Sequence to Function

Enformer: Transformers in regulatory genomics

Xinming, Ian

Enformer

ARTICLES
https://doi.org/10.1038/s41592-021-01252-x

nature methods

Check for updates

OPEN

Effective gene expression prediction from sequence by integrating long-range interactions

Žiga Avsec[1]✉, Vikram Agarwal[2,4], Daniel Visentin[1,4], Joseph R. Ledsam[1,3], Agnieszka Grabska-Barwinska[1], Kyle R. Taylor[1], Yannis Assael[1], John Jumper[1], Pushmeet Kohli[1]✉ and David R. Kelley[2]✉

David Kelly
PI at Calico Labs

Preview

Check for updates

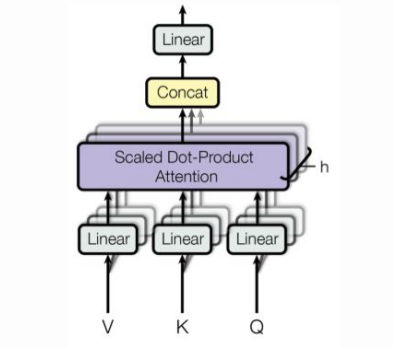
news & views

BIOINFORMATICS

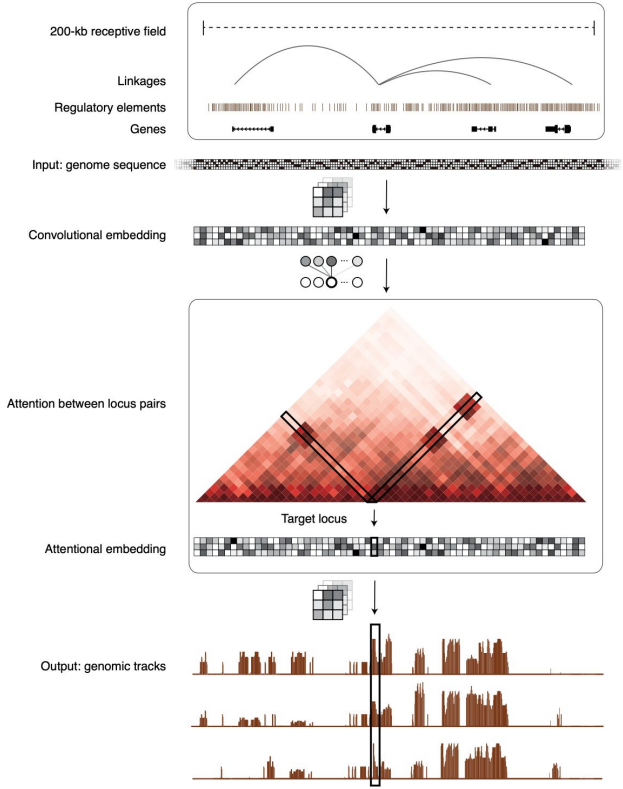
A wider field of view to predict expression

A gene sequence-to-expression machine learning model achieves improved accuracy by incorporating information about potential long-range interactions.

Yang Young Lu and William Stafford Noble



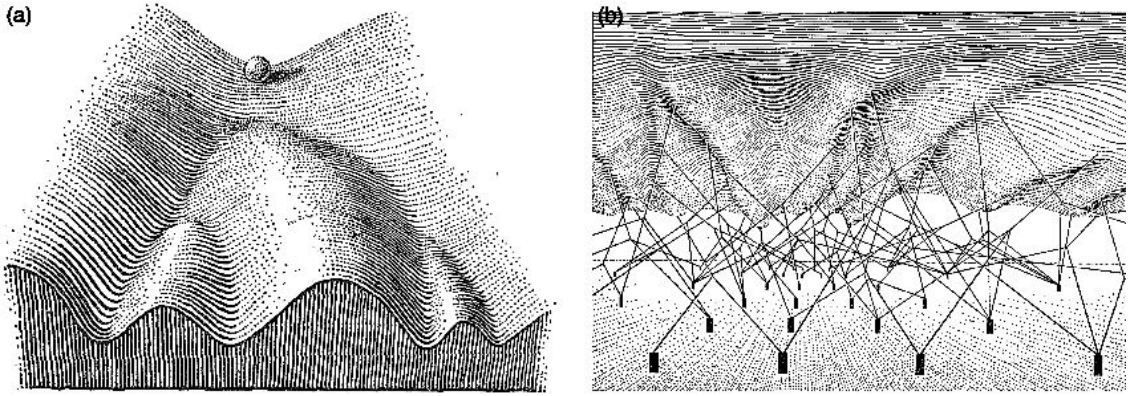
source: attention is all you need



Outline

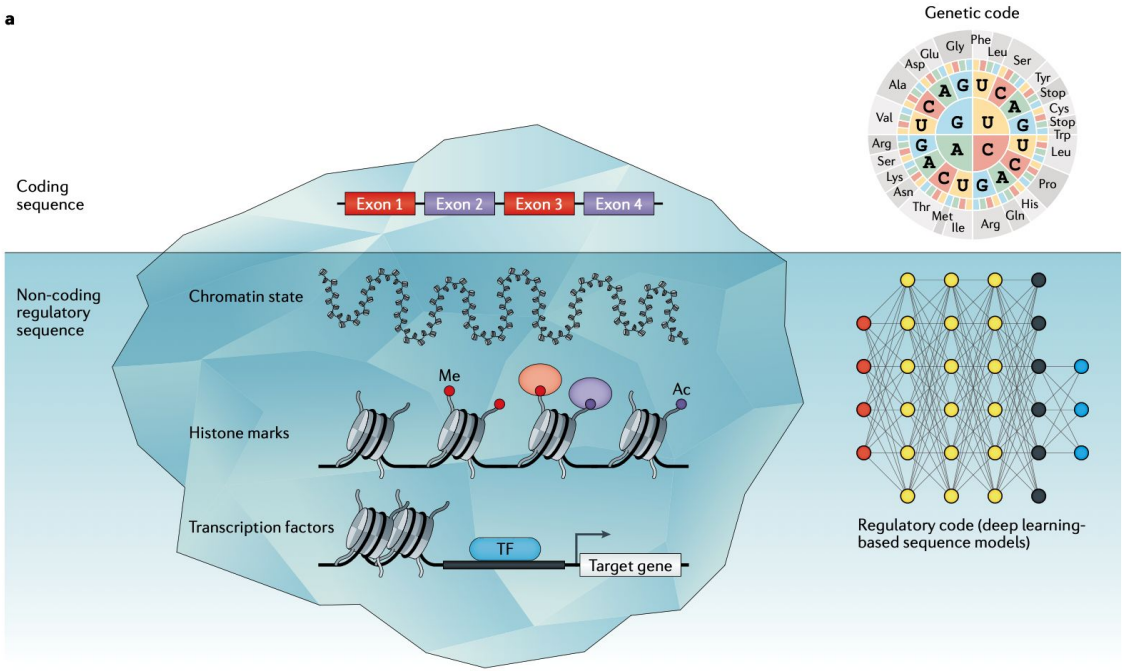
- Background
  - regulatory genomics
  - previous work
- Model: convolutions → self-attention
- Results
- Discussion

