# Units of language in the noncoding *cis-regulatory genome:*

- Sentences = *enhancers* = control levels of gene expression which define cell state
- Words = *transcription factor binding sites* = within enhancers control enhancer activity

*Chromatin accessibility* is a measurement that signals enhancer activity.

Given chromatin accessibility measurements, how do we predict which transcription factors are most likely to master regulators?

Goal: Learn and experimentally validate the relationship between DNA sequence and chromatin accessibility

**Problem:** Hashimoto et al. 2016 showed that cell type-specific accessible regions are more difficult to predict than regions that are universally accessible across many cell types.

**Solution:** *DeepAccess* An Ensemble of Boosted Convolutional Neural Networks
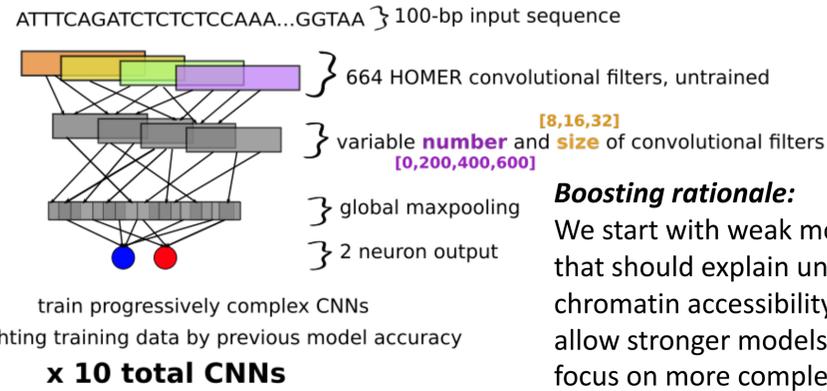
step 1: measure DNA accessibility, call 100-bp accessible regions



Accessibility Stem Cells (ESC)

Accessibility Definitive Endoderm (DE)
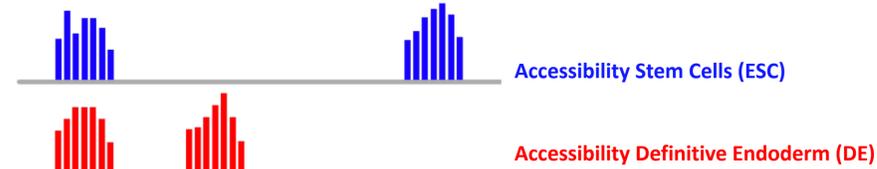
step 2: obtain DNA sequence from accessible regions with binary classification

Open ESC  Open DE  400,000 train (100,000 per class)  40,000 test (10,000 per class)

---

step 3: build ensemble of convolutional neural networks

ATTTCAGATCTCTCTCCAAA...GGTAA } 100-bp input sequence

664 HOMER convolutional filters, untrained

[8,16,32] variable **number** and **size** of convolutional filters [0,200,400,600]

global maxpooling

2 neuron output

train progressively complex CNNs
re-weighting training data by previous model accuracy

**x 10 total CNNs**

*Boosting rationale:* We start with weak models that should explain universal chromatin accessibility and allow stronger models to focus on more complex tasks.

