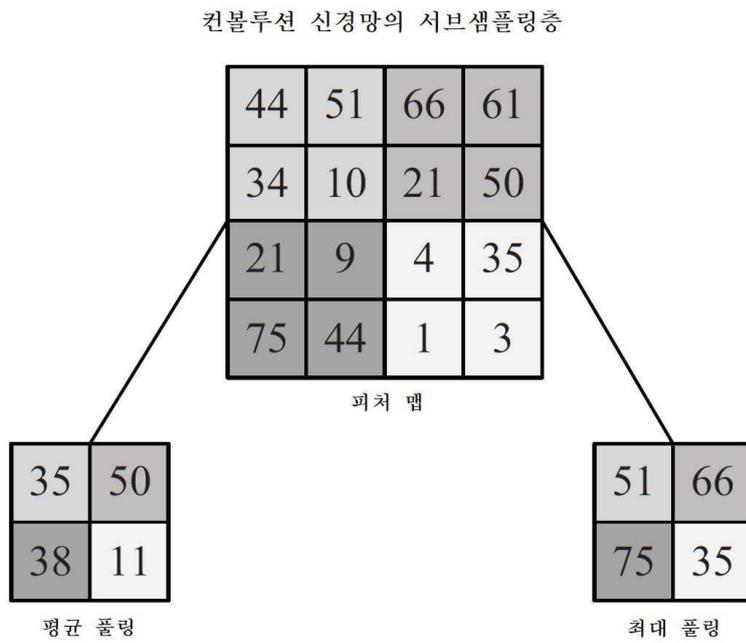
컨볼루션 신경망의 비선형층



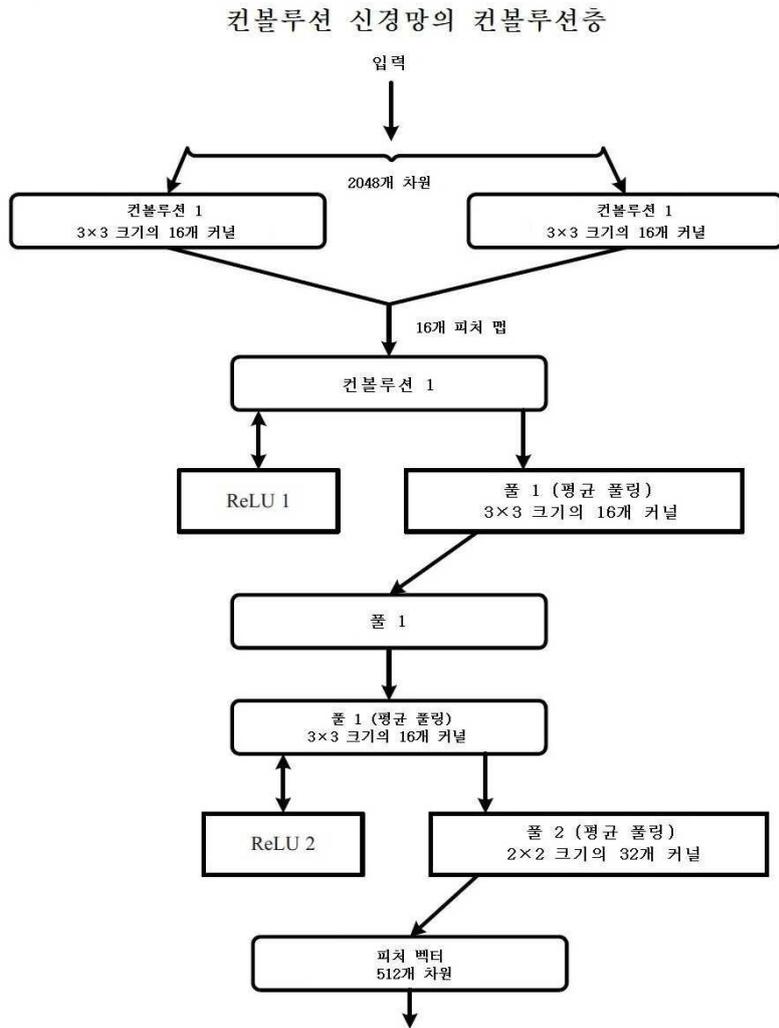
도면6



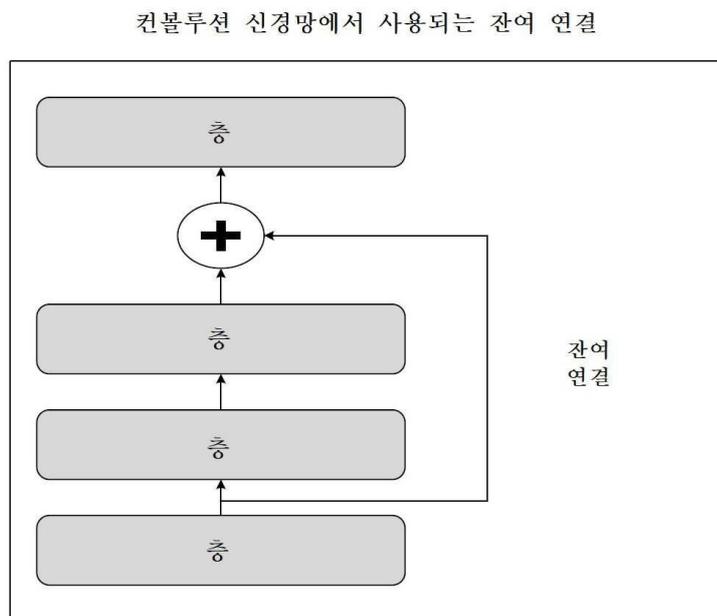
도면7



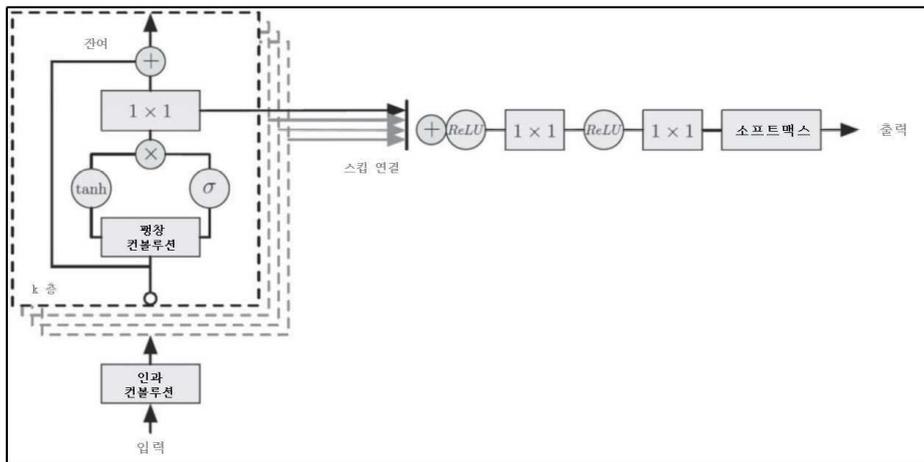
도면8



도면9

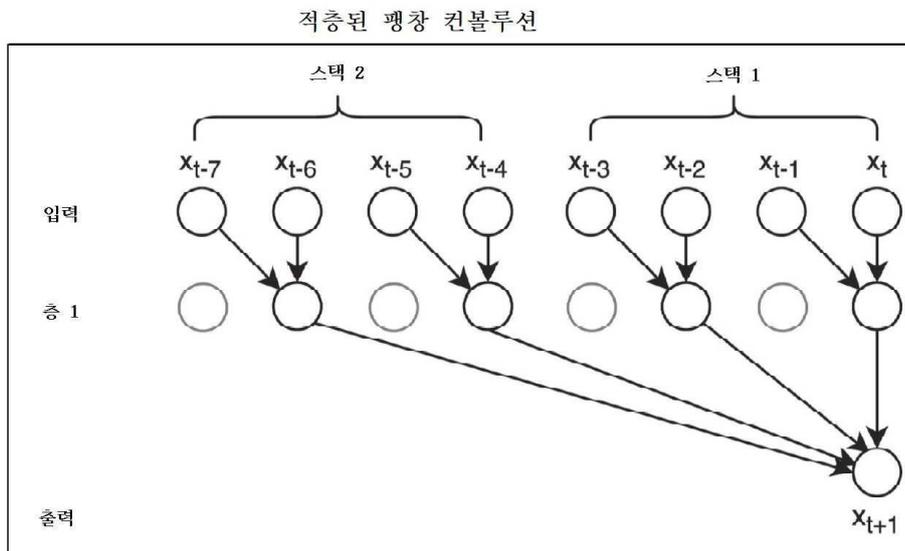


도면10



컨볼루션 신경망에서 사용되는 잔여 블록과 스킵 연결

도면11



도면12

컨볼루션 신경망을 이용한 일괄 정규화 순방향 패스

$$\begin{aligned} \mu_{\mathcal{B}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(\ell-1)} \\ \sigma_{\mathcal{B}}^2 &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^{(\ell-1)} - \mu_{\mathcal{B}})^2 \\ \hat{\mathbf{x}}^{(\ell-1)} &= \frac{\mathbf{x}^{(\ell-1)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \\ \mathbf{x}^{(\ell)} &= \gamma^{(\ell)} \hat{\mathbf{x}}^{(\ell-1)} + \beta^{(\ell)} \end{aligned}$$

도면13

일괄 정규화 - 컨볼루션 신경망과의 간섭

$$\begin{aligned} \hat{\mathbf{x}}^{(\ell-1)} &= \frac{\mathbf{x}^{(\ell-1)} - \mu_{\mathcal{D}}}{\sqrt{\sigma_{\mathcal{D}}^2 + \epsilon}} \\ \mathbf{x}_i^{(\ell)} &= \gamma^{(\ell)} \hat{\mathbf{x}}_i^{(\ell-1)} + \beta^{(\ell)} \end{aligned}$$

도면14

컨볼루션 신경망을 이용한 일괄 정규화 역방향 패스

$$\begin{aligned} \nabla_{\gamma^{(\ell)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{\mathbf{x}^{(\ell+1)}} \mathcal{L})_i \cdot \hat{\mathbf{x}}_i^{(\ell)} \\ \nabla_{\beta^{(\ell)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{\mathbf{x}^{(\ell+1)}} \mathcal{L})_i \end{aligned}$$

도면15

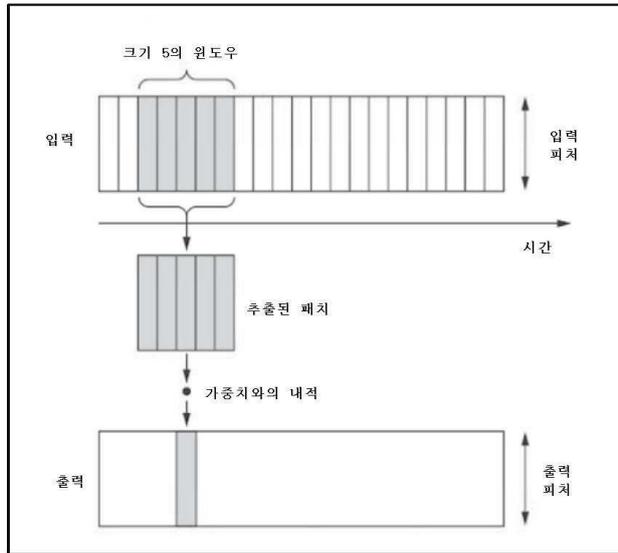
컨볼루션층의 일괄 정규화

```
conv_model.add(layers.Conv2D(32, 3, activation='relu')) ← 컨볼루션층 후
conv_model.add(layers.BatchNormalization())

dense_model.add(layers.Dense(32, activation='relu')) ← 조밀층 후
dense_model.add(layers.BatchNormalization())
```

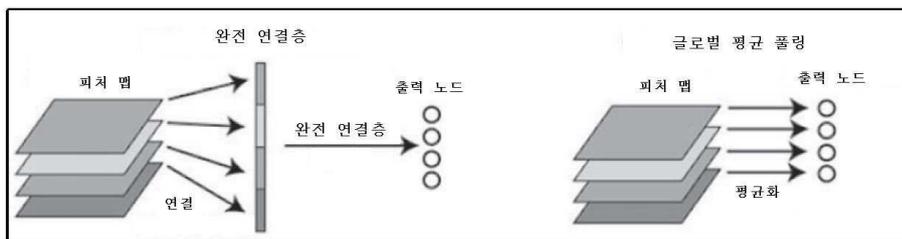
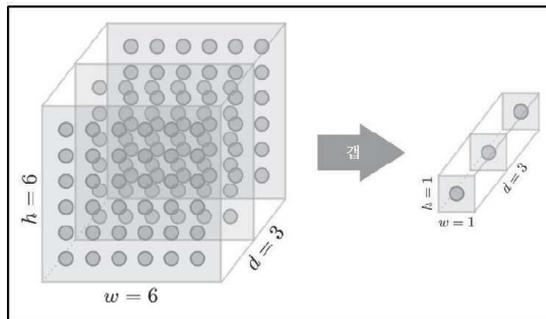
도면16

