KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)
Workshop for Life Scientists, Data Scientists,
and Bioinformatians

생물정보학&머신러닝 워크샵(온라인)

Noncoding variants and deep learning

이현주 _ GIST

SBi 한국생명정보학회
Korean Society for Bioinformatics

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2023

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists, Data Scientists, and Bioinformatians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의가 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크샵은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

강의개요

Noncoding variants and deep learning

악성종양 등의 복합 질환 환자의 DNA를 시퀀싱을 했을 때, 넌코딩 영역에서 많은 변이(noncoding variant)가 관찰되고 있다. Noncoding variants가 유전자의 발현이나 질병의 진행에 미치는 영향에 대한 연구는 질병을 이해하고, 이를 치료하기 위한 타겟을 선정하는데 중요하다. 최근에는 DNA 시퀀스에 기반하여 noncoding variant의 기능적 영향을 예측하기 위한 다양한 딥 러닝에 기반 방법론들이 개발되고 있다.

본 강의에서는 noncoding variant의 기능적 영향을 예측하기 위한 딥 러닝 방법론들을 소개하고, 이러한 방법론을 환자의 DNA 시퀀스에 적용하여, 질병 관련된 유전자들을 발굴한 연구들을 살펴본다. 본 강의를 통해서 DNA 시퀀스에 적용된 딥 러닝 기반 방법론들과 이를 생물학 지식으로 변환하는 연구들을 이해하는 것을 목표로 한다.

- Noncoding variants의 개요
- 딥 러닝 방법론의 DNA 시퀀스 적용
- Noncoding variants의 기능적 영향 예측
- 질병 관련 변이 예측 연구

* 강의 난이도: 중급

* 강의: 이현주 교수 (광주과학기술원 전기전자컴퓨터공학부)

Curriculum Vitae

Speaker Name: Hyunju Lee, Ph.D.

▶ **Personal Info**

Name	Hyunju Lee
Title	Professor
Affiliation	Gwangju Institute of Science and Technology

▶ **Contact Information**

Address	123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005
Email	hyunjulee@gist.ac.kr
Phone Number	062-715-2213

Research Interest

Bioinformatics, Machine learning, and Text Mining

Educational Experience

1997	B.S. in Computer Science, KAIST, South Korea
1999	M.A. in Computer Engineering, Seoul National University, South Korea
2006	Ph.D. in Computer Science, University of Southern California, USA

Professional Experience

2006-2007	Post-doc Research Fellow, Brigham and Women's Hospital and Harvard Medical School, USA
2007-	Full-time lecturer, Assistant, Associate, Full Professor, Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology

Selected Publications (5 maximum)

1. Yeonghun Lee and Hyunju Lee. Integrative reconstruction of cancer genome karyotypes using InfoGenomeR. Nature Communications, 12:2467, 2021.
2. Ho Jang and Hyunju Lee, Multiresolution correction of GC bias and application to identification of copy number alterations, Bioinformatics, 35(20), 2019.
3. Jeongkyun Kim, Jung-jae Kim, and Hyunju Lee, DigChem: Identification of disease-gene-chemical relationships from Medline abstracts, PLoS Computational Biology 15(5), 2019.
4. Jihee Soh, Hyejin Cho, Chan-Hun Choi, and Hyunju Lee, Identification and Characterization of MicroRNAs Associated with Somatic Copy Number Alterations in Cancer, Cancers, 10(12):475, 2018.
5. Bayarbaatar Amgalan and Hyunju Lee, DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method, Bioinformatics, 31(15), 2015.

SBi 한국생명정보학회 Korean Society for Bioinformatics

KSBi-BIML

Noncoding Variants and Deep Learning

이현주 (광주과학기술원 전기전자컴퓨터공학부)

Contents

- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants

Genomic variants

- Protein-coding regions make up around 1% of the human genome
- ENCODE suggests (Nature 489, 57–74 (2012))
  - 82% of the human genome was functionally important having biochemical activity.
  - ~20 % of the genome is associated with DNase hypersensitivity or transcription factor binding (common features for identifying regulatory region)
- How coding and noncoding variation can impact gene function

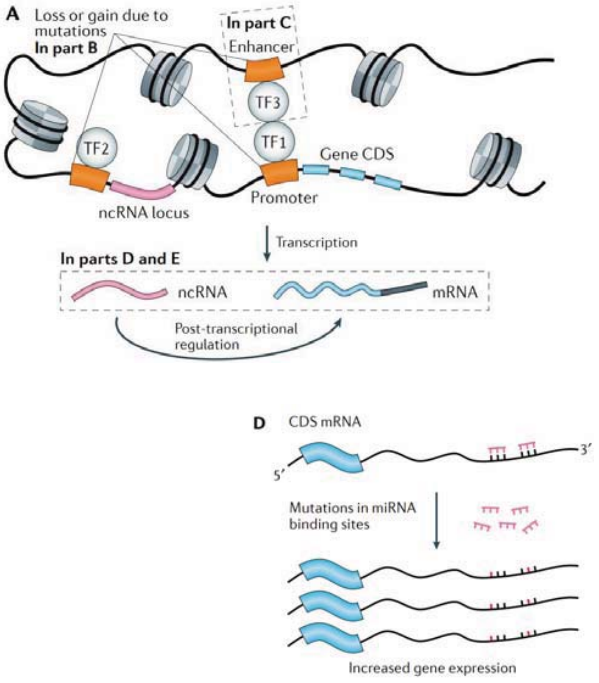
Variant Location	Transcript Map	Transcript Product	Transcript description	Potential Outcome
Coding (standard interpretation)			Synonymous/ Missense/ Nonsense	Homeostasis/ Altered Product/ Loss of function
Promoter/Enhancer/ Looping/cis-regulatory lncRNA			Over/ Under expression	Aberrant expression patterns
Splice Donor/Acceptor Branchpoint			Skipped exon/ Retained intron	Altered product Nonsense Mediated Decay

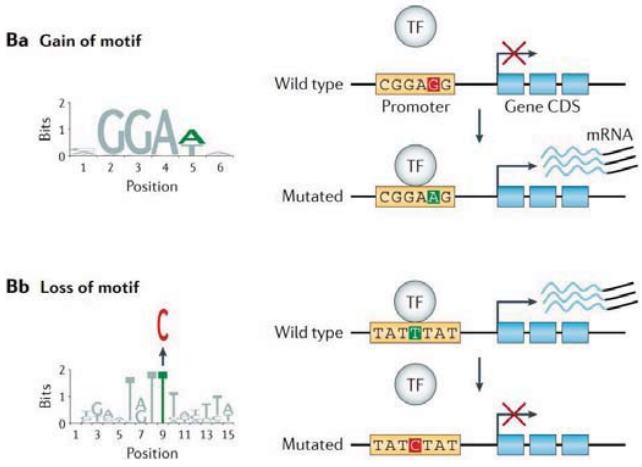
SBi 한국생명정보학회 Korean Society for Bioinformatics

Gloss and Dinger Experimental & Molecular Medicine (2018) 50:97

Noncoding variants

- Mutations in noncoding variants can lead to gain or loss of transcription





SBi 한국생명정보학회 Korean Society for Bioinformatics

Nature Reviews Genetics volume 17, pages93–108(2016)

Coding vs. noncoding variants

- Prediction of the effect of a coding variant on protein function
  - 'sorting tolerant from intolerant' (SIFT) algorithm
  - 'polymorphism phenotyping' (PolyPhen) tool
  - Protein sequences have been highly conserved throughout evolution
  - Based on a multiple-sequence alignment

- Regulatory elements
  - Conservation is a less important signal when interpreting variants
  - Effects of regulatory variants have quantitative rather than qualitative effects on gene expression
  - Same variant may have a larger or smaller effect in different tissues, at different developmental stages and even in different individuals.

Nature Methods volume 11, pages294–296(2014)

Contents

Computational methods to prioritize non-coding variants with functional effects

Tool	Year	Method used to build model
CADD	2014	Support vector machine
GWAVA	2014	Random forest algorithm
DeepSEA	2015	Deep learning, CNN
DanQ	2016	Deep learning, CNN, RNN
DeFine	2018	Deep learning, CNN
DeepFun	2021	Deep learning, CNN

Machine learning model (GWAVA)

- GWAVA: Genome-wide annotation of variants
  - Prioritization of noncoding variants by integrating various genomic and epigenomic annotations
  - https://www.sanger.ac.uk/tool/gwava/

Various genomic and epigenomic information → Modified Random Forest classifier → Binary classification (Disease-implicated SNVs vs. control SNVs)

(SNVs : single-nucleotide variants)

Nature Methods volume 11, pages294–296(2014)

Machine learning model (GWAVA)

- Disease-implicated SNVs
  - All variations annotated as 'regulatory mutations' from the public release of the Human Gene Mutation database (HGMD)
- Control sets
  - Common (minor allele frequency ≥1%) SNVs from the 1000 Genomes Project (1KG)
  - First set: a random selection of SNVs from across the genome in order to sample overall background.
  - Second set: matched for distance to the nearest TSS genome-wide.
    - HGMD variants are not distributed randomly across the genome; 75% lie within a 2 kilobase (kb) window around an annotated transcription start site (TSS)
  - Third set: all 1KG variants in the 1 kb surrounding each of the HGMD variants.

Nature Methods volume 11, pages294–296(2014) 9

Machine learning model (GWAVA)

- Genomic and epigenomic annotations
  - Open chromatin: DNase-seq data from ENCODE
  - Transcription factor binding: ChIP-seq peak calls for 124 TFs from ENCODE
  - Histone modifications: ChIP-seq peak calls for 12 modifications from ENCODE
  - RNA polymerase binding: ChIP-seq peak calls from ENCODE
  - CpG islands: Predictions from Ensembl
  - Genome segmentation: discrete states such as transcription start sites, gene ends, enhancers, transcriptional regulator CTCF-binding regions and repressed regions
  - Conservation: Genomic evolutionary rate profiling (GERP) scores from mammalian alignments
  - Human variation: Variants and allele frequencies 1000 Genomes Project phase 1 data
  - Genic context: distance from any base annotated as exonic, intronic, coding sequence, 5' or 3' untranslated region, splice site, or start or stop codon in any transcript

Nature Methods volume 11, pages294–296(2014) 10

Machine learning model (GWAVA)

- Genomic and epigenomic annotations
  - A large matrix with a row for each variant locus and a column for each possible annotation.
  - The column type depending on the annotation class

    (i) the number of cell lines in which the variant locus overlaps some annotation, such as DNase I hypersensitive sites and ChIPseq peaks

    (ii) a present-absent binary flag
    - Ex) whether this region is ever in an annotated intron

    (iii) a continuous value for genome-wide annotations
    - Ex) conservation and distance to the nearest TSS

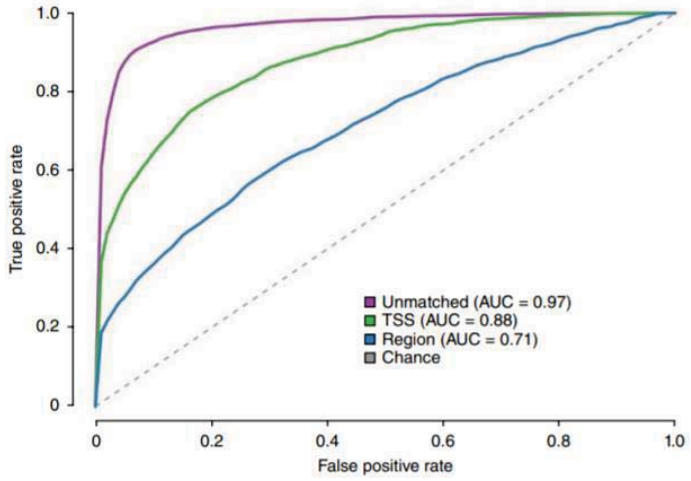
A part of example annotations

	chr	end	start	DNase	E2F1	H3K27ac	H3K27me3	cpg_island	gerp	tss_dist	...	TSS	INTRON	STOP	UTR3	...
rs111626726	chr3	1.5E+08	1.5E+08	12	0	12	1	1	3.18	447	...	6	1	0	0	...