**Problem:** How can we extract differentially accessible sequence features, such as transcription factor motifs?

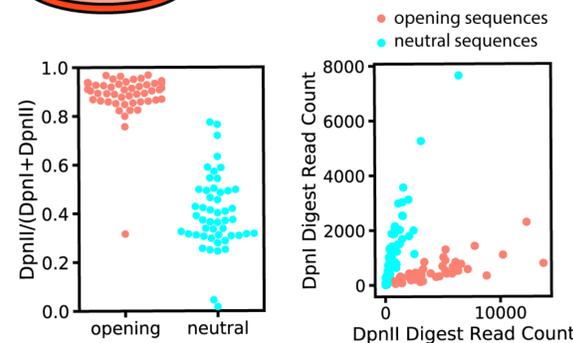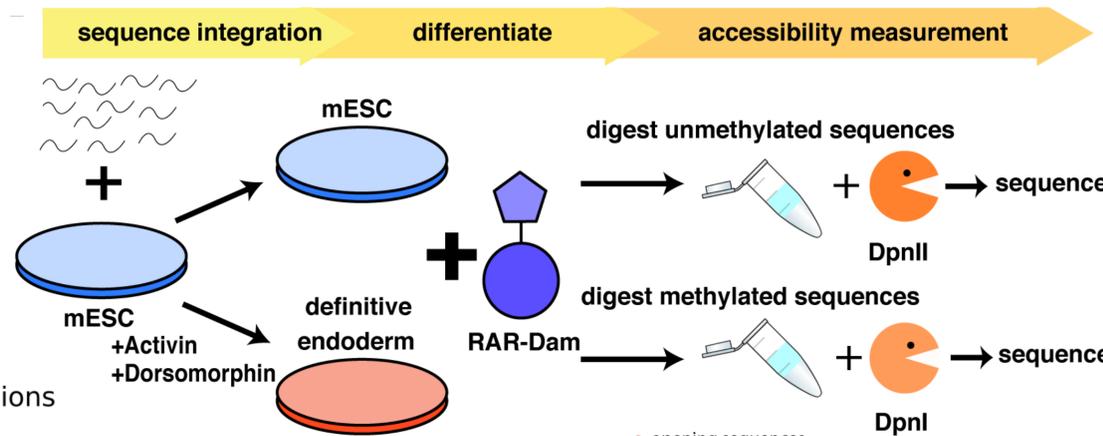**Solution:** Class discriminative saliency maps

$$S_c(X) = w^T X + b = \sum_{i=1}^{|X|} w_i x_i + b$$

$$w = \left.\frac{\partial S_{c_j} - S_{c_i}}{\partial X}\right|_{X_0} + \textit{noise in X}$$
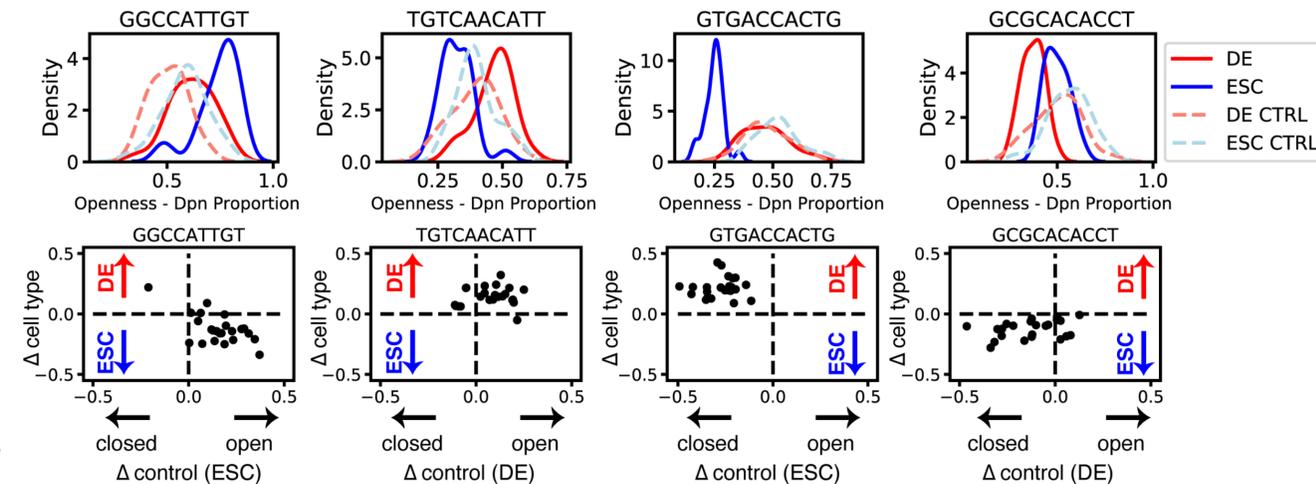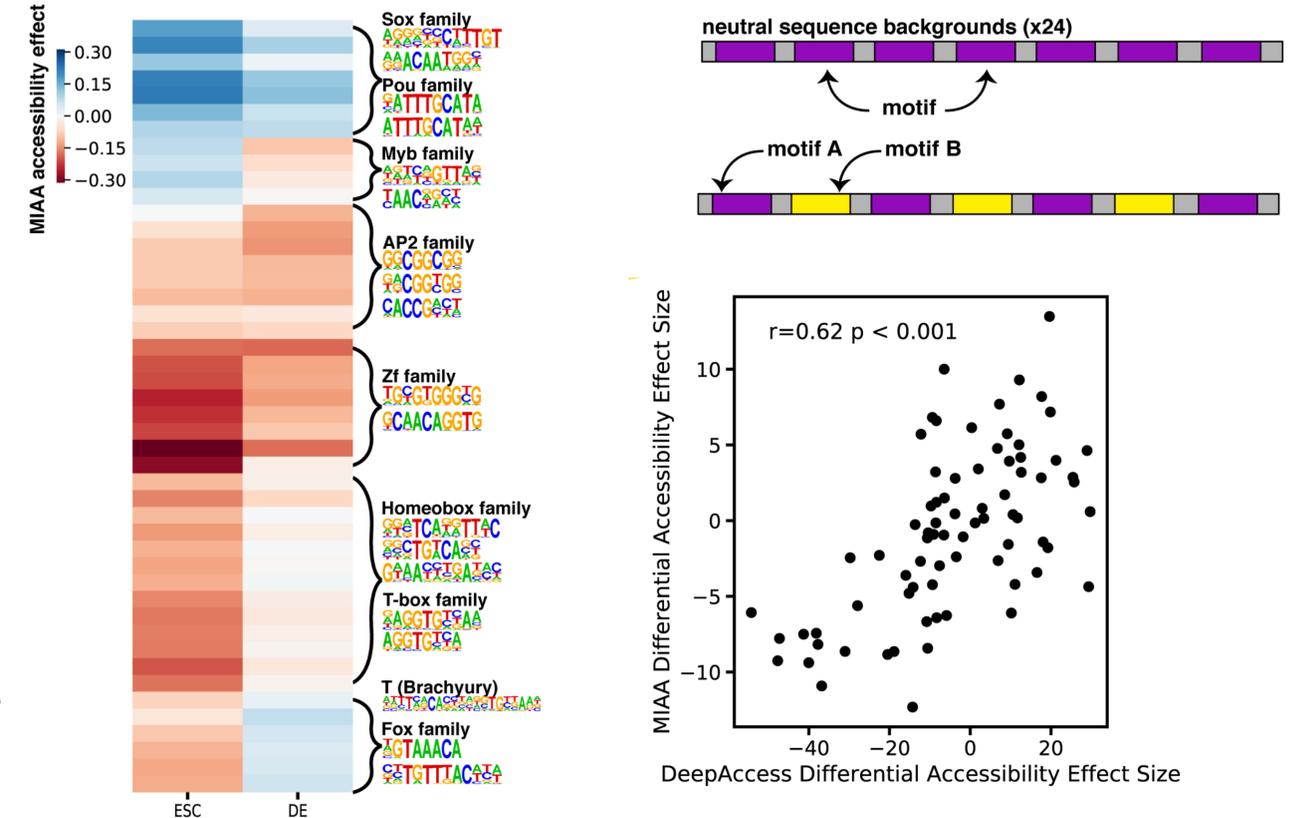*(SmoothGrad)*

