(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2021년10월18일
(11) 등록번호 10-2314219
(24) 등록일자 2021년10월12일

(51) 국제특허분류(Int. Cl.)
G16B 20/20 (2019.01) *G06K 9/62* (2006.01)
G06N 3/08 (2006.01) *G06N 7/00* (2006.01)
(52) CPC특허분류
G16B 20/20 (2019.02)
G06K 9/6257 (2013.01)
(21) 출원번호 10-2019-7036423
(22) 출원일자(국제) 2018년10월15일
심사청구일자 2019년12월26일
(85) 번역문제출일자 2019년12월09일
(65) 공개번호 10-2020-0011445
(43) 공개일자 2020년02월03일
(86) 국제출원번호 PCT/US2018/055881
(87) 국제공개번호 WO 2019/079182
국제공개일자 2019년04월25일
(30) 우선권주장
62/573,144 2017년10월16일 미국(US)
(뒷면에 계속)

(56) 선행기술조사문헌
Qiong Wei 외, The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics, PLOS ONE, 2013.07.09., Vol.8, Issue.7, pp1-12.
CA2894317 A1
Nilah M. Ioannidis 외, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants, Am J Hum Genet, 2016.09.22., Vol.99, pp877-885.
Karthik A Jagadeesh 외, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity, Nature Genetics, 2016.10.24., Vol.48, pp1581-1586.

(73) 특허권자
일루미나, 인코포레이티드
미국 캘리포니아 92122 샌디에고 일루미나 웨이 5200
(72) 발명자
순다람 락슈만
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 18 항 심사관 : 성경아

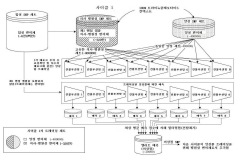
(54) 발명의 명칭 **심층 컨볼루션 신경망의 앙상블을 트레이닝하기 위한 반감독 학습**

(57) 요 약

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

(뒷면에 계속)

대 표 도



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 달리하며, 아트로스 컨볼루션 레이트는 잔여 블록의 그룹들 간에 달리한다. 트레이닝 데이터는, 양성 변이체 및 병원성 변이체로부터 생성되는 번역된 서열 쌍들의 양성 트레이닝 예와 병원성 트레이닝 예를 포함한다.

(52) CPC특허분류

G06K 9/6259 (2013.01)
G06K 9/6267 (2013.01)
G06N 3/08 (2013.01)
G06N 7/005 (2013.01)
G16B 40/00 (2019.02)

(72) 발명자

가오 홍

미국 캘리포니아주 92122 샌디에이고 5200 일루미나 웨이

맥레 제레미 프랜시스

미국 캘리포니아주 92122 샌디에이고 5200 일루미나 웨이

(30) 우선권주장

62/573,149 2017년10월16일 미국(US)
62/573,153 2017년10월16일 미국(US)
62/582,898 2017년11월07일 미국(US)

명 세 서

청구범위

청구항 1

병원성(pathogenic) 단일 뉴클레오타이드 다형성(single nucleotide polymorphism: SNP)을 식별하는 컴퓨터-구현 방법으로서,

컨볼루션 신경망(convolutional neural network: CNN)의 앙상블(ensemble)을 이용해서 사이클들에서 미표지(unlabeled) SNP를 처리하는 단계로서, 상기 사이클은 제1 위상(phase)과 제2 위상을 포함하고;

상기 미표지 SNP는 양성(benign)으로 표지되지 않고 엘리트 예측 병원성 SNP(elite predicted pathogenic SNP)에 첨가되지 않은 조합적으로 생성된 SNP 변이체를 포함하는, 상기 처리하는 단계;

상기 사이클의 제1 위상에서 처리하는 단계로서,

상기 미표지 SNP로부터 상기 앙상블에서의 CNN당 개별의 의사-병원성 트레이닝 예를 선택하여, 만약 엘리트 예측 병원성 SNP가 있다면, 엘리트 예측 병원성 SNP와 조합하여 사용하는 단계; 및

선택된 의사-병원성 트레이닝 예 및 만약 엘리트 예측 병원성 SNP가 있다면, 상기 엘리트 예측 병원성 SNP와 균형을 이룬, SNP의 공통 양성(benign) 트레이닝 예를 사용해서 상기 CNN의 앙상블을 트레이닝시킴으로써, 상기 앙상블에서의 상기 CNN의 파라미터를 업데이트하는 단계를 포함하는, 상기 사이클의 제1 위상에서 처리하는 단계;

상기 사이클의 제2 위상에서 처리하는 단계로서,

상기 CNN의 앙상블에 업데이트된 파라미터를 적용하여 상기 제1 위상에 연루되지 않은 미표지 SNP에 대해서 병원성 점수를 생성하는 단계;

상기 병원성 점수를 이용해서 이상점(outlier) SNP를 식별하는 단계; 및

상기 엘리트 예측 병원성 SNP에 상기 이상점 SNP를 축적시키는 단계를 포함하는, 상기 사이클의 제2 위상에서 처리하는 단계;

제1 사이클에서 상기 제1 위상 및 제2 위상 처리를 수행하는 단계로서,

엘리트 예측 병원성 SNP가 없이 시작하는 단계; 및

상기 제2 위상 처리에서 상기 엘리트 예측 병원성 SNP로서 축적된 식별된 이상점 SNP에서 종료되는 단계를 포함하는, 상기 제1 위상 및 제2 위상 처리를 수행하는 단계;

제2 사이클 및 후속의 사이클에서 상기 제1 위상 및 제2 위상 처리를 반복하는 단계로서,

이전의 사이클의 상기 제2 위상 처리에서 이상점으로서 식별된 상기 엘리트 예측 병원성 SNP에서 시작되는 단계; 및

상기 엘리트 예측 병원성 SNP에 추가로 식별된 이상점 SNP를 축적하고 종료되는 단계를 포함하는, 상기 제1 위상 및 제2 위상 처리를 반복하는 단계; 및

최종 사이클이 될 때까지 상기 사이클들을 반복하는 단계로서, 상기 최종 사이클 후에, 상기 엘리트 예측 병원성 SNP의 축적된 수는 상기 공통 양성 트레이닝 예의 수의 미리 결정된 범위 이내인, 상기 사이클들을 반복하는 단계를 포함하는, 컴퓨터-구현 방법.

청구항 2

제1항에 있어서, 상기 CNN의 앙상블은 각각 심층 컨볼루션 신경망(deep convolutional neural network: DCNN)인, 컴퓨터-구현 방법.

청구항 3

제2항에 있어서, 상기 앙상블은 4개 내지 16개의 DCNN을 포함하는, 컴퓨터-구현 방법.

청구항 4

제2항에 있어서, 상기 앙상블은 8개의 DCNN을 포함하는, 컴퓨터-구현 방법.

청구항 5

제2항에 있어서, 상기 사이클들 동안 에포크(epoch)로 상기 DCNN의 앙상블을 트레이닝하는 단계를 더 포함하는, 컴퓨터-구현 방법.

청구항 6

제2항에 있어서, 상기 점수를 사용하여 상기 DCNN의 앙상블로부터의 점수를 합산함으로써 현재 사이클의 이상점 SNP를 선택하는 단계를 더 포함하는, 컴퓨터-구현 방법.

청구항 7

제2항에 있어서, 상기 점수를 사용하여 상기 DCNN들의 앙상블에 의해 점수 매겨진 SNP들의 각각에 대해서 최대 평균값을 취함으로써 현재 사이클의 이상점 SNP들을 선택하는 단계를 더 포함하는, 컴퓨터-구현 방법.

청구항 8

제1항에 있어서, 현재 사이클 동안 교체 없는 샘플링은 현재 사이클 동안 의사-병원성 SNP의 상호배반적 별개의 트레이닝 세트를 야기시키는, 컴퓨터-구현 방법.

청구항 9

제1항에 있어서, 종료 조건은 미리 결정된 사이클 수인, 컴퓨터-구현 방법.

청구항 10

제9항에 있어서, 상기 미리 결정된 사이클 수는 21인, 컴퓨터-구현 방법.

청구항 11

제1항에 있어서, 상기 파라미터는 적어도 컨볼루션 필터 가중치 및 학습률인, 컴퓨터-구현 방법.

청구항 12

제1항에 있어서, 병원성 분류자로서 사용하기 위하여 상기 앙상블의 상기 CNN들 중 하나를 선택하는 단계를 더 포함하는, 컴퓨터-구현 방법.

청구항 13

제12항에 있어서, 선택된 CNN은 검증 샘플에 대하여 상기 앙상블의 다른 CNN의 예측을 뛰어넘은, 컴퓨터-구현 방법.

청구항 14

제1항에 있어서, 상기 사이클들에 걸쳐 완료된 엘리트 예측 병원성 SNP의 세트는 적어도 400000개의 엘리트 예측 병원성 SNP를 갖는, 컴퓨터-구현 방법.

청구항 15

제1항에 있어서, 각 사이클마다, 상기 엘리트 예측 병원성 SNP의 돌연변이율 편향을 방지하도록 상기 SNP의 공통 양성 트레이닝 예와 샘플링된 의사-병원성 SNP 간의 트라이뉴클레오타이드 컨텍스트를 매칭시키는 단계를 더 포함하는, 컴퓨터-구현 방법.

청구항 16

제1항에 있어서, 상기 조합적으로 생성된 SNP 변이체의 세트인 합성 세트로부터의 의사-병원성 SNP의 샘플링은

각각의 연속 사이클에서 5%만큼 감소되는, 컴퓨터-구현 방법.

청구항 17

메모리에 연결된 하나 이상의 프로세서를 포함하는 시스템으로서,

상기 메모리에는, 제1항 내지 제16항 중 어느 한 항에 따른, 병원성 단일 뉴클레오타이드 다형성(SNP)을 식별하는 컴퓨터-구현 방법을 위한 컴퓨터 명령어가 로딩되는, 시스템.

청구항 18

제1항 내지 제16항 중 어느 한 항에 따른, 병원성 단일 뉴클레오타이드 다형성(SNP)을 식별하는 컴퓨터-구현 방법을 위한 컴퓨터 프로그램 명령어가 부여된, 비일시적 컴퓨터 판독가능 저장 매체.

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

발명의 설명

기 술 분 야

[0001] **부록**

[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. 논문은 가장 먼저 언급된 참고문헌이다.

[0003] **우선권 출원**

[0004] 본 출원은, 미국 가특허 출원 제62/573,144호(대리인 문서번호 ILLM 1000-1/IP-1611-PRV)(출원일: 2017년 10월 16일, 발명의 명칭: "Training a Deep Pathogenicity Classifier Using Large-Scale Benign Training Data", 발명자: Hong Gao, Kai-How Farh, Laksshman Sundaram 및 Jeremy Francis McRae); 미국 가특허 출원 제62/573,149호(대리인 문서번호 ILLM 1000-2/IP-1612-PRV)(출원일: 2017년 10월 16 , 발명의 명칭: "Pathogenicity Classifier Based On Deep Convolutional Neural Networks (CNNS)", 발명자: Kai-How Farh, Laksshman Sundaram, Samskruthi Reddy Padigepati 및 Jeremy Francis McRae); 미국 가특허 출원 제62/573,153호(대리인 문서번호 ILLM 1000-3/IP-1613-PRV)(출원일: 2017년 10월 16일, 발명의 명칭: "Deep Semi-Supervised Learning that Generates Large-Scale Pathogenic Training Data", 발명자: Hong Gao, Kai-How Farh, Laksshman Sundaram 및 Jeremy Francis McRae); 및 미국 가특허 출원 제62/582,898호(대리인 문서번호 ILLM 1000-4/IP-1618-PRV)(출원일: 2017년 11월 7일, 발명의 명칭: "Pathogenicity Classification of

Genomic Data Using Deep Convolutional Neural Networks (CNNs)", 발명자: Hong Gao, Kai-How Farh 및 Laksshman Sundaram)의 우선권 또는 이점을 주장한다. 이들 가특허 출원은 모든 면에서 본 명세서에 참고로 원용된다.

[0005] **원용 문헌**

[0006] 이하의 것들은, 본 명세서에 그 전체가 기재된 것처럼 모든 면에서 참고로 원용되는 것이다:

[0007] PCT 특허출원번호 PCT/US2018/55840(대리인 문서번호 ILLM 1000-8/IP-1611-PCT)(출원일: 2018년 10월 15일, 발명의 명칭: "DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS", 발명자: Hong Gao, Kai-How Farh, Laksshman Sundaram 및 Jeremy Francis McRae), 후속하여 PCT 공보 WO____로서 공개됨.

[0008] PCT 특허출원번호 PCT/US2018/55878(대리인 문서번호 ILLM 1000-9/IP-1612-PCT)(출원일: 2018년 10월 15일, 발명의 명칭: "DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION", 발명자: Laksshman Sundaram, Kai-How Farh, Hong Gao, Samskruthi Reddy Padigepati 및 Jeremy Francis McRae), 후속하여 PCT 공보 WO____로서 공개됨.

[0009] 동시에 출원된 미국 정규출원(대리인 문서번호 ILLM 1000-5/IP-1611-US)(발명의 명칭: "DEEP LEARNING-BASED TECHNIQUES FOR TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS", 발명자: Hong Gao, Kai-How Farh, Laksshman Sundaram 및 Jeremy Francis McRae).

[0010] 미국 정규출원(대리인 문서번호 ILLM 1000-6/IP-1612-US)(발명의 명칭: "DEEP CONVOLUTIONAL NEURAL NETWORKS FOR VARIANT CLASSIFICATION", 발명자: Laksshman Sundaram, Kai-How Farh, Hong Gao 및 Jeremy Francis McRae).

[0011] 미국 정규출원(대리인 문서번호 ILLM 1000-7/IP-1613-US)(발명의 명칭: "SEMI-SUPERVISED LEARNING FOR TRAINING AN ENSEMBLE OF DEEP CONVOLUTIONAL NEURAL NETWORKS", 발명자: Laksshman Sundaram, Kai-How Farh, Hong Gao 및 Jeremy Francis McRae).

[0012] 문헌 1 - A. V. D. Oord, 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;

[0013] 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;

[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(skip connection), 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 convolution)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 연결(residual connection)을 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] SupplementaryTable1.txt 2018년 10월 2일 13 KB

[0041] SupplementaryTable2.txt 2018년 10월 2일 13 KB

[0042] SupplementaryTable3.txt 2018년 10월 2일 11 KB

[0043] SupplementaryTable4.txt 2018년 10월 2일 13 KB

[0044] SupplementaryTable6.txt 2018년 10월 2일 12 KB

[0045] SupplementaryTable7.txt 2018년 10월 2일 44 KB

[0046] SupplementaryTable13.txt 2018년 10월 2일 119 KB

[0047] SupplementaryTable18.txt 2018년 10월 2일 35 KB

[0048] SupplementaryTable20.txt 2018년 10월 2일 1027 KB

[0049] SupplementaryTable20Summary.txt 2018년 10월 2일 9 KB

[0050] SupplementaryTable21.txt 2018년 10월 2일 24 KB

[0051] SupplementaryTable21.txt 2018년 10월 2일 24 KB

[0052] SupplementaryTable18.txt 2018년 10월 4일 35 KB

[0053] DataFileS1.txt 2018년 10월 4일 138 MB

[0054] DataFileS2.txt 2018년 10월 4일 980 MB

[0055] DataFileS3.txt 2018년 10월 4일 1.01 MB

[0056] DataFileS4.txt 2018년 10월 4일 834 KB

[0057] Pathogenicity_prediction_model.txt 2018년 10월 4일 8.24 KB

[0058] 보충 표 1: 분석에 사용된 각각의 종의 변이체의 세부사항. 표에는 이러한 각 데이터 소스에 대한 파이프라인의 중간 결과가 포함된다. 이 표는 SupplementaryTable1.txt에 제공된다는 점에 주목한다.

[0059] 보충 표 2: 일반적인 인간 대립유전자 빈도로 다른 종에 존재하는 미스센스 변이체의 고갈(depletion). 인간과 다른 종 간에 IBS(identical-by-state)인 변이체를 사용하여 희귀 변이체(<0.1%)와 비교되는 흔한 변이체(>0.1%)에서의 미스센스:동의 비율(missense:synonymous ratio)을 기초로 고갈을 계산하였다. 이 표는 SupplementaryTable2.txt에 제공된다는 점에 주목한다.

[0060] 보충 표 3: 인간과 다른 포유동물 간의 >50% 평균 뉴클레오타이드 보존을 갖는 유전자로만 제한된, 일반적인 인간 대립유전자 빈도로 다른 종에 존재하는 미스센스 변이체의 고갈. 인간과 다른 종 간에 IBS인 변이체를 사용하여 희귀 변이체(<0.1%)와 비교되는 흔한, 즉, 공통(common) 변이체(>0.1%)에서의 미스센스:동의 비율을 기초로 고갈을 계산하였다. 이 표는 SupplementaryTable3.txt에 제공된다는 점에 주목한다.

[0061] 보충 표 4: 흔한 인간 대립유전자 빈도로 관련 종 쌍에서의 고정된 대체물로서 존재하는 미스센스 변이체의 고갈. 인간과 관련된 종 쌍 간에 IBS인 변이체를 사용하여 희귀 변이체(<0.1%)와 비교되는 공통 변이체(>0.1%)에서의 미스센스:동의 비율을 기초로 고갈을 계산하였다. 이 표는 SupplementaryTable4.txt에 제공된다는 점에 주목한다.

[0062] 보충 표 6: SCN2A 유전자의 도메인 특이적 주석. 윌콕슨 순위 합계 p-값은, 전체 단백질과 비교하여 특정 도메인에서의 PrimateAI 점수의 분기를 나타낸다. 굵게 강조 표시된 도메인은, 단백질의 약 7%를 차지하지만, 대부분 ClinVar 병원성 주석을 갖는다. 이것은, 도메인에 대한 평균 PrimateAI 점수와 잘 상관되며, PrimateAI 모델에 따라 상위 3개의 병원성 도메인이다. 이 표는 SupplementaryTable6.txt에 제공된다는 점에 주목한다.

[0063] 보충 표 7: 예상 미스센스:동의 비율에 대한 대립유전자 빈도의 영향을 계산하는 데 사용되는 미처리 계수치(count). 돌연변이율과 유전자 변환을 제어하도록 트라이뉴클레오타이드 컨텍스트를 사용하여, 인트론 영역의 변이체에 기초하여 동의 및 미스센스 변이체의 예상 계수치를 계산하였다. 이 표는 SupplementaryTables.xlsx에 제공된다는 점에 주목한다.

[0064] 보충 표 13: 3-상태 이차 구조(3-state secondary structure) 및 3-상태 용매 접근성 예측(3-state solvent accessibility prediction)을 위한 심층 학습 모델을 트레이닝하는 데 사용되는 단백질 데이터뱅크(PDB)로부터

의 단백질 이름의 리스트. 표의 열은, 단백질이 모델 트레이닝의 트레이닝/검증/테스트 단계에서 사용되는지의 여부를 나타낸다. 이 표는 SupplementaryTable13.txt에 제공된다는 점에 주목한다.

[0065] 보충 표 18: 단백질 절단 변이체(p<0.05)로부터만 계산된, DDD 연구에서의 질병 연관에 대해 명목상 유의한 605 개 유전자의 리스트. 이 표는 SupplementaryTable18.txt에 제공된다는 점에 주목한다.

[0066] 보충 표 20: 적어도 하나의 관찰된 DNM을 갖는 모든 유전자에 대해 유전자 당 드 노보 돌연변이(de novo mutation: DNM)의 농축(enrichment)에 대한 테스트 결과. 모든 DNM을 포함할 때 및 PrimateAI 점수가 0.803 미만인 미스센스 DNM을 제거한 후, P-값을 제공한다. FDR 보정된 P-값도 유사하게 제공한다. DDD 코호트 및 전체 메타 분석 코호트로부터 관찰된 단백질 절단(PTV) 및 미스센스 DNM의 계수치를 포함한다. 먼저 모든 미스센스 DNM을 포함할 때 그리고 두 번째로 PrimateAI 점수가 0.803 미만인 모든 미스센스 DNM을 제거한 후에 관측 및 예상 미스센스 DNM의 유사 계수치도 포함한다. 이 표는 SupplementaryTable20.txt 및 SupplementaryTable20Summary.txt에 제공된다는 점에 주목한다.

[0067] 보충 표 21: FDR이 <0.1인 유전자의 드 노보 돌연변이의 농축을 테스트한 결과. 관찰된 단백질 절단(PTV) 드 노보 돌연변이의 계수치, 및 다른 단백질-변경 드 노보 돌연변이의 계수치를, 한번은 모든 미스센스 드 노보 돌연변이를 갖고 그리고 한번은 손상된 미스센스 돌연변이만을 갖고서, 포함하고 있다. 모든 미스센스 부위를 포함할 때의 P-값과 낮은 점수의 미스센스 부위를 제외한 후의 P-값을 제공한다. 이 표는 SupplementaryTable21.txt에 제공된다는 점에 주목한다.

[0068] DataFileS1: 다른 종에 존재하는 모든 변이체의 리스트. ClinVar Significance 열에는 사용가능한 충돌하지 않는 ClinVar 주석이 포함된다. 이 표는 DataFileS1.txt에 제공된다는 점에 주목한다.

[0069] DataFileS2: 관련 종 쌍으로부터의 모든 고정된 치환물의 리스트. 이 표는 DataFileS2.txt에 제공된다는 점에 주목한다.

[0070] DataFileS3: 영장류가 갖는 보류된 양성 테스트 변이체 IBS의 리스트. 양성 테스트 변이체는 1개 이상의 영장류 종을 갖는 IBS인 비-인간 변이체이다. 이 표는 DataFileS3.txt에 제공된다는 점에 주목한다.

[0071] DataFileS4: 보류된 양성 테스트 변이체와 일치하는 영장류가 갖는 비표지 변이체 IBS의 리스트. 표지되지 않은 변이체는, 돌연변이율, 커버리지 편향(bias), 및 영장류 종과의 정렬성에 대하여 양성 테스트 변이체와 매칭된다. 이 표는 DataFileS4.txt에 제공된다는 점에 주목한다.

[0072] Pathogenicity_prediction_model: 일 구현예에 따라 개시된 기술을 가능하게 하는 파이썬(Python) 프로그래밍 언어의 코드. 이 코드 파일은 Pathogenicity_prediction_model.txt에 제공된다는 점에 주목한다.

[0073] 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 에뮬레이션을 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.

배 경 기 술

[0074] 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.

[0075] **기계 학습**

[0076] 기계 학습에서, 입력 변수는 출력 변수를 예측하는 데 사용된다. 입력 변수는, 종종 피처(feature)라고 하며, $X = (X_1, X_2, ..., X_k)$로 표현되며, 여기서 각 $X_i, i \in 1, ..., k$가 피처이다. 출력 변수는, 종종 응답 또는 종속 변수라고 칭하며, Y_i로 표현된다. Y와 대응 X 간의 관계는 다음과 같이 일반적으로 형태로 표현될 수 있다:

[0077] $$Y = f(X) + \in$$

[0078] 위 수학식에서, f는 피처 $(X_1, X_2, \ldots, X_k)$의 함수이고, ϵ는 랜덤 에러 항이다. 에러 항은 X와는 독립적이며 제로인 평균값을 갖는다.

[0079] 실제로, 피처 X는, Y를 갖지 않고서 또는 X와 Y 간의 정확한 관계를 몰라도 이용 가능하다. 에러 항은 제로인 평균값을 가지므로, 목적은 f를 추정하는 것이다.

[0080] $$\hat{Y} = \hat{f} = (X)$$

[0081] 위 수학식에서, $\hat{f}$는 ϵ의 추정이고, 이는, 종종 블랙 박스라고 간주되며, $\hat{f}$의 출력과 입력 간의 관계만이 알려져 있지만 왜 그렇게 기능하는지에 대한 답은 없음을 의미한다.

[0082] 함수 $\hat{f}$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 $\hat{f}$는 출력에 근접하도록 최적화된다. 비감독 학습에서는, 목적이 지표 없는 데이터로부터 숨겨진 구조를 찾는 것이다. 이 알고리즘은 입력 데이터의 정확도를 측정하지 않으므로, 감독 학습과 구별된다.

[0083] **신경망**

[0084] 도 1a는 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 피처 검출 뉴런의 다수의 층을 포함한다. 각 층은, 이전 층으로부터의 입력의 상이한 조합에 응답하는 많은 뉴런을 갖는다. 이들 층은, 제1 층이 입력 화상 데이터에서 프리미티브 패턴들의 세트를 검출하고 제2 층이 패턴 중 패턴을 검출하고 제3 층이 그러한 패턴 중 패턴을 검출하도록 구성된다.

[0085] 유전체학에서의 심층 학습의 적용에 대한 조사는 이하의 간행물에서 찾을 수 있다:

- T. Ching et al., Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org:142760, 2017;
- Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878;
- Park Y, Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825–826. (doi:10.1038/nbt.3313);
- Min, S., Lee, B. & Yoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016);
- Leung MK, Delong A, Alipanahi B et al. Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016; 및
- Libbrecht MW, Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32.

[0086]

도면의 간단한 설명

[0087] 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 1a는 다수의 층을 갖는 피드포워드 신경망(feed-forward neural network)의 일 구현예를 도시한다.

도 1b는 컨볼루션 신경망의 동작의 일 구현예를 도시한다.

도 1c는 개시된 기술의 일 구현예에 따라 컨볼루션 신경망을 트레이닝하는 블록도를 도시한다.

도 1d는 개시된 기술의 일 구현예에 따라 서브샘플링층(평균/최대 풀링)의 일 구현예이다.

도 1e는 개시된 기술의 일 구현예에 따라 ReLU 비선형 층의 일 구현예를 도시한다.

도 1f는 컨볼루션층의 2-층 컨볼루션의 일 구현예를 도시한다.

도 1g는 피처 맵 추가를 통해 다운스트림으로 이전 정보를 재주입하는 잔여 연결을 도시한다.

도 1h는 잔여 블록과 스킵 연결의 일 구현예를 도시한다.

도 1i는 일괄 정규화 순방향 패스(Forward Pass)를 도시한다.

도 1j는 테스트 시간에서의 일괄 정규화 변환을 도시한다.

도 1k는 일괄 정규화 역방향 패스를 도시한다.

도 1l은 컨볼루셔널 또는 밀집 연결층(densely connected layer)의 전후에 일괄 정규화층의 사용을 도시한다.

도 1m은 1D 컨볼루션의 일 구현예를 도시한다.

도 1n은 글로벌 평균 풀링(global average pooling: GAP)의 동작 방식을 도시한다.

도 1o는 확장 컨볼루션을 도시한다.

도 1p는 적층된 확장 컨볼루션의 일 구현예를 도시한다.

도 1q는 개시된 기술이 동작될 수 있는 예시적인 연산 환경을 도시한다.

도 2는, 본 명세서에서 "PrimateAI"라고 하는, 병원성 예측을 위한 심층 신경망의 예시적인 아키텍처를 도시한다.

도 3은 병원성 분류를 위한 심층 학습망 아키텍처인 PrimateAI의 개략도를 도시한다.

도 4a, 도 4b 및 도 4c는 병원성 예측 심층 학습 모델 PrimateAI의 예시적인 모델 아키텍처 세부사항을 도시하는 보충 표 16이다.

도 5 및 도 6은 단백질의 이차 구조 및 용매 접근성을 예측하는 데 사용되는 심층 학습망 아키텍처를 도시한다.

도 7a 및 도 7b는 3-상태 이차 구조 예측 심층 학습(DL) 모델에 대한 예시적인 모델 아키텍처 세부사항을 도시하는 보충 표 11이다.

도 8a 및 도 8b는 3-상태 용매 접근성 예측 심층 학습 모델에 대한 예시적인 모델 아키텍처 세부사항을 도시하는 보충 표 12이다.

도 9는 양성 및 병원성 변이체로부터 참조 단백질 서열 및 대체 단백질 서열을 생성하는 일 구현예를 도시한다.

도 10은 참조 단백질 서열 및 대체 단백질 서열을 정렬하는 일 구현예를 도시한다.

도 11은 위치 가중치 행렬(position weight matrix: PWM) 또는 위치-특정 점수매김 행렬(position-specific scoring matrix: PSSM)이라고도 불리는 위치 빈도 행렬(position frequency matrix: PFM)을 생성하는 일 구현예이다.

도 12, 도 13, 도 14 및 도 15는 이차 구조 및 용매 접근성 서브네트워크의 처리를 도시한다.

도 16은 변이체 병원성 분류자(variant pathogenicity classifier)의 동작을 도시한다. 본 명세서에서 사용되는 바와 같이, 변이체라는 용어는, 또한, 단일 뉴클레오타이드 다형성(또는 단일 염기 다형성)(single nucleotide polymorphism: SNP)을 가리키며 일반적으로는 단일 뉴클레오타이드 변이체(single nucleotide variant: SNV)를 가리킨다.

도 17은 잔여 블록을 도시한다.

도 18은 이차 구조 및 용매 접근성 서브네트워크의 신경망 아키텍처를 도시한다.

도 19는 변이체 병원성 분류자의 신경망 구조를 도시한다.

도 20은 주요 기능 도메인에 대해 주석이 달린 SCN2A 유전자의 각 아미노산 위치에서의 예측 병원성 점수를 도시한다.

도 21d는 트레이닝으로부터 보류된 10,000개의 공통 영장류 변이체의 테스트 세트에 대한 양성 결과를 예측할 때의 분류자들의 비교를 도시한다.

도 21e는, 상응하는 윌콕슨 순위-합 P 값을 갖는, 영향을 받지 않는 형제와 비교되는 해독 발달 장애(Deciphering Developmental Disorders: DDD) 환자에게서 발생하는 드 노보(de novo) 미스센스 변이체에 대한 PrimateAI 예측 점수의 분포를 도시한다.

도 21f는, DDD 케이스 대 대조군에서 드 노보 미스센스 변이체를 분리할 때의 분류자들의 비교를 도시한다. 윌콕슨 순위 합 테스트 P 값은 각 분류자에 대해 표시된다.

도 22a는, 드 노보 단백질 절단 변이체(P<0.05)에 대해 유의한 605개의 연관된 유전자 내의 DDD 코호트로부터 영향을 받은 개체에서 기대 이상의 드 노보 미스센스 돌연변이의 농축을 도시한다.

도 22b는, DDD 환자에게서 발생하는 드 노보 미스센스 변이체 대 605개의 연관된 유전자 내의 영향을 받지 않은 형제에 대한 PrimateAI 예측 점수의 분포를 상응하는 윌콕슨 순위 합 P 값과 함께 도시한다.

도 22c는, 605개 유전자 내의 대조군 대 케이스의 드 노보 미스센스 변이체를 분리할 때의 다양한 분류자의 비교를 도시한다.

도 22d는 각각의 분류자에 대해 표시된 곡선하 면적(area under curve: AUC)을 갖는 수신자 조작 특성 곡선에 도시된 다양한 분류자의 비교를 도시한다.

도 22e는 각 분류자에 대한 분류 정확도 및 곡선하 면적(AUC)을 도시한다.

도 23a, 도 23b, 도 23c 및 도 23d는 분류 정확도에 대한 트레이닝에 사용되는 데이터의 영향을 도시한다.

도 24는 공통 영장류 변이체의 확인에 대한 시퀀싱 커버리지의 영향에 대한 보정을 도시한다.

도 25a, 도 25b, 도 25c 및 도 26은 개시된 신경망에 의한 단백질 모티프의 인식을 도시한다. 도 26은 변이체에 대한 예측 심층 학습 점수에 대한 변이체의 내외부 각각의 위치를 교란시키는 영향을 도시하는 선 도표를 포함한다.

도 27은 가중치 모방 BLOSUM62 및 그랜덤(Grantham) 점수 행렬의 상관 패턴을 도시한다.

도 28a, 도 28b 및 도 28c는 심층 학습망 PrimateAI 및 다른 분류자의 성능 평가를 도시한다.

도 29a 및 도 29b는 4개의 분류자의 예측 점수의 분포를 도시한다.

도 30a, 도 30b 및 도 30c는, 605개의 질병 연관 유전자에서 병원성 변이체와 양성 변이체를 분리할 때 PrimateAI 망 및 다른 분류자의 정확도를 비교한다.

도 31a 및 도 31b는 인간 전문가에 의해 선별된 ClinVar 변이체에 대한 분류자 성능과 경험적 데이터 세트에 대한 성능 간의 상관을 도시한다.

도 32는, 단백질 데이터뱅크로부터의 주석이 달린 샘플에 대한 3-상태 이차 구조 및 3-상태 용매 접근성 예측 모델의 성능을 도시하는 보충 표 14이다.

도 33은, DSSP 데이터베이스로부터의 인간 단백질의 주석이 달린 이차 구조 표지를 사용하는 심층 학습망의 성능 비교를 도시하는 보충 표 15이다.

도 34는, 10,000개의 보류된 영장류 변이체에 대한 정확도 값 및 DDD 케이스의 드 노보 변이체에 대한 p 값 대본 발명자들이 평가한 20개의 분류자 각각에 대한 대조군을 도시하는 보충 표 17이다.

도 35는, 605개의 질환 연관 유전자로 제한되는 DOD 케이스의 드 노보 변이체의 상이한 분류자들 대 대조군 데이터세트의 성능의 비교를 도시하는 보충 표 19이다.

도 36은 개시된 반감독 학습자(semi-supervised learner)의 연산 환경을 도시한다.

도 37, 도 38, 도 39, 도 40 및 도 41은 개시된 반감독 학습의 다양한 사이클을 도시한다.

도 42는 반복적 균형맞춤된 샘플링 프로세스의 예시이다.

도 43은 양성 데이터세트를 생성하는 데 사용되는 연산 환경의 일 구현예를 도시한다.

도 44는 양성 인간 미스센스 SNP를 생성하는 일 구현예를 도시한다.

도 45는 인간 이종상동성 미스센스(human orthologous missense) SNP의 일 구현예를 도시한다. 비인간 종의 미스센스 SNP는 인간과 매칭되는 참조 코돈 및 대체 코돈을 갖는다.

도 46은 인간과 매칭되는 코돈을 갖는 비인간 영장류 종(예를 들어, 침팬지)의 SNP를 양성으로 분류하는 일 구현예를 도시한다.

도 47은 농축 점수를 계산하고 이를 비교하는 일 구현예를 도시한다.

도 48은 양성 SNP 데이터세트의 일 구현예를 도시한다.

도 49a, 도 49b, 도 49c, 도 49d 및 도 49e는 인간 대립유전자 빈도 스펙트럼에 걸친 미스센스/동의 비율을 도시한다.

도 50a, 도 50b, 도 50c 및 도 50d는 다른 종과 IBS인 미스센스 변이체에 대한 선별 선택을 도시한다.

도 51은 선별 선택이 없을 때 인간 대립유전자 빈도 스펙트럼에 걸쳐 예상되는 미스센스:동의 비율을 도시한다.

도 52a, 도 52b, 도 52c 및 도 52d는 CpG 및 비-CpG 변이체에 대한 미스센스:동의 비율을 도시한다.

도 53, 도 54 및 도 55는 6종의 영장류와 IBS인 인간 변이체의 미스센스:동의 비율을 도시한다.

도 56은 조사된 인간 코호트의 크기를 증가시킴으로써 발견된 새로운 공통 미스센스 변이체의 포화를 도시하는 시뮬레이션이다.

도 57은 게놈에서 상이한 보존 프로파일에 걸친 PrimateAI의 정확성을 도시한다.

도 58은 비인간 영장류에 존재하는 변이체 및 공통 인간 변이체로부터 표지된 양성 트레이닝 데이터세트에 대한 기여를 도시하는 보충 표 5이다.

도 59는 예상 미스센스:동의 비율에 대한 대립유전자 빈도의 영향을 도시하는 보충 표 8이다.

도 60은 ClinVar 분석을 도시하는 보충 표 9이다.

도 61은 일 구현예에 따라 ClinVar에서 발견된 다른 종으로부터의 미스센스 변이체의 수를 도시하는 보충 표 10이다.

도 62는 지적 장애가 있는 14개의 추가 후보 유전자의 발견의 일 구현예를 도시하는 표 1이다.

도 63은 ClinVar에서의 병원성 및 양성 변이체 간의 그랜덤 점수의 평균 차의 일 구현예를 도시하는 표 2이다.

도 64는 유전자당 농축 분석(per-gene enrichment analysis)의 일 구현예를 도시한다.

도 65는 게놈 전체 농축 분석(genome-wide enrichment analysis)의 일 구현예를 도시한다.

도 66은 개시된 기술을 구현하는 데 사용될 수 있는 컴퓨터 시스템의 단순화된 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0088] 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.

[0089] **도입부**

[0090] **컨볼루션 신경망**

[0091] 컨볼루션 신경망은 특수한 유형의 신경망이다. 조밀하게 연결된 층과 컨볼루션층 간의 근본적인 차이점은 다음과 같은데, 즉, 조밀하게 연결된 층은 입력 피처 공간에서 글로벌 패턴을 학습하는 반면, 컨볼루션층은 로컬 패

턴을 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.

[0092] 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.

[0093] 두 번째와 관련하여, 제1 컨볼루션층은 에지와 같은 작은 국소 패턴을 학습할 수 있고, 제2 컨볼루션층은 제1 컨볼루션층의 피처로 이루어진 큰 패턴 등을 학습한다. 이를 통해 컨볼루션 신경망이 점점 더 복잡해지고 추상적인 시각적 개념을 효율적으로 학습할 수 있다.

[0094] 컨볼루션 신경망은, 다른 많은 층에 배치된 인공 뉴런 층들을 그 층들을 종속시키는 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 피처들의 세트로부터 입력을 수신한다. 컨볼루션 신경망은, 동일한 피처 맵의 뉴런이 동일한 가중치를 가질 수 있기 때문에 동시에 학습한다. 이러한 국소 공유 가중치는, 다차원 입력 데이터가 컨볼루션 신경망에 진입할 때 컨볼루션 신경망이 피처 추출 및 회귀 또는 분류 프로세스에서 데이터 재구성의 복잡성을 피하도록 신경망의 복잡성을 감소시킨다.

[0095] 컨볼루션은, 2개의 공간 축(높이 및 폭)과 깊이 축(채널 축이라고도 함)을 갖는 피처 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 피처 맵으로부터 패치를 추출하고 이러한 패치 모두에 동일한 변환을 적용하여, 출력 피처 맵을 생성한다. 이러한 출력 피처 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.

[0096] 예를 들어, 제1 컨볼루션층은, 크기(28, 28, 1)의 피처 맵을 취하고 크기(26, 26, 32)의 피처 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 피처 맵이라는 용어의 의미이며, 깊이 축의 모든 치수는 피처(또는 필터)이며, 2D 텐서 출력([:, :, n])은 입력에 대한 이러한 필터의 응답의 2D 공간 맵이다.

[0097] 컨볼루션은, 두 개의 주요 파라미터에 의해 정의되는데, 즉, (1) 입력으로부터 추출된 패치의 크기 - 이들은 통상적으로 1×1, 3×3 또는 5×5이고, (2) 출력 피처 맵의 깊이 - 필터의 수는 컨볼루션에 의해 연산된다. 종종, 이들 컨볼루션은, 깊이 32에서 시작하여, 깊이 64로 계속되며, 깊이 128 또는 256에서 종료된다.

[0098] 컨볼루션은, 3D 입력 피처 맵 위로 3×3 또는 5×5 크기의 이들 윈도우를 슬라이딩하고, 모든 위치에서 정지하고, 주변 피처의 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 피처 맵의 모든 공간 위치는 입력 피처 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 1b에 상세히 설명되어 있다.

[0099] 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, 킨벌루션 층은, 내적 $W \cdot x + b$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 하

며, 즉, 중요한 패턴이 입력에 존재하는 경우, 컨볼루션 필터는 시퀀스의 위치에 관계없이 그 패턴을 학습한다.

[0100] **컨볼루션 신경망의 트레이닝**

[0101] 도 1c는 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.

[0102] 컨볼루션 신경망은, 실측 자료와 실제 출력 간의 차이에 기초하는 뉴런들 간의 가중치를 조정함으로써 트레이닝된다. 이것은 수학적으로 다음과 같이 설명된다:

[0103] $$\Delta w_i = x_i \delta$$

[0104] 여기서 δ = (실측 자료) - (실제 출력)

[0105] 일 구현예에서, 트레이닝 규칙은 다음과 같이 정의된다:

[0106] $$w_{nm} \leftarrow w_{nm} + \alpha(t_m - \varphi_m)a_n$$

[0107] 위 수학식에서, 화살표는 값의 업데이트를 나타내고, t_m은 뉴런 M의 목표 값이고, φ_m은 뉴런 m의 연산된 현재 출력이고, a_n은 입력 n이고, α는 학습률이다.

[0108] 트레이닝의 중간 단계는, 컨볼루션층을 사용하여 입력 데이터로부터 피처 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.

[0109] 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:

[0110] $$\varphi = f(h) = \frac{1}{1+e^{-h}}$$

[0111] 위 시그모이드 함수에서, h는 뉴런에 연산된 가중 합이다. 시그모이드 함수는 이하의 도함수를 갖는다:

[0112] $$\frac{\partial \varphi}{\partial h} = \varphi(1-\varphi)$$

[0113] 알고리즘은, 망의 모든 뉴런의 활성화를 연산하여, 순방향 패스를 위한 출력을 생성하는 것을 포함한다. 숨겨진 층의 뉴런 m의 활성화는 다음과 같이 기술된다:

[0114] $$\varphi_m = \frac{1}{1+e^{-hm}}$$

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

[0115] 이것은 아래와 같이 기술되는 활성화를 얻도록 모든 숨겨진 층에 대하여 행해진다:

$$\varphi_k = \frac{1}{1+e^{hk}}$$

[0116] $$h_k = \sum_{m=1}^{M} \varphi_m v_{mk}$$

[0117] 이어서, 층당 에러와 보정된 가중치를 계산한다. 출력에서의 에러는 다음과 같이 연산된다:

[0118] $$\delta_{ok} = (t_k - \varphi_k)\varphi_k(1-\varphi_k)$$

[0119] 숨겨진 층의 가중치는 다음과 같이 계산된다:

[0120] $$\delta_{hm} = \varphi_m(1-\varphi_m)\sum_{k=1}^{K} v_{mk}\delta_{ok}$$

[0121] 출력층의 가중치는 다음과 같이 업데이트된다:

[0122] $$v_{mk} \leftarrow v_{mk} + \alpha\delta_{ok}\varphi_m$$

[0123] 숨겨진 층의 가중치는 다음과 같이 학습률 α를 사용하여 업데이트된다:

[0124] $$v_{nm} \leftarrow w_{nm} + \alpha\delta_{hm}a_n$$

[0125] 일 구현예에서, 컨볼루션 신경망은, 그라디언트 하강 최적화를 이용하여 모든 층에 걸쳐 에러를 연산한다. 이러한 최적화에 있어서, 입력 피처 벡터 x와 예측 출력 $\hat{y}$에 대하여, 손실 함수는, 표적이 y인 경우, $\hat{y}$를 예측하는 비용에 대하여 l로서, 즉, $l(\hat{y}, y)$로서 정의된다. 예측 출력 $\hat{y}$은, 함수 f를 사용하여 입력 피처 벡터 x로부터 변환된다. 함수 f는 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, z는 입력 및 출력 데이터 쌍 (x, y)이다. 그라디언트 하강 최적화는 이하의 식에 따라 가중치를 업데이트함으로써 수행된다:

$$v_{t+1} = \mu v_t - \alpha \frac{1}{n}\sum_{i=1}^{N} \nabla w_t Q(z_t, w_t)$$

[0126] $$w_{t+1} = w_t + v_{t+1}$$

[0127] 위 수학식에서, α는 학습률이다. 또한, 손실은 데이터 쌍들의 세트에 대한 평균으로서 연산된다. 연산은, 선형 수렴시 학습률 α가 충분히 작을 때 종료된다. 다른 구현예에서, 그라디언트는, 연산 효율을 주입하도록 네스테로브(Nesterov)의 가속 그라디언트 및 적응형 그라디언트에 공급되는 선택된 데이터 쌍만을 사용하여 계산된다.

[0128] 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_t로부터만 연산함으로써, 손실 함수의 가중치에 대하여 그라디언트를 근사화한다:

$$v_{t+1} = \mu V - \alpha \nabla w Q(z_t, w_t)$$

[0129] $$w_{t+1} = w_t + v_{t+1}$$

[0130] 위 수학식에서, α는 학습률이고, μ는 모멘텀이고, t는 업데이트 전의 현재 가중 상태이다. SGD의 수렴 속도는, 학습률 α가 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.

[0131] **컨볼루션층**

[0132] 컨볼루션 신경망의 컨볼루션층은 피처 추출기로서 기능한다. 컨볼루션층은, 입력 데이터를 학습하고 계층적 피처로 분해시킬 수 있는 적응형 피처 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 n의 입력 벡터와 길이 m의 커널 g에 대해, f와 g의 컨볼루션 f*g는 다음과 같이 정의된다:

[0133] $$(f*g)(i)=\sum_{j=1}^{m} g(j)\cdot f(i-j+m/2)$$

[0134] 컨볼루션 동작은 입력 이미지 위로 커널을 슬라이딩하는 것을 포함한다. 커널의 각 위치에 대해, 커널과 입력 이미지의 중복 값이 승산되고 결과가 가산된다. 곱들의 합은, 커널이 중심에 있는 입력 이미지의 지점에서의 출력 이미지의 값이다. 많은 커널로부터의 상이한 출력 결과를 피처 맵이라고 한다.

[0135] 일단 컨볼루션층이 트레이닝되면, 이러한 컨볼루션층은 새로운 추론 데이터에 대한 인식 작업을 수행하는 데 적용된다. 컨볼루션층은, 트레이닝 데이터로부터 학습하므로, 명시적 피처 추출을 피하고 트레이닝 데이터로부터 은연중에 학습한다. 컨볼루션층은, 트레이닝 프로세스의 일부로서 결정 및 업데이트되는 컨볼루션 필터 커널 가중치를 사용한다. 컨볼루션층은, 상위 층에서 결합되는 입력의 상이한 피처들을 추출한다. 컨볼루션 신경망은 다양한 수의 컨볼루션층을 사용하며, 각 컨볼루션층은 커널 크기, 보폭, 패딩, 피처 맵의 수 및 가중치 등의 상이한 컨볼빙 파라미터를 갖는다.

[0136] **서브샘플링층**

[0137] 도 1D는 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.

[0138] 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:

[0139] $$\varphi_o=\max(\varphi_1,\varphi_2,\ldots,\varphi_N)$$

[0140] 위 수학식에서, N은 뉴런 세트 내의 요소들의 총수이다.

[0141] 평균 풀링에 있어서, 풀링 뉴런의 출력은, 이하에서 기술되는 바와 같이 입력 뉴런 세트와 함께 상주하는 입력 값들의 평균값이다:

[0142] $$\varphi_o=\frac{1}{N}\sum_{n=1}^{N}\varphi_n$$

[0143] 위 수학식에서, N은 입력 뉴런 세트 내의 요소들의 총수이다.

[0144] 도 1d에서, 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 4

개 값의 최대값은 완전한 정수 출력이다.

[0145] **비선형 층**

[0146] 도 1e는, 개시된 기술의 일 구현예에 따른 비선형 층의 일 구현예를 도시한다. 비선형 층은, 상이한 비선형 트리거 기능을 사용하여 각 숨겨진 층에서의 가능한 피처의 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.

[0147] 수학적으로 ReLU 활성화 함수는 다음과 같이 기술된다:

[0148] $$\varphi(h) = \max(h, 0)$$

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

[0149] 다른 구현예에서, 컨볼루션 신경망은, 다음과 같이 기술되는 연속 비포화 함수인 파워 유닛 활성화 함수를 사용한다:

[0150] $$\varphi(h) = (a + bh)^c$$

[0151] 위 수학식에서, a, b, c는 각각 시프트, 스케일, 및 파워를 제어하는 파라미터이다. 파워 활성화 함수는, c가 홀수이면 x와 y-비대칭 활성화를 생성할 수 있고 c가 짝수이면 y-대칭 활성화를 생성할 수 있다. 일부 구현예에서, 유닛은 비정류 선형 활성화를 생성한다.

[0152] 또 다른 구현예에서, 컨볼루션 신경망은, 이하의 로직 함수에 의해 기술되는 연속 포화 함수인 시그모이드 유닛 활성화 함수를 사용한다:

[0153] $$\varphi(h) = \frac{1}{1 + e^{-\beta h}}$$

[0154] 위 수학식에서, $\beta = 1$이다. 시그모이드 유닛 활성화 함수는, 네거티브 활성화를 생성하지 않으며, y축에 대해서만 비대칭이다.

[0155] **컨볼루션 예**

[0156] 도 1f는 컨볼루션층의 2-층 컨볼루션의 일 구현예를 도시한다. 도 1f에서, 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.

[0157] 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 얕은 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, 삼십(30)개 내지 오십(50)개의 뉴런을 갖는 다섯(5)개, 여섯(6)개 또는 여덟(8)개의 완전히 연결된 층을 갖는다.

[0158] **순방향 패스**

[0159] 피처 맵의 f개의 컨볼루션 코어들의 수에 대한 제k 피처 스냅 및 제l 컨볼루션층의 행 x, 열 y의 뉴런의 출력은

이하의 식에 의해 결정된다:

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

[0160]

[0161] 제k 피처 맵 및 제l 서브샘플층의 행 x, 열 y의 뉴런의 출력은 이하의 식에 의해 결정된다:

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

[0162]

[0163] 제l 출력층의 제i 뉴런의 출력은 이하의 식에 의해 결정된다:

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

[0164]

[0165] **역전파**

[0166] 출력층의 제k 뉴런의 출력 편차는 이하의 식에 의해 결정된다:

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

[0167]

[0168] 출력층의 제k 뉴런의 입력 편차는 이하의 식에 의해 결정된다:

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

[0169]

[0170] 출력층의 제k 뉴런의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

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

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

[0171]

[0172] 숨겨진 층의 제k 뉴런의 출력 편향은 이하의 식에 의해 결정된다:

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

[0173]

[0174] 숨겨진 층의 제k 뉴런의 입력 편향은 이하의 식에 의해 결정된다:

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

[0175]

[0176] 숨겨진 층의 k개 뉴런으로부터의 입력을 수신하는 이전 층의 제m 피처 맵의 행 x, 열 y의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

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

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

[0177]

[0178] 서브샘플층 S의 제m 피처 맵의 행 x, 열 y의 출력 편향은 이하의 식에 의해 결정된다:

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

[0179]

[0180] 서브샘플층 S의 제m 피처 맵의 행 x, 열 y의 입력 편향은 이하의 식에 의해 결정된다:

[0181] $$d(I_{x,y}^{S,m}) = \varphi'(v_k)d(O_{x,y}^{S,m})$$

[0182] 서브샘플층 S와 컨볼루션층 C의 제m 피처 맵의 행 x, 열 y의 가중치 및 편향 편차는 이하의 식에 의해 결정된다:

[0183] $$\Delta W^{S,m} = \sum_{x=0}^{fh}\sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m})O_{x,y}^{C,m}$$

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

[0184] 컨볼루션층 C의 제k 피처 맵의 행 x, 열 y의 출력 편향은 이하의 식에 의해 결정된다:

[0185] $$d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k})W^k$$

[0186] 컨볼루션층 C의 제k 피처 맵의 행 x, 열 y의 입력 편향은 이하의 식에 의해 결정된다:

[0187] $$d(I_{x,y}^{C,k}) = \varphi'(v_k)d(O_{x,y}^{C,k})$$

[0188] 제l 컨볼루션층 C의 제k 피처 맵의 제m 컨볼루션 코어의 행 r, 열 c의 가중치 및 편향 편차는 다음과 같다:

[0189] $$\Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh}\sum_{y=0}^{fw} d(I_{x,y}^{C,k})O_{x+r,y+c}^{l-1,m}$$

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

[0190] **잔여 연결**

[0191] 도 1g는 피처 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 괴롭히는 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 가정한 것이다. 이들의 크기가 다른 경우, 이전 활성화를 목표 형상으로 재구성하는 선형 변환을 사용할 수 있다. 잔여 연결에 대한 추가 정보는, K. He, X. Zhang, S. Ren, and J. Sun, "DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION," arXiv: 1512.03385, 2015에서 찾을 수 있으며, 이 문헌의 전문은 모든 면에서 본 명세서에 그 전체가 개시된 것처럼 참고로 원용된다.

[0192] **잔여 학습 및 스킵 연결**

[0193] 도 1h는 잔여 블록 및 스킵 연결의 일 구현예를 도시한다. 잔여 학습의 주요 아이디어는, 잔여 맵핑이 원래 매핑보다 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 시작 시, 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. 잔여 블록과 스킵 연결에 대한 추가 정보는, A. V. D. Oord, 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, 2016에서 찾을 수 있다.

[0194] 잔여 망으로부터의 이점을 통해, 심층 컨볼루션 신경망(convolutional neural network: CNN)을 쉽게 트레이닝할 수 있고, 이미지 분류 및 오브젝트 검출에 대한 정확도가 개선되었다. 컨볼루션 순방향 망은, 제l 층의 출력을 제(l+1) 층에 입력으로서 연결하며, 이는 이하의 층 천이 $x_l = H_l(x_{l-1})$를 발생시킨다. 잔여 블록은 식별 함수 $x_l = H_l(x_{l-1}) + x_{l-1}$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_l의 출력은 합산에 의해 결합되며, 이는 네트워크의 정보 흐름을 방해할 수 있다.

[0195] **팽창 컨볼루션**

[0196] 도 1o는 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 1o에 도시되어 있다. 입력들 사이의 거리는 팽창 인자에 의존한다.

[0197] **웨이브넷**

[0198] 웨이브넷은 원시 오디오 파형을 생성하기 위한 심층 신경망이다. 웨이브넷은, 저렴한 비용으로 비교적 큰 '시야'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.

[0199] 웨이브넷의 주요 빌딩 블록은 인과적 팽창 컨볼루션이다. 인과적 팽창 컨볼루션에 대한 확장으로서, 웨이브넷은, 도 1p에 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.

[0200]

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

[0201] 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.

[0202] **일괄 정규화**

[0203] 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 기능한다. 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. 일괄 정규화에 대한 추가 정보는, S. Ioffe and C. Szegedy, "BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT, " arXiv: 1502.03167, 2015에서

찾을 수 있으며, 이 문헌의 전문은 그 전체가 개시된 것처럼 본 명세서에 참고로 원용된다.

[0204] 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 구현예는 개시된 기술에 의해 사용될 수 있으며 도 1l에 도시되어 잇다. 일괄정규화층은 축 인수를 사용하며, 이러한 축 인수는 정규화되어야 하는 피처 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 틈새 사용의 경우에, 피처 축은 축 1이고, 일괄정규화에서의 축 인수는 1로 설정될 수 있다.

[0205] 일괄 정규화는, 입력을 피드포워딩하고 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 피처 맵의 상이한 요소- 즉, 활성화 -가 동일한 방식으로 정규화된다. 따라서, 미니-일괄의 모든 활성화는, 활성화마다 정규화되기보다는 모든 위치에서 정규화된다.

[0206] 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.

[0207] 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.

[0208] 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 피처 분포를 평균 0과 표준 편차 1을 갖도록 정규화한다.

[0209] **순방향 패스**

[0210] 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 1i에 도시되어 있다.

[0211] 도 1i에서, 각각 μ_β는 일괄 평균이고 σ_β^2는 일괄 분산이다. 학습된 스케일 및 시프트 파라미터는 각각 γ 및 β로 표시된다. 명확성을 위해, 일괄 정규화 절차는 활성화마다 본 명세서에 기재되고 상응하는 지수를 생략한다.

[0212] 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 1j에 도시된 바와 같이 표현될 수 있다. 도 1j에서, μ_D와 σ_D^2는 각각 일괄 통계라기보다는 모집단 평균과 분산을 나타낸다.

[0213] **역방향 패스**

[0214] 정규화는 미분가능한 동작이므로, 역방향 통과는 도 1k에 도시된 바와 같이 연산될 수 있다.

[0215] **1D 컨볼루션**

[0216] 1D 컨볼루션은, 도 1m에 도시한 바와 같이, 서열로부터 로컬 1D 패치 또는 서브시퀀스를 추출한다. 1D 컨볼루션

은 입력 서열의 시간 패치로부터 각 출력 타임스텝을 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.

[0217] **글로벌 평균 풀링**

[0218] 도 1n은 글로벌 평균 풀링(GAP)의 동작 방식을 도시한다. 점수 매김을 위해 마지막 층의 피치들의 공간 평균을 취함으로써 분류를 위해 완전히 연결된(fully connected: 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 폴링은 파라미터의 수를 감소시키고 완전히 연결된 층을 제거한다. 완전히 연결된 층들은 통상적으로 가장 파라미터가 많고 연결이 많은 층이며, 글로벌 평균 풀링은 비슷한 결과를 얻기 위해 훨씬 저렴한 비용의 접근 방식을 제공한다. 글로벌 평균 풀링의 주요 아이디어는, 각 마지막 층 피처 맵으로부터 평균값을 점수 매김을 위한 신뢰인자로서 생성하여 소프트맥스층에 직접 공급하는 것이다.

[0219] 글로벌 평균 풀링은 3가지 이점을 갖는데, 즉, (1) 글로벌 평균 풀링층에 추가 파라미터가 없으므로, 글로벌 평균 풀링층에서 과적합을 피하고, (2) 글로벌 평균 풀링의 출력이 전체 피처 맵의 평균이므로, 굴로벌 평균 풀링이 공간 변환에 더 강력하고, (3) 전체 망의 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.

[0220] 글로벌 평균 풀링은, 마지막 층에서 더 강한 피처가 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 피처 맵은, 신뢰 맵으로서 해석될 수 있으며, 피처 맵과 카테고리 간의 대응을 강제할 수 있다. 글로벌 평균 풀링은, 마지막 층 피처가 직접 분류에 대한 충분한 추상에 있는 경우 특히 효과적일 수 있지만, 멀티레벨 피처 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 분류자를 추가함으로써 가장 잘 수행된다.

[0221] **유전체학에서의 심층 학습**

[0222] 유전자 변이체는 많은 질환을 설명하는 데 도움이 될 수 있다. 모든 인간에게는 고유한 유전자 코드가 있으며, 개인 그룹 내에는 많은 유전자 변이체가 있다. 유해한 유전자 변이체의 대부분은 자연적인 선택에 의해 게놈으로부터 고갈되었다. 어떠한 유전자 변이체가 병원성이거나 유해할 가능성이 있는지를 식별하는 것이 중요하다. 이는, 연구자들이 유사 병원성 유전자 변이체에 집중하고 많은 질환의 진단 및 치료 속도를 가속화하는 데 도움을 줄 수 있다.

[0223] 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.

[0224] 지난 수십 년간 생화학 기술의 발전은, 그 어느 때보다 훨씬 적은 비용으로 게놈 데이터를 신속하게 생성하는 차세대 서열분석(NGS) 플랫폼으로 부상하였다. 이러한 압도적으로 많은 양의 서열분석된 DNA는 주석을 달기 어렵다. 감독되는 기계 학습 알고리즘은, 통상적으로 많은 양의 지표 있는 데이터를 사용할 수 있을 때 성능이 우수하다. 생물정보학 및 다른 많은 데이터가 풍부한 분야에서, 인스턴스를 지표 부착하는 프로세스는 비용이 많이 들지만, 지표가 없는 인스턴스는 저렴하고 쉽게 사용할 수 있다. 지표 있는 데이터의 양이 상대적으로 적고 지표가 없는 데이터의 양이 상당히 많은 상황에서는, 반감독 학습이 수동적 지표 부착에 대한 비용 효율적인 대안을 나타낸다.

[0225] 변이체의 병원성을 정확하게 예측하는 심층 학습 기반 병원성 분류자를 구성하기 위해 반감독 알고리즘을 사용할 기회가 발생한다. 인간 확인 편향이 없는 병원성 변이체의 데이터베이스가 생성될 수 있다.

[0226] 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.

[0227] 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 응급 심층 신경망이 제안되어 왔다.

[0228] 심층 신경망을 트레이닝하는 목적은 각 층의 가중치 파라미터를 최적화하는 것이며, 이는 간단한 피처들을 복잡한 피처들로 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.

[0229] 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.

[0230] 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.

[0231] 따라서, 변이체의 병원성을 예측하기 위한 강력한 연산 모델은 기본 과학 및 변환 연구 모두에 대하여 많은 이점을 가질 수 있다.

[0232] 일반적인 다형성은 자연적 선택의 세대에 의해 적합성이 테스트된 자연적 실험을 나타낸다. 인간 미스센스 및 동의 치환에 대하여 대립유전자 빈도 분포를 비교하여, 비인간 영장류 종에서 높은 대립유전자 빈도로 존재하는 미스센스 변이체는 그 변이체가 또한 인간 모집단에서 중립적 선택 하에 있음을 확실하게 예측한다는 것을 발견하였다. 대조적으로, 더 먼 종의 공통 변이체는, 진화 거리가 증가함에 따라 음성적 선택을 경험한다.

[0233] 본 발명자들은, 서열만을 사용하여 임상적 드 노보 미스센스 돌연변이를 정확하게 분류하는 반감독 심층 학습망을 트레이닝하도록 6종의 비인간 영장류 종으로부터의 공통 변이를 사용한다. 500종이 넘는 알려진 종을 이용하여, 영장류 계통은, 알려지지 않은 중요성을 가진 대부분의 인간 변이체의 영향을 체계적으로 모델링하는 데 충분한 공통 변이를 포함한다.

[0234] 인간 참조 게놈은, 7천만 개를 초과하는 잠재적 단백질-변경 미스센스 치환을 보유하며, 이들 대부분은 인간 건

강에 대한 영향이 특성화되지 않은 희귀한 돌연변이이다. 알려지지 않은 중요성의 이러한 변이체는, 임상 응용 분야에서 게놈 해석에 대한 도전을 제시하고, 인구 전체 스크리닝 및 개별화된 의약에 대한 서열분석의 장기적 채택의 장애물이다.

[0235] 다양한 인간 모집단에 걸쳐 공통 변이를 분류하는 것은 임상적 양성 변이를 식별하기 위한 효과적인 전략이지만, 현대 인간에서 이용 가능한 공통 변이는 인간 종의 먼 과거의 병목 현상에 의해 제한된다. 인간과 침팬지는 99%의 서열 동일성을 공유하는데, 이는 침팬지 변이체에 대하여 기능하는 자연적 선택이 인간에서 IBS인 변이체의 영향을 모델링할 가능성이 있음을 시사한다. 인간 모집단에서 중립적 다형성에 대한 평균 유착 시간은 종의 발산 시간의 일부이므로, 자연적으로 발생하는 침팬지 변이체는, 선택의 균형을 유지함으로써 유지되는 희소한 일배체형(haplotype)의 인스턴스를 제외하고는 인간 변이체와 중복되지 않는 돌연변이 공간을 대체로 탐색한다.

[0236] 60,706명의 인간으로부터의 응집된 엑솜 데이터의 최근 이용 가능성은, 미스센스 돌연변이 및 동의 돌연변이(synonymous mutation)에 대한 대립유전자 빈도 스펙트럼을 비교함으로써 이 가설을 테스트할 수 있게 한다. ExAC의 싱글톤 변이체(singleton variants)는 트라이뉴클레오타이드 컨텍스트를 사용하여 돌연변이율을 조정한 후 드 노보 돌연변이에 의해 예측된 예상 2.2:1 미스센스:동의 비율과 거의 일치하지만, 높은 대립유전자 빈도로, 관찰된 미스센스 변이체의 수는 자연적 선택에 의한 유해한 변이체의 필터링으로 인해 감소된다. 대립유전자 빈도 스펙트럼에 걸친 미스센스:동의 비율의 패턴은, 모집단 빈도가 0.1% 미만인 미스센스 변이체의 많은 부분이 약간 유해함을 나타내며, 즉, 모집단으로부터 즉각적인 제거를 보장할 만큼 병원성이 없고, 더욱 제한된 모집단 데이터에 대한 사전 관찰과 일치하는, 고 대립유전자 빈도로 존재할 수 있을 만큼 중립적이지 않음을 나타낸다. 이러한 연구 결과는, 선택 및 창시자 효과의 균형 맞춤에 의해 발생하는 소수의 잘 기록된 예외를 제외하고는, 침투성 유전자 질환에 대해 양성일 가능성이 있는 0.1% 내지 1%를 초과하는 대립유전자 빈도를 갖는 변이체를 필러링 제거하는 진단 실험실에 의한 광범위한 경험적 실습을 지원한다.

[0237] 공통 침팬지 변이체(침팬지 모집단 서열분석에서 두 번 이상 관찰됨)와 IBS인 인간 변이체의 서브세트로 이러한 분석을 반복하여, 미스센스:동의 비율이 대립유전자 빈도 스펙트럼에 걸쳐 대체로 일정하다는 것을 발견하였다. 침팬지 모집단에서 이러한 변이체의 높은 대립유전자 빈도는, 침팬지에서 자연적 선택의 체를 이미 통과했으며, 인간 모집단의 적합성에 대하여 해당하는 중립적 영향이 미스센스 변이체에 대한 선택적 압력이 두 종에서 매우 일치한다는 강력한 증거를 제공한다는 것을 나타낸다. 침팬지에서 관찰된 낮은 미스센스:동의 비율은, 조상 침팬지 모집단에서 더욱 큰 유효 모집단 크기와 일치하여, 침팬지 모집단이 약간 유해한 변이체를 더욱 효율적으로 필터링할 수 있게 한다.

[0238] 이와 대조적으로, 희귀 침팬지 변이체(침팬지 모집단 서열분석에서 한 번만 관찰됨)는, 높은 대립유전자 빈도에서 미스센스:동의 비율이 완만하게 감소됨을 나타낸다. 인간 변이 데이터로부터 동일한 크기의 코호트를 시뮬레이션함으로써, 이 크기의 코호트에서 한 번 관찰된 변이체의 64%만이, 코호트에서 여러 번 관찰된 변이체에 대한 99.8%에 비해 일반적인 모집단에서 0.1%보다 큰 대립유전자 빈도를 가질 것으로 추정되며, 이는 희귀 침팬지 변이체 모두가 선택 체를 통과한 것은 아님을 나타낸다. 전체적으로, 확인된 침팬지 미스센스 변이체의 16%가 일반 모집단에서의 대립유전자 빈도가 0.1% 미만이고 높은 대립유전자 빈도에서 음성 선택의 대상이 될 것으로 추정된다.

[0239] 다음에, 다른 비인간 영장류 종(보노보(Bonobo), 고릴라, 오랑우탄, 레서스(Rhesus) 및 마모셋)에서 관찰된 변이와 IBS인 인간 변이체를 특성화한다. 침팬지와 유사하게, 대립유전자 빈도에서의 미스센스 변동이 약간 고갈되는 것이 아니라 미스센스:동의 비율이 대립유전자 빈도 스펙트럼에 걸쳐 거의 동일함을 관찰하며, 이는 소수의 희귀 변이체(~5% 내지 15%)가 포함되어 있기 때문이라고 예상된다. 이러한 결과는, 미스센스 변이체에 대한 선택적 힘이 영장류 계통 내에서 적어도 3천 5백만년 전 인간 조상 계통에서 벗어난 것으로 추정되는 광비원류(new world monkey)와 대략 일치한다는 것을 암시한다.

[0240] 다른 영장류의 변이체와 IBS인 인간 미스센스 변이체는 ClinVar에서의 양성 결과를 위해 강하게 농축된다. 알려지지 않은 또는 충돌하는 주석이 있는 변이체를 배제한 후, 영장류 오솔로그(ortholog)가 있는 인간 변이체는, 일반적으로 미스센스 변이의 45%에 비해 ClinVar에서 양성 또는 유사 양성으로서 주석이 달릴 가능성이 약 95%인 것으로 관찰된다. 비인간 영장류로부터 병원성으로서 분류된 ClinVar 변이체의 작은 분획물은, 유사한 크기의 건강한 인간의 코호트로부터 희귀 변이체를 확인함으로써 관찰되는 병원성 ClinVar 변이체의 분획물과 비교할 수 있다. 병원성 또는 유사 병원성으로서 주석이 달린 이러한 변이체의 상당 부분은, 큰 대립유전자 빈도 데이터베이스가 출현하기 전에 해당 분류를 받았으며 오늘날 다르게 분류될 수 있음을 나타낸다.

[0241] 인간 유전체학 분야는 인간 돌연변이의 임상적 영향을 추론하기 위해 모델 유기체에 오랫동안 의존해 왔지만, 대부분의 유전적으로 다루기 쉬운 동물 모델까지의 긴 진화 거리는 이러한 발견이 인간에게 다시 일반화될 수 있는 정도에 대한 우려를 일으킨다. 인간과 더 먼 종에서의 미스센스 변이체에 대한 자연적 선택의 일치성을 조사하기 위해, 본 발명자들은 영장류 계통을 넘어 분석을 확장하여 4종의 추가 포유류 종(마우스, 돼지, 염소, 소)과 먼 종의 척추동물인 2개 종(닭, 제브라피시)으로부터의 대체로 일반적인 변이를 포함시키고 있다. 이전 영장류 분석과는 대조적으로, 희귀 대립유전자 빈도에 비해 흔한 대립유전자 빈도로, 특히, 더욱 큰 진화 거리에서 미스센스 변이체가 현저히 고갈됨을 관찰하였으며, 이는 더욱 먼 종의 일반적인 미스센스 변이체의 상당 분획물이 인간 모집단에서 음성 선택을 겪는다는 것을 나타낸다. 그럼에도 불구하고, 더욱 먼 척추동물에서의 미스센스 변이체의 관찰은, 자연적 선택에 의해 고갈된 일반적인 미스센스 변이체의 분획물이 베이스라인에서의 인간 미스센스 변이체에 대해 ~50% 고갈보다 훨씬 작기 때문에, 여전히 양성 결과의 가능성을 증가시킨다. 이러한 결과와 일관되게, 마우스, 개, 돼지 및 소에서 관찰된 인간 미스센스 변이체가, 전체적으로 영장류 변이의 경우 95% 및 ClinVar 데이터베이스의 경우 45%에 비해, ClinVar에서 양성 또는 유사 양성으로 주석 표시될 가능성이 약 85%인 것으로 나타났다.

[0242] 다양한 진화 거리에서 밀접하게 관련된 종들의 쌍의 존재는, 또한, 인간 모집단에서의 고정된 미스센스 치환의 기능적 결과를 평가할 기회를 제공한다. 포유류 가계도에서 밀접하게 관련된 종들의 쌍(가지 길이<0.1) 내에서, 고정된 미스센스 변동이 희귀 대립유전자 빈도와 비교하여 흔한 대립유전자 빈도로 고갈되어 있음을 관찰하였으며, 이는 종간 고정된 치환의 상당 부분이 영장류 계통 내에서도 인간에게서 비중립적임을 나타낸다. 미스센스 고갈의 크기의 비교는, 종간 고정된 치환이 종내 다형성보다 중립성이 현저히 낮다는 것을 나타낸다. 흥미롭게도, 밀접하게 관련된 포유류 간의 종간 변이는 종내의 공통 다형성에 비해 ClinVar에서 실질적으로 더 병원성이 아니고(83%는 양성 또는 유사 양성으로 주석 표시될 가능성이 있음), 이는 이러한 변화가 단백질 기능을 파괴하지 않고 오히려 특정 종 적응형 장점을 부여하는 단백질 기능의 조정을 반영함을 시사한다.

[0243] 임상 적용을 위한 알려지지 않은 유의미한 다수의 가능한 변이체 및 정확한 변이체 분류의 결정적 중요성은 기계 학습의 문제를 해결하기 위한 다수의 시도에 영감을 주었지만, 이러한 노력은 불충분한 양의 공통 인간 변이체 및 큐레이션 데이터베이스에서의 불확실한 주석 품질에 의해 크게 제한되어 왔다. 비인간 영장류 6종의 변이는, 공통 인간 변이와 중첩되지 않는 대략 양성 결과로 되는 300,000개를 초과하는 고유한 미스센스 변이체에 기여하며, 기계 학습 접근법에 사용될 수 있는 트레이닝 데이터세트의 크기를 크게 확대한다.

[0244] 다수의 인간 공학 피처 및 메타 분류자를 사용하는 초기 모델과는 달리, 본 발명에서는, 관심 변이체 옆에 있는 아미노산 서열 및 다른 종에서의 이종상동성 서열 정렬만을 입력으로서 취하는 간단한 심층 학습 잔여망을 적용한다. 단백질 구조에 관한 정보를 망에 제공하기 위해, 2개의 개별적인 망을 트레이닝하여 서열 단독으로부터 이차 구조 및 용매 접근성을 학습하고, 단백질 구조에 대한 영향을 예측하기 위해 이들을 더욱 심층의 학습망에 하위망으로서 통합한다. 서열을 출발점으로서 사용함으로써, 불완전하게 확인되거나 불일치하게 적용될 수 있는, 단백질 구조 및 기능적 도메인 주석에서의 잠재적 편향을 피한다.

[0245] 본 발명에서는, 망들의 앙상블을 초기에 트레이닝하여 돌연변이율 및 서열분석 커버리지와 일치하는 랜덤한 알려지지 않은 변이체 대 양성의 영장류 변이체를 분리함으로써 양성 표지를 갖는 변이체만을 포함하는 트레이닝 세트의 문제를 극복하도록 반감독 학습을 사용한다. 이러한 망들의 앙상블은, 병원성으로 더 예측되는 결과를 갖는 알려지지 않은 변이체를 향하여 바이어싱하고 각 반복 시 점진적인 단계를 취하여 모델이 준최적화된 결과로 미리 수렴하는 것을 방지함으로써 분류자의 다음 반복을 시딩(seed)하기 위해 미지 변이체들의 완전한 세트를 점수 매기고 선택에 영향을 주도록 사용된다.