컨볼루션 신경망에서 사용되는 1D 컨볼루션

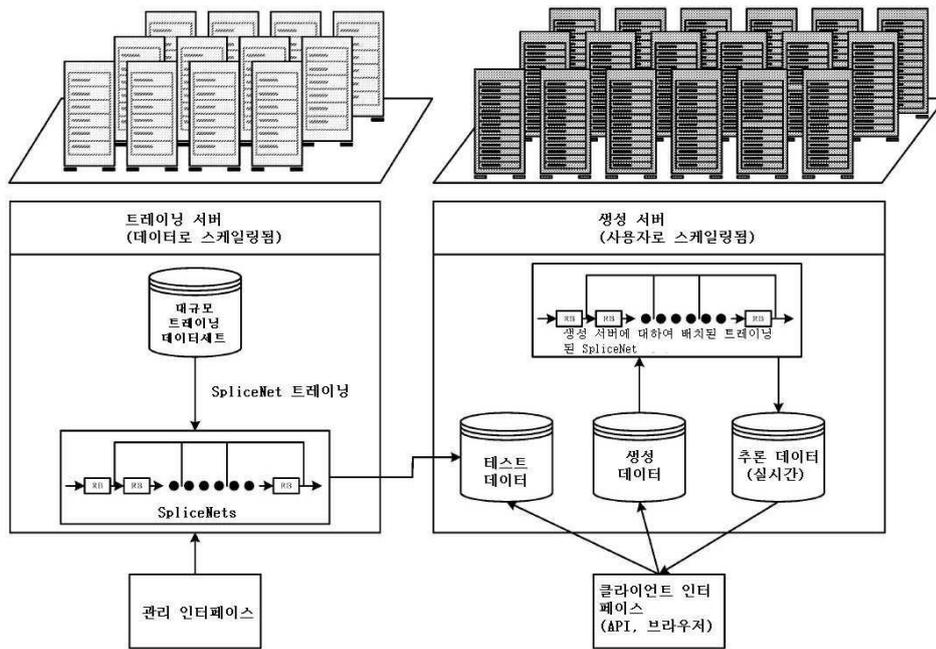


도면17

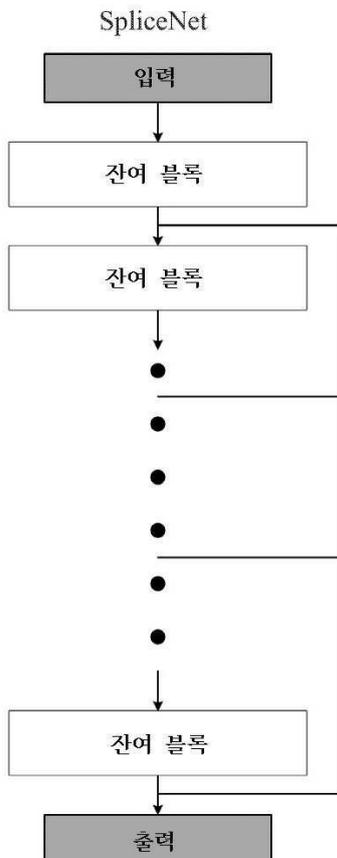
컨볼루션 신경망의 글로벌 평균 풀링(GAP)



