STANFORD SCHOOL of MEDICINE | Genetics

STANFORD COMPUTER SCIENCE

Interpretable, integrative deep learning models for regulatory genomics and epigenomics

Anshul Kundaje


Avanti Shrikumar


Nathan Boley


Johnny Israeli


Peyton Greenside


Chuan Sheng Foo

Predicting in-vivo TF ChIP-seq binding events at chromatin accessible sites

Output: Bound (+1) vs. not bound (0)

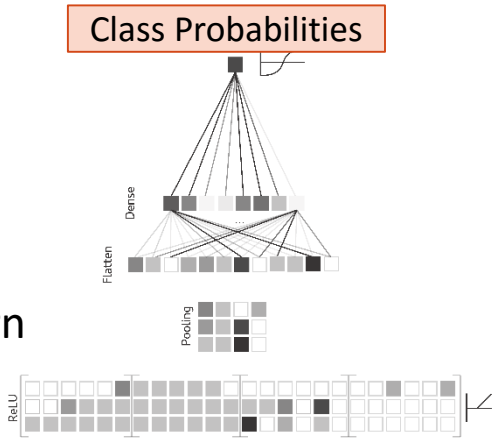
Binary classification problem

- **Positive set:** 1Kb sequences overlapping reproducible target TF ChIP-seq peaks in specific cell type

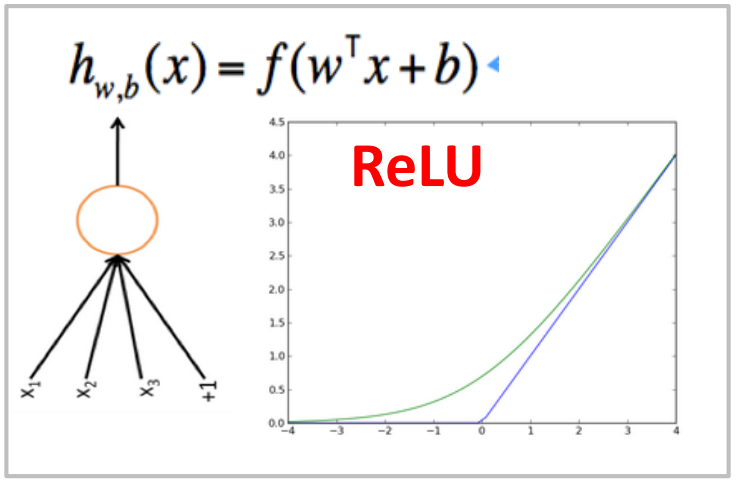
- **Negative set:** 1Kb sequences overlapping **all chromatin accessible sites** that do not overlap target TF ChIP-seq peaks