[0246] 공통 영장류 변이는, 또한, 메타-분류자의 증식으로 인해 객관적으로 평가하기 어려웠던 이전에 사용된 트레이닝 데이터와는 완전히 독립적인 기존 방법을 평가하기 위한 깨끗한 검증 데이터세트를 제공한다. 10,000개의 보류된 영장류 공통 변이체를 사용하여 4개의 다른 분류 알고리즘(Sift, Polyphen-2, CADD, M-CAP)을 이용하여 본 발명자들의 모델의 성능을 평가하였다. 모든 인간 미스센스 변이체의 대략 50%가 흔한 대립유전자 빈도로 자연적 선택에 의해 제거되기 때문에, 본 발명자들은 돌연변이율에 의해 10,000개의 보류된 영장류 공통 변이체와 일치한 랜덤하게 선택된 미스센스 변이체들의 세트에서의 각 분류자에 대한 50-백분위수 점수를 계산하였으며, 임계값을 사용하여 보류된 영장류 공통 변이체를 평가하였다. 본 발명의 심층 학습 모델의 정확도는, 인간 공통 변이체에 대해서만 트레이닝된 심층 학습망을 사용하거나 인간 공통 변이체와 영장류 변이체를 모두 사용하여, 이러한 독립적 검증 데이터세트에서의 다른 분류자보다 훨씬 우수하였다.

[0247] 최근의 트리오 서열분석 연구에서는, 신경발달 장애 환자 및 건강한 형제자매에서의 수천 개의 드 노보 돌연변

이를 목록화하였으며, 사례와 대조군에서 드 노보 미스센스 돌연변이를 분리하는 데 있어서 다양한 분류 알고리즘의 강도를 평가할 수 있게 한다. 4가지 분류 알고리즘 각각에 대해, 사례 대 대조군에서 각각의 드 노보 미스센스 변이체를 점수화하고, 두 분포 간의 차이에 대한 윌콕슨 순위 합 테스트의 p-값을 보고하였으며, 영장류 변이체($p \sim 10^{-33}$)에 대해 트레이닝된 심층 학습 방법이 이러한 임상 시나리오에서의 다른 분류자($p \sim 10^{-13}$ 내지 10^{-19})보다 훨씬 더 잘 수행되었음을 나타내었다. 이러한 코호트에 대해 보고된 기대치보다 드 노보 미스센스 변이체의 ~1.3배 농축 및 미스센스 변이체의 ~20%가 기능 손실 효과를 생성하는 이전 추정으로부터, 본 발명자들은, 완벽한 분류자가 $p \sim 10^{-40}$의 p 값을 갖는 두 개 클래스로 분리할 것으로 예상하는 바, 이는 본 발명자들의 분류자가 여전히 개선의 여지가 있음을 나타낸다.

[0248] 심층 학습 분류자의 정확도는 트레이닝 데이터세트의 크기에 따라 스케일링되고, 6종의 영장류 종의 각각으로부터의 변이 데이터는 분류자의 정확도를 높이는 데 독립적으로 기여한다. 현존하는 비인간 영장류 종의 수와 다양성은, 단백질 변경 변이체에 대한 선택적 압력이 영장류 계통 내에서 대부분 일치한다는 증거와 함께, 체계적 영장류 모집단 서열분석을, 임상 게놈 해석을 현재 제한하는 알려지지 않은 중요성의 수백만 개의 인간 변이체를 분류하기 위한 효과적인 전략으로서 시사한다. 504개의 알려진 비인간 영장류 종 중에서, 사냥과 서식지 손실로 인해 약 60%가 멸종 위기에 처해 있으며, 이렇게 독특한 대체불가한 종과 우리 자신에게 이익이 되는 전세계적 보존 노력에 대한 시급한 동기를 부여한다.

[0249] 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 계수치 대 예상 계수치를 계산할 수 있었다. 전반적으로, 본 발명자들은 엑손-인트론 경계로부터 >50nt의 거리에서 크립틱 스플라이스 돌연변이에서의 60% 고갈을 관찰한다. 감쇠된 신호는, 엑소좀과 비교하여 더 작은 표본 크기와 전체 게놈 데이터의 조합일 수 있으며, 심층적 인트로 변이체의 영향을 예측하기가 더 어려울 수 있다.

[0250] **용어**

[0251] 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 원용된다. 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.

[0252] 본 명세서에서 사용되는 바와 같이, 하기 용어들은 지시된 의미를 갖는다.

[0253] 염기는 뉴클레오타이드 염기 또는 뉴클레오타이드, A(아데닌), C(사이토신), T(티민) 또는 G(구아닌)를 가리킨다.

[0254] 본 출원은 "단백질" 및 "번역된 서열"이라는 용어를 호환 가능하게 사용한다.

[0255] 본 출원은 "코돈" 및 "염기 트리플렛"이라는 용어를 호환 가능하게 사용한다.

[0256] 본 출원은 "아미노산" 및 "번역된 유닛"이라는 용어를 호환 가능하게 사용한다.

[0257] 본 출원은 "변이체 병원성 분류자", "변이체 분류를 위한 컨볼루션 신경망-기반 분류자", "변이체 분류를 위한 심층 컨볼루션 신경망 기반 분류자"라는 구를 호환 가능하게 사용한다.

[0258] "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.

[0259] "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.

[0260] 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, 염소, 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.

[0261] "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, 즉, 판독물(read)은 이들 350개 뉴클레오타이드 내에 다수의 하위서열을 포함할 수 있다. 예를 들어, 샘플 리드는, 예를 들어, 20개 내지 50개의 뉴클레오타이드를 갖는 제1 및 제2 플랭킹 하위서열(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 일부)를 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, 표적 서열, 프라이머 서열, 측면 서열, 참조 서열 등)가 제공될 수 있다. "대립유전자"와 같은 다른 용어에는 유사한 대상을 구별하도록 다른 표지가 부여될 수 있다.

[0262] "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 WO07010252, PCT 출원 일련번호 PCTGB2007/003798, 및 미국 특허출원 공개공보 US 2009/0088327에 기재되어 있으며, 이들 각각은 본 명세서에 참고로 원용된다. 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.

[0263] "참조 게놈" 또는 "참조 서열"이라는 용어는, 대상으로부터 확인된 서열을 참조하는 데 사용될 수 있는, 부분적인지 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 10^5배 이상, 또는 적어도 약 10^6배 이상, 또는 적어도 약 10^7배 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.

[0264] "리드"라는 용어는, 뉴클레오타이드 샘플 또는 참조의 분획물을 기술하는 서열 데이터의 수집을 지칭한다. "리드"이라는 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.

[0265] 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 토런트 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.

[0266] "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.

[0267] "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"(raw)라는 용어는, 원시 분획물이 샘플 리드의 잠재적 변이체에 대응하고 이러한 잠재적 변이체를 인증 또는 확인하는 변이체를 나타내는지의 여부에 관계없이, 원시 분획물이 샘플 리드에서 시열 데이터와 일부 관계가 있는 서열 데이터를 포함한다는 것을 나타내는 데 사용된다. "원시 분획물"이라는 용어는, 분획물이 반드시 샘플 리드에서 변이체 콜(variant call)을 유효성 확인하는 지지 변이체를 포함한다는 것을 나타내지는 않는다. 예를 들어, 제1 변이체를 나타내기 위해 변이체 콜 애플리케이션에 의해 샘플 리드가 결정될 때, 변이체 콜 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.

[0268] "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 멤버쉽 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.

[0269] "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).

[0270] "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 콜은, DNA 샘플에서 낮은 빈도로 존재하는 변이체를 식별하기 위한 노력이다. 체세포 변이체 콜은 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.

[0271] "노이즈"라는 용어는, 서열분석 프로세스 및/또는 변이체 콜 애플리케이션에서의 하나 이상의 에러로 인한 잘못된 변이체 콜을 지칭한다.

[0272] "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.

[0273] 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 측정한 것이다.

[0274] "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.

[0275] "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.

[0276] 본 명세서에서 "임계값"이라는 용어는, 샘플, 핵산, 또는 그 일부(예를 들어, 리드)를 특성화하도록 컷오프로서 사용되는 숫자 또는 비숫자 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.

[0277] 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 계수치를 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 계수치의 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.

[0278] "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.

[0279] "리드 깊이"(통상적으로 "×"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×이고 목표 염기의 98%가 >100×를 커버한다고 말할 수 있다.

[0280] "염기 콜 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 콜은, 신뢰도 P-값이 0.01인 경우 올바른 것으로 간주될 수 있다. Q<20인 모든 염기 콜은 품질이 낮은 것으로 간주되어야 하며, 변이체를 지지하는 서열분석된 리드의 상당 부분이 품질이 낮은 것으로 식별된 임의의 변이체는 잠재적 위양성으로 간주되어야 한다.

[0281] "변이체 리드" 또는 "변이체 리드 수"라는 용어는 변이체의 존재를 지지하는 서열분석된 리드의 수를 지칭한다.

[0282] **서열분석 프로세스**

[0283] 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 콜을 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 원용된다.

[0284] 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)를 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.

[0285] 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 기판의 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.

[0286] 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 원용된다.

[0287] 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 제7,115,400호; 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 원용된다. 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 원용된다.

[0288] 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' 블록을 갖는 변형된 뉴클레오타이드를 이용하며, 이들 문헌 각각은 본 명세서에 참고로 원용된다. 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 입스위치에 소재하는 NEB사에 의해 공급되는 바와 같은 'USER', 부품 번호 M5505S), 열이나 알칼리에 대한 노출, 데옥시리보뉴클레오타이드로 달리 구성된 증폭 산물로 혼입된 리보뉴클레오타이드의 절단, 광화학적 절단, 또는 펩타이드 링커의 절단을 포함하여, 효소 또는 닉킹 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.

[0289] 이어서, 유동 세포를, 단일 뉴클레오타이드 첨가에 의해 각각의 목표 핵산에 혼성화된 프라이머를 확장시키는 조건 하에서 제거가능한 3' 블록 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 원용된다. 이어서, 확장된 핵산은 스캔 시약의 존재 하에서 형광 검출될 수 있다. 일단 형광이 검출되었다면, 사용된 블로킹 기에 적합한 디블로킹 시약을 사용하여 3' 블록을 제거할 수 있다. 각 블로킹 기에 유용한 예시적인 디블로킹 시약(deblock reagent)은 WO0004018497, US 2007/0166705 A1, 및 미국 특허번호 7,057,026에 기재되어 있으며, 이들 문헌 각각은 본 명세서에 참고로 원용된다. 디블로킹 시약을 세척하여, 목표 핵산을, 이제 추가의 뉴클레오타이드의 첨가를 위한 성분인 3'-OH기를 갖는 확장된 프라이머에 혼성화되게 한다. 따라서, 하나 이상의 동작 사이에서의 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.

[0290] 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.

[0291] 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 게놈에 걸쳐 이어질 수 있다. 일례로, 샘플 리드값은, STR이 의심되거나 SNP가 의심되는 그러한 유전 좌위와 같은 미리 정해진 유전 좌위에 관한 것이다.

[0292] 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.

[0293] 다음에, 하나 이상의 프로세서는 서열분석 데이터를 분석하여 잠재적 변이체 콜(들) 및 샘플 변이체 콜(들)의 샘플 변이체 빈도를 취득한다. 상기 동작은, 또한, 변이체 콜 애플리케이션 또는 변이체 콜러(variant caller)라고 칭할 수 있다. 따라서, 변이체 콜러는 변이체를 식별 또는 검출하고, 변이체 분류자는 검출된 변이체를 체세포 또는 생식세포로서 분류한다, 대안의 변이체 콜러는 본 발명의 구현예에 따라 이용될 수 있고, 여기서 상이한 변이체 콜러들은, 관심 샘플의 피처 등에 기초하여 수행되는 서열분석 동작의 유형에 기초하여 사용될 수 있다. 변이체 콜 애플리케이션의 비제한적인 일례는, https://github.com/Illumina/Pisces에 호스팅되고 Dunn, Tamsen & Berry, Gwenn & Emig-Agius, Dorothea & Jiang, Yu & Iyer, Anita & Udar, Nitin & Strφmberg, 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 Pisees™이 있으며, 이 문헌의 완전한 주제 전문은 명백하게 본 명세서에 참고로 원용된다.

[0294] 이러한 변이체 콜 애플리케이션은 다음과 같이 4개의 순차적으로 실행되는 모듈을 포함할 수 있다:

[0295] (1) 파이시즈 리드 스티치(Pisces Read Stitcher): BAM(동일한 분자의 리드 1과 리드 2)의 페어드 리드들을 컨센서스 리드로 스티칭함으로써 노이즈를 감소시킨다. 출력은 스티칭된 BAM이다.

[0296] (2) 파이시즈 변이체 콜러(Pisces Variant Caller): 작은 SNV, 삽입, 및 삭제를 콜한다. 파이시즈는, 리드 경계, 기본 필터링 알고리즘, 및 간단한 푸아송 기반 변이체 신뢰도 점수매김 알고리즘에 의해 분해된 변이체들을 병합하는 변이체 허탈 알고리즘을 포함한다. 출력은 VCF이다.

[0297] (3) 파이시즈 변이체 품질 재교정기(Pisces Variant Quality Recalibrator: VQR): 변이체 콜이 열적 손상 또는 FFPE 탈아민에 연관된 패턴을 압도적으로 추종하는 경우, VQR 단계는 의심되는 변이체 콜의 변이체 Q 점수를 다운그레이드한다. 출력은 조정된 VCF이다.

[0298] (4) 파이시즈 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 툴에 의한 기능적 결과를 더욱 정확하게 결정할 수 있다. 출력은 조정된 VCF이다.

[0299] 부가적으로 또는 대안적으로, 동작은, https://github.com/Illumina/strelka에 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 일루미나사에 의한 변이체 콜 애플리케이션 Strelka™을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 원용된다. 게다가, 부가적으로 또는 대안적으로, 동작은, https://github.com/Illumina/strelka에 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 콜 애플리케이션 Strelka2™을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 원용된다. 게다가, 부가적으로 또는 대안적으로, 동작은, https://github.com/Illumina/Nirvana/wiki에 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/콜 툴 Nirvana™을 이용할 수 있으며, 이러한 문헌의 주제 전문은, 명백하게 본 명세서에 참고로 원용된다.

[0300] 이러한 변이체 주석/콜 툴은, 아래와 같이 Nirvana에 개시된 알고리즘 기술 등의 상이한 알고리즘 기술을 적용할 수 있다:

[0301] 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, n은 트리의 간격의 수이고, k는 중복되는 간격의 수이다. 실제로, k는 대부분의 변이체에 대한 n에 비해 실제로 작기 때문에, 간격 트리의 유효 런타임은 O(k lg n)이다. 제1 중복 간격만 찾으면 되고 이어서 나아 있는 (k-1)개를 통해 열거 처리하도록 모든 간격이 정렬된 어레이로 저장되는 간격 어레이를 생성함으로써 O(lg n + k)으로 개선하였다.

[0302] 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 ClinGen과 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.

[0303] 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, 염색체 위치를 보충 주석이 상주하는 파일 위치에 맵핑하기 위해 훨씬 작은 인덱스가 생성된다. 인덱스는, 위치를 사용하여 이진 검색될 수 있는 (염색체 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 n은 데이터베이스의 항목 수이다.

[0304] d. VEP 캐시 파일

[0305] 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 프로파일되었는가?).

[0306] 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.

[0307] 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.

[0308] 이러한 변이체 주석 툴은 전처리를 포함할 수 있다. 예를 들어, Nirvana에는, ExAC, EVS, 1000 게놈 프로젝트,

dbSNP, ClinVar, Cosmic, DGV, 및 ClinGen과 같은 외부 데이터 공급원의 많은 주석이 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 주석의 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.

[0309] 적어도 일부 구현예에 따르면, 변이체 콜 애플리케이션은 저 빈도 변이체, 생식세포 콜 등에 대한 콜을 제공한다. 비제한적인 예로서, 변이체 콜 애플리케이션은 종양 전용 샘플 및/또는 종양-정상 쌍을 이룬 샘플에서 실행될 수 있다. 변이체 콜 애플리케이션은, 단일 뉴클레오타이드 변이(SNV), 다중 뉴클레오타이드 변이(MNV), 인델 등을 검색할 수 있다. 변이체 콜 애플리케이션은, 변이체를 식별하면서 서열분석 또는 샘플 준비 오류로 인한 불일치를 필터링한다. 각각의 변이체에 대해, 변이체 콜러는, 참조 서열, 변이체의 위치 및 잠재적 변이체 서열)(들)(예를 들어, A에서 C SNV로, 또는 AG에서 A 삭제로)을 식별한다. 변이체 콜 애플리케이션은, 샘플 서열(또는 샘플 분획물), 참조 서열/분획물, 및 변이체 콜을 변이체가 존재함을 나타내는 표시로서 식별한다. 변이체 콜 애플리케이션은, 원시 분획물을 식별할 수 있고, 원시 분획물의 지정, 잠재적 변이체 콜을 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.

[0310] 변이체 콜 애플리케이션은, .VCF 또는 .GVCF 파일과 같은 다양한 형식으로 콜을 출력할 수 있다. 단지 예로서, 변이체 콜 애플리케이션은 (예를 들어, MiSeq® 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.

[0311] 이어서, 하나 이상의 프로세서는 잠재적 변이체 콜과 관련하여 유효성확인 동작을 수행한다. 유효성확인 동작은 이하에 설명되는 바와 같이 품질 점수 및/또는 계층적 테스트의 층에 기초할 수 있다. 유효성확인 동작이 잠재적 변이체 콜을 인증하거나 검증하면, 유효성확인 동작은 (변이체 콜 애플리케이션으로부터) 변이체 콜 정보를 샘플 보고서 생성기에 전달한다. 대안으로, 유효성확인 동작이 잠재적 변이체 콜을 무효화 또는 실격화하는 경우, 유효성확인 동작은, 대응하는 표시(예를 들어, 음성 표시기, 무 콜 표시기, 무효 콜 표시기)를 샘플 보고서 생성기에 전달한다. 유효성확인 동작은, 또한, 변이체 콜이 정확하거나 무효 콜 지정이 정확하다는 신뢰도와 관련된 신뢰도 점수를 전달할 수 있다.

[0312] 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 콜을 제공하는 것, 유전자형 콜을 할 수 없음을 나타내는 것, 유전자형 콜의 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 보고서

의 인쇄, 또는 샘플 보고서의 표시를 포함할 수 있다.

[0313] 샘플 보고서는, 변이체 콜이 결정되었지만 유효성확인되지 않았음을 나타낼 수 있다. 변이체 콜이 무효한 것으로 결정되면, 샘플 보고서는 변이체 콜을 유효성확인하지 않는 결정의 근거에 관한 추가 정보를 나타낼 수 있다. 예를 들어, 보고서의 추가 정보는, 원시 분획물의 설명 및 원시 분획물이 변이체 콜을 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.

[0314] **변이체 콜 애플리케이션**

[0315] 본 명세서에 개시된 구현예들은 잠재적 변이체 콜을 식별하기 위해 서열분석 데이터를 분석하는 것을 포함한다. 변이체 콜은 이전에 수행된 서열분석 동작을 위해 저장된 데이터에 대해 수행될 수 있다. 추가적으로 또는 대안으로, 변이체 콜은 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 콜 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.

[0316] 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 사용자에 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.

[0317] 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 콜될 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 콜될 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 콜되는지를 결정하기 위해 후속 분석이 수행될 수 있다.

[0318] 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 콜 또는 잠정 좌위 콜이라고 칭할 수 있으며, 여기서 샘플 리드는 선택 서열과 상관되는 유전 좌위에 대하여 콜된다. 그러나, 전술한 바와 같이, 샘플 리드는 하나보다 많은 유전 좌위에 대하여 콜될 수 있다. 이러한 구현예에서, 잠재적 유전 좌위들 중 하나에 대해서만 샘플 리드를 콜하거나 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 콜된 유전 좌위의 유형에 기초할 수 있다.

[0319] 다음에, 잠재적 변이체 콜을 식별하도록 샘플 리드가 분석된다. 무엇보다도, 분석 결과는, 잠재적 변이체 콜, 샘플 변이체 빈도, 참조 서열, 및 변이체가 발생한 게놈 서열 내의 위치를 식별한다. 예를 들어, 유전 좌위가 SNP를 포함하는 것으로 알려진 경우, 유전 좌위를 콜된 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 원용된다.

[0320] 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 콜러는, 초기 변이체 콜러를 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 콜러는, 또한, 지지 변이체를 나타낸 원시 분획물 수의 계수치뿐만 아니라 지지 변이체가 발생한 원시 분획물 내의 위치도 식별한다. 예를 들어, 변이체 콜러는, 특정 위치 X에서 지지 변이체를 갖는 이중 스티치 분획물을 나타내도록 10개의 원시 분획물이 식별되었다는 표시를 출력할 수 있다. 변이체 콜러는, 또한, 특정 위치 Y에서 지지 변이체를 갖는 단일 언스티치 분획물을 나타내도록 원시 분획물의 5개 리스가 식별되었음을 출력할 수 있다. 변이체 콜러는, 또한, 참조 서열에 대응한 많은 원시 분획물을 출력할 수 있으므로, 다른 경우엔 관심 게놈 서열에서 잠재적 변이체 콜을 유효성확인하는 증거를 제공하는 지지 변이체를 포함하지 않았다.

[0321] 이어서, 지지 변이체가 발생한 위치뿐만 아니라 지지 변이체를 포함하는 원시 분획물의 계수치를 유지한다. 부가적으로 또는 대안적으로, (샘플 리드 또는 샘플 분획물의 잠재적 변이체 콜의 위치에 관한) 관심 위치에서 지지 변이체를 포함하지 않은 원시 분획물의 계수치를 유지할 수 있다. 부가적으로 또는 대안적으로, 참조 서열에 대응하고 잠재적 변이체 콜을 인증 또는 확인하지 않는 원시 분획물의 계수치를 유지할 수 있다. 결정된 정보는, 잠재적 변이체 콜을 지지하는 원시 분획물의 계수치와 유형, 원시 분획물의 지지 분산의 위치, 잠재적 변이체 콜을 지지하지 않는 원시 분획물의 수 등을 포함하여 변이체 콜 유효성확인 애플리케이션으로 출력된다.

[0322] 잠재적 변이체 콜이 식별되면, 프로세스는 잠재적 변이체 콜, 변이체 서열, 변이체 위치, 및 이에 연관된 참조 서열을 나타내는 표시를 출력한다. 변이체 콜은, 에러로 인해 콜 프로세스가 거짓 변이체를 식별할 수 있으므로 "잠재적" 변이체를 나타내도록 지정된다. 본 발명의 구현예에 따라, 잠재적 변이체 콜을 분석하여 거짓 변이체 또는 위양성을 감소 및 제거한다. 부가적으로 또는 대안적으로, 이 프로세스는, 샘플 리드에 연관된 하나 이상의 원시 분획물을 분석하고 원시 분획물에 연관된 해당 변이체 콜을 출력한다.

[0323] **양성(benign) 트레이닝 세트 생성**

[0324] 수백만 개의 인간 게놈과 엑솜이 서열분석되었지만, 이들의 임상 적용은, 질병을 유발하는 돌연변이를 양성 유전적 변이와 구별하기 어렵기 때문에 제한되어 있다. 여기서 다른 영장류 종들의 공통 미스센스 변이체가 인간에 있어서 대체로 임상적으로 양성이라는 것을 입증함으로써, 병원성 돌연변이를 제거 프로세스에 의해 체계적으로 식별할 수 있게 한다. 비인간 영장류 6종의 모집단 서열분석으로부터의 수십만 종의 공통 변이체를 이용하여, 88%의 정확도로 희귀병 환자의 병원성 돌연변이를 식별하고 게놈 전체의 중요도에서 지적 장애가 있는 14개의 새로운 후보 유전자를 발견할 수 있도록 하는 심층 신경망을 트레이닝한다. 추가 영장류 종으로부터 공통 변이를 분류하면, 불확실한 중요성의 수백만 변이체에 대한 해석을 개선하게 되며, 인간 게놈 서열분석의 임상적 효용이 더욱 진전될 것이다.

[0325] 진단 서열분석의 임상적 실행가능성은, 인간 모집단의 희귀한 유전적 변이체를 해석하고 질병 위험에 대한 해당 영향을 추론하는 어려움에 의해 제한된다. 건강 상태에 대한 해로운 영향 때문에, 임상적으로 중요한 유전적 변이체는 모집단에서 극히 드문 경향이 있고, 대부분의 경우, 인간의 건강에 미치는 영향은 결정되지 않았다. 불확실한 임상적 중요성이 있는 이러한 변이체의 많은 수와 희귀성은 개인화된 의학과 인구 전체의 건강 스크리닝

을 위한 서열분석의 채택에 엄청난 장애를 초래한다.

[0326] 대부분의 침투성 멘델리아 질병은 모집단에서의 유병률이 매우 낮기 때문에, 모집단에서 고 빈도로 변이체를 관찰하는 것은 양성 결과에 유리한 강력한 증거가 된다. 다양한 인간 모집단에 걸친 공통 변이를 검정하는 것은 양성 변이체를 분류하는 데 효과적인 전략이지만, 오늘날 인간의 공통 변이의 총량은 조상 다양성의 많은 부분이 손실된 우리 종의 최근 역사에서의 병목현상으로 인해 제한된다. 현재 인간에 대한 모집단 연구는, 지난 15,000년 내지 65,000년 내에 1만 명 미만의 개인으로 구성된 유효 인구 크기(N_e)로부터 주목할 만한 인플레이션을 나타내며, 공통 다형성의 작은 풀은 이러한 크기의 모집단의 변동에 대한 제한된 용량에서 유래된 것이다. 참조 게놈에서 7천만 개를 초과하는 잠재적 단백질 변화 미스센스 치환물 중에서, 1000명 중 대략 1명만이 0.1% 초과의 전체 모집단 대립유전자 빈도로 존재한다.

[0327] 현대의 인간 모집단 밖에서, 침팬지는, 다음으로 가장 가까운 현존하는 종을 포함하며, 99.4%의 아미노산 서열 동일성을 공유한다. 인간과 침팬지의 단백질 부호화 서열의 거의 동일성은, 침팬지 단백질-부호화 변이체에 대하여 동작하는 선별 선택이 또한 IBS인 인간 돌연변이의 적합성에 대한 결과를 모델링할 수 있음을 시사한다.

[0328] 중립 다형성이 조상 인간 계통(~4Ne 세대)을 지속하기 위한 평균 시간은, 종의 다양성 시간(~6백만 년 전)의 일부이기 때문에, 자연 발생 침팬지 변이는, 드물게 나타나는 균형맞춤 선택에 의해 유지되는 일배체형을 제외하고 우연한 경우를 제외하고는 대부분 중복되지 않는 돌연변이 공간을 탐구한다. IBS인 다형성이 두 종의 적합성에 유사하게 영향을 미치는 경우, 침팬지 모집단에서 높은 대립유전자 빈도로 변이체가 존재한다는 것은, 인간에게 양성 결과를 나타내는 것이며, 선별 선택에 의해 양성 결과가 확립된 것으로 알려진 변이체의 카탈로그를 확대한다.

[0329] **결과 - 다른 영장류의 공통 변이체는 인간에 있어서 대략 양성임**

[0330] 엑솜 집계 컨소시엄(ExAC)과 게놈 집계 데이터베이스(gnomAD)에서 수집된 123,136명의 인간을 포함하는 집계된 엑솜 데이터의 최근 가용성을 통해, 대립유전자 빈도 스펙트럼에 걸쳐 자연 선택이 미스센스 및 동의 돌연변이에 미치는 영향을 측정할 수 있다. 코호트에서 단 한 번만 관측되는 희귀 싱글톤 변이체는, 돌연변이율(도 49a, 도 51, 도 52a, 도 52b, 도 52c 및 도 52d)에 대한 트라이뉴클레오타이드 컨텍스트의 영향에 대해 조정한 후 드노보 돌연변이에 의해 예측되는 예상 2.2/1 미스센스/동의 비율과 밀접하게 일치하지만, 더 높은 대립유전자 빈도로 관찰된 미스센스 변이체의 수는, 자연 선택에 의한 유해 돌연변이의 숙청으로 인해 감소한다. 대립유전자 빈도가 증가함에 따라 미스센스/동의 비율의 점진적인 감소는, 건강한 개인에게서 관찰되었음에도 불구하고 약간의 해로운 결과를 갖는 0.1% 미만의 모집단 빈도의 미스센스 변이체의 상당 부분과 일치한다. 이러한 연구 결과는, 균형잡힌 선택과 창시자 효과로 인한 소수의 잘 문서화된 예외를 제외하고 0.1% 내지 ~1% 초과의 대립유전자 빈도로 침투성 유전 질병에 대해 양성일 수 있는 변이체들을 필터링하는 진단 연구소에 의한 광범위한 경험적 실시를 지지한다.

[0331] 본 발명자들은 24명의 무관한 개인의 코호트에서 2번 이상 샘플링된 공통 침팬지 변이체를 식별하였으며; 이들 변이체의 99.8%가 일반적인 침팬지 모집단(대립유전자 빈도(AF)>0.1%)에서 공통적이라고 추정하는데, 이는 이들 변이체가 이미 선별 선택의 체를 통과했음을 나타낸다. 본 발명자들은, 다수의 서열 정렬에 있어서 일대일 맵핑이 결여된 변이체와 함께, 확장된 주요 조직적합성 복합 영역을 균형맞춤 선택의 공지된 영역으로서 제외하고, 대응하는 IBS 인간 변이체(도 49b)에 대한 인간 대립유전자 빈도 스펙트럼을 조사하였다. 공통 침팬지 변이체와 IBS인 인간 변이체의 경우, 미스센스/동의 비율은 인간 대립유전자 빈도 스럼트펙(카이-제곱(χ^2) 테스트에 의한 P>0.5)에서 거의 일정하며, 이는 두 종의 미스센스 변이체에 대한 일치 선택 계수와 인간 모집단의 공통 침팬지 변이체에 대한 음성 선택의 부재와 일치한다. 공통 침팬지 변이체와 IBS인 인간 변이체에서 관찰되는 낮은 미스세스/동의 비율은 침팬지에서 더 큰 유효 모집단 크기(N_e~73,000)와 일치하여, 약간 유해한 변이를 더욱 효율적으로 필터링할 수 있게 한다.

[0332] 이와 대조적으로, 싱글톤 침팬지 변이체(코호트에서 한 번만 샘플링됨)에 대해, 본 발명자들은 흔한 대립유전자 빈도($P<5.8\times10^{-6}$; 도 49c)에서 미스센스/동의 비율의 현저한 감소를 관찰하였으며, 이는 0.1%보다 큰 대립유전자 빈도로 인간 모집단의 선별 선택에 의해 싱글톤 침팬지 미스센스 변이체의 24%가 필터링된다는 것을 나타낸다. 이러한 고갈은, 침팬지 싱글톤 변이체의 상당 부분이 적합성에 대한 유해한 영향이 두 종의 흔한 대립유전자 빈도에 영향을 끼치지 못한 희귀한 유해한 돌연변이라는 것을 나타낸다. 일반적인 침팬지 모집단에서 싱글톤 변이체의 69%만이 흔한(AF>0.1%) 것으로 추정한다.

[0333] 다음에, 본 발명자들은 6종의 비인간 영장류 종 중 적어도 하나에서 관찰된 변이와 IBS인 인간 변이체를 식별하였다. 6종 각각의 변이는 영장류 게놈 프로젝트(침팬지, 보노보, 고릴라 및 오랑우탄)로부터 확인되었거나 영장류 게놈 프로젝트(레수스, 마모셋)로부터 단일 뉴클레오타이드 다형성 데이터베이스(dbSNP)에 제출되었고, 각각의 종에 대해 관찰된 개체의 제한된 수 및 낮은 미스센스:동의 비율에 기초하는 공통 변이체를 주로 나타낸다(보충 표 1). 침팬지와 유사하게, 6종의 비인간 영장류 종으로부터의 변이체에 대한 미스센스/동의 비율이, 일반적인 대립유전자 빈도로 미스센스 변이의 경미한 고갈 이외에, 인간 대립유전자 빈도 스펙트럼에 걸쳐 대략 동일하다는 것을 발견하였으며(도 49d, 도 53, 도 54 및 도 55, 보충 데이터 파일 1), 이는 소수의 희귀 변이체(침팬지에서 0.1% 미만의 대립유전자 빈도로 ~16%, 및 다른 쫑에서는 더 적은 수의 개체로 인해 더 적음)가 포함됨으로 인한 것이라고 예상된다. 이 결과는, IBS 미스센스 변이체에 대한 선택 계수가 영장류 계통에서 적어도 3천 5백만년 전 인간 조상 계통으로부터 벗어난 것으로 추정되는 광비원류와 일치한다는 것을 시사한다.

[0334] 관찰된 영장류 변이체와 IBS인 인간 미스센스 변이체가 ClinVar 데이터베이스에서 양성 결과를 위해 강하게 농축됨을 발견하였다. 불확실한 유의미한 변이체와 상충되는 주석이 있는 변이체를 배제한 후, 적어도 하나의 비인간 영장류 종에 존재하는 ClinVar 변이체는, 일반적으로 ClinVar 미스센스 변이체의 35%에 비해 평균적으로 90%의 양성 또는 양성 가능성으로서 주석 처리된다($P<10^{-40}$; 도 49e). 영장류 변이체에 대한 ClinVar 주석의 병원성은, 큐레이션 편향을 줄이기 위해 대립유전자 빈도가 1%보다 큰 인간 변이체를 제외하고 비슷 한 크기 코호트의 건강한 인간 집단(~95% 양성 또는 양성 가능성 결과, P=0.07)을 샘플링하여 관찰된 것보다 약간 더 크다.

[0335] 인간 유전체학 분야는 인간 돌연변이의 임상적 영향을 추론하기 위해 모델 유기체에 오랫동안 의존해 왔지만, 대부분의 유전적으로 다루기 쉬운 동물 모델까지의 긴 진화 거리는 모델 유기체에 대한 연구 결과가 인간에게 일반화될 수 있는 정도에 대한 우려를 일으킨다. 본 발명자들은 본 발명자들의 분석을 영장류 계통 이상으로 확장하여 4개의 추가 포유류 종(마우스, 돼지, 염소, 및 소)과 2종의 먼 척추동물(닭과 제브라피쉬)의 대략 공통된 변이를 포함시켰다. 게놈 전체에 걸친 dbSNP의 변이에 대하여 충분히 확인된 종을 선택하였고, 2.2/1보다 훨씬 낮은 미스센스/동의 비율에 기초하여 이들이 대체로 일반적인 변이체임을 확인하였다. 본 발명자들의 영장류 분석과는 대조적으로, 더 먼 종의 변이와 IBS인 인간 미스센스 돌연변이는 흔한 대립유전자 빈도로 현저히 고갈되고(도 50a), 이러한 고갈의 크기는 더욱 긴 진화 거리에서 증가한다(도 50b와, 보충 표 2 및 표 3).

[0336] 인간에게는 유해하지만, 더 먼 종의 높은 대립유전자 빈도로 허용되는 미스센스 돌연변이는, IBS 미스센스 돌연변이에 대한 선택 계수가 인간 종보다 먼 종 간에 상당히 분기되었음을 나타낸다. 그럼에도 불구하고, 더욱 먼 포유류에서의 미스센스 변이체의 존재는, 일반적인 대립유전자 빈도로 자연 선택에 의해 고갈된 미스센스 변이체의 분율이 일반적으로 인간 미스센스 변이체에서 관찰되는 ~50% 고갈보다 작기 때문에 여전히 양성 결과의 가능성을 여전히 증가시킨다(도 49a). 이러한 결과와 일치하여, 마우스, 돼지, 염소, 및 소에서 관찰된 ClinVar 미스센스 변이체는 영장류 변이에 대한 90%에 비해 양성 또는 양성 가능성으로 주석이 달릴 가능성이 73%($P<2\times10^{-8}$; 도 50c)이고, ClinVar 데이터베이스 전체에 대해 35%인 것으로 나타났다.

[0337] 사육 아티팩트가 아닌 진화 거리가 선택 계수의 분기를 위한 주요 원동력임을 확인하기 위해, 광범위한 진화 거리의 범위에 걸쳐 종간 다형성 대신 밀접하게 관련된 종 쌍 간의 고정된 치환을 사용하여 분석을 반복하였다(도 50d, 보충 표 4 및 보충 데이터 파일 2). 종간 고정 치환과 IBS인 인간 미스센스 변이체의 고갈이 진화 분지 길이에 따라 증가하며, 사육에 노출된 것에 비해 야생 종에 대한 인식 가능한 차이가 없음을 발견하였다. 이는 파리(fly) 및 효모에서의 초기 연구와 일치하며, 이는 IBS 고정 미스센스 치환의 수가 다른 계통에서 우연히 예상되는 것보다 적다는 것을 발견하였다.

[0338] **변이체 병원성 분류를 위한 심층 학습망**

[0339] 개시된 기술은 변이체 병원성 분류를 위한 심층 학습망을 제공한다. 임상 적용을 위한 변이체 분류의 중요성은 문제를 해결하기 위해 감독된 기계 학습을 사용하려는 수많은 시도에 영감을 주었지만, 이러한 노력은 트레이닝을 위한 확실하게 표지된 양성 및 병원성 변이체를 포함하는 적절한 크기의 트루 데이터세트가 부족함으로 인해 방해를 받았다.

[0340] 인간 전문가 선별 변이체의 기존 데이터베이스는 전체 게놈을 나타내지 않으며, ClinVar 데이터베이스의 변이체의 ~50%는 단지 200개의 유전자(인간 단백질-코딩 유전자의 ~1%)로부터 온 것이다. 게다가, 체계적인 연구는, 많은 인간 전문가 주석이 의심스러운 지지 증거를 갖고 있고, 단 한 명의 환자에서만 관찰될 수 있는 희귀 변이체를 해석하기 어렵다는 점을 강조한다. 인간 전문가의 해석이 점점 더 엄격해졌지만, 분류 지침은 대부분 합의 관행을 중심으로 공식화되었으며 기존 경향을 강화할 위험이 있다. 인간의 해석 편향을 줄이기 위해, 최근 분류

자는 일반적인 인간 다형성 또는 고정된 인간 침팬지 치환에 대해 트레이닝되었지만, 이러한 분류자는 인간이 선별한 데이터베이스에 대해 트레이닝된 이전 분류자의 예측 점수를 해당 입력으로서 또한 사용한다. 이러한 다양한 방법의 성능에 대한 객관적인 벤치마킹은 독립적이고 편향이 없는 트루 데이터세트가 없는 경우에 달성하기 어려웠다.

[0341] 6종의 비인간 영장류(침팬지, 보노보, 고릴라, 오랑우탄, 레서스 및 마모셋)로부터의 변이는, 일반적인 인간 변이와 중복되지 않는 300,000개를 초과하는 고유한 미스센스 변이체에 기여하고, 선별 선택의 체를 거친 양성 결과의 일반적인 변이체를 대략 나타내며, 기계 학습 방법에 사용할 수 있는 트레이닝 데이터 세트를 크게 확대하였다. 평균적으로, 각 영장류 종은 ClinVar 데이터베이스 전체보다 많은 변이체(불확실한 유의미한 변이체 및 주석이 충돌하는 변이체를 배제한 후 2017년 11월 현재 ~42,000개의 미스센스 변이체)를 제공한다. 또한, 이 내용에는 사람의 해석에 대한 편향이 없다.

[0342] 공통 인간 변이체(AF>0.1%) 및 영장류 변이체(보충 표 5(도 58))를 포함하는 데이터세트를 사용하여, 관심 변이체를 플랭킹하는(즉, 관심 변이체에 측접하는) 아미노산 서열 및 다른 종에서의 이종상동성 서열 정렬을 입력으로서 취하는 새로운 심층 잔여망인 PrimateAI를 트레이닝하였다(도 2 및 도 3). 인간 공학 피처를 사용하는 기존 분류자와는 달리, 본 발명의 심층 학습망은 일차 서열로부터 피처를 직접 추출하는 것을 학습한다. 단백질 구조에 관한 정보를 통합하기 위해, 본 발명에서는, 서열 단독으로부터 2차 구조 및 용매 접근성을 예측하도록 별개의 네트워크를 트레이닝한 후 이들을 전체 모델의 서브네트워크로서 포함시켰다(도 5 및 도 6). 성공적으로 결정화된 소수의 인간 단백질이 주어지면, 일차 서열로부터 구조를 추론하는 것은, 불완전한 단백질 구조 및 기능적 도메인 주석으로 인한 편향을 피할 수 있다는 이점이 있다. 단백질 구조가 포함된 망의 전체 깊이는 대략 400,000개의 트레이닝 가능한 파라미터를 포함하는 36개의 컨볼루션층으로 이루어졌다.

[0343] 양성 표지를 갖는 변이체만을 사용하여 분류자를 트레이닝하도록, 본 발명에서는 주어진 돌연변이가 모집단에서 공통 변이체로서 관찰될 가능성이 있는지에 대한 예측 문제의 틀을 잡았다. 몇 개의 인자는 유해성에만 관심을 두고 있는 변이체를 대립유전자 빈도로 관찰할 확률에 영향을 미치며, 다른 인자로는, 돌연변이율, 서열분석 커버리지와 같은 기술적 아티팩트, 및 유전자 전환과 같은 중립적 유전적 드리프트에 영향을 미치는 인자가 있다.

[0344] 본 발명에서는, 양성 트레이닝 세트의 각 변이체를 ExAC 데이터베이스로부터의 123,136개의 엑솜에 존재하지 않은 미스센스 돌연변이와 일치시키고, 이들 각각에 대한 혼란 인자를 제어하고, 양성 변이체와 일치하는 대조군을 구별하도록 심층 학습망을 트레이닝하였다(도 24). 미표지 변이체의 수가 표지된 양성 트레이닝 데이터세트의 크기를 크게 초과함에 따라, 본 발명에서는 양성 트레이닝 데이터세트에 일치하는 미표지 변이체들의 상이한 세트를 각각 사용하는 8개의 망을 병렬로 트레이닝하여, 합의 예측을 얻었다.

[0345] 심층 학습망은, 일차 아미노산 서열만을 해당 입력으로서 사용하여, 전압-게이팅형 나트륨 채널 SCN2A(도 20), 간질, 자폐증, 및 지적 장애에 있는 주요 질환 유전자에 대해 도시된 바와 같이 유용한 단백질 기능성 도메인의 잔기에 높은 병원성 점수를 정확하게 할당한다. SCN2A의 구조는 4개의 상동 반복부를 포함하고, 각 반복부는 6개의 막횡단 나선(S1 내지 S6)을 포함한다. 막탈분극 시, 양으로 하전된 S4 막횡단 나선은 막의 세포 외측으로 이동하여, S5/S6 포어-형성 도메인이 S4-S5 링커를 통해 개방되게 한다. 임상적으로 초기 발병성 뇌병증에 연관된 S4, S4-S5 링커 및 S5 도메인의 돌연변이는, 망에 의해 유전자에서 가장 높은 병원성 점수를 갖는 것으로 예측되며, 건강한 모집단의 변이체에 대해 고갈된다(보충 표 6). 또한, 망이 도메인 내의 중요한 아미노산 위치를 인식하고 전사 인자의 DNA-접촉 잔기 및 효소의 촉매 잔기와 같이 이들 위치에서의 돌연변이에 가장 높은 병원성 점수를 할당한다는 것을 발견하였다(도 25a, 도 25b, 도 25c 및 도 26).

[0346] 심층 학습망이 기본 서열로부터의 기능과 단백질 구조에 대한 통찰력을 도출하는 방법을 더 잘 이해하기 위해, 학습망의 최초 3개 층으로부터 트레이닝 가능한 파라미터를 시각화하였다. 이들 층 내에서, 망은 그랜덤 점수와 같이 아미노산 거리의 기존 측정값에 근사하는 상이한 아미노산의 중량 간의 상관관계를 학습함을 관찰하였다(도 27). 이러한 초기 층의 출력은 이후 층에 대한 입력으로 되어서, 심층 학습망이 더욱 높은 데이터 표현을 점진적으로 구성할 수 있다.

[0347] 본 발명자들은, 트레이닝으로부터 보류된 10,000개의 공통 영장류 변이체를 사용하여 본 발명자들의 망의 성능을 기존의 분류 알고리즘과 비교하였다. 새로 발생하는 모든 인간 미스센스 변이체의 ~50%가 흔한 대립유전자 빈도로 선별 선택에 의해 필터링되기 때문에(도 49a), 돌연변이율 및 서열분석 커버리지에 의해 10,000개의 공통 영장류 변이체와 일치된 10,000개의 랜덤하게 선택된 변이체들의 세트에서 각각의 분류자에 대한 제50-백분위 점수를 결정하였고, 그 임계값에서 각 분류자의 정확도를 평가하였다(도 21d, 도 28a, 및 보충 데이터 파일 4). 본 발명의 심층 학습망(91% 정확도)은, 10,000개의 보류된 공통 영장류 변이체에 양성 결과를 할당할 때 다

른 분류자의 성능(다음으로 최고인 모델의 경우 80% 정확도)을 능가하였다.

[0348] 기존 방법들에 비해 개선된 점들의 대략 절반은 심층 학습망을 사용함으로써 발생하는 것이고, 절반은 인간 변이 데이터만으로 트레이닝된 망의 정확도와 비교할 때 영장류 변이로 트레이닝 데이터세트를 증대시키는 것으로부터 온 것이다(도 21d). 임상 시나리오에서 불확실한 유의미한 변이체의 분류를 테스트하기 위해, 본 발명에서는 신경 발달 장애 환자와 건강한 대조군에서 발생하는 드 노보 돌연변이를 구별하는 심층 학습망의 능력을 평가하였다. 유병률에 의해, 신경 발달 장애는 희귀한 유전자 질환의 가장 큰 카테고리들 중 하나를 구성하며, 최근의 트리오 서열분석 연구에서는 드 노보 미스센스 및 단백질 절단 돌연변이의 중심적 역할을 암시하였다.

[0349] 본 발명에서는, 개발 장애 해독(DDD) 코호트로부터의 영향을 받는 4,293명의 개체에서의 확실히 콜된 드 노보 미스센스 각각 대 사이몬 심플렉스 컬렉션(Simon's Simplex Collection: SSC) 코호트에서의 영향을 받지 않는 2,517명의 형제로부터의 드 노보 미스센스 변이체를 분류하였으며, 윌콕슨 순위-합 테스트로 두 개의 분포 간의 예측 점수의 차를 평가하였다(도 21e, 도 29a 및 도 29b). 심층 학습망은 이 작업에서 다른 분류자의 성능을 명백하게 능가한다($P<10^{-28}$; 도 21f 및 도 28b). 또한, 보류된 영장류 변이체 데이터세트 및 DDD 사례 대 대조군 데이터세트에 대한 다양한 분류자의 성능은 상관관계가 있었으며(Spearman $\rho=0.57$, $P<0.01$), 완전히 상이한 공급원과 방법을 사용하더라도 병원성을 평가하기 위한 두 개의 데이터세트 간에 양호한 일치를 나타내었다(도 30a).

[0350] 다음에, 본 발명에서는, 동일한 유전자 내의 양성 돌연변이 대 병원성 돌연변이를 분류할 때 심층 학습망의 정확성을 추정하고자 하였다. DDD 모집단이 일차 친척이 없는 영향을 받은 어린이의 인덱스 사례를 대부분 포함한다는 점을 고려할 때, 분류자가 드 노보 지배적 상속 모드를 갖는 유전자에 있어서 병원성을 선호함으로써 자신의 정확도를 팽창하지 않았음을 나타내는 것이 중요하다. 분석을, DDD 연구에서 질병 연관성에 대하여 명목상 중요한 605개 유전자로 제한하였으며, 단백질-절단 변이로부터만 계산하였다($P<0.05$). 이들 유전자 내에서, 드 노보 미스센스 돌연변이는 기대값(도 22a)에 비해 3/1로 농축되어, ~67%가 병원성임을 나타낸다.

[0351] 심층 학습망은, 동일한 유전자 세트의 내의 병원성 변이체와 양성 드 노보 변이체를 구별할 수 있었으며($P<10^{-15}$; 도 22b), 다른 방법들보다 큰 마진으로 성능을 능가하였다(도 22c 및 도 28c). 0.803 이상의 이진 컷오프에서(도 22d 및 도 30b), 88%의 분류 정확도에 대응하는 대조군에서의 드 노보 미스센스 돌연변이의 14%에 비해, 사례들의 드 노보 미스센스 돌연변이의 65%가 심층 학습망에 의해 병원성으로 분류된다(도 22e 및 도 30c). 신경 발달 장애에서 빈번한 불완전한 침투 및 가변 표현성을 고려할 때, 이 수치는, 부분 침투성 병원성 변이체가 대조군에 포함되어 있음으로 인해, 본 발명자들의 분류자의 정확도를 과소평가했을 수 있다.

[0352] **신규한 후보 유전자 발견**

[0353] 병원성 미스센스 돌연변이를 계층화하기 위해 ~0.803의 임계값을 적용하면, DDD 환자의 드 노보 미스센스 돌연변이가 1.5배에서 2.2배로 증가하고, 단백질-절단 돌연변이(2.5배)에 가까워지는 한편, 예상보다 많은 변이체의 총수의 1/3 미만을 포기하게 된다. 이것은 통계적 능력을 상당히 개선하여, 이전에는 초기 DDD 연구에서 게놈 전체의 유의성 임계값에 도달하지 못했던 지적 장애에서의 14개의 추가 후보 유전자를 발견할 수 있게 했다(표 1).

[0354] **인간 전문가 큐레이션과의 비교**

[0355] 본 발명에서는, ClinVar 데이터베이스로부터 최근의 인간 전문가에 의해 큐레이션된 변이체에 대한 다양한 분류자의 성능을 조사했지만, ClinVar 데이터세트에 대한 분류자의 성능은 보류된 영장류 변이체 데이터세트 또는 DDD 사례 대 대조군 데이터세트와 유의한 상관관계(각각 $P=0.12$ 및 $P=0.34$)가 없음을 발견하였다(도 31a 및 도 31b). 기존 분류자들이 인간 전문가 큐레이션으로부터의 편향을 갖고 있고, 이러한 인간 휴리스틱이 올바른 방향에 있는 경향이 있지만 최적이 아닐 수도 있다고 가정한다. 일례는, ClinVar에서 병원성 변이체와 양성 변이체 간의 그랜덤 점수의 평균 차이며, 이는 DDD 사례에서의 드 노보 변이체와 605개의 질병 연관 유전자 내의 대조군 간의 차의 두 배이다(표 2). 이에 비해, 인간 전문가 큐레이션은, 단백질 구조, 특히 다른 분자와 상호작용하도록 이용 가능한 표면에 노출된 잔기의 중요성을 충분히 이용하지 않는 것으로 보인다. ClinVar 병원성 돌연변이와 DDD 드 노보 돌연변이 모두가 예측된 용매 노출 잔기에 연관되어 있지만, 양성 및 병원성 ClinVar 변이체들 간의 용매 접근성의 차이는 DDD 사례 대 대조군에서 볼 수 있는 차이의 절반에 불과하다는 것을 관찰한다. 이러한 결과는 그랜덤 점수 및 보존과 같이 인간 전문가가 해석하는 데 더욱 간단한 인자를 선호하는 확인 편향을 시사한다. 인간이 큐레이션한 데이터베이스에 대해 트레이닝된 기계 학습 분류자는 이러한 경향을 강화

할 것으로 예상된다.

[0356] 본 발명의 결과는, 체계적인 영장류 모집단 서열분석이 현재 임상 게놈 해석을 제한하는 불확실한 유의성의 수백만의 인간 변이체를 분류하는 효과적인 전략임을 제시한다. 보류된 공통 영장류 변이체와 임상 변이체 모두에 대한 본 발명의 심층 학습망의 정확도는 네트워크를 트레이닝하는 데 사용되는 양성 변이체의 수에 따라 증가한다(도 23a). 또한, 6종의 비인간 영장류 종 각각의 변이체에 대한 트레이닝은 망의 성능 향상에 독립적으로 기여하는 반면, 더 먼 포유류의 변이체에 대한 트레이닝은 망의 성능에 부정적인 영향을 미친다(도 23b 및 도 23c). 이러한 결과는, 공통 영장류 변이체가 침투성 멘델리안 질환과 관련하여 인간에게서 대체로 양성이라는 주장을 지지하지만, 더 먼 종의 변이체에 대해서는 그렇지 않다.

[0357] 본 연구에서 조사된 비인간 영장류 게놈의 수는 서열분석된 인간 게놈과 솜의 수와 비교하여 적지만, 이러한 추가 영장류는 공통 양성 변이에 대한 불균형한 양의 정보에 기여한다는 점에 주목하는 것이 중요하다. ExAC를 사용한 시뮬레이션에 따르면, 겨우 수백 명의 개인(도 56) 후에 공통 인간 변이체(0.1% 초과의 대립유전자 빈도)의 발견이 빠르게 안정되고, 수백만 명에 대한 건강한 모집단 추가 서열분석이 희귀한 추가 변이체에 주로 기여함을 나타낸다. 대립유전자 빈도에 기초하여 임상적으로 대체로 양성인 것으로 알려진 일반적인 변이체와는 달리, 건강한 모집단에서의 희귀한 변이체는 불완전한 침투가 있는 우성 유전병 또는 열성 유전병을 유발할 수 있다. 각 영장류 종은 공통 변이체들의 상이한 풀을 갖고 있기 때문에, 각 종의 수십 개의 구성원을 서열분석하는 것은 영장류 계통에서 양성 미스센스 변이를 체계적으로 분류하는 효과적인 전략이다. 실제로, 이 연구에서 조사된 6명의 비인간 영장류 종에서의 134명의 개인은, ExAC 연구에 의한 123,136명의 인간보다 거의 4배 많은 공통 미스센스 변이체를 제공한다(보충 표 5(도 58)). 수백 명의 개인을 대상으로 한 영장류 모집단 서열분석 연구는, 야생 동물 보호 구역 및 동물원에 상주하는 상대적으로 적은 수의 관련이 없는 개인에게도 실용적일 수 있으므로, 야생 인구에 대한 방해를 최소화하며, 이는 비인간 영장류의 보존 및 윤리적 치료 측면에서 중요하다.

[0358] 오늘날의 인간 모집단은, 대부분의 비인간 영장류 종보다 유전자 다양성이 훨씬 낮으며, 침팬지, 고릴라, 긴팔원숭이와 같은 개체당 단일 뉴클레오타이드 변이체의 약 절반이며, 오랑우탄과 같은 개체당 많은 변이체의 1/3이다. 대부분의 비인간 영장류 종에 대한 유전적 다양성 수준은 알려져 있지 않지만, 현존하는 많은 수의 비인간 영장류 종은 가능한 양성 인간 미스센스 위치의 대부분이 적어도 하나의 영장류 종의 공통 변이체에 의해 덮일 가능성이 있다고 추정할 수 있게 하며, 병원성 변이체가 제거 프로세스에 의해 체계적으로 식별될 수 있게 한다(도 23d). 이러한 종의 서브세트만을 서열분석하더라도, 트레이닝 데이터 크기를 증가시킴으로써 기계 학습으로 미스센스 결과를 더욱 정확하게 예측할 수 있다. 마지막으로, 본 발명자들의 지견은 미스센스 변이에 중점을 두는 반면, 이 전략은 넌코딩 변이의 결과를 추론하는 데에도 적용될 수 있으며, 특히, 변이체가 IBS인지를 명확하게 결정하도록 인간과 영장류 게놈 간의 충분한 정렬이 있는 보존된 규제 지역에서 그러하다.

[0359] 504개의 공지된 비인간 영장류 종 중에서, 밀렵 및 광범위한 서식지 손실로 인해 약 60%가 멸종 위기에 처해 있다. 이들 종의 모집단 크기의 감소와 잠재적 멸종은 유전적 다양성의 대체불가능한 손실을 나타내며, 이러한 독특하고 대체 불가능한 종과 우리 자신 모두에게 이익이 될 전세계적 보존 노력에 대한 시급한 동기를 부여한다.

[0360] **데이터 생성 및 정렬**

[0361] 응용에서의 좌표는, 다중 서열 정렬을 사용하여 hg19에 맵핑된 다른 종들에서의 변이체에 대한 좌표를 포함하는 인간 게놈 구성 UCSC hg19/GRCh37을 지칭한다. 단백질-코딩 DNA 서열 및 99마리 척추동물 게놈 및 분지 길이의 다중 서열 정렬에 대한 정준 전사체를 UCSC 게놈 브라우저로부터 다운로드하였다.

[0362] 엑솜 집계 컨소시엄(ExAC)/게놈 집계 데이터베이스(gnomAD 엑솜) v2.0으로부터 인간 엑솜 다형성 데이터를 취득하였다. 24마리의 침팬지, 13마리의 보노보, 27마리의 고릴라, 10마리의 오랑우탄에 대한 전체 게놈 서열분석 데이터와 유전자형을 포함하는 유인원 게놈 서열분석 프로젝트로부터 영장류 변이 데이터를 취득하였다. 또한, 침팬지와 보노보에 대한 별개의 연구로부터의 35마리 침팬지의 변이를 포함했지만, 변이체 콜 방법의 차이로 인해, 이들을 모집단 분석으로부터 제외하고 심층 학습 모델의 트레이닝에만 사용하였다. 또한, 16개의 레서스 개체 및 9마리의 마모셋 개체를 사용하여 이들 종에 대한 초기 게놈 프로젝트에서의 변이를 분석하였지만, 개체 수준 정보는 이용할 수 없었다. dbSNP로부터 레서스, 마모셋, 돼지, 소, 염소, 마우스, 닭고기 및 제브라피쉬에 대한 변이 데이터를 취득하였다. dbSNP에는 추가 오랑우탄 변이체도 포함되었는데, 모집단 분석에 개별 유전자형 정보를 사용할 수 없었기 때문에, 심층 학습 모델을 트레이닝하는 데에만 사용하였다. 균형 맞춤 선택으로 인한 영향을 피하기 위해, 모집단 분석을 위해 확장된 주요 조직적합성 복합 영역(chr6: 28,477,797-33,448,354) 내에서 변이체를 배제하였다.

[0363] 인간의 단백질-코딩 영역에 대한 직교 일대일 맵핑을 보장하고 위유전자에 대한 맵핑을 방지하도록, 99마리 척추동물의 다수 종 정렬을 사용하였다. 참조/대체 배향에 있어서 변이체가 발생한 경우 변이체를 IBS인 것으로서 수용하였다. 변이체가 인간 및 다른 종 모두에서 동일한 예측된 단백질-코딩 결과를 갖는 것을 보장하기 위해, 코돈의 다른 두 개의 뉴클레오타이드가 미스센스 변이체 및 동의 변이체 둘 다에 대해 종들 간에 동일할 것을 요구하였다. 분석에 포함된 각 종으로부터의 다형성은 보충 데이터 파일 1에 나열되고, 자세한 메트릭은 보충 표 1에 표시되어 있다.

[0364] 4개의 대립유전자 빈도 카테고리(도 49a)의 각각에 대해, 96개의 가능한 트라이뉴클레오타이드 컨텍스트 각각에서의 동의 변이체 및 미스센스 변이체의 예상 수를 추정하고 돌연변이율을 보정하도록 인트로닉 영역에서의 변이를 사용하였다(도 51 및 보충 표 7, 표 8(도 59)). 또한 IBS CpG 디뉴클레오타이드 변이체 및 넌-CpG 디뉴클레오타이드 변이체를 개별적으로 분석하고 미스센스/동의 비율이 양측 클래스의 대립유전자 빈도 스펙트럼에서 평탄하다는 것을 검증하였으며, 이는 CpG 변이체 및 넌-CpG 변이체 둘 다에 대한 본 발명의 분석이 해당 돌연변이율의 큰 차이에도 불구하고 유지된다는 것을 나타낸다(도 52a, 도 52b, 도 52c 및 도 52d).

[0365] **다른 종에서 다형성과 IBS인 인간 미스센스 변이체의 고갈**

[0366] 다른 종에 존재하는 변이체가 인간의 흔한 대립유전자 빈도(>0.1%)로 허용될 수 있는지를 평가하기 위해, 다른 종의 변이와 IBS인 인간 변이체를 식별하였다. 각각의 변이체에 대해, 이들을 인간 모집단의 대립유전자 빈도에 기초하여 4가지 카테고리(싱글톤, 싱글톤 ~0.01% 초과, 0.01% 내지 ~0.1%, >0.1%) 중 하나에 할당하고, 희귀(<0.1%) 변이체와 흔한, 즉, 공통(>0.1%) 변이체 간의 미스센스/동의 비율(MSR)의 감소를 추정하였다. 흔한 인간 대립유전자 빈도(>0.1%)로의 IBS 미스센스 변이체의 고갈은, 인간의 흔한 대립유전자 빈도로 자연 선택에 의해 필터링 제거될 정도로 충분히 유해한 다른 종으로부터의 변이체의 분율을 나타낸다.

[0367] $$\text{고갈\%} = \frac{MSR_{\text{희귀}} - MSR_{\text{공통}}}{MSR_{\text{희귀}}}$$

[0368] 미스센스/동의 비율 및 고갈 백분율은 종마다 연산되어 도 50b 및 보충 표 2에 도시되어 있다. 또한, 침팬지 공통 변이체(도 49b), 침팬지 싱글톤 변이체(도 49c), 및 포유류 변이체(도 50a)의 경우, 희귀 변이체와 공통 변이체 간의 미스센스/동의 비율의 차가 유의한지를 테스트하도록 2×2 분할표에 대한 동질성의 카이 제곱 테스트(χ^2) 테스트를 수행하였다.

[0369] 서열분석은 유인원 게놈 프로젝트로부터 제한된 수의 개체에 대해서만 수행되었기 때문에, ExAC로부터의 인간 대립유전자 빈도 스펙트럼을 사용하여 일반적인 침팬지 모집단에서의 희귀한(<0.1%) 또는 흔한(>0.1%) 샘플링된 변이체의 분율을 추정하였다. ExAC 대립유전자 빈도를 기반으로 24명 인간의 코호트를 샘플링하고 이러한 코호트에서 한 번 또는 한 번보다 많게 관찰된 미스센스 변이체를 식별하였다. 한 번보다 많게 관찰된 변이체는 일반 모집단에서 99.8%의 공통 확률(>0.1%)을 보인 반면, 코호트에서 한 번만 관찰된 변이체는 일반 모집단에서 69%의 공통 확률을 보였다. 더욱 먼 포유류에서의 미스센스 변이체에 대하여 관찰된 고갈이 더 잘 보존된 유전자의 혼동 효과로 인한 것이 아님을 검증하기 위해, 상기 분석을 반복하여 인간에 비해 11마리의 영장류와 50마리의 포유류의 다중 서열 정렬에 있어서 >50% 평균 뉴클레오타이드 동일성을 갖는 유전자로만 제한하였다(보충 표 3 참조).

[0370] 이는 결과에 실질적으로 영향을 미치지 않으면서 분석으로부터 약 7%의 인간 단백질-코딩 유전자를 제거하였다. 또한, (dbSNP로부터 선택된 대부분의 종이 사육되었으므로) 사육 아티팩트 또는 변이체 콜에 의해 본 발명의 결과가 영향을 받지 않았음을 보증하기 위하여, 종내 다형성 대신 밀접하게 관련된 종들의 쌍으로부터 고정된 치환을 사용하여 분석을 반복하였다(도 50d, 보충 표 4, 및 보충 데이터 파일 2).

[0371] **인간, 영장류, 포유류, 및 기타 척추동물에 대한 다형성 데이터의 ClinVar 분석**

[0372] 다른 종과 IBS인 변이체의 임상적 영향을 조사하기 위해, 병원성에 대해 상충되는 주석이 있거나 불확실한 유의성 변이체로만 표지된 변이체를 제외하고 ClinVar 데이터베이스를 다운로드하였다. 보충 표 9에 제시된 필터링 단계에 따르면, 병원성 카테고리에는 총 24,853개의 미스센스 변이체가 있고, 양성 카테고리에는 17,775개의 미스센스 변이체가 있다.

[0373] 인간, 비인간 영장류, 포유류 및 다른 척추동물에서의 변이와 IBS인 병원성 및 양성 ClinVar 변이체의 수를 계수하였다. 인간에 대해서는, ExAC 대립유전자 빈도로부터 샘플링된 30명의 인간 코호트를 시뮬레이션하였다. 각

종에 대한 양성 및 병원성 변이체의 수는 보충 표 10에 나타낸다.

[0374] **모델 트레이닝을 위한 양성 변이체 및 미표지 변이체의 생성**

[0375] 기계 학습을 위해 인간 및 비인간 영장류로부터 주로 공통되는 양성 미스센스 변이체의 양성 트레이닝 데이터세트를 구성하였다. 데이터세트는, 공통 인간 변이체(>0.1% 대립유전자 빈도; 83,546개의 변이체), 및 침팬지, 보노보, 고릴라, 오랑우탄, 레서스 및 마모셋의 변이체(301,690개의 고유 영장류 변이체)를 포함한다. 각 공급원에 의해 제공되는 양성 트레이닝 변이체의 수는 보충 표 5에 나와 있다.

[0376] 트라이뉴클레오타이드 컨텍스트, 서열분석 커버리지, 및 종과 인간 간의 정렬성을 제어하도록 매칭된 미표지 변이체들의 세트와 표지있는 변이체들의 세트를 구별하도록 심층 학습망을 트레이닝하였다. 미표지 트레이닝 데이터세트를 취득하기 위해, 정준 코딩 영역에서 가능한 모든 미스센스 변이체로 시작하였다. ExAC로부터 123,136개의 엑솜에서 관찰된 변이체 및 시작 또는 정지 코돈의 변이체는 제외하였다. 총 68,258,623개의 미표지 미스센스 변이체가 생성되었다. 이것은, 서열분석 커버리지가 열악한 영역, 및 영장류 변이체에 대해 일치하지 않는 미표지 변이체를 선택할 때 인간과 영장류 게놈 간의 일대일 정렬이 없는 영역을 보정하도록 필터링되었다.

[0377] 표지된 양성 변이체들의 동일한 세트 및 미표지 변이체들의 랜덤하게 선택된 샘플링된 세트 8개를 사용하는 8개의 모델을 트레이닝하고 이들의 예측값의 평균을 취함으로써 컨센서스 예측을 얻었다. 또한, 유효성확인 및 테스트를 위해 랜덤하게 샘플링된 10,000개의 영장류 변이체 세트 2개를 따로 설정하였으며, 이것은 트레이닝으로부터 보류된 것이다(보충 데이터 파일 3). 이들 세트 각각에 대해, 트라이뉴클레오타이드 컨텍스트와 일치하는 10,000개의 미표지 변이체를 샘플링하였으며, 이는 상이한 분류 알고리즘(보충 데이터 파일 4)을 비교할 때 각 분류자의 임계값을 정규화하는 데 사용되었다. 다른 구현예에서는, 2 내지 500 범위의 더 적은 또는 추가 모델을 앙상블에서 사용할 수 있다.

[0378] 심층 학습망의 2가지 버전의 분류 정확도를 평가하였으며, 하나는 공통 인간 변이체만으로 트레이닝된 것이고, 하나는 공통 인간 변이체 및 영장류 변이체를 포함하는 완전한 양성 표지된 데이터 세트로 트레이닝된 것이다.

[0379] **심층 학습망의 아키텍처**

[0380] 각각의 변이체에 대해, 병원성 예측망은, 관심 변이체를 중심으로 51-길이 아미노산 서열을 입력으로서 취하고, 중심 위치에서 치환된 미스센스 변이체를 이차 구조 및 용매 접근성 망(도 2 및 도 3)의 출력으로서 취한다. 3개의 51 길이 위치 빈도 행렬은, 99마리의 척추동물의 다중 서열 정렬로부터 생성되며, 상기 3개는, 영장류 11마리에 대하여 하나, 영장류를 제외한 50마리 포유동물에 대하여 하나, 영장류와 포유류를 제외한 38마리의 척추동물에 대하여 하나이다.

[0381] 이차 구조 심층 학습망은, 각각의 아미노산 위치에서 알파-나선(H), 베타 시트(beta sheet)(B), 및 코일(C)의 3-상태 이차 구조를 예측한다(보충 표 11). 용매 접근성 망은, 각각의 아미노산 위치에서 매립된(B), 개재된(intermediate)(I) 및 노출된(E)인 3-상태 용매 접근성을 예측한다(보충 표 12). 양측 망은, 플랭킹 아미노산 서열만을 해당 입력으로서 취하고, 단백질 데이터뱅크에서 공지된 비-중복 결정 구조로부터의 표지를 사용하여 트레닝되었다(보충 표 13). 미리 트레이닝된 3-상태 이차 구조 및 3-상태 용매 접근성 망들에 대한 입력을 위해, 길이가 51 및 깊이 20인 99마리의 모든 척추동물에 대해 다중 서열 정렬로부터 생성된 단일 길이 위치 빈도 행렬을 사용하였다. 단백질 데이터뱅크로부터 알려진 결정 구조의 망을 미리 트레이닝한 후에, 이차 구조 및 용매 모델의 최종 2개 층을 제거하고, 망의 출력을 병원성 모델의 입력에 직접 연결하였다. 3-상태 이차 구조 예측 모델에 대해 달성된 최고의 테스트 정확도는 79.86%였다(보충 표 14). 예측된 구조 표지만을 사용하는 것 대 결정 구조를 갖는 대략 ~4,000개의 인간 단백질에 대해 DSSP-주석 표시된(단백질의 이차 구조 정의) 구조 표지를 사용할 때 신경망의 예측을 비교하는 경우 실질적인 차이가 없었다(보충 표 15).

[0382] 병원성 예측을 위한 본 발명의 심층 학습망(PrimateAI) 및 이차 구조 및 용매 접근성을 예측하기 위한 심층 학습망 모두는 잔여 블록의 구조를 채택하였다. PrimateAI의 상세한 아키텍처는 (도 3) 및 보충 표 16(도 4a, 도 4b, 및 도 4c)에 설명되어 있다. 이차 구조 및 용매 접근성을 예측하기 위한 망의 상세한 아키텍처는 도 6 및 보충 표 11(도 7a 및 도 7b) 및 보충 표 12(도 8a 및 도 8b)에 설명되어 있다.

[0383] **10,000개의 영장류 변이체의 보류 테스트 세트에 대한 분류자 성능의 벤치마킹**

[0384] 테스트 데이터세트의 10,000개의 보류된 영장류 변이체를 사용하여 심층 학습망 및 이전에 발표된 다른 20개의 분류자를 벤치마킹했으며, 이를 위해 dbNSFP 데이터베이스로부터 예측 점수를 취득하였다. 10,000개의 보류된 영장류 변이체 테스트 세트에서의 각각의 분류자의 성능도 도 28a에 제공된다. 상이한 분류자는 광범위하게 다

양한 점수 분포를 가졌기 때문에, 트라이뉴클레오타이드 컨텍스트에 의해 테스트 세트와 일치한 랜덤하게 선택된 10,000개의 미표지 변이체를 사용하여 각 분류자의 제50-백분위 임계값을 식별하였다. 방법들 간 공정한 비교를 보장하도록 해당 분류자의 제50-백분위 임계값에서 양성으로서 분류된 10,000개의 보류된 영장류 변이체 테스트 세트에서의 변이체의 분율에 대해 각 분류자를 벤치마킹하였다.

[0385] 각각의 분류자에 대해, 제50-백분위 임계값을 사용하여 양성으로서 예측된 보류된 영장류 테스트 변이체의 분율이 또한 도 28a 및 보충 표 17(도 34)에 도시되어 있다. 또한, PrimateAI의 성능이 변이체 위치에서 정렬된 종들의 수에 대하여 강력하고, 포유류로부터 충분한 보존 정보를 이용할 수 있는 한 일반적으로 성능이 우수함을 나타내는데, 이는 대부분의 단백질-코딩 서열에 대하여 참이다(도 57).

[0386] **DDD 연구로부터의 드 노보 변이체의 분석**

[0387] DDD 연구로부터 공개된 드 노보 변이체 및 SSC 자폐증 연구에서의 건강한 형제 대조군으로부터의 드 노보 변이체를 취득하였다. DDD 연구는 드 노보 변이체에 대한 신뢰 수준을 제공하며, 변이체 콜 에러로 인해 잠재적 위양성인 임계값 <0.1을 갖는 DDD 데이터세트로부터의 변이체를 제외하였다. 일 구현예에서, 전체적으로, DDD 영향을 받은 개인으로부터 3,512개의 미스센스 드 노보 변이체 및 건강한 대조군의 1,208개의 미스센스 드 노보 변이체를 가졌다. 99마리 척추동물 다중 서열 정렬을 위해 UCSC에 의해 사용된 정준 전사 주석은 DDD에 의해 사용된 전사 주석과 약간 달랐으며, 그 결과 미스센스 변이체의 총수에 약간의 차가 발생하였다. 자폐증 연구로부터 영향을 받지 않은 형제 대조군의 드 노보 미스센스 변이체 대 DDD 영향을 받는 개체의 드 노보 미스센스 변이체를 구별하는 능력에 대한 분류 방법을 평가하였다. 각 분류자에 대해, 두 분포에 대한 예측 점수들 간의 차에 대한 윌콕슨 순위-합 테스트로부터 P 값을 보고하였다(보충 표 17(도 34)).