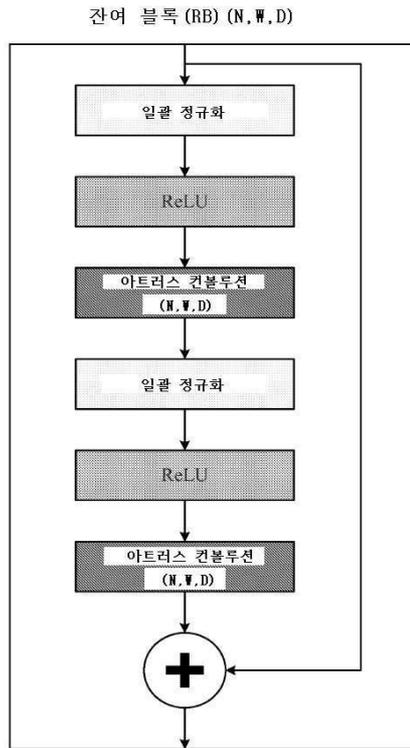
도면18



도면19

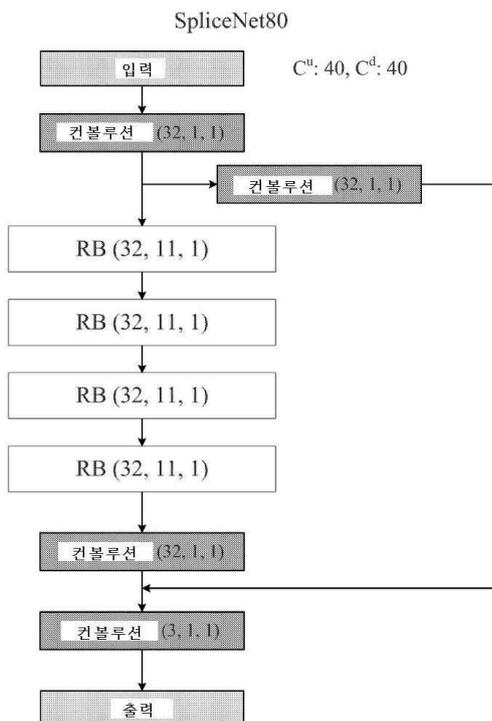


도면20

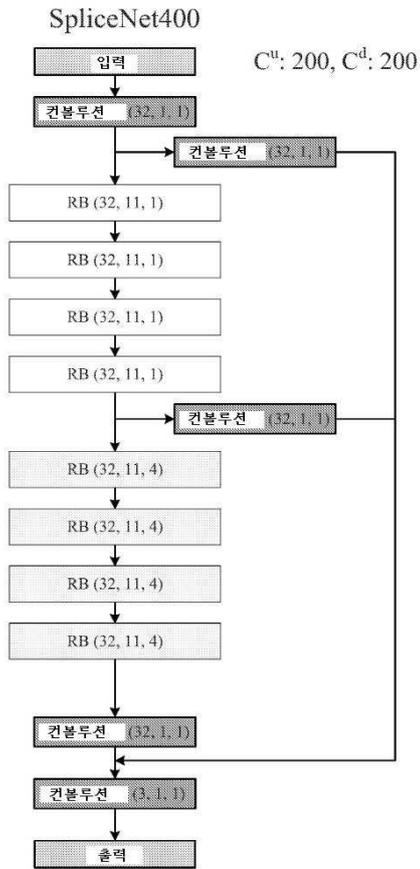


N: 컨볼루션 필터의 개수
 W: 컨볼루션 윈도우 크기
 D: 아트리스 컨볼루션 레이트

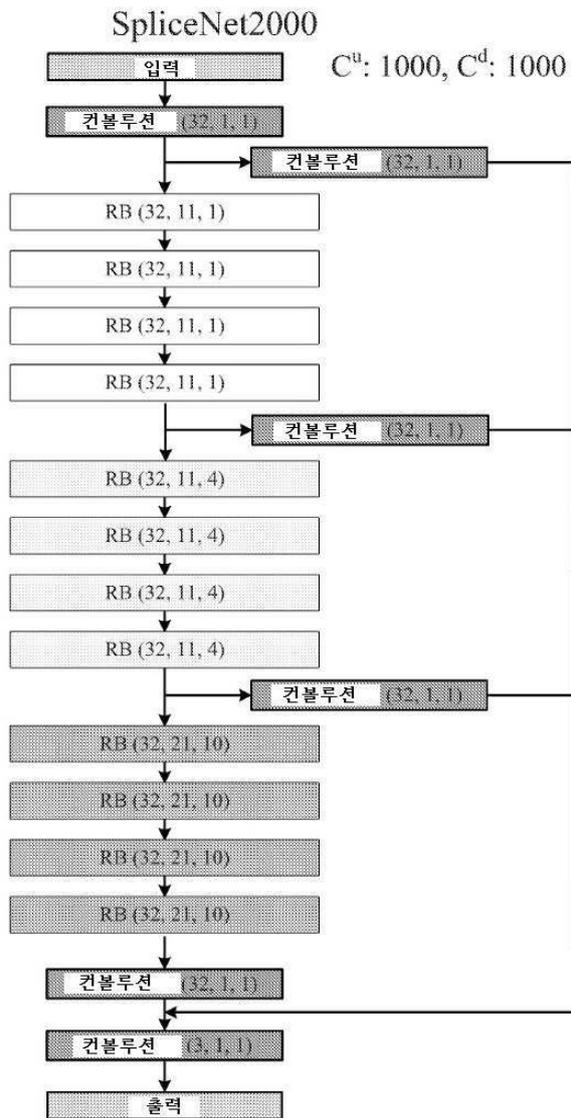
도면21



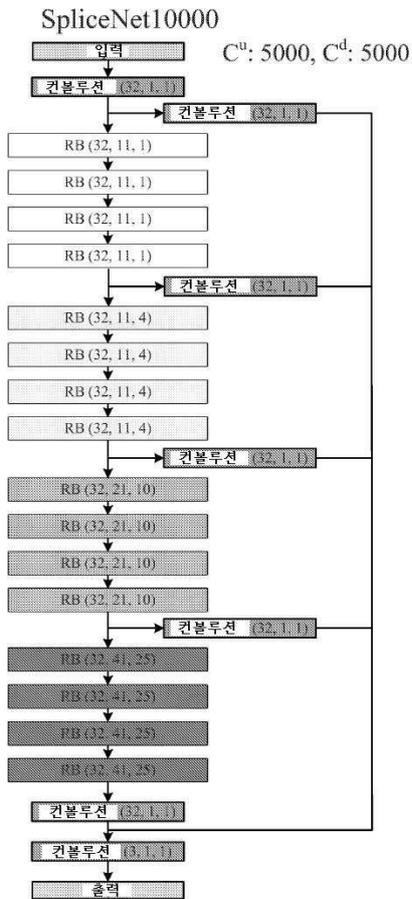
도면22



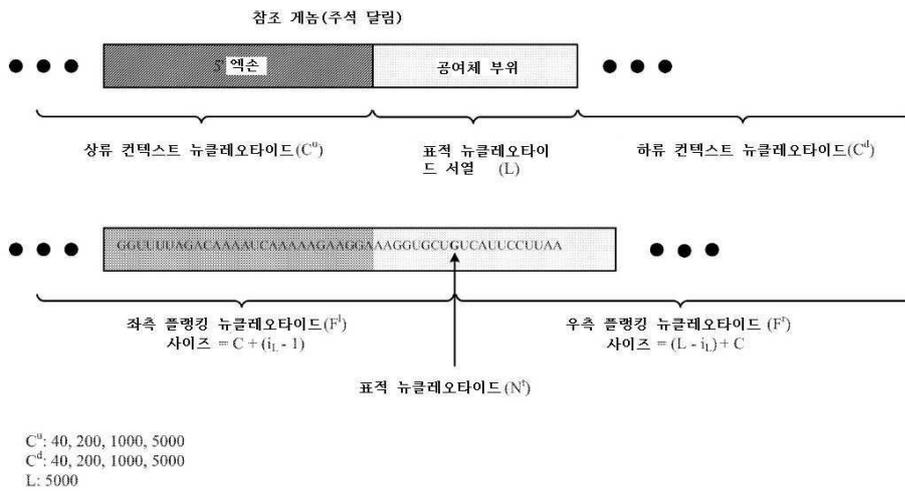
도면23



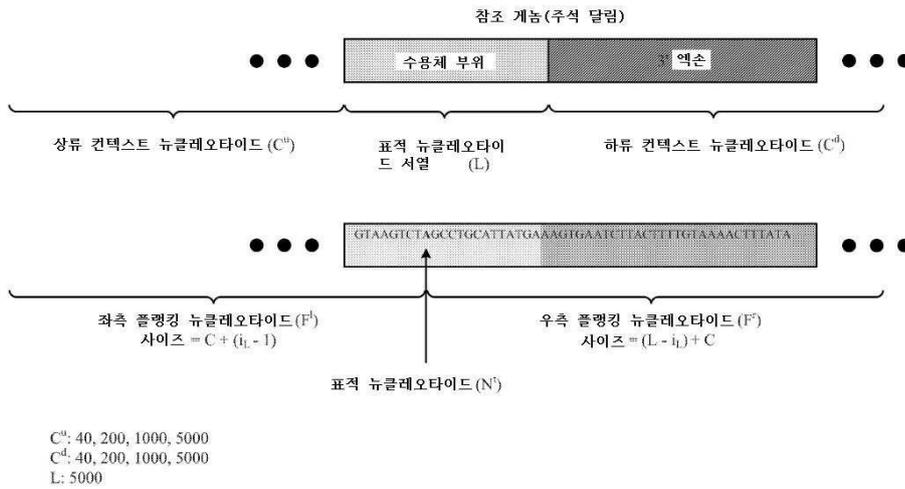
도면24



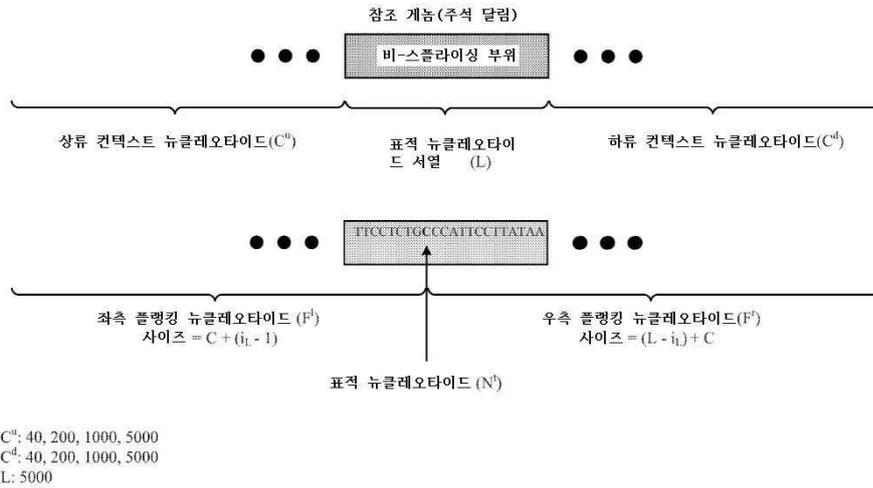
도면25



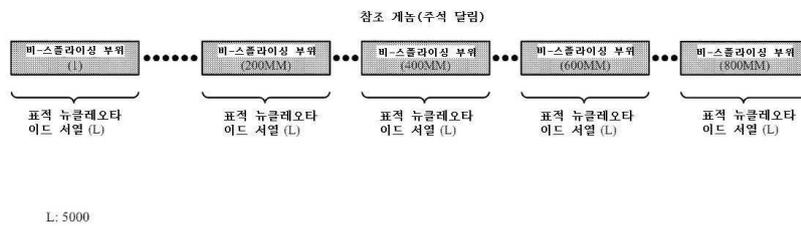
도면26



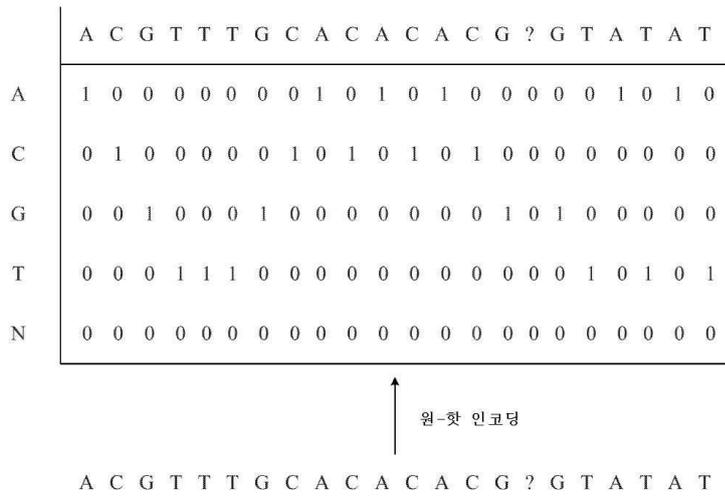
도면27



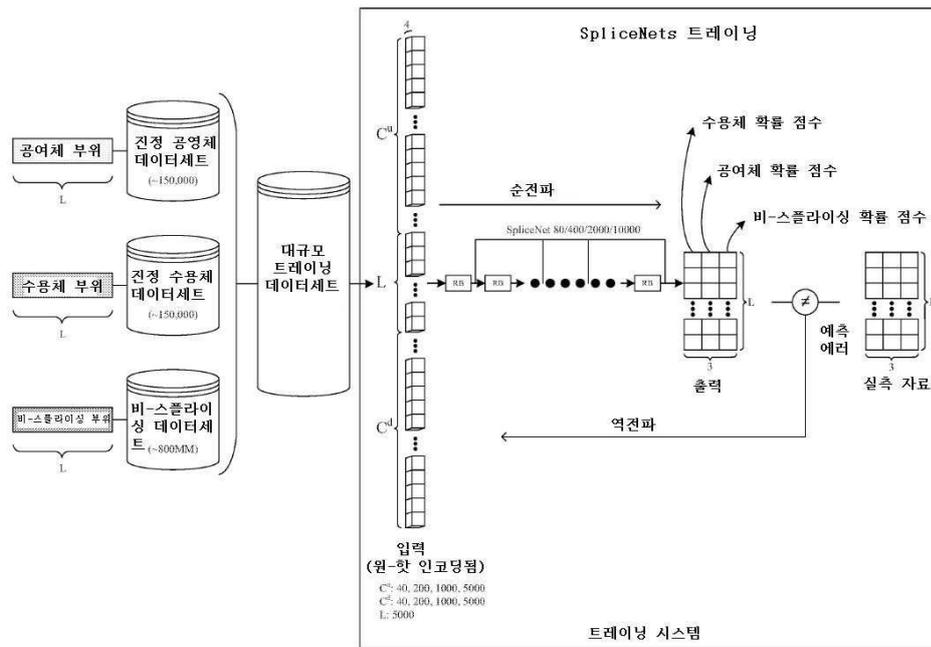
도면28



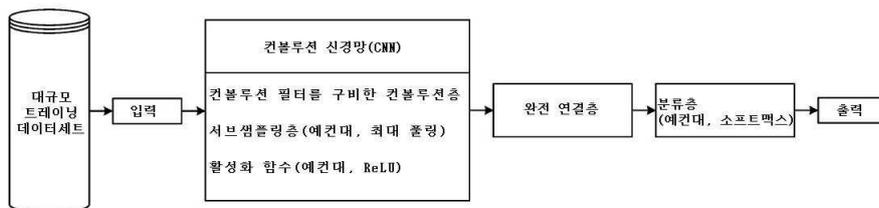
도면29



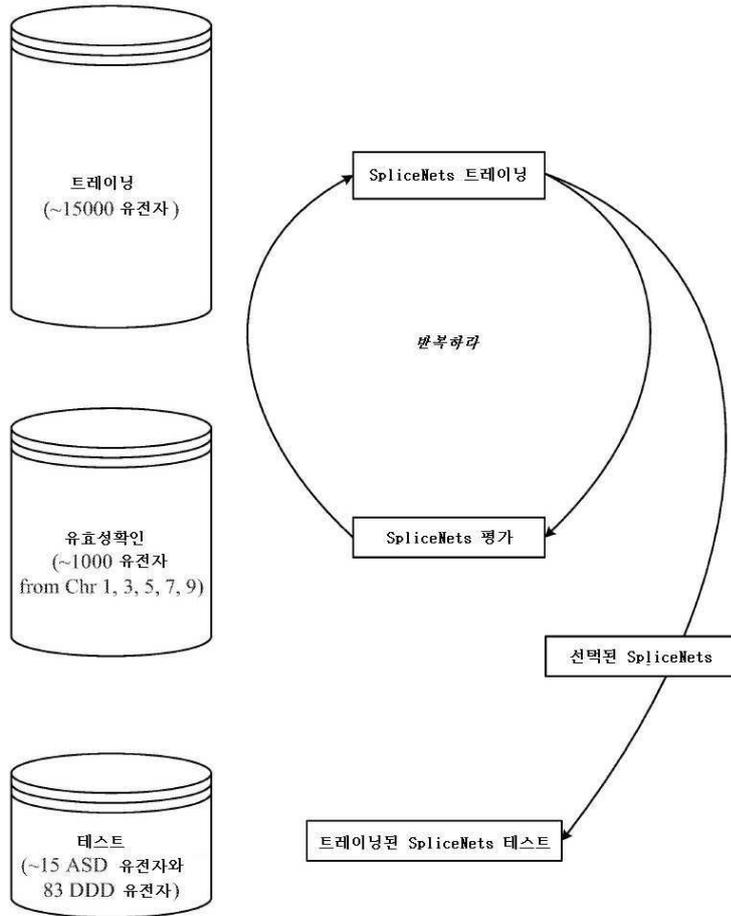
도면30



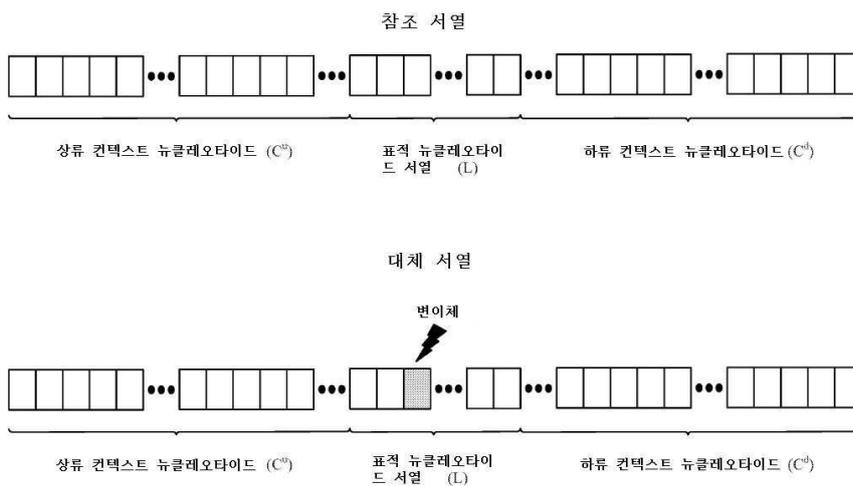
도면31



도면32

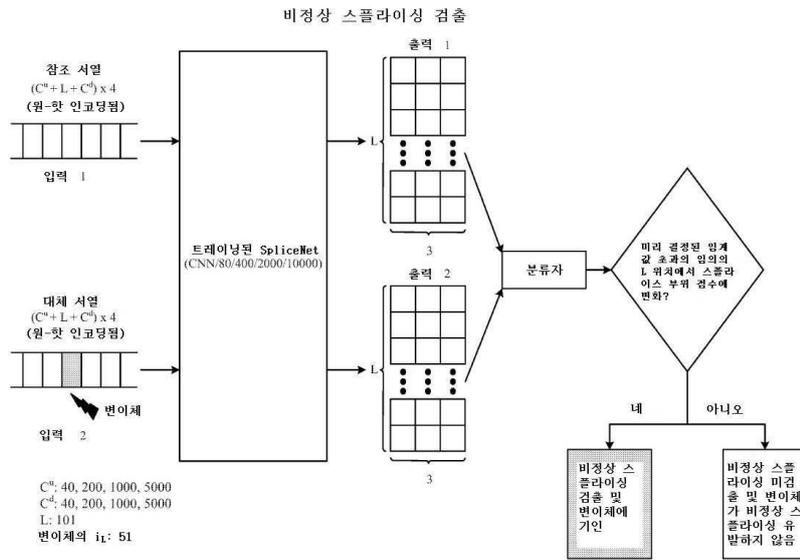


도면33

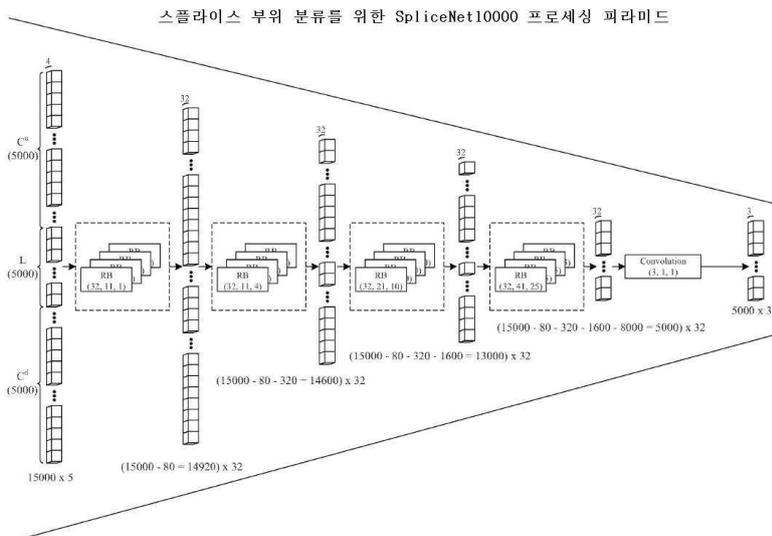


C^u: 40, 200, 1000, 5000
 C^d: 40, 200, 1000, 5000
 L: 101

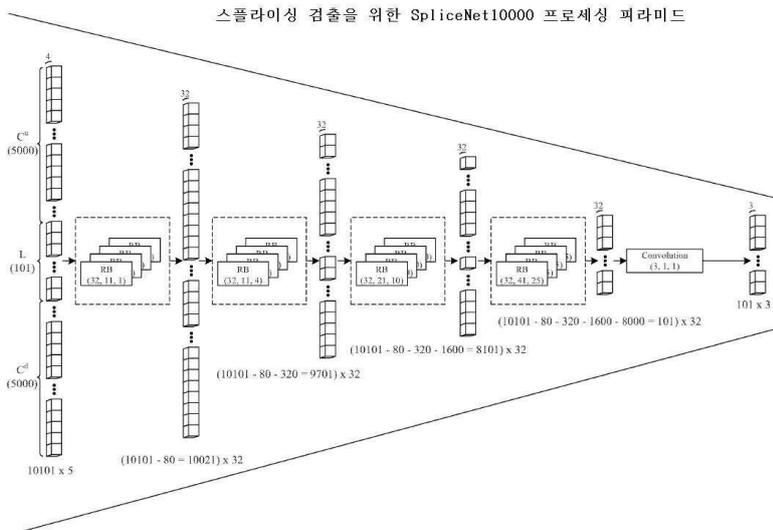
도면34



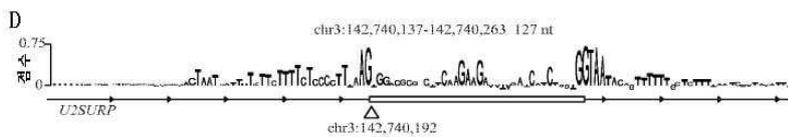