Regulatory genomics



Epigenetics landscape, Waddinton 1957

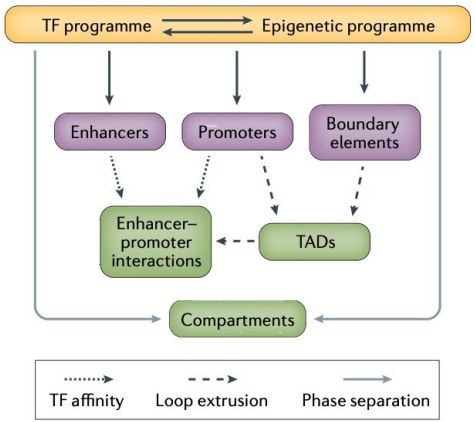
Background: non-coding regions

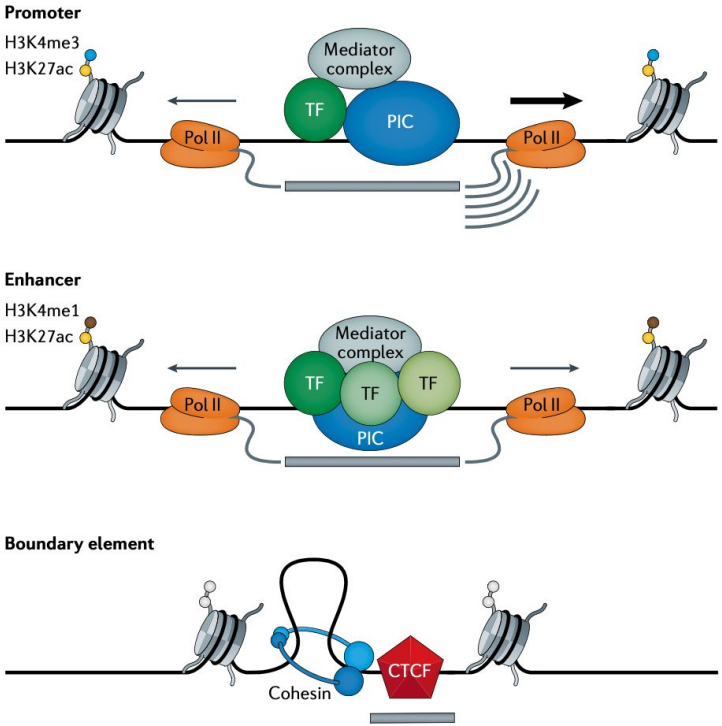


Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat Rev Genet* 1–17 (2021) doi:10.1038/s41576-021-00389-x.

Background: regulatory elements

- Promoter
- Enhancer
- Boundary element
  - Insulator (CTCF)

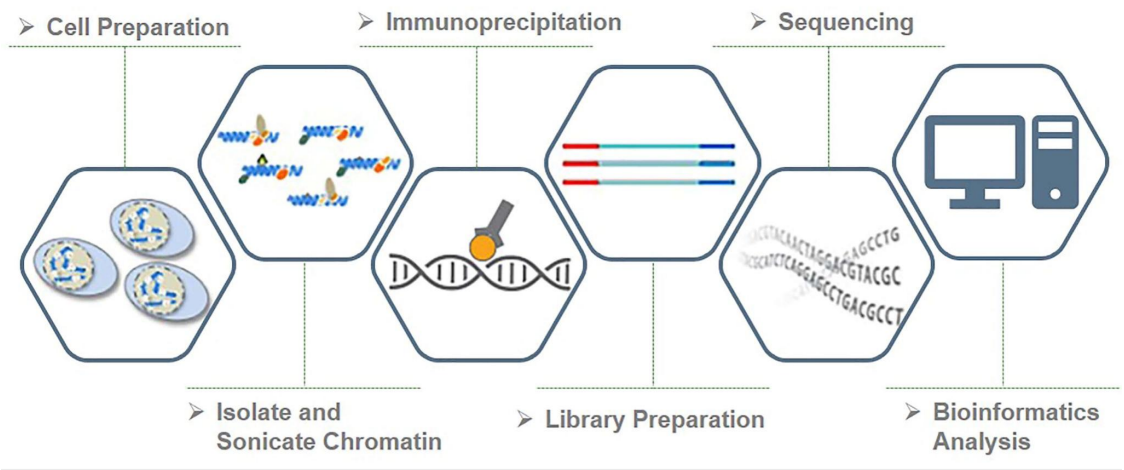




Oudelaar, A. M. & Higgs, D. R. The relationship between genome structure and function. *Nat Rev Genet* **22**, 154–168 (2021).