[0388] 동일한 질병 유전자 내의 양성 및 병원성 변이체을 구별할 때 다양한 분류자의 정확도를 측정하기 위해, DDD 코호트에서의 드 노보 단백질 절단 변이체에 농축된 605개 유전자의 서브세트에 대한 분석을 반복하였다($P<0.05$, 푸아송 정확 테스트)(보충 표 18). 이러한 605개 유전자 내에서, 기대 이상인 드 노보 미스센스 돌연변이의 3/1 농축에 기초하여 DDD 데이터세트에서의 드 노보 변이체의 2/3가 병원성이고 1/3이 양성이라고 추정하였다. 최소한의 불완전한 침투를 가정하였고 건강한 대조군에서의 드 노보 미스센스 돌연변이는 양성이었다. 각 분류자에 대해, 이들 데이터세트에서 관찰된 경험적 비율과 동일한 수의 양성 또는 병원성 예측값을 생성하는 임계값을 식별하였고, 이 임계값을 이진 컷오프로서 사용하여 사례 대 대조군에 있어서 드 노보 돌연변이를 구별할 때 각 분류자의 정확도를 추정하였다. 수신자 오퍼레이터 특성 곡선을 구성하기 위해, 드 노보 DDD 변이체의 병원성 분류를 진양성 콜로서 취급하고, 건강한 대조군에서의 드 노보 변이체의 분류를 위양성 콜인 병원성으로서 취급하였다. DDD 데이터세트에는 양성 드 노보 변이체의 1/3이 포함되므로, 이론적으로 완벽한 분류자에 대한 곡선하 면적(AUC)은 1보다 작다. 따라서, 양성 변이체와및 병원성 변이체를 완벽하게 분리하는 분류자는, DDD 환자의 드 노보 변이체의 67%를 진양성으로, DDD 환자의 드 노보 변이체의 33%를 위양성으로, 및 대조군의 드 노보 변이체의 100%를 진음성으로 분류하며, 0.837인 최대 가능 AUC를(도 29a 및 29b 및 보충 표 19(도 35))을 산출한다.

[0389] **신규한 후보 유전자 발견**

[0390] 드 노보 돌연변이의 관찰된 수를 널 돌연변이 모델 하에서 예상된 수와 비교함으로써 유전자의 드 노보 돌연변이의 농축을 테스트하였다. DDD 연구에서 수행된 농축 분석을 반복하고, PrimateAI 점수가 >0.803인 드 노보 미스센스 돌연변이만을 계수할 때 새롭게 게놈 전체에 중요한 유전자를 보고한다. PrimateAI 임계값>0.803(유전체 전체의 모든 가능한 미스센스 돌연변이의 대략 1/5)을 초과하는 미스센스 변이체의 분율에 의해 드 노보 손상 미스센스 변이체에 대한 게놈 전체의 기대값을 조정하였다. DDD 연구에 따르면, 각각의 유전자는 4가지 테스트를 필요로 하였고, 하나는 단백질 절단 농축을 테스트하는 것이고, 하나는 단백질-변경 드 노보 돌연변이의 농축을 테스트하는 것이며, 이들 테스트 모두는 DDD 코호트 및 신경발달 트리오 서열분석 코호트의 보다 큰 메타분석을 위한 것이었다. 단백질-변형 드 노보 돌연변이의 농축은, 피셔의 방법에 의해 코딩 서열 내의 미스센스 드 노보 돌연변이의 클러스터링의 테스트와 조합되었다(보충 표 20, 21). 각 유전자에 대한 P값은 4가지 테스트의 최소값으로부터 취해졌으며, 게놈 전체의 유의성은 $P<6.757\times10^{-7}$ ($\alpha=0.05$, 4가지 테스트에서 18,500개의 유전자)로 결정되었다.

[0391] **ClinVar 분류 정확도**

[0392] 기존 분류자의 대부분은 예컨대 ClinVar에 대해 트레이닝된 분류자의 예측 점수를 사용하여 ClinVar 콘텐츠에 대해 직접적으로 또는 간접적으로 트레이닝되기 때문에, ClinVar 데이터세트의 분석을 제한하여 2017년 이후 추

가된 ClinVar 변이체만을 사용하였다. 최근 ClinVar 변이체와 다른 데이터베이스 간에는 상당한 중복이 있었으므로, ExAC에서 흔한 대립유전자 빈도(>0.1%)로 발견되거나 인간 유전자 돌연변이 데이터베이스(HGMD), 라이덴 개방 변이 데이터베이스(LOVD), 또는 범용 단백질 자원(Uniprot)에 있는 변이체를 제거하기 위해 추가로 필터링을 행하였다. 불확실한 유의성으로서만 주석이 달린 변이체와 주석이 충돌하는 변이체를 배제한 후, 양성 주석이 있는 177개의 미스센스 변이체와 병원성 주석이 있는 969개의 미스센스 변이체가 남았다. 심층 학습망 및 다른 분류 방법을 모두 사용하여 이러한 ClinVar 변이체를 점수를 계산하였다. 각 분류자에 대해, 이들 데이터세트에서 관찰된 경험적 비율과 동일한 수의 양성 또는 병원성 예측을 생성하는 임계값을 식별하였고, 이 임계값을 이진 컷오프로서 사용하여 각 분류자의 정확도를 추정하였다(도 31a 및 도 31b).

[0393] **트레이닝 데이터 크기를 증가시키고 다른 트레이닝 데이터의 상이한 공급원을 사용하는 영향**

[0394] 심층 학습망의 성능에 대한 트레이닝 데이터 크기의 영향을 평가하기 위해, 385,236종의 영장류 및 일반적인 인간 변이체의 표지된 양성 트레이닝 세트로부터 변이체들의 서브세트를 랜덤하게 샘플링하고, 기저 심층 학습망 아키텍처를 동일하게 유지하였다. 각 개별 영장류 종의 변이체가 분류 정확도에 기여하는 반면 각 개별 포유류 종의 변이체가 분류 정확도를 낮추는 것을 나타내기 위하여, 일 구현예에 따르면, 83,546개의 인간 변이체와 각 종에 대해 랜덤하게 선택된 변이체를 포함하는 트레이닝 데이터세트를 사용하여 심층 학습망을 트레이닝하였고, 다시 기저 학습망 아키텍처를 동일하게 유지하였다. 트레이닝 세트에 추가한 변이체의 상수(23,380)는, 가장 적은 수의 미스센스 변이체를 갖는 종, 즉, 보노보에서 이용 가능한 변이체의 총수였다. 각 분류자의 중간 성능을 얻기 위해 트레이닝 절차를 5회 반복하였다.

[0395] **서열분석되는 영장류 모집단의 수의 증가와 함께 모든 가능한 인간 미스센스 돌연변이의 포화도**

[0396] ExAC에서 관찰된 인간 공통 미스센스 변이체(>0.1% 대립유전자 빈도)의 트라이뉴클레오타이드 컨텍스트에 기초하여 변이체를 시뮬레이션함으로써, 504개의 현존하는 영장류 종에 존재하는 공통 변이체에 의한 ~70백만 개의 가능한 인간 미스센스 돌연변이의 예상 포화도를 조사하였다. 각 영장류 종에 대해, 인간은 다른 영장류 종의 개체당 대략 절반의 변이체 수를 갖고 약 ~>0.1% 대립유전자 빈도로 선별 선택에 의해 인간 미스센스 변이체의 약 50%를 필터링하였기 때문에, 인간에서 관찰된 일반적인 미스센스 변이체의 수의 4배(대립유전자 빈도>0.1%인 ~83,500개의 미스센스 변이체)를 시뮬레이션하였다(도 49a).

[0397] 조사된 인간 코호트의 크기가 증가함에 따라 발견된 인간 공통 미스센스 변이체(>0.1% 대립유전자 빈도)의 분율을 모델링하기 위해(도 56), ExAC 대립유전자 빈도에 따라 유전자형을 샘플링하고 이들 시뮬레이션 코호트에서 관찰된 공통 변이체의 분율을 적어도 한번 보고하였다.

[0398] 일 구현예에서, PrimateAI 점수의 실제 적용을 위해, 대조군과 비교한 경우에 드 노보 변이체의 농축(도 21d), 및 열성 상속 모드를 갖는 유전자에서 유사 병원성에 대해서 0.7보다 높고 유사 양성에 대해서 0.5보다 낮은 임계값에 기초하여, >0.8의 임계값이 유사 병원성 분류를 위해 바람직하고, <0.6이 유사 양성이고, 지배적 상속 모드를 갖는 유전자에서 0.6 내지 0.8이 중간이다.

[0399] 도 2는 본 명세서에서 "PrimateAI"라고 칭하는 병원성 예측을 위한 심층 잔여망의 예시적인 아키텍처를 도시한다. 도 2에서, 1D는 1차원 컨볼루션층을 지칭한다. 예측된 병원성은 0(양성) 내지 1(병원성)까지의 크기로 된 것이다. 망은, 변이체를 중심으로 인간 아미노산(AA) 참조 서열 및 대체 서열(51 AA)을 입력으로서 취하고, 99마리 척추동물 종으로부터 계산된 위치 가중 행렬(PWM) 보존 프로파일, 및 이차 구조 및 용매 접근성 예측 심층 학습 망의 출력을 취하며, 이는 3-상태 단백질 이차 구조(나선-H, 베타 시트-B, 및 코일-C) 및 3-상태 용매 접근성(매립-B, 개재-I 및 노출-E)을 예측한다.

[0400] 도 3은 병원성 분류를 위한 심층 학습망인 PrimateAI의 개략도를 도시한다. 모델에 대한 입력은, 참조 서열 및 변이체가 치환된 서열 모두에 대한 플랭킹 서열의 51개 아미노산(AA), 영장류, 포유류, 및 척추동물 정렬로부터 3개의 51-AA-길이 위치-가중 행렬에 의해 표현된 보존, 및 미리 트레이닝된 이차 구조망 및 용매 접근성망(이 또한 51 AA 길이임)의 출력을 포함한다.

[0401] 도 4a, 도 4b 및 도 4c는 병원성 예측 심층 모델 PrimateAI의 예시적인 모델 아키텍처 세부사항을 나타내는 보충 표 16이다. 형상은 모델의 각 층에서의 출력 텐서의 형상을 특정하며, 활성화는 층의 뉴런에 제공되는 활성화이다. 모델에 대한 입력은, 변이체 주변의 플랭킹 아미노산 서열에 대한 위치-특이적 빈도 행렬(51 AA 길이, 20 깊이), 원-핫 인코딩된(one-hot encoded) 인간 참조 서열 및 대체 서열(51 AA 길이, 20 깊이), 및 이차 구조 및 용매 접근성 모델(51 AA 길이, 40 깊이)로부터의 출력이다.

[0402] 예시된 예는 1D 컨볼루션을 사용한다. 다른 구현예에서, 모델은, 2D 컨볼루션, 3D 컨볼루션, 팽창 또는 아트러

스 컨볼루션, 전치된 컨볼루션, 분리가능한 컨볼루션, 및 깊이별 분리가능한 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.

[0403] 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. x를 사용한 일괄 정규화층의 입력과 z를 사용한 출력을 이용하여, 일괄 정규화는 x에 대한 이하의 변환을 적용한다.

[0404] $$z = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \gamma + \beta$$

[0405] 일괄 정규화는, μ와 σ를 사용하여 입력 x에 평균-분산 정규화를 적용하고, 이를 γ와 β를 사용하여 선형으로 스케일링하고 시프트하며, μ와 σ는, 지수 이동 평균이라고 하는 방법을 사용하여 트레이닝 세트에 걸쳐 현재 층에 대하여 연산된다. 즉, 이들은 트레이닝 가능한 파라미터가 아니다. 대조적으로, γ와 β는 트레이닝 가능한 파라미터이다. 트레이닝 중에 계산된 μ와 및 σ의 값은 추론 동안 순방향 패스에 사용된다.

[0406] 도 5 및 도 6은 단백질의 이차 구조 및 용매 접근성을 예측하는 데 사용되는 심층 학습망 아키텍처를 도시한다. 모델에 대한 입력은, RaptorX 소프트웨어(단백질 데이터 서열에 대한 트레이닝용) 또는 99마리 척추동륨 정렬(인간 단백질 서열에 대한 트레이닝 및 추론용)에 의해 생성된 보존을 이용하는 위치 가중 행렬이다. 길이가 51개 AA인 최종 층에서 두 번째 층의 출력은 병원성 분류를 위한 심층 학습망에 대한 입력으로 된다.

[0407] 도 7a 및 도 7b는 3-상태 이차 구조 예측 심층 학습(DL) 모델에 대한 예시적인 모델 아키텍처 세부사항을 도시하는 보충 표 11이다. 형상은 모델의 각 층에서의 출력 텐서의 형상을 특정하며, 활성화는 층의 뉴런에 제공되는 활성화이다. 모델에 대한 입력은 변이체 주위의 플랭킹 아미노산 서열에 대한 위치-특이적 빈도 행렬(51 AA 길이, 20 깊이)였다.

[0408] 도 8a 및 도 8b는 3-상태 용매 접근성 예측 심층 학습 모델에 대한 예시적인 모델 아키텍처 세부사항을 나타내는 보충 표 12이다. 형상은 모델의 각 층에서의 출력 텐서의 형상을 특정하고, 활성화는 층의 뉴런에 제공되는 활성화이다. 모델에 대한 입력은, 변이체 주위의 플랭킹 아미노산 서열에 대한 위치-특이적 빈도 행렬 (51 AA 길이, 20 깊이)이었다.

[0409] 도 20은 주요 기능성 도메인에 대해 주석이 달린 SCN2A 유전자의 각 아미노산 위치에서 예측된 병원성 점수를 도시한다. 유전자에 따라 각 아미노산 위치에서 미스센스 치환에 대한 평균 PrimateAI 점수가 표시된다.

[0410] 도 21d는 트레이닝으로부터 보류된 10,000개의 공통 영장류 변이체의 시험 세트에 대한 양성 결과를 예측할 때 분류자들의 비교를 도시한다. y축은, 돌연변이율에 일치하는 10,000개의 랜덤한 변이체들의 세트에서 각 분류자의 임계값을 50-백분위 점수로 정규화한 후 양성으로 올바르게 분류된 영장류 변이체의 백분율을 나타낸다.

[0411] 도 21e는, 상응하는 윌콕슨 순위-합 P값과 함께, 영향을 받지 않는 형제와 비교하여 해독 발달 장애(DDD) 환자에게서 발생하는 드 노보 미스센스 변이체에 대한 PrimateAI 예측 점수의 분포를 도시한다.

[0412] 도 21F는 DDD 사례 대 대조군에서 드 노보 미스센스 변이체를 분리할 때 분류자들의 비교를 도시한다. 윌콕슨 순위-합 테스트 P 값이 각 분류자에 대해 표시된다.

[0413] 도 22a, 도 22b, 도 22c, 도 22d 및 도 22e는 P<0.05를 갖는 605개 DDD 유전자 내의 분류 정확도를 예시한다. 도 22a는 드 노보 단백질 절단 변형(P<0.05)에 대해 유의미한 605개의 관련 유전자 내에서 DDD 코호트로부터 영향을 받은 개체에서 기대 이상의 드 노보 미스센스 돌연변이의 농축을 도시한다. 도 22b는, DDD 환자 대 605개 관련 유전자 내에서 영향을 받지 않은 형제에게서 발생하는 드 노보 미스센스 변이체에 대한 PrimateAI 예측 점수의 분포를, 상응하는 윌콕슨 순위-합 P 값과 함께 도시한다.

[0414] 도 22c는, 605개 유전자 내의 대조군 대 사례에서의 드 노보 미스센스 변이체를 분리할 때의 다양한 분류자의 비교를 도시한다. y축은 각 분류자에 대한 윌콕슨 순위-합 테스트의 P값을 도시한다.

[0415] 도 22d는, 각각의 분류자에 대해 AUC가 표시된 수신자 오퍼레이터 특성 곡선에 도시된 다양한 분류자의 비교를 도시한다.

[0416] 도 22e는 각 분류자에 대한 분류 정확도 및 AUC를 도시한다. 도시된 분류 정확도는, 분류자가 도 22a에 도시된 농축에 기초하여 예상된 것과 동일한 수의 병원성 및 양성 변이체를 예측하는 임계값을 사용하는 진양성 및 진음성 에러율의 평균이다. DDD 드 노보 미스센스 변이체의 33%가 배경을 나타낸다는 사실을 고려하도록, 완벽한 분류자의 이용 가능한 최대 AUC는 점선으로 표시된다.

[0417] 도 23a, 도 23b, 도 23c 및 도 23d는 분류 정확도에 대한 트레이닝에 사용된 데이터의 영향을 도시한다. 심층 학습망은, 전체 데이터세트(385,236개의 변이체)까지 영장류 및 인간 공통 변이체의 수를 증가시키면서 트레이닝되었다. 도 23a에서, 각각의 망에 대한 분류 성능은, DDD 사례 대 대조군에서 10,000개의 보류된 영장류 변이체 및 드 노보 변이체의 정확도에 대해 벤치마킹된다.

[0418] 도 23b 및 도 23c는, 일 구현예에 따라 83,546개의 인간 공통 변이체 더하기 단일 영장류 또는 포유류 종으로부터의 23,380개의 변이체를 포함하는 데이터세트를 사용하여 트레이닝된 망의 성능을 도시한다. DDD 사례 대 대조군에서 10,000개의 보류된 영장류 변이체(도 23b) 및 도 노보 미스센스 변이체(도 23c)에 벤치마킹된 공통 변이의 상이한 공급원으로 트레이닝된 각 망에 대한 결과가 도시되어 있다.

[0419] 도 23d는 504개의 현존하는 영장류 종에서 IBS 공통 변이체(>0.1%)에 의한 모든 가능한 인간 양성 미스센스 위치의 예상 포화를 도시한다. y축은 하나 이상의 영장류 종에서 관찰된 인간 미스센스 변이체의 분율을 나타내며, CpG 미스센스 변이체는 녹색으로 표시되고, 모든 미스센스 변이체는 파란색으로 표시된다. 각 영장류 종에서의 공통 변이체를 시뮬레이션하기 위해, ExAC에서 공통 인간 변이체(>0.1% 대립유전자 빈도)에 대해 관찰된 트라이뉴클레오타이드 컨텍스트 분포와 일치하는, 대체가능한 모든 단일 뉴클레오타이드 치환 세트로부터 샘플링하였다.

[0420] 도 24는 공통 영장류 변이체의 확인에 대한 서열분석 커버리지의 영향에 대한 정정을 도시한다. 비인간 영장류 종에서 주어진 변이체를 관찰할 확률은, ExAC/gnomAD 엑솜 데이터세트에서 해당 위치의 서열분석 깊이와 반비례 상관관계가 있다. 대조적으로, 낮은 gnomAD 판독 깊이는, 서열분석된 다수의 인간 엑솜 공통 변이의 확인을 거의 보장하기 때문에. 해당 위치(>0.1% 대립유전자 빈도)에서 공통 인간 변이체를 관찰할 확률에 영향을 미치지 않았다. 망을 트레이닝하도록 각 영장류 변이체에 대해 일치하는 변이체를 선택할 때, 변이체를 선택할 확률은, 돌연변이율 및 유전자 변환을 제어하기 위한 트라이뉴클레오타이드 컨텍스트에 대한 매칭에 더하여, 서열분석 깊이의 영향에 대해 조정되었다.

[0421] 도 25a, 도 25b, 도 25c 및 도 26은 개시된 신경망에 의한 단백질 모티프의 인식을 도시한다. 도 25a, 도 25b, 도 25c와 관련하여, 단백질 도메인에 대한 신경망의 인식을 예시하기 위해, 3개의 상이한 단백질 도메인의 각 아미노산 위치에서 변이체에 대한 평균 PrimateAI 점수를 도시한다. 도 25a에서는, 반복 GXX 모티프에서 글리신을 갖는 COL1A2의 콜라겐 가닥이 강조된다. 콜라겐 유전자에서 임상적으로 확인된 돌연변이는, 콜라겐의 정상적인 조립을 방해하고 강한 지배적 음성 효과를 발휘하기 때문에, GXX 반복에서의 글리신의 미스센스 돌연변이에 대체로 기인한다. 도 25b에서는, IDS 설파타제 효소의 활성 부위가 강조되고, 이는 번역 후 포밀글리신으로 변형되는 시스테인을 활성화된 부위에서 포함한다. 도 25c에는, MYC 전사 인자의 bHLHzip 도메인이 도시되어 있다. 염기성 도메인은, 양으로 하전된 아르기닌 및 음으로 하전된 당-포스페이트 골격과 상호작용하는 라이신 잔기(강조됨)를 통해 DNA와 접촉한다. 류신-지퍼 도메인은 이량체화에 결정적인 7개 아미노산이 이격된 류신 잔기(강조됨)를 포함한다.

[0422] 도 26은 변이체에 대한 예측된 심층 학습 점수에 대한 변이체의 내외부의 각각의 위치를 교란시키는 효과를 도시하는 라인 플롯을 포함한다. 변이체 주위의 주변 아미노산(위치 -25 ~ +25)에서 입력을 체계적으로 제로화하였고 변이체의 신경망의 예상 병원성 변화를 측정하였다. 플롯은, 5,000개의 랜덤하게 선택된 변이체에 대해 각각의 주변 아미노산 위치에서 교란에 대하여 예측된 병원성 점수의 평균 변화를 도시한다.

[0423] 도 27은 가중치 모방 BLOSUM62 및 그랜덤 점수 행렬의 상관 패턴을 도시한다. 이차 구조 심층 학습망의 처음 3개 층에서 가중치의 상관 패턴은 BLOSUM62와 유사한 아미노산과 그랜덤 점수 행렬 간의 상관 관계를 나타낸다. 좌측 히트 맵은, 원-핫 표현을 사용하여 인코딩된 아미노산들 간에 이차 구조 심층 학습망의 2개의 초기 업샘플링층에 후속하는 제1 컨볼루션층으로부터의 파라미터 가중치의 상관 관계를 도시한다. 중간 히트 맵은 아미노산 쌍들 간의 BLOSUM62 점수를 나타낸다. 오른쪽 히트 맵은 아미노산 간의 그랜덤 거리를 나타낸다. 심층 학습 가중치와 BLOSUM62 점수 간의 피어슨 상관관계는 0.63이다($P=3.55\times10^{-9}$). 심층 학습 가중치와 그 랜덤 점수의 상관관계는 -0.59이다($P=4.36\times10^{-8}$). BLOSUM62와 그랜덤 점수 간의 상관관계는 -0.72이다($P=8.09\times10^{-13}$).

[0424] 도 28a, 도 28b, 및 도 28c는 심층 학습망 PrimateAI 및 다른 분류자의 성능 평가를 도시한다. 도 28a는, 트레

이닝 보류된 10,000개의 영장류 변이체의 테스트 세트에 대한 양성 결과를 예측할 때 심층 학습망 PrimateAI의 정확도, 및 SIFT, Polyphen-2, CADD, REVEL, M-CAP, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaSVM, MetaLR, MutPred, DANN, FATHMM-MKL_coding, Eigen, GenoCanyon, integrated_fitCons, 및 GERP를 포함하는 다른 분류자와의 비교를 도시한다. y축은, 돌연변이율 및 유전자 변환을 제어하도록 트라이뉴클레오타이드 컨텍스트에 대한 영장류 변이체와 일치하는 10,000개의 랜덤하게 선택된 변이체 세트를 사용하여 각 분류자의 임계값을 50-백분위 점수로 정규화하는 것에 기초하여 양성으로 분류된 영장류 변이체의 백분율을 나타낸다.

[0425] 도 28b는, 상기 열거된 20개의 기존 방법과 함께 DDD 사례 대 대조군에서 드 노보 미스센스 변이체들을 분리하는 데 있어서 PrimateAI 망의 성능의 비교를 도시한다. y축은 각 분류자에 대한 윌콕슨 순위-합 테스트의 P값을 나타낸다.

[0426] 도 28c는, 상기 열거된 20개 방법으로, 605개의 질환-관련 유전자 내에서 DDD 사례 대 영향을 받지 않은 대조군의 드 노보 미스센스 변이체들을 분리하는 데 있어서 PrimateAI 망의 성능의 비교를 나타낸다. y축은 각 분류자에 대한 윌콕슨 순위- 합 테스트의 P값을 나타낸다.

[0427] 도 29a 및 도 29b는 4개의 분류자의 예측 점수의 분포를 도시한다. SIFT, 폴리펜-2, CADD 및 REVEL을 포함한 4개 분류자의 예측 점수에 대한 히스토그램을, 해당 윌콕슨 순위-합 P-값을 갖는 DDD 사례 대 영향을 받지 않는 대조군에서 발생하는 드 노보 미스센스 변이체에 나타낸다.

[0428] 도 30a, 도 30b 및 도 30c는, 605개의 질병 관련 유전자에서 병원성 변이체와 양성 변이체를 분리할 때 PrimateAI 망과 다른 분류자의 정확도를 비교한다. 도 30a의 산점도는, DDD 사례 대 대조군(y 축)에 대한 각 분류자의 성능 및 보류된 영장류 데이터세트(x 축)에 대한 양성 예측 정확도를 도시한다. 도 30b는, 각각의 분류자에 대해 표시되는 곡선하 면적(AUC)과 수신자 오퍼레이터 특성(ROC) 곡선에 도시된 605개 유전자 내의 대조군 대 사례에서 드 노보 미스센스 변이체의 분리 시 상이한 분류자를 비교한다. 도 30c는 도 28a, 도 28b, 도 28c에 열거된 20개 분류자와 PrimateAI 망에 대한 분류 정확도를 도시한다. 도시된 분류 정확도는, 분류자가 도 22a의 농축에 기초하여 예상된 것과 동일한 수의 병원성 변이체 및 양성 변이체를 예측하는 임계값을 사용하는, 진양성률과 진음성율의 평균이다. DDD 사례에서는 드 노보 미스센스 변이체가 67% 병원성 변이체이고 및 33%가 양성이고 대조군에서는 드 노보 미스센스 변이체가 100% 양성이라고 가정하면, 완벽한 분류자에 대한 달성가능한 최대 AUC는 점선으로 표시된다.

[0429] 도 31a 및 도 31b는 인간 전문가에 의해 큐레이션된 ClinVar 변이체에 대한 분류자 성능과 경험적 데이터세트에 대한 성능 간의 상관 관계를 도시한다. 도 31a의 산점도는, 20개의 다른 분류자 각각에 대한 10,000개의 보류된 영장류 변이체(x축) 및 인간 단독 또는 인간+영장류 데이터로 트레이닝된 PrimateAI 망에서 ClinVar 변이체에 대한 분류 정확도(y축)를 도시한다. Spearman 상관 계수 rho 및 관련 P값이 표시된다. 평가를 분류자를 트레이닝하는 데 사용되지 않는 데이터로 제한하기 위해, 2017년 1월 내지 2017년 11월에 추가된 ClinVar 변이체만을 사용했으며 일반적인 인간 변이체는 ExAC/gnomAD(>0.1% 대립유전자 빈도)에서 제외했다. 표시된 ClinVar 분류 정확도는, 분류자가 ClinVar 데이터세트에서 관찰된 것과 동일한 수의 병원성 변이체 및 양성 변이체를 예측하는 임계값을 사용하는, 진양성 비율과 진음성 비율의 평균이다.

[0430] 도 31b의 산점도는, 인간만 또는 인간+영장류 데이터로 트레이닝된 PrimateAI 망과 20개의 다른 분류자 각각에 대하여 DDD 사례 대 대조군의 제어 완전 데이터세트(x축)에서 ClinVar 변이체에 대한 분류 정확도(y축)를 도시한다.

[0431] 도 32는, 트레이닝을 위한 6,367개의 비관련 단백질 서열, 검증을 위한 400개, 테스트를 위한 500개를 사용하여, 단백질 데이터뱅크로부터 주석이 달린 샘플에 대한 3-상태 이차 구조 및 3-상태 용매 접근성 예측 모델의 성능을 도시하는 보충 표 14이다. <25%의 서열 유사성을 갖는 단백질만이 단백질 데이터뱅크로부터 선택되었다. 3개의 클래스가 이차 구조 또는 용매 접근성에 대해 전체적으로 균형이 맞지 않으므로, 심층 학습망의 정확성을 성능 메트릭으로서 보고한다.

[0432] 도 33은, 예측된 이차 구조 표지를 사용하는 심층 학습망과 함께 이용 가능한 경우 DSSP 데이터베이스로부터 인간 단백질의 주석이 달린 이차 구조 표지를 사용하여 심층 학습망의 성능 비교를 도시하는 보충 표 15이다.

[0433] 도 34는, 10,000개의 보류된 영장류 변이체에 대한 정확도 값 및 평가한 20개 분류자의 각각에 대하여 DDD 사례 대 대조군의 드 노보 변이체의 p-값을 도시하는 보충 표 17이다. 인간 데이터만 갖는 PrimateAI 모델은, 공통 인간 변이체(모집단에서 >0.1%인 83.5K 변이체)만 포함하는 양성으로 주석 표시된 트레이닝 데이터세트를 사용

하여 트레이닝된 본 발명의 심층 학습망인 한편, 인간+영장류 데이터를 갖는 PrimateAI 모델은, 공통 인간 변이체와 영장류 변이체를 포함하는 385K 표지된 양성 변이체들의 풀 세트(full set)에 대하여 트레이닝된 본 발명의 심층 학습망이다.

[0434] 도 35는, 605개 질환 관련 유전자로 제한되는 DDD 사례 대 대조군 데이터세트의 드 노보 변이체에 대한 상이한 분류자들의 성능의 비교를 도시하는 보충 표 19이다. 상이한 방법들 간에 정규화를 행하도록, 각 분류자마다, 분류자가 DDD와 대조군 세트의 농축에 기초하여 예상된 것과 동일한 수의 병원성 변이체 및 양성 변이체를 예측하는 임계값을 식별하였다. 표시된 분류 정확도는 이러한 임계값에서 실제 진양성 비율과 진위성 비율의 평균이다.

[0435] 도 49a, 도 49b, 도 49c, 도 49d 및 도 49e는 인간 대립유전자 빈도 스펙트럼에 걸쳐 미스센스/동의 비율을 도시한다. 도 49a는, ExAC/gnomAD 데이터베이스로부터 123,136명의 인간에게서 관찰된 미스센스 변이체 및 동의 변이체가 대립유전자 빈도에 의해 4가지 카테고리로 나뉘어져 있음을 도시한다. 음영 처리된 회색 막대는 각 카테고리의 동의 변이체의 수를 나타내고, 진한 녹색 막대는 미스센스 변이체를 나타낸다. 각 막대의 높이는 각각의 대립유전자 빈도 카테고리에서 동의어 변이체의 수로 스케일링되고, 돌연변이율을 조정한 후 미스센스/동의 계수치 및 비율이 표시된다. 도 49b 및 도 49c는, 침팬지 공통 변이체(도 49b) 및 침팬지 단일 변이체(도 49c)와 IBS(identical-by-state)인 인간 미스센스 변이체 및 동의 변이체에 대한 대립유전자 빈도 스펙트럼을 도시한다. 희귀한 인간 대립유전자 빈도(<0.1%)와 비교하여 흔한 인간 대립유전자 빈도(>0.1%)에서의 침팬지 미스센스 변이체의 고갈은 동반되는 카이 제곱(χ^2) P값)과 함께 빨간 박스로 표시된다.

[0436] 도 49d는 비인간 영장류 종 중 적어도 하나에서 관찰되는 인간 변이체를 도시한다. 도 49e는, 영장류에서 관찰된 변이체(하단 행)와 비교하여 ExAC/gnomAD 대립유전자 빈도(중간 행)로부터 샘플링된 30명의 인간 코호트의 ClinVar 변이체와 비교하여, 전체 ClinVar 데이터베이스(상단 행)의 양성 미스센스 변이체 및 병원성 미스센스 변이체의 계수치를 도시한다. 불확실한 유의성으로만 주석이 달린 충돌하는 양성 및 병원성 주장 및 변이체는 배제하였다.

[0437] 도 50a, 도 50b, 도 50c 및 도 50d는 다른 종과 IBS인 미스센스 변이체에 대한 선별 선택을 도시한다. 도 50a는, 4종의 비영장류 포유류 종(마우스, 돼지, 염소 및 소)에 존재하는 변이체와 IBS인 인간 미스센스 변이체 및 동의 변이체에 대한 대립유전자 빈도 스펙트럼을 도시한다. 흔한 인간 대립유전자 빈도(>0.1%)로 미스센스 변이체들의 고갈은, 동반되는 카이제곱(χ^2) 테스트 P값과 함께 빨간 박스에 의해 표시된다.

[0438] 도 50b는, 분지 길이의 단위(뉴클레오타이드 위치당 평균 치환 수)로 표현된, 인간으로부터의 종의 진화 거리 대 흔한 인간 대립유전자 빈도(>0.1%)로 다른 종에서 관찰된 미스센스 변이체의 고갈을 도시하는 산점도이다. 각 종과 인간 사이의 총 가지 길이는 종의 명칭 옆에 표시된다. 단독 개체와 공통 변이체의 고갈값은, 관련 개체가 포함된 고릴라를 제외하고 변이체 빈도를 사용할 수 있는 종에 대해 표시된다.

[0439] 도 50c는, 영장류(중간 행)에서 관찰된 변이체와 비교된, 그리고 마우스, 돼지, 염소, 소(하단 행)에서 관찰되는 변이체와 비교된, ExAC/gnomAD 대립유전자 빈도(상단 행)로부터 샘플링된 30명의 인간 코호트에서의 양성 미스센스 변이체 및 병원성 미스센스 변이체의 계수치를 도시한다. 불확실한 유의성으로만 주석이 달린 충돌하는 양성 및 병원성 주장 및 변이체는 배제하였다.

[0440] 도 50d는, 흔한 인간 대립유전자 빈도(>0.1%) 대 인간으로부터 종의 진화 거리(평균 분지 길이의 단위로 표시됨)에서 밀접하게 관련된 종들의 쌍에서 관찰된 고정된 미스센스 치환의 고갈을 도시하는 산점도이다.

[0441] 도 51은 선별 선택이 없을 때 인간 대립유전자 빈도 스펙트럼에 걸쳐 예상되는 미스센스:동의 비율을 도시한다. 음영 처리된 회색 막대는 동의 변이체의 수를 나타내고, 짙은 녹색 막대는 미스센스 변이체의 수를 나타낸다. 점선은 동의 변이체에 의해 형성된 베이스라인을 도시한다. 각 대립유전자 빈도 카테고리에 대해 미스센스:동의 비율이 표시된다. 일 구현예에 따르면, 각각의 대립유전자 빈도 카테고리에서 예상되는 미스센스 및 동의 계수치는, 유전 변환의 GC 편향과 돌연변이율을 제어하는, 변이체의 트라이뉴클레오타이드 컨텍스트에 기초하여, 123,136개의 엑솜을 포함하는 ExAC/gnomAD 데이터세트로부터 인트론 변이체를 취하고 이들을 사용하여 4개의 대립유전자 빈도 카테고리 각각에 속할 것으로 예상되는 변이체의 분율을 추정함으로써 계산되었다.

[0442] 도 52a, 도 52b, 도 52c 및 도 52d는 CpG 변이체 및 넌-CpG 변이체에 대한 미스센스:동의 비율을 도시한다. 도 52a 및 도 52b는, ExAC/gnomAD 엑솜으로부터의 모든 변이체를 사용하여, 인간 대립유전자 빈도 스펙트럼에 걸친 CpG 변이체(도 52a) 및 넌-CpG 변이체(도 52a)에 대한 미스센스:동의 비율을 나타낸다. 도 52c 및 도 52d는, 침

팬지 공통 다형성과 IBS인 인간 변이체만으로 제한되는, 인간 대립유전자 빈도 스펙트럼에 걸친 CpG 변이체(도 52c) 및 넌-CpG 변이체(도 52d)에 대한 미스센스:동의 비율을 도시한다.

[0443] 도 53, 도 54 및 도 55는 6종의 영장류와 IBS인 인간 변이체의 미스센스:동의 비율을 도시한다. 미스센스:동의 비율의 패턴은, 침팬지, 보노보, 고릴라, 오랑우탄, 레서스 및 마모셋에 존재하는 변이와 IBS인 ExAC/gnomAD 변이체에 대한 인간 대립유전자 빈도 스펙트럼에 대한 것이다.

[0444] 도 56은, 조사된 인간 코호트의 크기를 증가시킴으로써 발견된 새로운 공통 미스센스 변이체의 포화를 나타내는 시뮬레이션이다. 시뮬레이션에서, 각 샘플의 유전자형을 gnomAD 대립유전자 빈도에 따라 샘플링하였다. 발견된 gnomAD 공통 변이체의 분율은 10 내지 100,000 크기의 각 샘플에 대해 100개의 시뮬레이션에 걸쳐 평균화된다.

[0445] 도 57은 게놈에서 상이한 보존 프로파일에 걸친 PrimateAI의 정확도를 도시한다. x축은 99-척추동물 정렬과 서열 주위에 51 AA의 백분율 정렬성을 나타낸다. y축은, 10,000개의 보류된 영장류 변이체의 테스트 데이터세트를 벤치마킹한, 각 보존 빈의 변이체에 대한 PrimateAI 정확도의 분류 성능을 나타낸다.

[0446] 도 58은 비인간 영장류에 존재하는 변이체 및 공통 인간 변이체로부터 표지된 양성 트레이닝 데이터세트에 대한 기여를 나타내는 보충 표 5이다.

[0447] 도 59는 예상 미스센스:동의 비율에 대한 대립유전자 빈도의 영향을 나타내는 보충 표 8이다. 돌연변이율 및 유전자 변환 편향을 제어하기 위해 트라이뉴클레오타이드 컨텍스트를 사용하여, 엑손 경계로부터 적어도 20-30nt 떨어진 인트론 영역에 있는 변이체의 대립유전자 빈도 스펙트럼에 기초하여 동의 변이체 및 미스센스 변이체의 예상 계수치를 계산하였다.

[0448] 도 60은 ClinVar 분석을 나타내는 보충 표 9이다. 일 구현예에 따르면, ClinVar 데이터베이스의 2017년 11월 빌드로부터 다운로드한 변이체를 필터링하여 주석이 충돌하는 미스센스 변이체를 제거하고 불확실한 유의미한 변이체를 배제하여, 17,775개의 양성 변이체와 24,853개의 병원성 변이체를 남겼다.

[0449] 도 61은, 일 구현예에 따르면, ClinVar에서 발견된 다른 종들로부터의 미스센스 변이체의 수를 나타내는 보충 표 10이다. 변이체는, 상응하는 인간 변이체와 IBS일 필요가 있고, 동일한 코딩 결과를 보장하도록 판독 프레임의 다른 두 위치에서 동일한 뉴클레오타이드를 가질 필요가 있었다.

[0450] 도 62는, 원래의 DDD 연구에서 이전에는 게놈 전체 유의성 임계값에 도달하지 않았던, 지적 장애가 있는 14개의 추가 후보 유전자의 발견의 일 구현예를 도시하는 표 1이다.

[0451] 도 63은 ClinVar에서 병원성 변이체와 양성 변이체 간의 그랜덤 점수의 평균 차의 일 구현예를 나타내는 표 2이며, 이는 DDD 사례 대 605개의 질환 관련 유전자 내의 대조군 간의 차의 2배이다.

[0452] **데이터 생성**

[0453] 논문에서 사용된 모든 좌표는 다른 종의 변이체에 대한 좌표를 포함하는 인간 게놈 빌드 UCSC hg19/GRCh37을 지칭하며, 이는 이 부문에 기술된 절차를 사용하여 다중 서열 정렬을 사용하여 hg19에 맵핑되었다. 인간과의 99마리 척추동물 게놈의 단백질-코딩 DNA 서열 및 다중 서열 정렬을, hg19 빌드를 위해 UCSC 게놈 브라우저로부터 다운로드하였다(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/alignments/knownCanonical.exonNuc.fa.gz). 다수의 정준 유전자 주석을 갖는 유전자의 경우, 가장 긴 코딩 전사체가 선택되었다.

[0454] 전세계 8개의 소집단으로부터 123,136 개체의 전체 엑솜 서열분석(WES) 데이터를 수집한 엑솜 집계 컨소시엄(ExAC)/게놈 집계 데이터베이스(gnomAD) v2.0으로부터 인간 엑솜 다형성 데이터를 다운로드 하였다(http://gnomad.broadinstitute.org/). ExAC VCF 파일에 주석이 달려있거나 정준 코딩 영역을 벗어난 기본 품질 제어 필터에 실패한 변이체는 제외하였다. 균형맞춤 선택으로 인한 영향을 피하기 위해, 영장류 분석을 위해 확장 MHC 영역(chr6: 28,477,797-33,448,354) 내의 변이체도 배제하였다. 유인원 게놈 서열분석 프로젝트는, (다운스트림 분석을 위해 붕괴된, 수마트라 아종의 5마리 및 보르네아 아종의 5마리를 포함하는) 24마리 침팬지, 13마리 보노보, 27마리 고릴라, 및 10마리 오랑우탄에 대한 전체 게놈 서열분석 데이터 및 유전자형을 제공한다. 침팬지와 보노보에 대한 연구는 추가 35마리 침팬지의 게놈 서열을 제공한다. 그러나, 이러한 추가 침팬지에 대한 변이체들은 유인원 게놈 서열분석 프로젝트와 동일한 방법을 사용하여 콜되지 않았기 때문에, 이들을 대립유전자 빈도 스펙트럼 분석에서 배지하고 심층 학습 모델 트레이닝에만 사용하였다. 이들 영장류 다양성 연구의 변이는 인간 참조(hg19)에 미리 맵핑되었다. 또한, 마모셋과 레서스의 경우, 16개의 레서스 개체 및 9개의 마모셋 개체는 이들 종의 게놈의 초기 서열분석의 변동을 분석하는 데 사용되었지만, 개체 수준 정보는

이용 가능하지 않다.

[0455] 유인원 게놈 서열분석 프로젝트 4는, (다운스트림 분석을 위해 붕괴된, 수마트라 아종으로부터 5마리 및 보르네아 아종으로부터 5마리를 포함하는) 24마리 침팬지, 13마리 보노보, 27마리 고릴라, 및 10마리 오랑우탄에 대한 전체 게놈 서열분석 데이터 및 유전자형을 제공한다. 침팬지와 보노보에 대한 연구는 추가 35마리 침팬지의 게놈 서열을 제공한다. 그러나, 이들 추가 침팬지에 대한 변이체는 유인원 게놈 서열분석 프로젝트와 동일한 방법을 사용하여 콜되지 않았기 때문에, 이들을 대립유전자 빈도 스펙트럼 분석에서 배제하고 심층 학습 모델 트레이닝에만 사용하였다. 이들 영장류 다양성 연구의 변이는 인간 참조(hg19)에 미리 맵핑되었다. 또한, 마모셋과 레서스의 경우, 16개의 레서스 개체 및 9개의 마모셋 개체는 이들 종 게놈의 초기 서열 분석의 변이를 분석하는 데 사용되었지만, 개인 수준 정보는 이용 가능하지 않다.

[0456] 다른 영장류 및 포유류와 비교하기 위해, 레서스, 마모셋, 돼지, 소, 염소, 마우스, 닭, 및 제브라피쉬를 포함하여 dbSNP로부터의 다른 종의 SNP도 다운로드하였다. dbSNP는, 대립유전자 빈도 스펙트럼 분석에 개별 유전자형 정보를 이용할 수 없었기 때문에, 추가 오랑우탄 변이체도 포함되어 있으며, 이것은 심층 학습망을 트레이닝하는 데에만 사용되었다. dbSNP는 해당 종에 대해 제한된 수의 변이체를 제공하므로, 개, 고양이, 또는 양 등의 다른 종을 버렸다.

[0457] 변이체를 인간에 맵핑하기 위해, 인간 단백질-코딩 영역에 대한 이종상동성 1:1 맵핑을 보장하도록 99마리 척추동물의 다중 종 정렬을 사용하였다. 이종상동성 다중 종 정렬을 이용하여 변이체를 맵핑하는 것은, 의사유전자 또는 역전사 서열에 의해 야기되는 아티팩트가 필수적이었으며, 이는 다대1 맵핑을 허용하는 liftOver 등의 도구를 사용하여 종들 간에 SNP를 직접 맵핑할 때에만 발생한다. dbSNP에서 종들의 게놈 빌드가 99-척추동물 다중 서열 정렬에서의 종들의 게놈 빌드와 일치하지 않은 경우, liftOver를 사용하여 변이체를 다중 서열 정렬에 사용된 게놈 빌드로 업데이트하였다. 변이체들이 참조/대체 배향으로 발생하였다면 이들을 IBS인 것으로서 수용하였으며, 예를 들어, 인간 참조가 G이고 대체 대립유전자가 A인 경우, 이는 참조가 A이고 대체 대립유전자가 G인 다른 종의 변이체와 IBS인 것으로 간주되었다. 변이체가 인간 및 다른 종 모두에서 동일한 예측된 단백질-코딩 결과를 갖는 것을 보장하기 위해, 코돈의 다른 두 개의 뉴클레오타이드가 미스센스 변이체 및 동의 변이체 둘 다에 대해 종들 간에 동일할 것을 요구하였다. 분석에 포함된 각 종으로부터의 다형성은 보충 데이터 파일 1에 나열되고, 자세한 메트릭은 보충 표 1에 표시된다.

[0458] 각각의 dbSNP 제출자 일괄로부터의 변이체들이 고품질이고 인간에 정확하게 정렬되었음을 보장하도록, 각각의 일괄에 대한 미스센스:동의 비율을 계산하여, 이것이 예상 비율 2.2:1보다 작음을 확인하였고, 대부분의 종은 1:1보다 낮은 비율, 특히 제브라피쉬와 마우스보다 낮은 비율을 가졌으며, 이는 매우 큰 유효 모집단 크기를 가질 것으로 예상된다. 추가 분석에서는 비정상적으로 높은 미스센스:동의 비율이 있는 소로부터의 SNP들의 일괄 2개(비율이 1.391인 snpBatch_COFACTOR_GENOMICS_1059634.gz, 비율이 2.568인 snpBatch_1000_BULL_GENOMES_1059190.gz)를 배제하였다. 나머지 소 일괄에 대한 평균 미스센스:동의 비율은 0.8:1이었다.

[0459] **미스센스:동의 비율, 돌연변이율, 유전자 드리프트, 및 GC-편향된 유전자 변환에 대한 대립유전자 빈도의 영향의 보정**

[0460] 선별 선택 동작에 더하여, 높은 대립유전자 빈도로 관찰된 인간 미스센스 변이체의 고갈은, 또한, 자연 선택과 관련이 없는 인자에 의해 영향을 받을 수 있다. 모집단에서 특정 대립유전자 빈도로 나타나는 중성 돌연변이의 확률은, 돌연변이율, 유전자 변환, 및 유전자 드리프트의 함수이며, 이들 인자는, 선택적 힘의 부재시 대립유전자 빈도 스펙트럼에 걸친 미스센스:동의 비율에 편향을 잠재적으로 유발할 수 있다.

[0461] 단백질-코딩 선택의 부재 하에서 각각의 대립유전자 빈도 카테고리에서 예상되는 미스센스:동의 비율을 연산하기 위해, 각각의 엑손의 31-50bp 상류에 있는 및 21-50bp 하류에 있는 인트론 영역 내의 변이체들을 선택하였다. 이들 영역은 확장된 스플라이스 모티프의 영향을 피하기에 충분히 먼 거리로 선택되었다. 이들 영역은 ExAC/gnomAD 엑솜에 대한 엑솜 포획 서열의 가장자리 근처에 있기 때문에, 변이체의 공정한 확인을 위해, 모든 chrX 영역을 제거하고 평균 리드 깊이가 <30인 영역을 배제하였다. 각각의 변이체 및 그의 바로 상류 및 하류 뉴클레오타이드는 64개의 트라이뉴클레오타이드 컨텍스트 중 하나에 속한다. 중간 뉴클레오타이드를 3개의 다른 염기로 돌연변이시키면, 총 64×3=192개의 트라이뉴클레오타이드가 가능하다. 트라이뉴클레오타이드 구성 및 이들의 역 보체가 동등하기 때문에, 효과적으로 96개의 트라이뉴클레오타이드 컨텍스트가 존재한다. 트라이뉴클레오타이드 컨텍스트가 돌연변이율에 매우 강한 영향을 미치고, GCbiased 유전자 변환에 미치는 영향이 적어서, 트라이뉴클레오타이드 컨텍스트가 이들 변수를 모델링하는 데 효과적이라는 것을 관찰하였다.

[0462] 이러한 인트론 지역 내에서, 192개의 트라이뉴클레오타이드 컨텍스트 및 대립유전자 빈도의 4개 카테고리(싱글톤, 싱글톤 초과 ~ 0.01%, 0.01% ~ 0.1%, >0.1%)에 기초하여 126,136개의 ExAC/gnomAD 엑솜으로부터 각 변이체를 취하여 4×192개 카테고리로 분리하였다. 4×192개 카테고리(대립유전자 빈도×트라이뉴클레오타이드 컨텍스트) 각각에서 관찰된 변이체의 수를, (세 가지 상이한 방식으로 인트론 서열의 각 뉴클레오타이드를 치환함으로써 취득된) 해당 트라이뉴클레오타이드 컨텍스트를 갖는 가능한 변이체의 총수로 나눔으로써 정규화하였다. 192개의 트라이뉴클레오타이드 컨텍스트 각각에 대해, 단백질 코딩 선택이 없을 때 4개의 대립유전자 빈도 카테고리의 각각에 속하는 변이체의 예상 분율을 수득하였다. 이것은, 트라이뉴클레오타이드 컨텍스트의 차이로 인한 돌연변이율, GC 편향 유전자 변환, 및 유전자 드리프트의 영향을 은연중에 모델링한다(보충 표 7).

[0463] 각각의 대립유전자 빈도 카테고리에서 예상되는 미스센스:동의 비율을 얻기 위해, 단일 뉴클레오타이드 치환에 의해 접근가능한 인간 게놈에서 가능한 동의 및 미스센스 돌연변이의 총수를 계수하고, 이들 각각을 192개의 트라이뉴클레오타이드 컨텍스트 중 하나에 할당하였다. 각 컨텍스트에 대해, 4×192개의 표를 사용하여 4개의 대립유전자 카테고리 각각 내에 속할 것으로 예상되는 변이체의 수를 계산하였다. 마지막으로, 192개의 트라이뉴클레오타이드 컨텍스트에 걸쳐 동의 및 미스센스 변이체의 수를 합산하여, 4개의 대립유전자 빈도 카테고리의 각각에서 동의 및 미스센스 변이체의 총 예상 수를 취득하였다(도 51 및 보충 표 8(도. 59)).

[0464] 기대되는 미스센스:동의 비율은, 대립유전자 주파수 빈도 스펙트럼에 걸쳐 거의 일정하였고, 자연 선택이 없을 때 드 노보 변이체에 대해 예상되는 2.23:1의 비율에 가까웠지만, 미스센스:동의 비율이 2.46:1로 예상되는 싱글톤 변이체는 예외였다. 이는, 단백질-코딩 선택적 압력(돌연변이율, 유전자 변환, 유전자 드리프트)과는 무관한 인자들의 작용으로 인해, ExAC/gnomAD에서 싱글톤 대립유전자 빈도 카테고리를 갖는 변이체가 기본적으로 드 노보 돌연변이의 해당 비율보다 약 10% 높은 미스센스:동의 비율을 가질 것으로 예상됨을 나타낸다. 이를 보정하기 위해, 대립유전자 빈도 분석에서 싱글톤에 대한 미스센스:동의 비율을 10%만큼 하향 조정하였다(도 49a, 49b, 49c, 49d, 49e, 및 50a, 50b, 50c, 및 50d). 이러한 작은 조정은, 영장류 및 다른 포유류(도 49a, 49b, 49c, 49d, 49e, 50a, 50b, 50c, 및 50d에 도시됨)에 존재하는 공통 인간 변이체에 대하여 추정된 미스센스 고갈을 대략 ~3.8% 감소시켰다. 싱글톤 변이체에 대한 더욱 높은 미스센스:동의 비율은, 전환 돌연변이보다 높은 돌연변이율(미스센스 변화를 생성할 가능성이 높음)로 인해 대립유전자 빈도가 더 높은 천이 돌연변이(동의 변화를 생성할 가능성이 높음) 때문이다.

[0465] 더욱이, 이것은 ExAC/gnomAD에서 싱글톤 변이체에 대해 2.33:1의 관찰된 미스센스:동의 비율을 설명하며, 이는 드 노보 돌연변이에 대한 예상 비율인 2.23:1을 초과한다. 미스센스:동의 비율에 대한 대립유전자 빈도 스펙트럼의 영향을 고려한 후, 이것은 사실상 예상과 비교하여 싱글톤 변이체의 5.3% 고갈을 반영하며, 이는 아마도 드 노보 지배적 상속 모드를 갖는 병원성 미스센스 돌연변이에 대한 선택 때문일 것이다. 실제로, 기능 손실 가능성이 높은(pLI>0.9) 일배체 부족 유전자만을 고려할 때, ExAC/gnomAD 싱글톤 변이체에 대한 미스센스:동의 비율은 2.04:1이며, 이는 일배체 부족 유전자 내에서 약 ~17%의 고갈을 나타낸다. 이 결과는, 미스센스 돌연변이의 20%가 어느 정도의 불완전한 침투를 가정할 때 기능 돌연변이의 손실과 동등하다는 이전 추정과 일치한다.

[0466] 또한, 돌연변이율의 큰 차이로 인해 인간 대립유전자 빈도 스펙트럼에 걸쳐 CpG 및 넌-CpG 변이체에 대한 미스센스:동의 비율을 구체적으로 조사하였다(도 52a, 도 52b, 도 52c 및 도 52d). CpG 돌연변이 및 넌-CpG 돌연변이에 대해, 침팬지 공통 다형성과 IBS인 인간 변이체가 대립유전자 빈도 스펙트럼에 걸쳐 거의 일정한 미스센스:동의 비율을 가짐을 확인하였다.

[0467] **<u>다른 종에서의 다형성과 IBS인 인간 미스센스 변이체의 고갈</u>**

[0468] 다른 종들로부터의 변이체가 인간의 흔한 대립유전자 빈도(>0.1%)로 허용될 수 있는지를 평가하기 위해, 다른 종에서의 변이와 IBS인 인간 변이체를 식별하였다. 각 변이체에 대해, 이러한 변이체를 인간 모집단의 해당하는 대립유전자 빈도(싱글톤, 싱글톤 초과 ~ 0.01%, 0.01% ~ 0.1%, >0.1%)를 기준으로 4가지 카테고리 중 하나에 할당하였고, 희귀(<0.1%) 변이체와 공통(>0.1%) 변이체 간의 미스센스:동의 비율(MSR)의 감소를 추정하였다. 흔한 인간 대립유전자 빈도(>0.1%)로 IBS인 미스센스 변이체의 고갈은, 인간의 흔한 대립유전자 빈도로 자연 선택에 의해 필터링될 정도로 충분히 유해한 다른 종으로부터의 변이체의 분율을 나타낸다.

[0469] $$\text{고갈\%} = \frac{MSR_{\text{희귀}} - MSR_{\text{공통}}}{MSR_{\text{희귀}}}$$

[0470] 미스센스:동의 비율 및 고갈 백분율은, 종마다 연산되었고, 도 50b 및 보충 표 2에 도시되어 있다. 또한, 침팬지 공통 변이체(도 49a), 침팬지 싱글톤 변이체(도 49c), 및 포유류 변이체(도 50a)에 대해, 희귀 변이체와과

공통 변이체 간의 미스센스:동의 비율의 차가 유의했는지를 테스트하도록 2×2 분할표에 대한 카이-제곱(χ^2) 균질성 테스트를 수행하였다.

[0471] 서열분석은, 유인원 다양성 프로젝트로부터 제한된 수의 개체에 대해서만 수행되었기 때문에, ExAC/gnomAD의 인간 대립유전자 빈도 스펙트럼을 사용하여 일반적인 침팬지 모집단의 희귀<0.1%) 또는 공통(>0.1%) 샘플링된 변이체의 분율을 추정하였다. ExAC/gnomAD 대립유전자 빈도를 기반으로 24명 개체의 코호트를 샘플링하였고, 이 코호트에서 한 번 또는 한 번 이상 관찰된 미스센스 변이체를 식별하였다. 한 번보다 많게 관찰된 변이체는 일반 모집단에서 99.8%의 공통 확률(>0.1%)을 보인 반면, 코호트에서 한 번만 관찰된 변이체는 일반 모집단에서 69%의 공통 확률을 보였다. 도 49b 및 도 49c에서는, 침팬지 싱글톤 변이체 중 일부가 희귀한 유해한 돌연변이인 결과로, 인간에게서 높은 대립유전자 빈도로 싱글톤 침팬지 변이체의 고갈을 관찰하지만, 공통 침팬지 변이체에 대해서는 그렇지 않다는 것을 도시한다. 24명 개체의 코호트에서 관찰된 침팬지 변이체의 대략 절반은 한 번만 관찰되었고, 대략 절반은 한 번보다 많게 관찰되었다.

[0472] 보다 먼 포유류에서 관찰된 미스센스 변이체에 대한 고갈이 보다 잘 보존된 유전자의 혼동 효과로 인한 것이 아니므로 더욱 정확하게 정렬됨을 확인하기 위해, 상기 분석을 반복하여, 인간과 비교하여 11마리의 영장류 및 50마리의 포유류의 다중 서열 정렬에서의 50% 초과의 평균 뉴클레오타이드 동일성을 갖는 유전자로만 제한하였다(보충 표 3 참조). 이는 결과에 실질적으로 영향을 미치지 않으면서 분석에서 약 7%의 인간 단백질 코딩 유전자를 제거하였다.

[0473] **영장류, 포유류, 및 먼 척추동물 간의 고정된 치환**

[0474] dbSNP 변형을 사용한 본 발명의 결과가 변이체 데이터 또는 (dbSNP로부터 선택된 대부분의 종이 사육되었으므로) 사유 아티팩트 문제에 의한 영향을 받지 않도록, 종간 다형성 대신 밀접하게 관련된 종들의 쌍들로부터 고정된 치환을 사용하여 분석을 또한 반복하였다. UCSC 게놈 브라우저로부터 척추동물 종 100마리(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh)의 계통 발생 트리를 분기 길이(위치당 뉴클레오타이드 치환의 평균 수)로 측정된 계통 길이와 함께 다운로드하였다. 추가 분석을 위해 밀접하게 관련된 종 쌍(분기 길이<0.25)을 선택하였다. 밀접하게 관련된 종 쌍 간의 고정된 치환을 식별하기 위해, 인간과의 99마리 척추동물 게놈의 다수의 서열 정렬뿐만 아니라 UCSC 게놈 브라우저로부터 인간과의 19마리 포유류(16마리 영장류) 게놈의 정렬에 대한 코딩 영역을 다운로드하였다. 보노보와 같은 영장류 종의 일부가 99마리 척추동물 정렬에 없었기 때문에 추가적인 19마리 포유류 다중 종 정렬이 필요하였다(http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way/alignments/knownCanonical.exonNuc.fa.gz). 전체적으로, 도 50d 및 보충 표 4에 열거된 바와 같이, 5개의 영장류 쌍을 포함하여 밀접하게 관련된 종 15쌍을 취득하였다.

[0475] 정준 코딩 영역 내에서 인간과 19마리 포유류 또는 99마리 척추동물 게놈의 다중 서열 정렬을 취하고, 보충 데이터 파일 2에 열거된 각각의 선택된 척추동물 쌍 간의 뉴클레오타이드 치환을 취득하였다. 이들 치환은 인간 게놈에 맵핑되었으며, 코돈 내의 다른 2개의 뉴클레오타이드가 인간과 다른 종 간에 변하지 않아야 했으며, 참조 또는 대체 배향으로 변이체를 수용했다. 관련 종 쌍으로부터 고정된 치환과 IBS인 인간 변이체를 사용하여, 희귀(<0.1%) 및 흔한(>0.1%) 대립유전자 빈도 카테고리의 변이체에 대한 미스센스:동의 비율을 계산하여, 보충 표 4에 나타낸 바와 같이, 음성 선택 하에서 고정된 치환의 분율을 취득하였다.

[0476] **인간, 영장류, 포유류, 및 다른 척추동물에 대한 다형성 데이터의 ClinVar 분석**

[0477] 다른 종과 IBS인 변이체의 임상적 영향을 조사하기 위해, ClinVar 데이터베이스에 대한 릴리스 변이체 요약을 다운로드하였다(ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_20171029.vcf.gz released on 02-Nov-2017)12. 데이터베이스는 hg19 게놈 빌드에서 324,698개의 변이체를 포함했으며, 이중 122,884개가 단백질-코딩 유전자의 본 발명의 리스트에 대한 미스센스 단일 뉴클레오타이드 변이체였다(보충 표 9). ClinVar 데이터베이스에 있는 대부분의 변이체는 미스센스 결과가 아니었으며 제외되었다. 다음에, 병원성에 대한 상충되는 해석으로 변이체를 필터링하였고, 양성, 유사 양성, 병원성, 및 유사 병원성의 주석이 있는 것들만을 유지했다. 양성 또는 유사 양성의 주석이 있는 변이체를 단일 카테고리로 병합하고, 병원성 또는 유사 병원성의 주석이 있는 변이체도 병합하였다. 보충 표 9에 제시된 필터링 단계에 따라, 병원성 카테고리에는 총 24,853개의 변이체가 있고 양성 카테고리에는 17,775개의 변이체가 있으며, 나머지는 알려지지 않은 유의성 또는 상충되는 주석의 변이체이기 때문에 제외되었다.

[0478] 인간 모집단에서 ClinVar 미스센스 변이체에 대한 베이스라인을 얻기 위해, ExAC/gnomAD 대립유전자 빈도로부터

샘플링된 30명의 개체의 코호트에서 ClinVar 미스센스 변이체를 조사하였다. 이 코호트 크기는, 영장류 다양성 프로젝트 연구에서 서열분석된 개체의 수를 대략적으로 반영하도록 선택되었다. 100개의 이러한 시뮬레이션으로부터 30명의 인간(도 49e)의 코호트에서 병원성 변이체와 양성 변이체의 평균 수를 보고한다. 큐레이터는 ClinVar에서 양성 결과로 공통 인간 변이체에 체계적으로 주석을 달았기 때문에, 이러한 큐레이션 편향을 피하기 위해 1% 초과의 대립유전자 빈도를 갖는 변이체를 배제하였다.

[0479] 영장류, 포유류 및 다른 척추동물에서의 변이와 IBS인 ClinVar 변이체를 분석하였다. 각 종에 대한 양성 변이체와 병원성 변이체의 수는 보충 표 10에 나와 있다. 인간, 영장류, 및 더 먼 포유류에 존재하는 ClinVar 변이체의 수의 요약은, 양성 변이체 대 병원성 변이체의 비율의 차에 대한 균일성의 카이-제곱(χ^2) 테스트로부터의 결과와 함께 도 49e 및 도 50b에 나타낸다.

[0480] **<u>모델 트레이닝을 위한 양성 변이체의 생성</u>**

[0481] 인간 모집단에서 공통적인 변이체는, 희귀한 경우의 설립자 효과 또는 균형맞춤 선택과는 별도로 대체로 중립적이기 때문에, 인간 해석의 편향에 영향을 받지 않는 기계 학습을 위한 양성 트레이닝 데이터세트로 적합하게 된다. 필터를 통과하지 않은 변이체를 제외하고 ExAC/gnomAD 데이터베이스(릴리스 v2.0)의 123,136개 엑솜의 대립유전자 빈도 데이터를 사용하여, 정준 단백질-코딩 전사물 내의 전체 모집단 대립유전자 빈도≥0.1%인 83,546개의 미스센스 변이체가 남게 되었다.

[0482] 영장류에 존재하는 변이체들이 인간에게서 대체로 양성임을 나타내는 본 발명자들의 초기 결과에 기초하여, 공통 인간 변이체(>0.1% 대립유전자 빈도), 유인원 다양성 프로젝트와 추가 영장류 서열분석으로부터의 침팬지, 보노보, 고릴라, 및 오랑우탄, 및 dbSNP로부터의 레서스, 오랑우탄 및 마모셋 변이체를 포함하는 기계 학습을 위한 양성 트레이닝 데이터세트를 생성하였다. 전체적으로, 일 구현예에 따르면 301,690개의 고유 영장류 변이체가 양성 트레이닝 세트에 추가되었다. 각 공급원에 의해 제공된 양성 트레이닝 변이체의 수는 보충 표 5에 나와 있다.

[0483] 주의할 점은, 대부분의 영장류 변이체가 각 모집단에서 공통적이지만, 이들 중 소수는 희귀한 변이체라는 점이다. 비인간 영장류 종은 서열분석된 한정된 개체 수를 가졌으므로, 확인된 변이체들의 세트가 일반적으로 공통 변이를 나타낼 것으로 예상된다. 실제로, 각각의 영장류 종으로부터의 변이체에 대한 미스센스:동의 비율이 드노보 돌연변이에 대해 예상되는 2.23:1 비율의 절반보다 작다는 것을 발견하였으며, 이는 선택 체를 미리 통과한 대부분의 공통 변이체임을 나타낸다. 또한, 침팬지 코호트에 대해, 확인된 변이체의 ~84%가 각각의 해당 모집단에서 흔한 대립유전자 빈도(>0.1%)로 존재하는 것으로 추정되었다. 새로 발생하는 미스센스 돌연변이의 ~50%가 흔한 인간 대립유전자 빈도(>0.1%)로 선별 선택에 의해 필터링되므로(도 49a), 이 수치는, ~16%의 희귀 변이체와 일치하여, 관찰된 영장류 변이와 IBS인 인간 미스센스 변이체의 8.8%의 관찰된 고갈을 나타낸다(도 49d).

[0484] 인간 미스센스 돌연변이의 ~20%가 기능 등가의 손실이라는 추정을 적용하면, 영장류 변이체는 3.2%의 완전 병원성 돌연변이, 91.2%의 양성 돌연변이(>0.1% 대립유전자 빈도로 허용됨), 및 유전자 기능을 완전히 파괴하지는 않지만 흔한 대립유전자 빈도(>0.1%)로 필터링될 정도로 해로운 5.6%의 중간 돌연변이를 포함할 것으로 예상된다. 이러한 트레이닝 데이터세트의 알려진 결함에도 불구하고, 심층 학습망의 분류 정확도는, 공통 인간 변이체만의 경우와 비교할 때, 공통 인간 변이체와 영장류 변이체를 모두 포함하는 양성 트레이닝 데이터세트에 대해 트레이닝될 때 훨씬 우수하였다. 따라서, 현재 분류 정확도에서, 이용 가능한 트레이닝 데이터의 양이 더욱 강력한 제한사항인 것으로 보인다. 더 많은 수의 개체가 각 영장류 종에서 서열분석됨에 따라, 더 높은 비율의 영장류 공통 변이체를 포함하는 트레이닝 데이터세트를 준비할 수 있어서, 트레이닝 데이터세트에서 병원성 변이체의 오염을 줄이고 분류 성능을 더욱 개선할 수 있다.

[0485] **<u>양성 트레이닝 데이터세트를 보완하기 위한 미표지 변이체의 생성</u>**

[0486] 모든 가능한 미스센스 변이체는, 해당 위치에서의 뉴클레오타이드를 다른 3개의 뉴클레오타이드로 치환함으로써 정준 코딩 영역의 각 염기 위치로부터 생성되었다. ExAC/gnomAD로부터 123,136개의 엑솜에서 관찰된 변이체 및 시작 또는 정지 코돈의 변이체는 제외하였다. 전체적으로, 68,258,623개의 미표지 변이체가 생성되었다. 각각의 미표지 변이체를 96개의 상이한 트라이뉴클레오타이드 컨텍스트 카테고리 중 하나에 할당하였다. 트라이뉴클레오타이드 컨텍스트에 의해 양성 데이터세트의 변이체와 일치하는 이러한 미표지 데이터세트의 변이체를 샘플링하고 양성 트레이닝 예와 미표지 트레이닝 예를 구별하도록 분류자를 트레이닝함으로써, 반감독 접근법을 사용하여 심층 학습망을 트레이닝하였다.

[0487] 미표지 **변이체의 추가 필터링**

[0488] 플랭킹 아미노산 서열과 함께 양성 변이체 및 미표지 변이체의 예를 제시함으로써, 심층 학습망은 돌연변이에 대해 내성이 강한 단백질 영역을 학습한다. 그러나, 단백질 서열의 영역 내에서의 공통 변이체의 부재는, 강력한 선별 선택에 기인할 수 있고 또는 변이체가 해당 영역에서 콜되는 것을 방지하는 기술적 아티팩트로 인한 것일 수 있다. 후자를 보정하기 위해, ExAC/gnomAD 데이터세트의 평균 커버리지가 1 미만인 영역으로부터 양성 데이터세트 및 미표지 데이터세트로부터의 변이체를 제거하였다. 유사하게, 트레이닝 동안 양성 데이터세트의 영장류 변이체에 미표지 변이체를 일치시킬 때, 해당 영장류가 다중 서열 정렬에 있어서 인간과 이종상동성 정렬 가능 서열을 갖지 않은 영역으로부터 미표지 변이체를 제외시켰다.

[0489] **유효성확인 및 테스트를 위한 보류된 영장류 변이체 및 영향을 받은 개체와 영향을 받지 않은 개체로부터의 드 노보 변이체**

[0490] 심층 학습망의 유효성확인 및 테스트를 위해, 유효성확인 및 테스트를 위해 10,000개의 영장류 변이체의 두 개 세트를 랜덤하게 샘플링하였으며, 이러한 세트는 트레이닝으로부터 보류되었다. 공통 인간 변이체(>0.1% 대립유전자 빈도)와 함께 영장류 변이체의 나머지는, 심층 학습망을 트레이닝하기 위한 양성 데이터세트로서 사용되었다. 또한, 유효성확인 세트 및 테스트 세트에 대해 보류된 영장류 변이체에 일치한 10,000개의 미표지 변이체의 두 개 세트를 샘플링하였다.

[0491] 유효성확인 세트에서의 10,000개의 보류된 영장류 변이체 및 일치된 10,000개의 미표지 변이체를 사용하여, 심층 학습망이 두 개 세트의 변이체를 구별하는 능력을 측정함으로써 트레이닝 과정 동안 심층 학습망의 성능을 감시하였다. 이는, 일단 망의 성능이 포화되었다면 트레이닝 중단점을 결정하고 과적합을 피할 수 있게 하였다.

[0492] 테스트 데이터세트에서의 10,000개의 보류된 영장류 변이체를 사용하여 다른 20개의 분류자뿐만 아니라 심층 학습망도 벤치마킹하였다. 상이한 분류자가 매우 가변적인 점수 분포를 가졌기 때문에, 이러한 미표지 변이체를 사용하여 각 분류자에 대한 50-백분위 임계값을 식별하였다. 방법들 간의 공정한 비교를 위해, 해당 분류자의 50-백분위 임계값에서 양성으로 분류된 10,000개의 보류된 영장류 변이체 테스트세트에 있는 변이체들의 분율에 대해 각 분류자를 벤치마킹하였다.

[0493] 신경 발달 장애가 있는 영향을 받은 개체의 드 노보 변이체 및 건강한 대조군에서의 드 노보 변이체를 사용하여 임상 환경에서 심층 학습망의 성능을 평가하기 위해, 발달 장애 해독(DDD) 연구에서 드 노보 변이체를 다운로드하고 사이몬 심플렉스 수집(SSC) 자폐증 연구에서 건강한 형제로부터 드 노보 변이체를 다운로드하였다. DDD 연구는 드 노보 변이체에 대한 신뢰 수준을 제공하며, 임계값이 <0.1인 DDD 데이터세트로부터의 변이체를 변이체 콜 에러로 인한 잠재적 위양성으로서 제외하였다. 전체적으로, DDD 영향을 받은 개체로부터 3,512개의 미스센스 드 노보 변이체 및 건강한 대조군의 1,208개의 미스센스 드 노보 변이체를 가졌다.

[0494] 후보 질환 유전자들의 패널 내에서 불확실한 유의성의 양성 변이체 및 병원성 변이체를 구별하는 실제 임상 시나리오를 보다 잘 모델링하기 위해, 분석을, 단백질-절단 변이로부터만 계산된, DDD 연구에서 질병에 연관된 605개 유전자 내의 드 노보 변이체($p<0.05$)로만 제한하였다(보충 표 18). 유전자-특이적 돌연변이율 및 고려된 염색체의 수를 고려하여 드 노보 돌연변이의 예상된 수의 귀무가설하에서 통계적 유의성을 연산함으로써, 단백질-절단 드 노보 돌연변이의 유전자-특이적 농축을 평가하였다. 공칭 P-값이 <0.05인 605개의 유전자를 선택하였다. 본 발명자들은 605개 유전자 내에서의 동의 및 미스센스 드 노보 돌연변이의 초과량을 관찰된 드 노보 돌연변이 빼기 예상된 드 노보 돌연변이의 차뿐만 아니라 관찰된 드 노보 돌연변이 대 예상된 드 노보의 계수치의 비율로서 계산하였다(도 22a). 이들 605개 유전자 내에서, DDD 영향을 받은 개체로부터 380개의 드 노보 미스센스 돌연변이를 관찰하였다(도 22a). 각 분류자에 대하여, 자체 분류를 포함하여, 변이체들의 작은 분율은, 예측되지 않았으며, 그 이유는 그 분율이 일반적으로 분류자에 의해 사용되는 동일한 전사 모델에 맵핑되지 않았기 때문이다. 따라서, 본 발명의 심층 학습망에 대하여, DDD 영향을 받은 개체로부터의 362개의 드 노보 미스센스 돌연변이 및 건강한 대조군으로부터의 65개의 드 노보 미스센스 돌연변이를 사용하여 도 22a, 도 22b, 도 22c, 도 22d, 도 22e에서 하향 분석을 수행하였다.