Nature Methods volume 11, pages294–296(2014) 11

Machine learning model (GWAVA)

- A modified version of the random forest algorithm
- Three classifiers using all available annotations to discriminate between the disease variants and variants from each of the three control sets

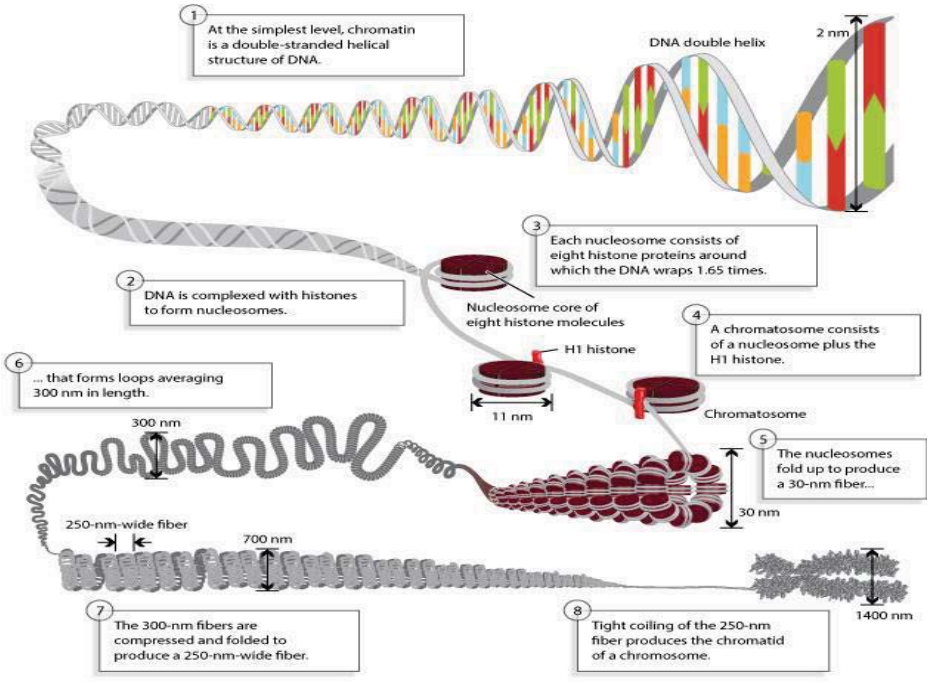


Nature Methods volume 11, pages294–296(2014) 12

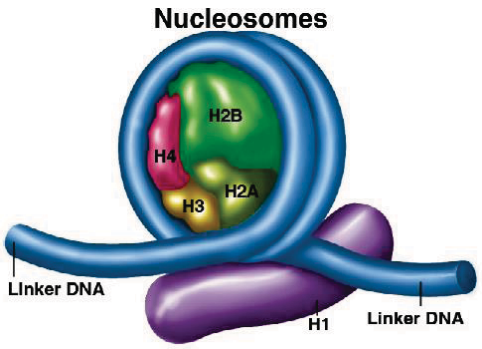
Contents

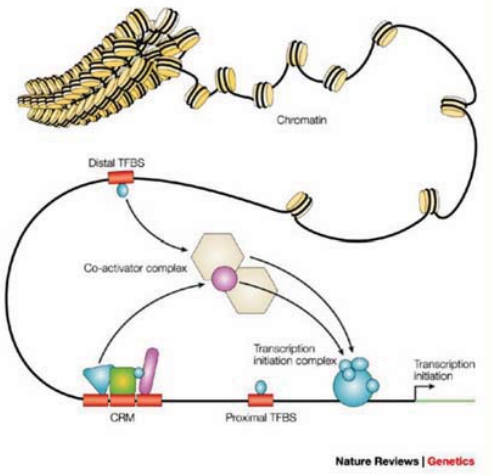
- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants

Chromosomes are composed of DNA tightly-wound around histones



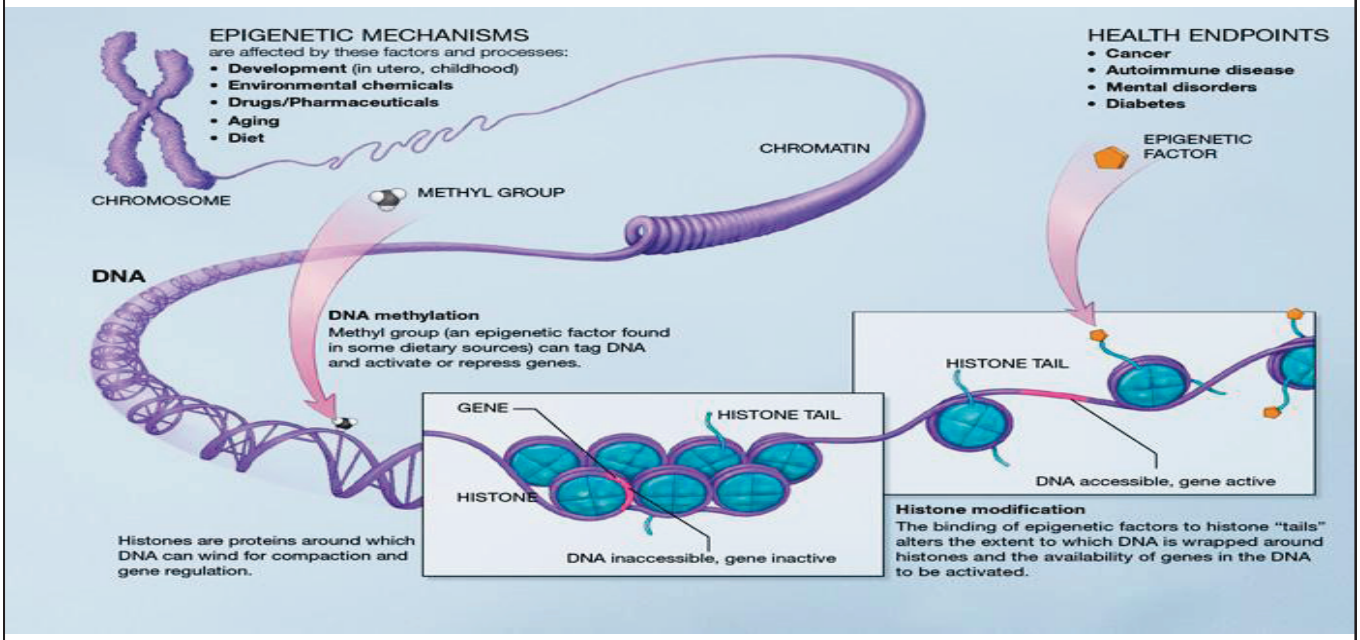
Histone and transcription





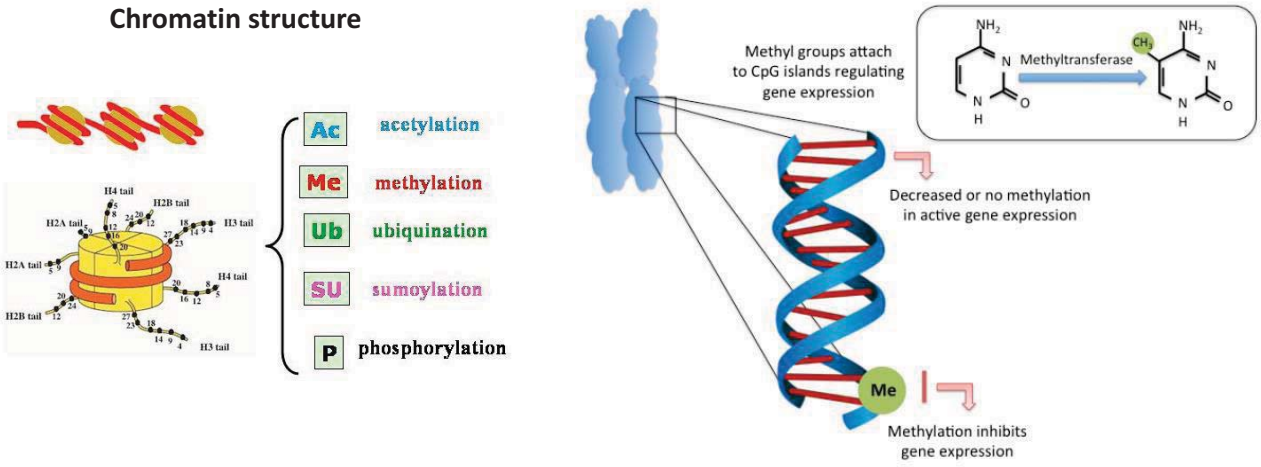
Histone proteins need to be modified and DNA needs to be released for transcription to take place.