*Motif extraction:* 10bp windows with highest average class discriminative saliency are selected as putative transcription factor motifs

**Problem:** How do we evaluate?

**Solution:** Multiplexed Integrated Accessibility Assay (MIAA)

sequence integration  differentiate  accessibility measurement



digest unmethylated sequences → DpnII → sequence
digest methylated sequences → DpnI → sequence

mESC  +Activin +Dorsomorphin → definitive endoderm  RAR-Dam



opening sequences
neutral sequences

---

## DeepAccess derived DNA sequence features can drive strongly cell type-specific chromatin accessibility



MIAA accessibility effect: 0.30 0.15 0.00 -0.15 -0.30

Sox family
Pou family
Myb family
AP2 family
Zf family
Homeobox family
T-box family
T (Brachyury)
Fox family

ESC   DE

neutral sequence backgrounds (x24)

motif

motif A  motif B



r=0.62 p < 0.001

MIAA Differential Accessibility Effect Size

DeepAccess Differential Accessibility Effect Size



GGCCATTGT  TGTCAACATT  GTGACCACTG  GCGCACACCT

DE  ESC  DE CTRL  ESC CTRL

Density — Openness - Dpn Proportion

Δ cell type — Δ control (ESC) / Δ control (DE) — closed / open

## What should I take away?

- DeepAccess allows us predict DNA motifs associated with cell type-specific chromatin accessibility
- Multiplexed Integrated Accessibility Assay provides a way to measure sequence effects in a controlled context

## Ask me about...

- How else can you use DeepAccess to understand the genome?
- Were there any surprising results you found with this new accessibility assay?