[0495] **서열분석된 영장류 모집단의 수의 증가와 함께 모든 가능한 인간 미스센스 돌연변이의 포화**

[0496] 504개의 현존하는 영장류 종에 존재하는 공통 변이체에 의해 모든 ~70M 가능한 인간 미스센스 돌연변이의 예상 포화를 조사하였다. 각 영장류 종에 대해, 인간은 다른 영장류 종의 개체당 대략 절반의 변이체 수를 갖고 약 ~>0.1% 대립유전자 빈도로 선별 선택에 의해 인간 미스센스 변이체의 약 50%를 필터링하였기 때문에, 인간에서 관찰된 일반적인 미스센스 변이체의 수의 4배(대립유전자 빈도>0.1%인 ~83,500개의 미스센스 변이체)를 시뮬레

이션하였다(도 49a). 96개의 트리클레오타이드 컨텍스트의 인간 공통 미스센스 변이체의 관찰된 분포에 기초하여 시뮬레이션된 변이체를 할당하였다. 예를 들어, 인간 공통 미스센스 변이체의 2%가 CCG>CTG 트라이뉴클레오타이드 컨텍스트에서 온 경우, 시뮬레이션된 변이체의 2%가 랜덤하게 샘플링된 CCG>CTG 돌연변이이어야 한다. 이것은 트라이뉴클레오타이드 컨텍스트를 사용하여 돌연변이율, 유전자 드리프트, 및 유전자 변환 편향의 영향을 제어하는 효과를 갖는다.

[0497] 도 23d의 곡선은, 각 영장류 종(>0.1% 대립유전자 빈도)에서 모든 공통 변이체를 확인한다고 가정할 때, 504개 영장류 종 중 임의의 종에 존재하는 공통 변이체에 의한 ~70M 가능한 인간 미스센스 돌연변이의 누적 포화도를 나타낸다. 도 49a로부터, 대략 ~50%의 인간 미스센스 돌연변이는 인간 및 다른 영장류가 흔한 대립유전자 빈도(>0.1%)로 상승하는 것을 방지하기 위해 충분히 해롭고, 따라서, 도 23d의 곡선은, 영장류 종의 수가 증가함에 따라 공통 영장류 변이에 의해 포화된 비유해성 인간 미스센스 돌연변이의 분율을 나타낸다. 504개 영장류 종에 있어서, 비유해성 인간 미스센스 돌연변이의 대다수가 포화될 것이며, 비유해성 CpG 돌연변이는 그들의 높은 돌연변이율로 인해 훨씬 적은 수의 종으로 포화되는 것으로 나타났다.

[0498] 조사된 인간 코호트의 크기가 증가함에 따라 발견된 인간 공통 미스센스 변이체(>0.1% 대립유전자 빈도)의 분율을 모델링하기 위해(도 36), gnomAD 대립유전자 빈도에 따라 유전자형을 샘플링하였다. 발견된 gnomAD 공통 미스센스 변이체의 분율은 10K 내지 100K의 각 샘플 크기에 대해 100개의 시뮬레이션에 걸쳐 평균화되었다.

[0499] **이차 구조 및 용매 접근성의 예측**

[0500] 병원성 예측을 위한 심층 학습망은, 이차 구조 및 용매 접근성 예측망을 위한 19개의 컨볼루션층 및 이차 구조와 용매 접근성 망의 결과를 입력으로서 취하는 주요 병원성 예측망을 위한 17개의 컨볼루션층을 포함하는 총 36개의 컨볼루션층을 포함한다. 대부분의 인간 단백질의 결정 구조는 알려져 있지 않기 때문에, 망이 일차 서열로부터 단백질 구조를 학습할 수 있도록 두 가지 모델을 트레이닝하였다. 두 모델 모두 도 6에 도시된 동일한 망 아키텍처 및 입력을 사용하였다. 이차 구조 및 용매 접근성 망에 대한 입력은, 99마리의 다른 척추동물과의 인간의 다중 서열 정렬로부터 보존 정보를 암호화하는 51 길이 × 20 아미노산 위치 빈도 행렬이다.

[0501] 이차 구조 망은, 3-상태 이차 구조, 즉, 알파 나선(H), 베타 시트(B), 및 코일(C)을 예측하도록 트레이닝된다. 용매 접근성 망은, 매립(B), 개재(I) 및 노출(E)의 3-상태 용매 접근성을 예측하도록 트레이닝된다. 양측 망 모두, 일차 서열을 입력으로서 사용하며 단백질 데이터뱅크의 알려진 결정 구조로부터의 표지를 사용하여 트레이닝되었다. 모델은 각 아미노산 잔기에 대한 하나의 상태를 예측한다.

[0502] **이차 구조와 용매 접근성의 예측을 위한 데이터 준비**

[0503] 모델 트레이닝을 위해 단백질 데이터뱅크로부터 관련이 없는 결정 구조를 사용하였다. 25% 초과의 서열 유사성을 갖는 아미노산 서열이 제거되었다. 전체적으로, 6,367개의 단백질 서열이 트레이닝에 사용되었고, 400개는 유효성확인에, 500개는 테스트에 사용되었다(보충 표 13). 아미노산 서열, 이차 구조 및 용매 접근성 표지를 포함하여 트레이닝에 사용된 데이터는 RaptorX 웹사이트 (http://raptorx.uchicago.edu/download/)에서 이용 가능하다.

[0504] 대부분의 용해된 결정 구조는 비인간 단백질이며, 따라서 이차 구조 및 용매 모델을 미리 트레이닝하도록, 인간-기반 다중 서열 정렬을 일반적으로 이용할 수 없었으므로, RaptorX 스위트(PSI-BLAST에 기초함)를 사용하여 관련 서열을 취득하였다. RaptorX에서 CNFsearch1.66_release 도구를 사용하여 단백질에 대한 다중 서열 정렬을 생성하였고, 99개의 가장 가까운 정렬로부터의 각 위치에 있는 아미노산을 계수하여 위치 빈도 행렬을 형성하였다. 예를 들어, 1u71.fasta 단백질에 대한 다중 서열 정렬을 검색하기 위해 RaptorX를 사용하는 특정 커맨드는 다음과 같다.

[0505]
```
% /buildFeature -i 1u71A.fasta -c 10 -o ./TGT/1u71A.tgt
% /CNFsearch -a 30 -q 1u71A
```

[0506] 데이터세트의 각각의 아미노산 위치에 대해, 플랭킹 51 아미노산에 상응하는 위치 빈도 행렬로부터 윈도를 취하고, 이를 이용하여 51-길이의 아미노산 서열의 중심에 있는 아미노산에 대한 이차 구조 또는 용매 접근성에 대한 표지를 예측하였다. 이차 구조 및 상대 용매 접근성에 대한 표지는, DSSP 소프트웨어를 사용하여 단백질의 공지된 3D 결정 구조로부터 직접 취득되었으며 일차 서열로부터의 예측을 필요로 하지 않았다. 병원성 예측 망의 일부로서 이차 구조와 용매 접근성 망을 통합하기 위해, 인간-기반 99마리 척추동물 다중 서열 정렬로부터 위치 빈도 행렬을 연산하였다. 이러한 두 가지 방법으로부터 생성된 보존 행렬은 대략적으로 유사하지만, 파라

미터 가중치의 미세 조정을 허용하기 위해 병원성 예측 트레이닝 동안 이차 구조 및 용매 접근성 모델을 통한 역전파를 가능하게 하였다.

[0507] **모델 아키텍처 및 트레이닝**

[0508] 단백질의 이차 구조와 상대 용매 접근성을 예측하기 위해 두 개의 개별적인 심층 컨볼루션 신경망 모델을 트레이닝하였다. 두 모델의 아키텍처와 입력 데이터는 동일하지만, 예측 상태가 다르다. 최상의 성능을 위해 모델을 최적화하도록 세부 하이퍼파라미터 검색을 수행하였다. 병원성 예측을 위한 본 발명의 심층 학습망과 이차 구조 및 용매 접근성을 예측하기 위한 심층 학습망 모두는 잔여 블록의 구조를 채택했으며, 이러한 구조는 이미지 분류의 성공으로 인해 널리 채택되어 왔다. 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 컨볼루션 레이어가 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. 상세한 아키텍처는 도 6 및 보충 표 11(도 7a 및 7b) 및 보충 표 12(도 8a 및 도 8b)에 제공되어 있다.

[0509] 잔여 블록에 이어서, 소프트맥스 층은, 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. 일단 망이 이차 구조 및 용매 접근성에 대해 미리 트레이닝되었다면, 망의 출력을 병원성 예측망의 입력으로서 직접 취하는 것 대신, 소프트맥스 층 이전의 층을 취하였고, 이에 따라 더욱 많은 정보가 병원성 예측망에 전달된다.

[0510] 3-상태 이차 구조 예측 모델에 대해 달성된 최상의 테스트 정확도는, DeepCNF 모델30에 의해 예측된 최신 정확도와 유사한 79.86%(보충 표 14)이다. 3-상태 용매 접근성 예측 모델에 대한 최상의 테스트 정확도는, 유사한 트레이닝 데이터세트에서 RaptorX에 의해 예측된 현재 최고의 정확도와 유사한 60.31%(보충 표 14)이다. 또한, 예측 구조 표지만을 사용하는 표준 PrimateAI 모델을 사용하는 것에 비교하여 결정 구조를 가진 대략 ~4000개의 인간 단백질에 대해 DSSP 주석 구조 표지를 사용할 때 신경망의 예측을 비교하였다. DSSP 주석이 있는 표지를 사용할 때 병원성 예측 정확도의 추가 개선을 취득하지 못했다(보충 표 15).

[0511] **병원성 예측을 위한 심층 학습 모델에 대한 입력 피처**

[0512] 병원성 예측망에 대한 트레이닝 데이터세트는, 필터링 후 385,236개의 표지된 양성 변이체 및 68,258,623개의 미표지 변이체를 포함한다. 각 변이체에 대해, 다음과 같은 입력 피처를 생성하였다. 각 변이체의 제1 입력 피처는, 변이체의 서열 컨텍스트를 심층 학습 모델에 제공하기 위해 hg19의 참조 서열로부터 취득된 변이체의 각 측면에 대한 자신의 51-길이 플랭킹 아미노산 서열, 즉, 25개 아미노산이다. 전체적으로, 이러한 플랭킹 참조 서열은 51개의 아미노산 길이이다. 경험적 관찰을 통해, 단백질 서열의 아미노산 표현이 뉴클레오타이드를 사용한 단백질 코딩 서열을 나타내는 것보다 효과적이라는 것을 발견하였다.

[0513] 제2 피처는, 변이체에 의해 중심 위치에서 대체 아미노산으로 치환된 51-길이 플랭킹 인간 아미노산 서열이다. 대체 플랭킹 서열은, 그 서열의 중간 위치가 참조 아미노산 대신 대체 아미노산을 포함한다는 점을 제외하고는 제1 피처의 참조 플랭킹 서열과 동일하다. 참조 인간 아미노산 서열과 대체 인간 아미노산 서열 모두는, 길이 51×20의 원-핫 인코딩된 벡터로 변환되었고, 여기서 각각의 아미노산은 0의 값을 갖는 19개의 아미노산 벡터 및 1의 값을 갖는 단일 아미노산의 벡터에 의해 표현된다.

[0514] 3개의 위치 빈도 행렬(PFM)은 변이체에 대한 99마리 척추동물의 다중 서열 정렬로부터 생성되는데, 하나는 11마리의 영장류에 대한 것이고, 하나는 영장류를 제외한 50마리 포유류에 대한 것이고, 하나는 영장류와 포유류를 제외한 38마리의 척추동물에 대한 것이다. 각각의 PFM은 치수 L×20을 갖고, L은 변이체 주위의 플랭킹 서열의 길이이다(본 발명의 경우에, L은 51개 아미노산을 나타낸다).

[0515] 미리 트레이닝된 3-상태 이차 구조 및 3-상태 용매 접근성 망으로의 입력을 위해, 또한 길이 51 및 깊이 20인 모든 99마리 척추동물에 대한 다중 서열 정렬로부터 생성된 단일 PFM 행렬을 사용하였다. 단백질 데이터뱅크로부터 알려진 결정 구조에 대한 망을 미리 트레이닝한 후, 이차 구조 및 용매 모델에 대한 최종 두 개의 층(글로벌 최대 풀링층 및 출력층)이 제거되었고, 이전 층의 출력의 51×40 형상이 병원성 예측망에 대한 입력으로서

사용되었다. 망의 구조적 층들을 통한 역전파를 허용하여 파라미터를 미세 조정할 수 있었다.

[0516] **반감독 학습**

[0517] 반감독 학습 알고리즘은, 트레이닝 프로세스에 있어서 표지된 인스턴스와 미표지 인스턴스를 모두 사용하기 때문에, 트레이닝에 사용할 수 있는 표지된 적은 양의 데이터만을 갖는 완전히 감독되는 학습 알고리즘보다 나은 성능을 달성하는 분류자들을 생성할 수 있다. 반감독 학습의 원칙은, 표지된 인스턴스만을 사용하는 감독 모델의 예측 기능을 강화하기 위해 미표지 데이터 내의 고유한 지식을 활용할 수 있어서 반감독 학습을 위한 잠재적 이점을 제공한다는 것이다. 소량의 표지된 데이터로부터 감독 분류자에 의해 학습된 모델 파라미터는, 미표지 데이터에 의해 보다 실제적인 분포(테스트 데이터의 분포와 더욱 유사함)로 조정될 수 있다.

[0518] 생물 정보학에서 널리 알려진 또 다른 과제는 데이터 불균형 문제이다. 데이터 불균형 현상은, 해당 클래스에 속하는 인스턴스가 드물거나(주목할 만한 사례) 취득하기 힘들기 때문에, 예측할 클래스들 중 하나가 데이터에 잘 나타나지 않을 때 발생한다. 아이러니하게도, 소수의 클래스는 통상적으로 특별한 사례에 연관될 수 있기 때문에 학습하는 것이 가장 중요하다.

[0519] 불균형 데이터 분포를 처리하기 위한 알고리즘 접근법은 분류자들의 앙상블에 기초한다. 제한된 양의 표지된 데이터는 자연스럽게 약한 분류자로 이어지지만, 약한 분류자들의 앙상블은 임의의 단일 구성요소 분류자의 성능을 능가하는 경향이 있다. 또한, 앙상블은, 통상적으로 여러 학습 모델에 연관된 노력과 비용을 유효성확인하는 인자에 의해 단일 분류자로부터 취득되는 예측 정확도를 개선한다. 직관적으로, 여러 분류자를 집계하면, 개별 분류자의 높은 변동성을 평균화하는 것이 또한 분류자의 과적합을 평균화하기 때문에, 과적합 제어가 향상된다.

[0520] 확실하게 표지된 병원성 변이체의 적절한 크기의 데이터세트의 부족으로 인해 반감독 학습 전략을 추구하였다. ClinVar 데이터베이스에는 300,000개가 넘는 엔트리가 있지만, 불확실한 유의성의 변이체를 제거한 후, 병원성에 대한 상충되는 해석으로 ~42,000개의 미스센스 변이체만이 남았다.

[0521] 체계적인 검토에 따르면, 또한, 이들 엔트리가 종종 주석이 달린 병원성을 지지하기에 불충분한 임상적 증거를 갖는다고 밝혀졌다. 더욱이, 인간에 의해 큐레이션된 데이터베이스에서의 대부분의 변이체는 유전자들의 매우 작은 세트 내에 있는 경향이 있어서, 이러한 변이체가 양성 트레이닝 데이터세트의 변이체와 일치하지 않게 되는데, 이는 인간 공통 변이체 또는 침팬지-인간 고정 치환을 사용하여 게놈 전체에서 확인된다. 데이터세트가 어떻게 다르게 확인되었는지를 고려할 때, 인간에 의해 큐레이션된 변이체를 병원성 세트로서 사용하고 게놈 전체 공통 변이체를 양성 세트로서 사용하여 감독 학습 모델을 트레이닝하면 상당한 편향이 도입될 수 있다.

[0522] 표지된 양성 변이체들의 한 세트와 조심스럽게 일치된 미표지 변이체들의 세트를 구별하여 편향을 제거하도록 심층 학습망을 트레이닝하였다. 일 구현예에 따르면, 385,236개의 표지된 양성 변이체들의 세트는, ExAC/gnomAD 데이터베이스로부터의 인간 공통 변이체(>0.1% 대립유전자 빈도) 및 6종의 비인간 영장류로부터의 변이체를 포함하였다.

[0523] (돌연변이율, 유전자 드리프트, 및 유전자 변환을 제어하기 위해) 트라이뉴클레오타이드 컨텍스트에서 양성 변이체와의 일치를 필요로 하고 변이체 확인에 대한 정렬성 및 서열 커버리지의 영향을 조정하는 미표지 변이체들의 세트를 샘플링하였다. 미표지 변이체의 수가 표지된 양성 변이체를 크게 초과하기 때문에, 표지된 양성 변이체들의 동일한 세트 및 미표지 변이체들의 랜덤하게 샘플링된 8개 세트를 사용하는 8개 모델을 트레이닝하고 이들의 예측값의 평균을 취함으로써, 컨센서스 예측을 얻었다.

[0524] 반감독 학습을 선택하는 동기는, 인간에 의해 큐레이션된 변이체 데이터베이스가 신뢰할 수 없고 노이즈 있으며, 특히, 신뢰할 수 있는 병원성 변이체가 없다는 점이다. gnomAD와 영장류 변이체들로부터 인간 공통 변이체로부터 신뢰할 수 있는 양성 변이체들의 세트를 취득하였다. 병원성 변이체의 경우, 알려지지 않은 변이체들(주석을 갖는 임상적 유의성이 없는 VUS 변이체)의 세트로부터 병원성 변이체들을 최초 샘플링하도록 반복적 균형맞춤 샘플링 방법을 채택하였다.

[0525] 샘플링 편향을 감소시키기 위해, 양성 변이체들의 동일한 세트 및 병원성 변이체들의 8개의 상이한 세트를 사용하는 8개의 모델의 앙상블을 트레이닝하였다. 처음에, 병원성 변이체를 나타내도록 알려지지 않은 변이체들을 랜덤하게 샘플링하였다. 그 다음, 반복하여, 모델의 앙상블을 사용해서 이전 트레이닝 사이클에 포함되지 않았던 알려지지 않은 변이체들의 세트를 점수 매긴다. 이어서, 최고 점수를 갖는 병원성 변이체들을 취득하여 이전 사이클에서 알려지지 않은 변이체들 중 5%를 교체한다. 필요한 것보다 25%보다 많은 최고 점수의 병원성 변이체를 유지하였고, 이에 따라 점수 있는 병원성 변이체의 8개의 상이한 세트를 샘플링하여 8가지 모델의 무작위성을 증가시키는 알려지지 않은 변이체들을 교체한다는 점에 주목한다. 이어서, 새로운 병원성 트레이닝 세트가

형성되고, 새로운 트레이닝 사이클이 실행된다. 이 프로세스는, 랜덤하게 샘플링된 초기의 알려지지 않은 변이체가 앙상블 모델에 의해 예측된 매우 확실한 병원성 변이체에 의해 모두 교체될 때까지 반복된다. 도 42는 반복적 균형맞춤 샘플링 프로세스를 예시한다.

[0526] **양성 트레이닝 세트 및 알려지지 않은 트레이닝 세트의 균형맞춤**

[0527] 양성 변이체와 일치하는 알려지지 않은 변이체의 샘플링 스킴은 본 발명의 모델 트레이닝의 편향을 줄이는 데 유용하다. 알려지지 않은 변이체가 랜덤하게 샘플링될 때, 심층 학습 모델은 종종 편향된 정보를 추출하고 사소한 솔루션을 제시한다. 예를 들어, 아미노산 치환 K->M이 양성 변이체보다 알려지지 않은 변이체에서 더 자주 발생하는 경우, 심층 학습 모델은 항상 K->M 치환을 병원성으로 분류하는 경향이 있다. 따라서, 두 가지 트레이닝 세트 간의 아미노산 치환 분포의 균형을 맞추는 것이 중요하다.

[0528] CpG 천이와 같은 더욱 높은 돌연변이성 클래스는 양성 공통 변이체에서 큰 표현 편향을 갖는다. 다른 영장류로부터의 이종상동성 변이체도 인간의 돌연변이율을 따르므로, 전체 양성 트레이닝 세트에서 높은 돌연변이성 클래스의 농축을 암시한다. 알려지지 않은 변이체의 샘플링 절차가 잘 제어되지 않고 균형이 맞지 않으면, 심층 학습 모델은, 변환이나 넌-CpG 변환과 같이 덜 표현된 클래스와 비교하여 CpG 변환을 양성으로 분류하는 경향이 있다.

[0529] 심층 학습 러닝 모델이 사소한 비생물학적 솔루션으로 수렴되는 것을 방지하기 위해, 양성 변이체 및 알려지지 않은 변이체의 트라이뉴클레오타이드 컨텍스트의 균형을 맞추는 것을 고려한다. 트라이뉴클레오타이드는 변이체 전의 염기, 변이체의 참조 염기, 및 변이체 후의 염기에 의해 형성된다. 변이체의 참조 염기는 다른 3개의 뉴클레오타이드로 변경될 수 있다. 전체적으로, 64×3개의 트라이뉴클레오타이드 컨텍스트가 있다.

[0530] **반복적 균형맞춤 샘플링**

[0531] **사이클 1**

[0532] 각각의 트라이뉴클레오타이드 컨텍스트에 대해 정확한 수의 양성 변이체를 일치시키도록 알려지지 않은 변이체들을 샘플링하였다. 다시 말하면, 제1 사이클에서는, 변이체의 트라이뉴클레오타이드 컨텍스트와 관련하여 양성 트레이닝 세트 및 병원성 트레이닝 세트를 반영하였다. 이러한 샘플링 방법론의 직관은, 양성 세트와 알려지지 않은 세트 간의 동일한 돌연변이율을 가진 변이체들이 동일하게 표현된다는 것이다. 이것은 모델이 돌연변이율에 기초하여 사소한 솔루션으로 수렴되는 것을 방지한다.

[0533] **사이클 2 내지 사이클 20**

[0534] 사이클 2에서는, 사이클 1로부터 트레이닝된 모델을 적용하여 사이클 1에 관여하지 않는 알려지지 않은 변이체들의 세트를 채점하고, 알려지지 않은 변이체들의 5%를 최상 예측된 병원성 변이체로 대체하였다. 이 세트는 모델에 의해 순수하게 생성되며, 이 세트에서 트라이뉴클레오타드 컨텍스트에 대한 균형맞춤을 적용하지 않았다. 트레이닝에 필요한 알려지지 않은 변이체의 나머지 95%는 양성 변이체에서 각 트라이뉴클레오타이드 컨텍스트의 계수치의 95%로 샘플링된다.

[0535] 직관은, 사이클 1이 완전히 일치된 트레이닝 세트를 사용하므로, 최상 예측 병원성 변이체가 어떠한 돌연변이율 편향 없이 생성된다는 것이다. 따라서, 이 세트에서 편향을 고려할 필요가 없다. 데이터의 나머지 95%는, 모델이 사소한 솔루션으로 수렴되는 것을 방지하도록 트라이뉴클레오타이드 컨텍스트 돌연변이율에 대해 여전히 제어된다.

[0536] 각각의 사이클에 대해, 교체된 알려지지 않은 변이체의 백분율은 5% 증가한다. 사이클 3의 경우, 알려지지 않은 변이체의 5%를 사이클 3의 모델로부터의 최상 예측된 병원성 변이체로 대체하였다. 누적식으로, 병원성 변이체의 분율은 10%로 증가하고, 트라이뉴클레오타이드-컨텍스트-반영된 알려지지 않은 변이체의 분율은 90%로 감소된다. 샘플링 프로세스는 나머지 사이클에 대하여 유사하다.

[0537] **사이클 21**

[0538] 최종 사이클인 사이클 21의 경우, 전체 병원성 트레이닝 세트는 순전히 심층 학습 모델로부터 예측된 최상 병원성 변이체를 포함한다. 모든 사이클에서 돌연변이율 편향을 명시적으로 제어하였으므로, 병원성 변이체는 트레이닝 데이터로서 사용하기에 신뢰할 만하며 돌연변이율 편향의 영향을 받지 않는다. 따라서, 트레이닝 절차의 최종 사이클은 병원성 예측을 위한 최종 심층 학습 모델을 생성한다.

[0539] 표지된 **양성 트레이닝 세트와 미표지 트레이닝 세트의 일치**

[0540] 미표지 변이체들의 균형맞춤 샘플링은 변이체의 유해성과 관련이 없는 편향을 제거하는 데 중요하다. 혼란스러운 효과에 대한 적절한 통제가 없다면, 심층 학습은 부주의하게 도입된 편향을 쉽게 찾아 내 클래스들을 구별할 수 있다. 인간 공통 변이체는 CpG 섬의 변이체와 같이 고도로 돌연변이 가능한 클래스의 변이체로 농축되는 경향이 있다. 마찬가지로, 영장류 다형성은, 또한, 인간 돌연변이율을 따르며, 이는 전체 양성 트레이닝 세트에서 고도로 돌연변이 가능한 변이체의 농축을 암시한다. 미표지 변이체의 샘플링 절차가 잘 통제되지 않고 균형이 맞지 않으면, 심층 학습망이 돌연변이율 편향에 의존하여 변이체를 분류하는 경향이 있으므로, 전환 또는 넌-CpG 천이 등의 덜 표현되는 클래스에 비해 CpG 천이를 양성으로 분류할 가능성이 높다. (이전에 논의된) 96개의 트라이뉴클레오타이드 컨텍스트 각각에서 표지된 양성 변이체와 정확히 동일한 수의 미표지 변이체를 샘플링하였다.

[0541] 표지된 양성 데이터 세트에서 영장류 변이체에 대해 표지되지 않은 변이체를 일치시킬 때, 해당 위치에 있는 해당 영장류 종의 변이체를 콜할 수 없었으므로, 미표지 변이체가 그 영장류 종이 다중 서열 정렬과 정렬되지 않은 인간 게놈 영역으로부터 선택되는 것을 허용하지 않았다.

[0542] 96개의 트라이뉴클레오타이드 컨텍스트 각각 내에서, 영장류 변이체에 대한 서열분석 커버리지를 보정하였다. 서열분석된 다수의 인간으로 인해, 인간 모집단의 공통 변이체는, 서열분석 커버리지가 낮은 영역에서도 잘 확인될 정도로 충분히 자주 관찰된다. 이는, 소수의 개체만이 서열분석되었기 때문에, 영장류 변이체에 대해서는 적용되지 않는다. ExAC/gnomAD 엑솜의 서열분석 커버지리에 기초하여 게놈을 10개의 빈으로 나눈다. 각 빈에 대해, 표지된 양성 데이터세트 대 미표지 데이터세트의 영장류 변이체의 분율을 측정하였다. 영장류 변이체가 선형 회귀를 사용하여 서열분석 커버리지에만 기초하여 표지된 양성 데이터세트의 구성원일 확률을 계산하였다(도 24). 표지된 양성 데이터세트에서 영장류 변이체와 일치하도록 미표지 변이체를 선택할 때, 회귀 계수를 사용하여 해당 위치에서 서열분석 커버지에 기초하여 변이체를 샘플링할 확률을 가중하였다.

[0543] **양성 변이체와 알려지지 않은 변이체의 생성**

[0544] **인간 모집단의 공통 변이체**

[0545] 최근 연구에 따르면, 인간 모집단의 공통 변이체는 일반적으로 양성이다. 일 구현예에 따르면, gnomAD는 표준 코딩 영역 내에서 소수 대립유전자 빈도(minor allele frequency: MAF)≥0.1%를 갖는 90,958개의 비동의 SNP를 제공한다. 필터를 통과한 변이체들은 유지된다. 인델은 제외된다. 시작 또는 정지 코돈에서 발생하는 변이체, 및 단백질-절단 변이체는 제거된다. 일 구현예에 따르면, 하위 집단을 면밀히 조사하면, 각 하위 집단 내에서 MAF≥0.1%인 미스센스 변이체의 총수는 245,360으로 증가한다. 이러한 변이체는 양성 변이체들의 트레이닝 세트의 일부를 형성한다.

[0546] **유인원의 공통 다형성**

[0547] 코딩 영역은 매우 보존적인 것으로 알려져 있으므로, 다형성이 유인원 모집단에서 고 빈도로 분리되는지를 가정하는 것이 간단하며, 이는 또한 인간 적합성에 약간의 영향을 미칠 수 있다. 유인원 게놈 프로젝트와 기타 연구로부터의 보노보, 침팬지, 고릴라 및 오랑우탄의 다형성 데이터는 dbSNP로부터의 마모셋 및 레서스의 SNP와 병합되었다.

[0548] **알려지지 않은 변이체의 생성**

[0549] 모든 가능한 변이체는, 해당 위치에서 뉴클레오타이드를 다른 3개의 뉴클레오타이드로 치환함으로써 정준 코딩 영역의 각각의 염기 위치로부터 생성된다. 새로운 코돈이 형성되어, 해당 위치에서 아미노산의 잠재적 변화를 초래한다. 동의 변경은 필터링된다.

[0550] gnomAD 데이터세트에서 관찰된 변이체는 제거된다. 시작 또는 정지 코돈에서 발생하는 변이체와 정지 코돈을 형성하는 변이체는 제거된다. 다수의 유전자 주석을 갖는 SNP의 경우, 정준 유전자 주석이 SNP의 주석을 나타내도록 선택된다. 일 구현예에 따르면, 전체적으로, 68,258,623개의 알려지지 않은 변이체가 생성된다.

[0551] **변이체의 추가 필터링**

[0552] 인간 게놈의 일부 영역에서는, 리드를 정렬하는 것이 어려운 것으로 알려져 있다. 그러한 영역을 포함하면, 트레이닝 및 테스트 데이터세트에 혼동 효과를 야기할 수 있다. 예를 들어, 선택적 압력이 높은 영역은 제한된 수의 다형성을 갖는 경향이 있다. 반면, 서열분석하기 어려운 영역도 다형성이 적다. 본 발명의 모델에 대한 이러한 혼란스러운 입력을 피하기 위해, gnomAD에 의해 서열분석되지 않은 유전자로부터 변이체를 제거하였다.

[0553] 일반적으로 양성 변이체는 다수의 종에 걸쳐 보존적인 경향이 있는 잘 서열분석된 영역에서 발견된다. 알려지지 않은 변이체는 게놈에 걸쳐 랜덤하게 샘플링되는데, 이것에는 불량하게 커버된 일부 영역이 포함된다. 이로 인해 양성 세트와 알려지지 않은 세트 간에 확인 편향이 발생한다. 편향을 줄이기 위해 gnomAD에서 리드 깊이 10 미만인 변이체를 필터링하였다. 또한, 모든 포유류 종에 걸쳐 플랭킹 서열 정렬에서 10% 초과의 결손 데이터를 갖는 모든 변이체를 필터링하였다.

[0554] **유효성확인 및 테스트를 위한 데이터**

[0555] 일 구현예에 따르면, 병원성 모델을 유효성확인하고 테스트하기 위해, 유효성확인 및 테스트를 위한 양성 변이체의 큰 풀로부터 10,000개의 양성 변이체의 두 세트를 각각 랜덤하게 샘플링하였다. 나머지 양성 변이체는 심층 학습 모델을 트레이닝하는 데 사용된다. 이들 변이체는, 인간 공통 변이체에 대해 일부 방법이 트레이닝됨에 따라 방법들 간의 공정한 비교를 보장하도록 이종상동성 영장류 변이체로부터 구체적으로 샘플링된다. 또한, 일 구현예에 따라 유효성확인 및 테스트를 위해 10,000개의 알려지지 않은 변이체들의 두 세트를 랜덤하게 샘플링하였다. 192개의 트라이뉴클레오타이드 컨텍스트 각각 내에서 알려지지 않은 변이체의 수가 유효성확인 및 테스트 세트에 대한 양성 변이체의 수와 일치함을 보장한다.

[0556] 자폐증 또는 발달 장애 질환(DDD) 및 이들의 영향을 받지 않는 형제와 영향을 받는 어린이의 드 노보 변이체를 사용하여 임상 설정에서 여러 방법의 성능을 평가하였다. 전체적으로, 일 구현예에 따르면, DDD 사례에서는 3821개의 미스센스 드 노보 변이체가 있고 자폐증이 있는 사례에서는 2736개의 미스센스 드 노보 변이체가 있다. 일 구현예에 따르면, 영향을 받지 않는 형제에 대하여 1231개의 미스센스 드 노보 변이체가 있다.

[0557] **심층 학습망 아키텍처**

[0558] 병원성 예측망은, 이차 구조 및 용매 접근성 망을 통해 5개의 직접 입력 및 2개의 간접 입력을 수신한다. 5개의 직접 입력은, 51-길이 아미노산 서열×20-깊이((20 개의 상이한 아미노산을 인코딩함)이며, 변이체가 없는 참조 인간 아미노산 서열(1a), 변이체가 치환된 대체 인간 아미노산 서열(1b), 영장류 종의 다중 서열 정렬로부터의 PFM(1c), 포유류 종의 다중 서열 정렬로부터의 PFM(1d), 및 더 먼 척추동물 종의 다중 서열 정렬로부터의 PFM(1e)을 포함한다. 이차 구조 및 용매 접근성 망들 각각은, 다중 서열 정렬 (1f) 및 (1g)로부터 PFM을 입력으로서 수신하고, 이들의 출력을 주 병원성 예측 망에 입력으로서 제공한다. 이차 구조 및 용매 접근성 망들은, 단백질 데이터뱅크에 대해 알려진 단백질 결정 구조에 대해 미리 트레이닝되었고, 병원성 모델 트레이닝 동안 역전파를 허용한다.

[0559] 5개의 직접 입력 채널은, 선형 활성화를 갖는 40개의 커널의 업샘플링 컨볼루션층을 통과한다. 인간 참조 아미노산 서열(1a)은, 영장류, 포유류, 및 척추동물 다중 서열 정렬로부터의 PFM과 병합된다(병합 1a). 유사하게, 인간 대체 아미노산 서열(1b)은, 영장류, 포유류, 및 척추동물 다중 서열 정렬로부터의 PFM과 병합된다(병합 1b). 이는 두 개의 병렬 트랙을 생성하며, 하나는 참조 서열이고 다른 하나는 변이체가 교체된 대체 서열용이다.

[0560] 참조 채널과 대체 채널 모두의 병합된 피처 맵(병합 1a 및 병합 1b)은, 일련의 6개의 잔여 블록(층 2a 내지 7a, 병합 2a 및 층 2b 내지 7b, 병합 2b)을 통과한다. 잔여 블록(병합 2a 및 병합 2b)의 출력은, 서로 연쇄화되어, 참조 채널과 대체 채널로부터의 데이터를 완전히 혼합하는 크기(51,80)(병합 3a, 병합 3b)의 피처 맵을 형성한다. 다음에, 데이터는, 부문 2.1에 정의된 바와 같이 각각 2개의 컨볼루션층을 포함하는 일련의 6개의 잔여 블록을 통해, 또는 1D 컨볼루션(층 21, 층 37, 층 47)을 통과한 후 2개의 모든 잔여 블록의 출력을 연결하는 스킵 연결을 통해 망을 병렬로 통과하기 위한 2개의 경로를 갖는다(병합 3 내지 9, 층 21과 34를 제외한 층 9 내지 46). 마지막으로, 병합된 활성화(병합 10)는 다른 잔여 블록(층 48 내지 53, 병합 11)으로 공급된다. 병합 11로부터의 활성화는, 필터 크기 1 및 시그모이드 활성화(층 54)를 이용하여 1D 컨볼루션에 주어지고, 이어서 변이체 병원성에 대한 망의 예측을 나타내는 단일 값을 선택하는 글로벌 최대 풀링층을 통과한다. 모델의 개략도는 도 3 및 보충 표 16(도 4a, 4b 및 4c)에 도시되어 있다.

[0561] **모델 개요**

[0562] 변이체 병원성을 예측하기 위해 반감독 심층 컨볼루션 신경망(CNN) 모델을 개발하였다. 모델에 대한 입력 피처는, 단백질 서열 및 보존 프로파일 플랭킹 변이체, 및 특정 유전자 영역에서의 미스센스 변이체의 고갈을 포함한다. 또한, 본 발명자들은 심층 학습 모델에 의해 변이체가 이차 구조 및 용매 접근성에 야기한 변화를 예측하고 이를 본 발명의 병원성 예측 모델에 통합하였다. 모델을 트레이닝하기 위해, 인간 하위 집단의 공통 변이체로부터의 양성 변이체 및 영장류로부터의 이종상동성 변이체를 생성하였다. 그러나, 병원성 변이체에 대한 신뢰

할만한 출처가 여전히 부족하다. 처음에 양성 변이체 및 알려지지 않은 변이체로 모델을 트레이닝한 다음, 반감독 반복적 균형맞춤 샘플링(IBS) 알고리즘을 사용하여, 알려지지 않은 변이체들을 신뢰도가 높은 것으로 예측된 병원성 변이체들의 세트로 점차 대체하였다. 마지막으로, 본 발명의 모델이 인간의 발달 장애 질환을 유발하는 드 노보 변이체를 양성 변이체와 구별하는 데 있어서 기존의 방법을 능가함을 입증하였다.

[0563] **잔여 블록의 채택**

[0564] 도 17은 잔여 블록을 도시한다. 병원성 예측의 본 발명의 심층 학습 모델과 이차 구조 및 용매 접근성을 예측하기 위한 심층 학습 모델은 모두 처음에는 예시된 잔여 블록의 정의를 채택한다. 잔여 블록의 구조는 이하의 도에 도시되어 있다. 입력층을 먼저 일괄 정규화한 후 비선형 활성화 "ReLU"가 이어진다. 이어서, 활성화는 1D 컨볼루션층을 통과한다. 1D 컨볼루션층으로부터의 이러한 중간 출력은, 다시 일괄 정규화되고 ReLU 활성화된 후, 다른 1D 컨볼루션층이 이어진다. 제2 1D 컨볼루션의 종료시, 해당 출력을 초기 입력과 병합한다. 이러한 아키텍처에서, 입력은 초기 상태로 유지되고, 잔여 연결에는 모델의 비선형 활성화가 없다.

[0565] 아트러스/팽창된 컨볼루션은 트레이닝 가능한 파라미터가 거의 없는 큰 수용장을 허용한다. 아트러스/팽창된 컨볼루션은, 입력값을 아트러스 컨볼루션율 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.

[0566] **본 발명의 모델의 신규성**

[0567] 본 발명의 방법은, 변이체의 병원성을 예측하기 위한 기존의 방법들과 3가지 방식으로 상이하다. 첫째, 본 발명의 방법은 반감독 심층 컨볼루션 신경망의 새로운 아키텍처를 채택한다. 둘째로, gnomAD 및 영장류 변이체로부터의 인간 공통 변이체로부터 신뢰할 수 있는 양성 변이체를 취득하는 한편, 동일한 인간에 의해 큐레이션된 변이체 데이터베이스를 사용하여 모델의 순환 트레이닝 및 테스트를 피하기 위해, 고도로 신뢰할 수 있는 병원성 트레이닝 세트가 반복적 균형맞춤 샘플링 및 트레이닝을 통해 생성된다. 셋째, 이차 구조 및 용매 접근성을 위한 심층 학습 모델이 본 발명의 병원성 모델의 아키텍처에 통합되었다. 구조 및 용매 모델로부터 취득되는 정보는 특정 아미노산 잔기에 대한 표지 예측으로 제한되지 않는다. 오히려, 판독 층이 구조 및 용매 모델로부터 제거되고, 미리 트레이닝된 모델이 병원성 모델과 병합된다. 병원성 모델을 트레이닝하는 동안, 미리 트레이닝된 구조 및 용매 층도 역전파되어 에러를 최소화한다. 이는 미리 트레이닝된 구조 및 용매 모델이 병원성 예측 문제에 초점을 맞추는 데 일조한다.

[0568] **이차 구조 및 용매 접근성 모델 트레이닝**

[0569] **데이터 준비**

[0570] 단백질의 3-상태 이차 구조 및 3-상태 용매 접근성을 예측하기 위해 심층 컨볼루션 신경망을 트레이닝하였다. PDB의 단백질 주석은 모델을 트레이닝하는 데 사용된다. 일 구현예에 따르면, 서열 프로파일과 25% 초과의 유사성을 갖는 서열은 제거된다. 일 구현예에 따르면, 전체적으로, 6,293개의 단백질 서열이 트레이닝에 사용되는데, 392개는 유효성확인에, 499개는 테스트에 사용된다.

[0571] 단백질에 대한 위치 특이적 점수매김 행렬(PSSM) 보존 프로파일은, UniRef90을 검색하기 위해 E-값 임계값인 0.001 및 3회 반복을 이용하여 PSI-BLAST를 실행함으로써 생성된다. 알려지지 않은 임의의 아미노산은 해당 이차 구조뿐만 아니라 공백으로서 설정된다. 또한, 모든 인간 유전자에 대해 유사한 파라미터 설정을 갖는 PSI-BLAST를 실행하여 PSSM 보존 프로파일을 수집한다. 이들 행렬은 구조 모델을 병원성 예측에 통합하는 데 사용된다. 이어서, 단백질 서열의 아미노산은 원-핫 인코딩 벡터로 전환된다. 단백질 서열 및 PSSM 행렬은, L×20의 행렬로 재구성되고, 여기서 L은 단백질의 길이이다. 이차 구조에 대한 3가지 예측된 라벨은 나선(H), 베타 시트(B), 및 코일(C)을 포함한다. 용매 접근성을 위한 3가지 라벨에는 매립(B), 개재(I) 및 노출(E)이 있다. 하나의 표지는 하나의 아미노산 잔기에 해당한다. 표지는 치수=3의 원-핫 인코딩 벡터로서 코딩된다.

[0572] **모델 아키텍처 및 트레이닝**

[0573] 단백질의 3-상태 이차 구조 및 3-상태 용매 접근성을 각각 예측하기 위해 2개의 엔드-투-엔드 심층 컨볼루션 신경망 모델을 트레이닝하였다. 두 모델은, 하나는 단백질 서열을 위한 입력 채널과 단백질 보존 프로파일을 위한 입력 채널인 두 개의 입력 채널을 포함하여 유사한 구성을 갖는다. 각각의 입력 채널은 L×20의 치수를 가지며,

여기서 L은 단백질의 길이를 나타낸다.

[0574] 각각의 입력 채널은 40개의 커널 및 선형 활성화를 갖는 1D 컨볼루션층(층 1a 및 1b)을 통과한다. 이 층은 입력 치수를 20에서 40으로 업샘플링하는 데 사용된다. 모델 전체에 걸쳐 다른 모든 층 40개의 커널을 사용한다. 2개의 층(1a 및 1b) 활성화는 40개 치수 각각에 걸쳐 값들을 합산함으로써 병합된다(즉, 병합 모드= "합"). 병합 노드의 출력은 1D 컨볼루션(층 2)의 단일 층을 통과한 다음 선형 활성화된다.

[0575] 층 2로부터의 활성화는 상기 정의된 바와 같이 일련의 9개의 잔여 블록(층 3 내지 11)을 통과한다. 층 3의 활성화는 층 4에 공급되고 층 4의 활성화는 층 5에 공급되는 방식으로 계속된다. 모든 세 번째 잔여 블록(층 5, 8 및 11)의 출력을 직접 합하는 스킵 연결도 있다. 이어서, 병합된 활성화는 ReLU 활성화와 함께 두 개의 1D 컨볼루션(층 12 및 13)에 공급된다. 층 13으로부터의 활성화는 소프트맥스 판독 층에 제공된다. 소프트맥스는 주어진 입력에 대한 3-클래스 출력의 확률을 연산한다.

[0576] 최상의 이차 구조 모델의 경우, 1D 컨볼루션은 1의 아트러스율을 갖는다. 용매 접근성 모델의 경우, 마지막 3개의 잔여 블록(층 9, 10 및 11)은 커널의 커버리지를 증가시키기 위해 2의 아트러스율을 갖는다. 단백질의 이차 구조는 근접한 아미노산의 상호작용에 크게 의존한다. 따라서, 커널 커버리지가 높은 모델은 성능을 약간 더 개선한다. 한편, 용매 접근성은 아미노산 간의 장거리 상호작용에 의해 영향을 받는다. 따라서, 아트러스 컨볼루션을 사용하는 커널의 커버리지가 높은 모델의 경우, 정확도가, 낮은 커버리지 모델의 정확도보다 2% 초과로 높다.

[0577] 이하의 표는, 일 구현예에 따라, 3-상태 이차 구조 예측 모델의 각 층에 대하여 활성화 및 파라미터에 관한 상세를 제공한다.

[0578]

층	유형	커널수, 윈도우 크기	형상	아트러스율	활성화
입력 서열(층 1a)	컨볼루션 1D	40,1	(L,40)	1	선형
입력 서열(층 1b)	컨볼루션 1D	40,1	(L,40)	1	선형
서열+PSSM 병합	병합(모드=연쇄화)		(L,80)		
층 2	컨볼루션 1D	40,5	(L,40)	1	선형
층 3	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 4	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 5	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 6	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 7	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 8	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 9	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 10	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 11	컨볼루션 1D	40,5	(L,40)	1	ReLU
활성화 병합	병합-층 5,8 및 11. 모드="합"		(L,40)		
층 12	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 13	컨볼루션 1D	40,5	(L,40)	1	ReLU
출력층	컨볼루션 1D	1,3	(L,3)		소프트맥스

[0579] 일 구현예에 따르면, 용매 접근성 모델의 상세는 이하의 표에 도시되어 있다.

층	유형	커널수, 윈도우 크기	형상	아트러스율	활성화
입력 서열(층 1a)	컨볼루션 1D	40,1	(L,40)	1	선형
입력 서열(층 1b)	컨볼루션 1D	40,1	(L,40)	1	선형
서열+PSSM 병합	병합(모드=연쇄화)		(L,80)		
층 2	컨볼루션 1D	40,5	(L,40)	1	선형
층 3	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 4	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 5	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 6	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 7	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 8	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 9	컨볼루션 1D	40,5	(L,40)	2	ReLU
층 10	컨볼루션 1D	40,5	(L,40)	2	ReLU
층 11	컨볼루션 1D	40,5	(L,40)	2	ReLU
활성화 병합	병합-층 5,8 및 11. 모드="합"		(L,40)		
층 12	컨볼루션 1D	40,5	(L,40)	1	ReLU
층 13	컨볼루션 1D	40,5	(L,40)	1	ReLU
출력층	컨볼루션 1D	1,3	(L,3)		소프트맥스

[0580]

[0581] 특정 아미노산 잔기의 이차 구조 클래스는 최대 예측 소프트맥스 확률에 의해 결정된다. 모델은, 역전파를 최적화하기 위한 ADAM 최적화기를 사용하여 전체 단백질 서열에 대하여 축적된 카테고리 교차 엔트로피 손실 함수로 트레이닝된다.

[0582] 3-상태 이차 구조 예측 모델에 대한 최상의 테스트 정확도는, 유사한 트레이닝 데이터세트에 대한 DeepCNF 모델에 의해 예측되는 최신 정확도와 유사하게 80.32%이다.

[0583] 3-상태 용매 접근성 예측 모델에 대한 최상의 테스트 정확도는, 유사한 트레이닝 데이터세트에 대한 RaptorX에 의해 예측되는 현재의 최신 정확도와 유사하게 64.83%이다.

[0584] 후술하는 바와 같이 미리 트레이닝된 3-상태 이차 구조 및 용매 접근성 예측 모델을 본 발명의 병원성 예측 모델에 통합하였다.

[0585] **변이체의 병원성을 예측하기 위한 트레이닝 모델**

[0586] **병원성 예측 모델에 대한 입력 피처**

[0587] 전술한 바와 같이, 병원성 예측 문제에 대하여, 병원성 모델을 트레이닝하기 위한 양성 변이체 트레이닝 세트 및 알려지지 않는 변이체 트레이닝 세트가 존재한다. 각 변이체에 대해, 이하의 입력 피처를 준비하여 모델에 공급하였다.

[0588] 각 변이체의 제1 입력 피처는, 플랭킹 아미노산 서열이며, 즉, 변이체의 서열 컨텍스트의 심층 학습 모델을 제공하도록 hg19의 참조 서열로부터 취득된 변이체의 각 측면 상의 25개 아미노산이다. 전체적으로, 이 플랭킹 참조 서열은 총 51개 아미노산의 길이를 갖는다.

[0589] 제2 피처는 변이체를 야기한 대체 아미노산이다. 참조-대체 아미노산 쌍을 직접 제공하는 대신, 대체 플랭킹 서열을 모델에 제공한다. 대체 플랭킹 서열은, 서열의 중간 위치가 참조 아미노산 대신에 대체 아미노산을 함유한다는 점을 제외하고는 제1 피처의 참조 플랭킹 서열과 동일하다.

[0590] 이어서, 양측 서열은 서열 길이 51×20의 원-핫 인코딩된 벡터로 전환되며, 여기서 각각의 아미노산은 20개의 0

또는 1의 벡터에 의해 표현된다.

[0591] 이어서, 변이체에 대한 99마리 척추동물의 다중 서열 정렬(MSA)로부터 3개의 위치 가중치 행렬(PWM)이 생성되고, 이러한 행렬의 하나는 12마리 영장류에 대한 것이고, 하나는 영장류를 제외한 47마리 포유류에 대한 것이고, 하나는 영장류와 포유류를 배제한 40마리 척추동물에 대한 것이다. 각 PWM의 치수는 L×20이며, 여기서 L은 변이체 주변의 플랭킹 서열의 길이이다(이 경우, L은 51개 아미노산을 나타냄). 이것은 종들의 각 카테고리에서 보이는 아미노산의 계수치를 포함한다.

[0592] 또한, psi 블라스트로부터 51개 아미노산의 변이체-플랭킹 서열에 대한 PSSM 행렬을 생성한다. 이것은 병원성 예측을 위한 3-상태 이차 구조와 용매 접근성 예측 모델의 통합에 사용된다.

[0593] 참조 서열(입력 1), 대체 서열(입력 2), 영장류(입력 3), 포유류(입력 4), 척추동물(입력 5), 및 3-상태 이차 구조와 용매 접근성 모델로부터의 정보로 병원성 모델을 트레이닝한다.

[0594] **심층 학습 모델 트레이닝**

[0595] 도 19는 심층 학습 모델 워크플로우의 개요를 제공하는 블록도이다. 병원성 트레이닝 모델은 5개의 직접 입력과 4개의 간접 입력을 포함한다. 5개의 직접 입력 피처는, 참조 서열(1a), 대체 서열(1b), 영장류 보존(1c), 포유류 보존(1d), 및 척추동물 보존(1e)을 포함한다. 간접 입력은, 참조 서열-기반 이차 구조(1f), 대체 서열-기반 이차 구조(1g), 참조 서열-기반 용매 접근성(1h), 및 대체 서열-기반 용매 접근성(1i)을 포함한다.

[0596] 간접 입력 1f 및 1g의 경우, 소프트맥스 층을 제외한 이차 구조 예측 모델의 미리 트레이닝된 층을 로딩한다. 입력 1f의 경우, 미리 트레이닝된 층은, 변이체에 대한 PSI-BLAST에 의해 생성된 PSSM과 함께 변이체에 대한 인간 참조 서열을 기초로 한다. 마찬가지로, 입력 1g에 대해, 이차 구조 예측 모델의 미리 트레이닝된 층들은 PSSM 행렬과 함께 입력으로서 인간 대체 서열에 기초한다. 입력 1h 및 1i는, 각각 변이체의 참조 서열 및 대안 서열에 대한 용매 접근성 정보를 포함하는 유사한 미리 트레이닝된 채널들에 해당한다.

[0597] 5개의 직접 입력 채널은 선형 활성화를 갖는 40개 커널의 업샘플링 컨볼루션층을 통과한다. 층(1a, 1c, 1d 및 1e)은 40개의 피처 치수에 걸쳐 합산된 값과 병합되어 층(2a)을 생성한다. 다시 말해, 참조 서열의 피처 맵은 세 가지 유형의 보존 피처 맵과 병합된다. 유사하게, 1b, 1c, 1d 및 1e는 40개의 피처 치수에 걸쳐 합산된 값과 병합되어 층 2b를 생성하며, 즉, 대체 서열의 피처는 3가지 유형의 보존 피처와 병합된다.

[0598] 층(2a 및 2b)은, ReLU의 활성화로 일괄 정규화되고, 각각 필터 크기 40(3a 및 3b)의 1D 컨볼루션층을 통과한다. 층(3a 및 3b)의 출력은 서로 연쇄화된 피처 맵들을 이용하여 1f, 1g, 1h, 1i와 병합된다. 다시 말해서, 보존 프로파일을 갖는 참조 서열의 피처 맵 및 보존 프로파일을 갖는 대체 서열은, 참조 서열 및 대체 서열의 이차 구조 피처 맵 및 참조 및 대체 서열의 용매 접근성 피처 맵(층 4)과 병합된다.

[0599] 층(4)의 출력은 6개의 잔여 블록(층 5, 6, 7, 8, 9, 10)을 통과한다. 마지막 3개의 잔여 블록은, 1D 컨볼루션에 대해 2의 아트러스율을 가지므로, 커널에 더 높은 커버리지를 제공한다. 층(10)의 출력은 필터 크기 1과 활성화 시그모이드(층 11)의 1D 컨볼루션을 통과한다. 층(11)의 출력은 변이체에 대한 단일 값을 선택하는 글로벌 최대 풀을 통과한다. 이 값은 변이체의 병원성을 나타낸다. 병원성 예측 모델의 일 구현예에 대한 세부 사항은 아래 표에 나타낸다.

층	유형	커널 수, 윈도우 크기	형상	아트러스율	활성화
참조 서열(1a)	보존 1D	40,1	(51,40)	1	선형
대체 서열(1b)	보존 1D	40,1	(51,40)	1	선형
영장류 보존(1c)	보존 1D	40,1	(51,40)	1	선형
포유류 보존(1d)	보존 1D	40,1	(51,40)	1	선형
척추동물보존(1e)	보존 1D	40,1	(51,40)	1	선형
참조 서열계 이차 구조(1f)	입력층		(51,40)		
대체 서열계 이차 구조(1g)	입력층		(51,40)		
참조 서열계 용매 접근성(1h)	입력층		(51,40)		
대체 서열계 용매 접근성(1i)	입력층		(51,40)		
참조 서열 병합(2a)	병합(모드=합)(1a,1c,1d,1e)		(51,40)		
대체 서열 병합(2b)	병합(모드=합)(1b,1c,1d,1e)		(51,40)		
3a	보존 1D(2a)	40,5	(51,40)	1	ReLU
3b	보존 1D(2b)	40,5	(51,40)	1	ReLU
4	병합(모드=연쇄화)(3a,3b,1f,1g,1h,1i)		(51,240)		
5	보존 1D	40,5	(51,40)	1	ReLU
6	보존 1D	40,5	(51,40)	1	ReLU
7	보존 1D	40,5	(51,40)	1	ReLU
8	보존 1D	40,5	(51,40)	1	ReLU
9	보존 1D	40,5	(51,40)	2	ReLU
10	보존 1D	40,5	(51,40)	2	ReLU
11	보존 1D	40,1	(51,1)	2	시그모이드
12	글로벌 최대 풀링		1		
출력층	활성화 층		1		시그모이드

[0600]

[0601] **앙상블**

[0602] 일 구현예에서, 본 발명의 방법의 각 사이클에 대해, 동일한 양성 데이터세트와 8개의 서로 다른 알려지지 않은 데이터세트를 학습하는 8개의 서로 다른 모델을 실행하였고, 8개의 모델에 걸쳐 평가 데이터세트의 평균을 평균화하였다. 알려지지 않은 변이체의 랜덤하게 샘플링된 다수의 세트를 모델에 제시하는 경우 샘플링 편차를 줄일 수 있고 잘 제어할 수 있다.

[0603] 또한, 앙상블들의 앙상블 접근법의 채택은 본 발명의 평가 데이터세트에 대한 본 발명의 모델의 성능을 개선할 수 있다. CADD는, 10개 모델의 앙상블을 사용하고, 10개 모델 모두에 걸쳐 평균 점수를 얻어 변이체를 채점한다. 여기서, 유사한 앙상블 접근법을 사용하려 하였다. 하나의 앙상블을 사용하여 결과를 벤치마킹한 다음 앙상블의 수를 증가시켜 성능을 다시 평가하였다. 각 앙상블에는 동일한 양성 데이터세트와 8개의 서로 다른 알려지지 않은 데이터세트를 학습하는 8개의 모델이 있다는 점에 주목한다. 상이한 앙상블의 경우, 난수 생성기의 시드 값이 다르므로 난수 변이체 세트가 서로 다르게 도출된다.

[0604] 일 구현예에 따른 상세한 결과가 아래 표에 도시되어 있다.

앙상블 수	DDD 데이터세트(앙상블의 평균의 평균)에 대한 로그(p값)	DDD 데이터세트(앙상블의 평균의 중앙)에 대한 로그(p값)
1	3.4e-27	3.4e-27
5	2.74e-29	3.8e-29
10	2.98e-29	2.55e-29
15	4.06e-29	3.88e-29
20	3.116e-29	3.05e-29
25	3.77e-29	3.31e-29
30	3.81e-29	3.34e-29

[0605]

[0606] 하나의 앙상블과 비교하여, 5개 앙상블 및 10개 앙상블은, DDD 데이터세트를 사용하여 평가할 때 더욱 중요한 p-값을 생성하였다. 그러나, 앙상블 수를 증가시키면, 성능이 더 개선되지 않아서, 앙상블에 대한 포화를 나타낸다. 앙상블은 알려지지 않은 다양한 변이체로 샘플링 편향을 감소시킨다. 그러나, 양성 클래스와 병원성 클래스 간의 192개의 트라이뉴클레오타이드 컨텍스트를 일치시킬 필요가 있었으며, 이는 본 발명의 샘플링 공간을 실질적으로 제한하여 빠른 포화로 이어졌다. 앙상블 접근법이 모델 성능을 크게 개선하고 모델에 대한 이해를 더욱 풍부하게 한다고 결론내렸다.

[0607] **병원성 모델을 트레이닝하기 위한 초기 중단**

[0608] 신뢰할 수 있는 주석이 달린 병원성 변이체 샘플이 없기 때문에, 모델 트레이닝에 대한 정지 기준을 정의하기가 어렵다. 모델 평가에 있어서 병원성 변이체 사용을 피하기 위해, 일 구현예에서는, 이종상동성 영장류의 10,000개의 양성 유효성확인 변이체와 10,000개의 트라이뉴클레오타이드-컨텍스트 일치된 알려지지 않은 변이체를 사용하였다. 모델의 각 에포크(epoch)를 트레이닝한 후, 양성 유효성확인 변이체와 알려지지 않은 유효성확인 변이체를 평가하였다. 윌콕슨 순위-합 테스트를 사용하여 양측 유효성확인 변이체 세트의 확률 분포의 차를 평가하였다.

[0609] 테스트의 p-값은, 양성 변이체를 알려지지 않은 변이체들의 세트와 구별하는 능력이 개선됨에 따라 더욱 중요해진다. 모델 트레이닝의 임의의 5번의 연속된 기간 동안 두 분포를 구별하는 모델의 능력에 있어서 개선이 관찰되지 않으면 트레이닝을 중단한다.

[0610] 이전에는, 트레이닝으로부터 10,000개의 보류된 영장류 변이체의 두 개의 개별 세트를 따로 분리하였으며, 유효성확인 세트 및 테스트 세트라고 칭하였다. 모델 트레이닝 동안 조기 중단을 평가하기 위해 트라이뉴클레오타이드와 일치된 10,000개의 보류된 영장류 변이체와 10,000개의 미표지 변이체의 유효성확인 세트를 사용하였다. 각 트레이닝 에포크 후에, 윌콕슨 순위-합 테스트를 사용하여 예측 점수의 분포의 차를 측정하여, 표지된 양성 유효성확인 세트의 변이체와 미표지 일치하는 대조군을 구별하는 심층 신경망의 능력을 평가하였다. 과적합성을 방지하기 위해 5번의 연속적인 트레이닝 에포크 후에 더 이상의 개선이 관찰되지 않으면 트레이닝을 중단하였다.

[0611] **분류자 성능의 벤치마킹**

[0612] 심층 학습망의 두 개 버전의 분류 정확도를 평가하였으며, 하나는 공통 인가 변이체만으로 트레이닝된 것이고, 하나는 공통 인간 변이체와 영장류 변이체 모두를 포함하는 전체 양성 표지된 데이터세트로 트레이닝된 것이며, 이하의 분류자들도 포함하였다: SIFT, Polyphen-2, CADD, REVEL, M-CAP, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaSVM, MetaLR, MutPred, DANN, FATHMM-MKL_coding, Eigen, GenoCanyon, and GERP++ 13,32-48. 다른 분류자들의 각각에 대한 점수를 취득하도록, dbNSFP 49로부터 모든 미스센스 변이체에 대한 점수를 다운로드하였고(https://sites.google.com/site/jpopgen/dbNSFP), 10,000개의 보류된 영장류 변이체 테스트 세트, 및 DDD 사례 대 대조군의 드 노보 변이체에 대한 방법을 벤치마킹하였다. SIFT, Polyphen-2 및 CADD를 주요 논문에 포함시킬 것으로서 선택하였으며, 그 이유는 이들이 가장 널리 사용되는 방법들이기 때문이며, REVEL이, 상이한 평가 모드들에 걸쳐, 평가한 20개의 기존의 분류자 중 최고 중 하나로서 두각을 나타내기 때문이다. 평가한 모든 분류자의 성능은 도 28a에 제공되어 있다.

[0613] 심층 학습망의 성능에 대한 가용 트레이닝 데이터 크기의 영향을 평가하기 위해, 385,236종의 영장류 및 공통 인간 변이체의 표지된 양성 트레이닝 세트로부터 랜덤하게 샘플링함으로써 도 6의 각 데이터 포인트에서 심층

학습망을 트레이닝하였다. 분류자의 성능에 있어서 랜덤 노이즈를 줄이기 위해, 초기 파라미터 가중치의 무작위 인스턴스화를 매번 사용하여 이러한 트레이닝 절차를 5회 수행했으며, 도 6에서 10,000개의 보류된 영장류 변이체 및 DDD 사례 대 대조군 데이터세트의 중앙값 성능을 나타내었다. 우연하게도, 385,236으로 표지된 양성 변이체의 전체 데이터세트를 갖는 중앙 분류자의 성능은, DDD 데이터세트에 대한 논문 나머지에서 사용했던 성능보다 약간 우수했다(윌콕슨 순위-합 테스트에 의한 $P<10^{-28}$ 대신 $P<10^{-29}$). 각 개별 영장류 종으로부터의 변이체가 분류 정확도에 기여하는 반면 각 개별 포유류 종의 변이체들은 분류 정확도를 낮추는 것을 나타내기 위해, 일 구현예에 따르면, 83,546개의 인간 변이체와 각 종에 대해 랜덤하게 선택된 변이체를 포함하는 트레이닝 데이터세트를 사용하여 심층 학습망을 트레이닝하였다. 트레이닝 세트에 추가한 변이체들의 상수(23,380)는, 미스센스 변이체가 가장 적은 종에서 이용 가능한 변이체, 즉, 보노보의 수이다. 노이즈를 줄이기 위해, 트레이닝 절차를 다시 5번 반복하였고, 분류자의 중간 성능을 보고했다.

[0614] **모델 평가**

[0615] 일 구현예에서, 반복적 균형맞춤 샘플링 절차에 따라 21 사이클의 심층 학습망 모델을 트레이닝하였다. 본 발명자들은 본 발명의 분류자의 성능을 평가하기 위해 두 가지 유형의 평가를 수행하였다. 또한, 본 발명자들은 두 개 메트릭으로 본 발명의 모델을 Polyphen2, SIFT, 및 CADD와 비교하고 임상 주석에 대한 본 발명의 모델의 적용 가능성을 평가하였다.

[0616] **방법 1: 양성 테스트 세트 정확도**

[0617] 일 구현예에서, 8개의 상이한 트레이닝된 모델의 앙상블을 사용하여 해당 예측된 확률을 연산함으로써 10,000개의 양성 변이체와 알려지지 않은 변이체를 평가하였다. 또한, 위에서 언급한 다른 기존 방법들에 의해 점수를 매긴 해당 예측 확률을 취득하였다.

[0618] 이어서, 본 발명자들은 평가에 사용된 각각의 방법에 대해 알려지지 않은 테스트 변이체에 걸쳐 예측된 확률의 중앙값을 취득하였다. 중앙값 점수를 사용하여, 각 방법에 의해 사용된 양성 변이체 및 병원성 변이체의 주석에 따라 중앙값 초과 또는 미만의 점수를 갖는 양성 변이체의 수를 발견하였다. SIFT, CADD, 및 본 발명의 방법은 병원성 변이체를 1로 양성 변이체를 0으로 표지하였다. 따라서, 중앙값 미만의 점수를 갖는 양성 변이체의 수를 계수하였다. Polyphen은 반대 주석을 사용하며 중앙값을 초과하는 양성 변이체의 수를 계수하였다. 중앙값 초과/미만의 점수를 갖는 양성 변이체의 수를 양성 변이체의 총수로 나눈 비율이, 양성 변이체의 예측 정확도를 나타낸다.

[0619] 양성 정확도 = 중앙값 초과의(미만의) 양성 변이체의 총수 / 양성 변이체의 총수

[0620] 이 평가 방법에 대한 본 발명자들의 추론은, gnomAD의 변이체의 선택적 압력 분석에 의존한다. gnomAD의 싱글톤의 경우, 미스센스 변이체 대 동의 변이체의 비율은 ~2.26:1이다. gnomAD의 공통 변이체(MAF>0.1%)의 경우, 미스센스 대 동의 비율은 ~1.06:1이다. 이것은, 랜덤하게 알려지지 않은 변이체들의 세트 중에서, 약 50%가 자연 선택에 의해 제거될 것으로 예상되고 나머지 50%는 경미한 경향이 있으며 모집단에서 공통일 가능성이 있음을 나타낸다.

[0621]

방법	보류된 양성 세트의 정확도
Polyphen	0.74
SIFT	0.69
CADD	0.77
본 발명의 모델	0.85

[0622] 위의 표에 나타낸 바와 같이, 본 발명의 방법은 두 번째로 우수한 방법 CADD보다 8%를 초과하여 뛰어나다. 이것은 양성 변이체를 분류하는 본 발명의 모델의 능력이 크게 개선되었음을 나타낸다. 이러한 입증은 본 발명의 모델의 능력을 입증하지만, 이하의 방법 2는 임상 해석을 위한 임상 데이터세트에 대한 본 발명의 모델의 유용성을 나타낸다.

[0623] **방법 2: 임상 데이터세트 평가**

[0624] 일 구현예에서는, 발달 장애 질환(DDD) 사례-대조군 데이터세트를 포함하는 임상 데이터세트에 대한 이들 병원성 예측 방법을 평가하였다. DDD 데이터세트는, 영향을 받은 어린이로부터의 3,821개의 드 노보 미스센스 변이체 및 영향을 받지 않은 형제로부터의 1,231개의 드 노보 미스센스 변이체를 포함한다. 본 발명의 가설은, 영향을 받은 아이들의 드 노보 변이체들이 영향을 받지 않은 형제들의 드 노보 변이체들보다 해로운 경향이 있다는 것이다.

[0625] 임상 테스트 데이터세트가 병원성 변이체들을 명확하게 표지하지 않으므로, 그러한 방법들의 성능을 측정하기 위해 (영향을 받은 및 영향을 받지 않은) 드 노보 변이체들의 두 개 세트 간의 사이의 분리를 사용하였다. 윌콕슨 순위-합 테스트를 적용하여 이러한 드 노보 변이체들의 두 개 세트가 얼마나 잘 분리되는지를 평가하였다.

[0626]

방법	DDD 데이터세트에 대한 -log10(p-값)
Polyphen	15.02
SIFT	13.52
CADD	13.83
DL	28.35

[0627] 위 표에 따르면, 본 발명의 반감독 심층 학습 모델은, 변이체의 영향을 받은 드 노보 세트와 영향을 받지 않은 드 노보 세트를 구분하는 데 훨씬 더 나은 성능을 제공한다. 이것은 본 발명의 모델이 기존의 방법들보다 임상 해석에 더 적합하다는 것을 나타낸다. 또한, 이는, 게놈 서열 및 보존 프로파일으로부터 피처를 추출하는 일반적인 접근법이 사람 큐레이션된 데이터세트를 기반으로 하는 수동으로 제작된 피처보다 우수함을 유효성확인한다.

[0628] **10,000개 영장류 변이체의 보류된 테스트 세트에 대한 양성 예측 정확도**

[0629] 테스트 데이터세트의 10,000개의 보류된 영장류 변이체를 사용하여 심층 학습망 및 다른 20개의 분류자를 벤치마킹하였다. 상이한 분류자는 광범위하게 다양한 점수 분포를 가졌기 때문에, 트라이뉴클레오타이드 컨텍스트에 의해 테스트 세트와 일치한 랜덤하게 선택된 10,000개의 미표지 변이체를 사용하여 각 분류자에 대한 50-백분위 임계값을 식별하였다. 방법들 간의 공정한 비교를 보장하기 위해, 해당 분류자의 50-백분위 임계값에서 양성으로 분류된 10,000개의 보류된 영장류 변이체 테스트 세트에서의 변이체들의 분율에 대해 각 분류자를 벤치마킹하였다.

[0630] 양성 변이체를 식별하기 위해 50-백분위를 사용하는 것에 대한 본 발명자들의 추론은, ExAC/gnomAD 데이터세트에서의 미스센스 변이체에 대해 관찰된 선택적 압력에 기초한다. 싱글톤 대립유전자 빈도로 발생하는 변이체의 경우, 미스센스:동의 비율은 ~2.2:1인 반면, 공통 변이체(>0.1% 대립유전자 빈도)의 경우, 미스센스:동의 비율은 ~1.06:1이다. 이것은, 대략 50%의 미스센스 변이체가 흔한 대립유전자 빈도로 자연 선택에 의해 제거될 것으로 예상되고, 나머지 50%는 유전 드리프트를 통해 모집단에서 공통일 가능성이 있을 정도로 순하다는 것을 나타낸다.

[0631] 각 분류자에 대해, 50-분위 임계값을 사용하여 양성으로 예측되는 보류된 영장류 테스트 변이체의 분율이 도시되어 있다(도 28a 및 보충 표 17(도 34)).

[0632] **DDD 연구로부터의 드 노보 변이체의 분석**

[0633] DDD 영향을 받은 개체에서의 드 노보 미스센스 변이체와 영향을 받지 않은 형제 대조군에서의 드 노보 미스센스 변이체를 구별하는 능력에 대한 분류 방법을 벤치마킹하였다. 각 분류자에 대해, 두 분포에 대한 예측 점수 간의 차의 윌콕슨 순위-합 테스트로부터의 p-값을 보고하였다(도 28b 및 28c 및 보충 표 17(도 34)).

[0634] 모델의 성능을 분석하기 위한 본 발명의 2개의 메트릭이 상이한 공급원 및 방법론으로부터 도출된다는 점을 고려하여, 2개의 상이한 메트릭에 대한 분류자의 성능이 상관되는지를 테스트하였다. 실제로, 이들 두 가지 메트릭은, 보류된 영장류 테스트세트에 대한 양성 분류 정확도와 DDD 사례 대 대조군에서의 드 노보 미스센스 변이체에 대한 윌콕슨 순위-합 p-값 사이에 스피어맨 $\rho=0.57(P<0.01)$과 상관 관계가 있음을 발견하였다. 이것은, 분류자를 벤치마킹하기 위해 DDD 사례 대 대조군 p-값과 보류된 영장류 테스트 설정 정확도 간에 양호한 일치가 있음을 나타낸다(도 30a).

[0635] 또한, 심층 학습망이 질환에 연관된 유전자의 발견을 도울 수 있는지를 테스트하였다. 드 노보 돌연변이의 관찰된 수를 널 돌연변이 모델 하에서 예상되는 수와 비교함으로써 유전자에서의 드 노보 돌연변이의 농축을 시험하였다.

[0636] 모든 미스센스 드 노보 돌연변이로부터의 결과 대 점수>0.803을 갖는 미스센스 돌연변이로부터의 결과를 비교하는 심층 학습망의 성능을 조사하였다. 모든 미스센스 드 노보의 테스트에서는 디폴트 미스센스율을 사용한 반면, 필터링된 미스센스 드 노보의 테스트에서는 >0.803의 점수를 갖는 부위로부터 계산된 미스센스 돌연변이율을 사용하였다. 각 유전자는 4개 테스트를 필요로 했으며, 하나는 단백질 절단 농축을 테스트하는 것이고, 하나는 단백질-변형 드 노보 돌연변이의 농축을 테스트하는 것이고, 이들 양측은 단지 DDD 코호트를 위한 그리고 신경발달 트리오 서열분석 코호트의 더욱 큰 메타 분석을 위한 테스트이었다. 단백질-변형 드 노보 돌연변이의 농축을 피셔의 방법에 의해 코딩 서열 내에서의 미스센스 드 노보 돌연변이의 클러스터링의 테스트와 조합하였다(보충 표 20 및 표 21). 각 유전자에 대한 p- 값은 4가지 테스트 중 최소값으로부터 취해졌으며, 게놈 전체의 유의성은 $P<6.757\times10^{-7}$($\alpha=0.05$, 4가지 테스트를 이용한 18,500개의 유전자)로 결정되었습니다.