Epigenetic mechanisms



http://commonfund.nih.gov/epigenomics/figure.aspx

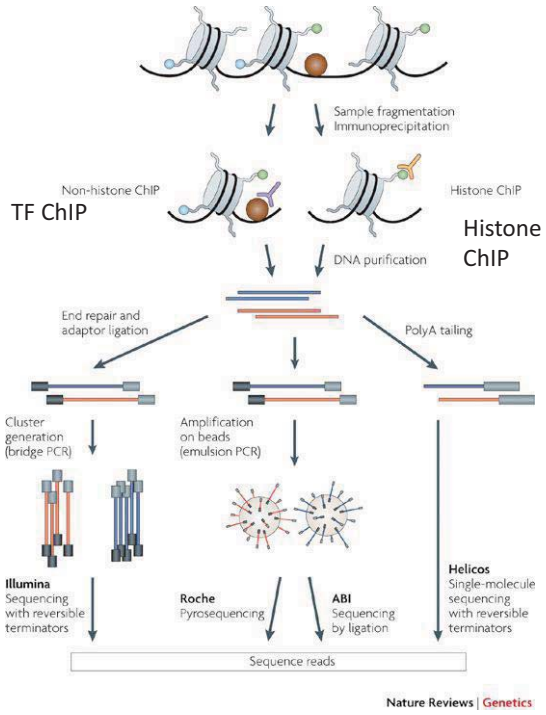
Histone modification DNA methylation



Nevin C and Carroll M, J Hum Genet Clin Embryol 2015, 1:004

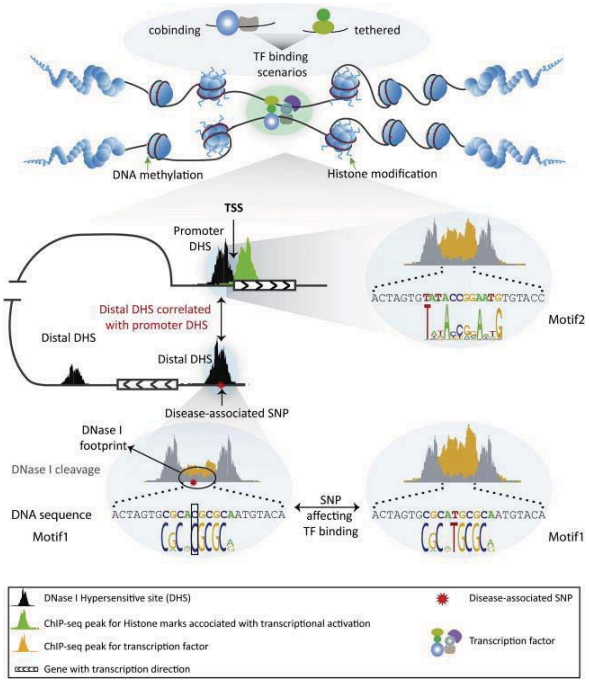
Chromatin Immunoprecipitation

- Chromatin Immunoprecipitation (ChIP): a technique that permits to “freeze” the protein-DNA bonds inside the cell nucleus, and the extraction of the DNA bound by a specific protein
- Antibodies are used to select specific proteins or nucleosomes which enriches for DNA-fragments that are bound to these proteins or nucleosomes
- Selected fragments can be either hybridized to a microarray (**ChIP-chip)** or sequenced on modern NGS platform **(ChIP-seq).**
- Thus, we can extract DNA bound in vivo by
  - Modified histones
  - Specific transcription factors
  - RNA Pol II



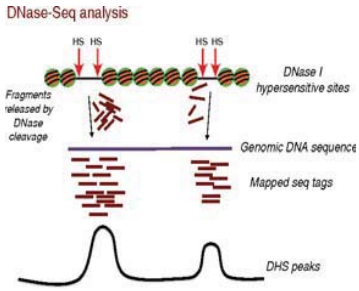
Regulation

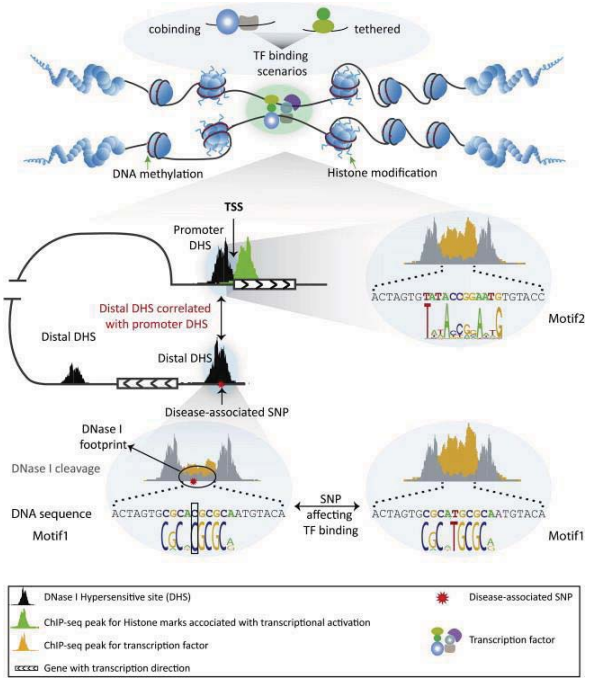
- Transcription factors (TFs)
  - Regulate gene transcription by binding to specific DNA elements such as promoters, enhancers, silencers.



Regulation

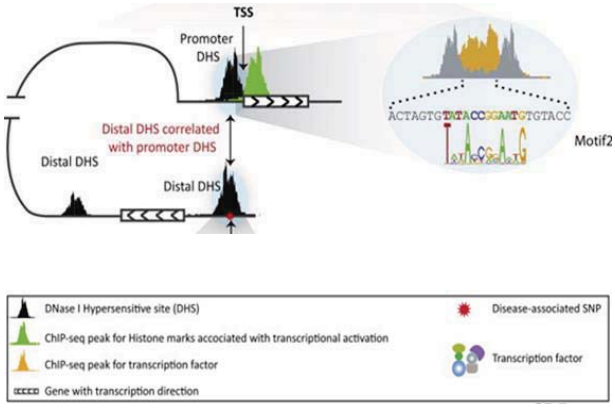
- Chromatin accessibility
  - Hallmark of regulatory DNA regions
  - characterized by DNase I hypersensitivity (DHS)
  - DHSs are regions of chromatin that are sensitive to cleavage by the DNase I enzyme.





DNase I footprinting

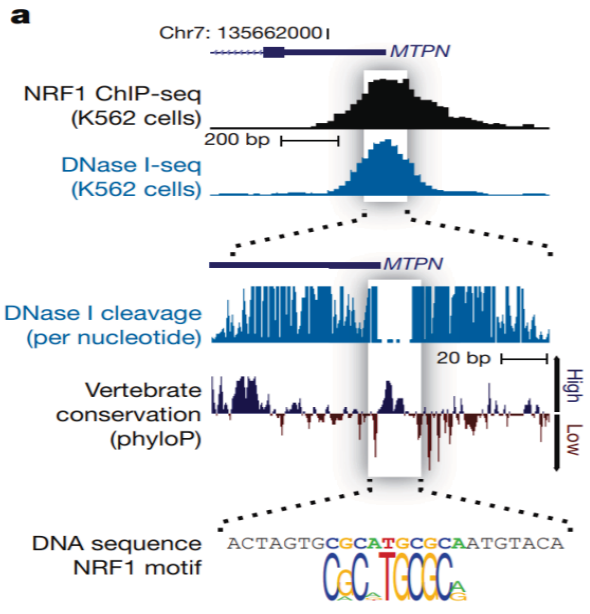
- DNase I footprinting detects DNA sequences that are protected from cleavage by DNaseI because they are bound by regulatory factors.



Genomics Proteomics Bioinformatics 11 (2013) 135 - 141 21

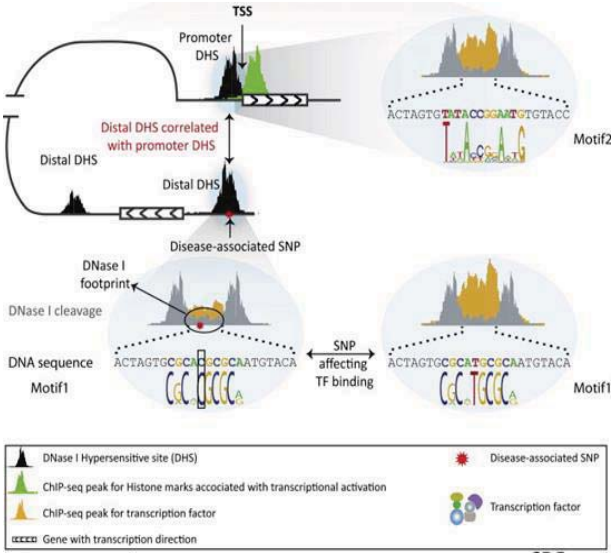
DNase I footprinting

- DNase I footprinting of K562 cells identifies the individual nucleotides within the MTPN promoter that are bound by NRF1.



22

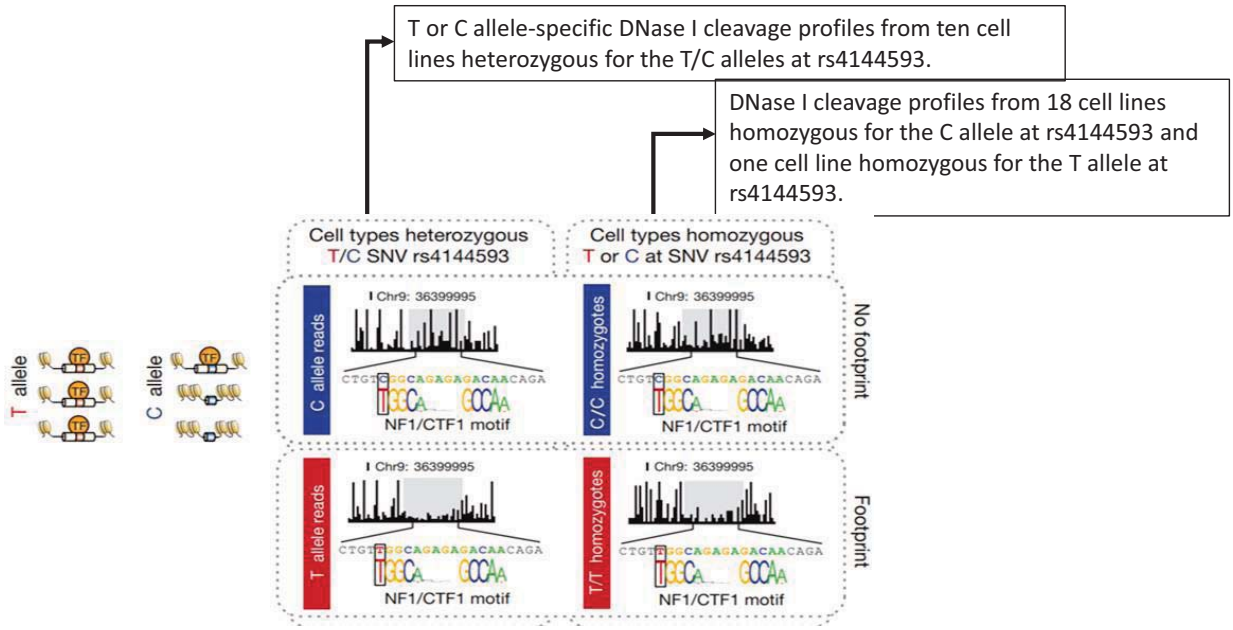
Noncoding variants and TF binding



Genomics Proteomics Bioinformatics 11 (2013) 135 - 141 23

Noncoding variants and TF binding

- DNase I footprints mark sites of in vivo protein occupancy.
- Effect of T/C SNV rs4144593 on protein occupancy and chromatin accessibility.

