Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning (DeepBind)

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey

Presenter: Jack Lanchantin

University of Virginia

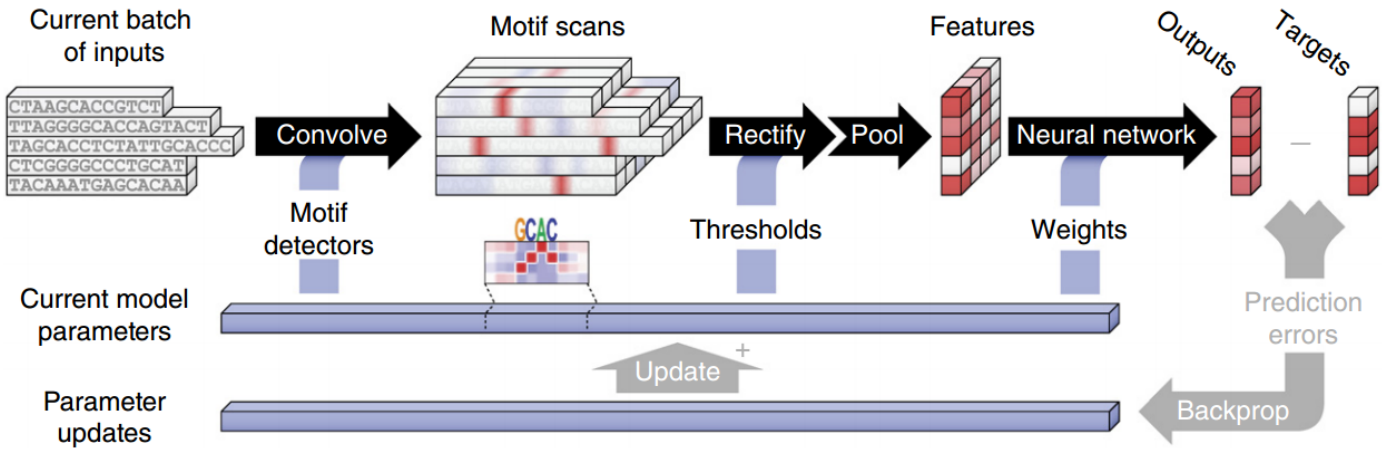
https://qdata.github.io/deep2Read/

Nature Biotech, 2015

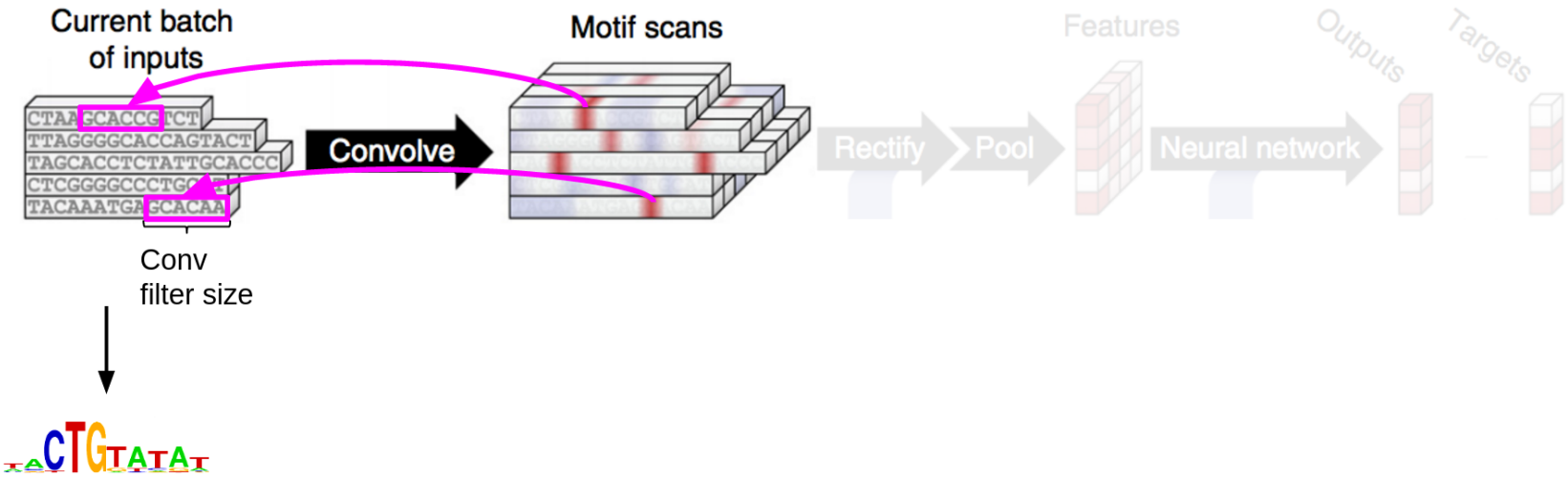