[0637] **605개 DDD 질환-연관된 유전자 내의 수신자 오퍼레이터 특성 및 분류 정확도 연산**

[0638] 심층 학습망이 실제로 새로운 유전적 상속 방식을 갖는 유전자의 병원성을 선호하기보다는 동일한 유전자 내의 병원성 변이체 및 양성 변이체를 구별하는지를 테스트하기 위해, (드 노보 단백질-절단 변이 단독을 사용하여 계산된) DDD 코호트에서 p-값<0.05인 신경 발달 질환에 연관된 605개 유전자의 세트를 식별하였다(보충 표 18). DDD 및 대조군 데이터세트에서의 605개 유전자의 변이체의 확률 분포를 분리하는 능력에 있어서 모든 분류자의 윌콕슨 순위-합 p-값을 보고한다(도 28c 및 보충 표 19(도 35)).

[0639] 이러한 605개 유전자의 세트 내에서, 돌연변이율 단독에 의해 예상되는 것의 3배인 드 노보 미스센스 변이체에 대한 농축 비를 관찰한다. 이는, DDD 영향을 받은 환자의 드 노보 미스센스 변이체가 약 67%의 병원성 변이체 및 33% 백그라운드 변이체를 포함하는 반면, 건강한 대조군의 드 노보 미스센스 변이체는 불완전한 침투를 제외하고는 대부분 백그라운드 변이체로 구성되어 있음을 나타낸다.

[0640] 병원성 변이체 및 양성 변이체를 완벽하게 구별하는 분류자에 대한 가능한 최대 AUC를 계산하기 위해, 605개 유전자 내에서 영향을 받은 개체의 드 노보 미스센스 변이체의 67%만이 병원성이며, 나머지는 백그라운드임을 고려하였다. 수신자 오퍼레이터 특성 곡선을 구성하기 위해, 드 노보 DDD 변이체의 분류를 진양성 콜로서 병원성으로 취급하였고, 건강한 대조군에서의 드 노보 변이체의 분류를 위양성 콜로서 병원성으로 취급하였다. 따라서, 완벽한 분류자는 DDD 환자의 드 노보 변이체의 67%를 진양성으로 분류하고, DDD 환자의 드 노보 변이체의 33%를 위음성으로 분류하고, 대조군에서의 드 노보 변이체의 100%를 진음성으로 분류할 것이다. 수신자 오퍼레이터 곡선의 시각화는, 직선에 의해 플롯의 (0%, 0%) 및 (100%, 100%) 모서리에 연결된 67% 진양성률 및 0% 위양성률을 갖는 단일 포인트만을 표시하여, 양성 돌연변이 및 병원성 돌연변이의 완전한 구별을 갖는 분류자에 대해 최대 AUC 0.837을 산출한다(도 30b 및 보충 표 19(도 35)).

[0641] 결합된 DDD 및 건강한 대조군 데이터세트에서의 605개 유전자 내의 병원성 변이체의 예상 분율을 추정함으로써 이진 역치에서 병원성 변이체 및 양성 변이체를 분리하기 위한 심층 학습망의 분류 정확도를 계산하였다. DDD 데이터세트는 기대 이상의 과도한 249개의 드 노보 미스센스 변이체와 함께 379개의 드 노보 변이체를 포함하고, 대조군 데이터 세트는 65개의 드 노보 변이체를 포함하였으며, 총 444개의 변이체 중 249개의 병원성 변이체를 예상하였다(도 22a). 이 예상 비율에 따라 444개의 드 노보 미스센스 변이체를 양성 또는 병원성 카테고리로 분리한 각 분류자에 대한 임계값을 선택하였고, 이를 이진 컷오프로서 사용하여 각 분류자의 정확도를 평가하였다. 본 발명의 심층 학습 모델의 경우, 이 임계값은 0.803 이상의 컷오프에서 달성되었으며, 실제 진양성 비율은 65%, 위양성 비율은 14%였다. DDD 개체에서 ~33% 백그라운드 변이체의 존재에 대해 조정된 분류 정확도를 계산하기 위해, 백그라운드였던 드 노보 DDD 변이체의 33%가 건강한 대조군에서 관찰된 것과 동일한 위양성 비율로 분류될 것이라고 가정했다. 이는 DDD 데이터세트에서 진양성 분류 이벤트의 14%×0.33=4.6%에 해당하며 실제로 백그라운드 변이체로부터의 위양성이다. 심층 학습망의 조정된 진양성 비율은 (65%-4.6%)/67% = 90%로 추정된다. 심층 학습망에 대해 88%인, 진양성 비율과 진음성 비율의 평균을 보고한다(도 30c 및 보충 표 19(도 35)). 이 추정값은, 신경 발달 장애에 있어서 불완전한 침투의 높은 유병률로 인해 분류자의 실제 정확도를 과소평가할 가능성이 높다.

[0642] **ClinVar 분류 정확도**

[0643] 기존 분류자의 대부분은 ClinVar에 대해 트레이닝되고, ClinVar에서 직접 학습하지 않은 분류자라도, ClinVar에

서 학습된 분류자의 예측 점수를 사용하여 영향을 받을 수 있다. 또한, 대립유전자 빈도가 양성 결과를 변이체에 할당하기 위한 기준의 일부이기 때문에, 공통 인간 변이체는 양성 ClinVar 결과에 대해 매우 농축된다.

[0644] 다른 분류 방법들이 이전 년도들에 공개되었으므로, 2017년에 추가된 ClinVar 변이체만을 사용함으로써 분석에 적합해지도록 ClinVar 데이터세트의 원형도를 최소화하였다. 2017 ClinVar 변이체 중에서도, ExAC에 있어서 흔한 대립유전자 빈도(>0.1%)로 임의의 변이체, 또는 HGMD, LSDB 또는 Uniprot에 존재하는 임의의 변이체를 제외했다. 이러한 모든 변이체를 필터링하고 불확실한 유의미한 변이체 및 주석이 충돌하는 변이체를 배제한 후, ClinVar에는, 양성 주석이 있는 177개의 변이체와 병원성 주석이 있는 969개의 변이체가 남았다.

[0645] 심층 학습망과 기존 방법을 모두 사용하여 모든 ClinVar 변이체를 채점하였다. 이 데이터세트 내에서 양성 변이체 및 병원성 변이체의 관찰된 비율에 따라 ClinVar 변이체들을 양성 또는 병원성 카테고리로 분리한 각 분류자의 임계값을 선택하였으며, 이를 이진 컷오프로서 사용하여 각 분류자의 정확도를 평가하였다. 각 분류자에 대한 진양성 비율과 진음성 비율의 평균을 보고한다(도 31a 및 도 31b). ClinVar 데이터세트에 대한 분류자의 성능은, 10,000개의 보류된 영장류 변이체에 대한 분류 정확도 또는 DDD 사례 대 대조군 데이터세트에 대한 윌콕슨 순위-합 p-값에 대한 분류자들의 성능과 유의한 상관 관계가 없었다(도 31a 및 도 31b).

[0646] 기존 분류자들이 인간 전문가의 행동을 정확하게 모델링하지만, 경험적 데이터에서 병원성 돌연변이와 양성 돌연변이를 구별하기 위해 인간 휴리스틱이 완전히 최적이 아닐 수 있다고 가정한다. 이러한 일례가 그랜덤 점수이며, 이는 아미노산 치환의 유사성 또는 비유사성을 특성화하기 위한 거리 메트릭을 제공한다. 완전한 ClinVar 데이터세트 내에서 병원성 변이체 및 양성 변이체에 대한 평균 그랜덤 점수(~42,000개 변이체)를 연산하였고, 이를 605개 유전자 내의 영향을 받은 개체 및 영향을 받지 않은 개체의 드 노보 변이체에 대한 평균 그랜덤 점수와 비교하였다. DDD 영향을 받은 개체에 ~33%의 백그라운드 변이체의 존재를 보정하도록, DDD 사례와 대조군 간의 그랜덤 점수 차를 50%만큼 증가시켰으며, 이는 ClinVar의 병원성 변이체 및 양성 변이체 간의 차보다 여전히 작았다. 한 가지 가능성은, 인간 전문가가 아미노산 치환 거리와 같이 측정하기 쉬운 측정에 너무 많은 가중치를 부여하는 반면, 단백질 구조와 같은 인자에는 가중치를 덜 부여하여 인간 전문가가 정량화하는 것이 더욱 어렵다는 점이다.

[0647] **심층 학습 모델의 해석**

[0648] 기계 학습 알고리즘이 문제를 해결하는 방법을 이해하는 것은 종종 어려운 일이다. 변형 병원성을 예측하기 위해 추출하도록 학습한 피처를 이해하기 위해 심층 학습망의 초기 층들을 시각화하였다. 미리 트레이닝된 3-상태 이차 구조 예측 모델의 처음 3개 층(제1 컨볼루션층이 있고 이어서 두 개의 업샘플링층이 있음) 내의 상이한 아미노산에 대한 상관 계수를 계산했으며, 컨볼루션층의 가중치가 BLOSUM62 행렬 또는 그랜덤 거리와 매우 유사한 피처를 학습한다는 것을 나타내었다.

[0649] 상이한 아미노산 간의 상관 계수를 연산하기 위해, 이차 구조 모델에서 3개의 업샘플링층(층 1a, 1b, 및 1c)보다 선행하는 제1 컨볼루션층의 가중치로 시작하였다. 3개의 층 간에 행렬 곱셈을 수행하였으며, 차원(20, 5, 40)을 갖는 행렬을 생성하였고, 여기서 20은 아미노산의 수이고, 5는 컨볼루션층의 윈도우 크기이고, 40은 커널의 수이다. 마지막 두 개의 차원을 평평하게 함으로써 차원(20,200)을 갖도록 행렬을 재형상화하여, 20개의 아미노산 각각에 작용하는 가중치가 200-길이 벡터로 표현된 행렬을 취득하였다. 20개 아미노산 간의 상관 행렬을 계산하였다. 각 차원은 각 아미노산을 나타내므로, 상관 계수 행렬을 계산함으로써, 아미노산 간의 상관을 계산하고, 트레이닝 데이터로부터 학습한 것에 기초하여 심층 학습망에 얼마나 유사하게 보이는지를 계산한다. 상관 계수 행렬의 시각화는, 도 27에 도시되어 있고(아미노산은 BLOSUM62 행렬 순서에 의해 정렬됨), 소수성 아미노산(메티오닌, 이소류신, 류신, 발린, 페닐알라닌, 티로신, 트립토판) 및 친수성 아미노산(아스파라긴, 아스파르트산, 글루타민, 아르기닌, 및 리신)을 포함하는 두 개의 두드러진 클러스터를 도시한다. 이러한 초기 층의 출력은 이후 층들의 입력으로 되므로, 심층 학습망이 점점 더 복잡한 계층적 데이터 표현을 구성할 수 있게 한다.

[0650] 해당 예측의 신경망에 의해 사용되는 아미노산 서열의 윈도우를 예시하기 위해, 5000개의 랜덤하게 선택된 변이체 내와 주변의 각 위치를 교란하여 변이체에 대한 예측 PrimateAI 점수에 대한 영향을 관찰하였다(도 25b). 변이체 둘레에 있는(+25 내지 -25) 근처 아미노산 위치에서 입력들을 체계적으로 제로 아웃하였고, 신경망의 변이체의 예측 병원성의 변화를 측정하였고, 5000개 변이체에 걸쳐 변화의 평균 절대값을 도시하였다. 변이체 근처의 아미노산은, 대략 대칭 분포에서 변이체로부터의 거리가 증가함에 따라 점차적으로 테일링되는 가장 큰 효과를 갖는다. 중요하게도, 모델은, 단백질 모티프를 인식하는 데 필요하듯이, 변이체의 위치에서의 아미노산뿐만 아니라 윈도우 창으로부터의 정보를 이용하는 것에도 기초하여 예측을 행한다. 비교적 컴팩트한 크기의 단백질 서브도메인과 일관되게, 윈도우의 크기를 51개 초과의 아미노산으로 확장하는 것이 정확도를 추가로 개선하지

않았음을 경험적으로 관찰하였다.

[0651] 정렬에 대한 심층 학습 분류자의 민감도를 평가하기 위해, 변이체 분류의 정확도에 대한 정렬 깊이의 영향을 다음과 같이 조사하였다. 정렬에서 종의 수를 기준으로 데이터를 5개의 빈으로 나누고, 각 빈의 망 정확도를 평가하였다(도 57). (도 21d에서와 같이, 그러나 각각의 빈에 대해 개별적으로 수행된) 트라이뉴클레오타이드 컨텍스트에 대해 일치된 랜덤하게 선택된 돌연변이로부터 보류된 양성 돌연변이들의 세트를 분리할 때 망의 정확도가 최상위 3개의 빈에서 가장 강하고 최하위 2개의 빈에서 상당히 약하다는 것을 것을 발견하였다. 99마리 척추동물 다종 정렬은, 11마리의 비인간 영장류, 50마리의 포유류, 및 38마리의 척추 동물을 포함하며, 하단의 2개의 빈은 다른 비포유류 포유류로부터의 희박한 정렬 정보를 갖는 단백질을 나타낸다. 심층 학습망은, 정렬 정보가 영장류와 포유류 전체에 걸쳐 확장될 때 강력하고 정확하며, 더욱 먼 척추동물로부터의 보존 정보는 덜 중요하다.

[0652] **정준 코딩 영역의 정의**

[0653] 정준 코딩 영역을 정의하기 위해, DNA 서열(CDS) 영역을 코딩하기 위해 인간과 99마리 척추동물 게놈의 다중 정렬(knownCanonical.exonNuc.fa.gz)을 UCSC 게놈 브라우저로부터 다운로드하였다. 인간의 경우, 엑손 좌표는 빌드(Build) hg19에 있다. 엑손들은 병합되어 유전자를 형성한다. 오토솜 및 chrX 상의 유전자들은 유지된다. 비동종 유전자를 제거하였으며, 여기서, 동종 유전자들의 리스트는 NCBI ftp://ftp.ncbi.nih.gov/pub/HomoloGene/current/homologene.data로부터 다운로드된다. 유전자 주석이 여러 개인 SNP의 경우, SNP의 주석을 나타내기 위해 가장 긴 전사가 선택된다.

[0654] **인간, 유인원, 및 포유류 다형성 데이터**

[0655] 전세계 8개 하위 집단으로부터 123,136명 개체의 전체 엑솜 서열분석 데이터를 수집한, 최근 대규모 연구인 게놈 집계 데이터베이스(gnomAD)로부터, 인간 엑솜 다형성 데이터를 다운로드하였다. 이어서, 필터를 통과하고 정준 코딩 영역에 해당하는 변이체를 추출하였다.

[0656] 유인원 게놈 서열분석 프로젝트는, 24마리의 침팬지, 13마리의 보노보, 27마리의 고릴라, 및 (5마리의 수마트라 오랑우탄 및 5마리의 보르네오 오랑우탄을 포함하는) 10마리의 오랑우탄의 전체 게놈 서열분석 데이터를 제공한다. 침팬지와 보노보에 대한 연구는 추가 25마리 유인원의 WGS를 제공한다. 모든 서열분석 데이터가 hg19에 맵핑됨에 따라, 본 연구로부터 VCF 파일을 다운로드하고 정준 코딩 영역 내에서 변이체를 직접 추출하였다.

[0657] 다른 유인원 및 포유류와 비교하기 위해, 본 발명자들은 레서스, 마모셋, 돼지, 소, 염소, 마우스, 및 닭을 포함하여 dbSNP로부터 다른 몇 종의 SNP를 또한 다운로드하였다. dbSNP는 해당 종에 대해 제한적인 수의 변이체를 제공하므로, 개, 고양이 또는 양과 같은 다른 종들을 버렸다. 처음에는 각 종의 SNP를 hg19로 올렸다. 변이체의 약 20%가 유사 유전자 영역에 맵핑되어 있다고 밝혀졌다. 이어서, 본자들은 정준 코딩 영역의 100마리 척추동물의 다중 정렬 파일로부터 엑손 좌표를 취득하였으며, 그러한 영역들 내의 변이체를 추출하였다. 이어서, 추출된 SNP를 hg19로 올렸다. 변이체들이 정렬로부터 종들의 다른 게놈 빌드에 있는 경우, 먼저 정렬의 게놈 빌드에 변이체를 들어올렸다.

[0658] 소 SNP 데이터는 다양한 연구에서 나오므로, dbSNP로부터 모든 대량의 소 변이체(VCF 파일>100MB의 16개 일괄)을 다운로드하였고, 각각의 일괄에 대하여 미스센스 대 동의 비율을 연산함으로써 소 SNP의 상이한 일괄들의 품질을 평가하였다. 미스센스 대 동의 비율의 중앙값은 0.781이고, 중앙 절대 편차(MAD)는 0.160이다(평균은 0.879, SD는 0.496). 이상점 비율(outlier ratio)을 갖는 2개의 일괄(1.391의 비율을 갖는 snpBatch_1000_BULL_GENOMES_1059190.gz 및 2.568의 비율을 갖는 snpBatch_COFACTOR_GENOMICS_1059634.gz)은 추가 분석에서 제외되었다.

[0659] **유인원과 포유류의 다형성의 특성 평가**

[0660] 유인원 SNP의 유용성을 입증하기 위해, 싱글톤과 일반 SNP의 수(대립유전자 빈도(AF)>0.1%)의 비율을 측정하는 농축 점수를 고려하였다. 동의 변이체들은, 양성이며 일반적으로 선택 압력 없이 중립적으로 진화하는 것으로 알려져 있다. 해로운 미스센스 변이체는 자연 선택에 의해 점진적으로 제거되므로, 대립유전자 빈도 분포는 동의 변이체들에 비해 희귀 변이체들을 초과하는 경향이 있다.

[0661] 영장류, 포유류, 및 가금류에서 관찰된 SNP와 중복되는 그러한 gnomAD SNP에 초점을 맞추었다. 종당 동의어와 미스센스 변이체의 수를 계수하였다. 미스센스 변이체의 경우, "미스센스 동일"이라고 칭하는 다른 종의 동일한 아미노산 변화를 공유하는 변이체들 및 "미스센스 상이"라고 칭하는 다른 종의 상이한 아미노산 변화를 갖는 변

이체들을 두 개의 유형으로 추가로 분류하였다. 이어서, 종당 농축 점수를 싱글톤 대 공통 변이체의 수의 비율로서 연산하였다.

[0662] 또한, 각각의 종에 대한 동의 변이체와 미스센스 동일한 변이체 간의 농축 점수를 비교하기 위해 카이-제곱(χ^2) 2×2 분할표에 대한 균질성의 테스트를 수행하였다. 모든 영장류는 동의 변이체와 미스센스 동일 변이체 간에 농축 점수에서 유의미한 차이가 없는 반면, 소, 마우스, 마우스 및 닭은 유의미한 차이를 나타낸다.

[0663] 결과는 유인원에서 동일한 아미노산 변화를 공유하는 그러한 SNP가 동의 SNP와 매우 유사한 농축 점수를 갖는 경향이 있음을 나타내었고, 이는 인간 건강에 순한 영향을 미치는 경향이 있음을 암시한다. 상이한 아미노산 변화를 갖거나 유인원에 부재한 것들은 동의 SNP의 농축 점수로부터 현저하게 벗어난 농축 점수를 갖는다. 비영장류 종으로부터의 미스센스 다형성은, 또한, 동의 변이체와 다른 대립유전자 빈도 분포를 갖는다. 결론은, 유인원에서 동일한 아미노산 변화를 공유하는 SNP가 양성 변이체의 트레이닝 세트에 추가될 수 있다는 것이다.

[0664] 본 발명의 가정은, 대부분의 변이체가 상태에 따라 동일한(IBD)에 의해 생성되지 않고 독립적으로 유도된다는 것이다. 따라서, 이들의 농축 점수의 다른 거동을 평가하기 위해 IBD SNP에서의 희귀한 변이체의 농축 분석을 수행하였다. IBD SNP는, 침팬지, 보노보, 고릴라, S. 오랑우탄 및 B. 오랑우탄을 포함하여 인간과 두 마리 이상의 유인원 종에서 나타나는 인간 SNP로서 정의된다. 이어서, 싱글톤의 수를 공통 변이체의 수(AF>0.1%)로 나눈 것으로서 정의되는 농축 점수는, 중립으로 간주되고 비교를 위한 베이스라인으로서 기능하는 미스센스 변이체 및 동의 변이체에 대해 별도로 계산된다.

[0665] **포유류 종들 간의 고정 치환**

[0666] **고정 치환의 농축 분석**

[0667] 또한, 종간 치환의 희귀한 변이체 농축 분석을 연구하였다. UCSC 게놈 브라우저로부터 100마리 척추동물 종의 계통발생 트리를 다운로드하였다(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh). 이어서, 쌍별 계통 발생 거리를 계산하고 밀접하게 관련된 종 쌍들을 선택하였다(거리<0.3). 영장류 종 쌍을 취득하기 위해, UCSC 게놈 브라우저로부터 CDS 영역에 대한 인간과 19마리 포유류(16마리 영장류) 게놈의 정렬 hg38을 다운로드하였다. 4마리의 영장류 쌍이 13마리의 척추동물 쌍에 추가되었다. 다음 표는, 일 구현예에 따라 밀접하게 관련된 종들의 다수의 쌍의 유전자 거리를 나타낸다.

[0668]

종 1	종 2	거리
침팬지	보노보	0.00799195
레서스	게먹이원숭이	0.00576807
마모셋	다람쥐원숭이	0.0680206
긴코원숭이	금빛원숭이	0.01339514
개코원숭이	사바나원숭이	0.045652
말	흰코뿔소	0.084397
양	염소	0.1
마우스	쥐	0.176098
친칠라	브러시꼬리쥐	0.16
황금두더지	텐렉	0.265936
중국햄스터	골든햄스터	0.08
데이비드 미오티스 박쥐	작은박쥐	0.08254
돌고래	범고래	0.074368
세이커매	송골매	0.2
닭	칠면조	0.126972
푸른바다거북	비단거북	0.2
자라	무른갑가시자라	0.2

[0669] 정준 코딩 영역 내에서 인간과 19마리 포유류 또는 99마리 척추동물 게놈의 다수의 정렬을 취하고, 각각의 선택된 척추동물 쌍들 간의 뉴클레오타이드 치환을 취득하였다. 이들 치환은 gnomAD로부터의 인간 엑솜 SNP에 맵핑

되었으며, 종 쌍과 인간 변이체 간에 동일한 코돈 변화를 필요로 하였다. 변이체들을, 동의 변이체, 다른 종에서 동일한 아미노산 변화를 공유하는 미스센스 변이체, 및 다른 종에서 상이한 아미노산 변화를 갖는 미스센스 변이체인 3가지 유형으로 분류하였다. 농축 점수는 종 쌍당 각 클래스에 대해 연산되었다.

[0670] **종내 및 종간 다형성의 비교**

[0671] 종내 변이체 및 종간 변이체가 이들 종에 대해 이용 가능하기 때문에, 침팬지, 레서스, 마모셋, 염소, 마우스 및 닭을 포함하여 종내 및 종간 다형성의 비교를 수행하기 위해 6종이 선택되었다. 종내 및 종간 변이체의 농축 점수의 비교는 2개의 2×2 분할표의 교차비의 비교와 유사하다. Woolf 테스트는 일반적으로 분할표들 간의 교차비의 동질성을 평가하는 데 적용된다. 따라서, Woolf 테스트를 이용하여 종내 및 종간 다형성 간의 농축 점수의 차를 평가하였다.

[0672] **유전자당 농축 분석**

[0673] 도 64는 유전자당 농축 분석의 일 구현예를 도시한다. 일 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, 병원성인 것으로 결정된 변이체의 병원성을 확인하는 유전자당 농축 분석을 구현하도록 구성된다. 유전자 장애를 갖는 개체의 코호트로부터 샘플링된 특정 유전자의 경우, 유전자당 농축 분석은, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 적용하여 병원성인 특정 유전자의 후보 변이체를 식별하는 것, 후보 변이체의 관찰된 트라이뉴클레오타이드 돌연변이율을 합산하고 그 합을 전염 계수치 및 코호트의 크기와 곱하는 것에 기초하여 특정 유전자에 대한 돌연변이의 베이스라인 수를 결정하는 것, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 적용하여 병원성인 특정 유전자의 드 노보 미스센스 변이체를 식별하는 것, 및 돌연변이의 베이스라인 수를 드 노보 미스센스 변이체의 계수치와 비교하는 것을 포함한다. 비교 출력에 기초하여, 유전자당 농축 분석은, 특정 유전자가 유전 장애에 연관되어 있음을 드 노보 미스센스 변이체가 병원성임을 확인한다. 일부 구현예에서, 유전 장애는 자폐 스펙트럼 장애(autism spectrum disorder: ASD)이다. 다른 구현예에서, 유전 장애는 발달 지연 장애(developmental delay disorder: DDD)이다.

[0674] 도 64에 도시된 예에서, 특정 유전자의 5개의 후보 변이체는, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자에 의해 병원성으로 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 전염/염색체 계수치(2)와 곱한 것에 기초하여 10^{-5}인 것으로 결정된다. 이것은 이어서 드 노보 변이체 계수치(3)와 비교된다.

[0675] 일부 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, p-값을 생성하는 통계 테스트를 출력으로서 사용하여 비교를 수행하도록 구성된다.

[0676] 다른 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, 돌연변이의 베이스라인 수를 드 노보 미스센스 변이체의 계수치와 비교하고 비교의 출력에 기초하여 특정 유전자가 유전 장애와 연관되지 않으며 드 노보 미스센스 변이체가 양성임을 확인하도록 구성된다.

[0677] **게놈 전체 농축 분석**

[0678] 도 65는 게놈 전체 농축 분석의 일 구현예를 도시한다. 다른 일 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, 병원성으로 결정된 변이체의 병원성을 확인하는 게놈 전체 농축 분석을 구현하도록 구성된다. 게놈 전체 농축 분석은, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 적용하여 건강한 개체의 코호트로부터 샘플링된 복수의 유전자 내의 병원성인 드 노보 미스센스 변이체들의 제1 세트를 식별하는 것, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 적용하여 유전 장애가 있는 개체의 코호트로부터 샘플링된 복수의 유전자 내의 병원성인 드 노보 미스센스 변이체들의 제2 세트를 식별하는 것, 및 제1 및 제2 세트의 각 계수치를 비교하고, 비교의 출력에 기초하여 드 노보 변이체들의 제2 세트가 유전 장애가 있는 개체의 코호트에 농축되어 있고 이에 따라 병원성을 확인하는 것을 포함한다. 일부 구현예에서, 유전 장애는 자폐 스펙트럼 장애(약칭 ASD)이다. 다른 구현예에서, 유전 장애는 발달 지연 장애(약칭 DDD)이다.

[0679] 일부 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, p-값을 출력으로서 생성하는 통계 테스트를 사용하여 비교를 수행하도록 구성된다. 일 구현예에서, 비교는 각각의 코호트 크기에 의해 추가로 파라미터화된다.

[0680] 일부 구현예에서, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자는, 또한, 제1 및 제2 세트의 각각의 계수치

를 비교하고, 비교의 출력에 기초하여 드 노보 미스센스 변이체들의 제2 세트가 유전 장애를 가진 개체의 코호트에 농축되어 있지 않으며 이에 따라 양성임을 확인하도록 구성된다.

[0681] 도 65에 도시된 예에서, 건강한 코호트에서의 돌연변이율(0.001) 및 영향을 받는 코호트에서의 돌연변이율(0.004)은 개체별 돌연변이 양(4)과 함께 도시되어 있다.

[0682] **구체적인 구현예**

[0683] 변이체 병원성 분류자를 구성하기 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 설명의 일부 구현에서 누락된 부분은 이전 부문에서 설명한 조합들을 제한하는 것으로 간주되어서는 안 되며, 이러한 설명은 이하의 각 구현예에 참조로 통합되는 것이다.

[0684] 개시된 기술의 시스템 구현예는 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는, 게놈 서열(예를 들어, 뉴클레오타이드 서열)에서 스플라이스 부위를 식별하는 스플라이스 부위 검출기를 트레이닝하기 위한 컴퓨터 명령어가 로딩된다.

[0685] 도 48 및 도 19에 도시된 바와 같이, 시스템은 메모리에 연결된 다수의 프로세서 상에서 실행되는 컨볼루션 신경망 기반 변이체 병원성 분류자를 트레이닝한다. 시스템은, 양성 변이체 및 병원성 변이체로부터 생성된 단백질 서열 쌍의 양성 트레이닝 예 및 병원성 트레이닝 예를 사용한다. 양성 변이체는, 인간과 일치하는 참조 코돈 서열을 공유하는 대체 비-인간 영장류 코돈 서열에서 발생하는 비-인간 영장류 미스센스 변이체 및 공통 인간 미스센스 변이체를 포함한다. "단백질 서열 쌍"이라는 어구는 참조 단백질 서열 및 대체 단백질 서열을 지칭하며, 여기서 참조 단백질 서열은 참조 트리플릿 뉴클레오타이드 염기(참조 코돈)에 의해 형성된 참조 아미노산을 포함하고, 대체 단백질 서열은, 대체 단백질 서열이 참조 단백질 서열의 참조 아미노산을 형성하는 참조 트리플릿 뉴클레오타이드 염기(참조 코돈)에서 발생하는 변이체의 결과로서 생성되도록 대체 트리플릿 뉴클레오타이드 염기(대체 코돈)에 의해 형성된 대체 아미노산을 포함한다. 변이체는 SNP, 삽입, 또는 결실일 수 있다.

[0686] 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 법정 클래스 세트에 대하여 반복되지 않는다. 독자는, 이 부문에서 식별되는 피처를 다른 법정 클래스의 기본 피처와 쉽게 결합할 수 있는 방법을 이해할 것이다.

[0687] 도 44에 도시된 바와 같이, 공통 인간 미스센스 변이체는, 적어도 100000명의 인간으로부터 샘플링된 인간 모집단 변이체 데이터세트에 걸쳐 0.1% 초과의 소수 대립유전자 빈도(약칭 MAF)를 갖는다.

[0688] 도 44에 도시된 바와 같이, 샘플링된 인간은 상이한 인간 하위집단에 속하고, 공통 인간 미스센스 변이체는 각각의 인간 하위집단 변이체 데이터세트 내에서 0.1%보다 큰 MAF를 갖는다.

[0689] 인간 하위집단은, 아프리카/아프리카계 미국인(African/African American: AFR), 미국인(American: AMR), 애쉬케나지 유대인(Ashkenazi Jewish: ASJ), 동아시아인(East Asian: EAS), 핀란드인(Finnish: FIN), 비핀란드계 유럽인(Non-Finnish European: NFE), 남아시아인(South Asian: SAS) 및 기타(Others: OTH)를 포함한다.

[0690] 도 43 및 도 44에 도시된 바와 같이, 비인간 영장류 미스센스 변이체는, 침팬지, 보노보, 고릴라, B. 오랑우탄, S. 오랑우탄, 레서스 및 마모셋을 포함하는 복수의 비인간 영장류 종으로부터의 미스센스 변이체를 포함한다.

[0691] 도 45 및 도 46에 도시된 바와 같이, 농축 분석에 기초하여, 시스템은, 양성 변이체 중 특정 비인간 영장류 종의 미스센스 변이체를 포함시키기 위해 특정 비인간 영장류 종을 수용한다. 농축 분석은, 특정 비인간 영장류 종에 대해, 특정 비인간 영장류 종의 동의 변이체의 제1 농축 점수를 특정 비인간 영장류 종의 미스센스 동일 변이체의 제2 농축 점수와 비교하는 것을 포함한다.

[0692] 도 45는 인간 이종상동성 SNP의 일 구현예를 도시한다. 비인간 종의 미스센스 SNP는 인간과 일치하는 참조 코돈 및 대체 코돈을 갖는다. 도 45에 도시된 바와 같이, 미스센스 동일 변이체는, 인간과 일치하는 참조 코돈 서열 및 대안 코돈 서열을 공유하는 미스센스 변이체이다.

[0693] 도 46 및 도 47에 도시된 바와 같이, 제1 농축 점수는, 0.1%보다 큰 MAF를 갖는 공통 동의 변이체에 비해 0.1% 미만의 MAF를 갖는 희귀 동의 변이체의 비율을 결정함으로써 생성된다. 제2 농축 점수는, 0.1%보다 큰 MAF를 갖

는 공통 미스센스 동일 변이체에 비해 0.1% 미만의 MAF를 갖는 희귀 미스센스 동일 변이체의 비율을 결정함으로써 생성된다. 희귀 변이체에는 싱글톤 변이체가 포함된다.

[0694] 도 46 및 도 47에 도시된 바와 같이, 제1 농축 점수와 제2 농축 점수 간의 차는 소정의 범위 내에 있으며, 양성 변이체 중 특정 비인간 영장류의 미스센스 변이체를 포함시키기 위해 특정 비인간 영장류 종을 수용하는 것을 추가로 포함한다. 소정의 범위 내에 있는 차는, 미스센스 동일 변이체가 동의 변이체와 동일한 정도의 자연 선택 하에 있고 따라서 동의 변이체로서 양성임을 나타낸다.

[0695] 도 48에 도시된 바와 같이, 시스템은, 양성 변이체 중 비인간 영장류 종의 미스센스 변이체를 포함시키기 위해 복수의 비인간 영장류 종을 수용하도록 농축 분석을 반복적으로 적용한다. 시스템은, 또한, 각각의 비인간 영장류 종에 대한 동의 변이체들의 제1 농축 점수와 미스센스 동일 변이체들의 제2 농축 점수를 비교하기 위한 균질성의 카이 제곱 테스트를 포함한다.

[0696] 도 48에 도시된 바와 같이, 비인간 영장류 미스센스 변이체의 계수치는 적어도 100000이고, 비인간 영장류 미스센스 변이체의 계수치는 385236이다. 공통 인간 미스센스 변이체의 계수치는 50000 이상이다. 공통 인간 미스센스 변이체의 계수치는 83546이다.

[0697] 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.

[0698] 개시된 기술의 다른 시스템 구현예는 단일 뉴클레오타이드 다형성(약칭 SNP) 병원성 분류자를 구성하는 것을 포함한다. 이 시스템은, 양성 SNP 및 병원성 SNP에 의해 발현되는 아미노산 서열의 양성 트레이닝 예 및 병원성 트레이닝 예를 사용하여, 메모리에 연결된 다수의 프로세서 상에서 실행되는 컨볼루션 신경망 기반 SNP 병원성 분류자를 트레이닝한다. 양성 트레이닝 예는 아미노산 서열 쌍으로 표현된 제1 및 제2 뉴클레오타이드 서열 세트를 포함하며, 각각의 아미노산 서열은, 상류 아미노산과 하류 아미노산이 측면에 있는 중심 아미노산을 포함한다. 각각의 아미노산 서열 쌍은, 참조 뉴클레오타이드 서열에 의해 발현되는 아미노산의 참조 서열 및 SNP를 함유하는 대체 뉴클레오타이드 서열에 의해 발현되는 대체 아미노산 서열을 포함한다.

[0699] 도 9에 도시된 바와 같이, 제1 세트는 인간 뉴클레오타이드 서열 쌍들을 포함하며, 각 쌍은, SNP를 함유하고 인간 모집단 내에서 공통인 것으로 여겨지는 소수 대립유전자 빈도(MAF)를 갖는 인간 대체 뉴클레오타이드 서열을 포함한다. 제2 세트는 비인간 영장류 대체 뉴클레오타이드 서열과 쌍을 이루는 비인간 영장류 참조 뉴클레오타이드 서열을 포함한다. 비인간 영장류 참조 뉴클레오타이드 서열은 이종상동성 인간 뉴클레오타이드 참조 서열을 갖는다. 비인간 영장류 대체 뉴클레오타이드 서열은 SNP를 포함한다.

[0700] 제1 시스템 구현을 위한 이러한 구체적인 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0701] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서예 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는 전술한 시스템의 동작을 수행하는 방법을 포함할 수 있다.

[0702] 도 48 및 도 19에 도시된 바와 같이, 개시된 기술의 제1 방법 구현예는 변이체 병원성 분류자를 구성하는 단계를 포함하고, 방법은 다음을 포함한다. 이 방법은, 양성 변이체 및 병원성 변이체로부터 생성된 단백질 서열 쌍의 양성 트레이닝 예 및 병원성 트레이닝 예를 사용하여, 메모리에 연결된 다수의 프로세서 상에서 실행되는 컨볼루션 신경망 기반 변이체 병원성 분류자를 트레이닝하는 단계를 더 포함한다. 양성 변이체는, 인간과 일치하는 참조 코돈 서열을 공유하는 대체 비인간 영장류 코돈 서열에서 발생하는 비인간 영장류 미스센스 변이체 및 공통 인간 미스센스 변이체를 포함한다.

[0703] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0704] 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 명령어를 실행하여 상술한 방법을 수행하도록 동작가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0705] 도 48 및 도 19에 도시된 바와 같이, 개시된 기술의 제2 방법 구현은 단일 뉴클레오타이드 다형성(SNP) 병원성

분류자를 구성하는 단계를 포함한다. 상기 방법은, 양성 SNP 및 병원성 SNP에 의해 발현되는 아미노산 서열의 양성 트레이닝 예 및 병원성 트레이닝 예를 사용하여, 메모리에 연결된 다수의 프로세서 상에서 실행되는 컨볼루션 신경망 기반 SNP 병원성 분류자를 트레이닝하는 단계를 더 포함한다. 양성 트레이닝 예는 아미노산 서열 쌍으로 표현된 뉴클레오타이드 서열들의 제1 및 제2 세트를 포함하고, 각각의 아미노산 서열은 업스트림 및 다운스트림 아미노산이 옆에 있는 중심 아미노산을 포함하고, 각각의 아미노산 서열 쌍은, 참조 뉴클레오타이드 서열에 의해 표현된 아미노산의 참조 서열 및 SNP를 함유하는 대체 뉴클레오타이드 서열에 의해 표현된 아미노산의 대체 서열을 포함한다. 제1 세트는 인간 뉴클레오타이드 서열 쌍들을 포함하고, 각 쌍은, SNP를 함유하고 인간 모집단 내에서 흔한 것으로 간주되는 소수 대립유전자 빈도(MAF)를 갖는 인간 대체 뉴클레오타이드 서열을 포함한다. 제2 세트는 비인간 영장류 대체 뉴클레오타이드 서열과 쌍을 이루는 비인간 영장류 참조 뉴클레오타이드 서열을 포함한다. 비인간 영장류 참조 뉴클레오타이드 서열은 이종상동성 인간 뉴클레오타이드 참조 서열을 가지며, 비 인간 영장류 대체 뉴클레오타이드 서열은 SNP를 포함한다.

[0706] 제2 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0707] 다른 구현예는, 상술한 방법을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는, 메모리 및 메모리에 저장된 명령어를 실행하여 상술한 방법을 수행하도록 동작가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0708] 이차 구조 및 용매 접근성 분류자를 갖는 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 사용하기 위한 시스템, 방법, 및 제조 물품을 설명한다. 구현예의 하나 이상의 피치는 기본 구현예와 결합될 수 있다. 상호 배타적이지 않은 구현예들은 결합가능하도록 교시된다. 구현예의 하나 이상의 피치는 다른 구현예와 결합될 수 있다. 본 개시 내용은 주기적으로 사용자에게 이러한 선택사항을 상기시킨다. 이러한 선택사항을 반복하는 설명의 일부 구현에서 누락된 부분은 이전 부문에서 설명된 조합을 제한하는 것으로서 간주되어서는 안 되며, 이러한 설명은 이하의 각 구현에 참조로 통합된다.

[0709] 개시된 기술의 시스템 구현예는 메모리에 연결된 하나 이상의 프로세서를 포함한다. 메모리에는, 이차 구조 및 용매 접근성 분류자를 갖는 심층 컨볼루션 신경망 기반 변이체 병원체 분류자를 실행하기 위한 컴퓨터 명령어가 로딩된다.

[0710] 시스템은, 단백질 서열 내의 아미노산 위치에서 3-상태 이차 구조를 예측하도록 트레이닝된, 메모리에 연결된 다수의 프로세서 상에서 실행되는 제1 이차 구조 서브네트워크를 포함한다. 시스템은, 단백질 서열 내의 아미노산 위치에서 3-상태 용매 접근성을 예측하도록 트레이닝된, 메모리에 연결된 다수의 프로세서 상에서 실행되는 제2 용매 접근성 서브네트워크를 더 포함한다.

[0711] 3-상태 이차 구조는, 복수의 DNA 이차 구조 상태 알파 나선(H), 베타 시트(B), 및 코일(C) 중 하나를 지칭한다.

[0712] 3-상태 용매 접근성은, 매립된(B), 개재된(I) 및 노출된(E) 복수의 단백질 용매 접근성 상태 중 하나를 지칭한다.

[0713] 다수의 프로세서 중 적어도 하나에서 실행되는 위치 빈도 행렬(약칭 PFM) 생성기는, 영장류, 포유류, 및 영장류와 포유류를 제외한 척추동물의 3개의 서열 그룹에 적용되어, 영장류 PFM, 포유류 PFM을 생성하고, 척추 동물 PFM을 생성한다.

[0714] 다시 말하면, 이것은, 영장류 서열 데이터에 PFM 생성기를 적용하여 영장류 PFM을 생성하는 것, 포유류 서열 데이터에 PFM 생성기를 적용하여 포유류 PFM을 생성하는 것, 및 영장류와 포유류를 포함하지 않는 척추동물 서열 데이터에 PFM 생성기를 적용하여 척추동물 PFM을 생성하는 것을 포함한다.

[0715] 입력 프로세서는, 각각의 방향으로 25개 이상의 아미노산이 상류측과 하류측에 있는 표적 변이체 아미노산을 갖는 변이체 아미노산 서열을 수용하고, 단일 뉴클레오타이드 변이체는 표적 변이체 아미노산을 생성한다. 다수의 프로세서 중 적어도 하나에서 실행되는 보충 데이터 할당기는, 변이체 아미노산 서열과 정렬된, 각 방향으로 적어도 25개 아미노산이 상류측과 하류측에 있는 표적 참조 아미노산을 갖는 참조 아미노산 서열을 할당한다. 이어서, 참조 아미노산 서열에 대한 제1 및 제2 서브네트워크에 의해 생성된 참조 상태 분류를 할당한다. 이후, 보충 데이터 할당기는, 변이체 아미노산 서열에 대해 제1 및 제2 서브네트워크에 의해 생성된 변이체 상태 분류를 할당한다. 마지막으로, 참조 아미노산 서열과 정렬된 영장류 PFM, 포유류 PFM, 및 척추 동물 PFM을 할당한다.

[0716] 본 명세서의 맥락에서, "정렬된"이란 어구는, 참조 아미노산 서열 또는 대체 아미노산 서열의 각 아미노 위치에 대하여 영장류 PFM, 포유류 PFM, 및 척추동물 PFM을 위치별로 결정하고, 아미노산 위치가 참조 아미노산 서열 또는 대체 아미노산 서열에서 발생할 때 동일한 순서로 위치별 또는 서수적 위치 기반에 대한 결정의 결과를 인코딩하고 저장하는 것을 가리킨다.

[0717] 상기 시스템은, 또한, 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 할당된 참조 및 변이체 상태 분류, 및 할당된 PFM에 기초하여 변이체 아미노산 서열을 양성 또는 병원성으로 분류하도록 트레이닝된, 다수의 프로세서 상에서 실행되는 심층 컨볼루션 신경망을 포함한다. 시스템은 변이체 아미노산 서열에 대한 병원성 점수를 적어도 보고하는 출력 프로세서를 포함한다.

[0718] 개시된 이러한 시스템 구현예 및 다른 시스템들은 이하의 피처 중 하나 이상을 선택적으로 포함한다. 시스템은, 또한, 개시된 방법들과 관련하여 설명된 피치들을 포함할 수 있다. 간결성을 위해, 시스템 피치들의 대체 조합은 개별적으로 열거되지 않는다. 시스템, 방법, 및 제조 물품에 적용되는 피처들은 기본 피처들의 각 법정 클래스에 대해 반복되지 않는다. 독자는 이 부문에서 식별된 피처들을 다른 법정 클래스의 기본 피처들과 쉽게 결합할 수 있는 방법을 이해할 것이다.

[0719] 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 포함하는 시스템은, 또한, 병원성 점수에 기초하여 단일 뉴클레오타이드 변이체를 양성 또는 병원성으로 분류하도록 구성된다.

[0720] 시스템은 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 포함하며, 심층 컨볼루션 신경망은, 적어도 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 할당된 변이체 이차 구조 상태 분류, 할당된 참조 이차 구조 상태 분류, 할당된 변이체 용매 접근성 상태 분류, 할당된 참조 용매 접근성 상태 분류, 할당된 영장류 PFM, 할당된 포유류 PFM, 및 할당된 척추동물 PFM을 입력으로서 병렬로 수용한다.

[0721] 이 시스템은, 일괄 정규화층, ReLU 비선형성 층, 및 차원 변경층을 사용하여 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 할당된 영장류 PFM, 할당된 포유류 PFM, 및 할당된 척추동물 PFM을 전처리하도록 구성된다. 시스템은, 또한, 전처리된 특성화들을 합산하고, 그 합을, 할당된 변이체 이차 구조 상태 분류, 할당된 참조 이차 구조 상태 분류, 할당된 변이체 용매 접근성 상태 분류, 및 할당된 참조 용매 접근성 상태 분류와 연쇄화하여 연쇄화된 입력을 생성하도록 구성된다. 시스템은, 차원 변경층을 통해 연쇄화된 입력을 처리하고 처리된 연쇄화된 입력을 수용하여 심층 컨볼루션 신경망의 잔여 블록을 개시한다.

[0722] 심층 컨볼루션 신경망은, 가장 낮은 것에서 가장 높은 것까지 순서대로 배열된 잔여 블록들의 그룹들을 포함한다. 심층 컨볼루션 신경망은, 비선형 활성화 없이 잔여 블록의 수, 스킵 연결의 수 및 잔여 연결의 수에 의해 파라미터화된다. 심층 컨볼루션 신경망은 선행 입력의 공간 및 특징 치수를 재형상화하는 차원 변경층을 포함한다.

[0723] 시스템은, 또한, 영장류, 포유류, 및 척추동물에 걸쳐 정렬된 참조 아미노산 서열로 보존되는 표적 참조 아미노산으로부터 표적 변이체 아미노산을 생성하는 단일 뉴클레오타이드 변이체를 병원성으로 분류하게끔 트레이닝하도록 구성된다.

[0724] 보존은 표적 참조 아미노산의 기능적 유의성(functional significance)을 나타내고 PFW로부터 결정된다. 시스템은, 또한, 변이체 아미노산 서열과 참조 변이체 아미노산 서열 간의 상이한 이차 구조를 야기하는 단일 뉴클레오타이드 변이체를 병원성으로 분류하게끔 트레이닝하도록 구성된다.

[0725] 시스템은, 또한, 변이체 아미노산 서열과 참조 변이체 아미노산 서열 간의 상이한 용매 접근성을 야기하는 단일 뉴클레오타이드 변이체를 병원성으로 분류하게끔 트레이닝하도록 구성된다.

[0726] PFM은, 위치별로, 다른 종들의 정렬된 단백질 서열들에 걸쳐 인간 단백질 서열에서의 아미노산 발생 빈도를 위치별로 결정함으로써 다른 종들의 정렬된 단백질 서열들에 걸쳐 인간 단백질 서열에서의 아미노산의 보존을 나타낸다.

[0727] 이차 구조의 3개 상태는 나선, 시트 및 코일이며, 제1 이차 구조 서브네트워크는, 입력 단백질 서열 및 입력 단백질 서열 내의 아미노산 위치와 정렬된 영장류 PFM, 포유류 PFM, 및 척추동물 PFM을 수용하고 각각의 아미노산 위치에서 3-상태 이차 구조를 예측하도록 트레이닝된다. 용매 접근성의 3개 상태는 노출, 매립 및 개재이다.

[0728] 제2 용매 접근성 서브네트워크는, 입력 단백질 서열 및 입력 단백질 서열 내의 아미노산 위치와 정렬된 영장류 PFM, 포유류 PFM, 및 척추동물 PFM을 수용하고 각각의 아미노산 위치에서 3-상태 용매 접근성을 예측하도록 트레이닝된다. 입력 단백질 서열은 참조 단백질 서열이다. 입력 단백질 서열은 대체 단백질 서열이다. 제1 이차

구조 서브네트워크는 최저에서 최고로 순차적으로 배열된 잔여 블록들의 그룹들을 포함한다. 제1 이차 구조 서브네트워크는 비선형 활성화 없이 잔여 블록의 수, 스킵 연결의 수 및 잔여 연결의 수에 의해 파라미터화된다.

[0729] 제1 이차 구조 서브네트워크는 선행 입력의 공간 및 피처 치수를 재형상화하는 차원 변경층을 포함한다. 제2 용매 접근성 서브네트워크는 최저에서 최고로 순차적으로 배열된 잔여 블록들의 그룹들을 포함한다. 제2 용매 접근성 서브네트워크는 비선형 활성화 없이 잔여 블록의 수, 스킵 연결의 수 및 잔여 연결의 수에 의해 파라미터화된다. 제2 용매 접근성 서브네트워크는 선행 입력의 공간 및 피처 치수를 재형상화하는 차원 변경층을 포함한다.

[0730] 각각의 잔여 블록은, 적어도 하나의 일괄 정규화층, 적어도 하나의 정류된 선형 유닛(약칭 ReLU) 층, 적어도 하나의 차원 변경층, 및 적어도 하나의 잔여 연결을 포함한다. 각각의 잔여 블록은, 2개의 일괄 정규화층, 2개의 ReLU 비선형성 층, 2개의 차원 변경층, 및 1개의 잔여 연결을 포함한다.

[0731] 심층 컨볼루션 신경망, 제1 이차 구조 서브네트워크, 및 제2 용매 접근성 서브네트워크는 각각 최종 분류층을 포함한다. 최종 분류층은 시그모이드 기반 층이다. 최종 분류층은 소프트맥스 기반 층이다.

[0732] 시스템은, 또한, 심층 컨볼루션 신경망과의 협력을 위해 제1 이차 구조 서브네트워크 및 제2 용매 접근성 서브네트워크의 최종 분류층을 제거하도록 구성된다.

[0733] 시스템은, 또한, 심층 컨볼루션 신경망의 트레이닝 동안, 서브네트워크로의 역전파 에러 및 서브네트워크 가중치 업데이트를 포함하여 병원성 분류에 대해 제1 이차 구조 서브네트워크 및 제2 용매 접근성 서브네트워크를 트레이닝하도록 구성된다.

[0734] 제2 용매 접근성 서브네트워크는 적어도 아트러스 컨볼루션층을 포함한다. 시스템은, 또한, 발달 지연 장애(약칭 DDD)-유발 변이체를 병원성으로 분류하도록 추가로 구성된다. 변이체 아미노산 서열 및 참조 아미노산 서열은 측면 아미노산을 공유한다. 시스템은, 또한, 원-핫 인코딩을 사용하여 입력을 심층 컨볼루션 신경망으로 인코딩하도록 구성된다.

[0735] 도 1q는 개시된 기술이 동작될 수 있는 예시적인 연산 환경을 도시한다. 심층 컨볼루션 신경망, 제1 이차 구조 서브네트워크, 및 제2 용매 접근성 서브네트워크는 하나 이상의 트레이닝 서버에서 트레이닝된다. 트레이닝된 심층 컨볼루션 신경망, 제1 트레이닝된 이차 구조 서브네트워크, 및 트레이닝된 제2 용매 접근성 서브네트워크는 요청 클라이언트로부터 입력 서열을 수신하는 하나 이상의 생성 서버에 배치된다. 생성 서버는, 심층 컨볼루션 신경망, 제1 이차 구조 서브네트워크, 및 제2 용매 접근성 서브네트워크 중 적어도 하나를 통해 입력 서열을 처리하여 클라이언트에 전송되는 출력을 생성한다.

[0736] 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.

[0737] 개시된 기술의 또 다른 시스템 구현예는, 메모리에 연결된 다수의 프로세서 상에서 실행되는 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 포함한다. 이 시스템은, 영장류 PFM 및 포유류 PFM을 생성하도록 영장류와 포유 동물의 두 개의 서열기에 적용되는, 다수의 프로세서 중 적어도 하나에서 실행되는 위치 빈도 행렬(약칭 PFM) 생성기를 포함한다. 상기 시스템은, 또한, 각 방향으로 25개 이상의 아미노산이 상류측과 하류측에 있는 표적 변이체 아미노산을 갖는 변이체 아미노산 서열을 수용하는 입력 프로세서를 포함하며, 여기서 단일 뉴클레오타이드 변이체는 표적 변이체 아미노산을 생성한다. 이 시스템은 또한 다수의 프로세서 중 적어도 하나에서 실행되는 보충 데이터 할당자를 포함하며, 이러한 할당자는, 변이체 아미노산 서열과 정렬되는, 각각의 방향으로 적어도 25개의 아미노산이 상류측과 하류측에 있는 표적 참조 아미노산을 갖는 참조 아미노산 서열을 할당한다. 또한, 참조 아미노산 서열과 정렬된 영장류 PFM 및 포유류 PFM을 할당한다. 시스템은, 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 및 할당된 PFM을 기초로 하여 변이체 아미노산 서열을 양성 또는 병원성으로 분류하도록 트레이닝된, 다수의 프로세서 상에서 실행되는 심층 컨볼루션 신경망을 더 포함한다. 마지막으로, 시스템은 변이체 아미노산 서열에 대해 적어도 병원성 점수를 보고하는 출력 프로세서를 포함한다.

[0738] 개시된 본 시스템 구현예 및 다른 시스템은 다음 피처들 중 하나 이상을 선택적으로 포함한다. 시스템은, 또한, 개시된 방법과 관련하여 설명된 피처를 포함할 수 있다. 간결성을 위해, 시스템 기능의 대체 조합은 개별적으로 열거되지 않는다. 시스템, 방법, 및 제조 물품에 적용되는 피처는 염기 피처의 각 법정 클래스에 대해 반복되지 않는다. 독자는 이 부문에서 식별된 피처를 다른 법정 클래스의 기본 피처와 쉽게 결합할 수 있는 방법을 이해

할 것이다.

[0739] 시스템은, 또한, 단일 뉴클레오타이드 변이체를 병원성 점수에 기초하여 양성 또는 병원성으로 분류하도록 구성된다. 심층 컨볼루션 신경망은, 병행하여, 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 할당된 영장류 PFM, 및 할당된 포유류 PFM을 수용하고 처리한다. 시스템은, 또한, 영장류 및 포유류에 걸쳐 참조 아미노산 서열에 보존된 표적 참조 아미노산으로부터 표적 변이체 아미노산을 생성하는 단일 뉴클레오타이드 변이체를 병원성으로 분류하게끔 트레이닝하도록 구성된다. 보존은 표적 참조 아미노산의 기능적 유의성을 나타내며 PFW로부터 결정된다.

[0740] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0741] 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.

[0742] 개시된 기술의 제1 방법 구현예는, 단백질 서열 내의 아미노산 위치에서 3-상태 이차 구조를 예측하도록 트레이닝된, 메모리에 연결된 다수의 프로세서 상에서 제1 이차 구조 서브네트워크를 실행하는 단계를 포함한다. 단백질 서열 내의 아미노산 위치에서 3-상태 용매 접근성을 예측하도록 트레이닝된, 메모리에 연결된 다수의 프로세서 상에서 제2 용매 접근성 서브네트워크를 실행한다. 다수의 프로세서 중 하나 이상에서 영장류, 포유류, 및 영장류와 포유류를 제외한 척추동물의 3개의 서열 그룹에 적용되는 위치 빈도 행렬(PFM) 생성기를 실행하여 영장류 PFM, 포유류 PFM, 및 척추동물 PFM을 생성한다. 각 방향으로 적어도 25개 아미노산이 상류측과 하류측에 있는 표적 변이체 아미노산을 갖는 입력 프로세서는, 변이체 아미노산 서열을 수용한다. 단일 뉴클레오타이드 변이체는 표적 변이체 아미노산을 생성한다. 다수의 프로세서 중 적어도 하나에서 실행되는 보충 데이터 할당기는, 변이체 아미노산 서열과 정렬된, 각 방향으로 적어도 25개의 아미노산이 상류측과 하류측에 위치하는 표적 참조 아미노산을 갖는 참조 아미노산 서열을 할당한다. 보충 데이터 할당기는, 또한, 참조 아미노산 서열에 대한 제1 및 제2 서브네트워크에 의해 생성되는 참조 상태 분류를 할당한다. 보충 데이터 할당기는, 또한, 변이체 아미노산 서열에 대해 제1 및 제2 서브네트워크에 의해 생성되는 변이체 상태 분류를 할당한다. 보충 데이터 할당기는, 참조 아미노산 서열과 정렬된 영장류 PFM, 포유류 PFM, 및 척추동물 PFM을 할당한다. 다수의 프로세서 상에서 실행되는 심층 컨볼루션 신경망은, 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 할당된 참조 및 변이체 상태 분류, 및 할당된 PFM의 처리에 기초하여 변이체 아미노산 서열을 양성 또는 병원성으로 분류하도록 트레이닝된다. 출력 프로세서를 통해 변이체 아미노산 서열에 대한 적어도 병원성 점수를 보고한다.

[0743] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0744] 다른 구현예는 전술한 방법을 수행하도록 프로세서예 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는, 메모리 및 메모리에 저장된 명령어를 실행하여 상술한 방법을 수행하도록 동작가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0745] 개시된 기술의 제2 방법 구현예는, 심층 컨볼루션 신경망 기반 변이체 병원성 분류자를 메모리에 연결된 다수의 프로세서 상에서 실행하는 것을 포함한다. 다수의 프로세서 중 적어도 하나 상에서 영장류 PFM과 포유류 PFM을 생성하도록 영장류와 포유류의 2개의 서열 그룹에 적용되는 위치 빈도 행렬(PFM) 생성기를 실행한다. 입력 프로세서에서는, 각 방향으로 25개 이상의 아미노산이 상류측과 하류측에 있는 표적 변이체 아미노산을 갖는 변이체 아미노산 서열을 수용한다. 단일 뉴클레오타이드 변이체는 표적 변이체 아미노산을 생성한다. 다수의 프로세서 중 하나 이상에서 보충 데이터 할당자를 실행하여, 변이체 아미노산 서열과 정렬된 각 방향으로 25개 이상의 아미노산이 상류측과 하류측에 있는 표적 참조 아미노산을 갖는 참조 아미노산 서열을 할당하고, 영장류 PFM과 포유류 PFM은 참조 아미노산 서열과 정렬된다. 변이체 아미노산 서열, 할당된 참조 아미노산 서열, 및 할당된 PFM을 처리하는 것에 기초하여 변이체 아미노산 서열을 양성 또는 병원성으로 분류하도록 트레이닝된 다수의 프로세서 상에서 심층 컨볼루션 신경망을 실행한다. 출력 프로세서에서는 변이체 아미노산 서열에 대한 병원성 점수를 보고한다.

[0746] 제2 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 방법 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0747] 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 상술한 방법을 수행하도록 동작가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0748] 개시된 기술의 또 다른 시스템 구현예는, 단일 뉴클레오타이드 다형성(SNP) 병원성 분류자를 트레이닝하기 위한 대규모 병원성 트레이닝 데이터를 생성하는 시스템을 포함한다.

[0749] 도 19에 도시된 바와 같이, 시스템은, 양성 SNP의 트레이닝 세트 및 조합적으로 생성된 SNP의 합성 세트로부터 선발제거된(culled) 엘리트 예측 병원성(elite predicted pathogenic) SNP의 트레이닝 세트를 사용하여, 메모리에 연결된 다수의 프로세서 상에서 실행되는 SNP 병원성 분류자를 트레이닝한다. 본 명세서의 컨텍스트에서, 엘리트 예측 병원성 SNP는, 앙상블에 의해 출력되는 바와 같이 해당 평균 또는 최대 병원성 점수에 기초하여 각 사이클의 종료 시 SNP가 생성/선택되는 것이다. "엘리트"라는 용어는, 유전자 알고리즘 어휘에서 빌려 왔으며, 유전자 알고리즘 간행물에서 통상적으로 제공되는 의미를 갖고자 하는 것이다.

[0750] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 시스템은, 예측된 SNP가 없는 것으로 시작하고 합성 세트로부터 이상점 SNP를 선발제거함으로써 예측된 SNP의 풀 세트를 누적하여 엘리트 세트를 주기적으로 반복하여 구축한다. 합성 세트는, 양성 세트에 존재하지 않는 조합적으로 생성된 SNP이고 이상점 SNP가 엘리트 세트에 포함되도록 합성 세트로부터 반복적으로 선발제거됨에 따라 세트 멤버십이 감소되는 의사-병원성(pseudo-pathogenic) SNP를 포함한다. 본 발명과 관련하여, "선발제거"이라는 용어는, 이전 모집단을 새로운 모집단으로 필터링, 대체, 업데이트, 또는 선택하는 것을 의미한다. "선발제거"이라는 용어는, 유전자 알고리즘 어휘에서 빌려 왔으며, 유전자 알고리즘 간행물에 통상적으로 주어진 의미를 갖고자 하는 것이다.

[0751] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 시스템은, SNP 병원성 분류자들의 앙상블을 트레이닝하고 적용하여 합성 세트로부터 이상점 SNP를 주기적으로 반복적으로 선발제거한다. 이것은, 양성 SNP의 공통 트레이닝 세트, 엘리트 예측 병원성 SNP의 공통 트레이닝 세트, 및 교체 없이 합성 세트로부터 샘플링된 의사-병원성 SNP의 별개의 트레이닝 세트를 사용하여 앙상블을 트레이닝하는 것을 포함한다. 이것은, 또한, 트레이닝된 앙상블을 적용하여 현재 사이클에서 앙상블을 트레이닝하는 데 사용되지 않았던 합성 세트 중 적어도 일부의 SNP를 채점하고 점수를 이용하여 채점된 SNP로부터 현재 사이클의 이상점 SNP를 선택하여 공통 엘리트 세트에 축적함으로써, 트레이닝된 앙상블을 적용하여 합성 세트로부터 이상점 SNP를 선발제거하고 공통 엘리트 세트에 선발제거된 이상점 SNP를 축적하는 것을 포함한다.

[0752] 본 명세서의 컨텍스트에서, "의사-병원성 SNP"는, 트레이닝 목적으로 병원성으로 표지되고 트레이닝 동안 교체 없이 합성적으로 생성된 변이체로부터 샘플링된 SNP이다.

[0753] 또한, 엘리트 예측 병원성 SNP의 트레이닝 세트는 여러 사이클에 걸쳐 반복적으로 구성된다.

[0754] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 시스템은, 트레이닝에 의해 도출된 분류자 파라미터, 사이클에 걸쳐 그리고 공통 양성 세트의 미리 결정된 확산 내에 완료된 공통 엘리트 세트, 및 SNP 병원성 분류자를 트레이닝하기 위한 공통 양성 세트를 메모리에 저장한다.

[0755] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시한 바와 같이, 엘리트 예측 병원성 SNP는 앙상블에 의해 예측된 SNP의 상위 5%이다. 일부 구현예에서는, 최상 점수의 SNP의 고정된 수, 예컨대 20000이 있다.

[0756] SNP 병원성 분류자 및 SNP 병원성 분류자의 앙상블 각각은 심층 컨볼루션 신경망(약칭 DCNN)이다. 앙상블에는 4개 내지 16개의 DCNN이 포함된다. 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시한 바와 같이, 앙상블은 8개의 DCNN을 포함한다.

[0757] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시한 바와 같이, 시스템은, 사이클 동안 에포크에서 DCNN의 앙상블을 트레이닝하며, 유효성확인 샘플에 대한 예측이 양성 및 병원성 예측의 이산 확률 분포 클러스터를 형성할 때 특정 사이클에 대한 트레이닝을 마친다.

[0758] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 시스템은, 점수를 사용하여 DCNN의 앙상블로부터의 점수를 합산함으로써 현재 사이클의 이상점 SNP를 선택한다.

[0759] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 점수를 사용하여 시스템은, DCNN의 앙상블에 의해 채점된 각각의 SNP에 대한 최대 평균값을 취함으로써 현재 사이클의 이상점 SNP를 선택한다.

[0760] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 현재 사이클 동안 교체 없는 샘플링에 의해,

샘플링은 현재 사이클 동안 의사-병원성 SNP의 분리된 개별 트레이닝 세트를 초래한다.

[0761] 시스템은 종료 조건에 도달할 때까지 사이클을 계속한다. 종료 조건은 미리 결정된 횟수의 사이클일 수 있다. 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 소정의 사이클 수는 21이다.

[0762] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에서, 종결 조건은 엘리트 예측 병원성 세트 크기가 양성 세트 크기의 미리 결정된 확산 내에 있을 때이다.

[0763] 분류자 파라미터는 적어도 컨볼루션 필터 가중치 및 학습률일 수 있다.

[0764] 시스템은, 앙상블 내의 SNP 병원성 분류자들 중 하나를 SNP 병원성 분류기로서 선택할 수 있다. 선택된 SNP 병원성 분류자는, 최종 사이클에서 평가된 유효성 확인 샘플 상의 앙상블에서 다른 SNP 병원성 분류자를 예상하지 못한 것일 수 있다.

[0765] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 사이클에 걸쳐 완료된 공통 엘리트 세트는 400000개 이상의 엘리트 예측 병원성 SNP를 가질 수 있다.

[0766] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 각 사이클에서, 시스템은, 양성 SNP와 샘플링된 의사-병원성 SNP 간의 트라이뉴클레오타이드 컨텍스트를 일치시켜 엘리트 예측 병원성 SNP에서의 돌연변이율 편향을 방지할 수 있다.

[0767] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 합성 세트로부터의 의사-병원성 SNP의 샘플링은 각각의 연속 사이클에서 5%만큼 감소될 수 있다.

[0768] 도 37, 도 38, 도 39, 도 40, 도 41 및 도 42에 도시된 바와 같이, 시스템은, 트레이닝을 위해 현재 사이클에서 샘플링된 의사-병원성 SNP, 엘리트 예측 병원성 SNP, 및 트레이닝을 위해 현재 사이클에서 사용되는 양성 SNP에 의해, 합성 SNP를 필터링할 수 있다.

[0769] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0770] 다른 구현예는, 상술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는, 메모리 및 메모리에 저장된 명령을 실행하여 상술한 시스템의 동작을 수행하도록 동작가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0771] 개시된 기술의 또 다른 구현예는, 도 36에 도시된 바와 같이 컨볼루션 신경망(약칭 CNN) 기반 반감독 학습자를 포함한다.

[0772] 도 36에 도시된 바와 같이, 반감독 학습자는, 양성 트레이닝 세트 및 병원성 트레이닝 세트에 대해 반복적으로 트레이닝되는, 메모리에 연결된 다수의 프로세서 상에서 실행되는 CNN의 앙상블을 포함할 수 있다.

[0773] 도 36에 도시된 바와 같이, 반감독 학습자는, 트레이닝된 앙상블의 합성 세트에 대한 평가에 기초하여 병원성 트레이닝 세트의 세트 크기를 점진적으로 증대시키는 프로세서들 중 적어도 하나에서 실행되는 세트 증강기를 포함할 수 있다.

[0774] 각각의 반복에서, 평가는 세트 증강기에 의해 병원성 트레이닝 세트에 추가된 엘리트 예측 병원성 세트를 생성한다.

[0775] 반감독 학습자는, 단일 뉴클레오타이드 다형성(약어 SNP) 병원성 분류자를 구성하고 트레이닝하도록 CNN, 증강 병원성 트레이닝 세트 및 양성 트레이닝 세트 중 적어도 하나를 사용하는 빌더(builder)를 포함할 수 있다.

[0776] 제1 시스템 구현을 위한 이러한 특정 구현 부문에서 논의된 각각의 피처는 이 시스템 구현에 동일하게 적용된다. 전술한 바와 같이, 모든 시스템 피처는 여기서 반복되지는 않으며 참조로 반복되는 것으로 간주되어야 한다.

[0777] 다른 구현예는, 전술한 시스템의 동작을 수행하도록 프로세서에 의해 실행가능한 명령어를 저장하는 비일시적 컴퓨터 판독가능 저장 매체를 포함할 수 있다. 또 다른 구현예는, 메모리 및 메모리에 저장된 명령을 실행하여 상술된 시스템의 동작을 수행하도록 동작 가능한 하나 이상의 프로세서를 포함하는 시스템을 포함할 수 있다.

[0778] 전술한 기술은 개시된 기술의 제조 및 이용을 가능하게 하기 위해 제공된다. 개시된 구현예들에 대한 다양한 수

정이 명백 할 것이고, 본 명세서에 정의된 일반적인 원리들은 개시된 기술의 사상 및 범위를 벗어나지 않고 다른 구현예 및 응용분야에 적용될 수 있다. 따라서, 개시된 기술은, 도시된 구현으로 제한하고자 하는 것이 아니라, 본 명세서에 개시된 원리 및 특징과 일치하는 가장 넓은 범위에 따라야 한다. 개시된 기술의 범위는 첨부된 청구범위에 의해 정의된다.

[0779] **컴퓨터 시스템**

[0780] 도 66은 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.

[0781] 일 구현예에서, 양성 데이터세트 생성기, 변이체 병원성 분류자, 이차 구조 분류자, 용매 접근성 분류자, 및 반감독 학습자와 같은 신경망들은 저장 서브시스템 및 사용자 인터페이스 입력 장치에 통신가능하게 연결된다.

[0782] 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.

[0783] 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.

[0784] 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.

[0785] 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.

[0786] 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.

[0787] 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 변화함으로 인해, 도 66에 도시된 컴퓨터 시스템의 설명은, 개시된 기술을 예시하기 위한 특정한 일례를 의도한 것일 뿐이다. 도 66에 도시된 컴퓨터 시스템보다 많거나 적은 구성요소를 갖는 컴퓨터 시스템의 다른 많은 구성이 가능하다.

[0788] 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.

[0789] **부록**

[0790] 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. Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F. McRae, Yanjun Li, Jack A. Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, Jinbo Xu, Serafim Batzloglou, Xiaolin Li & Kyle Kai-How Farh. Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics* volume 50, pages1161–1170 (2018). Accessible at https://www.nature.com/articles/s41588-018-0167-z.
2. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469-476, doi:10.1038/nature13127 (2014).
3. Rehm, H. L., J. S. Berg, L. D. Brooks, C. D. Bustamante, J. P. Evans, M. J. Landrum, D. H. Ledbetter, D. R. Maglott, C. L. Martin, R. L. Nussbaum, S. E. Plon, E. M. Ramos, S. T. Sherry, M. S. Watson. ClinGen--the Clinical Genome Resource. *N. Engl. J. Med.* 372, 2235-2242 (2015).
4. Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 12, 745–755 (2011).
5. Rehm, H. L. Evolving health care through personal genomics. *Nature Reviews Genetics* 18, 259–267 (2017).
6. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17, 405-424, doi:10.1038/gim.2015.30 (2015).
7. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291, doi:10.1038/nature19057 (2016).
8. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201-206, doi:10.1038/nature18964 (2016).
9. Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* 526, 68-74, doi:10.1038/nature15393 (2015).
10. Liu, X., X. Jian, E. Boerwinkle. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* 32, 894–899 (2011).
11. Chimpanzee Sequencing Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69-87, doi:10.1038/nature04072 (2005).
12. Takahata, N. Allelic genealogy and human evolution. *Mol Biol Evol* 10, 2-22 (1993).
13. Asthana, S., Schmidt, S. & Sunyaev, S. A limited role for balancing selection. *Trends Genet* 21, 30-32, doi:10.1016/j.tig.2004.11.001 (2005).

[0791]

14. Leffler, E. M., Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J.D. Wall, G. Sella, P. Donnelly. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339, 1578-1582 (2013).
15. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat Genet* 46, 944-950, doi:10.1038/ng.3050 (2014).
16. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96-98 (1973).
17. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet* 17, 502-510 (2001).
18. Whiffin, N., E. Minikel, R. Walsh, A. H. O'Donnell-Luria, K. Karczewski, A. Y. Ing, P. J. Barton, B. Funke, S. A. Cook, D. MacArthur, J. S. Ware. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine* 19, 1151–1158 (2017).
19. Prado-Martinez, J. *et al.* Great ape genome diversity and population history. *Nature* 499, 471-475 (2013).
20. Klein, J., Satta, Y., O'HUigin, C. & Takahata, N. The molecular descent of the major histocompatibility complex. *Annu Rev Immunol* 11, 269-295, doi:10.1146/annurev.iy.11.040193.001413 (1993).
21. Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).
22. de Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354, 477-481, doi:10.1126/science.aag2602 (2016).
23. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* 469, 529-533 (2011).
24. Rhesus Macaque Genome Sequencing Analysis Consortium *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316, 222-234, doi:10.1126/science.1139247 (2007).
25. Worley, K. C., W. C. Warren, J. Rogers, D. Locke, D. M. Muzny, E. R. Mardis, G. M. Weinstock, S. D. Tardif, K. M. Aagaard, N. Archidiacono, N. A. Rayan. The common marmoset genome provides insight into primate biology and evolution. *Nature Genetics* 46, 850-857 (2014).
26. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29, 308-311 (2001).
27. Schrago, C. G. & Russo, C. A. Timing the origin of New World monkeys. *Mol Biol Evol* 20, 1620-1625, doi:10.1093/molbev/msg172 (2003).
28. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44, D862-868, doi:10.1093/nar/gkv1222 (2016).
29. Brandon, E. P., Idzerda, R. L. & McKnight, G. S. Targeting the mouse genome: a compendium of knockouts (Part II). *Curr Biol* 5, 758-765 (1995).
30. Lieschke, J. G., P. D. Currie. Animal models of human disease: zebrafish swim into view. *Nature Reviews Genetics* 8, 353-367 (2007).