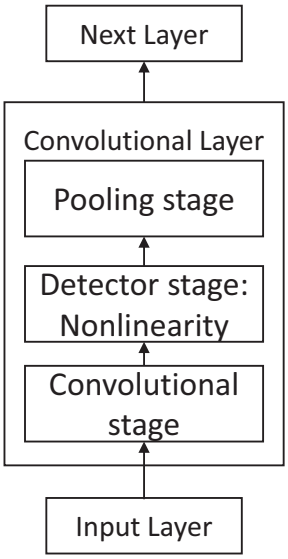


Neph, S. et al. Nature 489, 83–90 (2012) 24

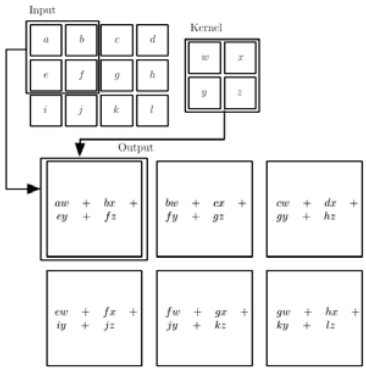
Contents

- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants
  - Convolutional neural network

A typical convolutional neural network layer

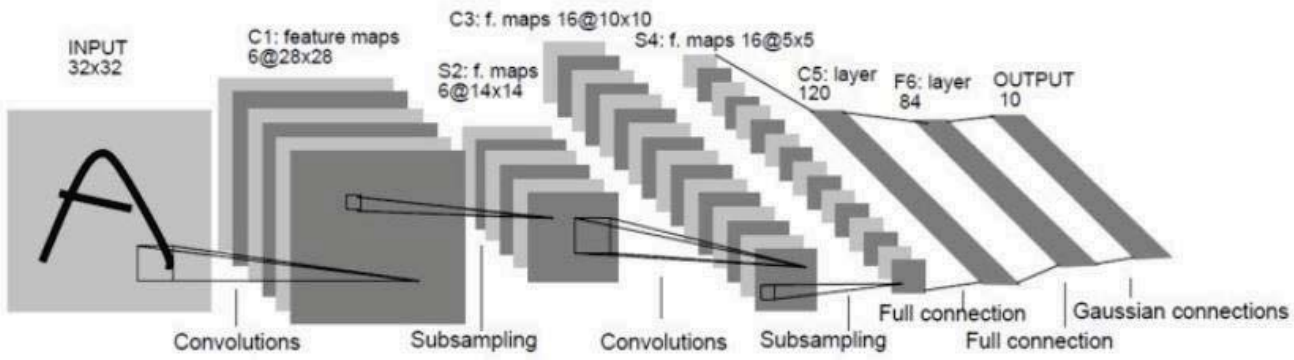


- Convolution stage



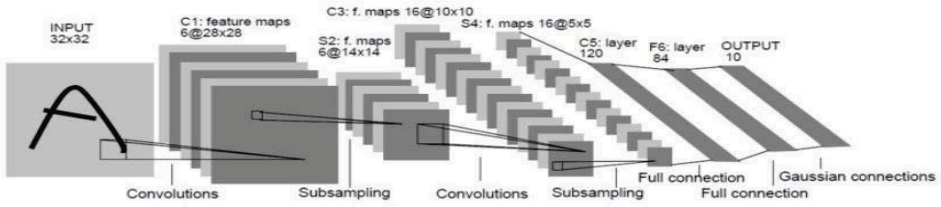
- Nonlinearity function
  - Rectified linear unit (ReLU)
  - Tanh, etc.
- Pooling stage
  - Max pooling
  - Average pooling, etc.

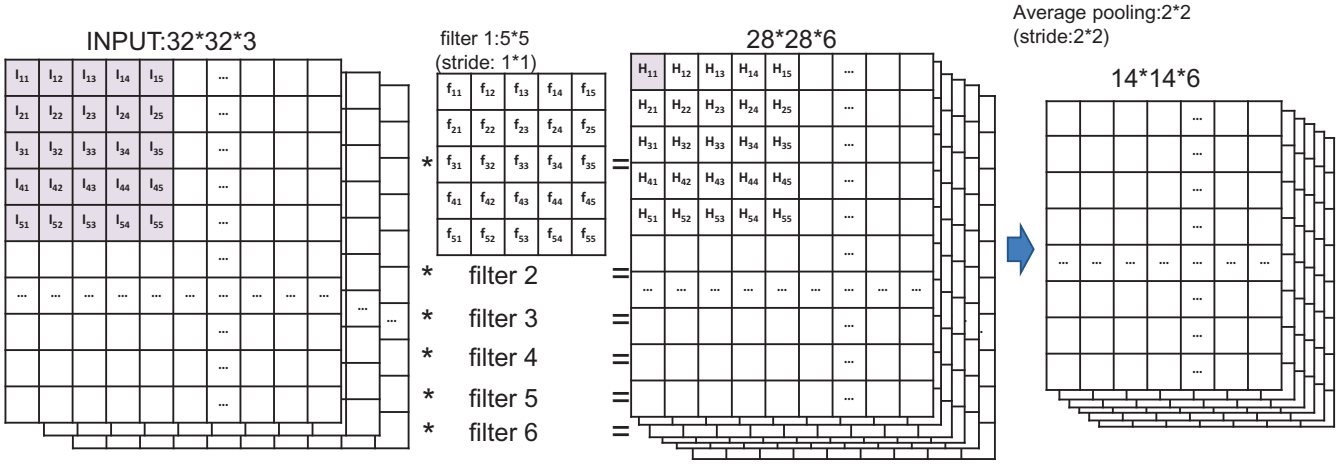
LeNet-5 (1998): An example of 2-D convolution



LeCun, Y.; Bottou, L.; Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86(11): 2278 - 2324.

LeNet-5 (1998): An example of 2-D convolution



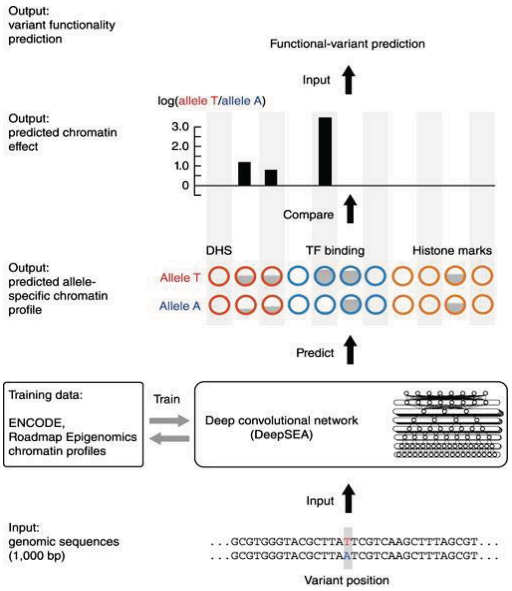


Contents

- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants
  - Convolutional neural network
  - DeepSea: Predicting effects of noncoding variants with deep learning–based sequence model

Sequence-based algorithmic framework DeepSEA (deep learning–based sequence analyzer)

- Goal: Predict with single-nucleotide sensitivity the effects of noncoding variants on transcription factor (TF) binding, DNA accessibility and histone marks of sequences
  1. Simultaneously predict large-scale chromatin-profiling data, including TF binding, DNase I sensitivity and histone-mark profiles
  2. Predicting allele-specific chromatin profile and chromatin effect
  3. Those predictions are used to estimate functional effects of noncoding variants



Nat Methods. 2015 October; 12(10): 931–934

Datasets

- Genome-wide chromatin profiles
  - From the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects
  - 690 TF binding profiles for 160 different TFs, 125 DNase I hypersensitivity (DHS) profiles and 104 histone mark profiles (a total of 919 peak sets). (Supplementary Table 1)
  - 521.6 Mbp of the genome (17%) were found to be bound by at least one measured TF and were used as a regulatory information–rich and challenging set for training the DeepSEA regulatory code model

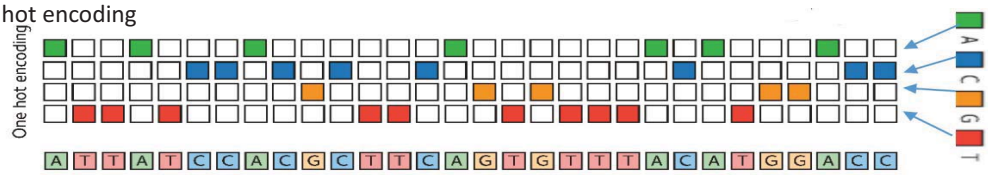
Table 1 Summary of ENCODE experiments

Experiment	Description
DNA methylation	In 82 human cell lines and tissues: A549, Adrenal gland, AG04449, AG04450, AG09309, AG09319, AG10803, AoSMC, BE2 C, BJ, Brain, Breast, Caco-2, CMK, ECC-1, Fibrobl, GM06990, GM12878, GM12891, GM12892, GM19239, GM19240, H1-hESC, HAEpiC, HCF, HCM, HCPEpiC, HCT-116, HEEpiC, HEK293, HeLa-S3, Hepatocytes, HepG2, HIPEpiC, HL-60, HMEC, HNPCEpiC, HPAEpiC, HRCEpiC, HRE, HRPEpiC, HSMM, HTR8svn, IMR90, Jurkat, K562, Kidney, Left Ventricle, Leukocyte, Liver, LNCaP, Lung, MCF-7, Melano, Myometr, NB4, NH-A, NHBE, NHDF-neo, NT2-D1, Osteoblasts, Ovcar-3, PANC-1, Pancreas, PanIslets, Pericardium, PFSK-1, Placenta, PrEC, ProgFib, RPTEC, SAEC, Skeletal muscle, Skin, SkMC, SK-N-MC, SK-N-SH, Stomach, T-47D, Testis, U87, UCH-1 and Uterus
TF ChIP-seq	A total of 119 TFs: ATF3, BATF, BCLAF1, BCL3, BCL11A, BDP1, BHLHE40, BRCA1, BRF1, BRF2, CCNT2, CEBPB, CHD2, CTBP2, CTCF, CTCFL, EBF1, EGR1, ELF1, ELK4, EP300, ESRRA, ESR1, ETS1, E2F1, E2F4, E2F6, FOS, FOSL1, FOSL2, FOXA1, FOXA2, GABPA, GATA1, GATA2, GATA3, GTF2B, GTF2F1, GTF3C2, HDAC2, HDAC8, HMGN3, HNF4A, HNF4G, HSF1, IRF1, IRF3, IRF4, JUN, JUNB, JUND, MAFF, MAFK, MAX, MEF2A, MEF2C, MXI1, MYC, NANOG, NFE2, NFKB1, NFYA, NFYB, NRF1, NR2C2, NR3C1, PAX5, PBX3, POLR2A, POLR3A, POLR3G, POU2F2, POU5F1, PPARGC1A, PRDM1, RAD21, RDBP, REST, RFX5, RXRA, SETDB1, SIN3A, SIRT6, SIX5, SMARCA4, SMARCB1, SMARCC1, SMARCC2, SMC3, SPI1, SP1, SP2, SREBF1, SRF, STAT1, STAT2, STAT3, SUZ12, TAF1, TAF7, TAL1, TBP, TCF7L2, TCF12, TFAP2A, TFAP2C, THAP1, TRIM28, USF1, USF2, WRNIP1, YY1, ZBTB7A, ZBTB33, ZEB1, ZNF143, ZNF263, ZNF274 and ZZZ3
Histone ChIP-seq	A total of 12 types: H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2 and H4K20me1
DNase-seq	In 125 cell types or treatments: 8988T, A549, AG04449, AG04450, AG09309, AG09319, AG10803, AoAF, AoSMC/serum_free_media, BE2_C, BJ, Caco-2, CD20, CD34, Chorion, CLL, CMK, Fibrobl, FibroP, Gliobla, GM06990, GM12864, GM12865, GM12878, GM12891, GM12892, GM18507, GM19238, GM19239, GM19240, H7-hESC, H9ES, HAc, HAEpiC, HA-h, HA-sp, HBMEC, HCF, HCFaa, HCM, HConF, HCPEpiC, HCT-116, HEEpiC, HeLa-S3, HeLa-S3_IFNa4h, Hepatocytes, HepG2, HESC, HFF, HFF-Myc, HGF, HIPEpiC, HL-60, HMEC, HMF, HMVEC-dAd, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Ad, HMVEC-dLy-Neo, HMVEC-dNeo, HMVEC-LBl, HMVEC-LLy, HNPCEpiC, HPAEC, HPAF, HPDE6-E6E7, HPdLF, HPF, HRCEpiC, HRE, HRGEC, HRPEpiC, HSMM, HSMMemb, HSMMtube, HTR8svn, Huh-7, Huh-7.5, HUVEC, HVMF, iPS, Ishikawa_Estr, Ishikawa_Tamox, Jurkat, K562, LNCaP, LNCaP_Andr, MCF-7, MCF-7_Hypox, Medullo, Melano, MonocytesCD14+, Myometr, NB4, NH-A, NHDF-Ad, NHDF-neo, NHEK, NHLF, NT2-D1, Osteobl, PANC-1, PanIsletD, PanIslets, pHTE, PrEC, ProgFib, PrEC, RPTEC, RWPE1, SAEC, SKMC, SK-N-MC, SK-N-SH_RA, Stellate, T-47D, Th0, Th1, Th2, Urothelia, Urothelia_UT189, WERI-Rb-1, WI-38 and WI-38_Tamox
DNase footprint	In 41 cell types: AG10803, AoAF, CD20+, CD34+ Mobilized, fBrain, fHeart, fLung, GM06990, GM12865, HAEpiC, HA-h, HCF, HCM, HCPEpiC, HEEpiC, HepG2, H7-hESC, HFF, HIPEpiC, HMF, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Neo, HMVEC-LLy, HPAF, HPdLF, HPF, HRCEpiC, HSMM, Th1, HVMF, IMR90, K562, NB4, NH-A, NHDF-Ad, NHDF-neo, NHLF, SAEC, SkMC and SK-N-SH RA
MNase-seq	In GM12878 and K562
3C-carbon copy (5C)	In GM12878, K562, HeLa-S3 and H1-hESC
GWAS SNP targeting	296 noncoding GWAS SNPs were assigned a target promoter