Overview

1. Predict TFBSs in vivo using ChIP-Seq Data with 1-layer CNN
2. Extract TF-specific motifs using CNN kernel activations
3. Analyze SNPs with mutation maps by varying individual nucleotides

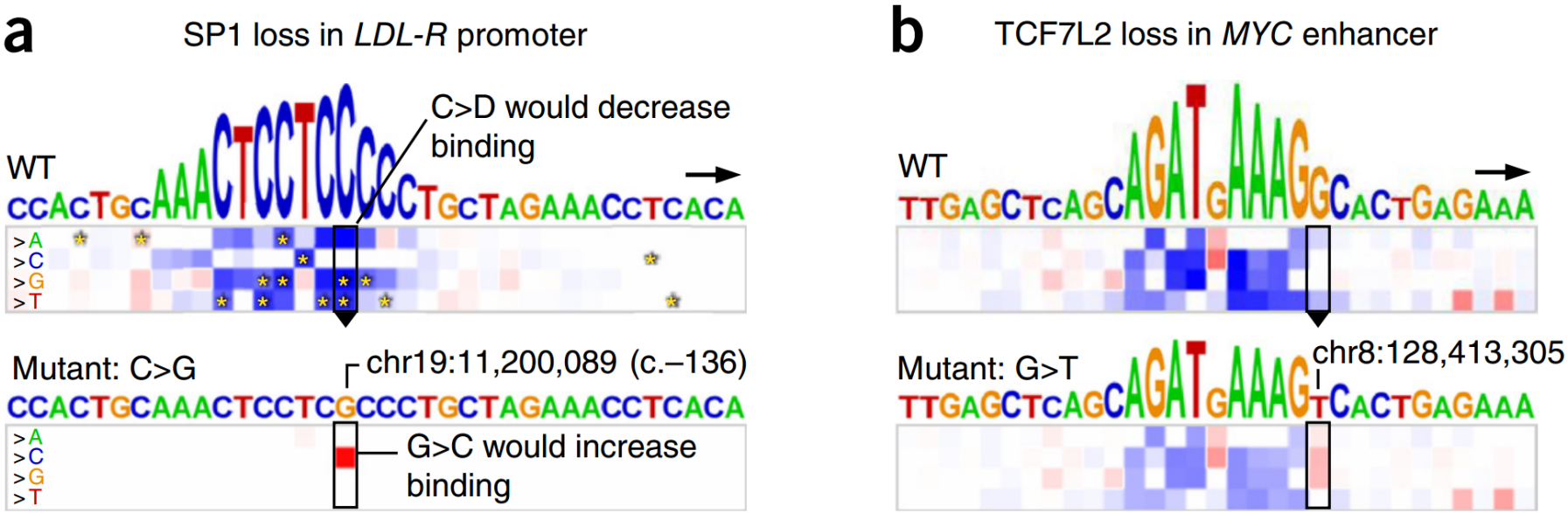
1. Predict TFBSs in vivo using ChIP-Seq Data with CNN