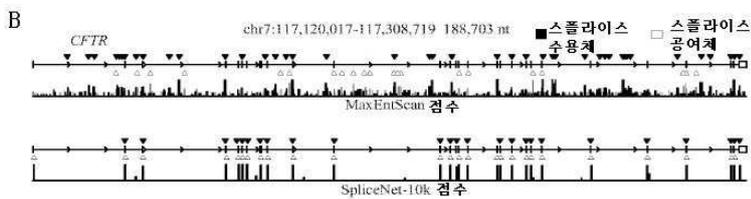
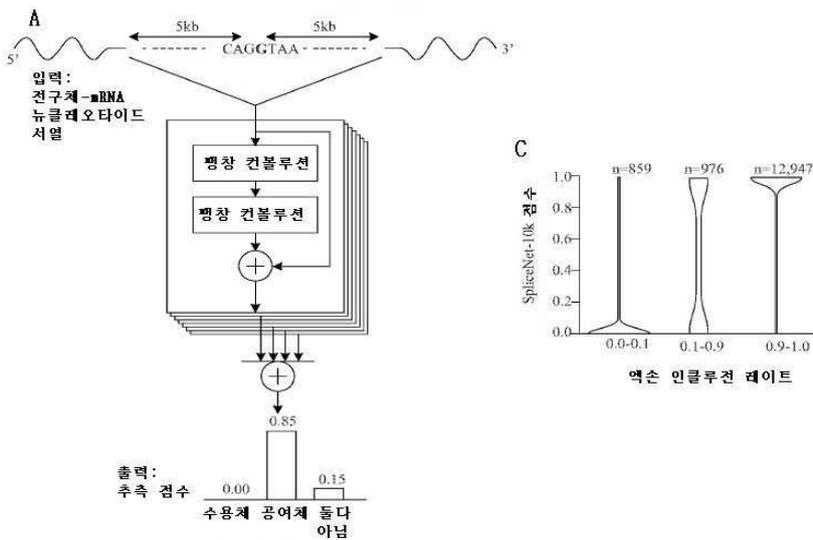
도면35



도면36



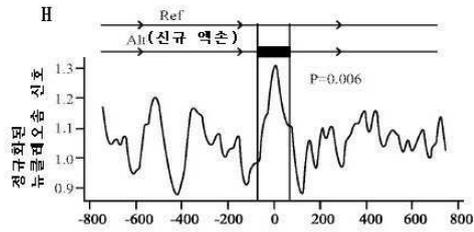
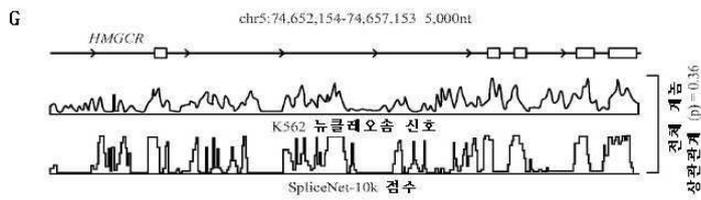
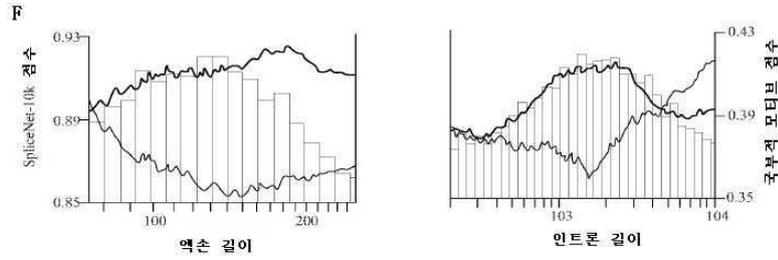
도면37ad



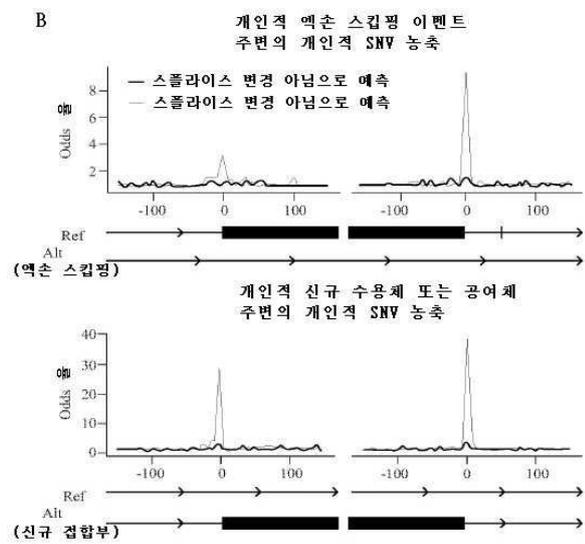
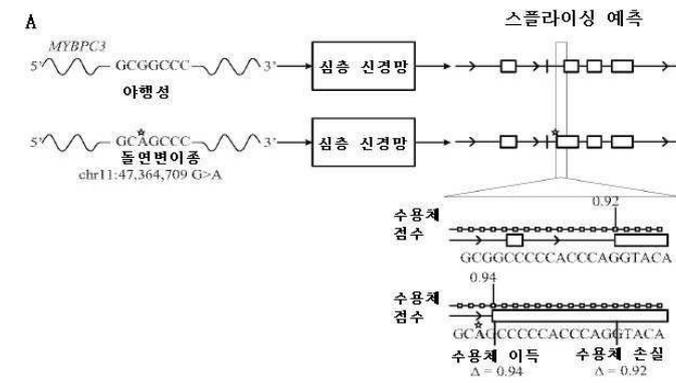
도면37eh

E

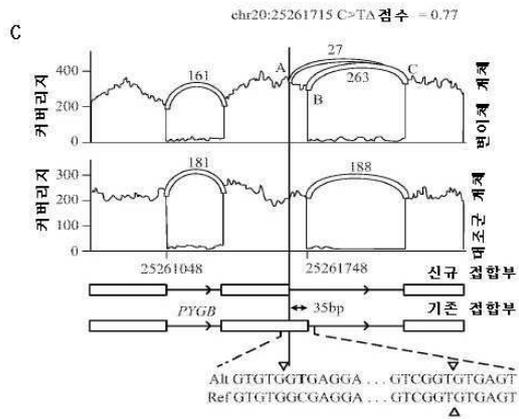
	Top-k 정확도	PR-AUC
SpliceNet-80nt	0.57	0.60
SpliceNet-400nt	0.90	0.95
SpliceNet-2k	0.93	0.97
SpliceNet-10k	0.95	0.98
GeneSplicer	0.30	0.23
MaxEntScan	0.22	0.15
NNSplice	0.22	0.15



도면38ab

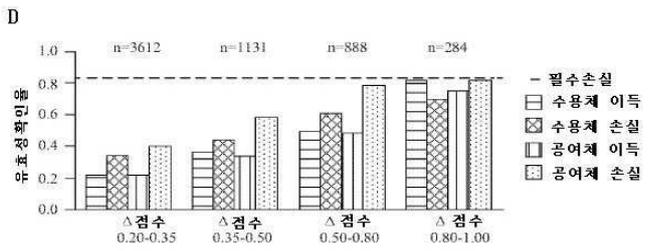
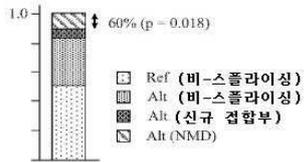


도면38cd

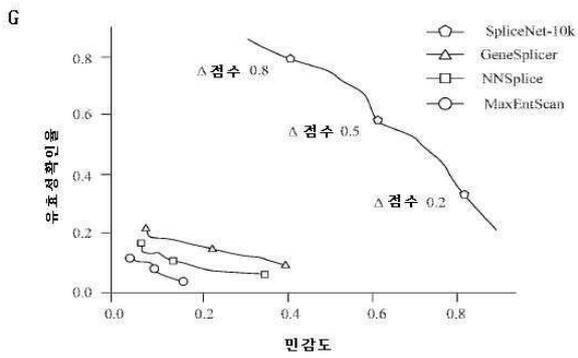
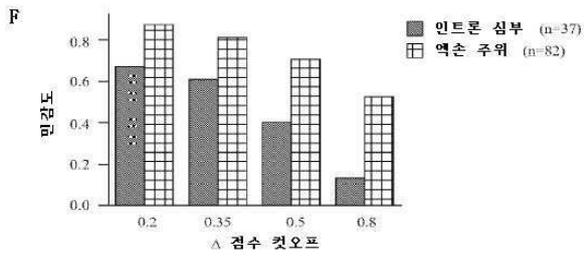
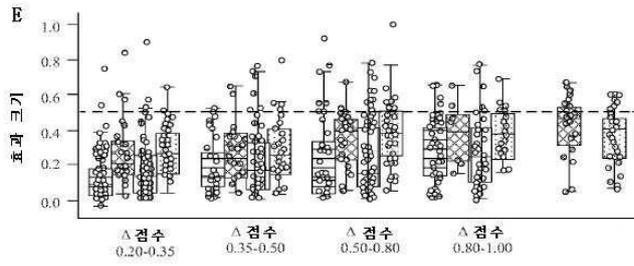


신규 접합부의 상대적 사용

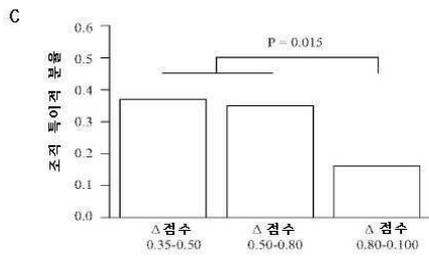
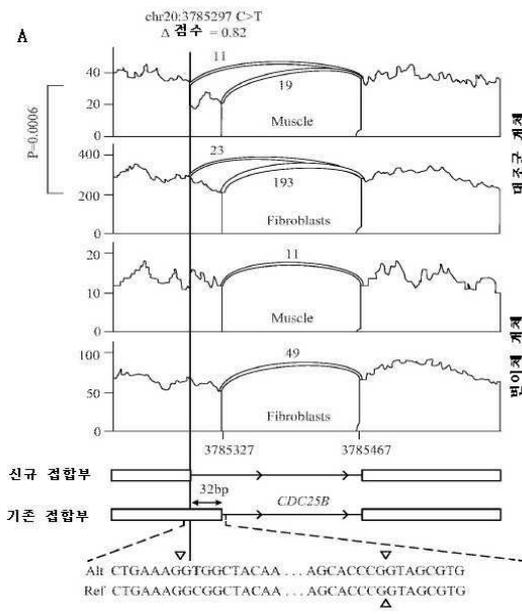
$$\left(\frac{AC}{AC+BC} \right)_{mut} - \left(\frac{AC}{AC+BC} \right)_{ctrl} = \frac{27}{27+263} - 0 = 0.09$$