Nat Methods. 2015 October; 12(10): 931–934

Datasets for chromatin profile prediction

- Input
  - From 521,6 Mbp sequences (the human GRCh37 **reference genome**)
  - 1,000-bp DNA sequence
    - Centered on each 200-bp bin
    - 400-bp flanking regions at the two sides for extra contextual information
  - One hot encoding



- Output
  - 919 chromatin features
  - A chromatin feature was labeled 1 if more than half of the 200-bp bin is in the peak region and 0 otherwise.
  - Example:
    - Whether DNase-seq in a cell-line T-47D has a peak in the 200-bp bin
    - Whether TF FOXA1 in a brain cell-line has a peak in the 200-bp bin

Nat Methods. 2015 October; 12(10): 931–934

Training and Test sets

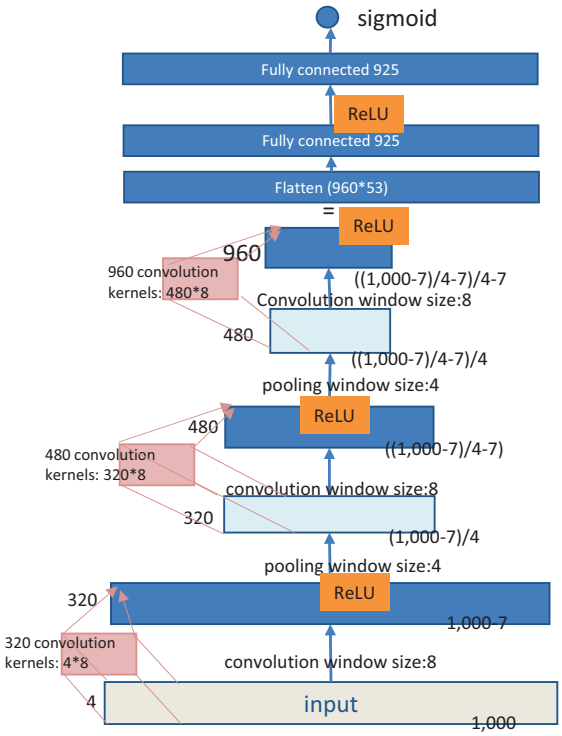
- Test: Chromosome 8 and 9
- Validation:
  - 4,000 samples on chromosome 7 spanning the genomic coordinates 30,508,751–35,296,850.
  - Hyperparameter selection
- Training: Rest of the autosomes

Nat Methods. 2015 October; 12(10): 931–934

DeepSEA model configuration

- Model Architecture

1. Convolution layer (320 kernels. Window size: 8. Step size: 1)
2. Pooling layer (Window size: 4. Step size: 4)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1)
4. Pooling layer (Window size: 4. Step size: 4)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer



Nat Methods. 2015 October; 12(10): 931–934

DeepSEA model configuration

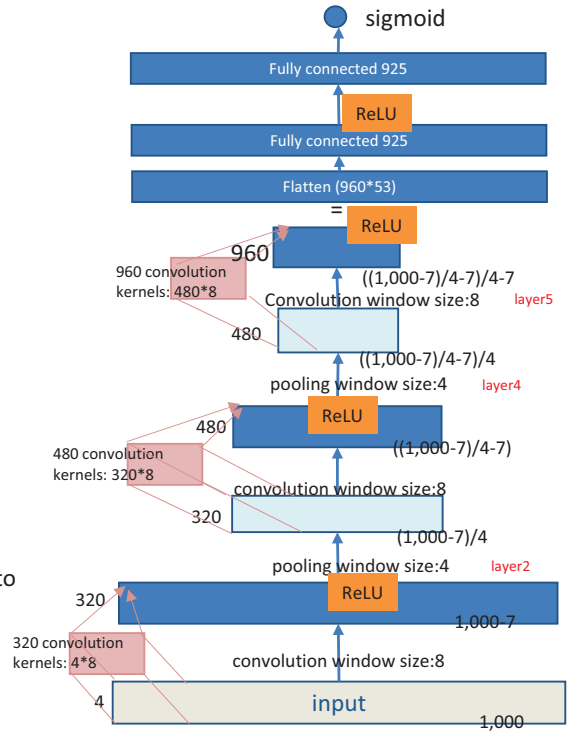
- Training of the DeepSEA model.

$$\text{objective} = \text{NLL} + \lambda_1 \|W\|_2^2 + \lambda_2 \|H^{-1}\|_1$$

$$\text{NLL} = -\sum_s \sum_t \log(Y_t^s f_t(X^s) + (1 - Y_t^s)(1 - f_t(X^s)))$$

  - s : index of training samples
  - t : index of chromatin features.
  - Y_t^s: 0,1 label for sample s, chromatin feature t.
  - $f_t\ (X^s)$: the predicted probability output of the model for chromatin feature t given input X^s.

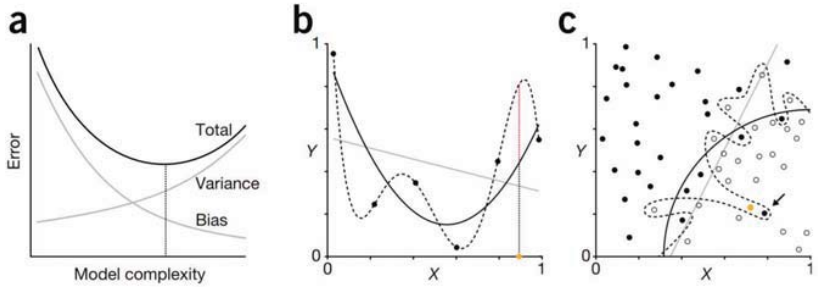
- Regularization Parameters:
  - L2 regularization (λ_1): 5e-07
  - L1 sparsity (λ_2): 1e-08
  - Dropout proportion (proportion of outputs randomly set to 0):
    - Layer 2: 20%, Layer 4: 20%, Layer 5: 50%, All other layers: 0%



Nat Methods. 2015 October; 12(10): 931–934

Regularization

- When model complexity increases, generally bias decreases and variance increases
- Minimize the total error.



(b) Polynomial fits to data simulated from a third-order polynomial underlying a model with normally distributed noise.
- Underfitting (gray diagonal line, linear fit), reasonable fitting (black curve, third-order polynomial) and overfitting (dashed curve, fifth-order polynomial).

(c) Two-class classification (open and solid circles)
- Underfitted (gray diagonal line), reasonable (black curve) and overfitted (dashed curve) decision boundaries.
- The overfit is influenced by an outlier (arrow) and would classify the new point (orange circle) as solid, which would probably be an error.

NATURE METHODS | VOL.13 NO.9 | SEPTEMBER 2016 | 703

Regularization (L1 norm and L2 norm)

- To reduce its generalization error but not its training error

$$\underset{w}{\operatorname{argmin}}((Xw - Y)^T (Xw - Y) + \lambda \mathrm{reg}(w)) = \underset{w}{\operatorname{argmin}}(J(w) + \lambda \mathrm{reg}(w))$$

L2 regularization

$\mathrm{reg}(w) = ||w||_2^2$

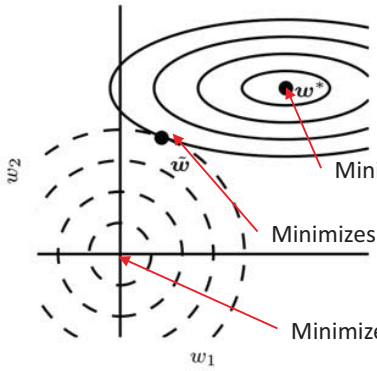
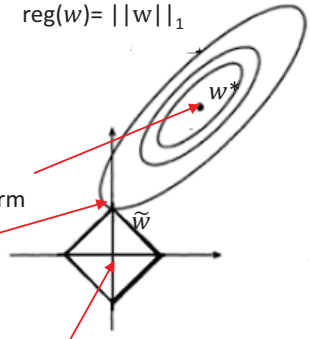