[0792]

31. Sittig, L. J., P. Carbonetto, K. A. Engel, K. S. Krauss, C. M. Barrios-Camacho, A. A. Palmer. Genetic background limits generalizability of genotype-phenotype relationships. *Neuron* 91, 1253-1259 (2016).
32. Bazykin, G. A. *et al.* Extensive parallelism in protein evolution. *Biol Direct* 2, 20, doi:10.1186/1745-6150-2-20 (2007).
33. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* 11, 863-874, doi:10.1101/gr.176601 (2001).
34. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 7, 248-249, doi:10.1038/nmeth0410-248 (2010).
35. Chun, S., J. C. Fay. Identification of deleterious mutations within three human genomes. *Genome research* 19, 1553-1561 (2009).
36. Schwarz, J. M., C. Rödelsperger, M. Schuelke, D. Seelow. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 7, 575–576 (2010).
37. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 39, e118, doi:10.1093/nar/gkr407 (2011).
38. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 24, 2125-2137, doi:10.1093/hmg/ddu733 (2015).
39. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* 14 Suppl 3, S3, doi:10.1186/1471-2164-14-S3-S3 (2013).
40. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 7, e46688, doi:10.1371/journal.pone.0046688 (2012).
41. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 47, 276-283, doi:10.1038/ng.3196 (2015).
42. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536-1543, doi:10.1093/bioinformatics/btv009 (2015).
43. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761-763, doi:10.1093/bioinformatics/btu703 (2015).
44. Bell, C. J., D. L. Dinwiddie, N. A. Miller, S. L. Hateley, E. E. Ganusova, J. Midge, R. J. Langley, L. Zhang, C. L. Lee, R. D. Schilkey, J. E. Woodward, H. E. Peckham, G. P. Schroth, R. W. Kim, S. F. Kingsmore. Comprehensive carrier testing for severe childhood recessive diseases by next generation sequencing. *Sci. Transl. Med.* 3, 65ra64 (2011).

[0793]

45. Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310-315 (2014).
46. Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am J Hum Genet* 99, 595-606, doi:10.1016/j.ajhg.2016.07.005 (2016).
47. Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 99, 877-885, doi:10.1016/j.ajhg.2016.08.016 (2016).
48. Jagadeesh, K. A., A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, G. Bejerano. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics* 48, 1581-1586 (2016).
49. Grimm, D. G. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human mutation* 36, 513-523 (2015).
50. He, K., X. Zhang, S. Ren, J. Sun. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.
51. Heffernan, R. *et al.* Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 5, 11476, doi:10.1038/srep11476 (2015).
52. Wang, S., J. Peng, J. Ma, J. Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports* 6, 18962-18962 (2016).
53. Harpak, A., A. Bhaskar, J. K. Pritchard. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. *PLoS Genetics* 12 (2016).
54. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* 475, 353-358 (2011).
55. Shen, H. *et al.* Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution. *Science* 355, eaal4326, doi:10.1126/science.aal4326 (2017).
56. Nakamura, K. *et al.* Clinical spectrum of SCN2A mutations expanding to Ohtahara syndrome. *Neurology* 81, 992-998, doi:10.1212/WNL.0b013e3182a43e57 (2013).
57. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89, 10915-10919 (1992).
58. Li, W. H., C. I. Wu, C. C. Luo. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution* 21, 58-71 (1984).
59. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 185, 862-864 (1974).

[0794]

60. LeCun, Y., L. Bottou, Y. Bengio, P. Haffner. in *Proceedings of the IEEE* 2278-2324.
61. Vissers, L. E., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* 17, 9-18, doi:10.1038/nrg3999 (2016).
62. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* 485, 242-245, doi:10.1038/nature11011 (2012).
63. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237-241, doi:10.1038/nature10945 (2012).
64. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209-215, doi:10.1038/nature13772 (2014).
65. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519, 223-228, doi:10.1038/nature14135 (2015).
66. Deciphering Developmental Disorders Study. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* 542, 433-438, doi:10.1038/nature21062 (2017).
67. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* 515, 216-221, doi:10.1038/nature13908 (2014).
68. Zhu, X., Need, A. C., Petrovski, S. & Goldstein, D. B. One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. *Nat Neurosci* 17, 773-781, doi:10.1038/nn.3713 (2014).
69. Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, M. Przeworski. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS biology* 10, e1001388 (2012).
70. Estrada, A. *et al.* Impending extinction crisis of the world's primates: Why primates matter. *Science advances* 3, e1600946 (2017).
71. Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A.M. Zahler, D. Haussler. The human genome browser at UCSC. *Genome Res.* 12, 996-1006 (2002).
72. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res* 45, D626-D634, doi:10.1093/nar/gkw1134 (2017).
73. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-2637, doi:10.1002/bip.360221211 (1983).
74. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* 39, D411-419, doi:10.1093/nar/gkq1105 (2011).
75. He, K., Zhang, X., Ren, S. & Sun, J. in *European Conference on Computer Vision*. 630-645 (Springer).
76. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48, 214-220, doi:10.1038/ng.3477 (2016).

[0795]

77. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25, 2744-2750, doi:10.1093/bioinformatics/btp528 (2009).
78. Lu, Q. *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 5, 10576, doi:10.1038/srep10576 (2015).
79. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34, 57-65, doi:10.1002/humu.22225 (2013).
80. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).
81. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* 37, 235-241, doi:10.1002/humu.22932 (2016).
82. Jain, S., White, M. & Radivojac, P. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2066-2072.
83. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367, 1921-1929, doi:10.1056/NEJMoa1206524 (2012).
84. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285-299, doi:10.1016/j.neuron.2012.04.009 (2012).
85. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* 485, 246-250, doi:10.1038/nature10989 (2012).
86. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674-1682, doi:10.1016/S0140-6736(12)61480-9 (2012).
87. Epi, K. C. *et al.* De novo mutations in epileptic encephalopathies. *Nature* 501, 217-221, doi:10.1038/nature12439 (2013).
88. Euro, E.-R. E. S. C., Epilepsy Phenome/Genome, P. & Epi, K. C. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* 95, 360-370, doi:10.1016/j.ajhg.2014.08.013 (2014).
89. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511, 344-347, doi:10.1038/nature13394 (2014).
90. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* 19, 1194-1196, doi:10.1038/nn.4352 (2016).
76. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48, 214-220, doi:10.1038/ng.3477 (2016).

[0796]

77. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25, 2744-2750, doi:10.1093/bioinformatics/btp528 (2009).
78. Lu, Q. *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 5, 10576, doi:10.1038/srep10576 (2015).
79. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 34, 57-65, doi:10.1002/humu.22225 (2013).
80. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).
81. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* 37, 235-241, doi:10.1002/humu.22932 (2016).
82. Jain, S., White, M. & Radivojac, P. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2066-2072.
83. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367, 1921-1929, doi:10.1056/NEJMoa1206524 (2012).
84. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285-299, doi:10.1016/j.neuron.2012.04.009 (2012).
85. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* 485, 246-250, doi:10.1038/nature10989 (2012).
86. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674-1682, doi:10.1016/S0140-6736(12)61480-9 (2012).
87. Epi, K. C. *et al.* De novo mutations in epileptic encephalopathies. *Nature* 501, 217-221, doi:10.1038/nature12439 (2013).
88. Euro, E.-R. E. S. C., Epilepsy Phenome/Genome, P. & Epi, K. C. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am J Hum Genet* 95, 360-370, doi:10.1016/j.ajhg.2014.08.013 (2014).
89. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511, 344-347, doi:10.1038/nature13394 (2014).
90. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* 19, 1194-1196, doi:10.1038/nn.4352 (2016).

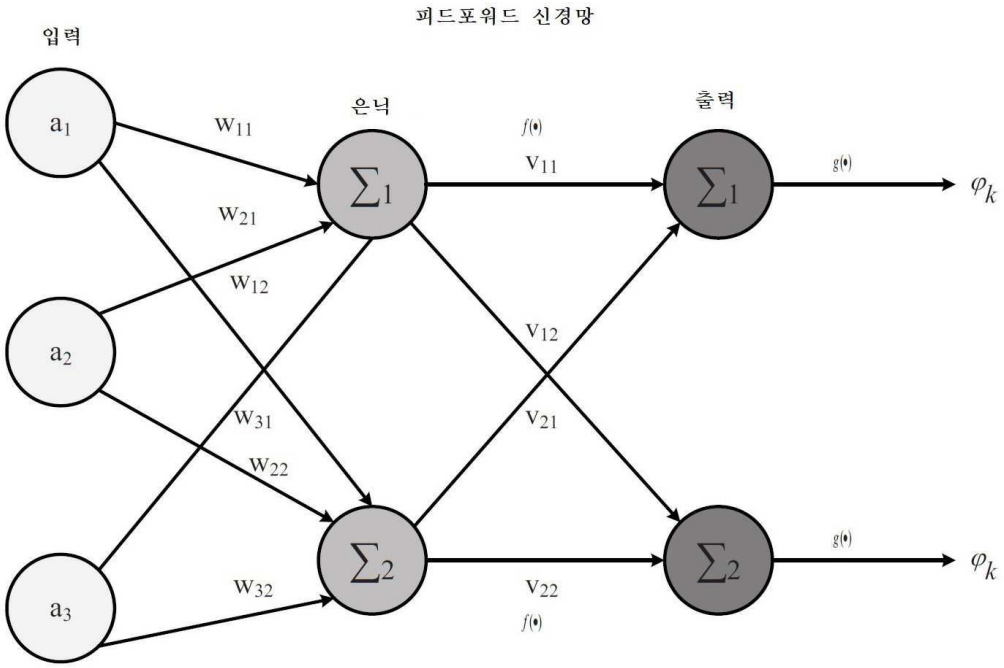
[0797]

91. Famiglietti, M. L. *et al.* Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum Mutat* 35, 927-935, doi:10.1002/humu.22594 (2014).
92. Horaitis, O., Talbot, C. C., Jr., Phommarinh, M., Phillips, K. M. & Cotton, R. G. A database of locus-specific databases. *Nat Genet* 39, 425, doi:10.1038/ng0407-425 (2007).
93. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133, 1-9, doi:10.1007/s00439-013-1358-4 (2014).

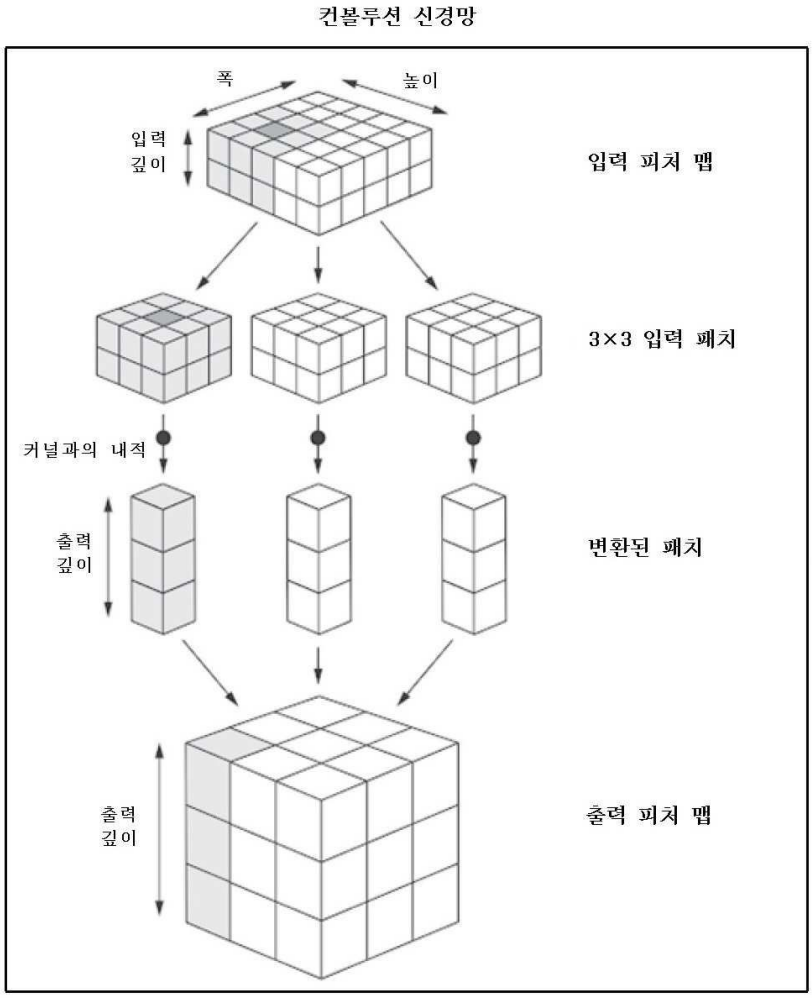
[0798]

도면

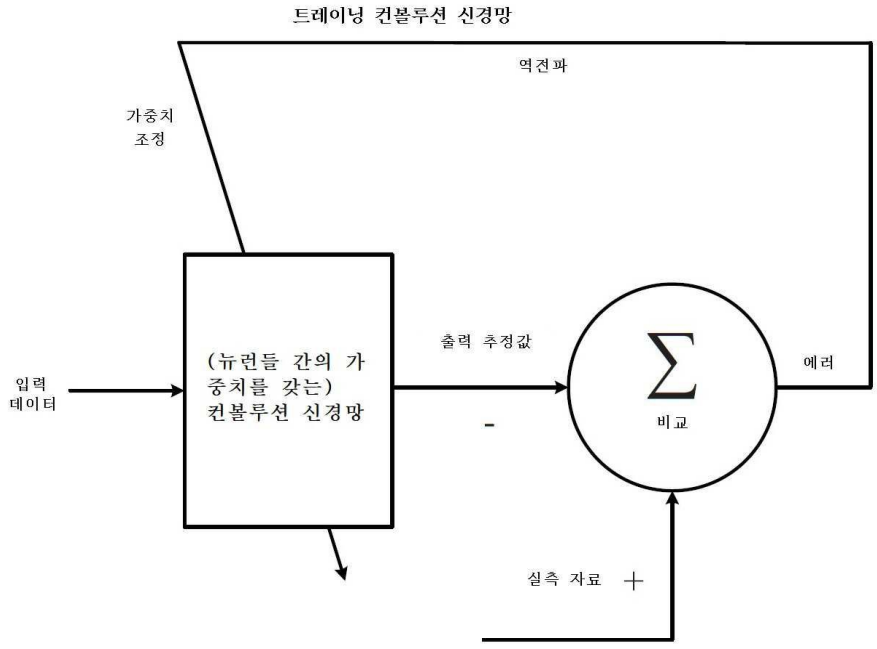
도면1a


피드포워드 신경망
입력
은닉
출력
a_1
a_2
a_3
W_{11}
W_{21}
W_{12}
W_{31}
W_{22}
W_{32}
$\sum 1$
$\sum 2$
$f(\bullet)$
V_{11}
V_{12}
V_{21}
V_{22}
$f(\bullet)$
$g(\bullet)$
$g(\bullet)$
φ_k
φ_k

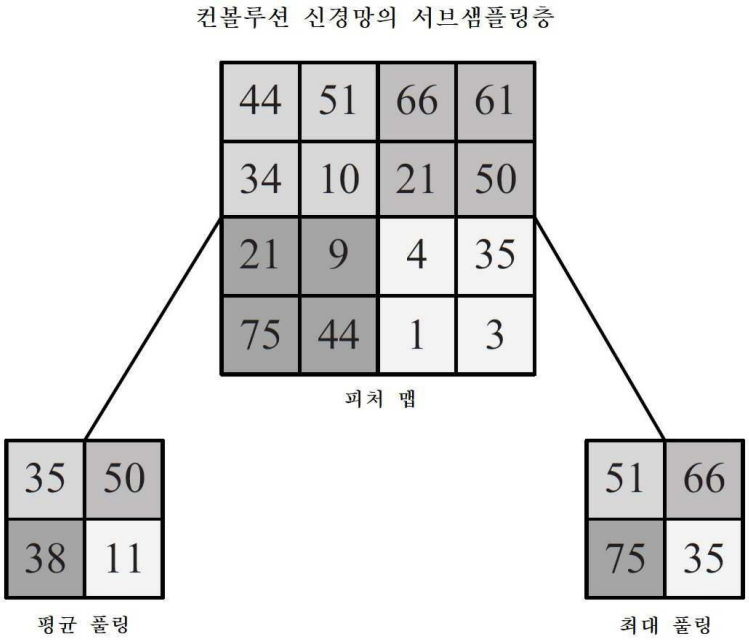
도면1b


컨볼루션 신경망
폭
높이
입력
깊이
입력 피처 맵
3×3 입력 패치
커널과의 내적
출력
깊이
변환된 패치
출력
깊이
출력 피처 맵

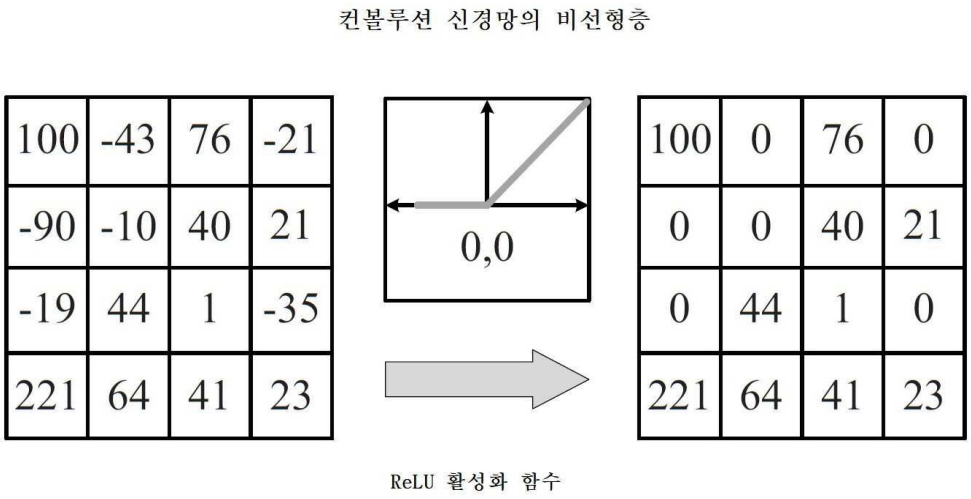
도면1c


트레이닝 컨볼루션 신경망
역전파
가중치
조정
입력
데이터
(뉴런들 간의 가
중치를 갖는)
컨볼루션 신경망
출력 추정값
-
∑
비교
에러
실측 자료 +

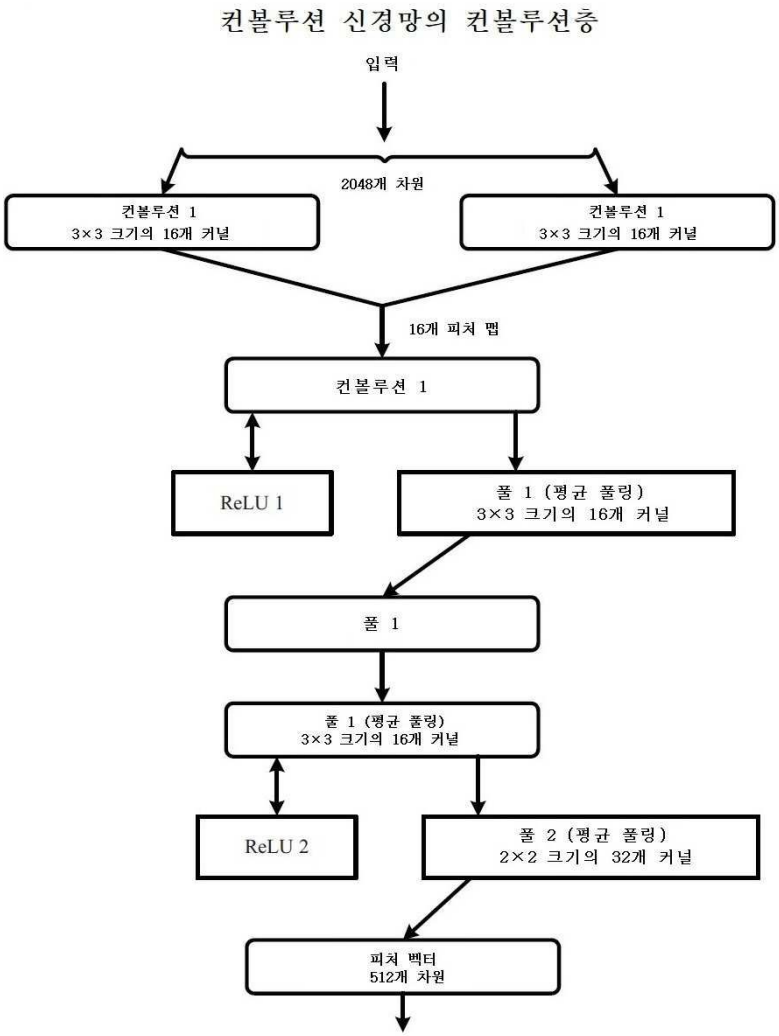
도면1d


컨볼루션 신경망의 서브샘플링층
44 51 66 61
34 10 21 50
21 9 4 35
75 44 1 3
피처 맵
35 50
38 11
평균 풀링
51 66
75 35
최대 풀링

도면1e

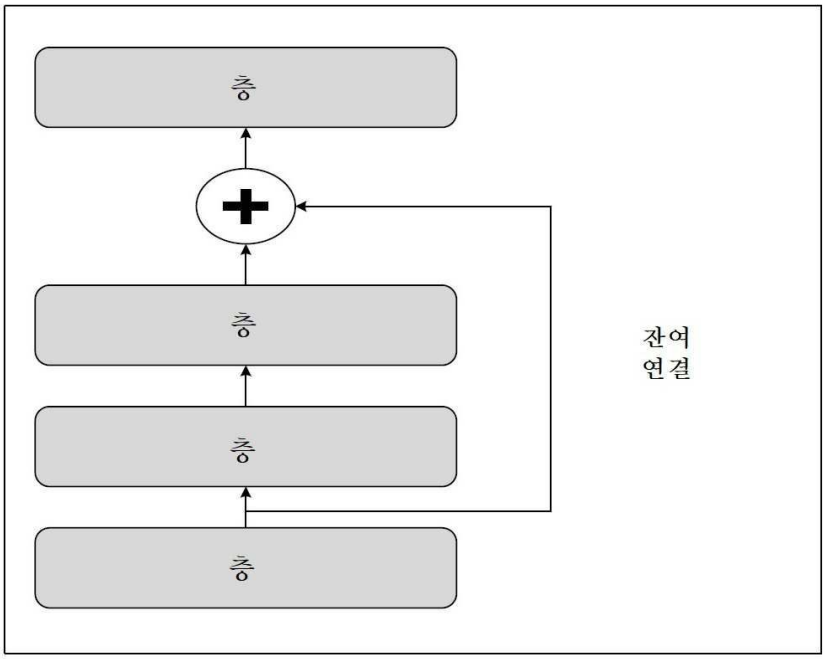

컨볼루션 신경망의 비선형층
100 -43 76 -21
-90 -10 40 21
-19 44 1 -35
221 64 41 23
0,0
100 0 76 0
0 0 40 21
0 44 1 0
221 64 41 23
ReLU 활성화 함수

도면1f

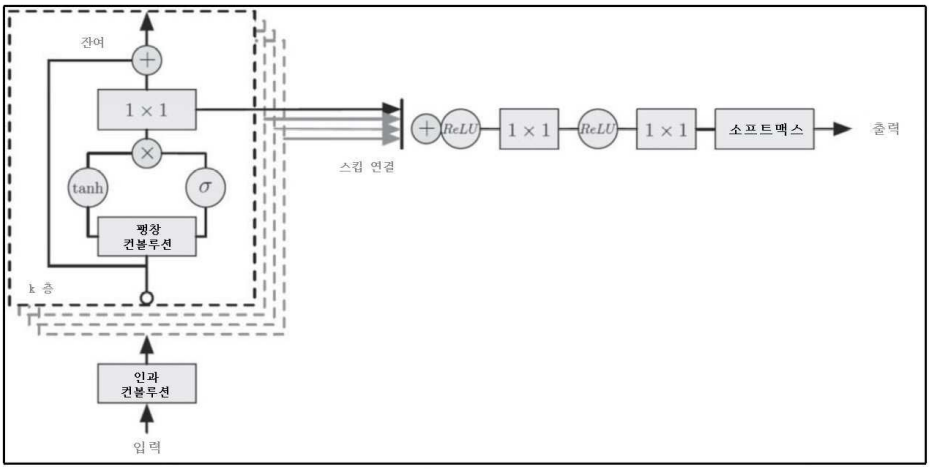

컨볼루션 신경망의 컨볼루션층
입력
2048개 차원
컨볼루션 1
3×3 크기의 16개 커널
컨볼루션 1
3×3 크기의 16개 커널
16개 피처 맵
컨볼루션 1
ReLU 1
풀 1 (평균 풀링)
3×3 크기의 16개 커널
풀 1
풀 1 (평균 풀링)
3×3 크기의 16개 커널
ReLU 2
풀 2 (평균 풀링)
2×2 크기의 32개 커널
피처 벡터
512개 차원

도면1g

컨볼루션 신경망에서 사용되는 잔여 연결



도면1h



컨볼루션 신경망에서 사용되는 잔여 블록과 스킵 연결

도면1i

컨볼루션 신경망을 이용한 일괄 정규화 순방향 패스

$$\mu_{\mathcal{B}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^{(\ell-1)}$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i^{(\ell-1)} - \mu_{\mathcal{B}})^2$$

$$\hat{\boldsymbol{x}}^{(\ell-1)} = \frac{\boldsymbol{x}^{(\ell-1)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$\boldsymbol{x}^{(\ell)} = \boldsymbol{\gamma}^{(\ell)} \hat{\boldsymbol{x}}^{(\ell-1)} + \boldsymbol{\beta}^{(\ell)}$$

도면1j

일괄 정규화 - 컨볼루션 신경망과의 간섭

$$\hat{\boldsymbol{x}}^{(\ell-1)} = \frac{\boldsymbol{x}^{(\ell-1)} - \mu_{\mathcal{D}}}{\sqrt{\sigma_{\mathcal{D}}^2 + \epsilon}}$$

$$\boldsymbol{x}_i^{(\ell)} = \boldsymbol{\gamma}^{(\ell)} \hat{\boldsymbol{x}}_i^{(\ell-1)} + \boldsymbol{\beta}^{(\ell)}$$

도면1k

컨볼루션 신경망을 이용한 일괄 정규화 역방향 패스

$$\nabla_{\boldsymbol{\gamma}^{(\ell)}} \mathcal{L} = \sum_{i=1}^{n} \left(\nabla_{\boldsymbol{x}^{(\ell+1)}} \mathcal{L}\right)_i \cdot \hat{\boldsymbol{x}}_i^{(\ell)}$$

$$\nabla_{\boldsymbol{\beta}^{(\ell)}} \mathcal{L} = \sum_{i=1}^{n} \left(\nabla_{\boldsymbol{x}^{(\ell+1)}} \mathcal{L}\right)_i$$

도면11

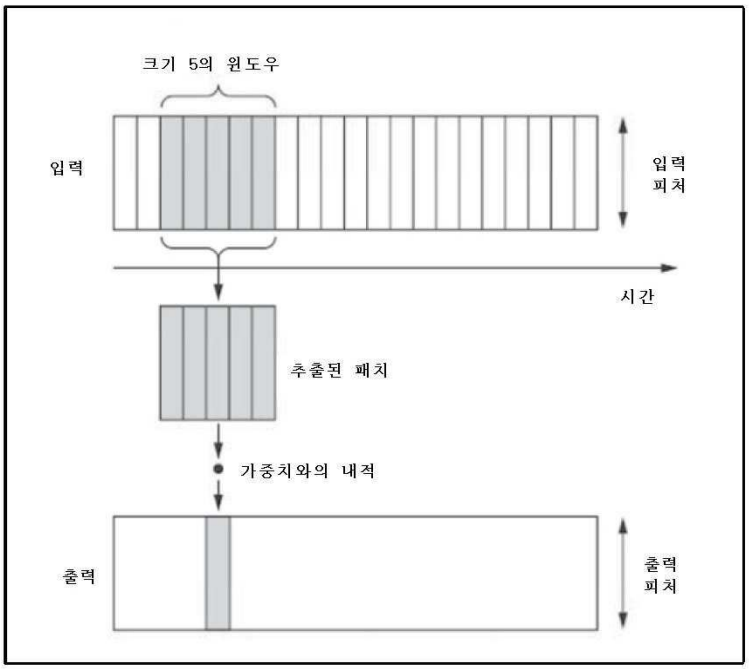
컨볼루션층의 일괄 정규화

```
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())
```

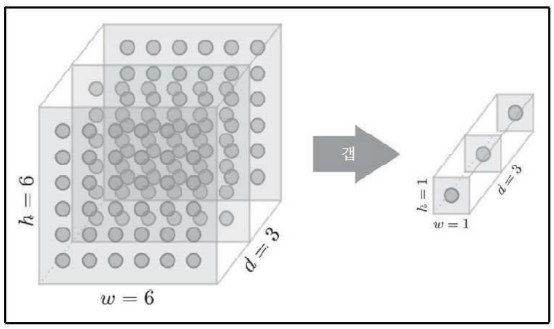
도면1m

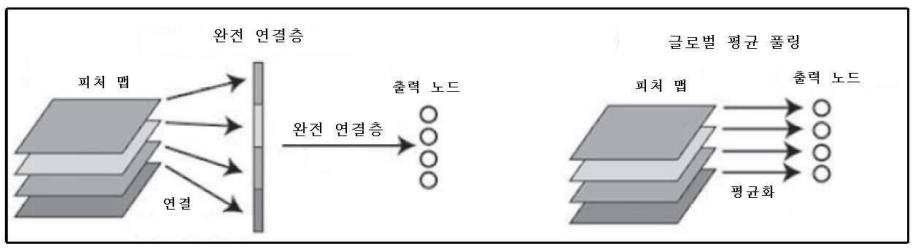
컨볼루션 신경망에서 사용되는 1D 컨볼루션



도면1n

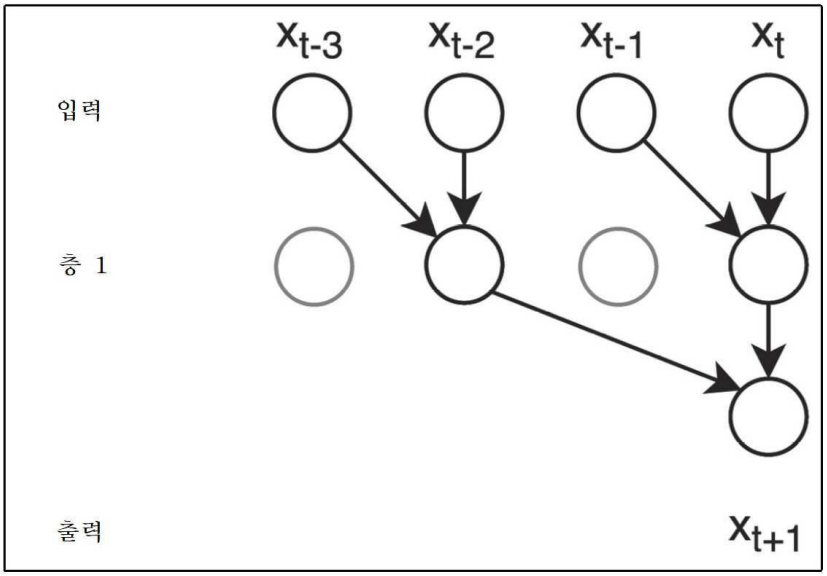
컨볼루션 신경망의 글로벌 평균 풀링(GAP)





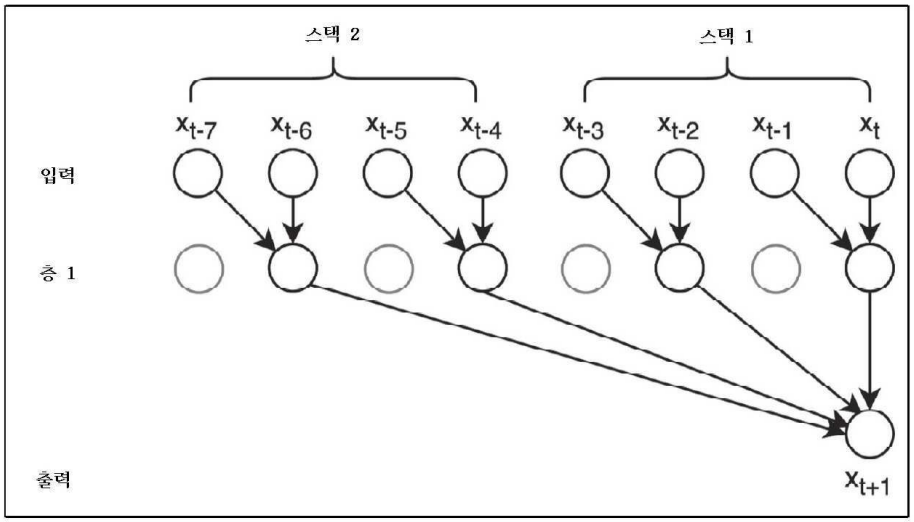
도면1o

팽창 컨볼루션

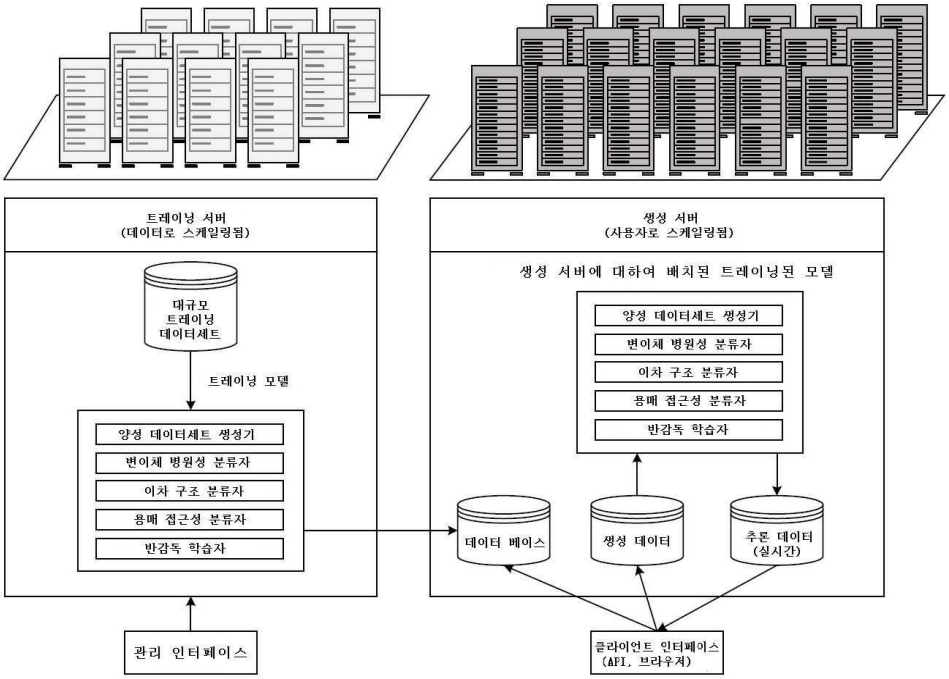


도면1p

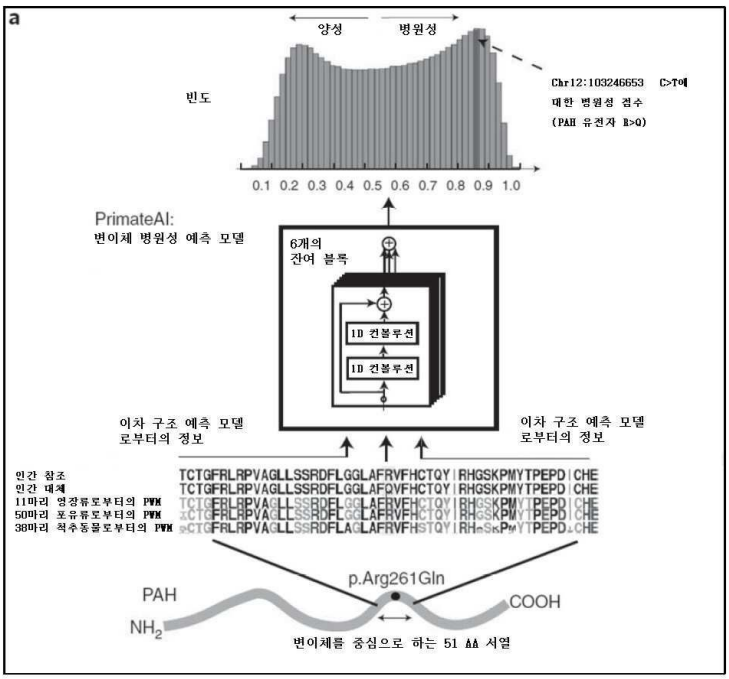
적층된 팽창 컨볼루션



도면1q

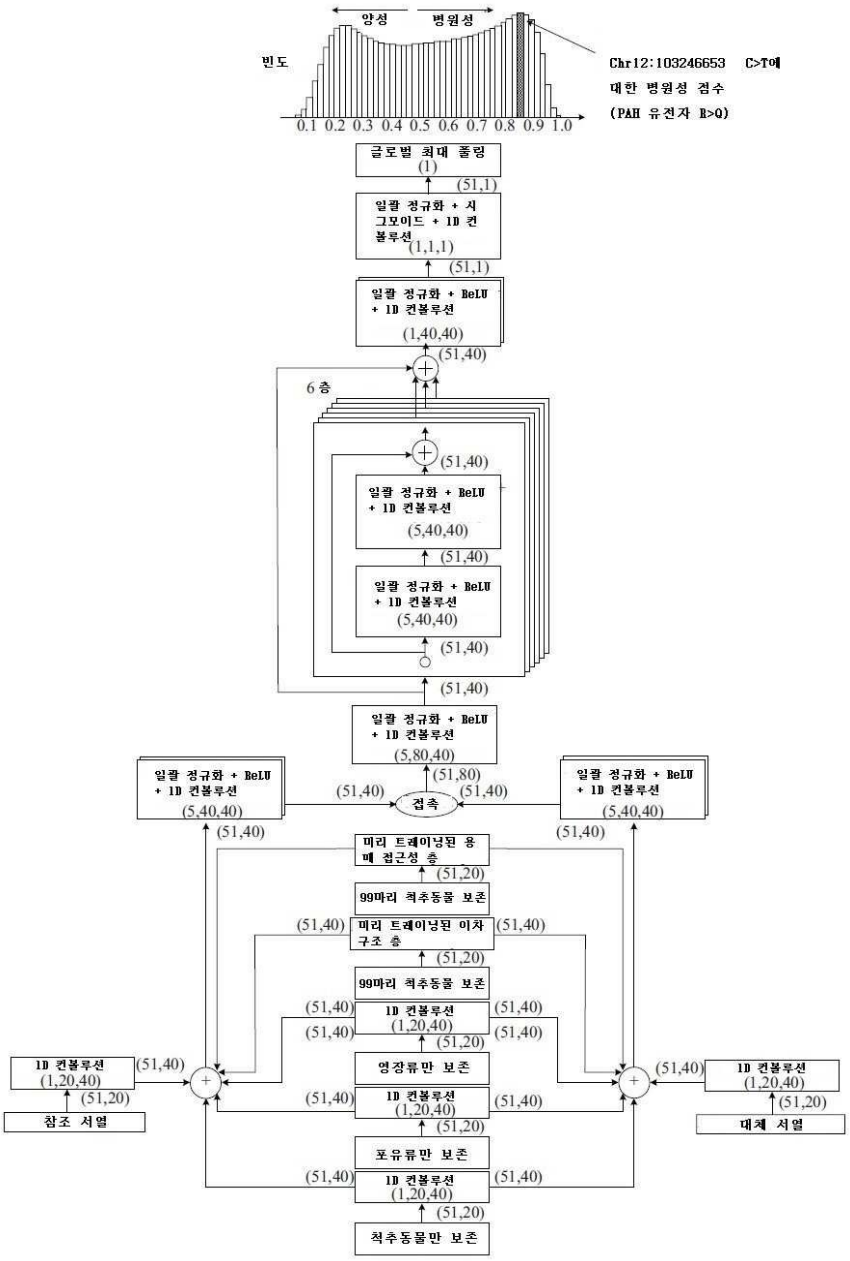


도면2



PAH – 페닐알라닌 수산화효소
NH_2 – 아미노 말단
COOH – 카복실 말단

도면3


양성
병원성
빈도
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Chr12:103246653 C>T에
대한 병원성 점수
(PAH 유전자 R>Q)
글로벌 최대 폴링
(1)
(51,1)
일괄 정규화 + 시
그모이드 + 1D 컨
볼루션
(1,1,1)
(51,1)
일괄 정규화 + ReLU
+ 1D 컨볼루션
(1,40,40)
(51,40)
6 층
(51,40)
일괄 정규화 + ReLU
+ 1D 컨볼루션
(5,40,40)
(51,40)
일괄 정규화 + ReLU
+ 1D 컨볼루션
(5,40,40)
(51,40)
(51,40)
일괄 정규화 + ReLU
+ 1D 컨볼루션
(5,80,40)
(51,80)
일괄 정규화 + ReLU
+ 1D 컨볼루션
(5,40,40)
(51,40)
(51,40)
(51,40)
접속
(51,40)
(51,40)
미리 트레이닝된 용
매 접근성 층
(51,20)
99마리 척추동물 보존
(51,40)
(51,40)
미리 트레이닝된 이차
구조 층
(51,20)
99마리 척추동물 보존
(51,40)
(51,40)
1D 컨볼루션
(1,20,40)
(51,40)
(51,40)
(51,20)
영장류만 보존
1D 컨볼루션
(1,20,40)
(51,40)
(51,40)
(51,20)
포유류만 보존
(51,40)
1D 컨볼루션
(1,20,40)
(51,40)
(51,20)
척추동물만 보존
1D 컨볼루션
(1,20,40)
(51,40)
(51,20)
참조 서열
(51,40)
1D 컨볼루션
(1,20,40)
(51,20)
대체 서열

도면4a

층	층에 대한 입력	유형	커널의 수 및 윈도우 크기	형상	활성화
층 1a	참조 서열	컨볼루션 1D	40,1	(51,40)	선형
층 1b	대체 서열	컨볼루션 1D	40,1	(51,40)	선형
층 1c	영장류 보존	컨볼루션 1D	40,1	(51,40)	선형
층 1d	포유류 보존	컨볼루션 1D	40,1	(51,40)	선형
층 1e	척추동물 보존	컨볼루션 1D	40,1	(51,40)	선형
이미 트레이닝 된 이차 구조 예측 모델 (트레이닝가능 망)	영장류 + 포유류 + 척추동물 보존	-	-	(51,40)	-
미리 트레이닝된 용매 접근성 모델 (트레이닝가능 망)	영장류 + 포유류 + 척추동물 보존	-	-	(51,40)	-
층 1f	이미 트레이닝된 구조 예측 모델 출력	컨볼루션 1D	40,1	(51,40)	선형
층 1g	미리 트레이닝된 용매 접근성 모델 출력	컨볼루션 1D	40,1	(51,40)	선형
병합	층 1a, 층 1c, 층 1d, 층 1e, 층 1f, 층 1g	추가	-	(51,40)	-
병합	층 1b, 층 1c, 층 1d, 층 1e, 층 1f, 층 1g	추가	-	(51,40)	-
층 2a	병합 1a	일괄 정규화	-	(51,40)	-
층 3a	층 2a	활성화	-	(51,40)	Relu
층 4a	층 3a	컨볼루션 1D	40,5	(51,40)	선형
층 5a	층 4a	일괄 정규화	-	(51,40)	-
층 6a	층 5a	활성화	-	(51,40)	Relu
층 7a	층 6a	컨볼루션 1D	40,5	(51,40)	선형
층 2b	병합 1b	일괄 정규화	-	(51,40)	-
층 3b	층 2b	활성화	-	(51,40)	Relu
층 4b	층 3b	컨볼루션 1D	40,5	(51,40)	선형

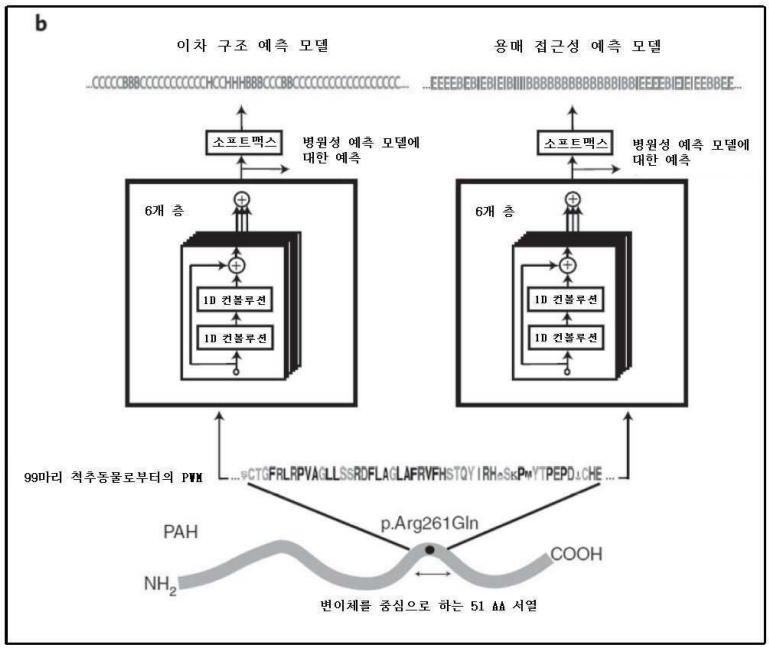
도면4b

층 5b	층 4b	일괄 정규화	-	(51,40)	-
층 6b	층 5b	활성화	-	(51,40)	Relu
층 7b	층 6b	컨볼루션 1D	40,5	(51,40)	선형
병합 2a	층 7a, 층 7b	연쇄화	-	(51,80)	-
병합 2b	층 7a, 층 7b	연쇄화	-	(51,80)	-
층 8a	병합 2a	컨볼루션 1D	40,5	(51,40)	선형
층 8b	병합 2b	컨볼루션 1D	40,5	(51,40)	선형
층 9	층 8a	일괄 정규화	-	(51,40)	-
층 10	층 9	활성화	-	(51,40)	Relu
층 11	층 10	컨볼루션 1D	40,5	(51,40)	선형
층 12	층 11	일괄 정규화	-	(51,40)	-
층 13	층 12	활성화	-	(51,40)	Relu
층 14	층 13	컨볼루션 1D	40,5	(51,40)	선형
병합 3	층 8a, 층 14	추가	-	(51,40)	-
층 15	병합 3	일괄 정규화	-	(51,40)	-
층 16	층 15	활성화	-	(51,40)	Relu
층 17	층 16	컨볼루션 1D	40,5	(51,40)	선형
층 18	층 17	일괄 정규화	-	(51,40)	-
층 19	층 18	활성화	-	(51,40)	Relu
층 20	층 19	컨볼루션 1D	40,5	(51,40)	선형
병합 4	병합 3, 층 20	추가	-	(51,40)	-
층 21	병합 4	컨볼루션 1D	40,1	(51,40)	선형
층 22	병합 4	일괄 정규화	-	(51,40)	-
층 23	층 22	활성화	-	(51,40)	Relu
층 24	층 23	컨볼루션 1D	40,5	(51,40)	선형
층 25	층 24	일괄 정규화	-	(51,40)	-
층 26	층 25	활성화	-	(51,40)	Relu
층 27	층 26	컨볼루션 1D	40,5	(51,40)	선형
병합 5	병합 4, 층 27	추가	-	(51,40)	-
층 28	병합 5	일괄 정규화	-	(51,40)	-
층 29	층 28	활성화	-	(51,40)	Relu
층 30	층 29	컨볼루션 1D	40,5	(51,40)	선형
층 31	층 30	일괄 정규화	-	(51,40)	-
층 32	층 31	활성화	-	(51,40)	Relu

도면4c

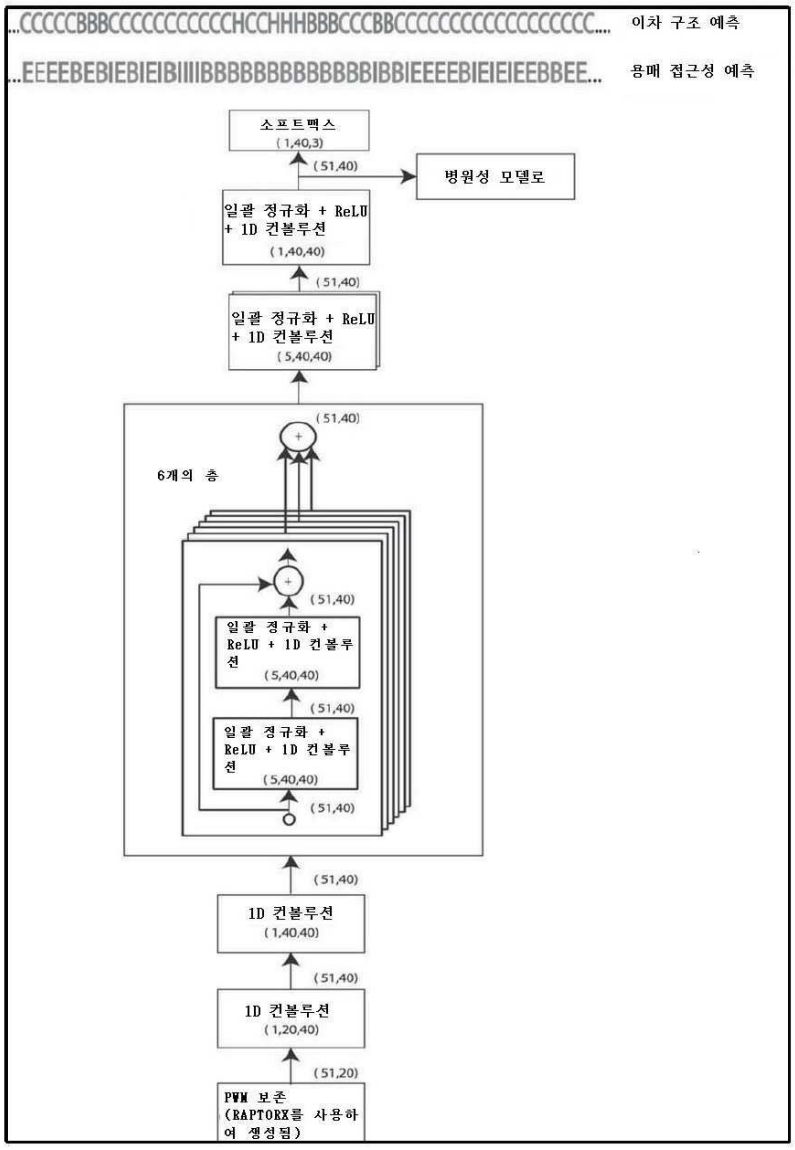
층 33	층 32	컨볼루션 1D	40,5	(51,40)	선형
병합6	병합5, 층 33	추가	-	(51,40)	-
층 34	병합6	컨볼루션 1D	40,1	(51,40)	선형
층 35	병합6	일괄 정규화	-	(51,40)	-
층 36	층 35	활성화	-	(51,40)	Relu
층 37	층 36	컨볼루션 1D	40,5	(51,40)	선형
층 38	층 37	일괄 정규화	-	(51,40)	-
층 39	층 38	활성화	-	(51,40)	Relu
층 40	층 39	컨볼루션 1D	40,5	(51,40)	선형
병합7	병합6, 층40	추가	-	(51,40)	-
층 41	병합7	일괄 정규화	-	(51,40)	-
층 42	층 41	활성화	-	(51,40)	Relu
층 43	층 42	컨볼루션 1D	40,5	(51,40)	선형
층 44	층 43	일괄 정규화	-	(51,40)	-
층 45	층 44	활성화	-	(51,40)	Relu
층 46	층 45	컨볼루션 1D	40,5	(51,40)	선형
병합8	병합7, 층46	추가	-	(51,40)	-
층 47	병합8	컨볼루션 1D	40,1	(51,40)	선형
병합9	층 8b, 층 21, 층 34, 층 47	추가	-	(51,40)	-
층 48	병합10	일괄 정규화	-	(51,40)	-
층 49	층 48	활성화	-	(51,40)	Relu
층 50	층 49	컨볼루션 1D	40,1	(51,40)	선형
층 51	층 50	일괄 정규화	-	(51,40)	-
층 52	층 51	활성화	-	(51,40)	Relu
층 53	층 52	컨볼루션 1D	40,1	(51,40)	선형
병합10	병합9, 층 53	추가	-	(51,40)	-
층 54	병합10	컨볼루션 1D	1,1	(51,1)	시그모이드
출력 층	층 54	글로벌 최대 풀링	-		1 -

도면5



PAH – 페닐알라닌 수산화효소
NH_2 – 아미노 말단
COOH – 카복실 말단

도면6


...CCCCCBBBCCCCCCCCCCCHCCHHHBBBCCCBBCCCCCCCCCCCCCCCCCC.... 이차 구조 예측
...EEEEBEBIEBIEIBIIIIBBBBBBBBBBBBBBIBBIEEEEBIEIEIEEBBEE... 용매 접근성 예측
소프트맥스
(1,40,3)
(51,40)
병원성 모델로
일괄 정규화 + ReLU + 1D 컨볼루션
(1,40,40)
(51,40)
일괄 정규화 + ReLU + 1D 컨볼루션
(5,40,40)
(51,40)
6개의 층
(51,40)
일괄 정규화 + ReLU + 1D 컨볼루션
(5,40,40)
(51,40)
일괄 정규화 + ReLU + 1D 컨볼루션
(5,40,40)
(51,40)
(51,40)
1D 컨볼루션
(1,40,40)
(51,40)
1D 컨볼루션
(1,20,40)
(51,20)
PWM 보존
(RAPTORX를 사용하여 생성됨)

도면7a

층	층에 대한 입력	유형	커널의 수 및 윈도우 크기	형상	활성화
층 1a	입력 PSFM	컨볼루션 1D	40,1	(51,40)	선형
층 2a	층 1a	컨볼루션 1D	40,1	(51,40)	선형
층 1b	입력PSFM	컨볼루션 1D	40,1	(51,40)	선형
층 2b	층 1b	컨볼루션 1D	40,1	(51,40)	선형
층 3	층 2a	일괄 정규화	-	(51,40)	-
층 4	층 3	활성화	-	(51,40)	Relu
층 5	층 4	컨볼루션 1D	40,5	(51,40)	선형
층 6	층 5	일괄 정규화	-	(51,40)	-
층 8	층 7	활성화	-	(51,40)	Relu
층 9	층 8	컨볼루션 1D	40,5	(51,40)	선형
병합1	층 2a, 층 9	추가	-	(51,40)	-
층 10	병합 1	일괄 정규화	-	(51,40)	-
층 11	층 10	활성화	-	(51,40)	Relu
층 12	층 11	컨볼루션 1D	40,5	(51,40)	선형
층 13	층 12	일괄 정규화	-	(51,40)	-
층 14	층 13	활성화	-	(51,40)	Relu
층 15	층 14	컨볼루션 1D	40,5	(51,40)	선형
병합2	병합1, 층 15	추가	-	(51,40)	-
층 16	병합2	컨볼루션 1D	40,1	(51,40)	선형
층 17	병합2	일괄 정규화	-	(51,40)	-
층 18	층 17	활성화	-	(51,40)	Relu
층 19	층 18	컨볼루션 1D	40,5	(51,40)	선형
층 20	층 19	일괄 정규화	-	(51,40)	-
층 21	층 20	활성화	-	(51,40)	Relu
층 22	층 21	컨볼루션 1D	40,5	(51,40)	선형
병합3	병합2, 층22	추가	-	(51,40)	-

도면7b

층 23	병합3	일괄 정규화	-	(51,40)	-
층 24	층 23	활성화	-	(51,40)	Relu
층 25	층 24	컨볼루션 1D	40,5	(51,40)	선형
층 26	층 25	일괄 정규화	-	(51,40)	-
층 27	층 26	활성화	-	(51,40)	Relu
층 28	층 27	컨볼루션 1D	40,5	(51,40)	선형
병합4	병합3 , 층 28	추가	-	(51,40)	-
층 29	병합4	컨볼루션 1D	40,1	(51,40)	선형
층 30	병합4	일괄 정규화	-	(51,40)	-
층 31	층 30	활성화	-	(51,40)	Relu
층 32	층 31	컨볼루션 1D	40,5	(51,40)	선형
층 33	층 32	일괄 정규화	-	(51,40)	-
층 34	층 33	활성화	-	(51,40)	Relu
층 35	층 34	컨볼루션 1D	40,5	(51,40)	선형
병합5	병합4, 층 35	추가	-	(51,40)	-
층 36	병합5	일괄 정규화	-	(51,40)	-
층 37	층 36	활성화	-	(51,40)	Relu
층 38	층 37	컨볼루션 1D	40,5	(51,40)	선형
층 39	층 38	일괄 정규화	-	(51,40)	-
층 40	층 39	활성화	-	(51,40)	Relu
층 41	층 40	컨볼루션 1D	40,5	(51,40)	선형
병합6	병합5, 층 41	추가	-	(51,40)	-
층 42	병합6	컨볼루션 1D	40,1	(51,40)	선형
병합7	층 2b, 층 16, 층 29, 층 42	추가	-	(51,40)	-
층 43	병합7	일괄 정규화	-	(51,40)	-
층 44	층 43	활성화	-	(51,40)	Relu
층 45	층 44	컨볼루션 1D	40,1	(51,40)	선형
층 46	층 45	일괄 정규화	-	(51,40)	-
층 47	층 46	활성화	-	(51,40)	Relu
층 48	층 47	컨볼루션 1D	40,1	(51,40)	선형
병합8	병합7, 층 48	추가	-	(51,40)	-
출력 층	병합8	컨볼루션 1D	3,1	(51,3)	소프트맥스

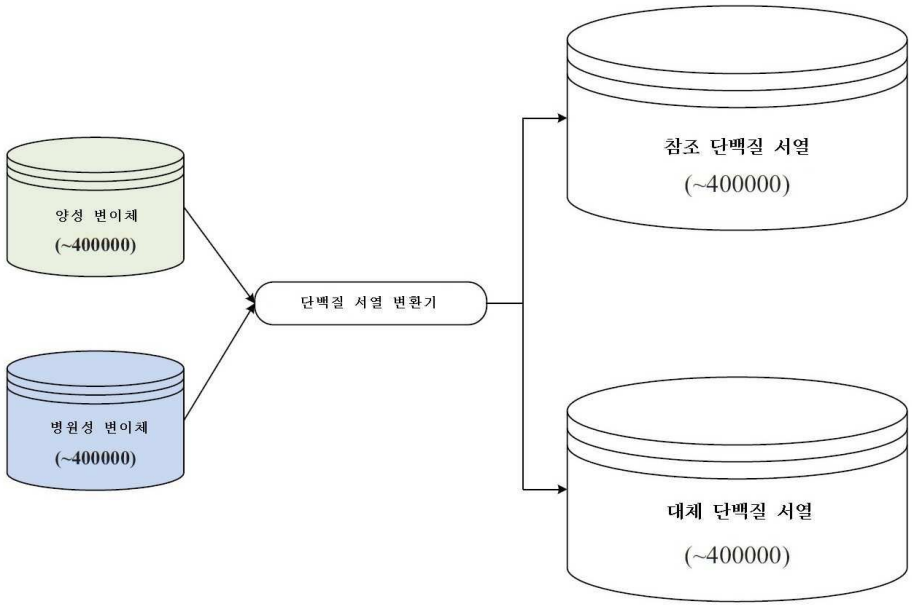
도면8a

층	층에 대한 입력	유형	커널의 수 및 윈도우 크기	형상	활성화
층 1a	입력 PSFM	컨볼루션 1D	40,1	(51,40)	선형
층 2a	층 1a	컨볼루션 1D	40,1	(51,40)	선형
층 1b	입력 PSFM	컨볼루션 1D	40,1	(51,40)	선형
층 2b	층 1b	컨볼루션 1D	40,1	(51,40)	선형
층 3	층 2a	일괄 정규화	-	(51,40)	-
층 4	층 3	활성화	-	(51,40)	Relu
층 5	층 4	컨볼루션 1D	40,5	(51,40)	선형
층 6	층 5	일괄 정규화	-	(51,40)	-
층 8	층 7	활성화	-	(51,40)	Relu
층 9	층 8	컨볼루션 1D	40,5	(51,40)	선형
병합1	층 2a, 층 9	추가	-	(51,40)	-
층 10	병합1	일괄 정규화	-	(51,40)	-
층 11	층 10	활성화	-	(51,40)	Relu
층 12	층 11	컨볼루션 1D	40,5	(51,40)	선형
층 13	층 12	일괄 정규화	-	(51,40)	-
층 14	층 13	활성화	-	(51,40)	Relu
층 15	층 14	컨볼루션 1D	40,5	(51,40)	선형
병합2	병합1, 층 15	추가	-	(51,40)	-
층 16	병합2	컨볼루션 1D	40,1	(51,40)	선형
층 17	병합2	일괄 정규화	-	(51,40)	-
층 18	층 17	활성화	-	(51,40)	Relu
층 19	층 18	컨볼루션 1D	40,5	(51,40)	선형
층 20	층 19	일괄 정규화	-	(51,40)	-
층 21	층 20	활성화	-	(51,40)	Relu
층 22	층 21	컨볼루션 1D	40,5	(51,40)	선형
병합3	병합2, 층 22	추가	-	(51,40)	-

도면8b

층 23	병합3	일괄 정규화	-	(51,40)	-
층 24	층 23	활성화	-	(51,40)	Relu
층 25	층 24	컨볼루션 1D	40,5	(51,40)	선형
층 26	층 25	일괄 정규화	-	(51,40)	-
층 27	층 26	활성화	-	(51,40)	Relu
층 28	층 27	컨볼루션 1D	40,5	(51,40)	선형
병합4	병합3, 층 28	추가	-	(51,40)	-
층 29	병합4	컨볼루션 1D	40,1	(51,40)	선형
층 30	병합4	일괄 정규화	-	(51,40)	-
층 31	층 30	활성화	-	(51,40)	Relu
층 32	층 31	컨볼루션 1D	40,5	(51,40)	선형
층 33	층 32	일괄 정규화	-	(51,40)	-
층 34	층 33	활성화	-	(51,40)	Relu
층 35	층 34	컨볼루션 1D	40,5	(51,40)	선형
병합5	병합4, 층 35	추가	-	(51,40)	-
층 36	병합5	일괄 정규화	-	(51,40)	-
층 37	층 36	활성화	-	(51,40)	Relu
층 38	층 37	컨볼루션 1D	40,5	(51,40)	Linear
층 39	층 38	일괄 정규화	-	(51,40)	-
층 40	층 39	활성화	-	(51,40)	Relu
층 41	층 40	컨볼루션 1D	40,5	(51,40)	선형
병합6	병합5, 층 41	추가	-	(51,40)	-
층 42	병합6	컨볼루션 1D	40,1	(51,40)	선형
병합7	층 2b, 층 16, 층 29, 층 42	추가	-	(51,40)	-
층 43	병합7	일괄 정규화	-	(51,40)	-
층 44	층 43	활성화	-	(51,40)	Relu
층 45	병합44	컨볼루션 1D	40,1	(51,40)	선형
층 46	층 45	일괄 정규화	-	(51,40)	-
층 47	층 46	활성화	-	(51,40)	Relu
층 48	층 47	컨볼루션 1D	40,1	(51,40)	선형
병합8	병합7, 층 48	추가	-	(51,40)	-
출력 층	병합8	컨볼루션 1D	3,1	(51,3)	소프트맥스

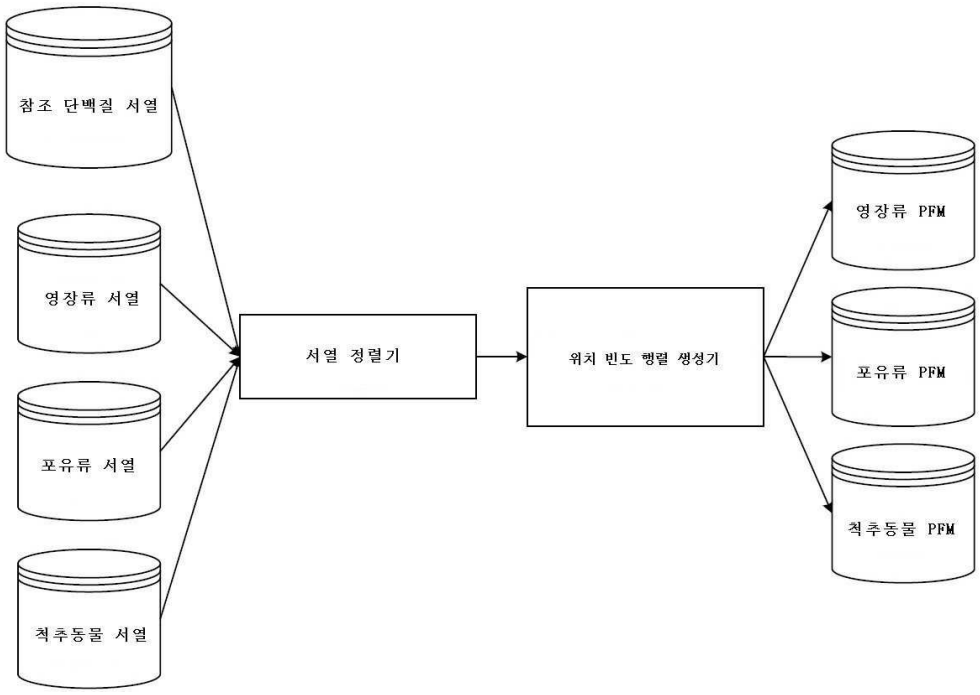
도면9



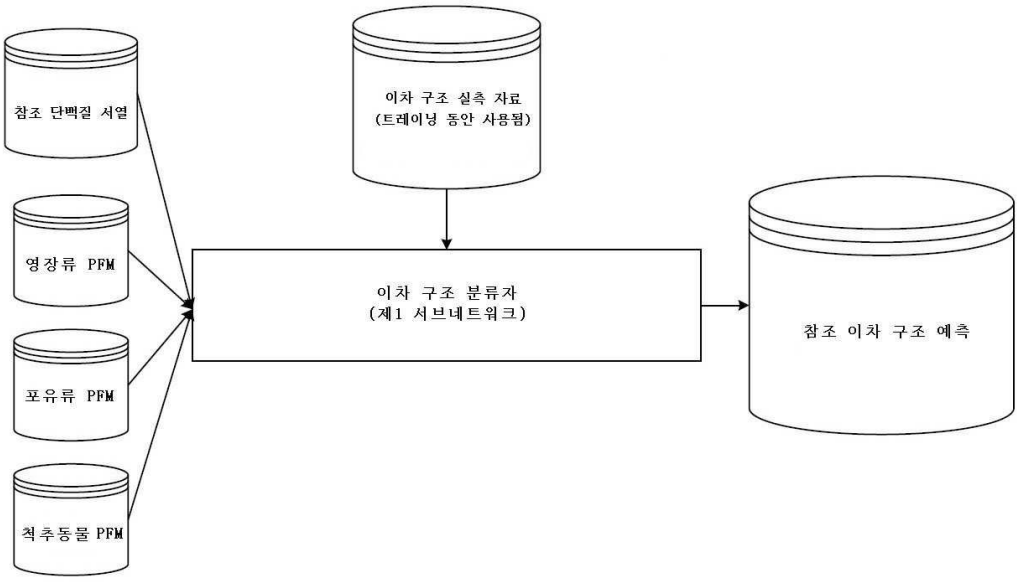
도면10



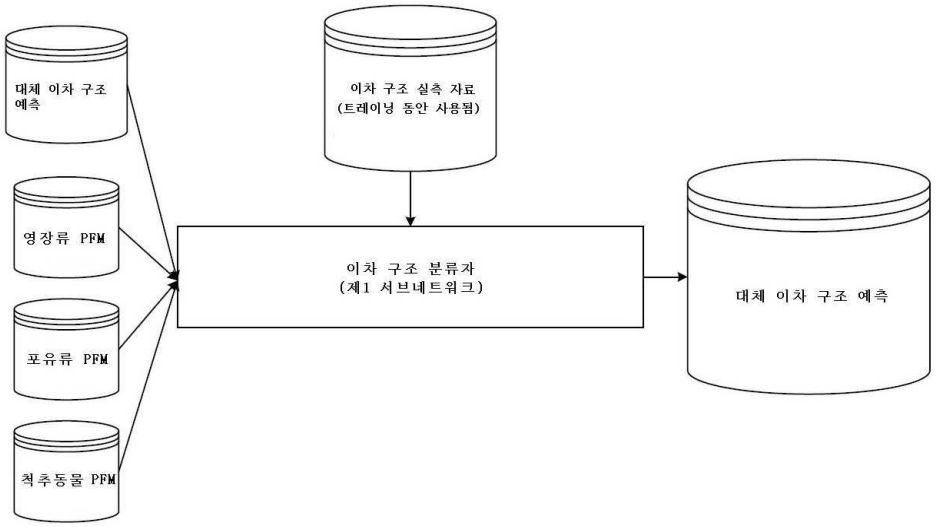
도면11



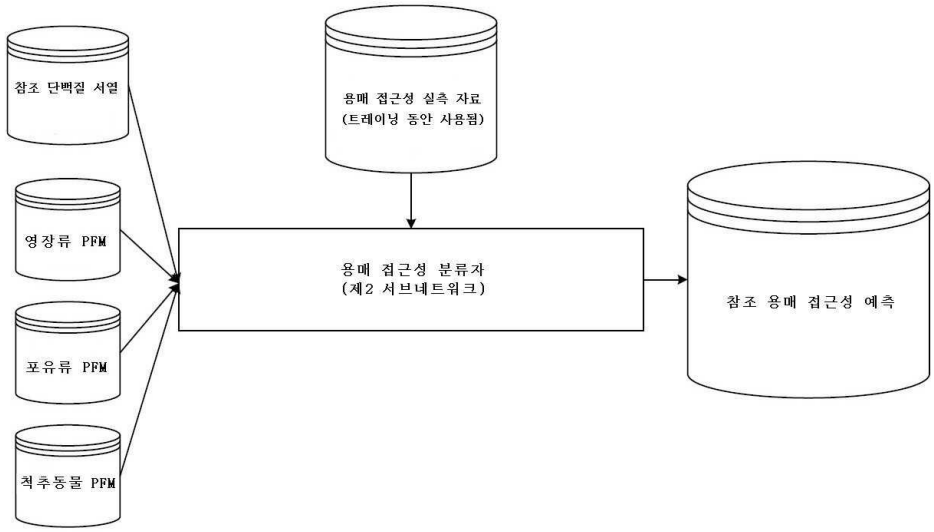
도면12



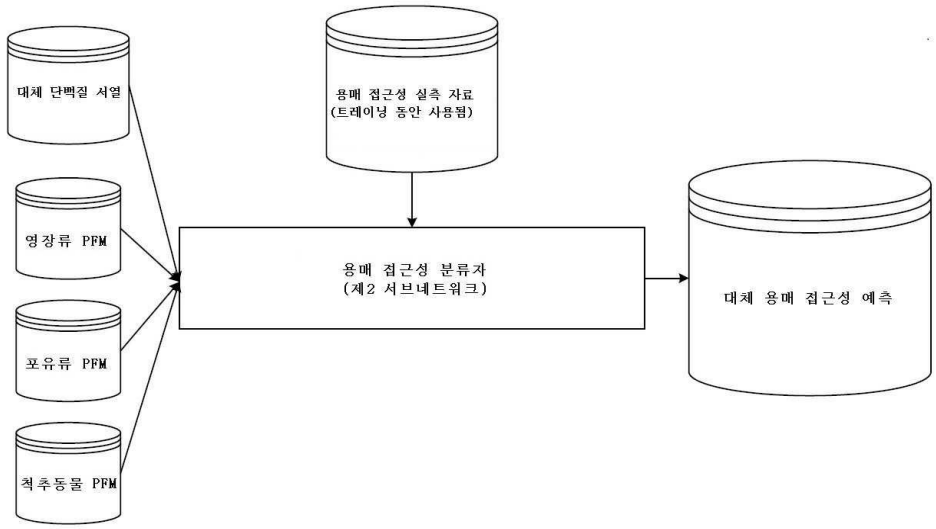
도면13


대체 이차 구조 예측
이차 구조 실측 자료 (트레이닝 동안 사용됨)
영장류 PFM
포유류 PFM
척추동물 PFM
이차 구조 분류자 (제1 서브네트워크)
대체 이차 구조 예측