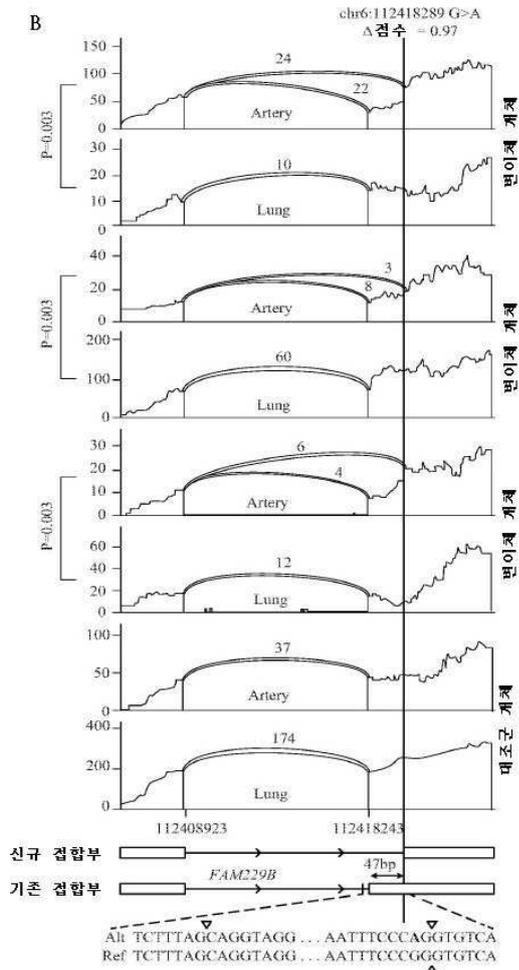
도면38eg



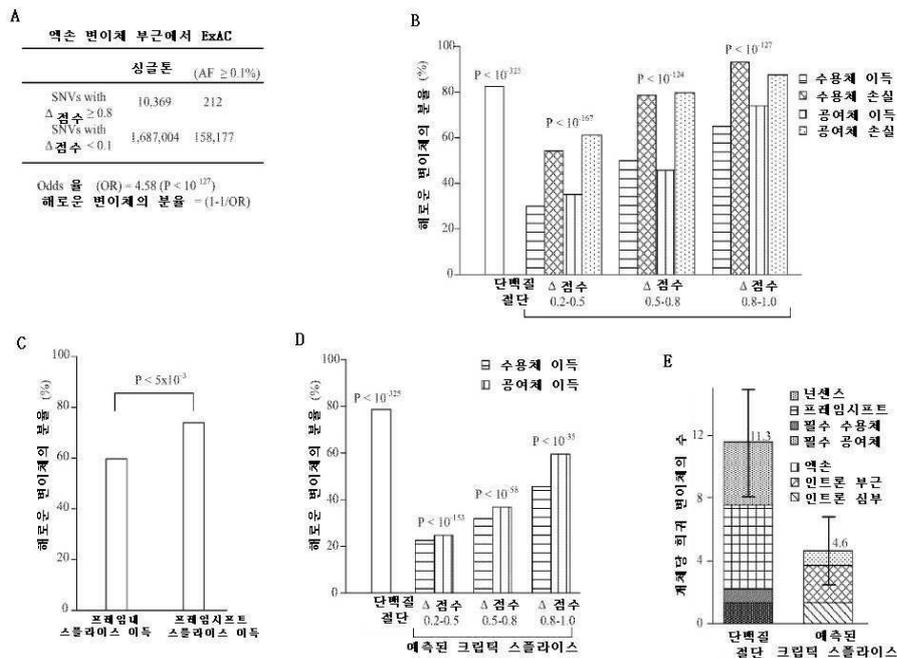
도면39ac



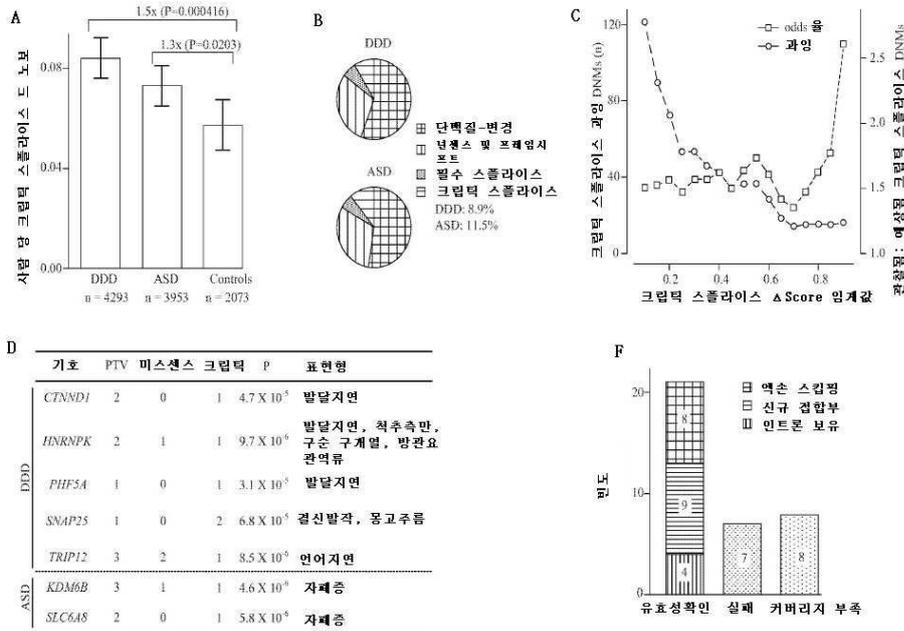
도면39b



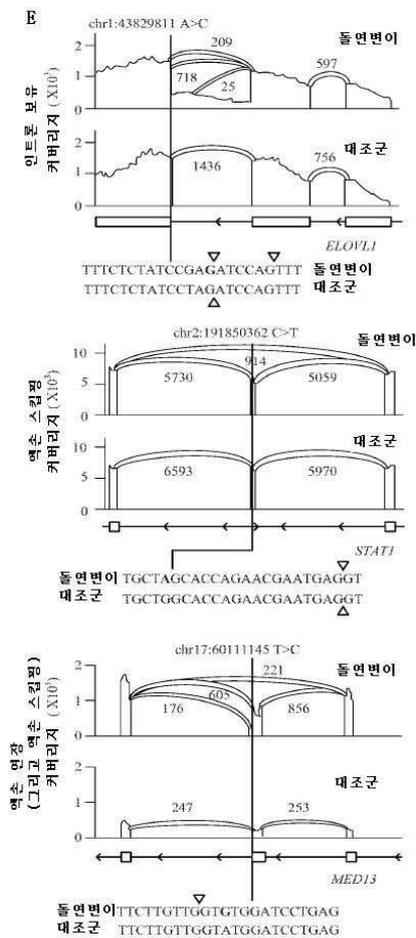
도면40



도면41af



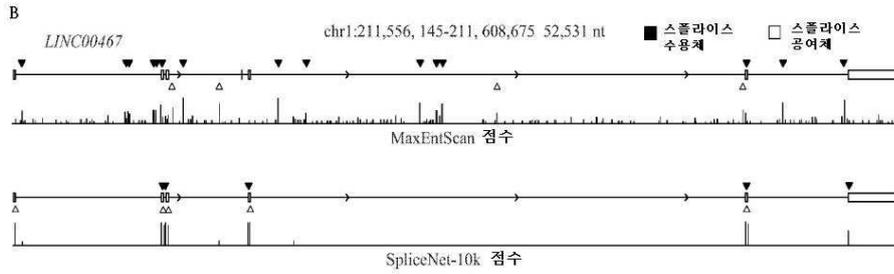
도면41e



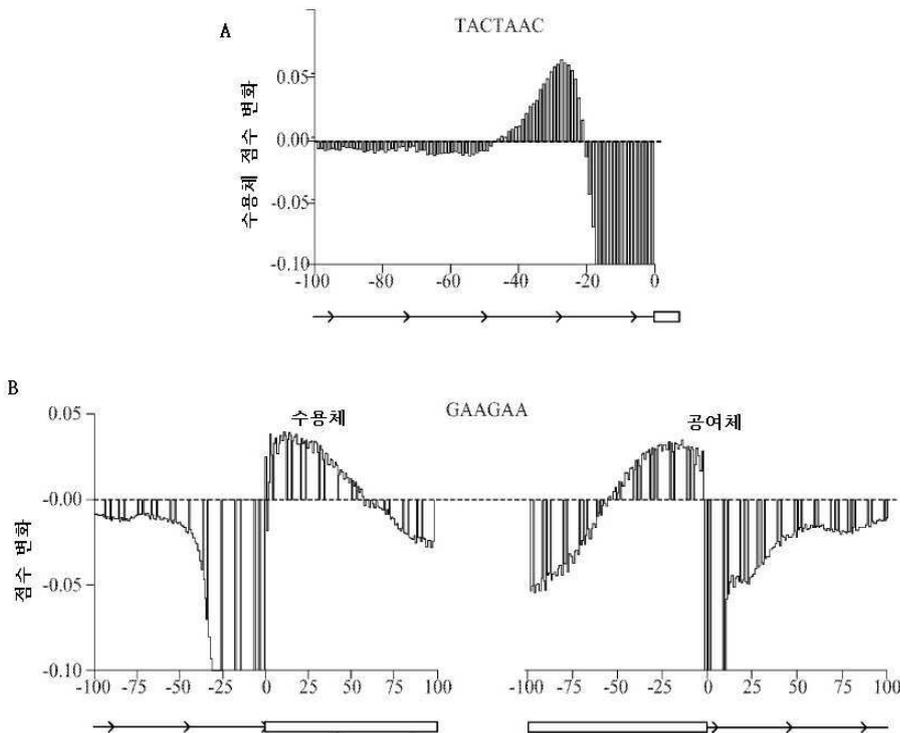
도면42

A

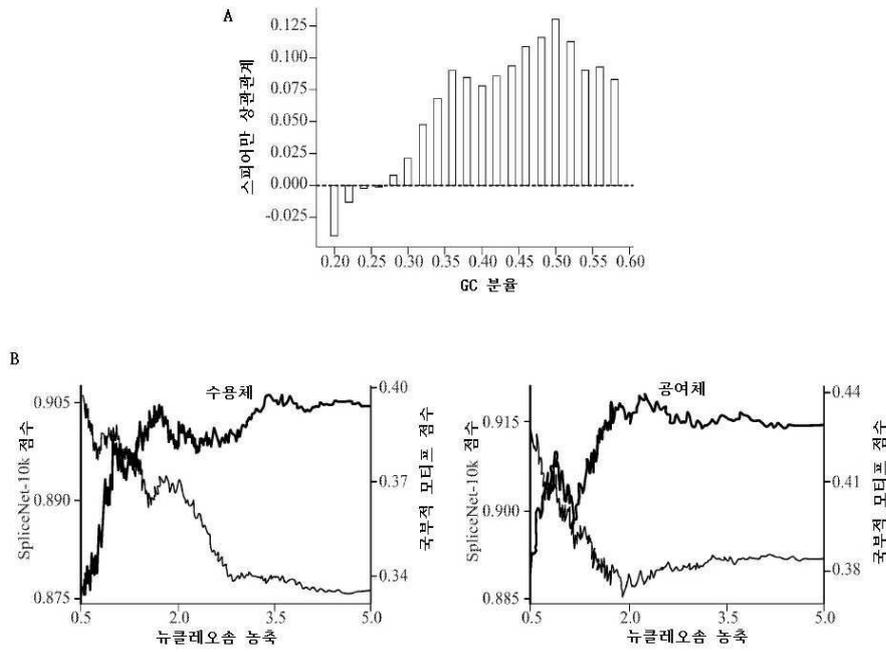
	Top-k 정확도		정밀-재호출	
	수용체	공여체	수용체	공여체
SpliceNet-80nt	0.5129	0.5110	0.5195	0.5273
SpliceNet-400nt	0.6848	0.7221	0.7320	0.7982
SpliceNet-2k	0.8090	0.8405	0.8742	0.9047
SpliceNet-10k	0.8223	0.8586	0.8896	0.9107
GeneSplicer	0.3238	0.3381	0.2234	0.2982
MaxEntScan	0.2932	0.3687	0.2233	0.2937
NNSplice	0.2655	0.3687	0.1702	0.3003