2. Extract TF-specific motifs using CNN kernel activations



3. Analyze SNPs with mutation maps by varying individual nucleotides



Predicting effects of noncoding variants with deep learning based sequence model (DeepSEA)

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey

Jian Zhou, Olga G Troyanskaya

University of Virginia
https://qdata.github.io/deep2Read/

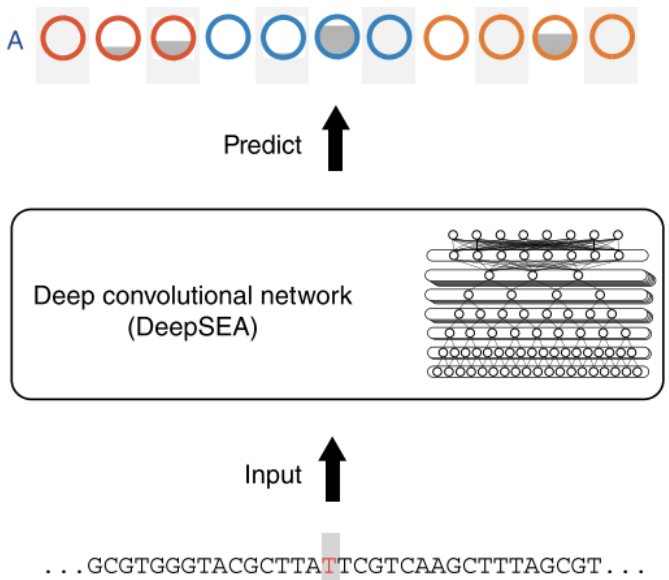
Nature Methods, 2015

Overview

1. Train and predict TFBS, HM, and DHS from raw sequence
2. Use chromatin predictions to predict variant effects of two separate allele input sequences