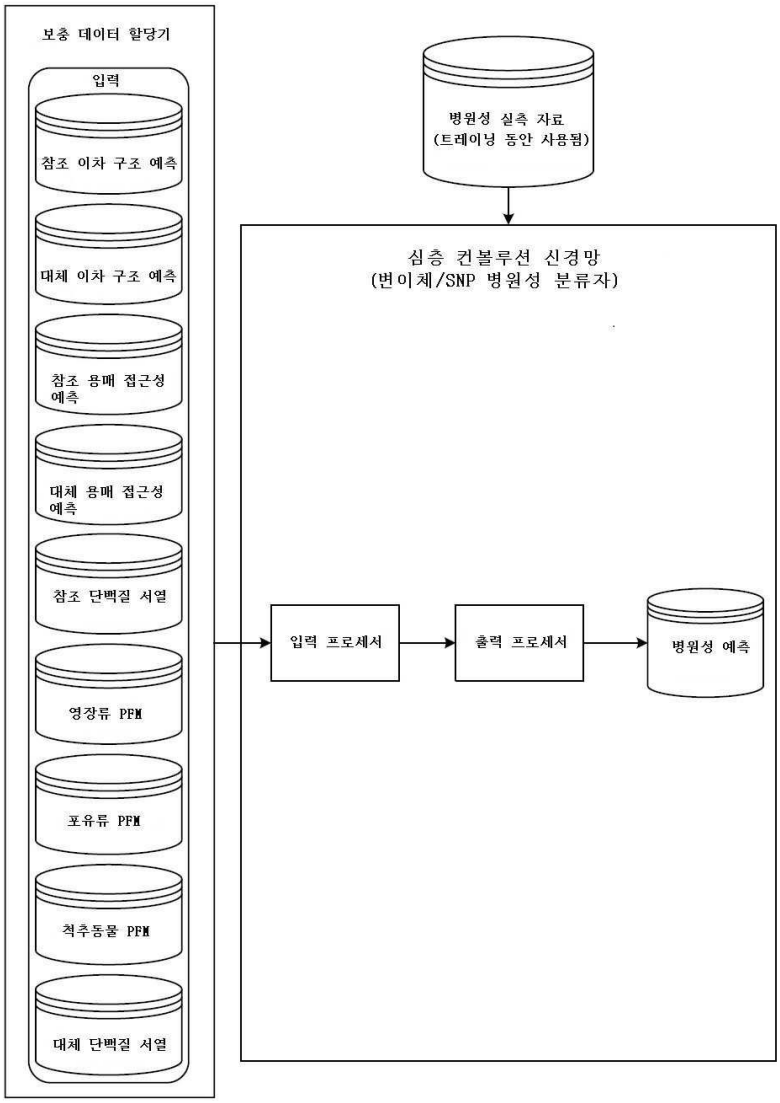
도면14


참조 단백질 서열
용매 접근성 실측 자료 (트레이닝 동안 사용됨)
영장류 PFM
포유류 PFM
척추동물 PFM
용매 접근성 분류자 (제2 서브네트워크)
참조 용매 접근성 예측

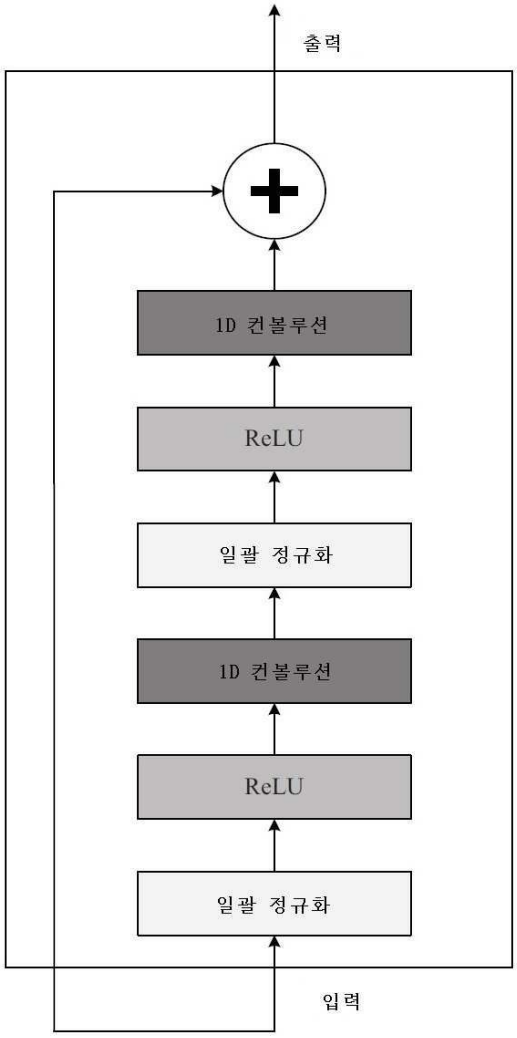
도면15


대체 단백질 서열
용매 접근성 실측 자료
(트레이닝 동안 사용됨)
영장류 PFM
용매 접근성 분류자
(제2 서브네트워크)
대체 용매 접근성 예측
포유류 PFM
척추동물 PFM

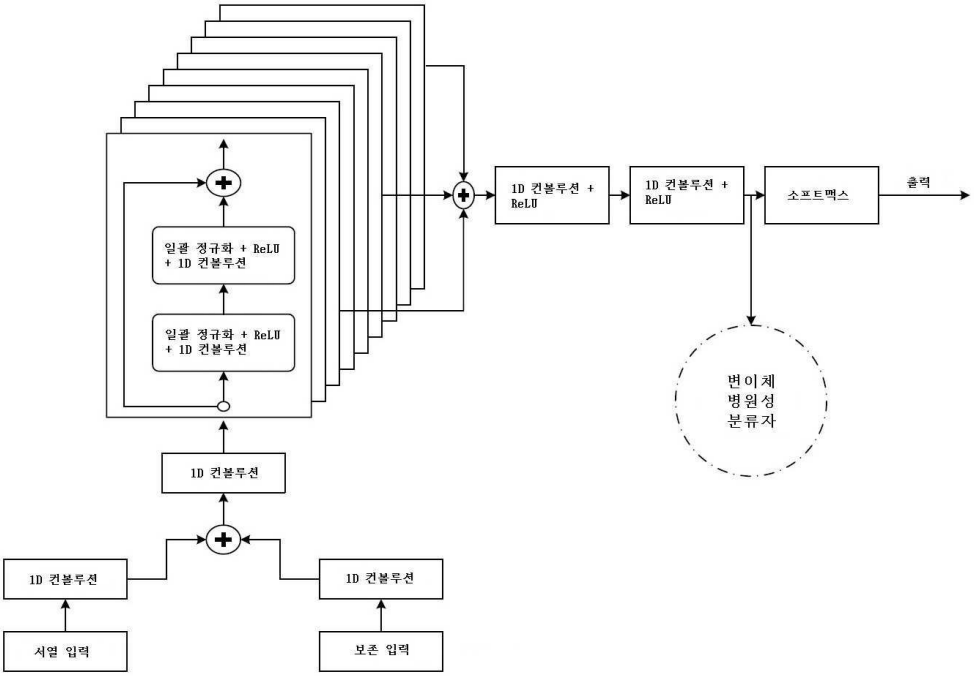
도면16


보충 데이터 할당기
입력
참조 이차 구조 예측
대체 이차 구조 예측
참조 용매 접근성 예측
대체 용매 접근성 예측
참조 단백질 서열
영장류 PFM
포유류 PFM
척추동물 PFM
대체 단백질 서열
병원성 실측 자료
(트레이닝 동안 사용됨)
심층 컨볼루션 신경망
(변이체/SNP 병원성 분류자)
입력 프로세서
출력 프로세서
병원성 예측

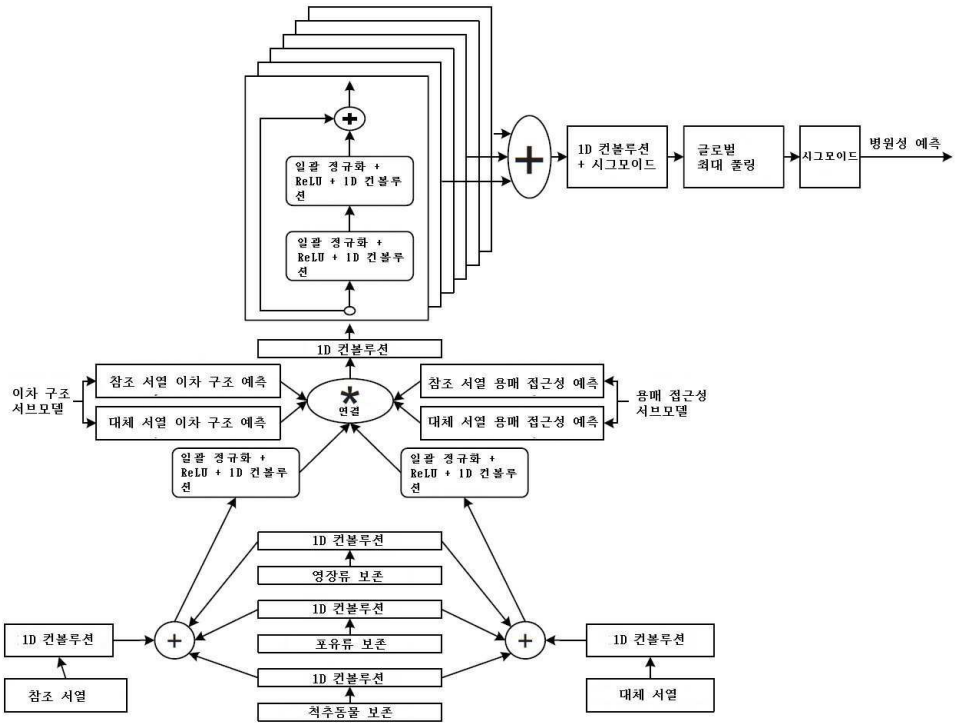
도면17



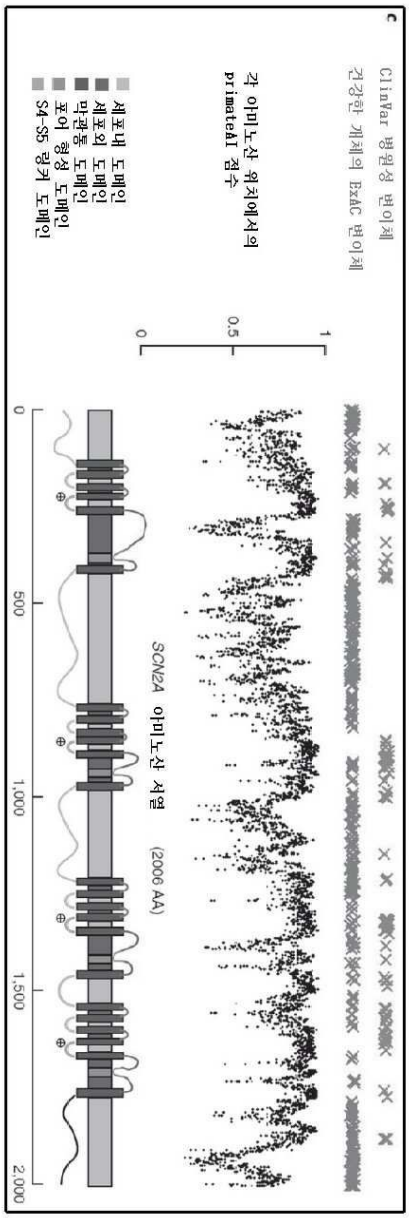
도면18


일괄 정규화 + ReLU + 1D 컨볼루션
일괄 정규화 + ReLU + 1D 컨볼루션
1D 컨볼루션
1D 컨볼루션
1D 컨볼루션
서열 입력
보존 입력
1D 컨볼루션 + ReLU
1D 컨볼루션 + ReLU
소프트맥스
출력
변이체 병원성 분류자

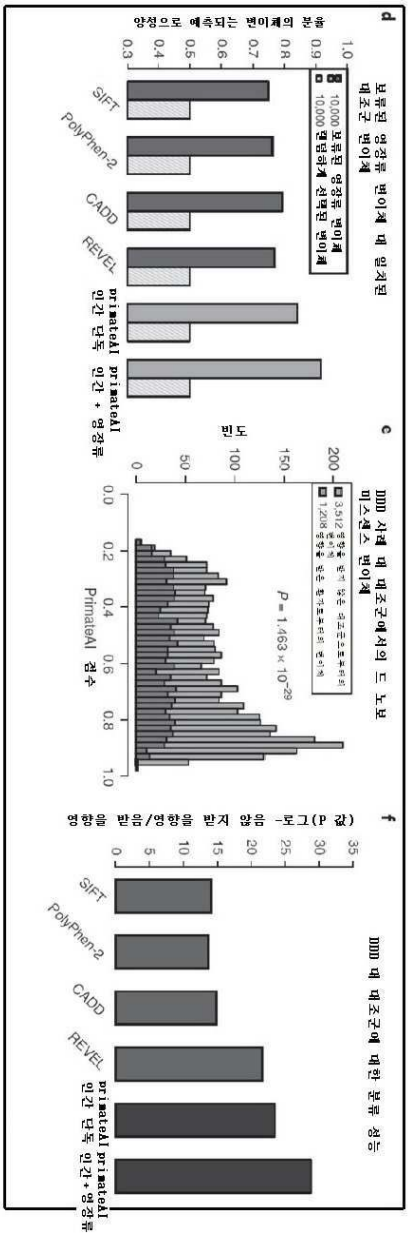
도면19


일괄 정규화 + ReLU + 1D 컨볼루션
일괄 정규화 + ReLU + 1D 컨볼루션
1D 컨볼루션
1D 컨볼루션 + 시그모이드
글로벌 최대 풀링
시그모이드
병원성 예측
이차 구조 서브모델
참조 서열 이차 구조 예측
대체 서열 이차 구조 예측
연결
참조 서열 용매 접근성 예측
대체 서열 용매 접근성 예측
용매 접근성 서브모델
일괄 정규화 + ReLU + 1D 컨볼루션
일괄 정규화 + ReLU + 1D 컨볼루션
1D 컨볼루션
영장류 보존
1D 컨볼루션
포유류 보존
1D 컨볼루션
척추동물 보존
1D 컨볼루션
참조 서열
1D 컨볼루션
대체 서열

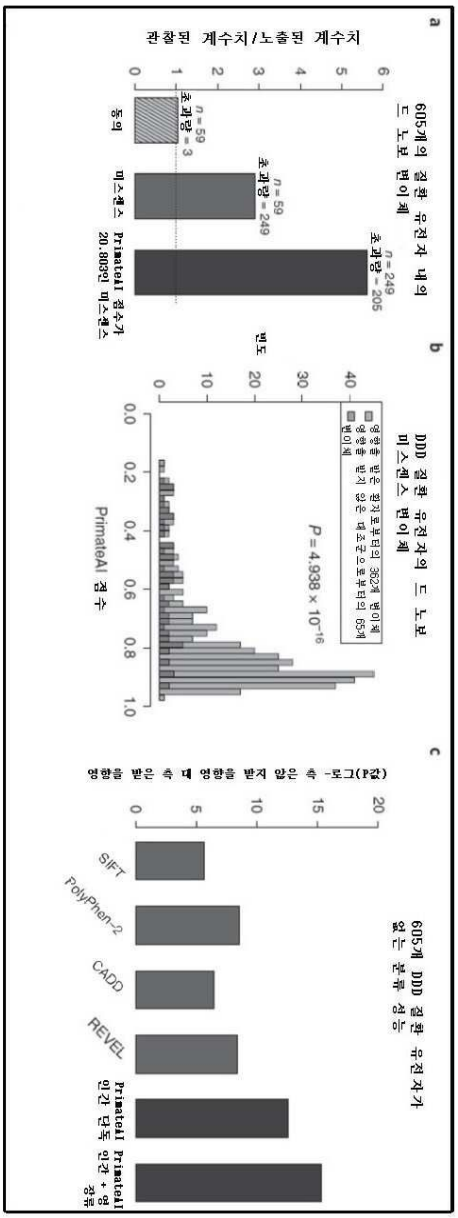
도면20


ClinVar 병원성 변이체
건강한 개체의 ExAC 변이체
각 아미노산 위치에서의 primateAI 점수
1
0.5
0
세포내 도메인
세포외 도메인
막관통 도메인
포어 형성 도메인
S4-S5 링커 도메인
0
500
1,000
1,500
2,000
SCN2A 아미노산 서열 (2006 AA)

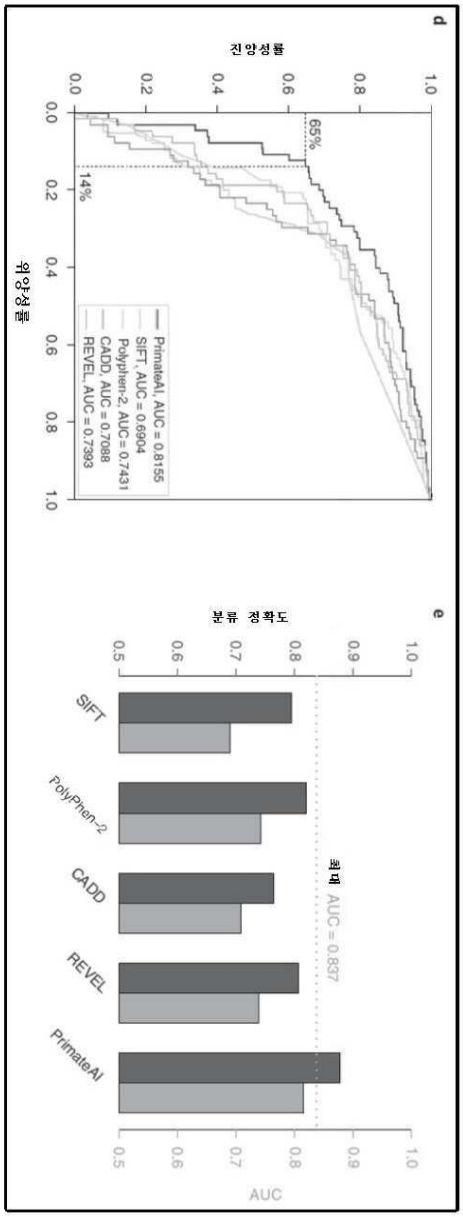
도면21


d
보류된 양성을 변이체 대 일치된 대조군 변이체
양성으로 예측되는 변이체의 분율
10,000 명 중 양성을 변이체
10,000 명 중 일치된 대조군 변이체
SIFT
PolyPhen-2
CADD
REVEL
primateAI 인간 단독
primateAI 인간 + 영장류
e
DDD 사례 대 대조군에서의 드 노보 미스센스 변이체
빈도
PrimateAI 점수
P = 1.463 × 10-29
f
DDD 대 대조군에 대한 분류 성능
영향을 받음/영향을 받지 않음 -로그(P 값)
SIFT
PolyPhen-2
CADD
REVEL
primateAI 인간 단독
primateAI 인간 + 영장류

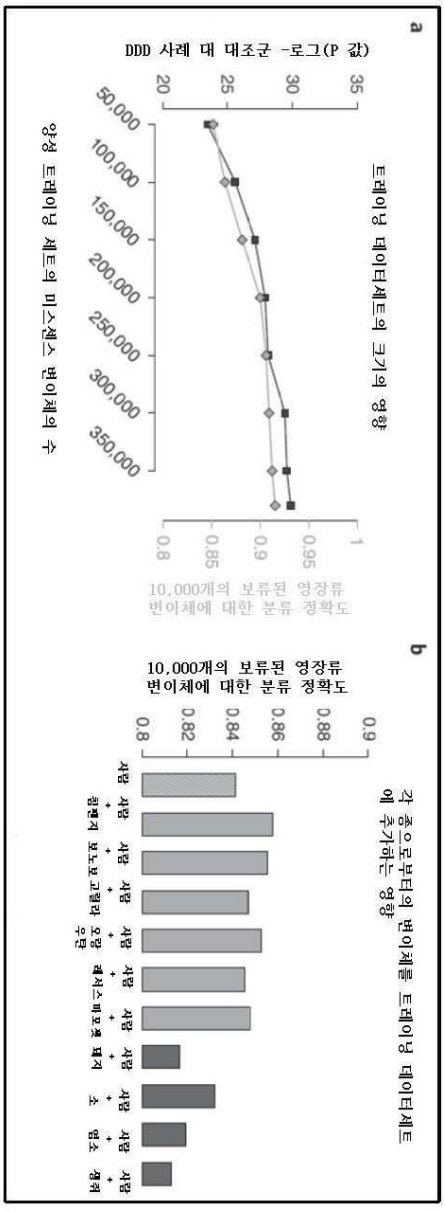
도면22ac


a
관찰된 계수치/노출된 계수치
605개의 질환 유전자 내의 드 노보 변이체
동의
미스센스
PrimateAI 점수가 20.803인 미스센스
n = 59
초과량 = 3
n = 59
초과량 = 249
n = 249
초과량 = 205
b
DDD 질환 유전자의 드 노보 미스센스 변이체
빈도
PrimateAI 점수
P = 4.938 × 10^-16
c
605개 DDD 질환 유전자가 없는 분류 성능
영향을 받은 측 대 영향을 받지 않은 측 -로그(P값)
SIFT
PolyPhen-2
CADD
REVEL
PrimateAI 인간 단독
PrimateAI 인간 + 영장류

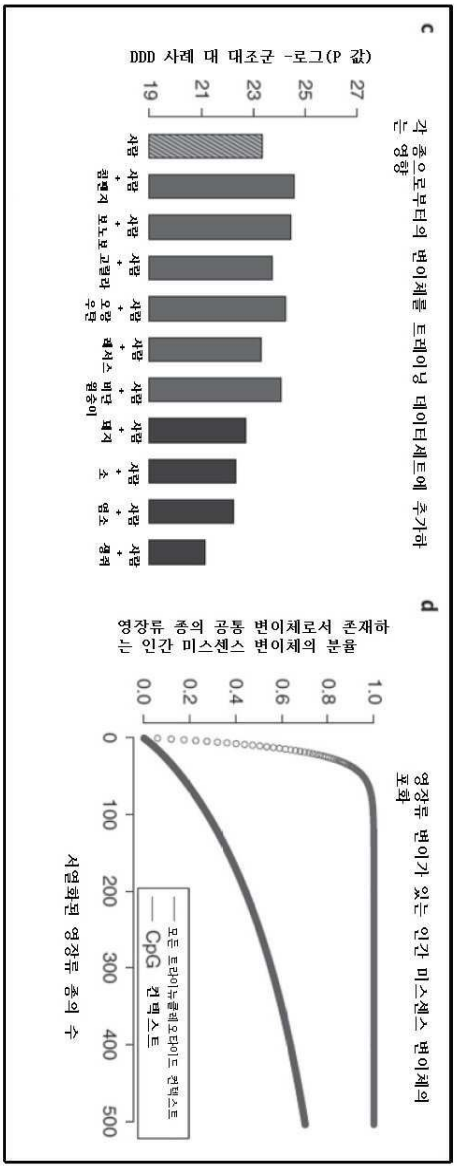
도면22de


d
진양성률
위양성률
65%
14%
PrimateAI, AUC = 0.8155
SIFT, AUC = 0.6904
Polyphen-2, AUC = 0.7431
CADD, AUC = 0.7088
REVEL, AUC = 0.7393
e
분류 정확도
SIFT
PolyPhen-2
CADD
REVEL
PrimateAI
최대 AUC = 0.837
AUC

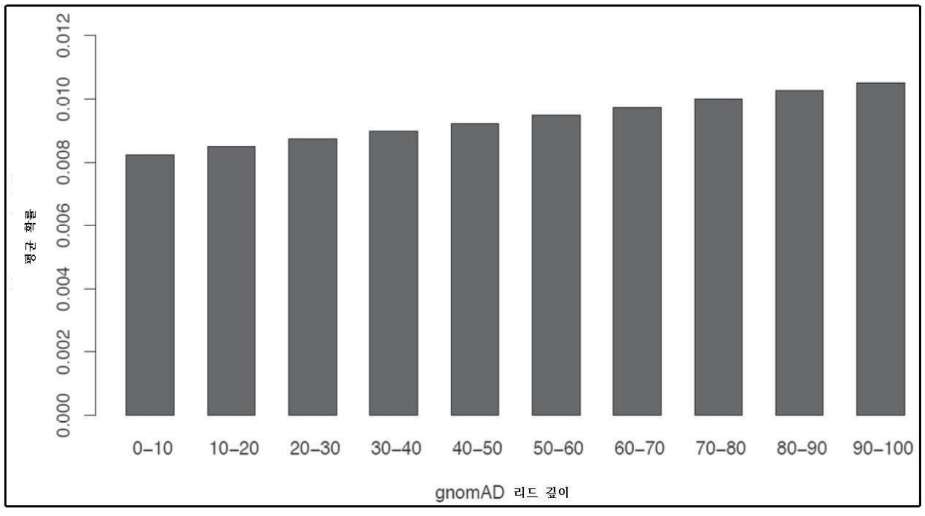
도면23ab


a
DDD 사례 대 대조군 -로그(P 값)
20
25
30
35
50,000
100,000
150,000
200,000
250,000
300,000
350,000
양성 트레이닝 세트의 미스센스 변이체의 수
트레이닝 데이터세트의 크기의 영향
0.8
0.85
0.9
0.95
1
10,000개의 보류된 영장류
변이체에 대한 분류 정확도
b
10,000개의 보류된 영장류
변이체에 대한 분류 정확도
0.8
0.82
0.84
0.86
0.88
0.9
각 종으로부터의 변이체를 트레이닝 데이터세트에 추가하는 영향
사람
사람 + 침팬지
사람 + 붉은털원숭이
사람 + 돼지
사람 + 소
사람 + 개
사람 + 생쥐
사람 + 랫트
사람 + 생쥐 + 랫트
사람 + 침팬지 + 붉은털원숭이 + 돼지 + 소 + 개 + 생쥐 + 랫트

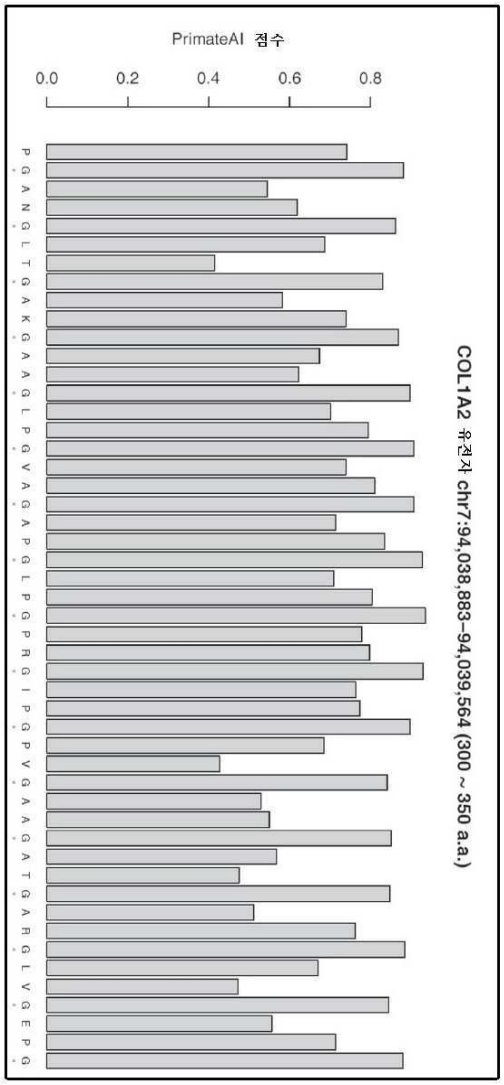
도면23cd


c
DDD 사례 대 대조군 -로그(P 값)
19
21
23
25
27
각 종으로부터의 변이체를 트레이닝 데이터세트에 추가하는 영향
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
사람 + 사람
d
영장류 종의 공통 변이체로서 존재하는 인간 미스센스 변이체의 분율
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
서열화된 영장류 종의 수
모든 트라이뉴클레오타이드 컨텍스트
CpG 컨텍스트
영장류 변이가 있는 인간 미스센스 변이체의 포화

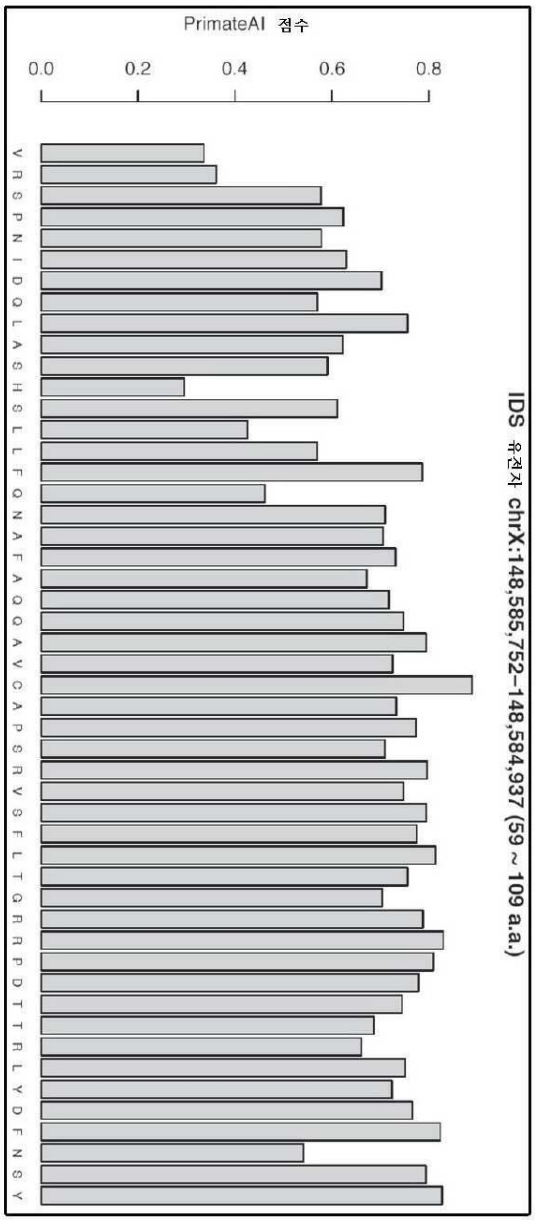
도면24


0.000
0.002
0.004
0.006
0.008
0.010
0.012
돌연변이율
0–10
10–20
20–30
30–40
40–50
50–60
60–70
70–80
80–90
90–100
gnomAD 리드 깊이

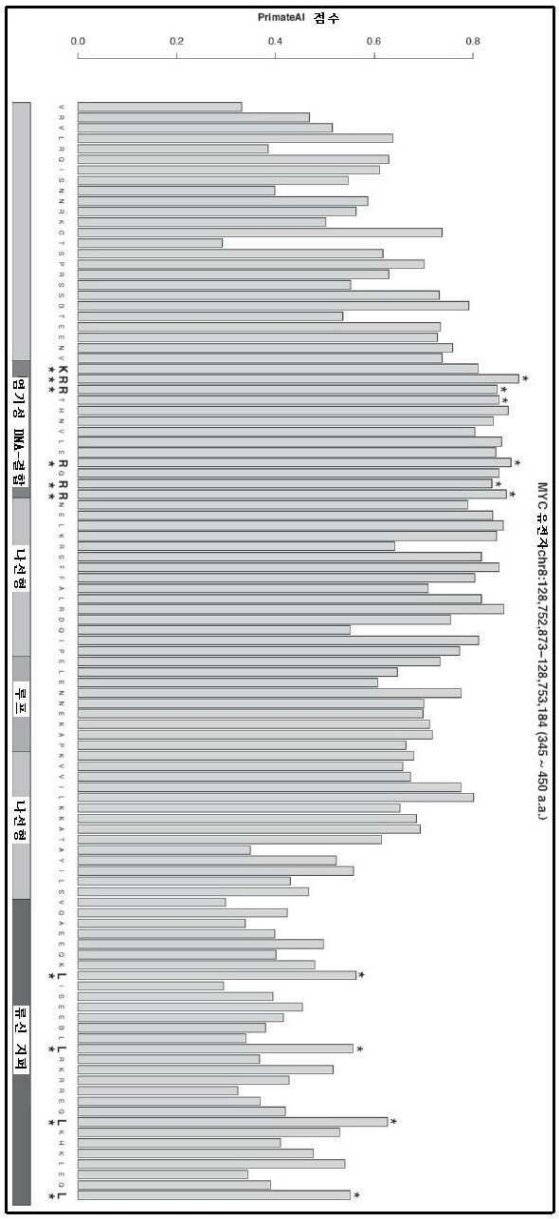
도면25a


PrimateAI 점수
0.0
0.2
0.4
0.6
0.8
COL1A2 유전자 chr7:94,038,883-94,039,564 (300 ~ 350 a.a.)

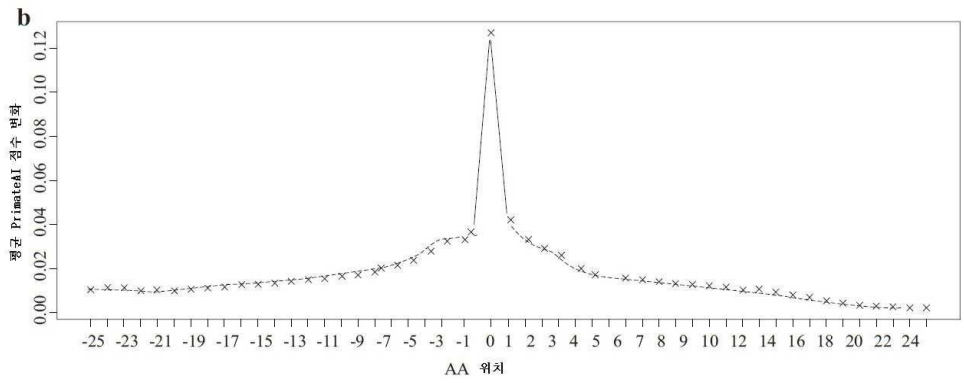
도면25b


PrimateAI 점수
0.0
0.2
0.4
0.6
0.8
V R S P N I D Q L A S H S L L F Q N A F A Q Q A V C A P S R V S F L T G R R P D T T R L Y D F N S Y
IDS 유전자 chrX:148,585,752-148,584,937 (59 ~ 109 a.a.)

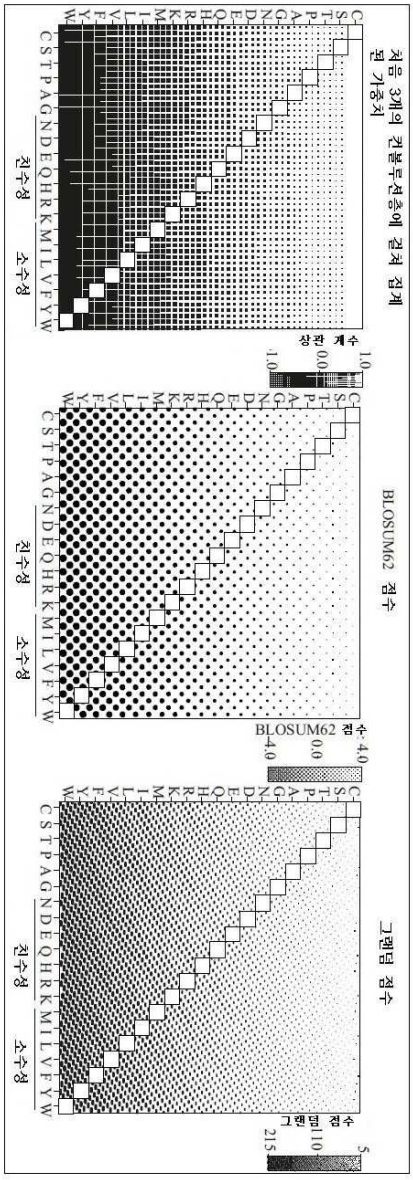
도면25c


PrimateAI 점수
0.0
0.2
0.4
0.6
0.8
MYC유전자chr8:128,752,873-128,753,184 (345 ~ 450 a.a.)
염기성 DNA-결합
나선형
코일
나선형
분산 지역

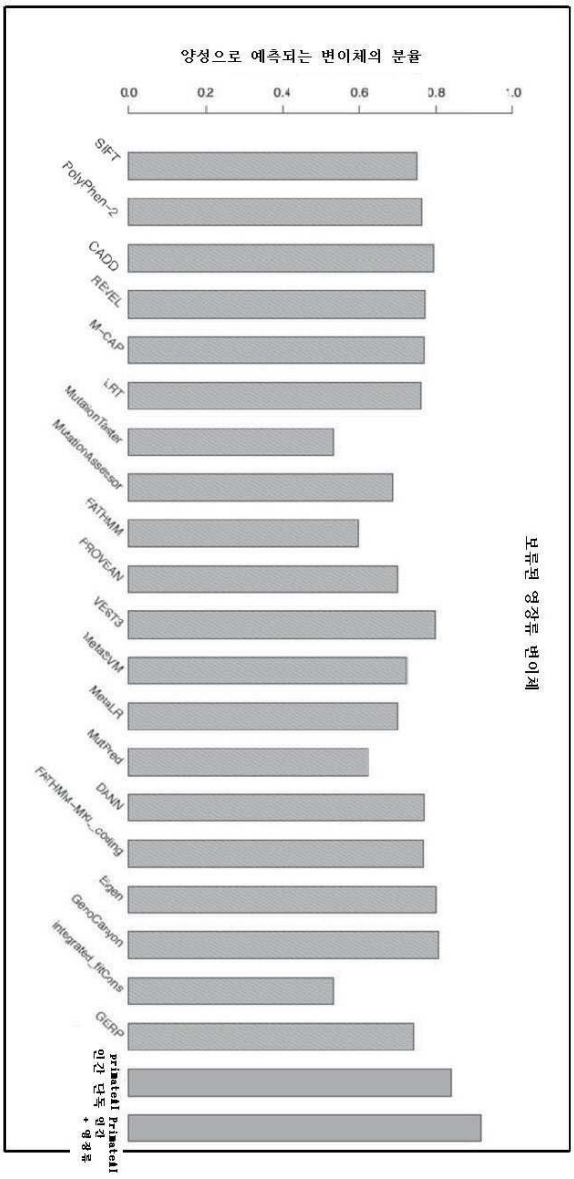
도면26


b
평균 PrimateAI 점수 변화
0.00
0.02
0.04
0.06
0.08
0.10
0.12
-25 -23 -21 -19 -17 -15 -13 -11 -9 -7 -5 -3 -1 0 1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24
AA 위치

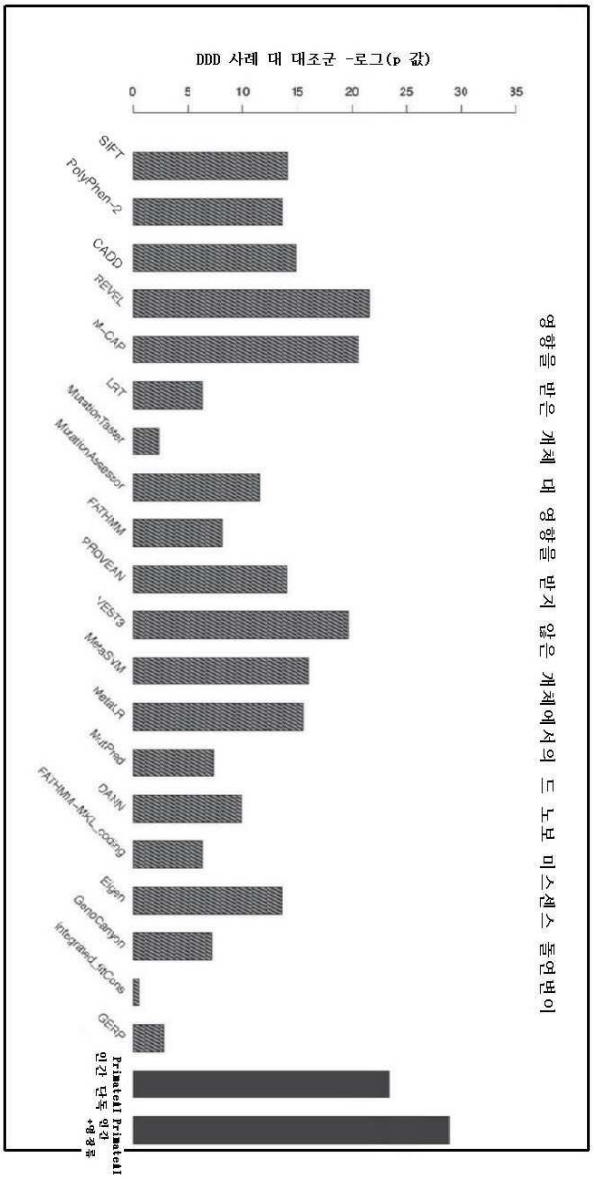
도면27



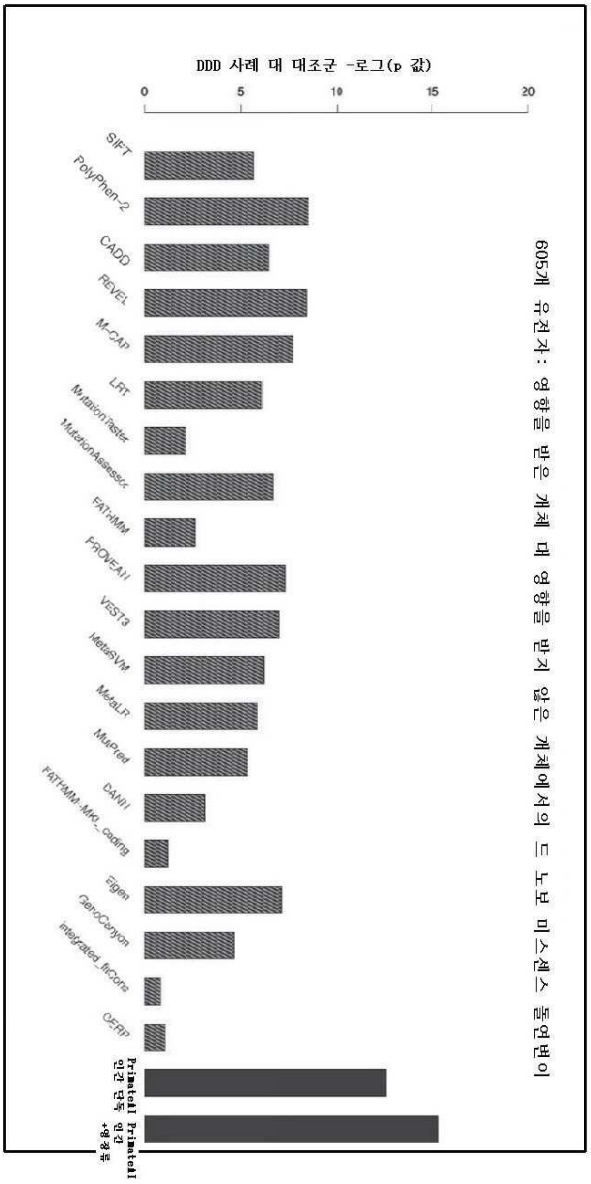
도면28a


양성으로 예측되는 변이체의 분율
0.0
0.2
0.4
0.6
0.8
1.0
SIFT
PolyPhen-2
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL_coding
Eigen
GenoCanyon
integrated_fitCons
GERP
primateDL
Primate1
인간 단독
인간 + 유인원
표준된 양성을 변이체

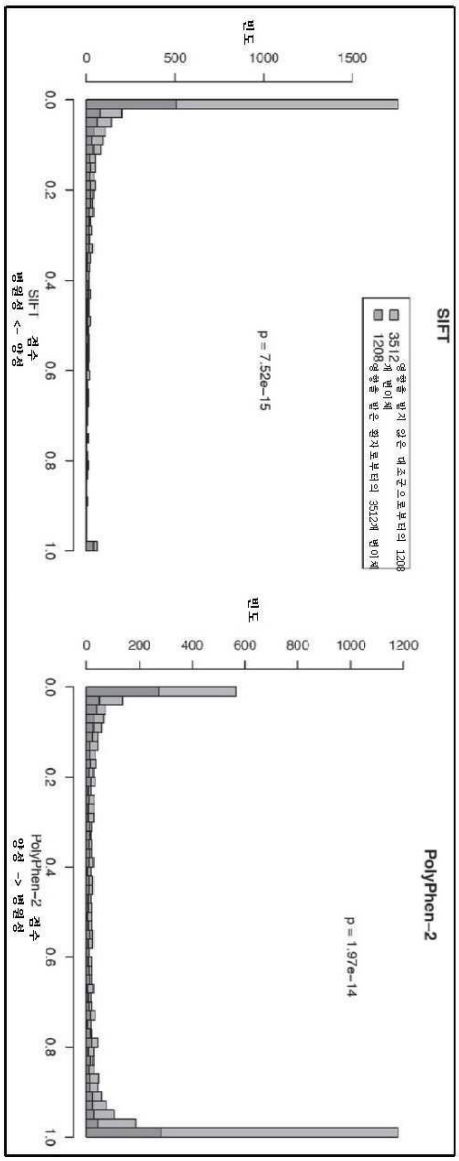
도면28b


DDD 사례 대 대조군 -로그(p 값)
0
5
10
15
20
25
30
35
SIFT
PolyPhen-2
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL_coding
Eigen
GenoCanyon
Integrated_fitCons
GERP
Primate1 Primate1
인간 단독 인간
+영장류
영향을 받은 개체 대 영향을 받지 않은 개체에서의 드 노보 미스센스 돌연변이

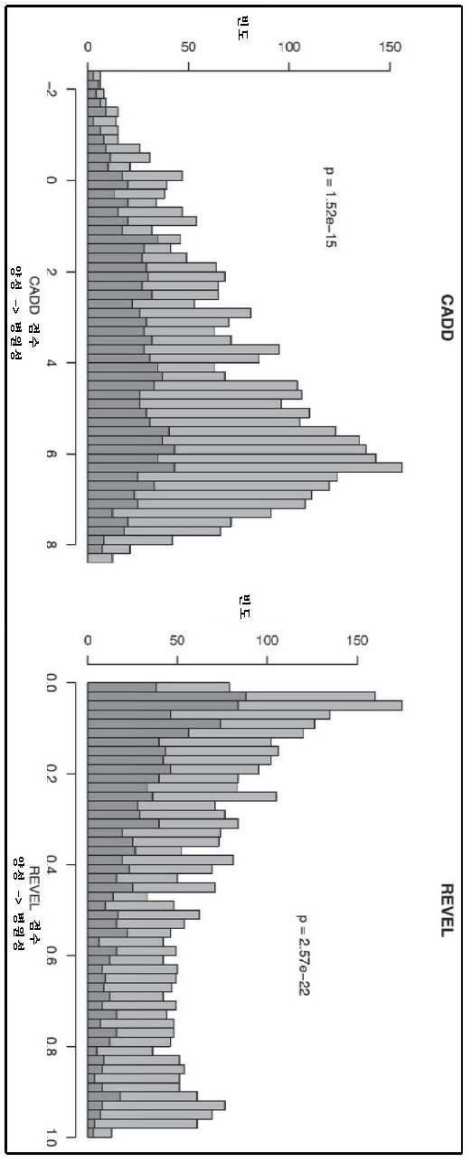
도면28c


DDD 사례 대 대조군 -로그(p 값)
0
5
10
15
20
SIFT
PolyPhen-2
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL_coding
Eigen
GenoCanyon
integrated_fitCons
GERP
Primate
Primate
605개 유전자: 영향을 받은 개체 대 영향을 받지 않은 개체에서의 드 노보 미스센스 돌연변이

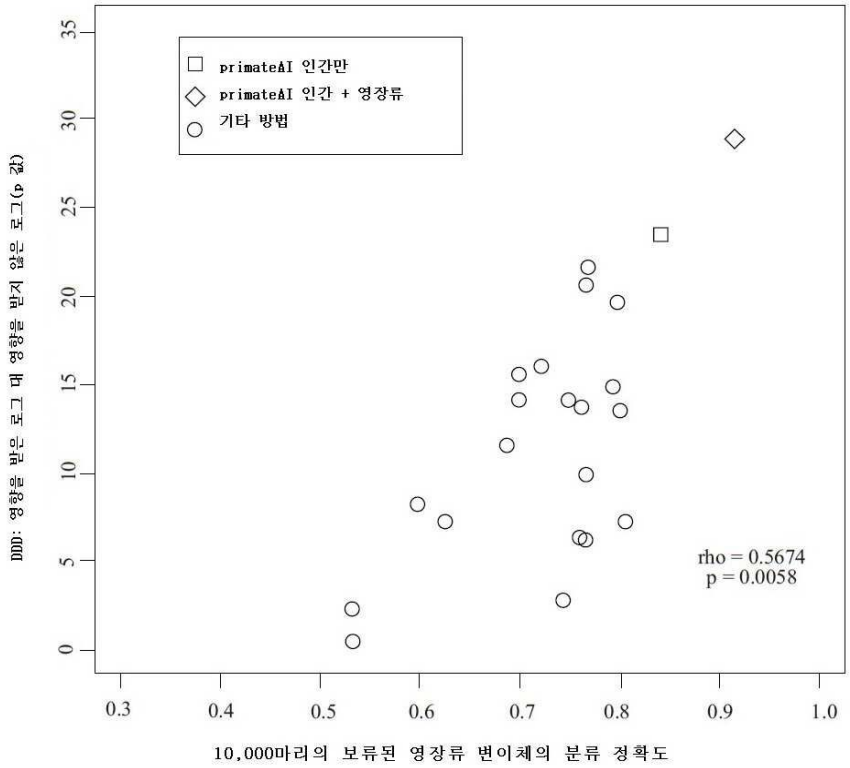
도면29a


SIFT
p = 7.52e-15
351개 변이체
1208개 변이체
빈도
SIFT 점수
병원성 -> 양성
PolyPhen-2
p = 1.97e-14
PolyPhen-2 점수
양성 -> 병원성

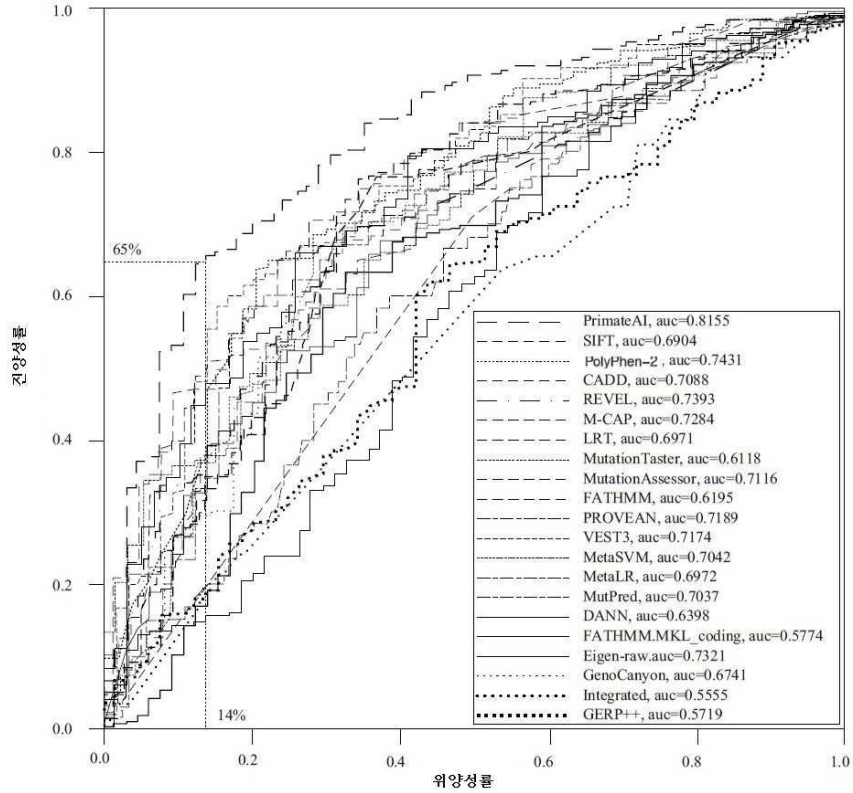
도면29b


CADD
p = 1.52e-15
빈도
CADD 점수
양성 <-> 병원성
REVEL
p = 2.57e-22
빈도
REVEL 점수
양성 <-> 병원성

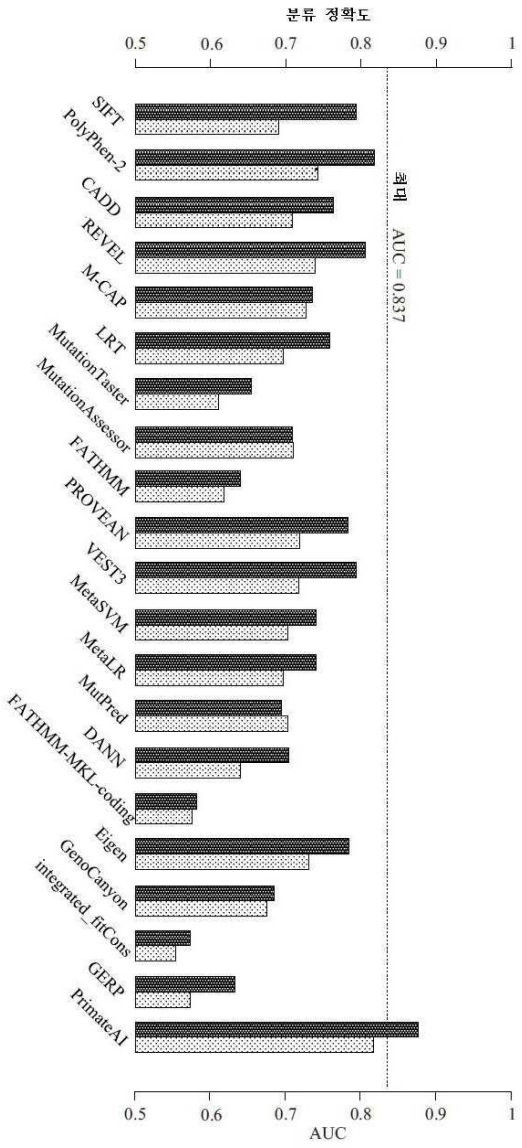
도면30a


primateAI 인간만
primateAI 인간 + 영장류
기타 방법
rho = 0.5674
p = 0.0058
DDD: 영향을 받은 로그 대 영향을 받지 않은 로그(p 값)
10,000마리의 보류된 영장류 변이체의 분류 정확도

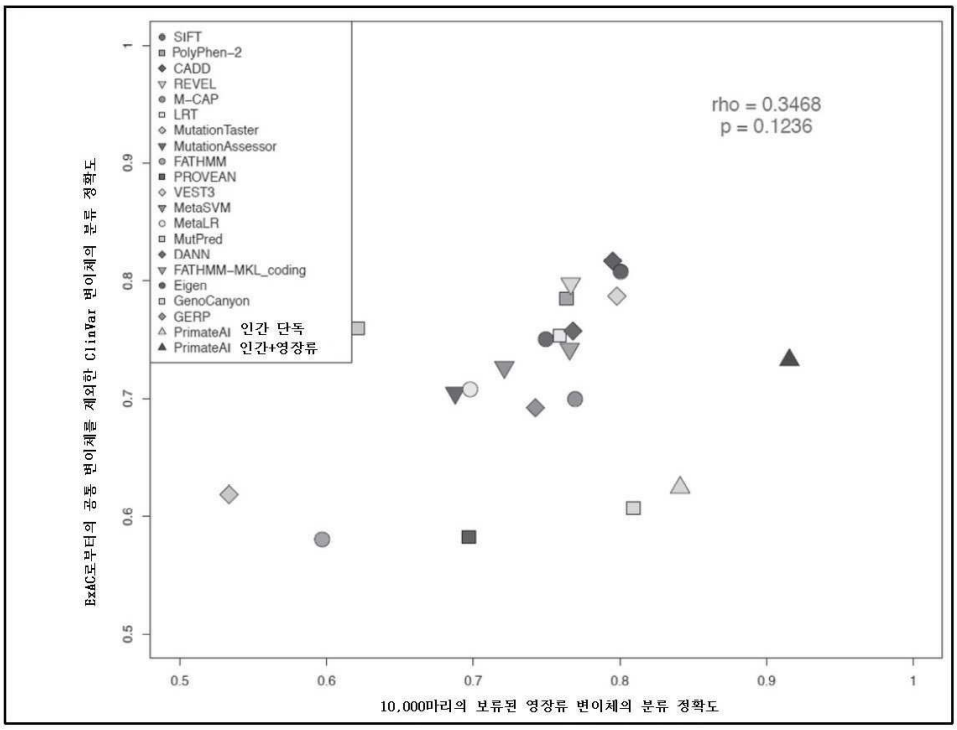
도면30b


65%
14%
PrimateAI, auc=0.8155
SIFT, auc=0.6904
PolyPhen-2 , auc=0.7431
CADD, auc=0.7088
REVEL, auc=0.7393
M-CAP, auc=0.7284
LRT, auc=0.6971
MutationTaster, auc=0.6118
MutationAssessor, auc=0.7116
FATHMM, auc=0.6195
PROVEAN, auc=0.7189
VEST3, auc=0.7174
MetaSVM, auc=0.7042
MetaLR, auc=0.6972
MutPred, auc=0.7037
DANN, auc=0.6398
FATHMM.MKL_coding, auc=0.5774
Eigen-raw.auc=0.7321
GenoCanyon, auc=0.6741
Integrated, auc=0.5555
GERP++, auc=0.5719
진양성률
위양성률

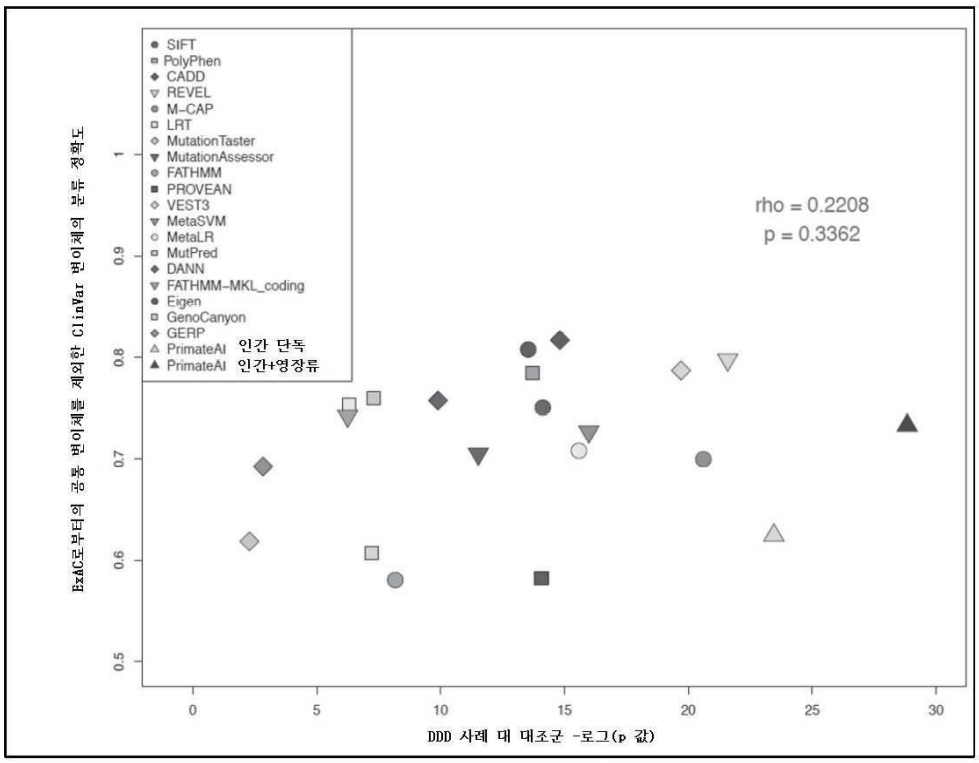
도면30c


분류 정확도
0.5
0.6
0.7
0.8
0.9
1
SIFT
PolyPhen-2
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL-coding
Eigen
GenoCanyon
integrated_fitCons
GERP
PrimateAI
최대 AUC = 0.837
AUC

도면31a


SIFT
PolyPhen-2
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL_coding
Eigen
GenoCanyon
GERP
PrimateAI 인간 단독
PrimateAI 인간+영장류
rho = 0.3468
p = 0.1236
ExAC로부터의 공통 변이체를 제외한 ClinVar 변이체의 분류 정확도
10,000마리의 보류된 영장류 변이체의 분류 정확도

도면31b


SIFT
PolyPhen
CADD
REVEL
M-CAP
LRT
MutationTaster
MutationAssessor
FATHMM
PROVEAN
VEST3
MetaSVM
MetaLR
MutPred
DANN
FATHMM-MKL_coding
Eigen
GenoCanyon
GERP
PrimateAI 인간 단독
PrimateAI 인간+영장류
rho = 0.2208
p = 0.3362
ExAC로부터의 공통 변이체를 제외한 ClinVar 변이체의 분류 정확도
DDD 사례 대 대조군 -로그(p 값)

도면32

	트레이닝 정확도	유효성확인 정확도	테스트 정확도
3-상태 이차 구조	80.32 %	79.17 %	79.86 %
3-상태 용매 접근성 구조	64.83 %	60.53 %	60.31 %

도면33

모델	복제물	보류된 영장류 변이체 정확도	DDD 사례 대 대조군 -로그(p 값)	DDD 605개 질화 유전자 -로그(p 값)
DL 예측 이차 구조 표지를 사용	1	0.914	28.60	15.61
	2	0.919	32.32	15.94
	3	0.915	29.48	16.38
	4	0.917	30.35	15.62
	5	0.916	29.11	16.22
	중앙값	0.916	29.48	15.94
이용가능한 경우 DL 예측 이차 구조 표지를 인간 단백질의 DSSP 이차 구조 표지로 교체	1	0.916	29.48	15.94
	2	0.913	31.72	16.24
	3	0.915	31.09	15.69
	4	0.915	30.49	16.09
	5	0.915	30.57	14.79
	중앙값	0.915	30.79	16.09

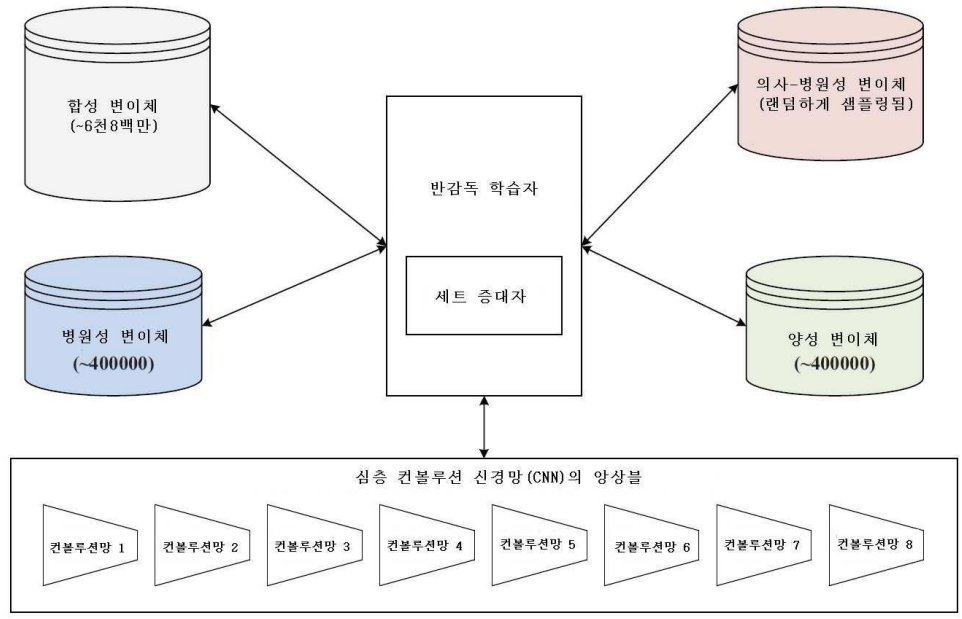
도면34

	보류된 영장류 변이체 정확도	DDD 사례 대 대조군 -로그(p 값)
SIFT	0.7490	14.1233
PolyPhen-2	0.7618	13.7044
CADD	0.7939	14.8167
M-CAP	0.7667	20.6078
LRT	0.7598	6.2960
MutationTaster	0.5332	2.2811
MutationAssessor	0.6879	11.5202
FATHMM	0.5981	8.1690
PROVEAN	0.6995	14.0747
VEST3	0.7978	19.7068
MetaSVM	0.7215	15.9829
MetaLR	0.6991	15.5842
REVEL	0.7686	21.5892
MutPred	0.6240	7.3044
DANN	0.7666	9.8976
MKL_coding	0.7657	6.2469
Eigen	0.8003	13.5364
GenoCanyon	0.8058	7.2196
integrated_fitCons	0.5331	0.4954
GERP	0.7430	2.8285
PrimateAI 인간 단독	0.8411	23.4536
PrimateAI 인간+영장류	0.9156	28.8346

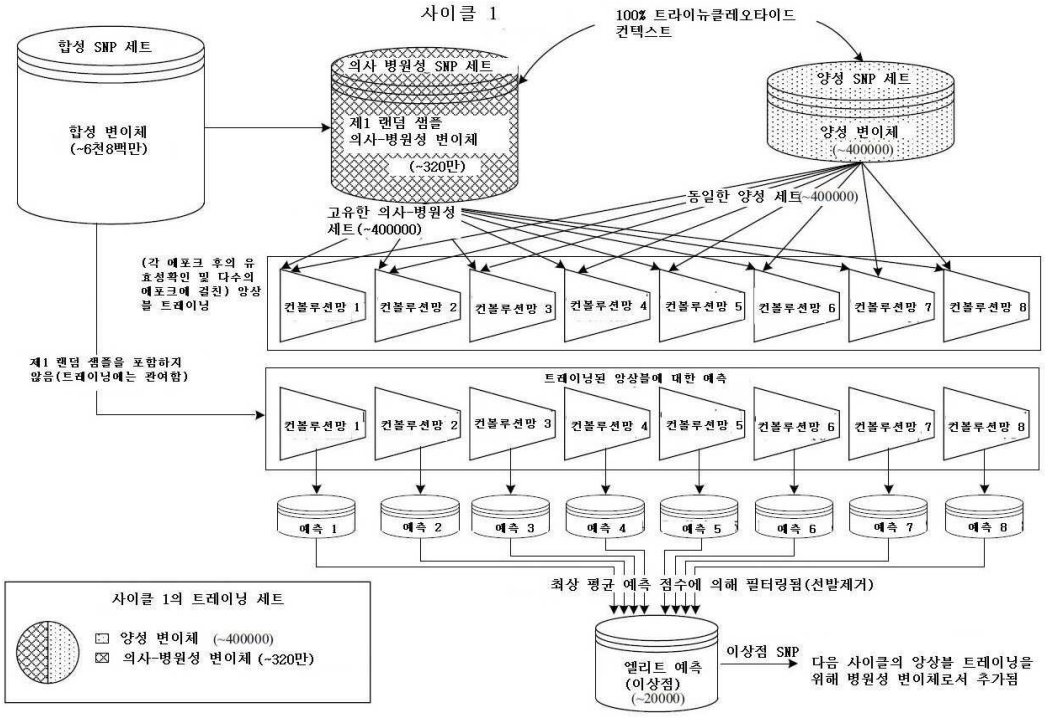
도면35

	로그 (P 값)	AUC	임계값	분류 정확도
SIFT	5.6793	0.6903	0.001	0.7949
PolyPhen-2	8.5522	0.7430	0.956	0.8197
CADD	6.4772	0.7087	5.040	0.7643
M-CAP	7.7143	0.7283	0.081	0.7348
LRT	6.0285	0.6970	1.00E-06	0.7604
MutationTaster	2.1366	0.6117	1.000	0.6564
MutationAssessor	6.6779	0.7115	2.240	0.7096
FATHMM	2.5979	0.6194	-1.260	0.6401
PROVEAN	7.3447	0.7188	-4.310	0.7833
VEST3	7.0357	0.7173	0.679	0.7953
MetaSVM	6.2556	0.7041	-0.3371	0.7406
MetaLR	5.8707	0.6971	0.349	0.7407
REVEL	8.4115	0.7392	0.456	0.8056
MutPred	5.3370	0.7036	0.526	0.6949
DANN	3.1588	0.6397	0.996	0.7046
MKL_coding	1.1649	0.5773	0.965	0.5840
Eigen	7.1659	0.7320	0.577	0.7844
GenoCanyon	4.6763	0.6740	0.999	0.6875
integrated_fitCons	0.8315	0.5554	0.706	0.5736
GERP	1.0330	0.5718	5.060	0.6330
PrimateAI 인간+영장류	15.3064	0.8154	0.802	0.8772

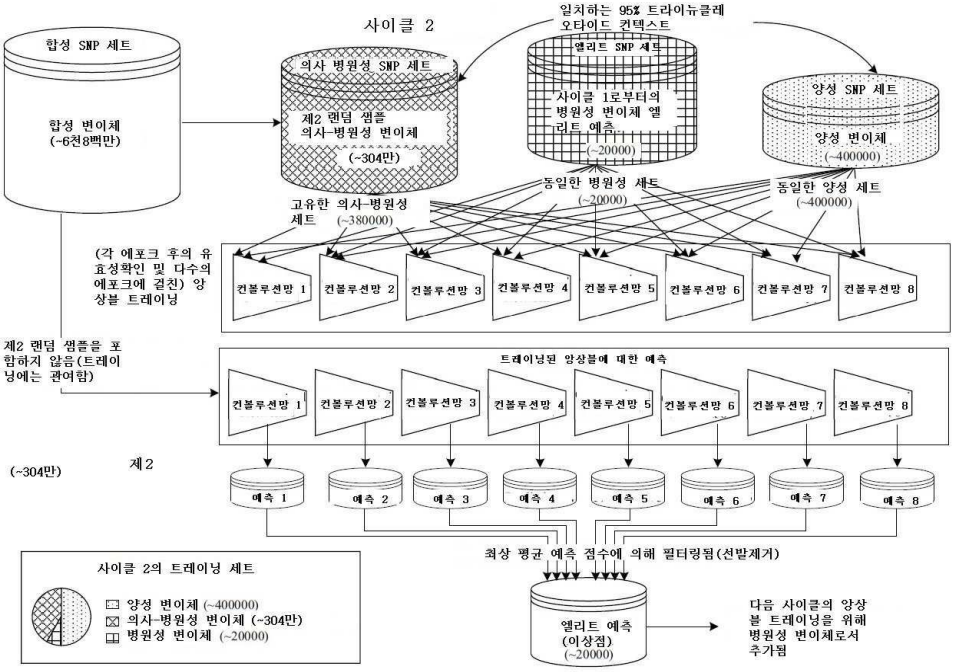
도면36


합성 변이체
(~6천8백만)
반감독 학습자
의사-병원성 변이체
(랜덤하게 샘플링됨)
세트 증대자
병원성 변이체
(~400000)
양성 변이체
(~400000)
심층 컨볼루션 신경망(CNN)의 앙상블
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
컨볼루션망 8

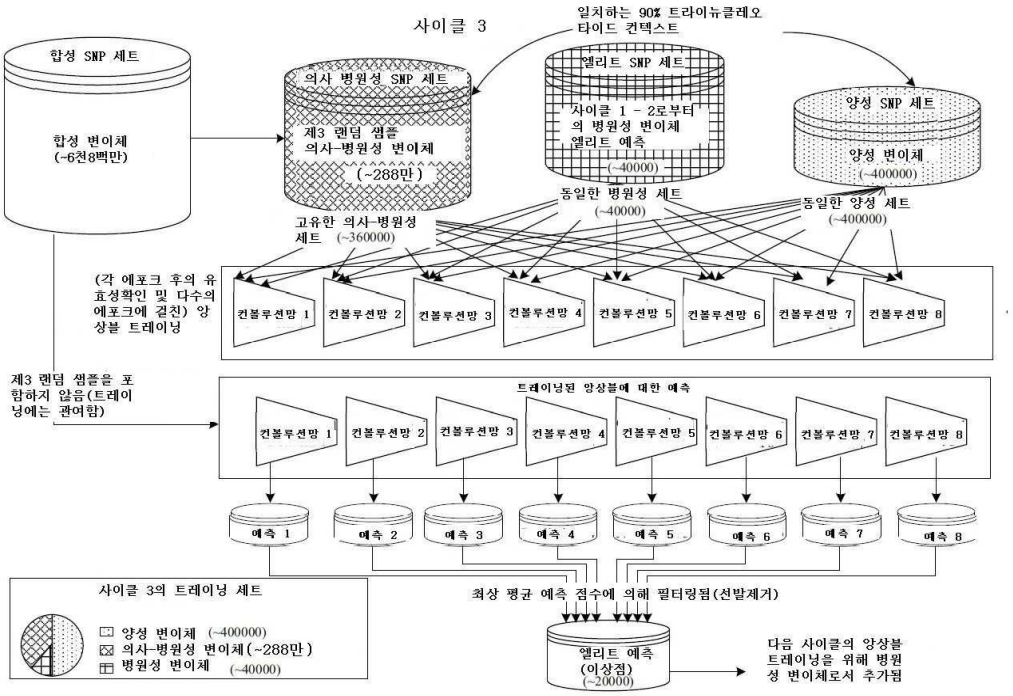
도면37


사이클 1
100% 트라이뉴클레오타이드
컨텍스트
합성 SNP 세트
합성 변이체
(~6천8백만)
의사 병원성 SNP 세트
제1 랜덤 샘플
의사-병원성 변이체
(~320만)
양성 SNP 세트
양성 변이체
(~400000)
동일한 양성 세트(~400000)
고유한 의사-병원성
세트(~400000)
(각 에포크 후의 유
효성확인 및 다수의
에포크에 걸친) 앙상
블 트레이닝
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
컨볼루션망 8
제1 랜덤 샘플을 포함하지
않음(트레이닝에는 관여함)
트레이닝된 앙상블에 대한 예측
예측 1
예측 2
예측 3
예측 4
예측 5
예측 6
예측 7
예측 8
최상 평균 예측 점수에 의해 필터링됨(선발제거)
사이클 1의 트레이닝 세트
양성 변이체 (~400000)
의사-병원성 변이체 (~320만)
엘리트 예측
(이상점)
(~20000)
이상점 SNP
다음 사이클의 앙상블 트레이닝을
위해 병원성 변이체로서 추가됨