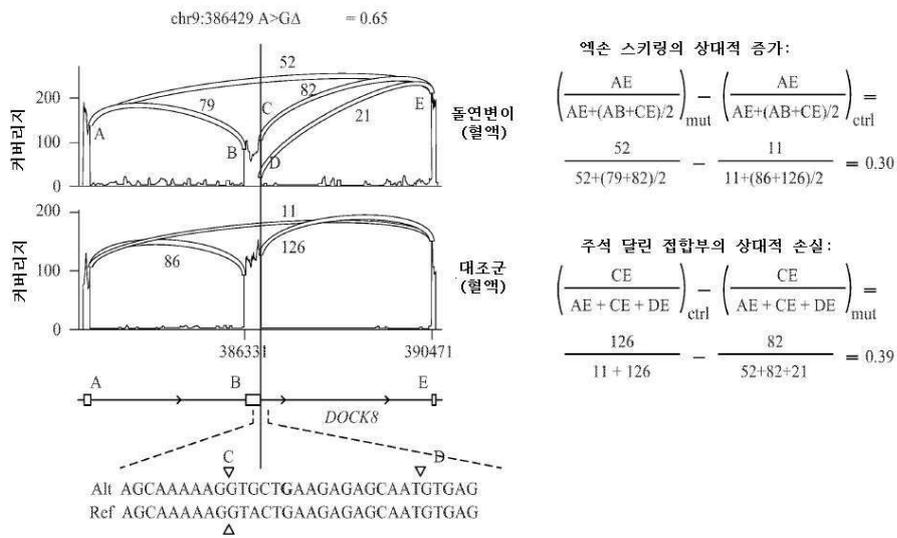
도면43



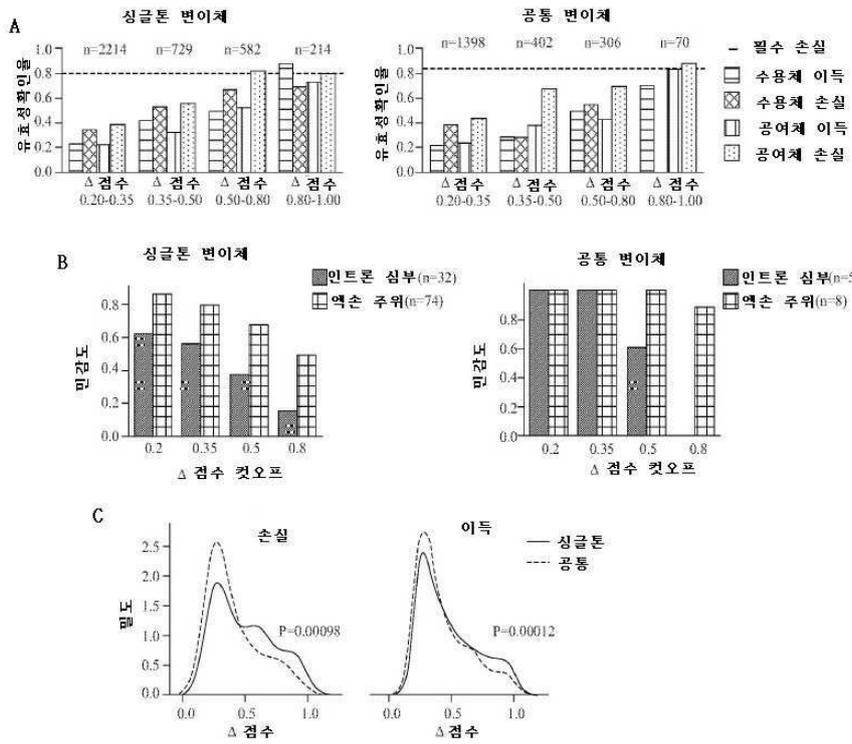
도면44



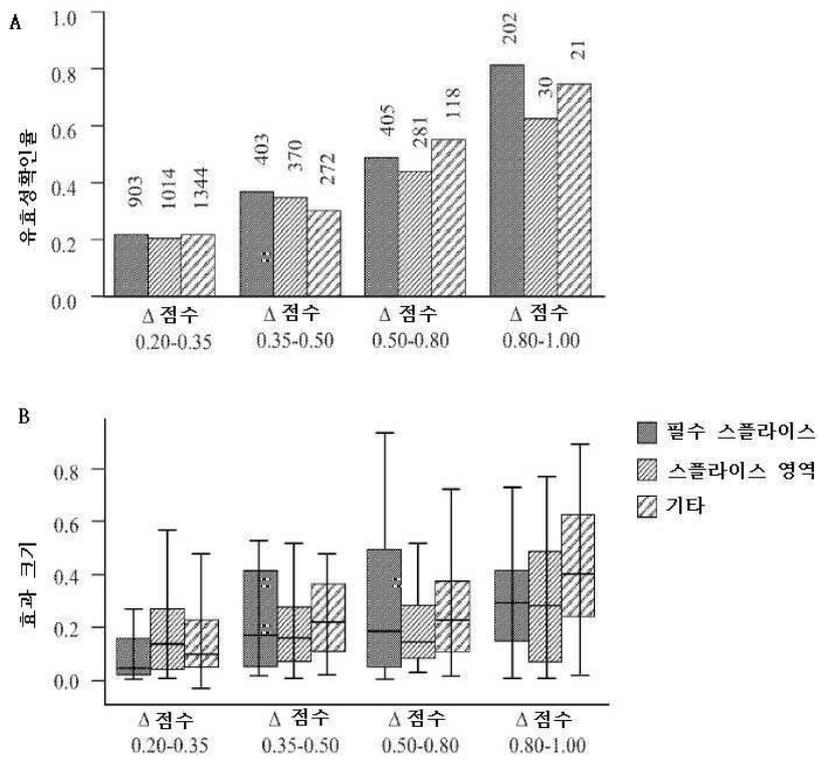
도면45



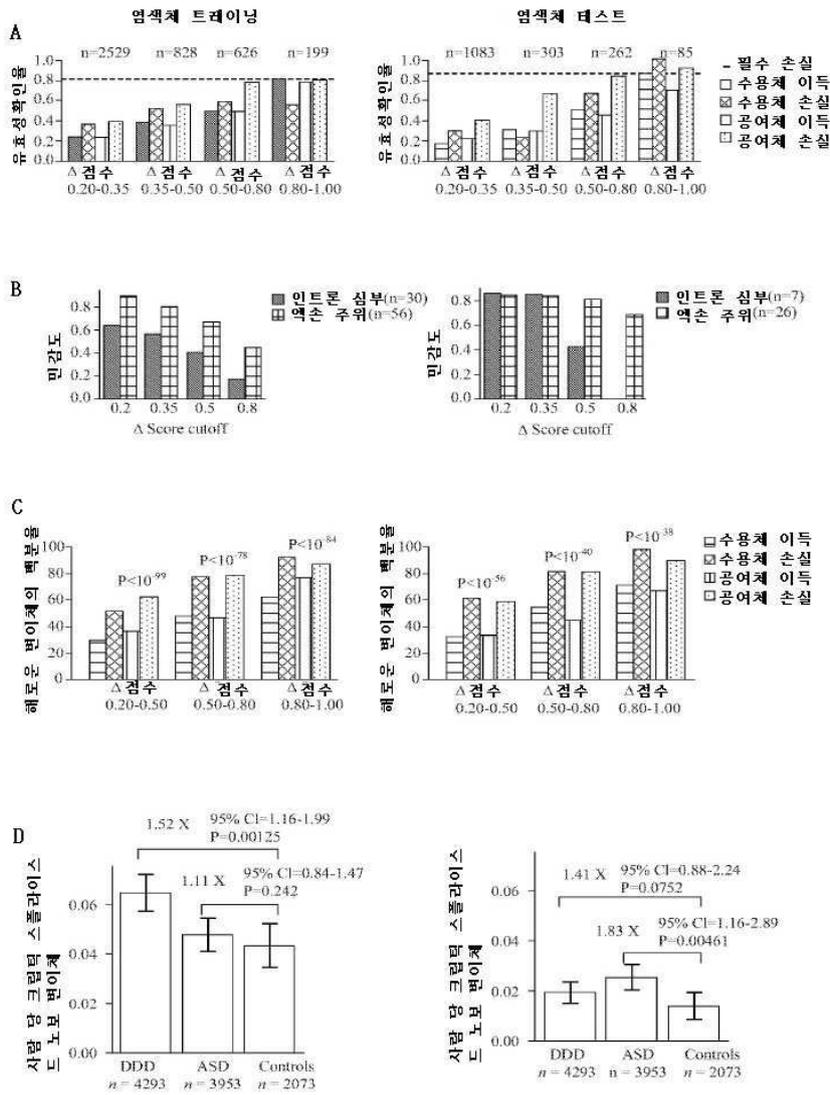
도면46



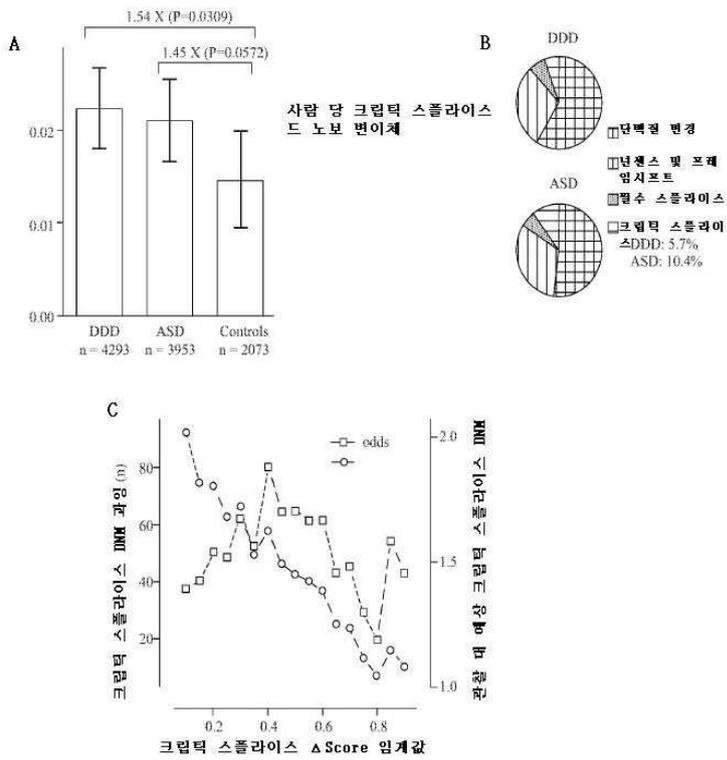
도면47



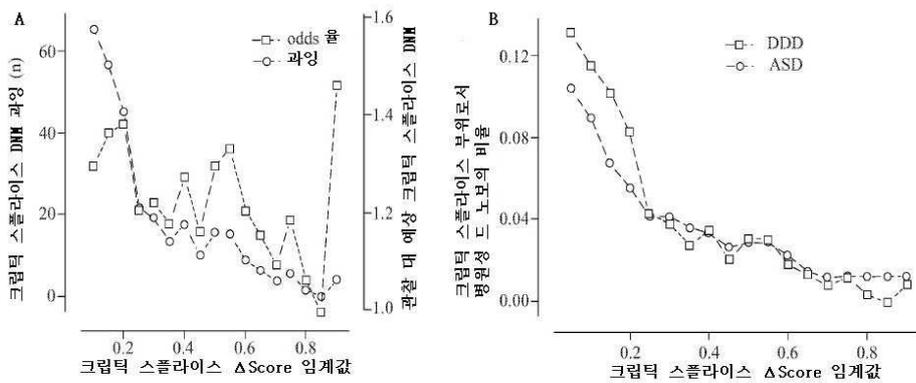
도면48



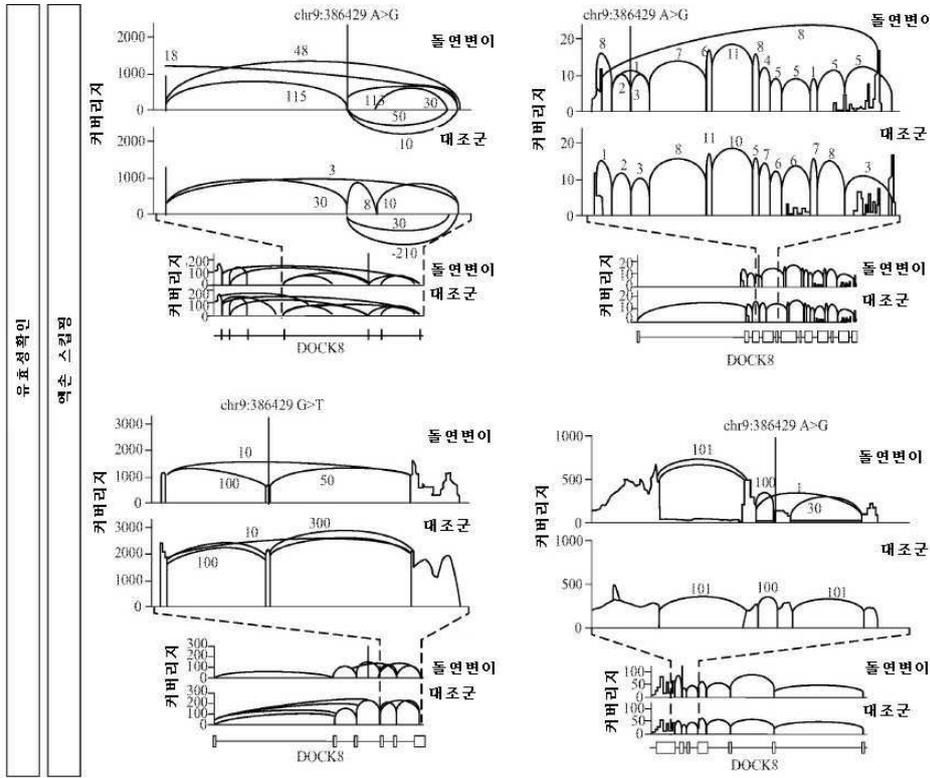
도면49



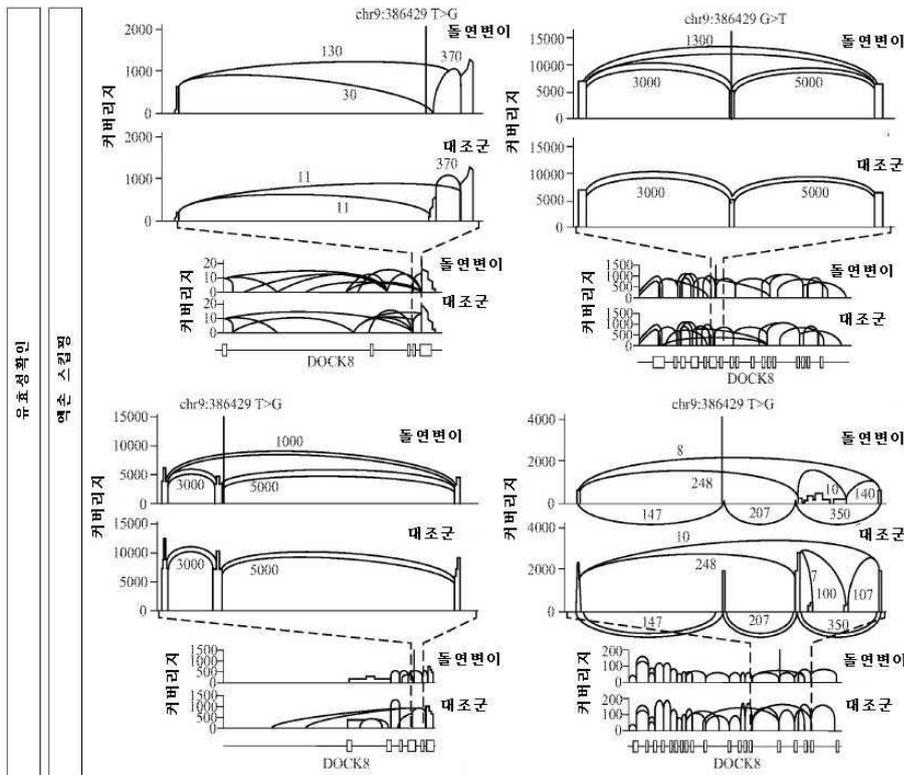
도면50



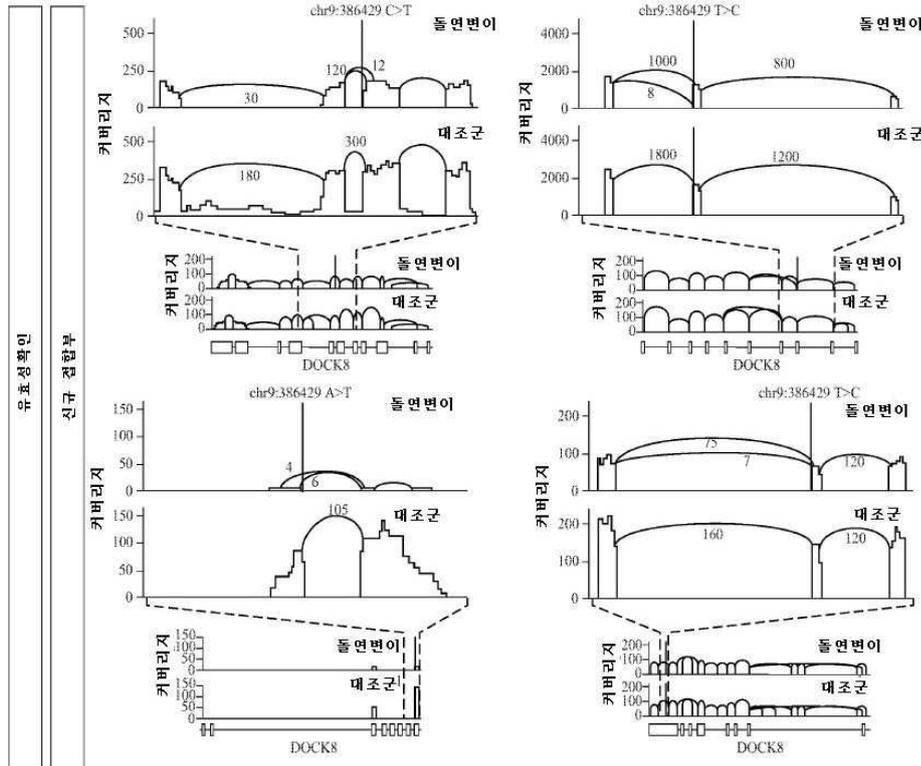
도면51a



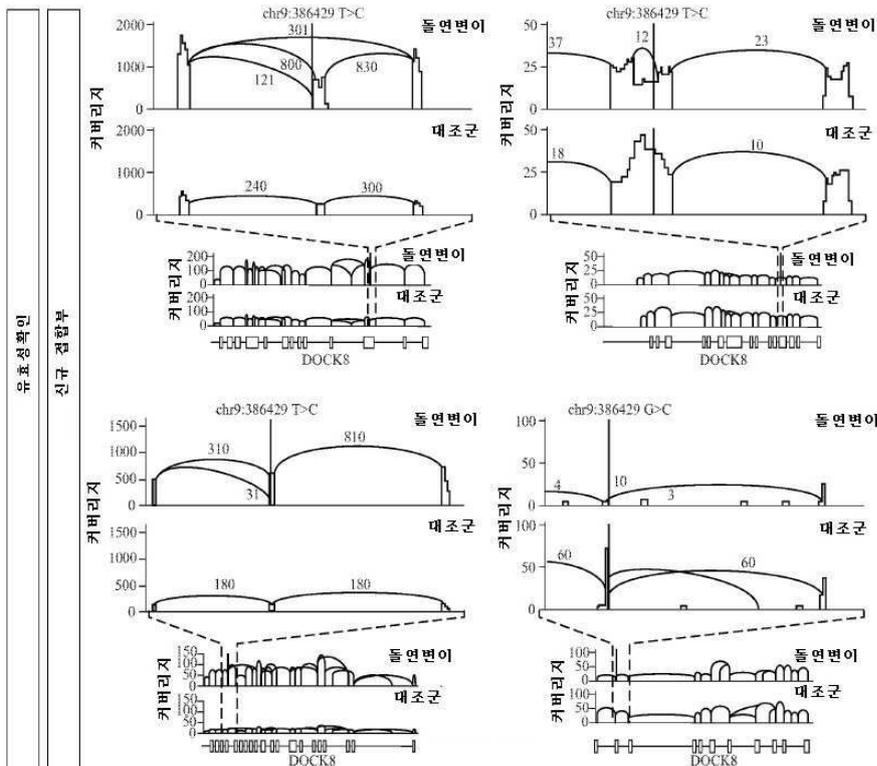
도면51b



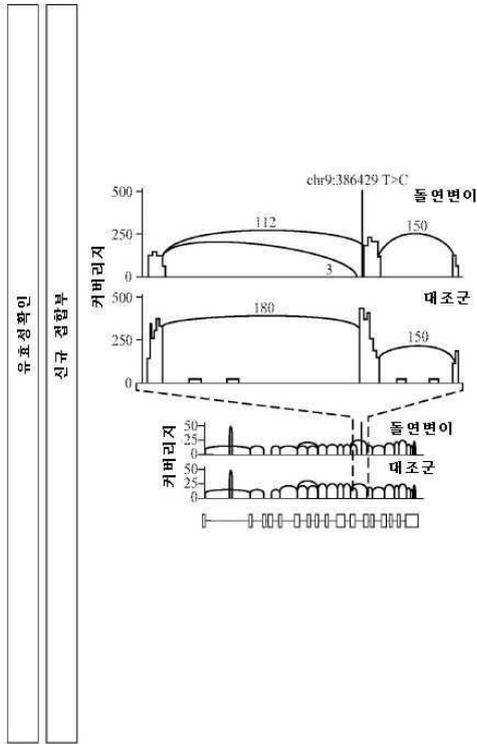
도면51c



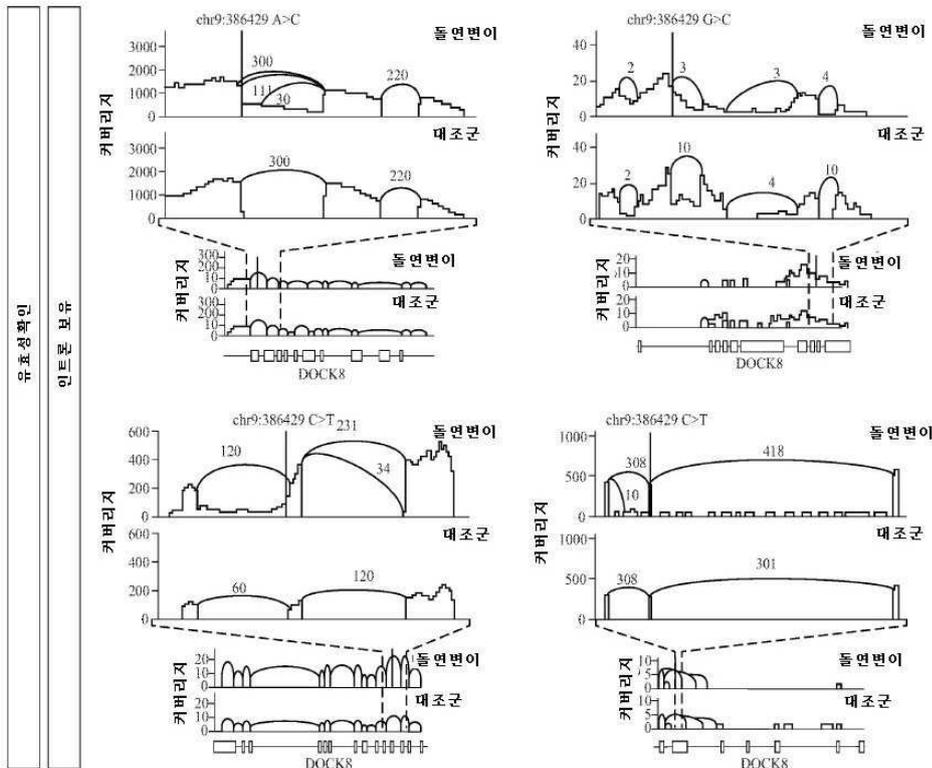
도면51d



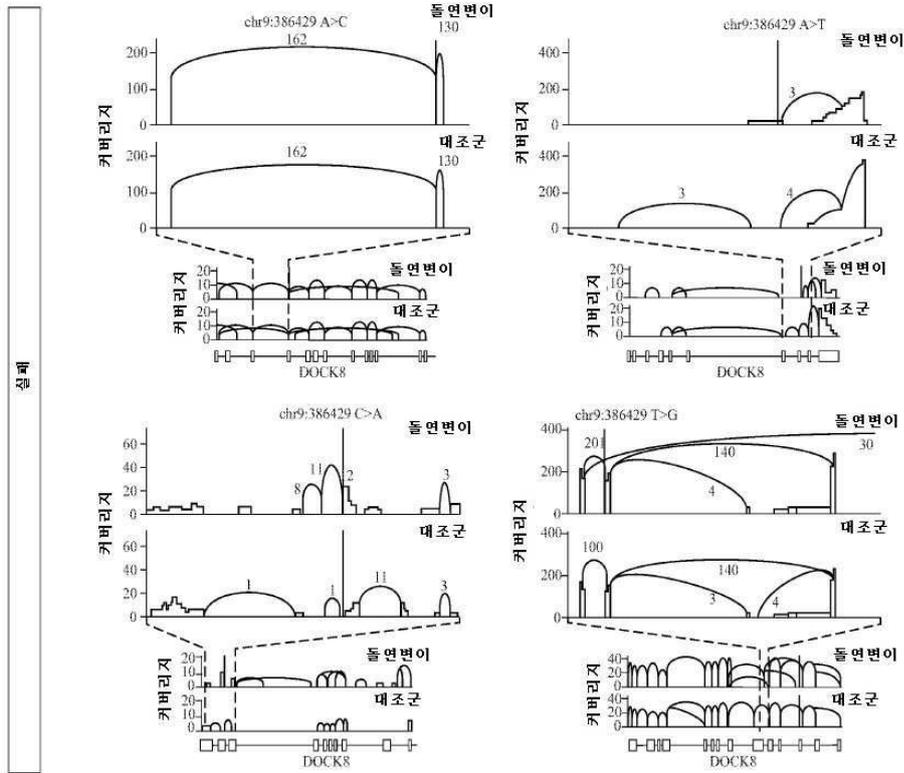
도면51e



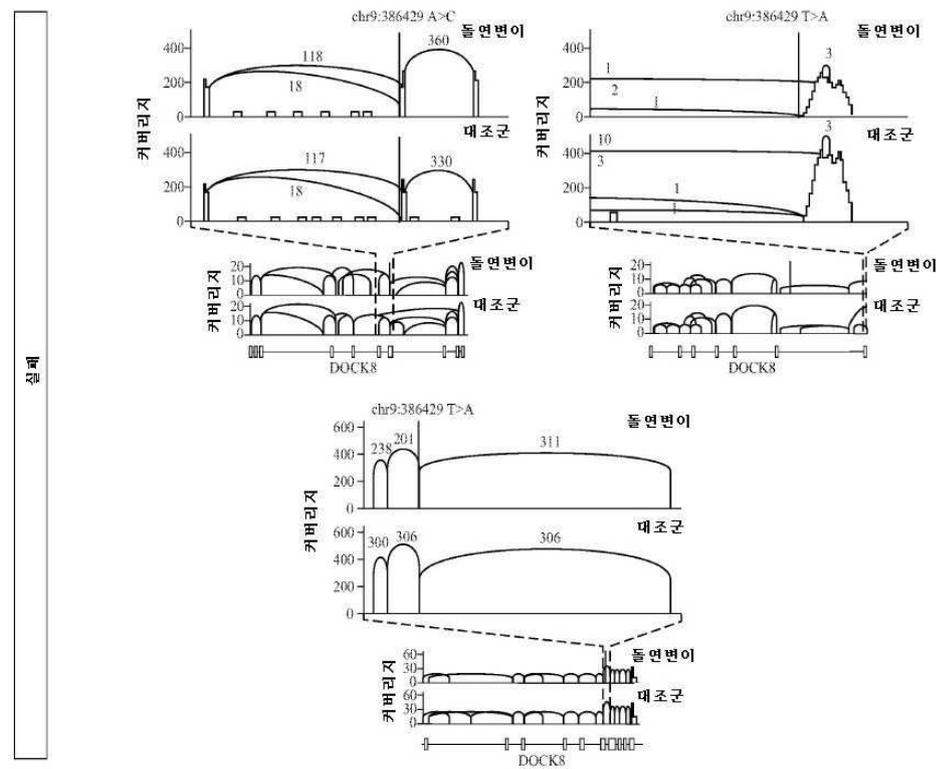
도면51f



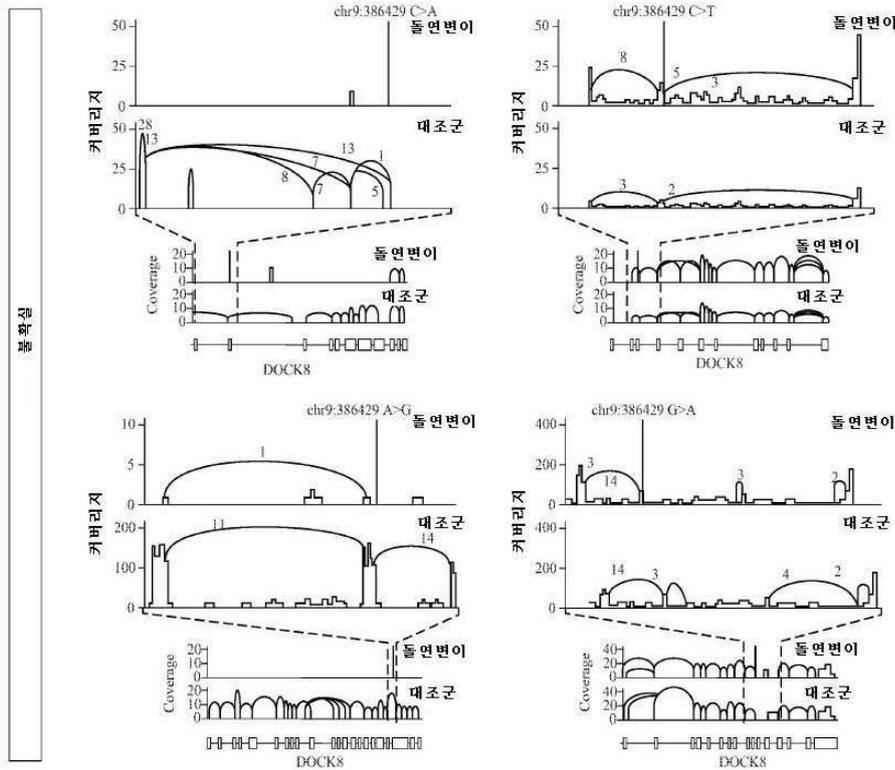
도면51g



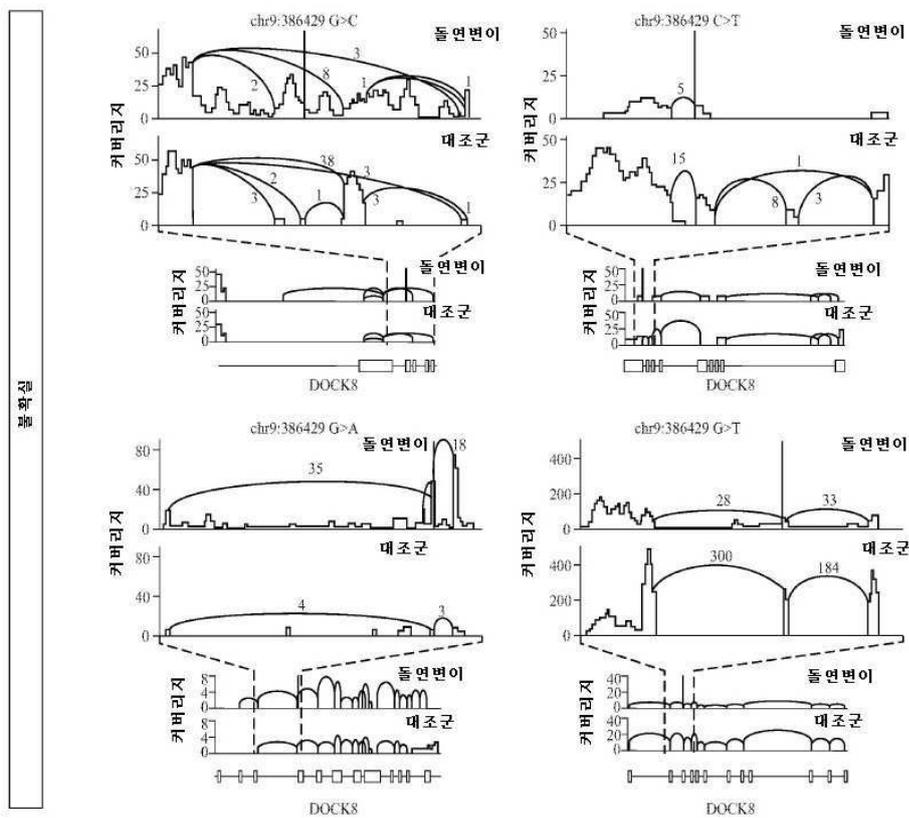
도면51h