Input: Raw DNA sequence

Convolutional neural network (CNN) learning from raw DNA sequence

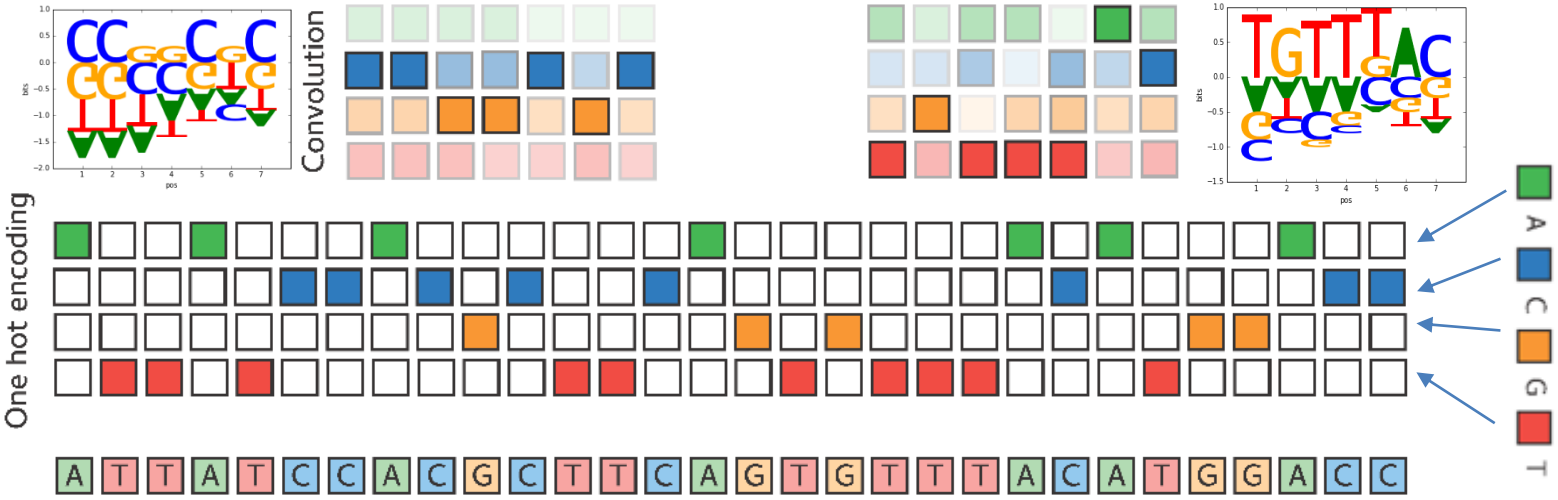


Higher conv. layers learn motif combinations

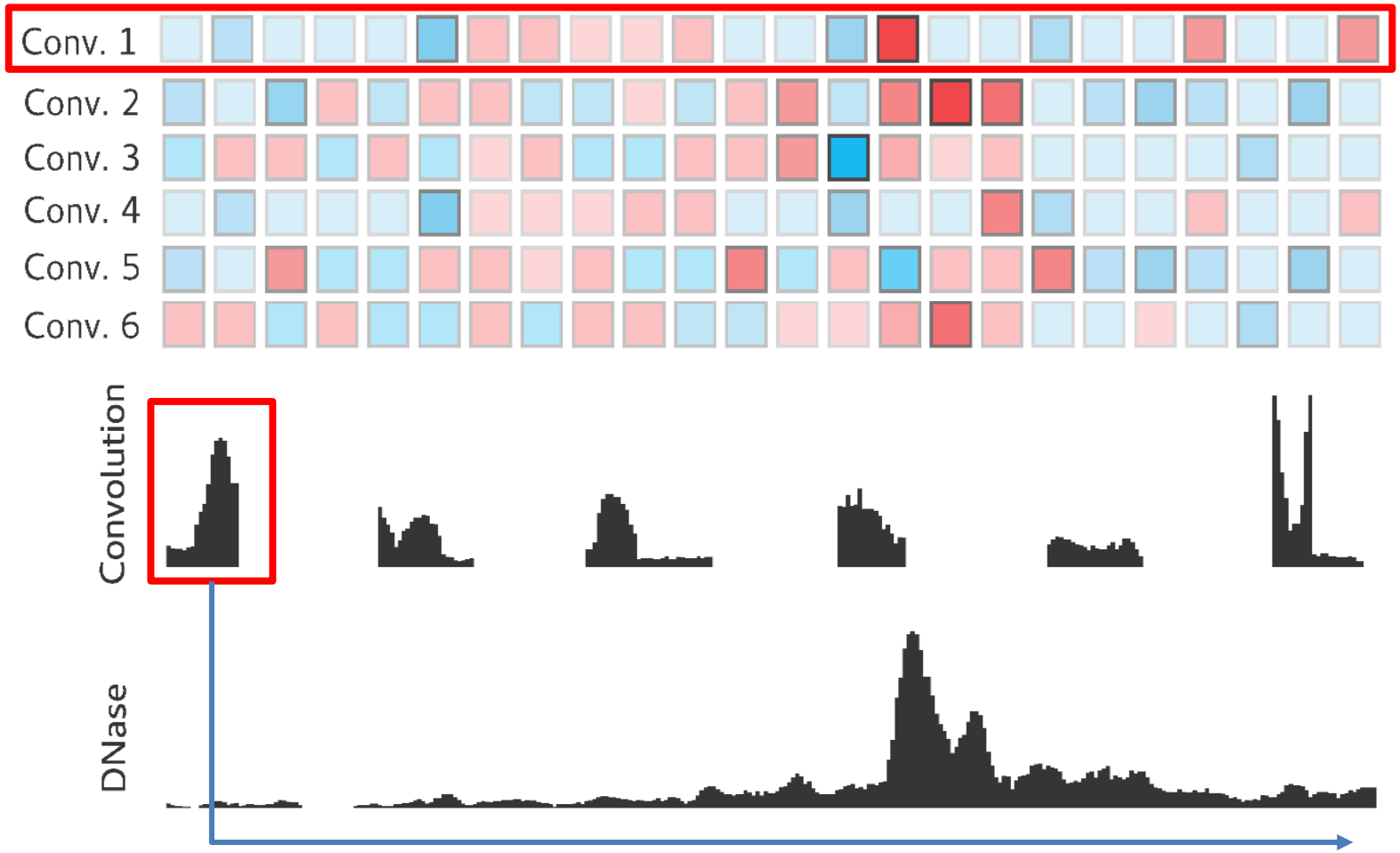


Score sequence using filters

Convolutional layers learn motif (PWM) like filters



CNN for learning from 1D genomic signal profiles (e.g. Dnase-seq, MNase-seq, ATAC-seq)