Figure 7.1

L1 sparsity

$\mathrm{reg}(w) = ||w||_1$



Minimizes data term

Minimizes combination

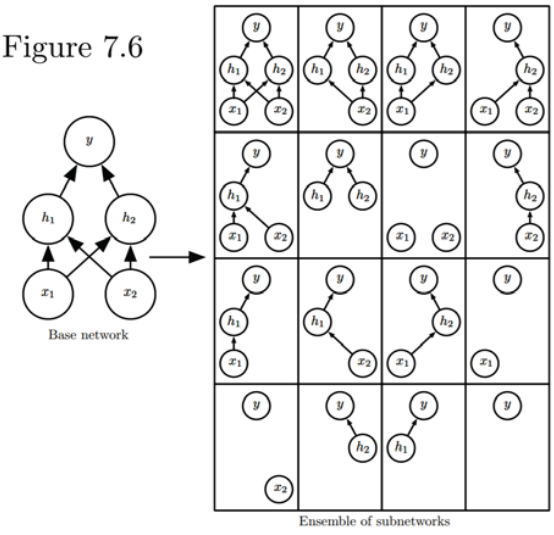
Minimizes regularization

- L1: Encourages sparsity
- Squared L2: Encourages small weights

Regularization for Deep Learning

- Dropout

Figure 7.6



DeepSEA model configuration

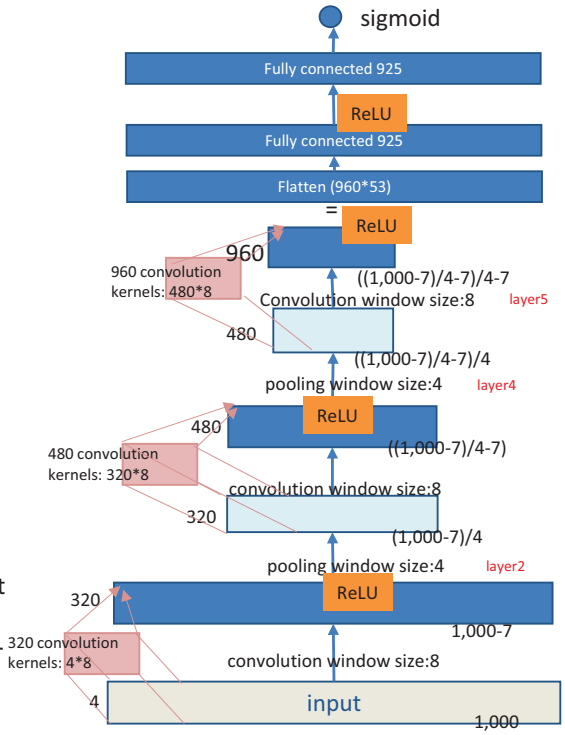
- Training of the DeepSEA model.

$$\text{objective} = \text{NLL} + \lambda_1 \|W\|_2^2 + \lambda_2 \|H^{-1}\|_1$$

$$\text{NLL} = -\sum_s \sum_t \log(Y_t^s f_t(X^s) + (1 - Y_t^s)(1 - f_t(X^s)))$$

  - s : index of training samples
  - t : index of chromatin features.
  - Y_t^s: 0,1 label for sample s, chromatin feature t.
  - $f_t\ (X^s)$: the predicted probability output of the model for chromatin feature t given input X^s.

- Regularization Parameters:
  - L2 regularization (λ_1): 5e-07
  - L1 sparsity (λ_2): 1e-08
  - Dropout proportion (proportion of outputs randomly set to 0):
    - Layer 2: 20%, Layer 4: 20%, Layer 5: 50%, All other layers: 0%



Nat Methods. 2015 October; 12(10): 931–934

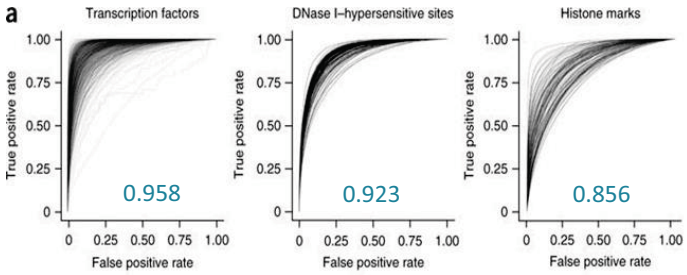
model.lua

```
require 'torch'
require 'nn'
require 'cunn'
require 'math'

nfeats = 4
width = trainData.data:size(3)
height = 1
ninputs = nfeats*width*height
nkernels = {320,480,960}

model = nn.Sequential()

model:add(nn.SpatialConvolutionMM(nfeats, nkernels[1], 1, 8, 1, 1, 0):cuda())
model:add(nn.Threshold(0, 1e-6):cuda())
model:add(nn.SpatialMaxPooling(1,4,1,4):cuda())
model:add(nn.Dropout(0.2):cuda())

model:add(nn.SpatialConvolutionMM(nkernels[1], nkernels[2], 1, 8, 1, 1, 0):cuda(
))
model:add(nn.Threshold(0, 1e-6):cuda())
model:add(nn.SpatialMaxPooling(1,4,1,4):cuda())
model:add(nn.Dropout(0.2):cuda())

model:add(nn.SpatialConvolutionMM(nkernels[2], nkernels[3], 1, 8, 1, 1, 0):cuda())
model:add(nn.Threshold(0, 1e-6):cuda())
model:add(nn.Dropout(0.5):cuda())

nchannel = math.floor((math.floor((width-7)/4.0)-7)/4.0)-7
model:add(nn.Reshape(nkernels[3]*nchannel))
model:add(nn.Linear(nkernels[3]*nchannel, noutputs))
model:add(nn.Threshold(0, 1e-6):cuda())
model:add(nn.Linear(noutputs , noutputs):cuda())
model:add(nn.Sigmoid():cuda())

print(model)
```

Nat Methods. 2015 October; 12(10): 931–934

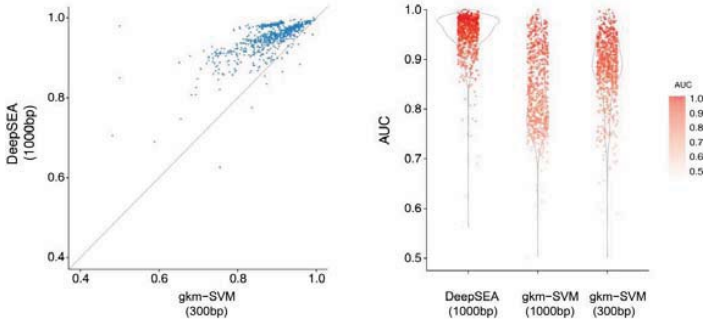
Chromatin profile prediction performance

Receiver operating characteristic (ROC) curves



Performance comparison with gkm-SVM for TF binding site prediction

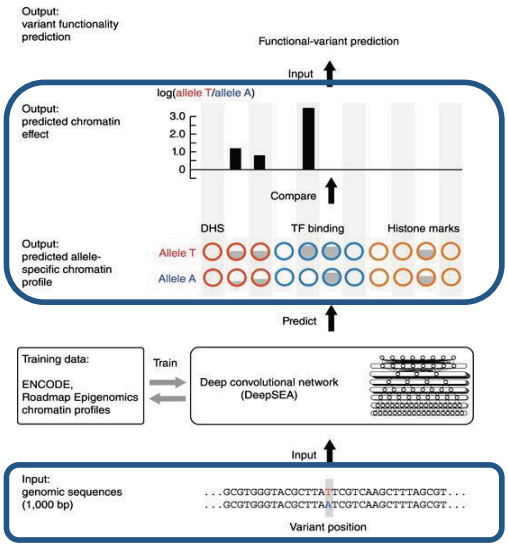
DeepSEA outperformed k-mer SVM for most TFs.



Gapped *k*-mer SVM did not gain performance from increasing size of context sequences

Nat Methods. 2015 October; 12(10): 931–934

Chromatin effects of single-nucleotide alteration in noncoding sequence



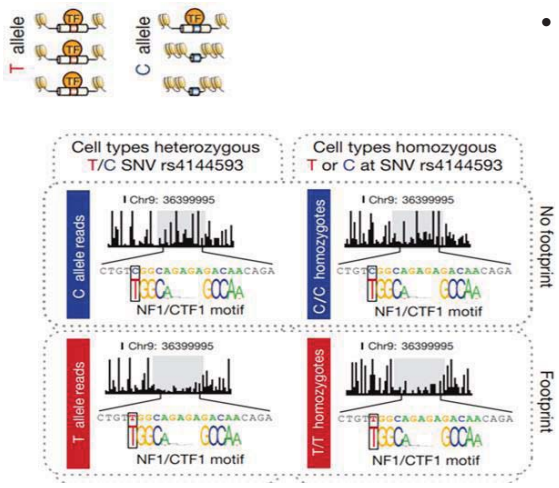
- Computational mutation scanning to assess the effect of mutating every base of the input sequence
- The effect of a base substitution on a specific chromatin feature prediction

$$\log_2\left(\frac{P_0}{1-P_0}\right) - \log_2\left(\frac{P_1}{1-P_1}\right)$$

P_0: probability predicted for the original sequence
P_1: probability predicted for the mutated sequence

Nat Methods. 2015 October; 12(10): 931–934

Chromatin effects of single-nucleotide alteration in noncoding sequence



Neph, S. et al. Nature 489, 83–90 (2012).

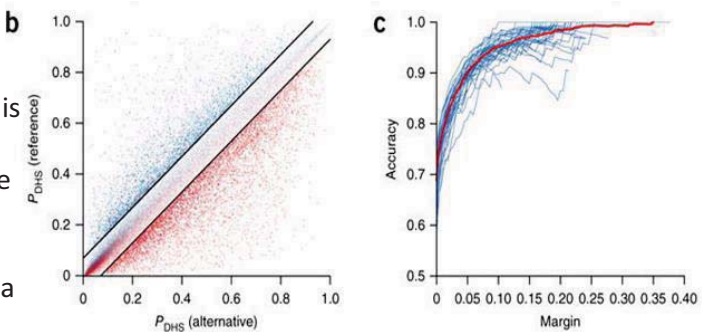
- Evaluation data
  - Allelic imbalance information from digital genomic footprinting (DGF) DNase-seq data on ENCODE cell lines.
  - Allelic imbalance: one allele is observed in DNase-seq data significantly more often than the other allele at a heterozygous site for a single-cell-type sample
  - 57,407 allelically imbalanced SNPs from 35 cell types with DHS predictors
    - 28,918 reference allele–biased variants
    - 28,489 alternative allele– biased variants

Nat Methods. 2015 October; 12(10): 931–934

Performance for predictions for DNase I–sensitive alleles

(b)
- Y-axis: predicted prob. that reference allele is DHS
- X-axis: predicted prob. that alternative allele is DHS
- Red dot: experimentally determined alternative allele–biased variant by DGF data
- Blue dot: experimentally determined reference allele–biased variant by DGF data

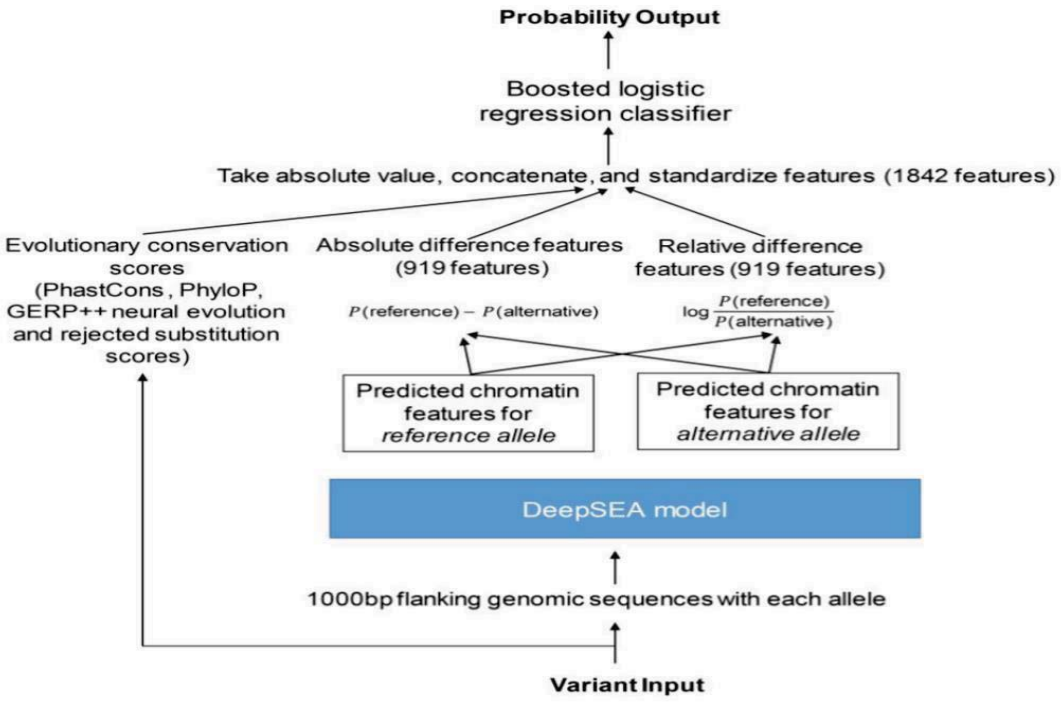
- Black lines: the margin, or the threshold of predicted probability differences between the two alleles for classifying high-confidence predictions (margin = 0.07 for this plot).