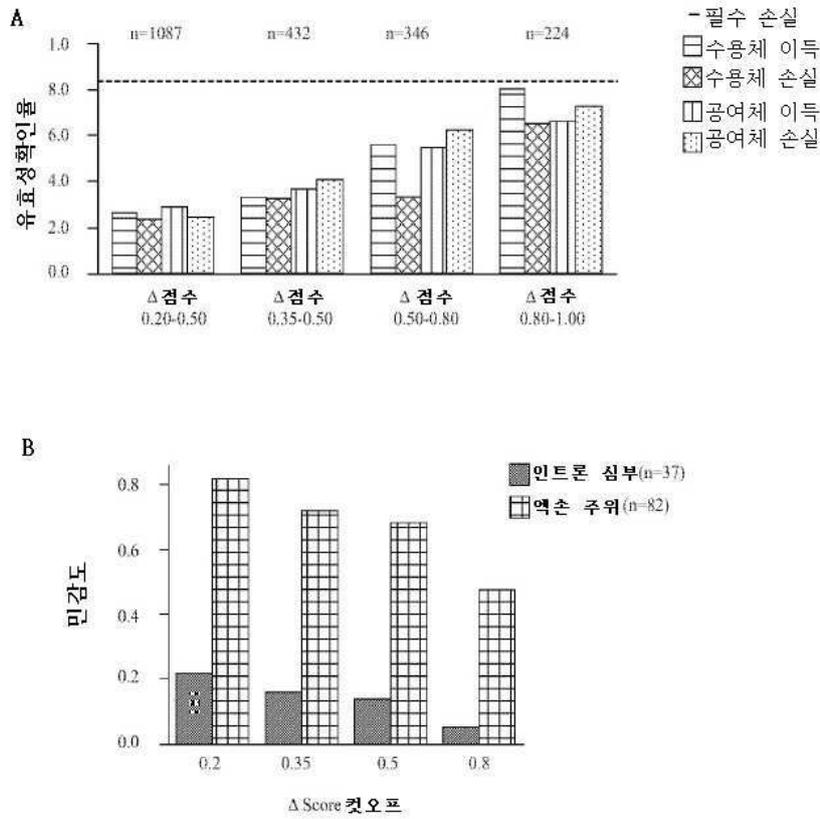
도면51i



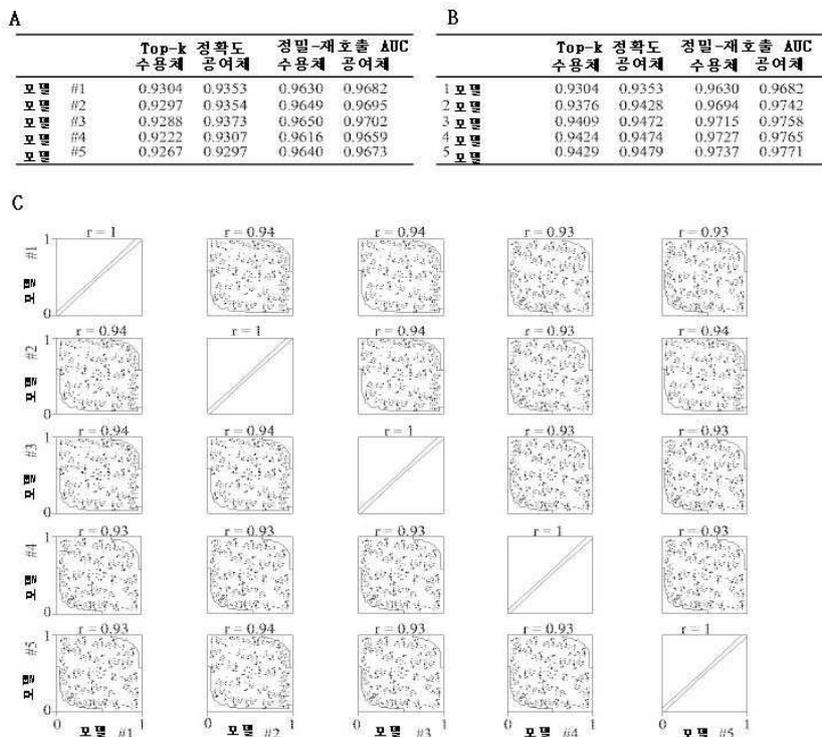
도면51j



도면52



도면53



도면54

A SpliceNet-10k				
엑손 밀도	Top-k 정확도		정밀-재호출 AUC	
	수용체	공여체	수용체	공여체
1	0.9165	0.9292	0.9592	0.9659
2	0.9307	0.9357	0.9645	0.9674
3	0.9519	0.9512	0.9816	0.9822
4	0.9556	0.9540	0.9833	0.9845
≥5	0.9579	0.9665	0.9826	0.9872

B MaxEntScan				
엑손 밀도	Top-k 정확도		정밀-재호출 AUC	
	수용체	공여체	수용체	공여체
1	0.1834	0.2487	0.1140	0.1660
2	0.2180	0.3124	0.1512	0.2375
3	0.2616	0.3348	0.1964	0.2773
4	0.2950	0.3820	0.2307	0.3347
≥5	0.3583	0.4652	0.3103	0.4314

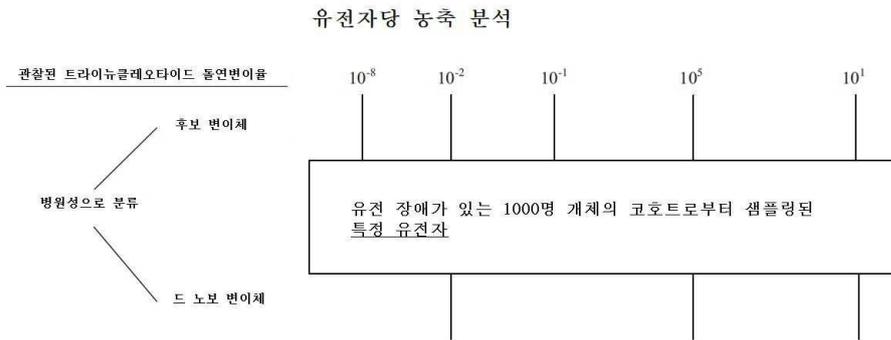
도면55

조직	변이체 개체	대조군 개체
피부 - 태양 노출(다리 하부)	T6MO	OXRK
근육 - 골격, 변형 섬유모세포	WHSB	WK11
동맥 - 경골, 폐	XOT4, S341, POMQ	WFON
전혈	TMMY	T8EM

도면56

이득				손실			
SpliceNet-10k	GeneSplicer	NNSplice	MaxEntScan	SpliceNet-10k	GeneSplicer	NNSplice	MaxEntScan