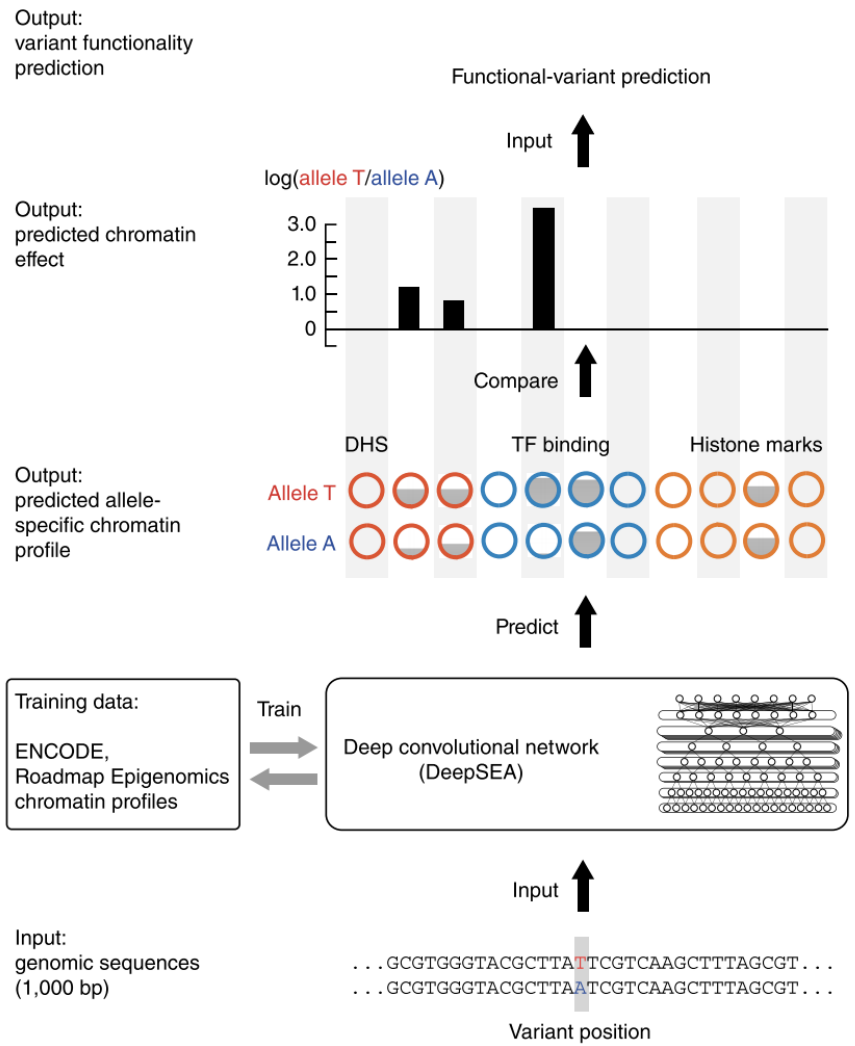
Task 1

- Input: 1000 length sequence
- Output: binary classification of: 690 TF binding profiles (160 different TFs), 125 DHS profiles and 104 histone mark profiles (total of 919 chromatin features).



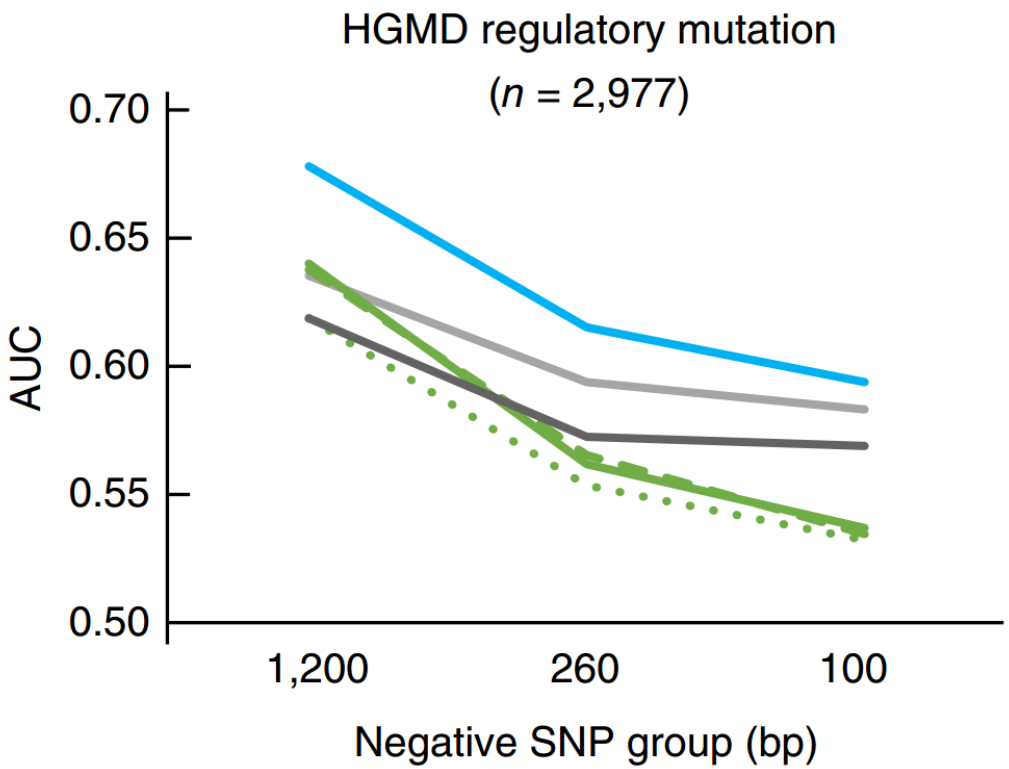
Task 2

- Input: Two 1000 length sequences (Allele A and Allele B)
- Output: Functional variant = yes/no



Results

- TFBS median AUC = 0.958
- DHS median AUC = 0.923
- HM median AUC = 0.856



Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey

David R. Kelley, Jasper Snoek, and John L. Rinn

University of Virginia
https://qdata.github.io/deep2Read/

Genome Research, 2016

Overview

1. Predict DHS signals from raw sequence using CNN
2. Extract motifs from filters - match filters to known TF motifs using TomTom
3. Predict the effect of SNP mutations on DHS signals