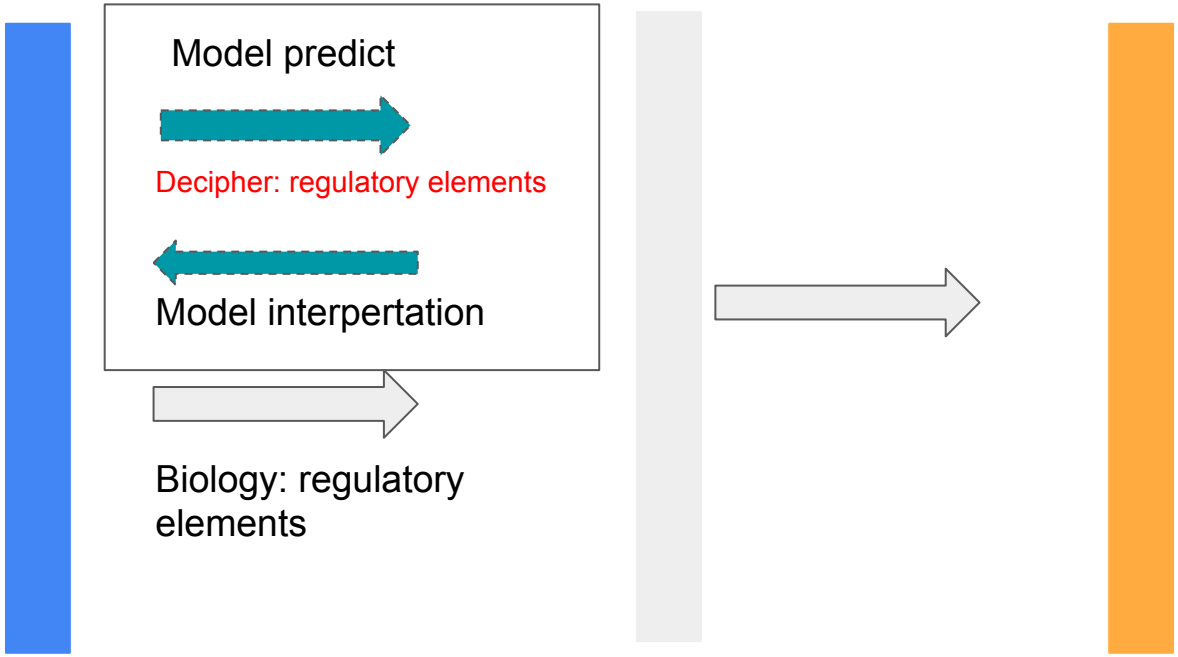
Background - Genomics tracks

- DNase/ATAC
  - Chormotin accessibility
- CAGE
  - gene expression
- Chip Histone
  - Histone marker
- Chip TF
  - TF binding



Chip-seq: https://www.encodeproject.org/chip-seq/transcription_factor/

Computational regulatory genomics



genotype/sequence

genomics/epigenetics Phenotype
Genomics track (Dnase, Histone marker)

Phenotype (Disease, function)

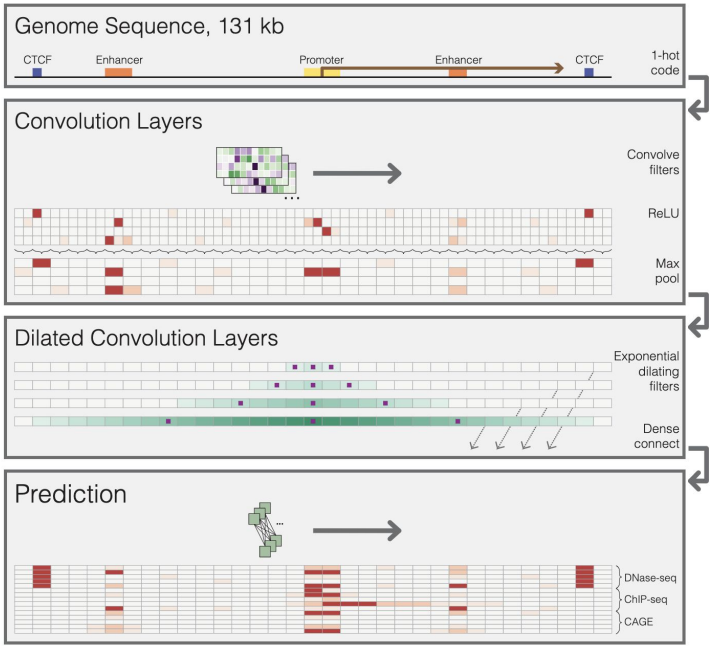
Previous work in computational regulatory genomics

- The state of the art: CNN
  - DeepBind, DeepSEA, Basenji, etc

Table 1 | **Methods for transcriptional/biochemical impact**

Model overview			**Input data**
Model	**Method**	**Prediction task**	
DeepSEA[25,26]	Deep learning, CNN	Chromatin, TF binding	TF, HM, DHS
gkm-SVM/ delta-SVM[28]	SVM	Chromatin, TF binding	TF, HM, DHS
DanQ[82]	Deep learning, CNN, BLSTM	Chromatin, TF binding	TF, DHS, HM
Basset[87]	Deep learning, CNN	Chromatin accessibility	DHS
DeepCpG[89]	Deep learning, CNN, GRU	CpG state	Bisulfite sequencing
ExPecto[95]	Deep learning, CNN, linear regression	Expression prediction	TF, HM, DHS, RNA-seq
Basenji[95]	Deep learning, CNN	Expression prediction	TF, DHS, HM, CAGE peaks
BPNet, DeepLIFT, TFModisco[84]	Deep learning, CNN	TF binding	TF
ChromDragoNN[83]	Deep learning, CNN, ResNet	Chromatin accessibility	DHS, RNA-seq
Xpresso[83,94]	Deep learning, CNN	Expression prediction	CAGE peaks, gene annotations
AMBER[40]	Auto machine learning, RNN, deep learning, CNN	Chromatin, TF binding	TF, HM, DHS

Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat Rev Genet* 1–17 (2021) doi:10.1038/s41576-021-00389-x.