(c) Accuracy.
- Blue line: performance for a different cell type
- Red line: overall performance on allelically imbalanced variants for all 35 cell types

Nat Methods. 2015 October; 12(10): 931–934

Functional SNP prioritization



Nat Methods. 2015 October; 12(10): 931–934

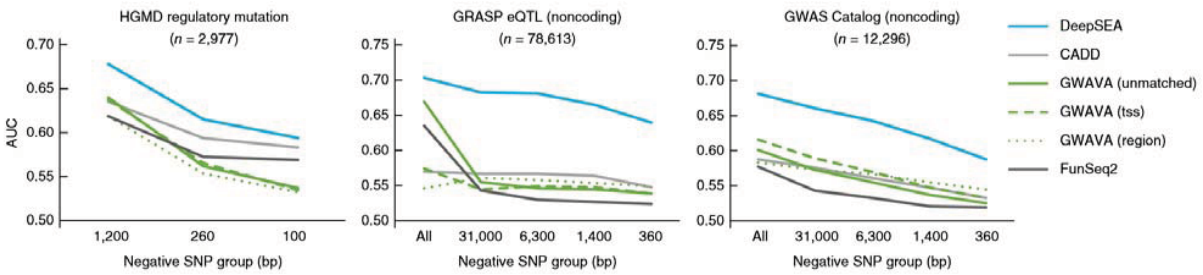
Data for functional SNP prioritization

- Positive standards
  - Human Gene Mutation Database (HGMD) annotated noncoding regulatory mutations
  - Noncoding eQTLs from the GRASP (Genome-Wide Repository of Associations between SNPs and Phenotypes) database
  - Noncoding trait-associated SNPs identified in GWAS studies from the US National Human Genome Research Institute's GWAS Catalog
- Negative standards
  - Several sets of negative SNPs with different distances to positive standard SNPs
  - Closest 1000 Genomes SNPs in the full set, 25% random subset and 5% random subset of 1000 Genomes SNPs with minor allele frequency greater than 0.01.
  - More…

Nat Methods. 2015 October; 12(10): 931–934

Performance of functional SNP prioritization

AUC values for tenfold cross-validation



- *x* axes: average distances of negative-variant groups to a nearest positive variant
- All: randomly selected negative 1000 Genomes SNPs

Nat Methods. 2015 October; 12(10): 931–934

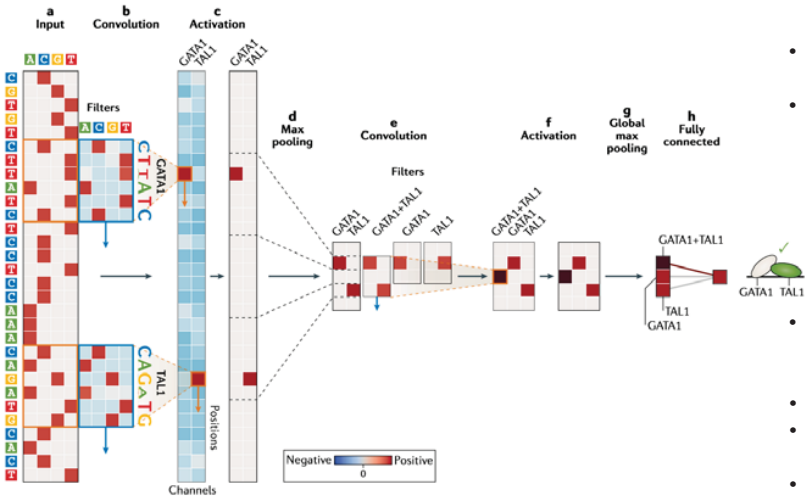
Contents

- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants
  - Convolutional neural network
  - DeepSea: Predicting effects of noncoding variants with deep learning–based sequence model
  - DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences

Nucleic Acids Research, 2016, Vol. 44, No. 11 e107

Recall) CNN and modelling TF binding sites

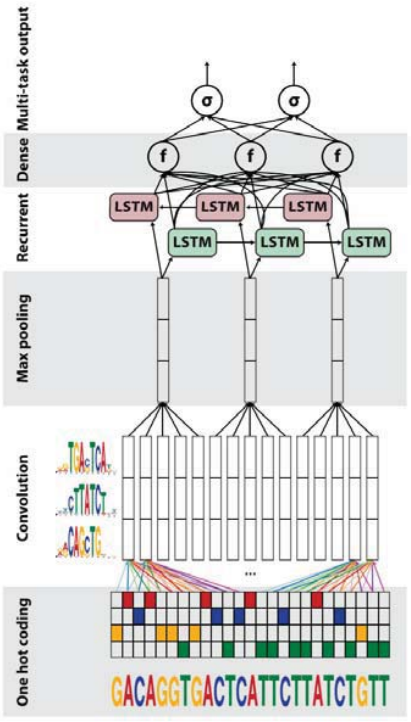
- CNN predicts the binding affinity of the TAL1–GATA1 transcription factor complex.



Nature reviews genetics volume 20:389 July 2019

- **a**: One-hot encoding of the DNA sequence.
- **b**: First convolutional layer scans the input sequence using filters, which are exemplified by position weight matrices of the GATA1 and TAL1 transcription factors.
- **c**: Negative values are truncated to 0 using ReLU activation function.
- **d**: In the max pooling operation, contiguous bins of the activation map are summarized by taking the maximum value for each channel in each bin.
- **e**: The second convolutional layer scans the sequence for pairs of motifs and for instances of individual motifs.
- **f**: ReLU activation function is applied.
- **g**: The maximum value across all positions for each channel is selected.
- **h**: A fully connected layer is used to make the final prediction.

DanQ model

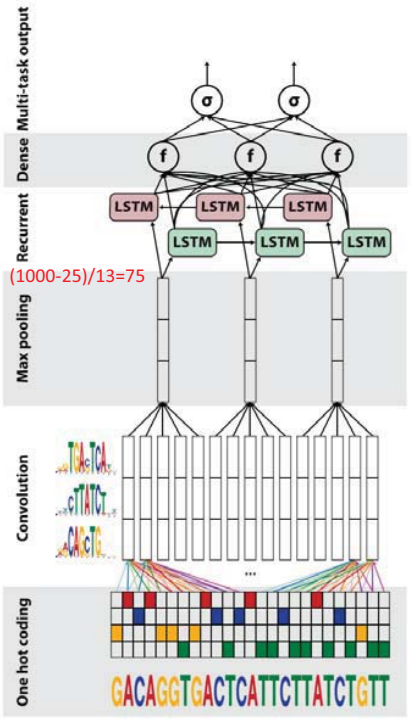


- Graphical illustration of the DanQ model
  - Input sequence
    - One hot encoded into a 4-row bit matrix.
  - Convolution layer with rectifier activation
    - Acts as a motif scanner across the input matrix
    - Produces an output matrix with a row for each convolution kernel and a column for each position in the input.
  - Max pooling
    - Reduces the size of the output matrix along the spatial axis, preserving the number of channels.

```
model = Sequential()
model.add(Convolution1D(input_dim=4,
                        input_length=1000,
                        nb_filter=320,
                        filter_length=26,
                        border_mode="valid",
                        activation="relu",
                        subsample_length=1))

model.add(MaxPooling1D(pool_length=13, stride=13))
model.add(Dropout(0.2))
```

Nucleic Acids Research, 2016, Vol. 44, No. 11 e107

DanQ model



- Graphical illustration of the DanQ model
- BLSTM layer
  - Considers the orientations and spatial distances between the motifs.
- Two fully connected layers
  - A dense layer of rectified linear unit
  - Sigmoid non-linear transformation to a vector that serves as probability predictions of the epigenetic marks to be compared via a loss function to the true target vector.

- The rationale for BLSTM layer
  - Motifs can follow a regulatory grammar
  - *in vivo* spatial arrangements and frequencies of combinations of motifs,
  - A feature associated with tissue-specific functional elements such as enhancers

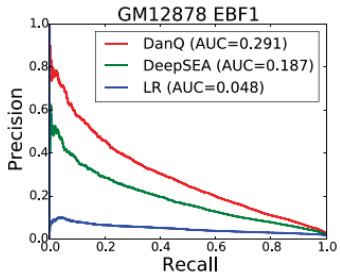
```
forward_lstm = LSTM(input_dim=320, output_dim=320, return_sequences=True)
backward_lstm = LSTM(input_dim=320, output_dim=320, return_sequences=True)
brnn = Bidirectional(forward=forward_lstm, backward=backward_lstm, return_sequences=True)
model.add(brnn)
model.add(Dropout(0.5))
model.add(Flatten())
model.add(Dense(input_dim=75*640, output_dim=925))
model.add(Activation('relu'))
model.add(Dense(input_dim=925, output_dim=919))
model.add(Activation('sigmoid'))
```

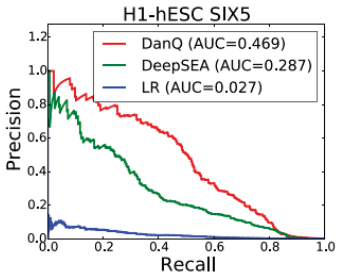
Note (1000-25)/13=75

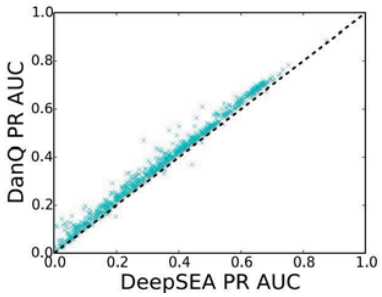
Nucleic Acids Research, 2016, Vol. 44, No. 11 e107

Performance comparison

- Training, validation and testing sets were downloaded from the DeepSEA website
- Input: reference sequence
- Output:A length 919 binary target vector from 919 ChIP-seq and DNase-seq peak sets from uniformly processed ENCODE and Roadmap Epigenomics data releases
- A better metric to measure the performance is the area under precision-recall curve (PR AUC)
- PR AUC metric is less prone to inflation by the class imbalance than the ROC AUC metric is





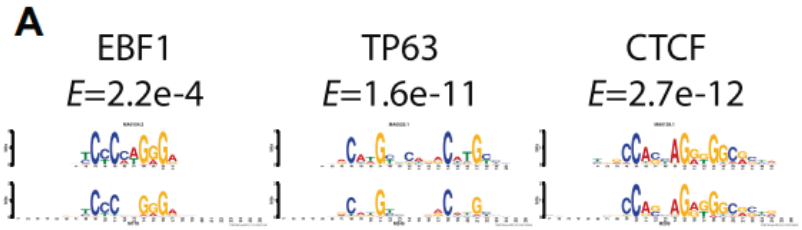


- LR models achieve a PR AUC below 5% for the two examples
- 97.6% of all DanQ PR AUC scores surpass DeepSEA PR AUC scores

Nucleic Acids Research, 2016, Vol. 44, No. 11 e107

Position frequency matrices, or motifs

- Convert the kernels from the convolution layer of the DanQ models to position frequency matrices, or motifs.
- Align these motifs to known motifs using the TOMTOM algorithm.
- Of the 320 motifs learned by the DanQ model, 166 significantly match known motifs ($E < 0.01$).