The Genetics of Transcription Factor DNA Binding Variation

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey

Bart Deplancke, Daniel Alpern, and Vincent Gardeux

University of Virginia
https://qdata.github.io/deep2Read/

Cell, July 2016

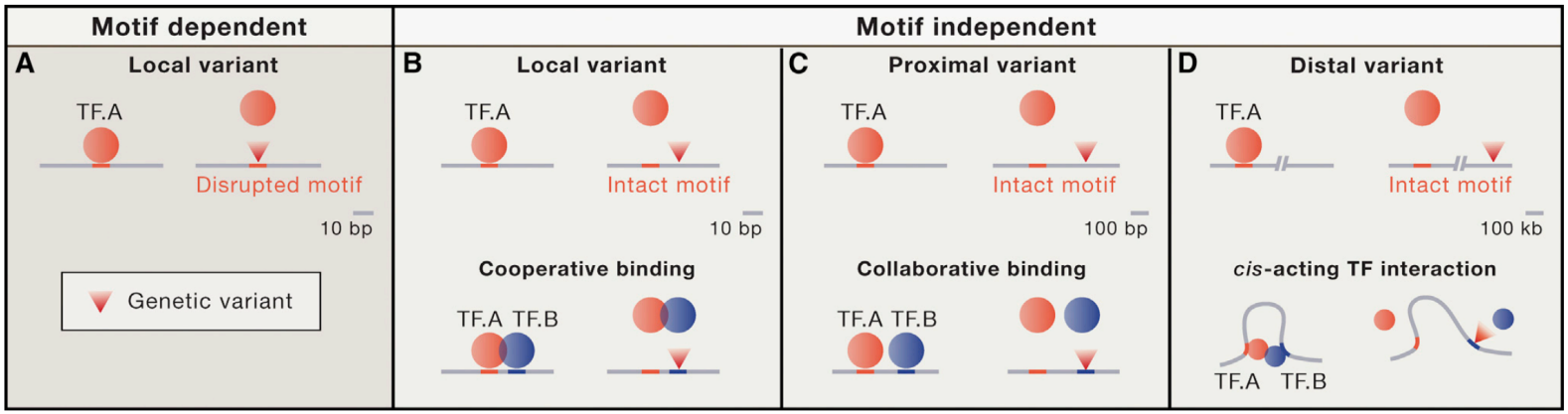
Identification of Causal Mutation and Affected Gene

- TF-DNA interactions are key drivers of phenotypic variation.
- However, the majority of TF binding changes are not driven by sequence variations in the TF motif of interest.
- It is wrong to assume we can match a GWAS SNP to the closest TFBS. Variants could be associated with genes $>$ 1Mbp away.