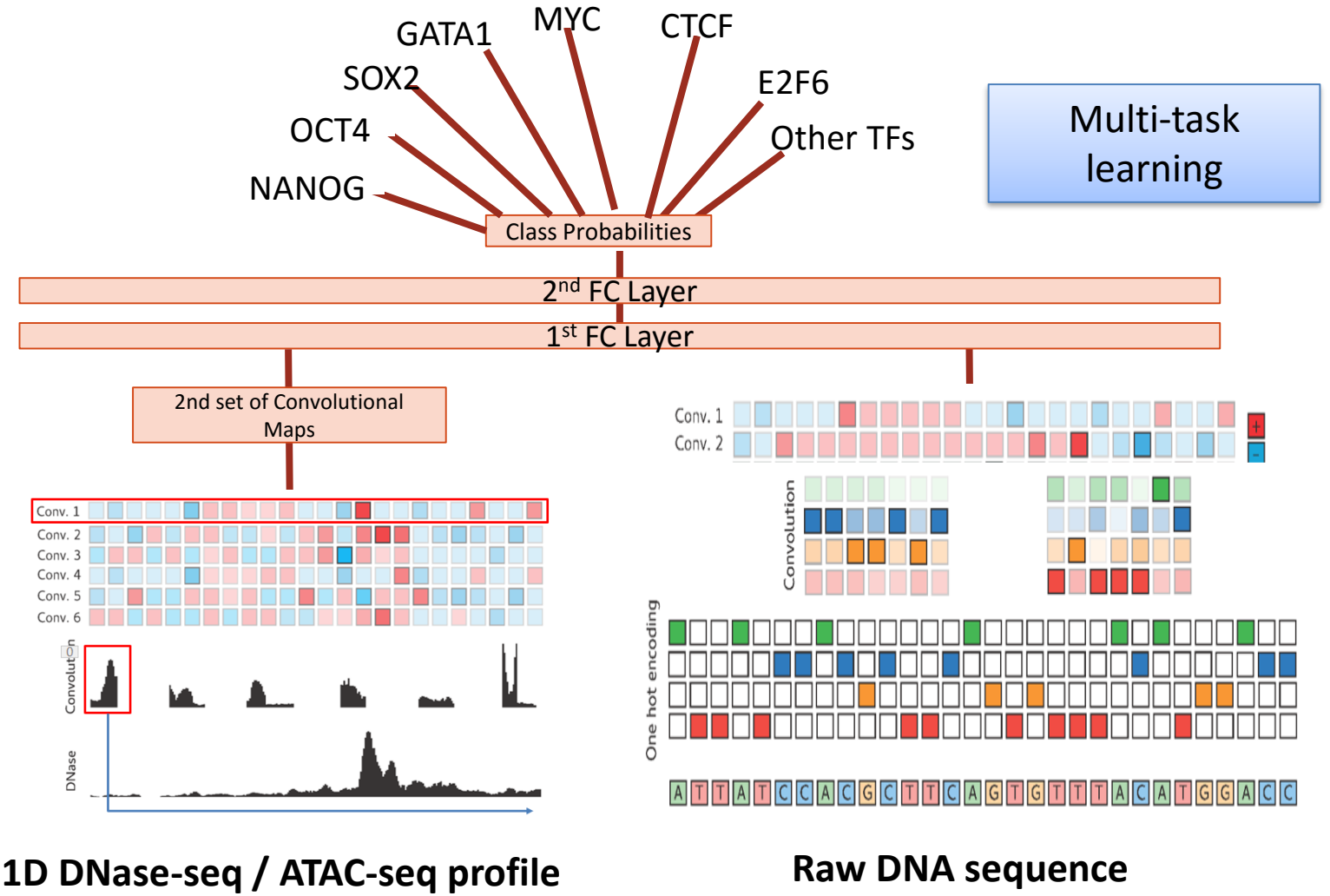
Scan DNase profile using filter

THE CHROMPUTER

Integrating **multiple inputs** (1D profiles, sequence) to simultaneously **predict multiple outputs**