0.1	103.5145	0.9024	0.901	0.1	1.5183	0.1063	0
0.2	105.6838	0.9288	0.9171	0.2	2.6295	0.2315	0.0068
0.3	107.1061	0.9542	0.9419	0.3	3.4707	0.3554	0.0312
0.4	108.0655	0.967	0.9582	0.4	4.0856	0.4625	0.0646
0.5	108.9011	0.9739	0.9694	0.5	4.6085	0.5349	0.1003
0.6	109.6138	0.9781	0.9759	0.6	5.2064	0.6345	0.1624
0.7	110.388	0.9817	0.9826	0.7	6.4456	0.7573	0.2838
0.8	111.2775	0.9839	0.9875	0.8	8.1053	0.8202	0.4271
0.9	112.6826	0.9858	0.9915	0.9	10.2586	0.8529	0.641

도면57



$$\sum (\text{관찰된 트라이뉴클레오타이드 돌연변이율}) = 10^{-5}$$

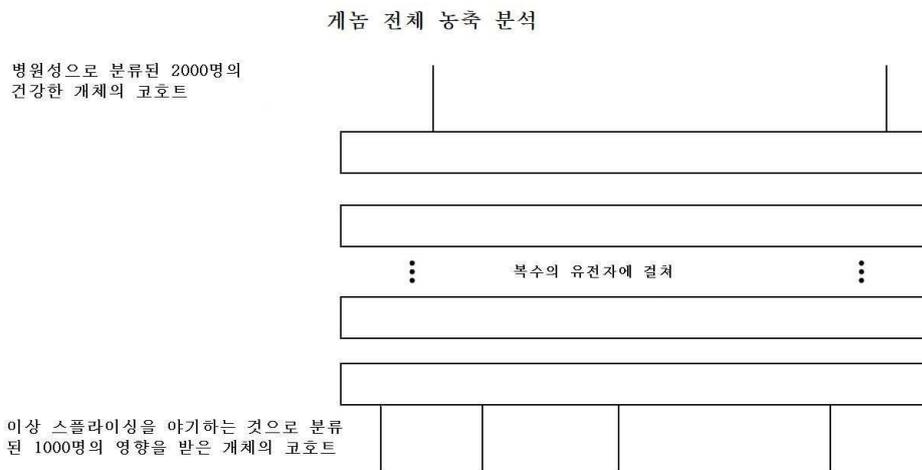
돌연변이의 베이스라인 수 = $2 \times 1000 \times 10^{-5} = 0.2$

↑ 염전(염색체) 카운트

↓ 코호트 크기

드 노보 변이체 카운트 = 3

도면58



건강한 코호트의 돌연변이율 = $2/2000 = 0.001$

영향을 받은 코호트의 돌연변이율 = $4/1000 = 0.004$

개체당 돌연변이율 = $0.004/0.001 = 4$

도면59