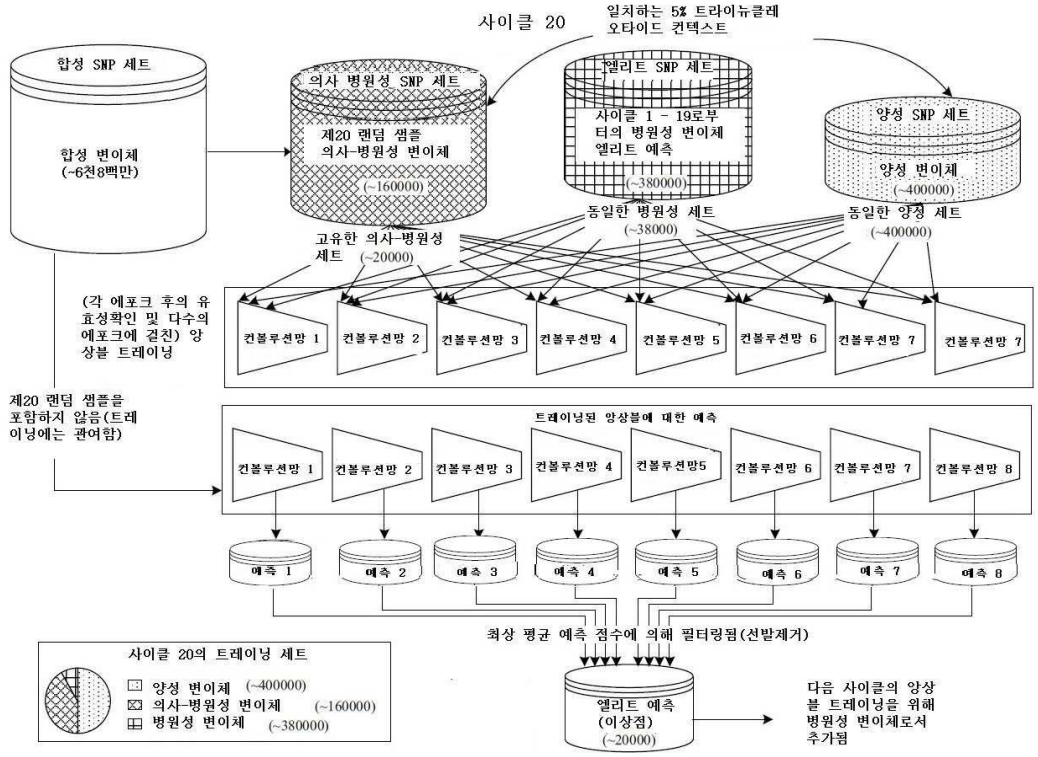
도면38


사이클 2
일치하는 95% 트라이뉴클레오타이드 컨텍스트
합성 SNP 세트
합성 변이체 (~6천8백만)
의사 병원성 SNP 세트
제2 랜덤 샘플 의사-병원성 변이체 (~304만)
엘리트 SNP 세트
사이클 1로부터의 병원성 변이체 엘리트 예측
(~20000)
양성 SNP 세트
양성 변이체 (~400000)
동일한 병원성 세트 (~20000)
동일한 양성 세트 (~400000)
고유한 의사-병원성 세트 (~380000)
(각 에포크 후의 유효성확인 및 다수의 에포크에 걸친) 앙상블 트레이닝
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
컨볼루션망 8
제2 랜덤 샘플을 포함하지 않음(트레이닝에는 관여함)
트레이닝된 앙상블에 대한 예측
(~304만)
제2
예측 1
예측 2
예측 3
예측 4
예측 5
예측 6
예측 7
예측 8
최상 평균 예측 점수에 의해 필터링됨(선발제거)
사이클 2의 트레이닝 세트
양성 변이체 (~400000)
의사-병원성 변이체 (~304만)
병원성 변이체 (~20000)
엘리트 예측 (이상점) (~20000)
다음 사이클의 앙상블 트레이닝을 위해 병원성 변이체로서 추가됨

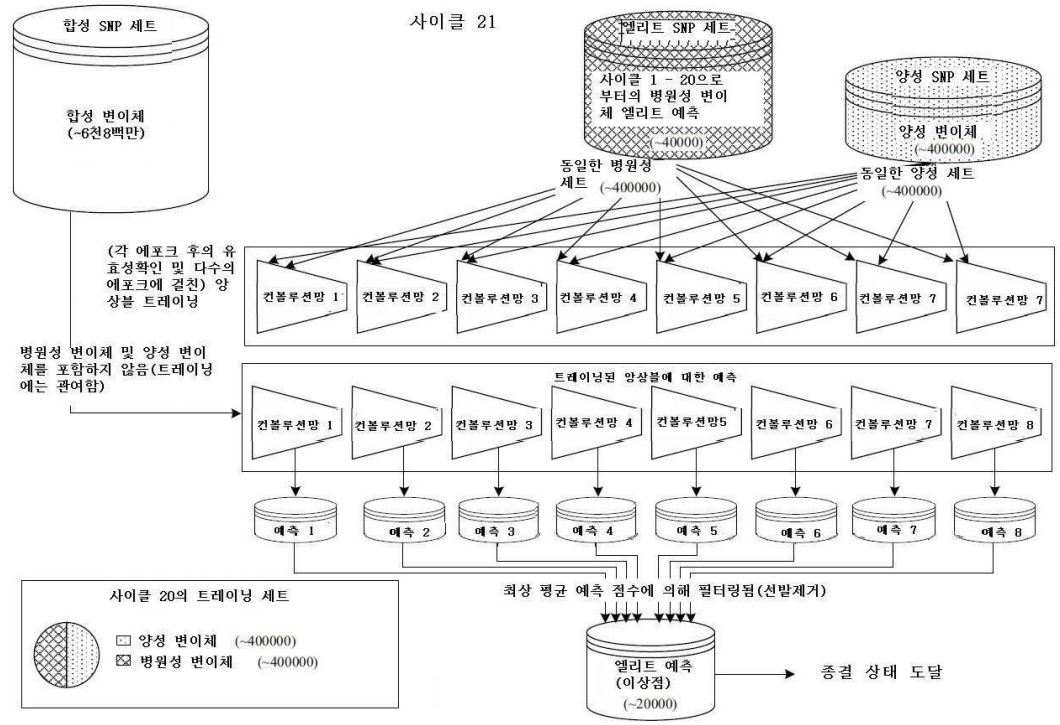
도면39


사이클 3
일치하는 90% 트라이뉴클레오타이드 컨텍스트
합성 SNP 세트
합성 변이체 (~6천8백만)
의사 병원성 SNP 세트
제3 랜덤 샘플 의사-병원성 변이체 (~288만)
엘리트 SNP 세트
사이클 1 - 2로부터의 병원성 변이체 엘리트 예측
(~40000)
양성 SNP 세트
양성 변이체 (~400000)
동일한 병원성 세트 (~40000)
동일한 양성 세트 (~400000)
고유한 의사-병원성 세트 (~360000)
(각 에포크 후의 유효성확인 및 다수의 에포크에 걸친) 앙상블 트레이닝
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
컨볼루션망 8
제3 랜덤 샘플을 포함하지 않음(트레이닝에는 관여함)
트레이닝된 앙상블에 대한 예측
예측 1
예측 2
예측 3
예측 4
예측 5
예측 6
예측 7
예측 8
최상 평균 예측 점수에 의해 필터링됨(선발제거)
사이클 3의 트레이닝 세트
양성 변이체 (~400000)
의사-병원성 변이체 (~288만)
병원성 변이체 (~40000)
엘리트 예측 (이상점) (~20000)
다음 사이클의 앙상블 트레이닝을 위해 병원성 변이체로서 추가됨

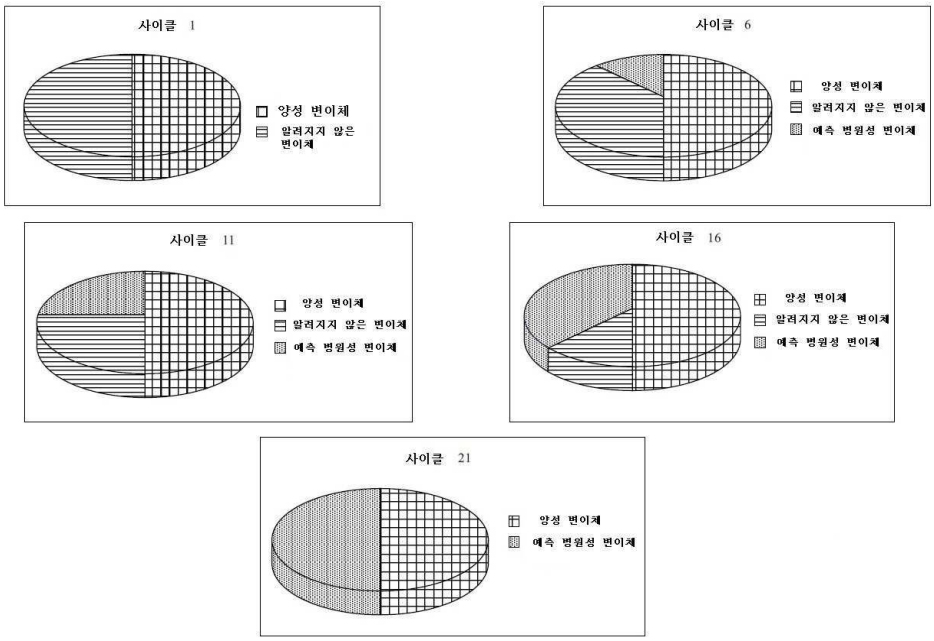
도면40


사이클 20
일치하는 5% 트라이뉴클레오타이드 컨텍스트
합성 SNP 세트
합성 변이체 (~6천8백만)
의사 병원성 SNP 세트
제20 랜덤 샘플 의사-병원성 변이체 (~160000)
엘리트 SNP 세트
사이클 1 - 19로부터의 병원성 변이체 엘리트 예측 (~380000)
양성 SNP 세트
양성 변이체 (~400000)
동일한 병원성 세트 (~38000)
동일한 양성 세트 (~400000)
고유한 의사-병원성 세트 (~20000)
(각 에포크 후의 유효성확인 및 다수의 에포크에 걸친) 앙상블 트레이닝
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
제20 랜덤 샘플을 포함하지 않음(트레이닝에는 관여함)
트레이닝된 앙상블에 대한 예측
컨볼루션망 8
예측 1
예측 2
예측 3
예측 4
예측 5
예측 6
예측 7
예측 8
최상 평균 예측 점수에 의해 필터링됨(선발제거)
사이클 20의 트레이닝 세트
양성 변이체 (~400000)
의사-병원성 변이체 (~160000)
병원성 변이체 (~380000)
엘리트 예측 (이상점) (~20000)
다음 사이클의 앙상블 트레이닝을 위해 병원성 변이체로서 추가됨

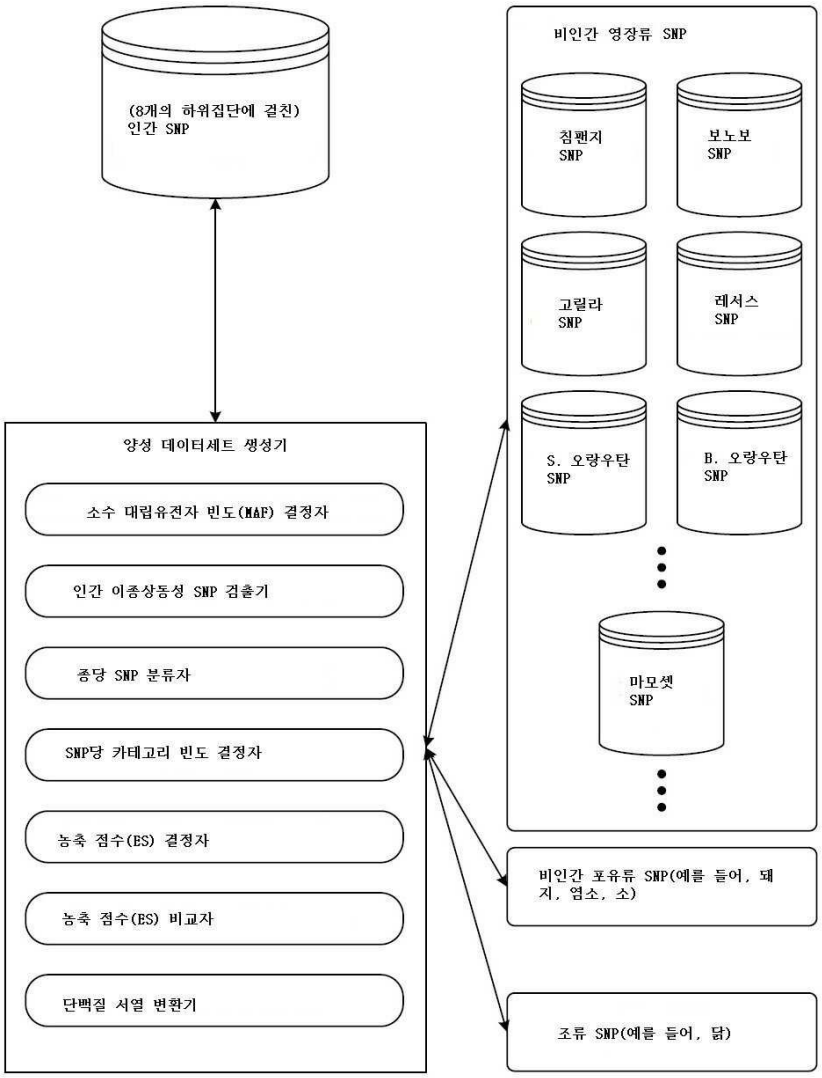
도면41


사이클 21
합성 SNP 세트
합성 변이체 (~6천8백만)
엘리트 SNP 세트
사이클 1 - 20으로부터의 병원성 변이체 엘리트 예측 (~40000)
양성 SNP 세트
양성 변이체 (~400000)
동일한 병원성 세트 (~400000)
동일한 양성 세트 (~400000)
(각 에포크 후의 유효성확인 및 다수의 에포크에 걸친) 앙상블 트레이닝
컨볼루션망 1
컨볼루션망 2
컨볼루션망 3
컨볼루션망 4
컨볼루션망 5
컨볼루션망 6
컨볼루션망 7
병원성 변이체 및 양성 변이체를 포함하지 않음(트레이닝에는 관여함)
트레이닝된 앙상블에 대한 예측
컨볼루션망 8
예측 1
예측 2
예측 3
예측 4
예측 5
예측 6
예측 7
예측 8
최상 평균 예측 점수에 의해 필터링됨(선발제거)
사이클 20의 트레이닝 세트
양성 변이체 (~400000)
병원성 변이체 (~400000)
엘리트 예측 (이상점) (~20000)
종결 상태 도달

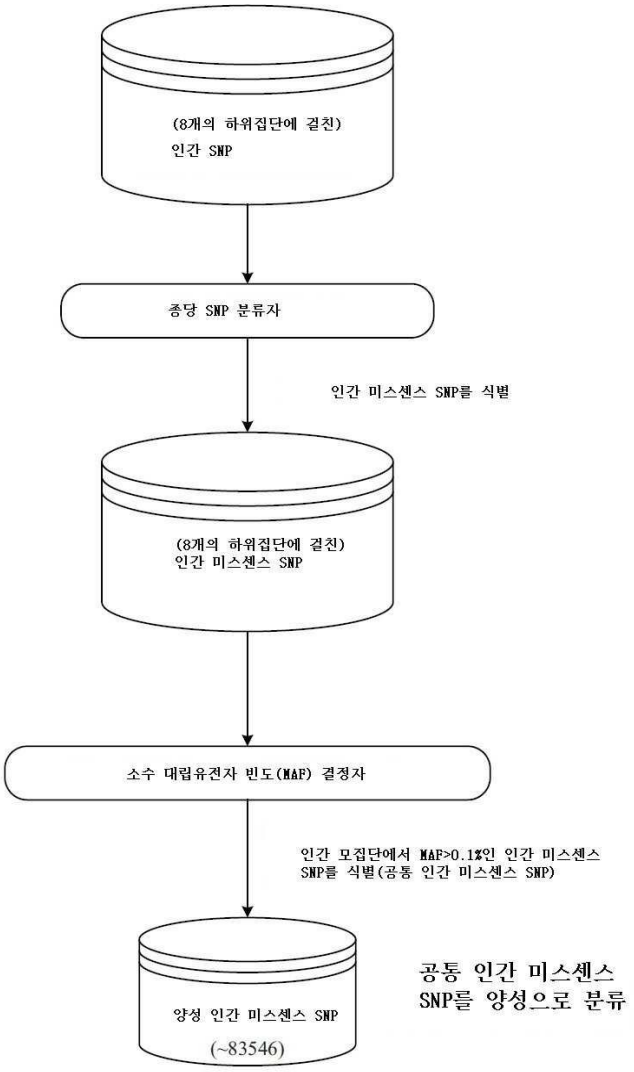
도면42


사이클 1
양성 변이체
알려지지 않은 변이체
사이클 6
양성 변이체
알려지지 않은 변이체
예측 병원성 변이체
사이클 11
양성 변이체
알려지지 않은 변이체
예측 병원성 변이체
사이클 16
양성 변이체
알려지지 않은 변이체
예측 병원성 변이체
사이클 21
양성 변이체
예측 병원성 변이체

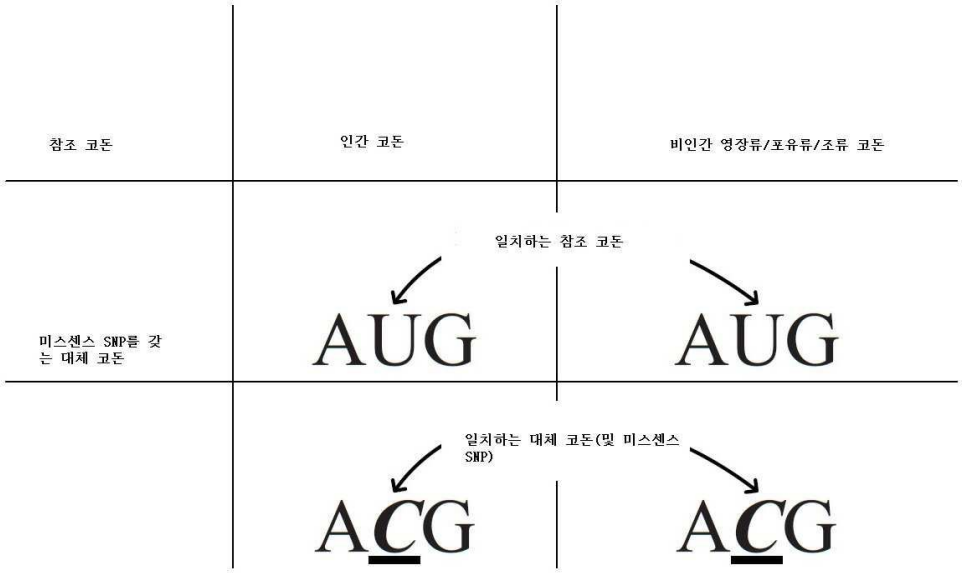
도면43


(8개의 하위집단에 걸친)
인간 SNP
양성 데이터세트 생성기
소수 대립유전자 빈도(MAF) 결정자
인간 이종상동성 SNP 검출기
종당 SNP 분류자
SNP당 카테고리 빈도 결정자
농축 점수(ES) 결정자
농축 점수(ES) 비교자
단백질 서열 변환기
비인간 영장류 SNP
침팬지
SNP
보노보
SNP
고릴라
SNP
레서스
SNP
S. 오랑우탄
SNP
B. 오랑우탄
SNP
마모셋
SNP
비인간 포유류 SNP(예를 들어, 돼지, 염소, 소)
조류 SNP(예를 들어, 닭)

도면44

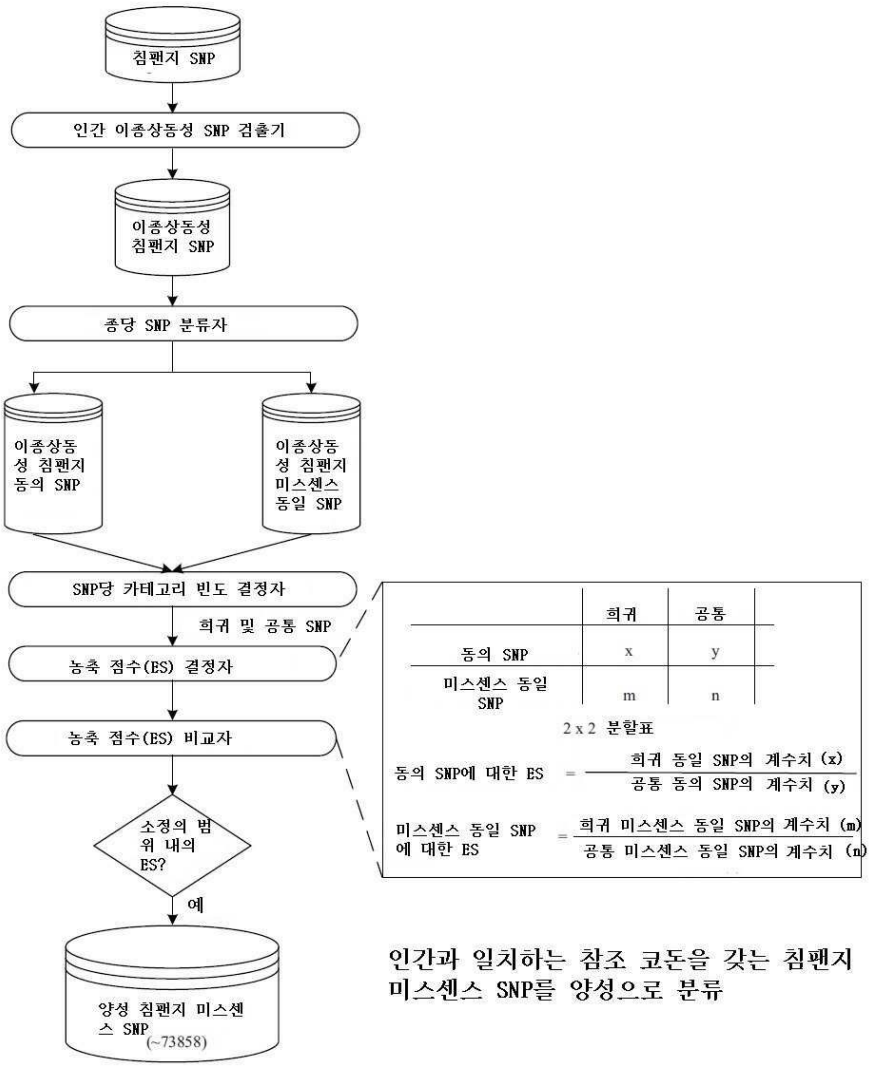

(8개의 하위집단에 걸친)
인간 SNP
종당 SNP 분류자
인간 미스센스 SNP를 식별
(8개의 하위집단에 걸친)
인간 미스센스 SNP
소수 대립유전자 빈도(MAF) 결정자
인간 모집단에서 MAF>0.1%인 인간 미스센스
SNP를 식별(공통 인간 미스센스 SNP)
양성 인간 미스센스 SNP
(~83546)
공통 인간 미스센스
SNP를 양성으로 분류

도면45



인간 이종상동성 미스센스 SNP: 인간과 일치하는 참조 및 대체 코돈을 갖는 비인간 종의 미스센스 SNP

도면46


침팬지 SNP
인간 이종상동성 SNP 검출기
이종상동성 침팬지 SNP
종당 SNP 분류자
이종상동성 침팬지 동의 SNP
이종상동성 침팬지 미스센스 동일 SNP
SNP당 카테고리 빈도 결정자
희귀 및 공통 SNP
농축 점수(ES) 결정자
농축 점수(ES) 비교자
소정의 범위 내의 ES?
예
양성 침팬지 미스센스 SNP (~73858)
희귀
공통
동의 SNP
x
y
미스센스 동일 SNP
m
n
2 x 2 분할표
동의 SNP에 대한 ES = 희귀 동일 SNP의 계수치 (x) / 공통 동의 SNP의 계수치 (y)
미스센스 동일 SNP에 대한 ES = 희귀 미스센스 동일 SNP의 계수치 (m) / 공통 미스센스 동일 SNP의 계수치 (n)
인간과 일치하는 참조 코돈을 갖는 침팬지 미스센스 SNP를 양성으로 분류

도면47

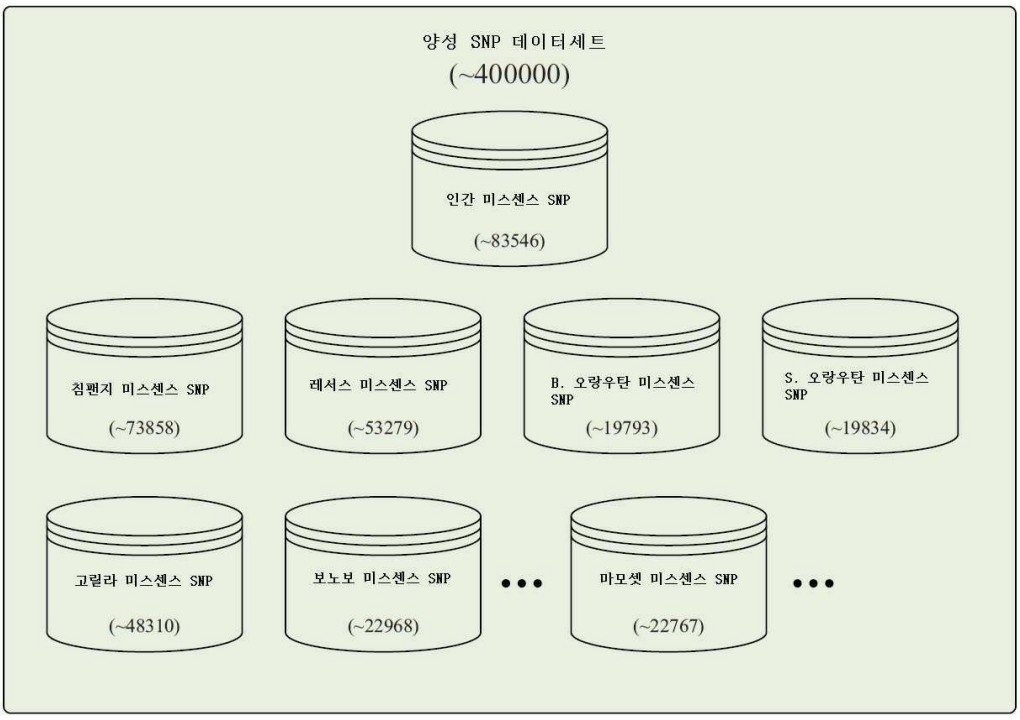
	희귀	공통
동의 SNP	x	y
미스센스 동일 SNP	m	n

2 x 2 분할표

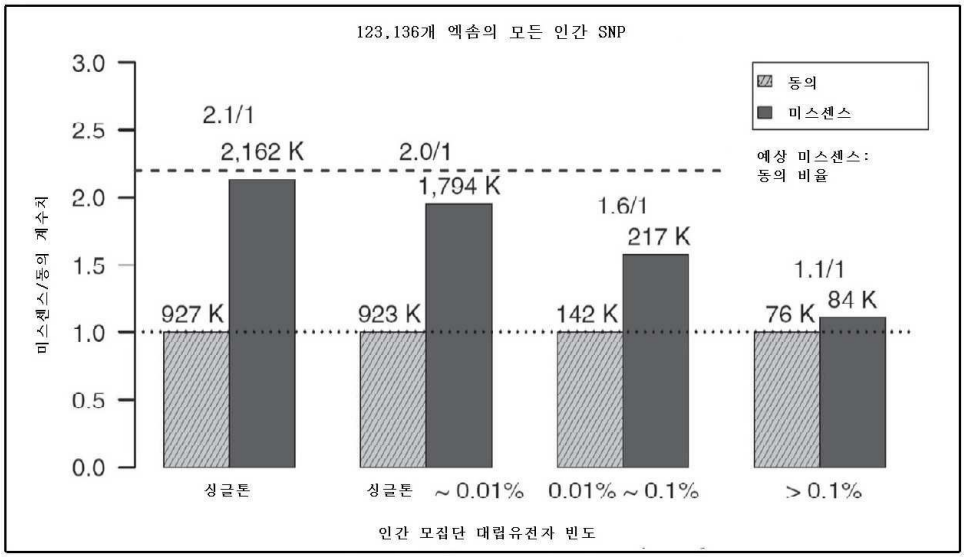
$$\text{동의 SNP에 대한 ES} = \frac{\text{희귀 동일 SNP의 계수치 (x)}}{\text{공통 동의 SNP의 계수치 (y)}}$$

$$\text{미스센스 동일 SNP에 대한 ES} = \frac{\text{희귀 미스센스 동일 SNP의 계수치 (m)}}{\text{공통 미스센스 동일 SNP의 계수치 (n)}}$$

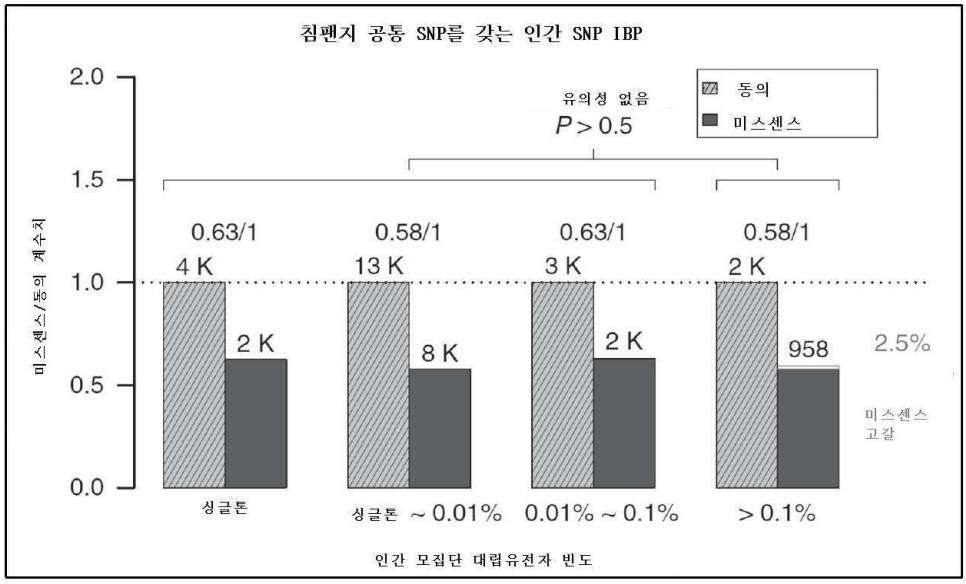
도면48



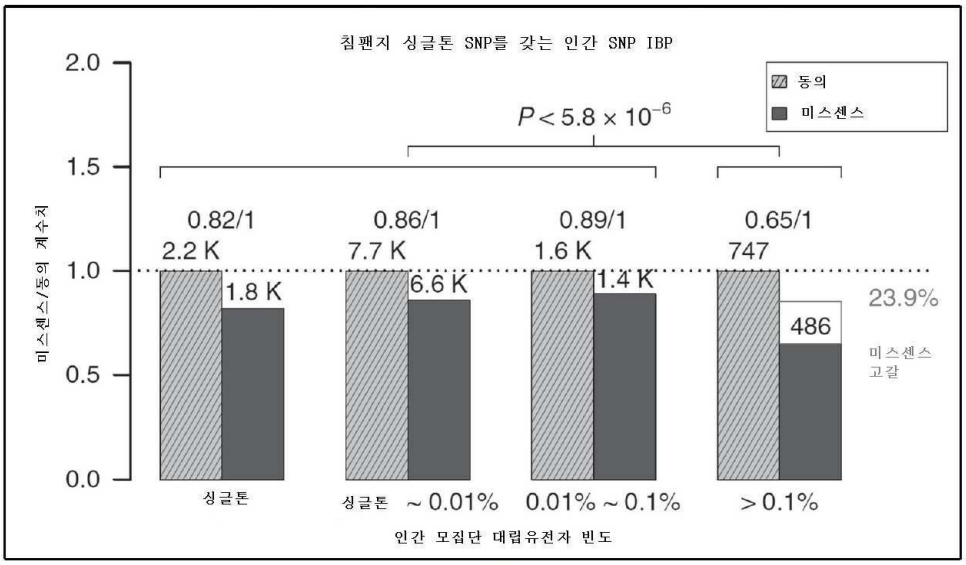
도면49a


123,136개 엑솜의 모든 인간 SNP
동의
미스센스
예상 미스센스:
동의 비율
미스센스/동의 계수치
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.1/1
2,162 K
927 K
2.0/1
1,794 K
923 K
1.6/1
217 K
142 K
1.1/1
84 K
76 K
싱글톤
싱글톤 ~0.01%
0.01% ~ 0.1%
> 0.1%
인간 모집단 대립유전자 빈도

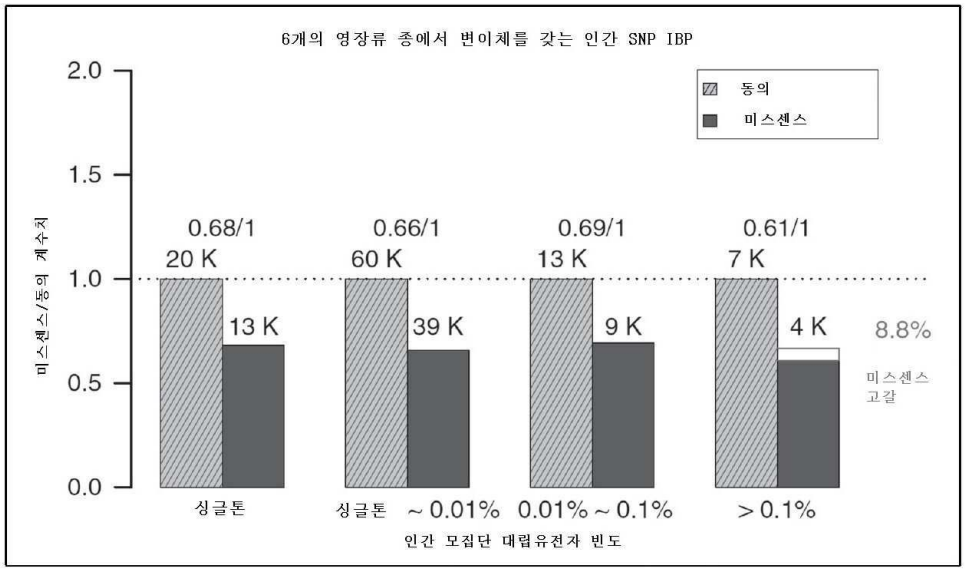
도면49b


침팬지 공통 SNP를 갖는 인간 SNP IBP
유의성 없음
P > 0.5
동의
미스센스
미스센스/동의 계수치
2.0
1.5
1.0
0.5
0.0
0.63/1
4 K
2 K
0.58/1
13 K
8 K
0.63/1
3 K
2 K
0.58/1
2 K
958
2.5%
미스센스
고갈
싱글톤
싱글톤 ~0.01%
0.01% ~ 0.1%
> 0.1%
인간 모집단 대립유전자 빈도

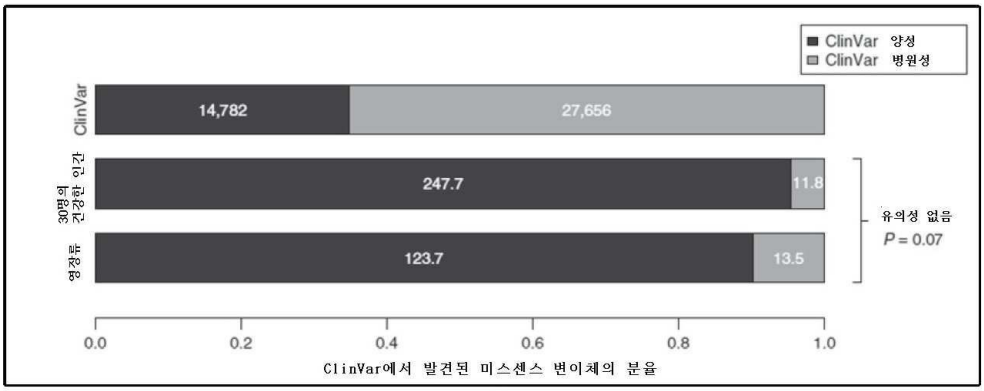
도면49c



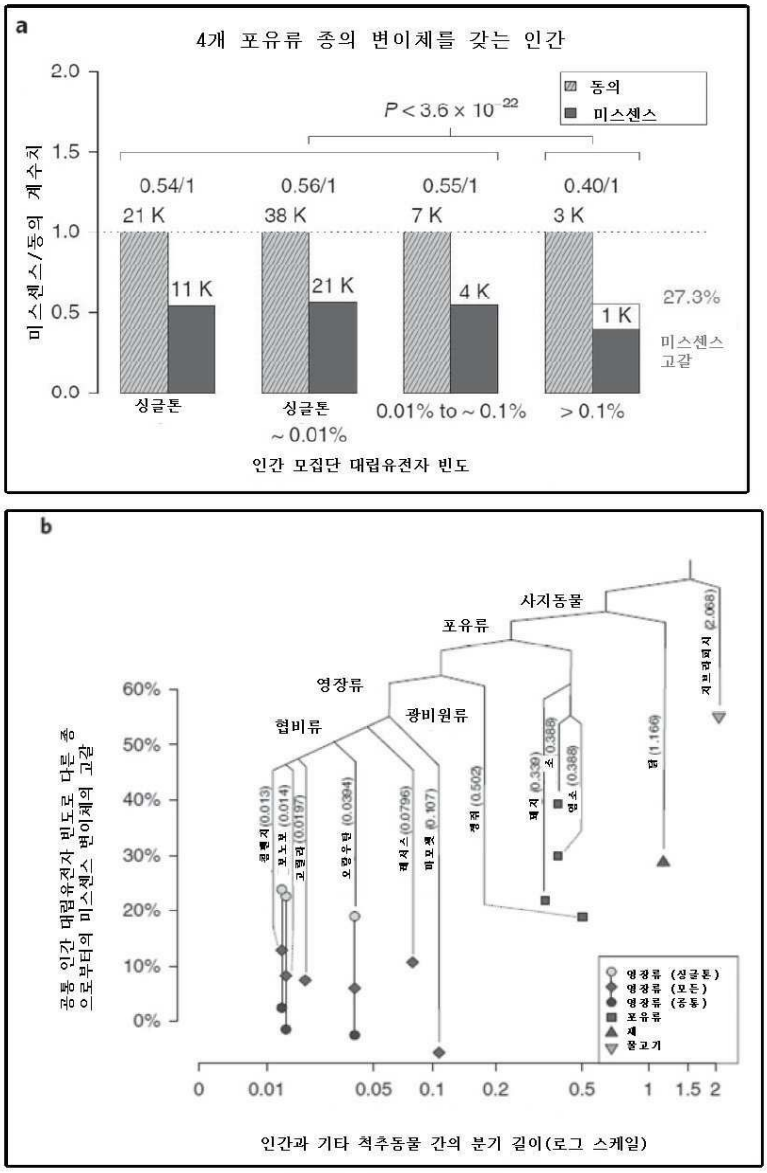
도면49d



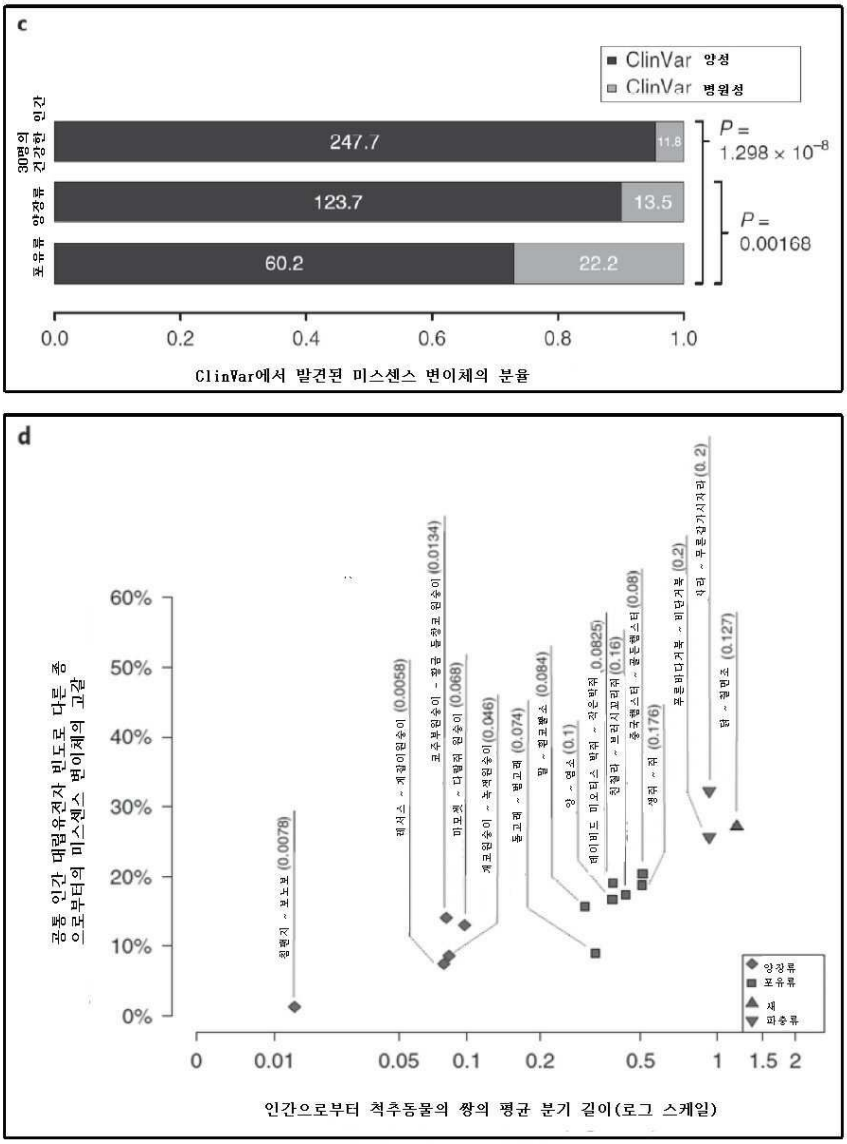
도면49e



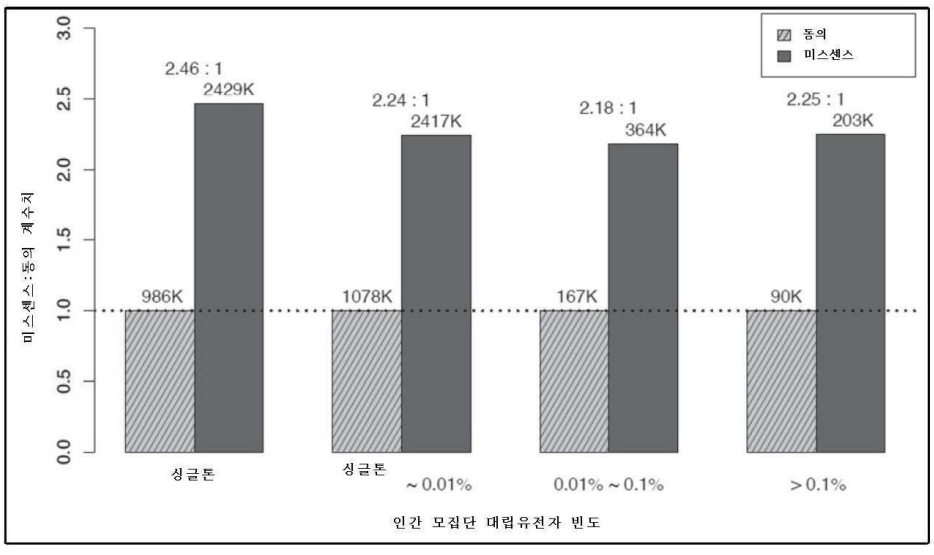
도면50ab


a
4개 포유류 종의 변이체를 갖는 인간
동의
미스센스
$P < 3.6 \times 10^{-22}$
미스센스/동의 계수치
2.0
1.5
1.0
0.5
0.0
0.54/1
21 K
11 K
0.56/1
38 K
21 K
0.55/1
7 K
4 K
0.40/1
3 K
1 K
27.3%
미스센스 고갈
싱글톤
싱글톤 ~0.01%
0.01% to ~0.1%
> 0.1%
인간 모집단 대립유전자 빈도
b
사지동물
포유류
영장류
광비원류
협비류
공통 인간 대립유전자 빈도로 다른 종 으로부터의 미스센스 변이체의 고갈
60%
50%
40%
30%
20%
10%
0%
0
0.01
0.05
0.1
0.2
0.5
1
1.5
2
영장류 (싱글톤)
영장류 (모든)
영장류 (공통)
포유류
새
황고기
인간과 기타 척추동물 간의 분기 길이(로그 스케일)

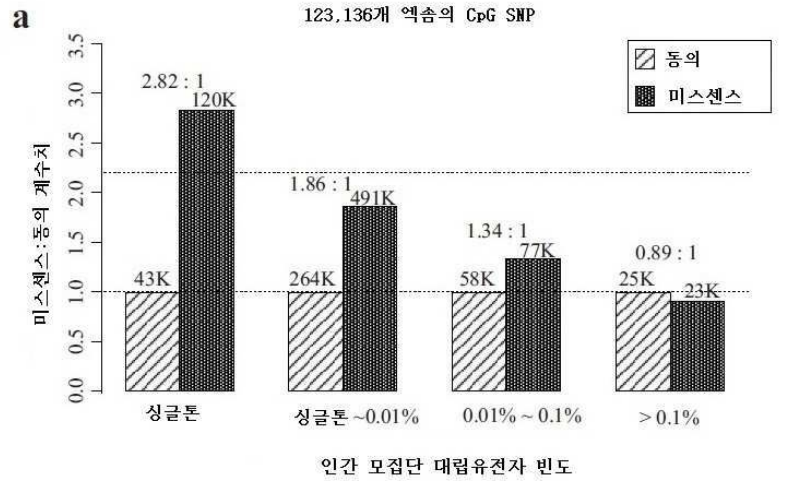
도면50cd

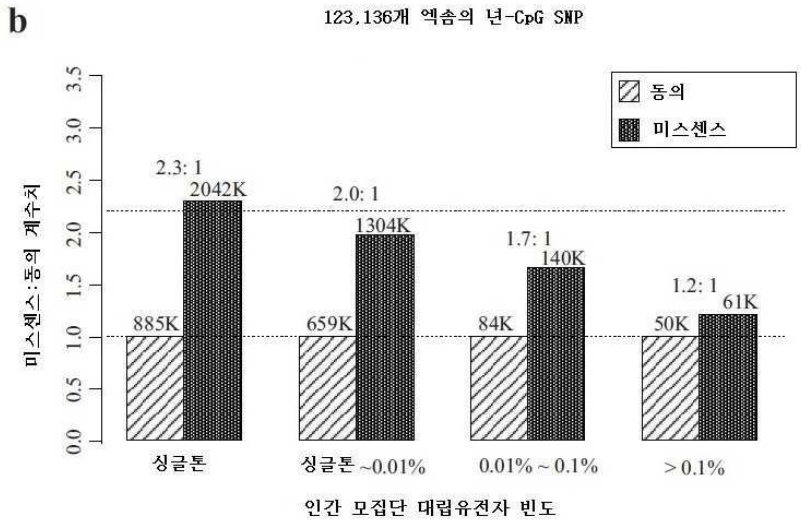

c
ClinVar 양성
ClinVar 병원성
30명의 건강한 인간
247.7
11.8
P = 1.298 × 10⁻⁸
포유류 양장류
123.7
13.5
P = 0.00168
60.2
22.2
0.0
0.2
0.4
0.6
0.8
1.0
ClinVar에서 발견된 미스센스 변이체의 분율
d
공통 인간 대립유전자 빈도로 다른 종으로부터의 미스센스 변이체의 고갈
60%
50%
40%
30%
20%
10%
0%
0
0.01
0.05
0.1
0.2
0.5
1
1.5
2
양장류
포유류
새
파충류
인간으로부터 척추동물의 쌍의 평균 분기 길이(로그 스케일)

도면51


동의
미스센스
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2.46 : 1
2429K
986K
2.24 : 1
2417K
1078K
2.18 : 1
364K
167K
2.25 : 1
203K
90K
싱글톤
싱글톤
~0.01%
0.01% ~ 0.1%
>0.1%
인간 모집단 대립유전자 빈도

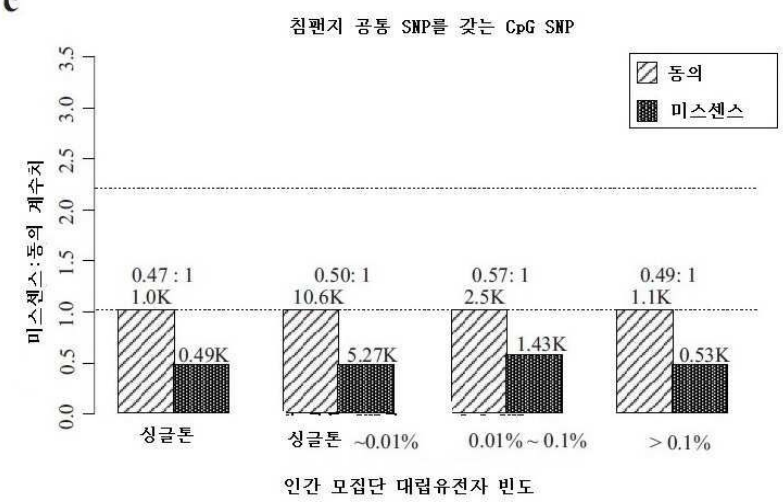
도면52ab


a
123,136개 엑솜의 CpG SNP
동의
미스센스
미스센스:동의 계수치
2.82 : 1
120K
43K
1.86 : 1
491K
264K
1.34 : 1
77K
58K
0.89 : 1
25K
23K
싱글톤
싱글톤~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도

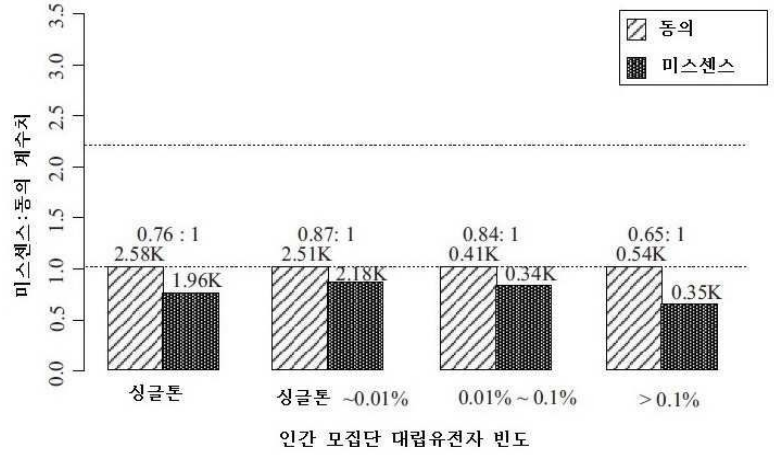

b
123,136개 엑솜의 넌-CpG SNP
동의
미스센스
미스센스:동의 계수치
2.3: 1
2042K
885K
2.0: 1
1304K
659K
1.7: 1
140K
84K
1.2: 1
61K
50K
싱글톤
싱글톤~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도

도면52cd

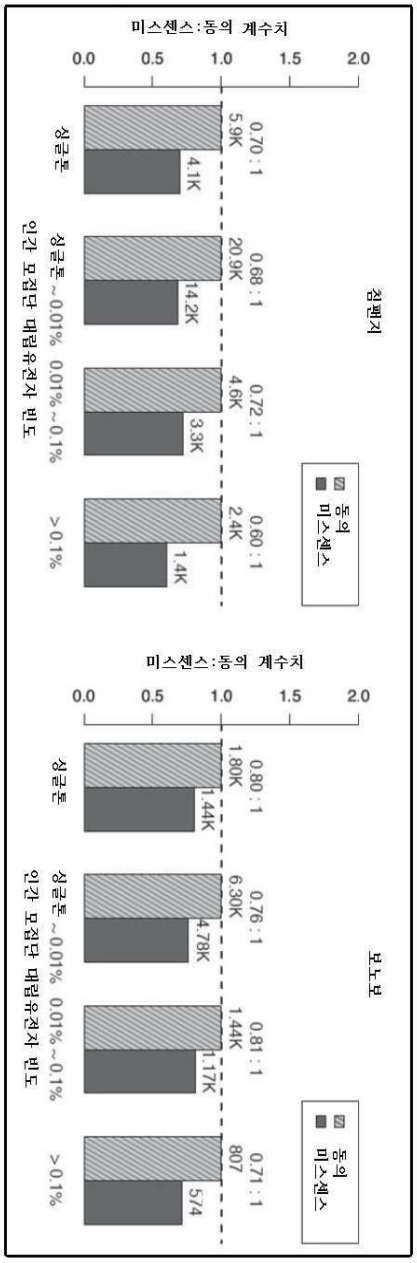
c


침팬지 공통 SNP를 갖는 CpG SNP
동의
미스센스
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.47 : 1
1.0K
0.49K
0.50: 1
10.6K
5.27K
0.57: 1
2.5K
1.43K
0.49: 1
1.1K
0.53K
싱글톤
싱글톤 ~0.01%
0.01% ~ 0.1%
> 0.1%
인간 모집단 대립유전자 빈도

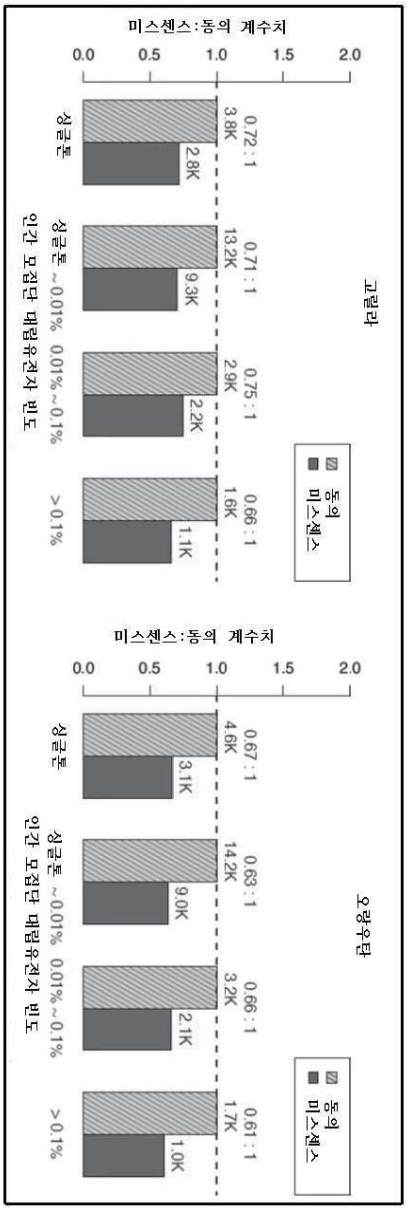
d


침팬지 공통 SNP를 갖는 넌-CpG SNP
동의
미스센스
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.76 : 1
2.58K
1.96K
0.87: 1
2.51K
2.18K
0.84: 1
0.41K
0.34K
0.65: 1
0.54K
0.35K
싱글톤
싱글톤 ~0.01%
0.01% ~ 0.1%
> 0.1%
인간 모집단 대립유전자 빈도

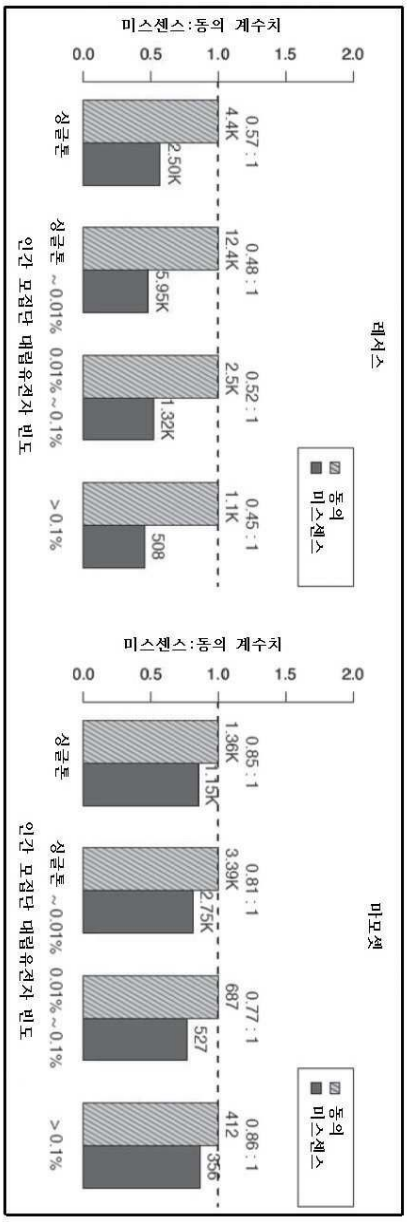
도면53


미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
싱글톤~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도
5.9K
4.1K
0.70 : 1
20.9K
14.2K
0.68 : 1
4.6K
3.3K
0.72 : 1
2.4K
1.4K
0.60 : 1
동의
미스센스
침팬지
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
싱글톤~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도
1.80K
1.44K
0.80 : 1
6.30K
4.78K
0.76 : 1
1.44K
1.17K
0.81 : 1
807
574
0.71 : 1
동의
미스센스
보노보

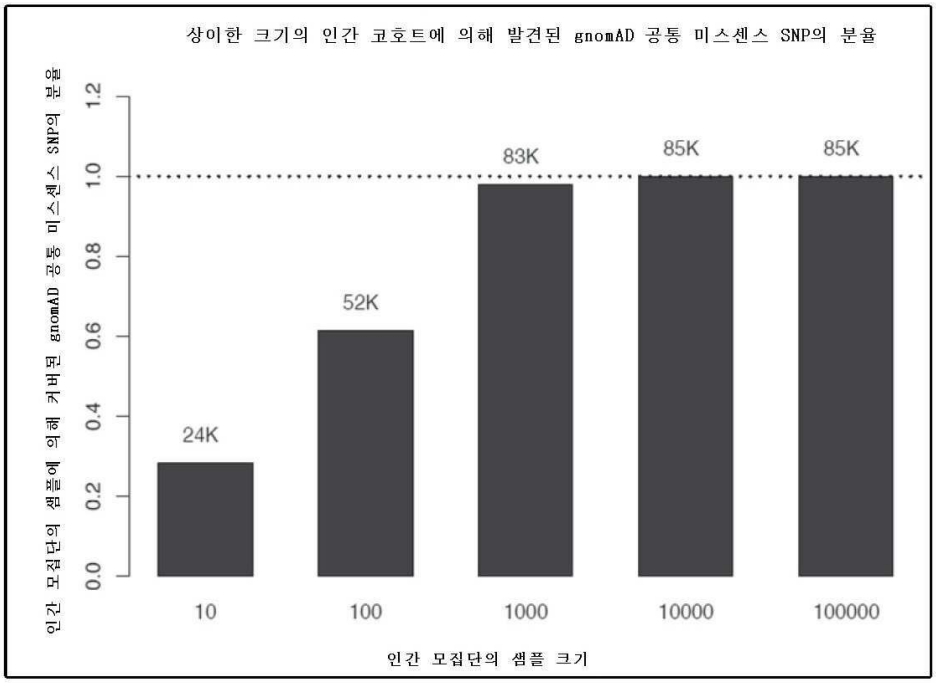
도면54


미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
싱글톤 ~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도
3.8K
2.8K
0.72 : 1
13.2K
9.3K
0.71 : 1
2.9K
2.2K
0.75 : 1
1.6K
1.1K
0.66 : 1
동의
미스센스
고릴라
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
싱글톤 ~0.01%
0.01%~0.1%
>0.1%
인간 모집단 대립유전자 빈도
4.6K
3.1K
0.67 : 1
14.2K
9.0K
0.63 : 1
3.2K
2.1K
0.66 : 1
1.7K
1.0K
0.61 : 1
동의
미스센스
오랑우탄

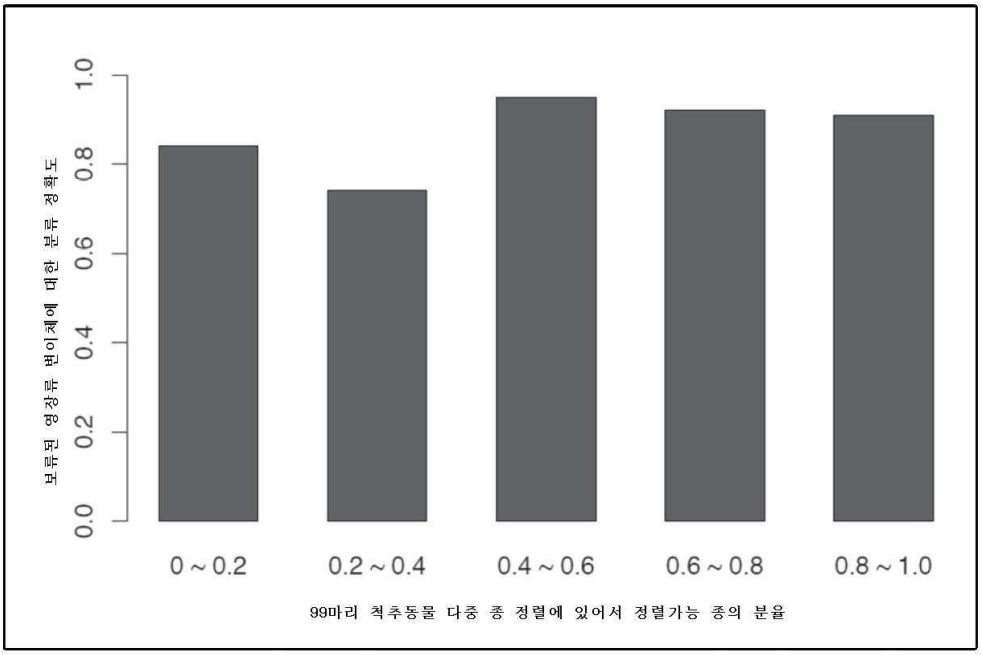
도면55


레서스
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
0.57 : 1
4.4K
2.50K
싱글톤 ~0.01%
0.48 : 1
12.4K
5.95K
0.01%~0.1%
0.52 : 1
2.5K
1.32K
>0.1%
0.45 : 1
1.1K
508
동의
미스센스
인간 모집단 대립유전자 빈도
마모셋
미스센스:동의 계수치
0.0
0.5
1.0
1.5
2.0
싱글톤
0.85 : 1
1.36K
1.15K
싱글톤 ~0.01%
0.81 : 1
3.39K
2.75K
0.01%~0.1%
0.77 : 1
687
527
>0.1%
0.86 : 1
412
356
동의
미스센스
인간 모집단 대립유전자 빈도

도면56


상이한 크기의 인간 코호트에 의해 발견된 gnomAD 공통 미스센스 SNP의 분율
인간 모집단의 샘플에 의해 커버된 gnomAD 공통 미스센스 SNP의 분율
0.0
0.2
0.4
0.6
0.8
1.0
1.2
24K
52K
83K
85K
85K
10
100
1000
10000
100000
인간 모집단의 샘플 크기

도면57


보류된 영장류 변이체에 대한 분류 정확도
0.0
0.2
0.4
0.6
0.8
1.0
0 ~ 0.2
0.2 ~ 0.4
0.4 ~ 0.6
0.6 ~ 0.8
0.8 ~ 1.0
99마리 척추동물 다중 종 정렬에 있어서 정렬가능 종의 분율

도면58

종(및 개체 수)	공급원	인간 코돈과 일치하는 미스센스 변이체의 수	기여한 고유한 미스센스 변이체의 수
인간 공통 변이체 (MAF > 0.1%, 123136 개체)	gnomAD / ExAC database	83,546	83,546
침팬지 (24)	Prado-Martinez *et al.*, Nature (2013)	76,293	73,858
보노보 (13)	Prado-Martinez *et al.*, Nature (2013)	25,277	22,968
고릴라 (27)	Prado-Martinez *et al.*, Nature (2013)	52,052	48,310
오랑우탄, 보르네오 아종 (5)	Prado-Martinez *et al.*, Nature (2013)	21,621	19,793
오랑우탄, 수마트라 아종 (5)	Prado-Martinez *et al.*, Nature (2013)	27,567	19,834
오랑우탄(개체 데이터는 이용 불가) (35)	de Manuel *et al.*, Science (2016)	75,580	30,327
레서스(16, 개체 데이터는 이용불가)	dbSNP	18,627	10,554
마모셋(9, 개체 데이터는 이용불가)	dbSNP	62,393	53,279
총	dbSNP	25,048	22,767
			385,236

도면59

123,136개 엑솜의 인간 모집단 대립유전자 빈도	동의 변이체의 예상 수	미스센스 변이체의 예상 수	미스센스 :동의 비율
싱글톤	986,141.22	2,429,287.85	2.46
싱글톤 ~0.01%	1,077,630.37	2,416,751.61	2.24
0.01%~0.1%	166,849.26	364,292.75	2.18
>0.1%	90,221.42	202,918.38	2.24

도면60

	관심 ClinVar 변이체의 총 수
초기 ClinVar 파일	666,403
hg19 레코드만을 유지	324,698
단일 뉴클레오타이트 변이체만을 추출	264,623
넌-코딩 변이체를 제거	186,769
중단 코돈을 생성하거나 방해하는 모든 동의 변화 및 미스센스 변화를 무시	122,884
알려지지 않은 유의성의 변이체 및 주석이 충돌하는 변이체를 제외	42,438

도면61

종	CLinVar에서 양성 주석이 있는 변이체의 수	CLinVar에서 병원성 주석이 있는 변이체의 수
침팬지	218	21
보노보	85	7
고릴라	167	23
오랑우탄	160	12
레서스	87	11
마모셋	25	7
생쥐	30	4
돼지	74	30
소	77	39
염소	60	16
닭	9	8
지브라피시	2	6

도면62

표 1. PrimateAI 점수≥0.803인 미스센스 드 노보 돌연변이(DNM)만을 고려할 때 지적 장애에 있어서 게놈 전체 유의성을 달성하는 추가 유전자

HGNC 부호	단백질 전달 변이체	미스센스		P 값		다수 개체에서 관찰되는 표현형 이상
		PrimateAI 점수 ≥0.803	모든 미스센스	PrimateAI 점수 ≥0.803	모든 미스센스	
ACTL6B	0	3	3	1.5×10^{-7}	2.4×10^{-6}	소두증
EBF3	3	3	3	5.2×10^{-8}	5.4×10^{-6}	성장 지연증, 눈 이상, 사시, 운동실조
EFTUD2	2	4	4	1.5×10^{-7}	1.5×10^{-5}	소두증, 처진 귀, 소이증, 후비공폐쇄증
HECW2	1	8	8	2.8×10^{-10}	6.7×10^{-7}	발작, 근질환, 비정상적 두개관
KDM6A	2	3	3	2.3×10^{-7}	9.8×10^{-6}	눈꺼풀, 치아 이상, 근력저하
KIF5C	0	3	3	3.0×10^{-7}	2.8×10^{-6}	뇌 형성 부전
MAP2K1	0	5	5	3.1×10^{-8}	2.7×10^{-6}	격리증, 처진 귀, 양수과다증
PPP1CB	0	6	6	1.5×10^{-8}	1.6×10^{-6}	이마 비정상, 단신
PRKD1	0	6	6	8.6×10^{-8}	1.7×10^{-5}	피부, 디지털, 및 심장 이상, 드문 머리털
SOX11	1	3	3	3.1×10^{-7}	2.4×10^{-5}	원시, 손톱 형성 부전
TBR1	4	4	4	1.3×10^{-10}	4.2×10^{-7}	자폐증 행동
TLK2	3	5	5	4.7×10^{-9}	6.3×10^{-7}	코, 눈꺼풀 비정상, 경사진 눈꺼풀 틈
TRIP12	6	2	4	1.4×10^{-7}	5.4×10^{-7}	관절 이완
U2AF2	0	4	4	2.6×10^{-7}	1.2×10^{-5}	발작, 눈, 구개음, 인중 비정상

단백질 전달 및 미스센스 DNM의 카운트를 제공한다. PrimateAI 점수≥0.803인 미스센스 돌연변이만을 이용해서 통계적 테스트가 실행되었을 때 그리고 모든 미스센스 돌연변이에 대하여 반복되었을 때 유전자 농축에 대한 P값이 도시되어 있다.

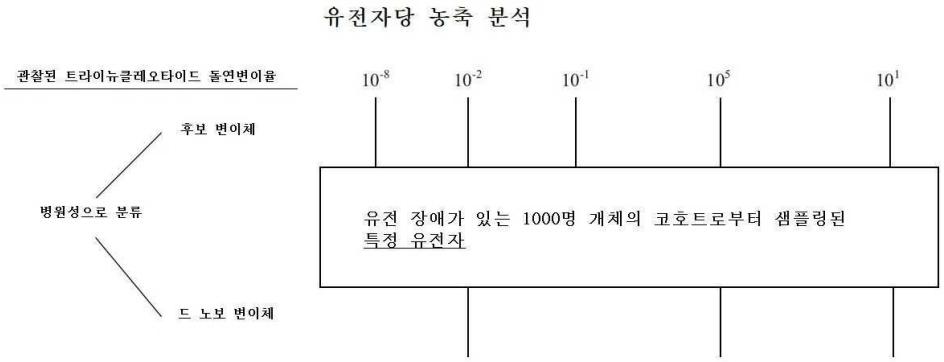
도면63

표 2. DDD 사례 대 대조군에 있어서 그랜덤 점수의 차, 단백질 표면-노출, 및 ClinVar의 인간 전문가에 의한 주석이 있는 변이체와 드 노보 변이체 간의 아미노산 서열 보존의 비교

	그랜덤 점수	단백질 표면-노출	서열 보존
ClinVar 병원성 변이체	91.1	0.53	0.87
ClinVar 양성 변이체	67.4	0.41	0.54
인간 전문가의 주석의 차	**+23.7**	+0.12	**+0.33**
DDD 환자의 드 노보 변이체	84.9	0.51	0.90
건강한 대조군의 드 노보 변이체	72.7	0.29	0.73
영향을 받은 개체 대 영향을 받지 않은 개체 간의 차	+12.2	**+0.22**	+0.17

ClinVar 데이터베이스에서 비충돌 주석을 갖는 미스센스 돌연변이에 대한 그리고 605개의 질환 연관 유전자 내에서 DDD 사례 대 대조군에서 존재하는 드 노보 변이체에 대한 평균 점수가 도시되어 있다. 단백질 표면-노출은 용매 접근성 신경망에 의해 노출된 잔기로서 예측되는 아미노산의 분율을 반영하며, 서열 보존은, 100마리 척추동물의 정렬에 있어서 서열 동질성을 갖는 아미노산의 분율을 나타낸다. 굵은 수치는, 경험적 데이터에 비교되는 인간 전문가가 선호하는 휴리스틱에서의 차를 강조한다.

도면64

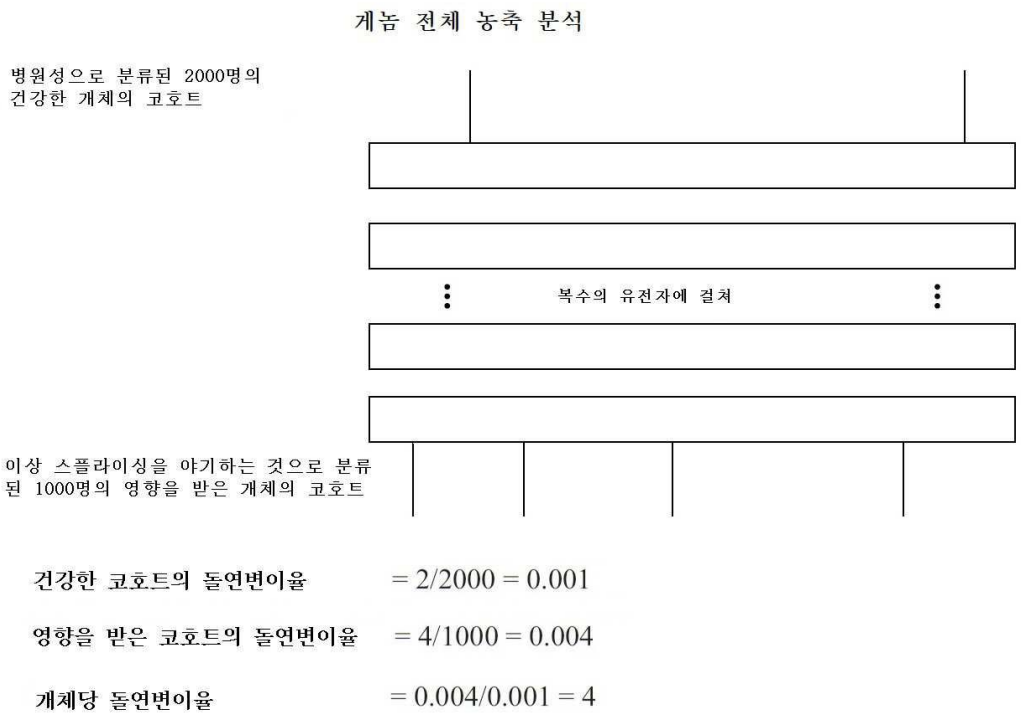


$$\sum(\text{관찰된 트라이뉴클레오타이드 돌연변이율}) = 10^{-5}$$

$$\text{돌연변이의 베이스라인 수} = \underbrace{2}_{\text{염전(염색체) 계수치}} \times 1000 \times \underbrace{10^{-5}}_{\text{코호트 크기}} = 0.2$$

드 노보 변이체 계수치 = 3

도면65



건강한 코호트의 돌연변이율 = 2/2000 = 0.001

영향을 받은 코호트의 돌연변이율 = 4/1000 = 0.004

개체당 돌연변이율 = 0.004/0.001 = 4

도면66