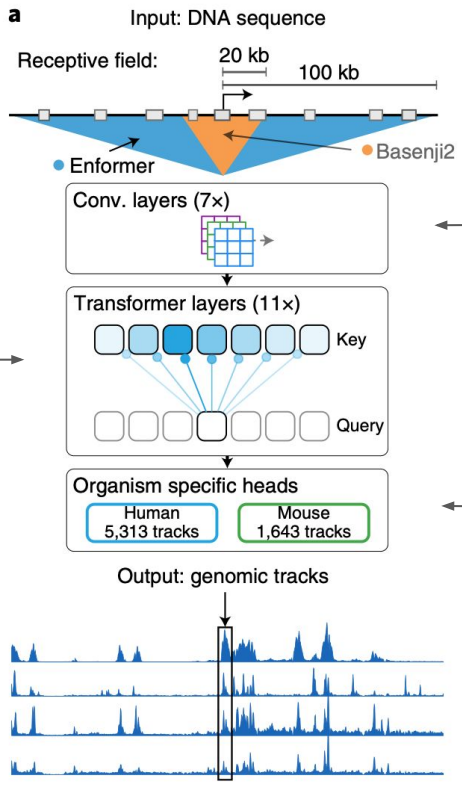
Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res 28, 739–750 (2018).

Model

- A self-attention based sequence-to-sequence model (transformer)
- Enformer = enhancer + transformer

Architecture

Architecture



Capture TF motifs, reduce length (196 kbp → 1,536 bp)

Share information across **entire** sequence

Separate output parameters for mouse, human sequences

Predictions represent 128 bp bins (length 896)

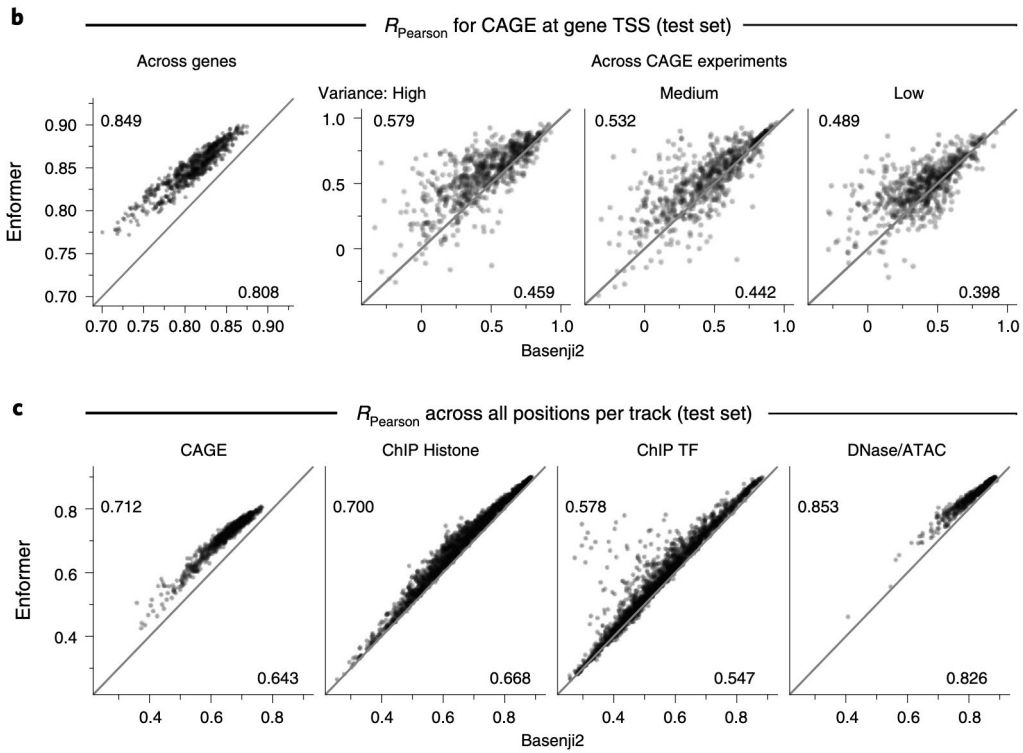
Transformer layers

- Standard self-attention operations: keys, queries, values
- Attention based on keys/queries, and positional embeddings
- Modified positional embeddings
  - Standard NLP approach did not scale to large distances
  - Additive with key/query dot product (as in TransformerXL)
  - Combined several distance functions (exponential, central mask, Gamma)