Nathan Boley

Johnny Israeli

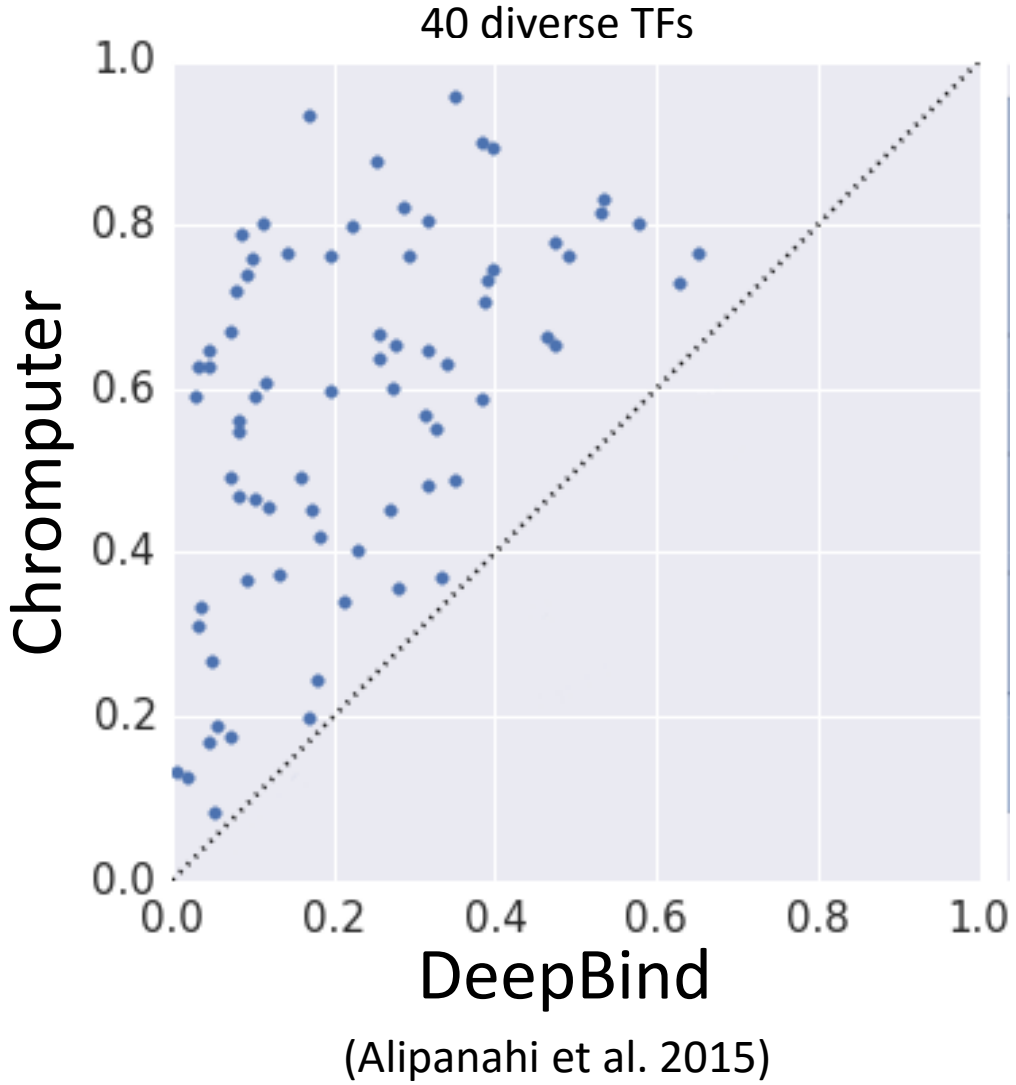


1D DNase-seq / ATAC-seq profile

Raw DNA sequence

Model performance: cross cell-type prediction (held-out cell type + chromosome)