- Top: EBF1, TP63 and CTCF motif logos from JASPAR
- Bottom: three convolution kernels

Nucleic Acids Research, 2016, Vol. 44, No. 11 e107

Contents

- Introduction to noncoding variants
- Computational methods to prioritize noncoding variants
- Genomic and epigenomic information
- Deep learning methods to prioritize noncoding variants
  - Convolutional neural network
  - DeepSea: Predicting effects of noncoding variants with deep learning–based sequence model
  - DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences
  - DeepFun: Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations

Nucleic Acids Research, 2021, 49(1): 53-66

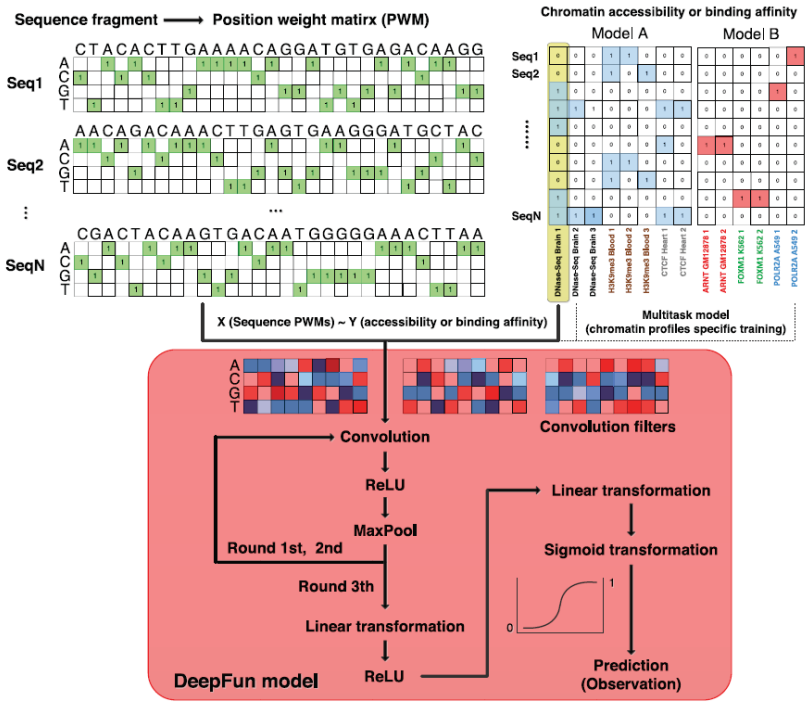
DeepFun

- Assess the functional impact of a non-coding variant and its impact in a tissue- and cell type-specific manner
- Increased epigenetics tracks from ENCODE and Roadmap (6 May 2019)
  - 7870 chromatin features
    - 1548 DNase I accessibility, 1536 histone mark and 4795 transcription factor binding profiles.
  - Removal of technical or biological replicates
    - DeepFun incorporates a total of 117 DNase-seq, 360 histone modification, and 795 TF binding profiles

vs. DeepSea

- A total of 919 peak sets (125 DNase I hypersensitivity profiles, 104 histone mark profiles, 690 TF binding profiles for 160 different TFs)

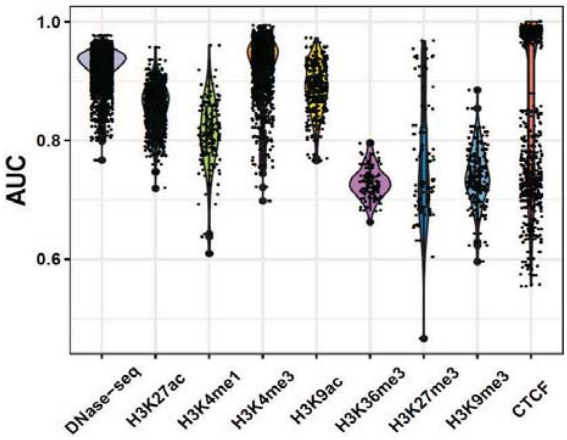
CNN model



- One hot encoding sequence
  - 1,000 bp sequence
  - Four-low binary matrix
- Three-layer CNN architecture
  - 300 convolution filters in the 1st layer
  - 2nd and 3rd layers were operated on the output of the prior layer
- Two fully connected neural network layers with 30% dropout rate
- A fully connected sigmoid transformation layer
- Predicting activity (accessibility or binding affinity) probability

SBi 한국생명정보학회 Korean Society for Bioinformatics Nucleic Acids Research, 2021, 49(1): 53-66

Performances



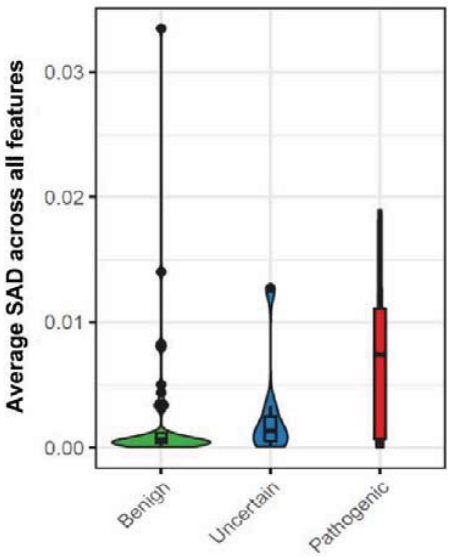
- Random selection of 80%, 10%, and the remaining 10% for training, validation, and for testing.
- A median AUC of 0.933 over all DNase-seq assays
- A median AUC at 0.80 for all TFs assays, ranging from 0.64 (*ZC3H11A*) to 0.98 (*SP4*)

Nucleic Acids Research, 2021, 49(1): 53-66

Prioritizing regulatory variants

- SNP Activity Difference (SAD)
  - *Alt – Ref*
  - *Ref:* predicted activity probability for the reference allele/original sequence (ranging 0 ~ 1)
  - *Alt:* predicted activity probability for the alternative allele/mutated sequence (ranging 0 ~ 1)
  - Variants with a higher positive SAD : alternative allele increases the epigenetic signal compared to the reference allele
  - Variants with a negative SAD value: decrease the epigenetic signal

Application to non-coding variants in ClinVar database

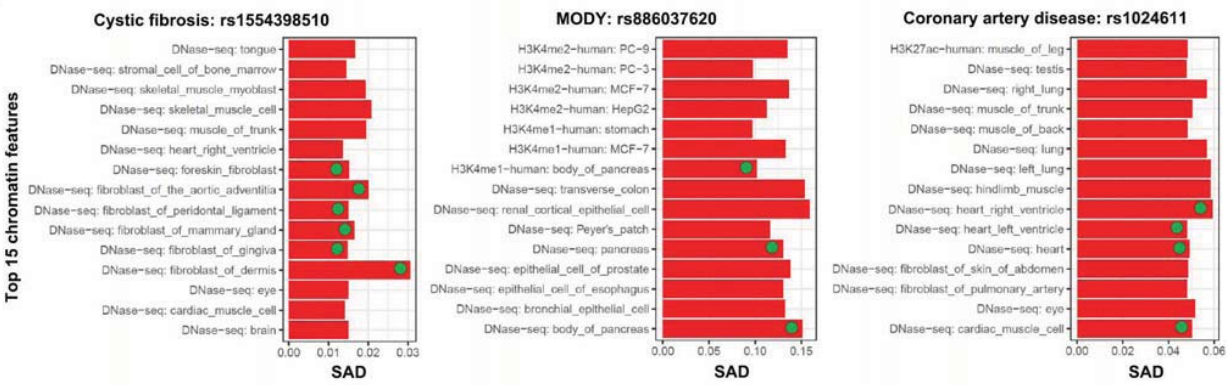


Nucleic Acids Research, 2021, 49(1): 53-66

SBi 한국생명정보학회 Korean Society for Bioinformatics

Prioritizing regulatory variants

- Prioritize non-coding causal variants in a tissue specific fashion.
- Top 15 chromatin features related for three non-coding variants



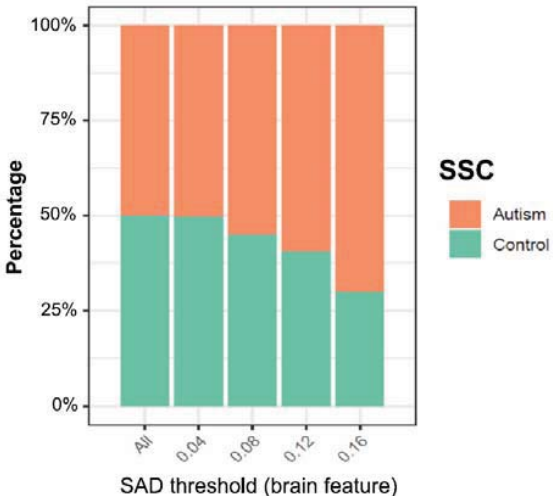
- Cystic fibrosis: Most fibroblast tissues related DNase-seq profiles were associated with rs1554398510, especially in fibroblast of dermis.
- Maturity-onset diabetes of the young (MODY): Both DNase-seq and H3K4me1 profiles in pancreas tissue had strong association with rs886037620.
- Coronary atery disease: The impact of rs1024611 was the strongest in heart and cardiac muscle tissue

Nucleic Acids Research, 2021, 49(1): 53-66

59

Prioritizing regulatory variants

- Autism *de novo* mutations from Simons Simplex Collection (SSC) cohort
  - 2600 simplex families
  - Each family has one child affected by ASD and unaffected parents and siblings.
- All non-coding variants are grouped into unaffected and affected siblings.
- Consider the average SAD scores of the non-coding variants over all brain tissues.
- With increasing SAD thresholds
  - The percentage of variants in patient siblings increases
  - The percentage in health siblings decreases.



60

Summary

- Noncoding variants
- Computational methods to prioritize noncoding variants based on genomic and epigenomic information
  - GWAVA: Genome-wide annotation of variants
- Deep learning methods based on genomic sequence
  - DeepSea
  - DanQ
  - DeepFun
- If you are interested, see studies in related topics.
  - DeepC: predicting 3D genome folding using megabase-scale transfer learning (Nature Methods 17:1118–1124(2020))
  - Predicting 3D genome folding from DNA sequence with Akita (Nature Methods 17:1111–1117(2020))