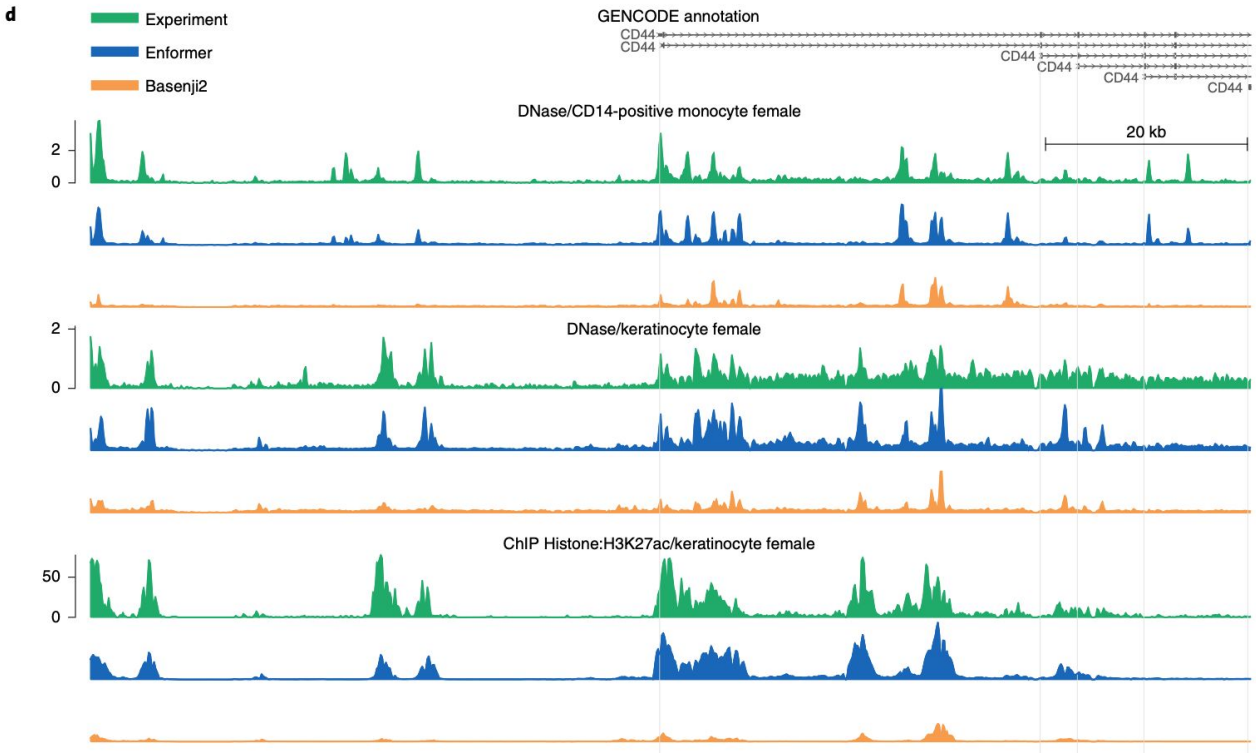
Key differences

- Relative to DNA sequence models: large receptive field via self-attention, more channels, longer input (vs. Basenji)
- Relative to NLP: positional embeddings designed for long-range interactions

Results


b
$R_{Pearson}$ for CAGE at gene TSS (test set)
Across genes
Across CAGE experiments
Variance: High
Medium
Low
Enformer
Basenji2
0.849
0.808
0.579
0.459
0.532
0.442
0.489
0.398
c
$R_{Pearson}$ across all positions per track (test set)
CAGE
ChIP Histone
ChIP TF
DNase/ATAC
0.712
0.643
0.700
0.668
0.578
0.547
0.853
0.826

Results (qualitative)



Ablations

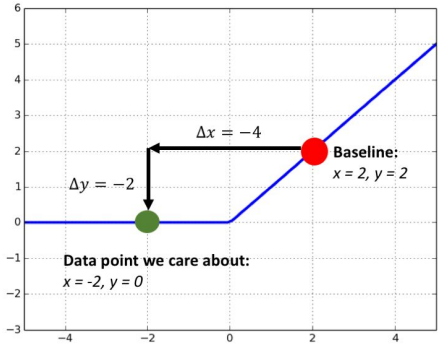
How to attribute performance improvement?

- Enformer outperforms Basenji (dilated convolutions)
- Consistent across various model sizes, # layers, # data points
- Reducing receptive field size hurts performance
- Standard positional embeddings hurt performance