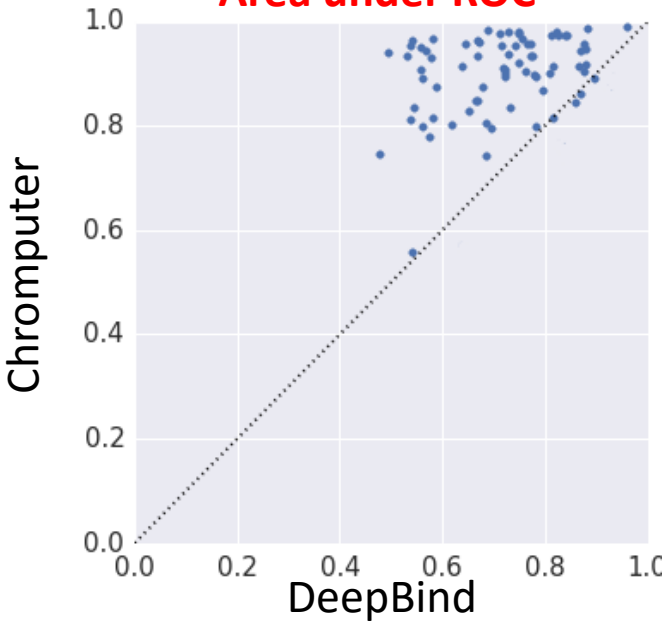
Area under Precision-Recall Curve



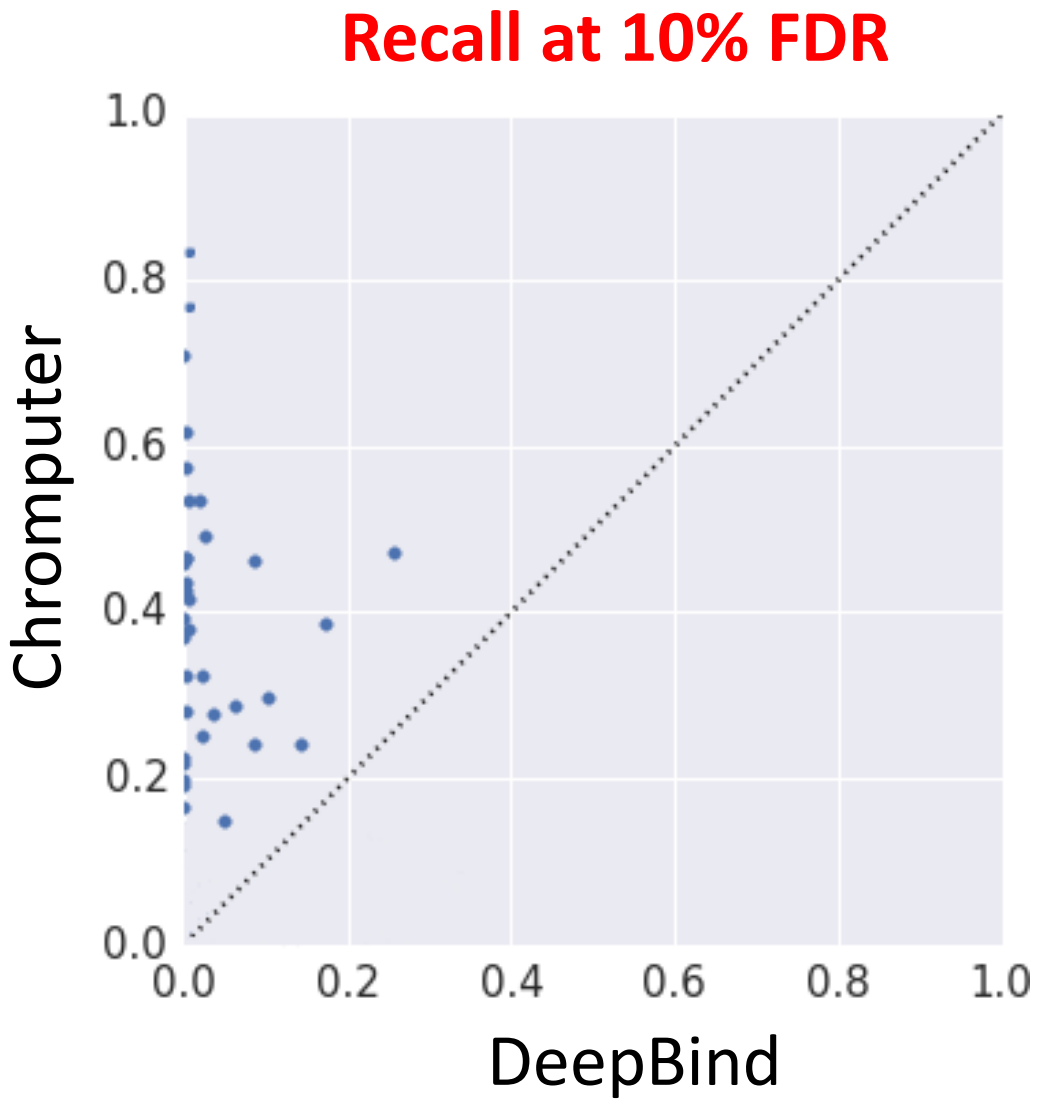
(Alipanahi et al. 2015)

- Prediction task is highly unbalanced (5-10x more negatives than positives)
- **auROC is highly misleading for unbalanced data!**

Area under ROC



Model performance: cross cell-type prediction (held-out cell type + chromosome)



- Why does DeepBind do so poorly in this setting?
  - Trains on dinucleotide shuffled negatives (not representative of relevant genomic background)
  - **Negative set matters**

Model interpretation

- Q's we will try to answer:
  - **Motif discovery:** Primary motifs and cofactor motifs?
  - **Learn sequence grammars:** homotypic/heterotypic co-binding events, density and spacing of motifs
  - **Heterogeneity:** Are there different subsets of TF binding events with distinct sequence grammars?
  - From low resolution (~100-500 bp) peaks to **high-resolution point binding events**

DeepLIFT: Predictive power of features in Deep Neural Networks

Avanti Shrikumar

Which neurons/filters are predictive?

Which nucleotides in input sequence are contributing to binding

DeepLIFT: Predictive power of features in Deep Neural Networks

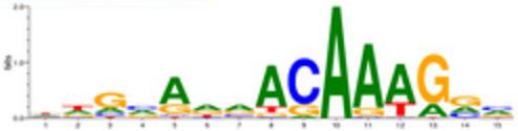
- Decomposition of contribution of each input to immediate outputs
  - ReLU networks: **piece-wise linear**
  - Recursively apply (with chain rule) to get contribution of any input to any output
  - Can be **computed efficiently** with a single backpropagation (unlike in-silico mutagenesis)
  - **Less susceptible to buffering effects** than in-silico mutagenesis
- Technical details:
  - Importance of any input to any output = gradient * input
  - Expands on classical sensitivity analysis proposed in Simonyan et al. 2014

Current motif discovery approaches produce multiple partially redundant motifs (e.g. Nanog)

MEME/HOMER
(Kheradpour et al.)