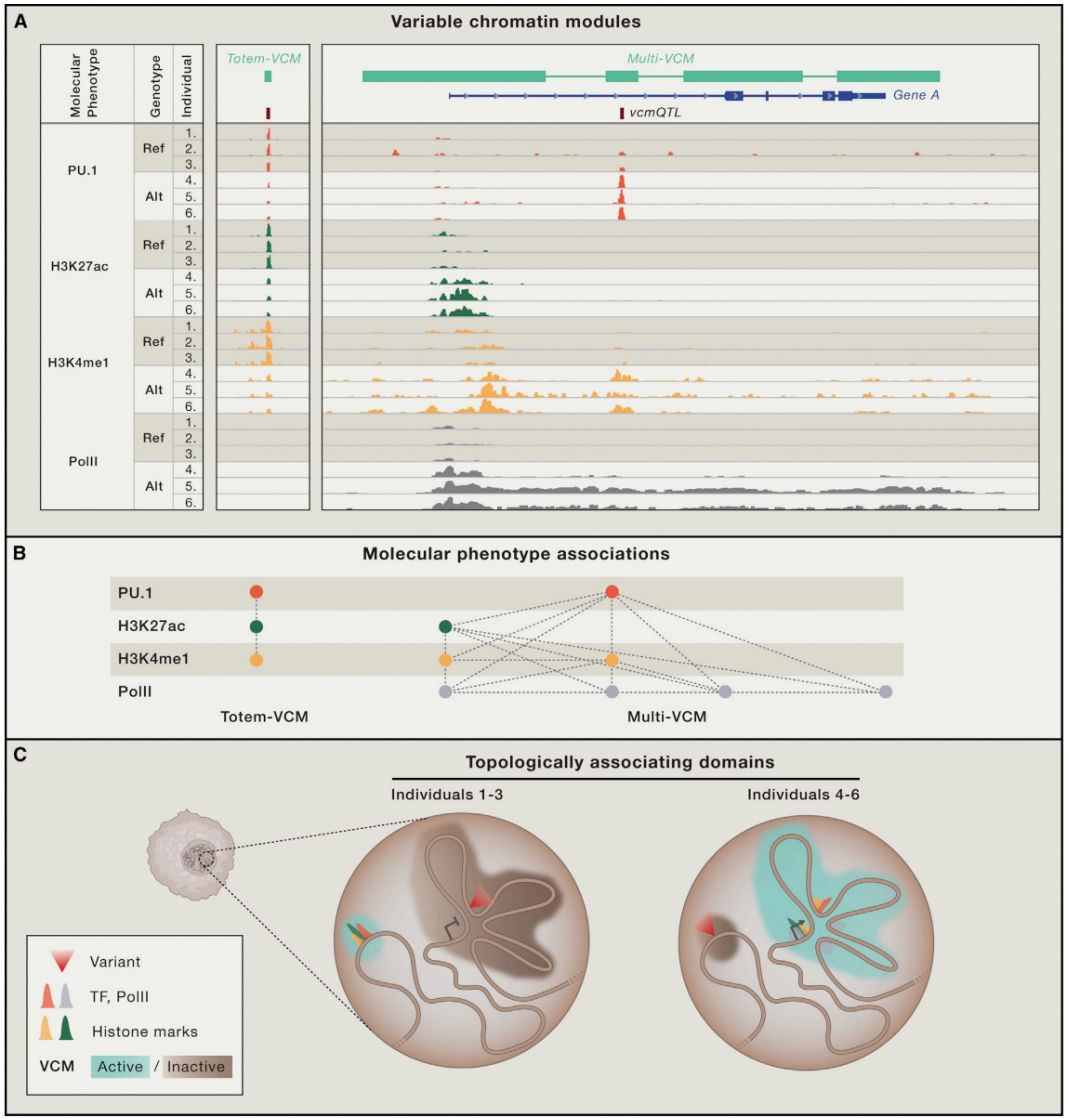
Problem: Incomplete TF Motif Catalog

- Over 1/3 of human TFs are devoid of consensus motifs
- Co-binding has largely been underrepresented in predictive models.
- A variant located either proximally ($<$200 bp) or distally to the focal motif affects the binding of the respective TF

Co-binding


Motif dependent
Motif independent
A
Local variant
TF.A
Disrupted motif
10 bp
Genetic variant
B
Local variant
TF.A
Intact motif
10 bp
Cooperative binding
TF.A TF.B
C
Proximal variant
TF.A
Intact motif
100 bp
Collaborative binding
TF.A TF.B
D
Distal variant
TF.A
Intact motif
100 kb
cis-acting TF interaction
TF.A TF.B

Co-binding


A
Variable chromatin modules
Molecular Phenotype
Genotype
Individual
Totem-VCM
Multi-VCM
Gene A
vcmQTL
PU.1
H3K27ac
H3K4me1
PolII
Ref
Alt
B
Molecular phenotype associations
PU.1
H3K27ac
H3K4me1
PolII
Totem-VCM
Multi-VCM
C
Topologically associating domains
Individuals 1-3
Individuals 4-6
Variant
TF, PolII
Histone marks
VCM
Active / Inactive