interpretation pf the model

- Gradient based
- Attention based

$$y = ReLU(x) = max(0, x)$$

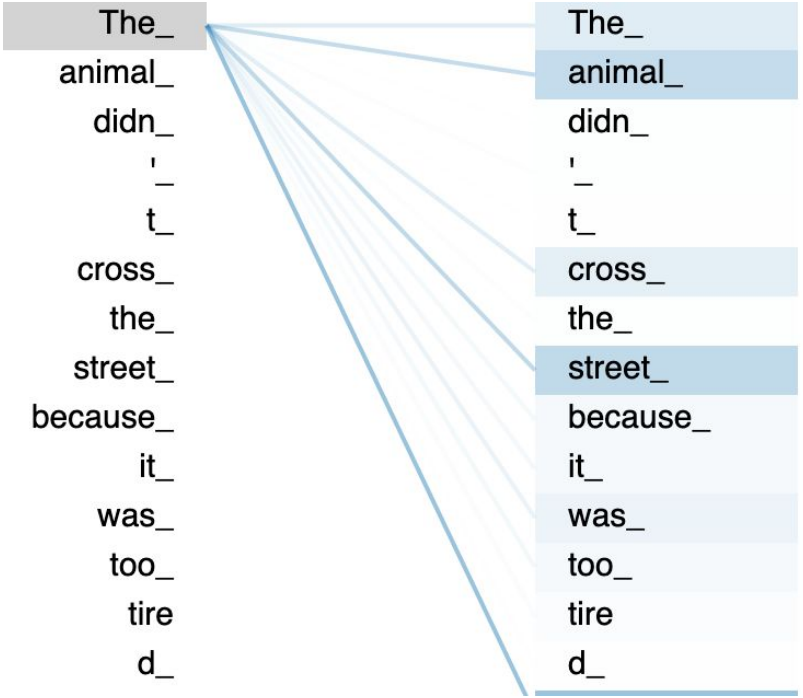


1. Calculating the slope

$$\frac{\Delta y}{\Delta x} = \frac{-2}{-4} = 0.5$$

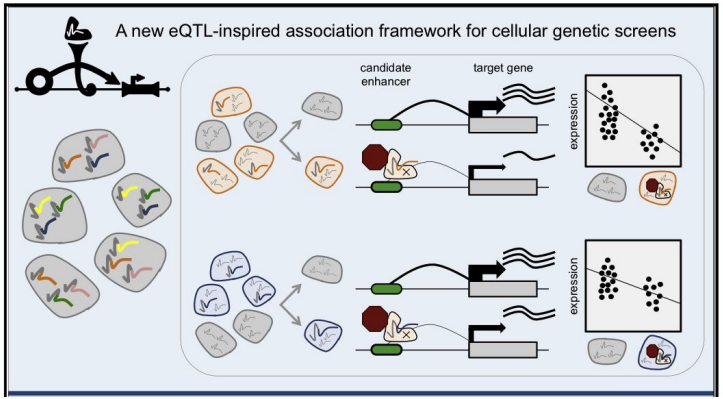
2. Finding the feature importance

$$\Delta x \times \frac{\Delta y}{\Delta x} = -4 \times 0.5 = -2$$



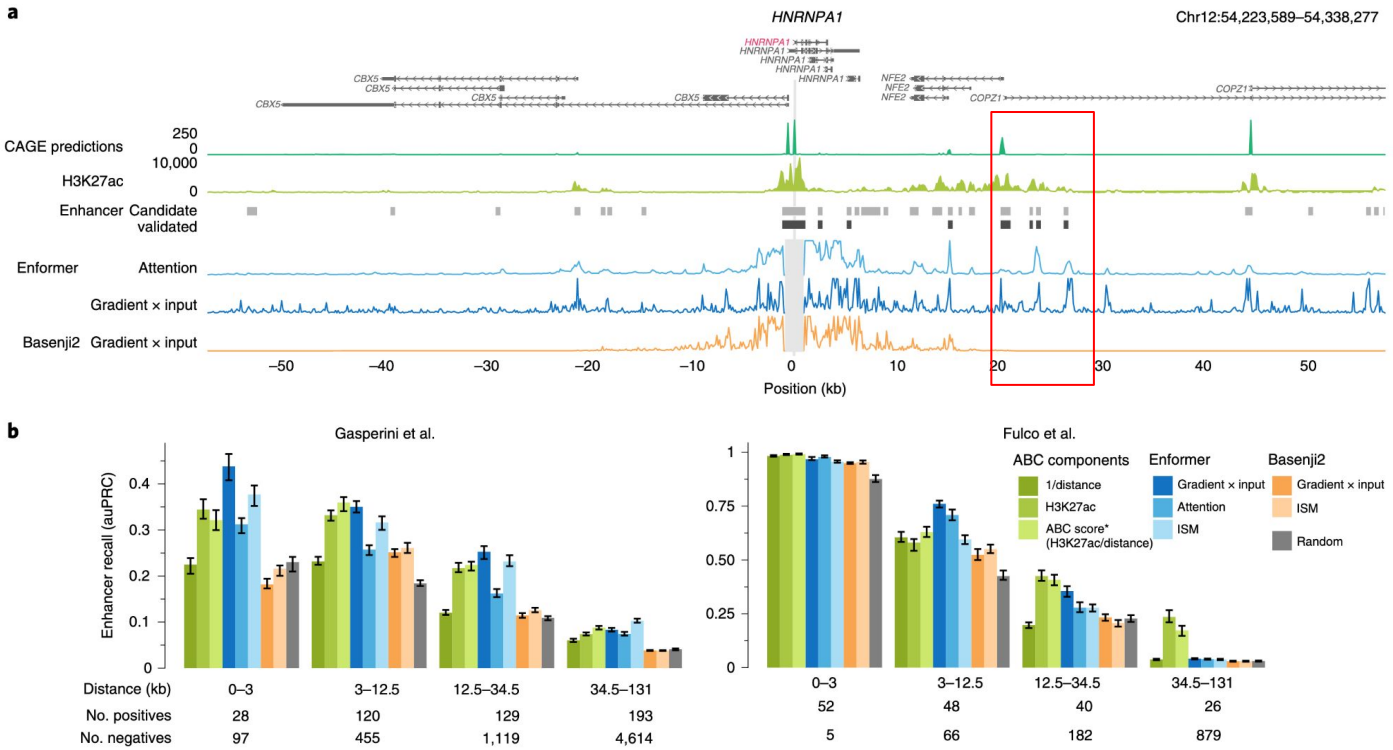
Enhancer prioritization

- CRISPRi assay perturbing the enhancer of interest while measuring the expression change of the gene



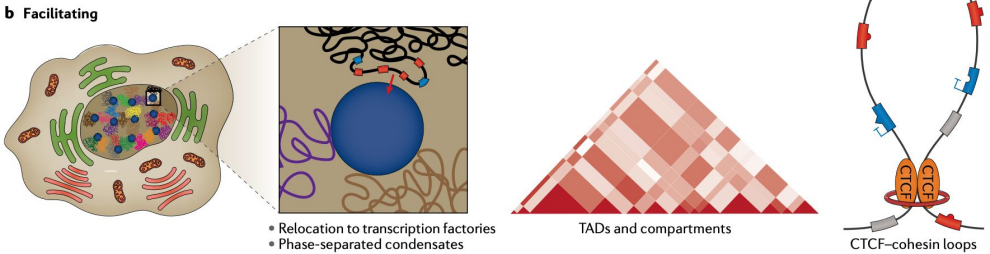
Gasperini, M. et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 377-390.e19 (2019).