NANOG_disc1:

NANOG_disc2:

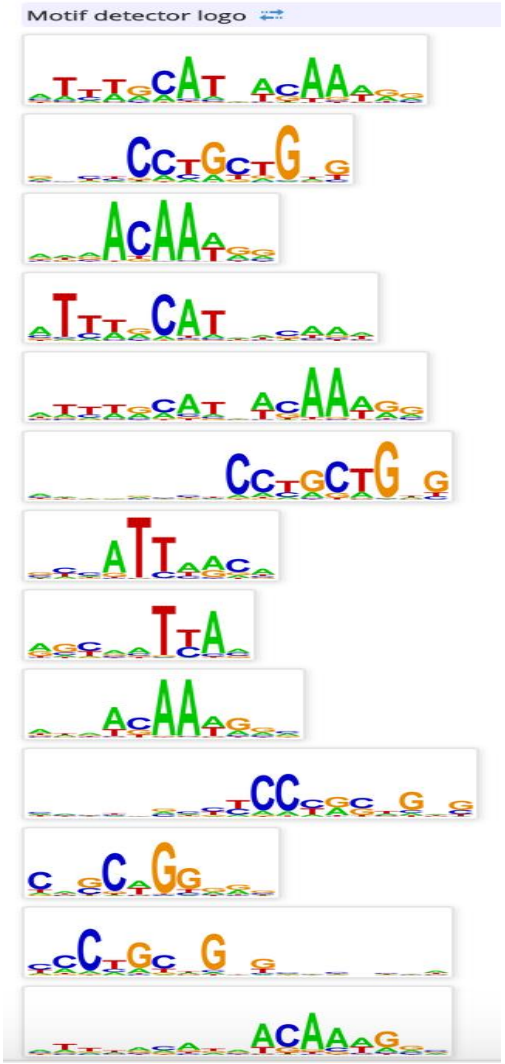


NANOG_disc3:

NANOG_disc4:

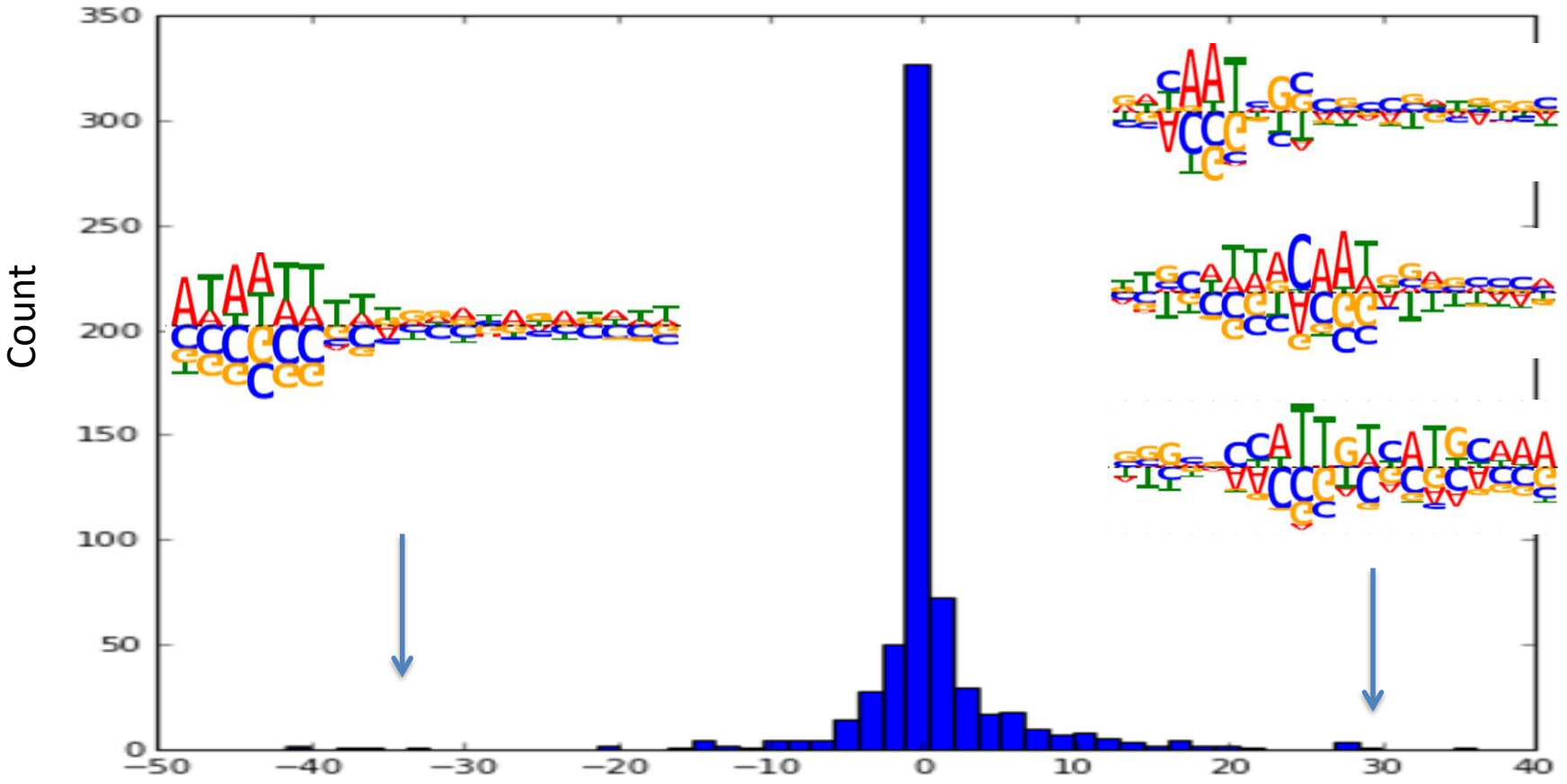
NANOG_known1:

DeepBind (Alipanahi et al.)



Motif discovery using DeepLIFT

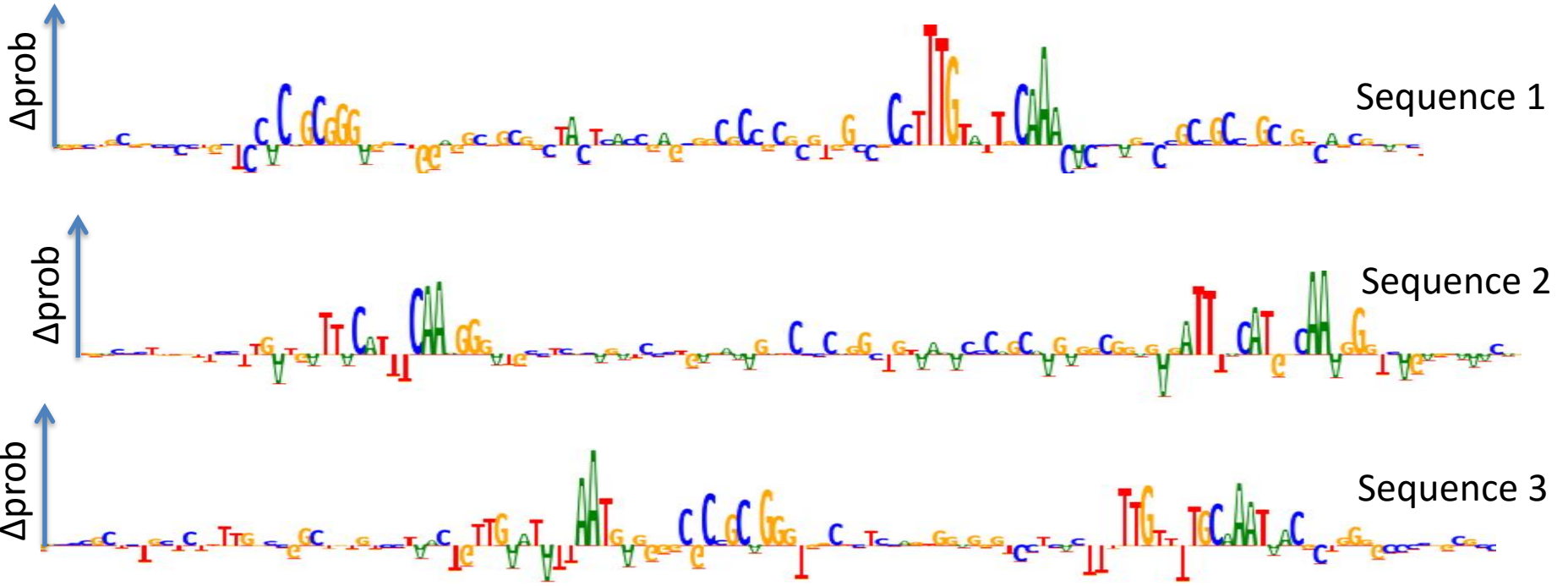
Each PWM like ‘filter’ in convolutional layers gets a deepLIFT score



Histogram of DeepLIFT scores of motif detectors from Nanog model

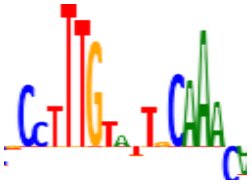
How do we combine the aggregate contribution of multiple filters at individual sequences?

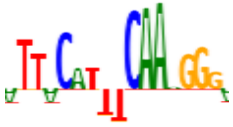
Insight: filter contributions are resolved at the nucleotide level



How do we combine the aggregate contribution of multiple filters at individual sequences?

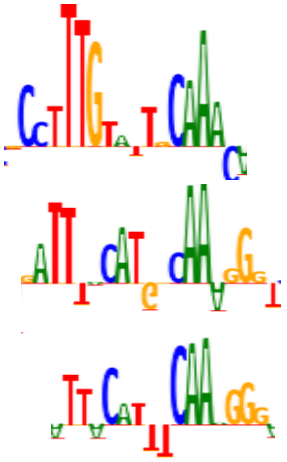
Insight: filter contributions are resolved at the nucleotide level





How do we combine the aggregate contribution of multiple filters at individual sequences?