Enformer attends to cell-type-specific enhancers

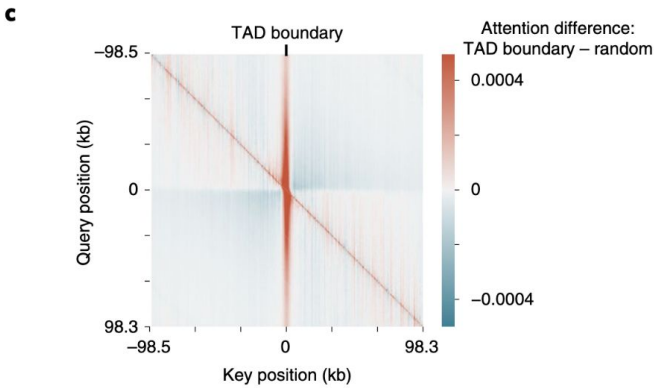


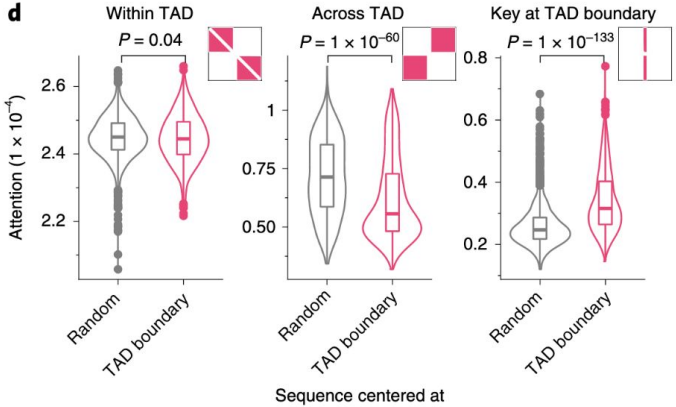
Insulator

- TAD: Topological Assoiaction Domain

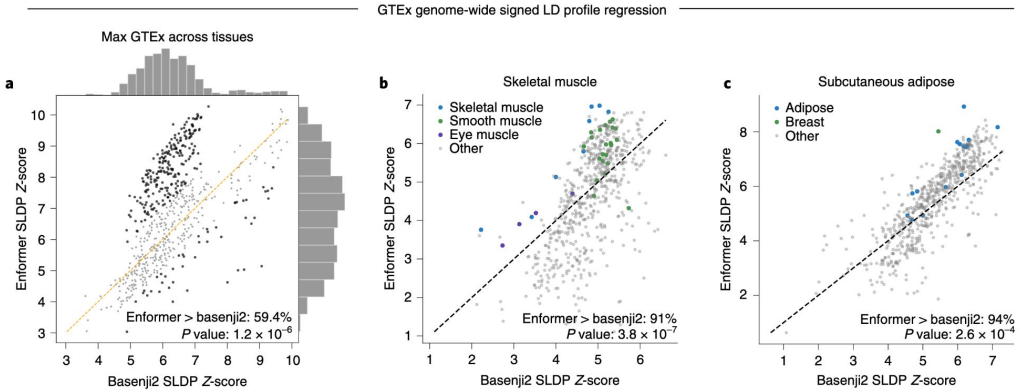


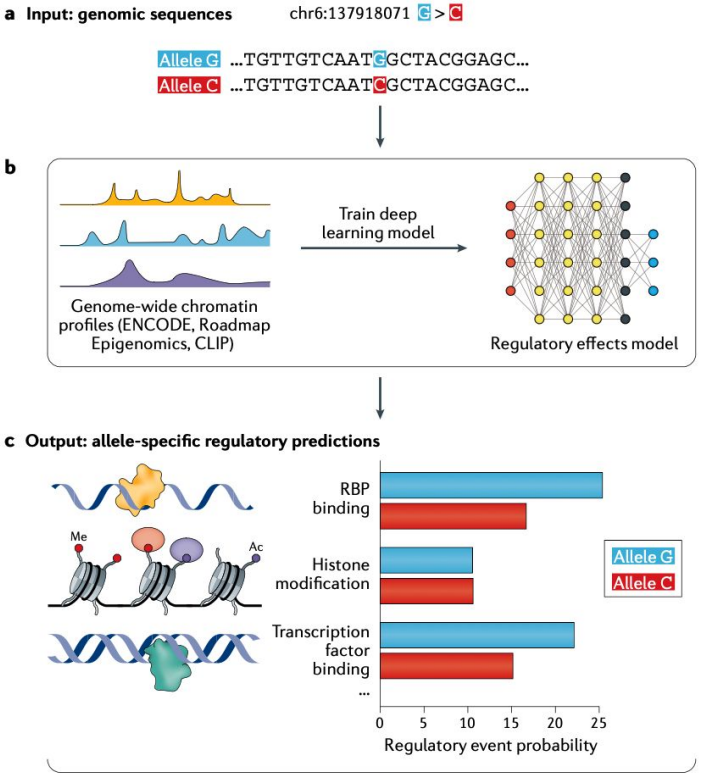
Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. Nat Rev Genet 20, 437–455 (2019).





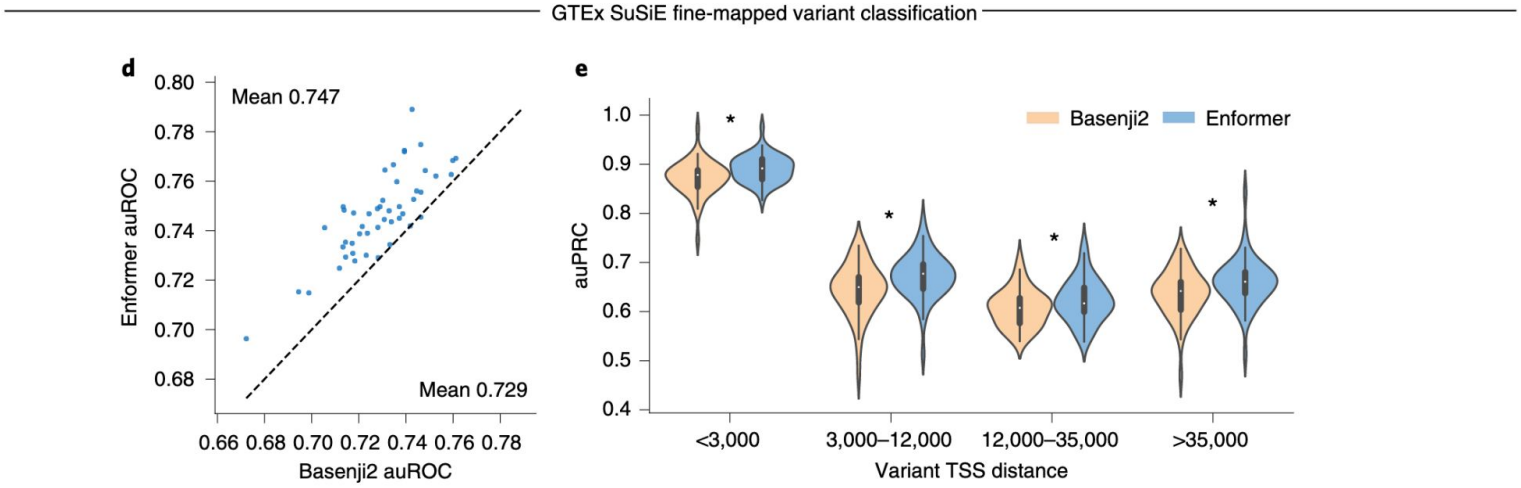
Variant effect prediction on eQTL data.



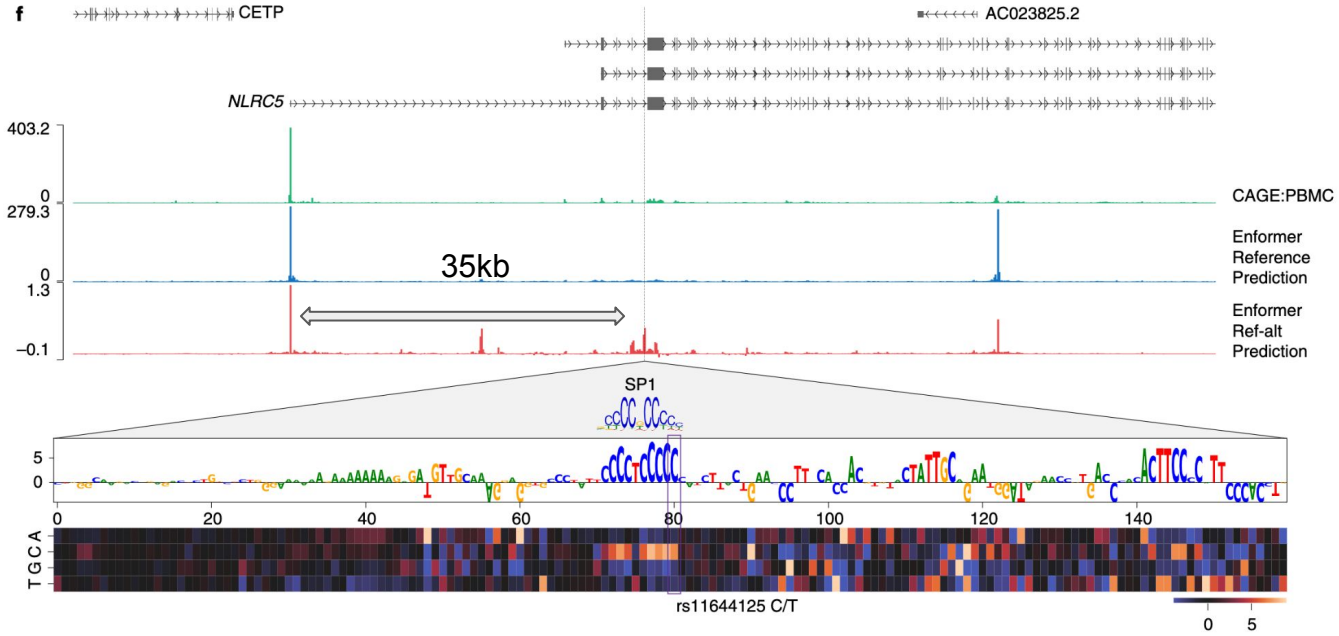


Wong, A. K., Sealfon, R. S. G., Theesfeld, C. L. & Troyanskaya, O. G. Decoding disease: from genomes to networks to phenotypes. *Nat Rev Genet* 1–17 (2021) doi:10.1038/s41576-021-00389-x.

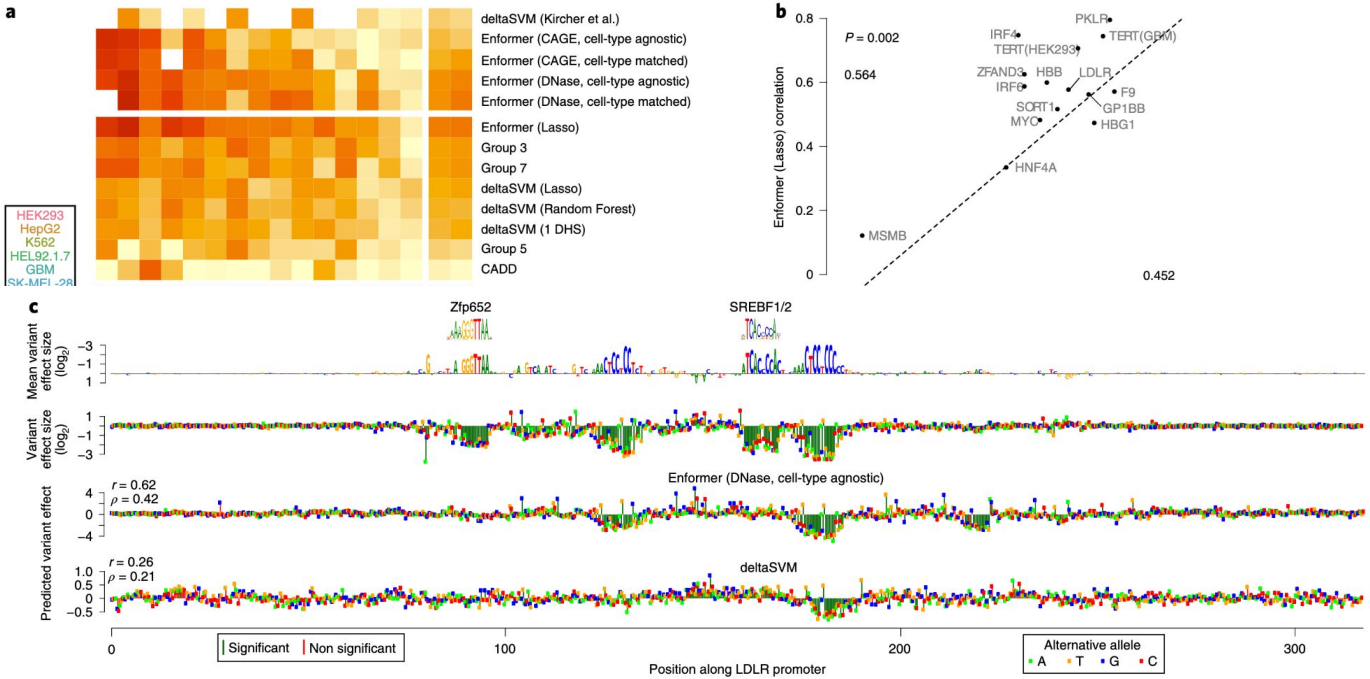
Variant effect prediction on eQTL data.



Variant effect prediction on eQTL data.



Enformer improves MPRA mutation effect prediction



Take-home message

- Self-attention can extend the receptive field
- Self-attention can capture long-distance interection
- Enformer can help us build a more comprehensive regulatory map than before

Discussion

- Cis-regulatory, how about trans-regulatory?
- Covariates matrix - like gene expression(Expecto)?
- Opportunities for large pre-trained models in genomics?