Insight: filter contributions are resolved at the nucleotide level



How do we combine the aggregate contribution of multiple filters at individual sequences?

Insight: filter contributions are resolved at the nucleotide level

4 main non-redundant agglomerated DeepLIFT motifs

Aggregate DeepLIFT gradients

Underlying PWM

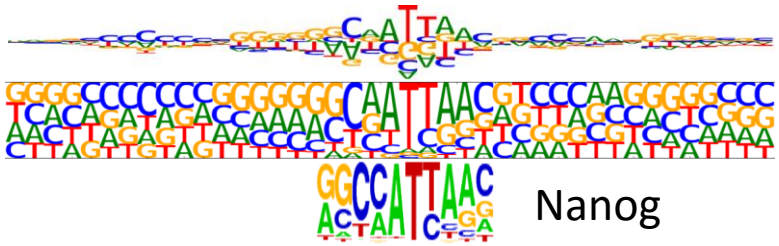
Homer match (Heinz et al.)

Oct-Sox-Nanog

Aggregate DeepLIFT gradients

Underlying PWM

Homer match (Heinz et al.)

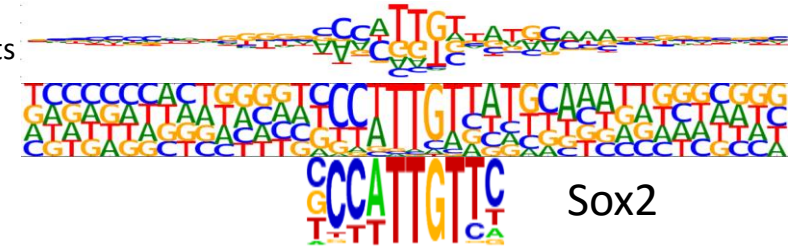


Nanog

Aggregate DeepLIFT gradients

Underlying PWM

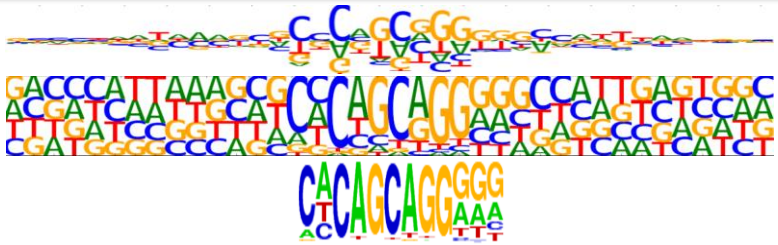
Homer match (Heinz et al.)



Sox2

Aggregate DeepLIFT

Underlying PWM



4 main non-redundant agglomerated DeepLIFT motifs

Aggregate DeepLIFT gradients

Underlying PWM

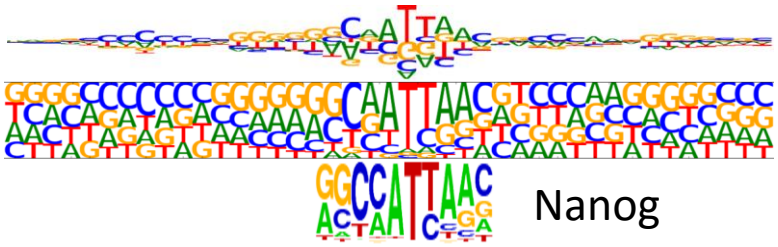
Homer match (Heinz et al.)

Oct-Sox-Nanog

Aggregate DeepLIFT gradients

Underlying PWM

Homer match (Heinz et al.)



Nanog

Agg

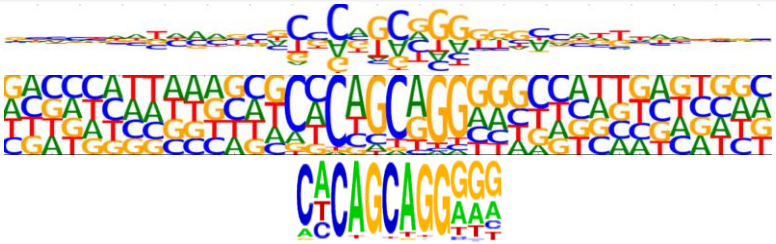
Zic3 Is Required for Maintenance of Pluripotency in Embryonic Stem Cells

Linda Shushan Lim,[*†] Yuin-Han Loh,[†‡] Weiwei Zhang,[‡§] Yixun Li,[∥] Xi Chen,[‡§] Yinan Wang,[‡§] Manjiri Bakre,[*] Huck-Hui Ng,[‡§] and Lawrence W. Stanton[*§]

(Heinz et al.)

Aggregate DeepLIFT

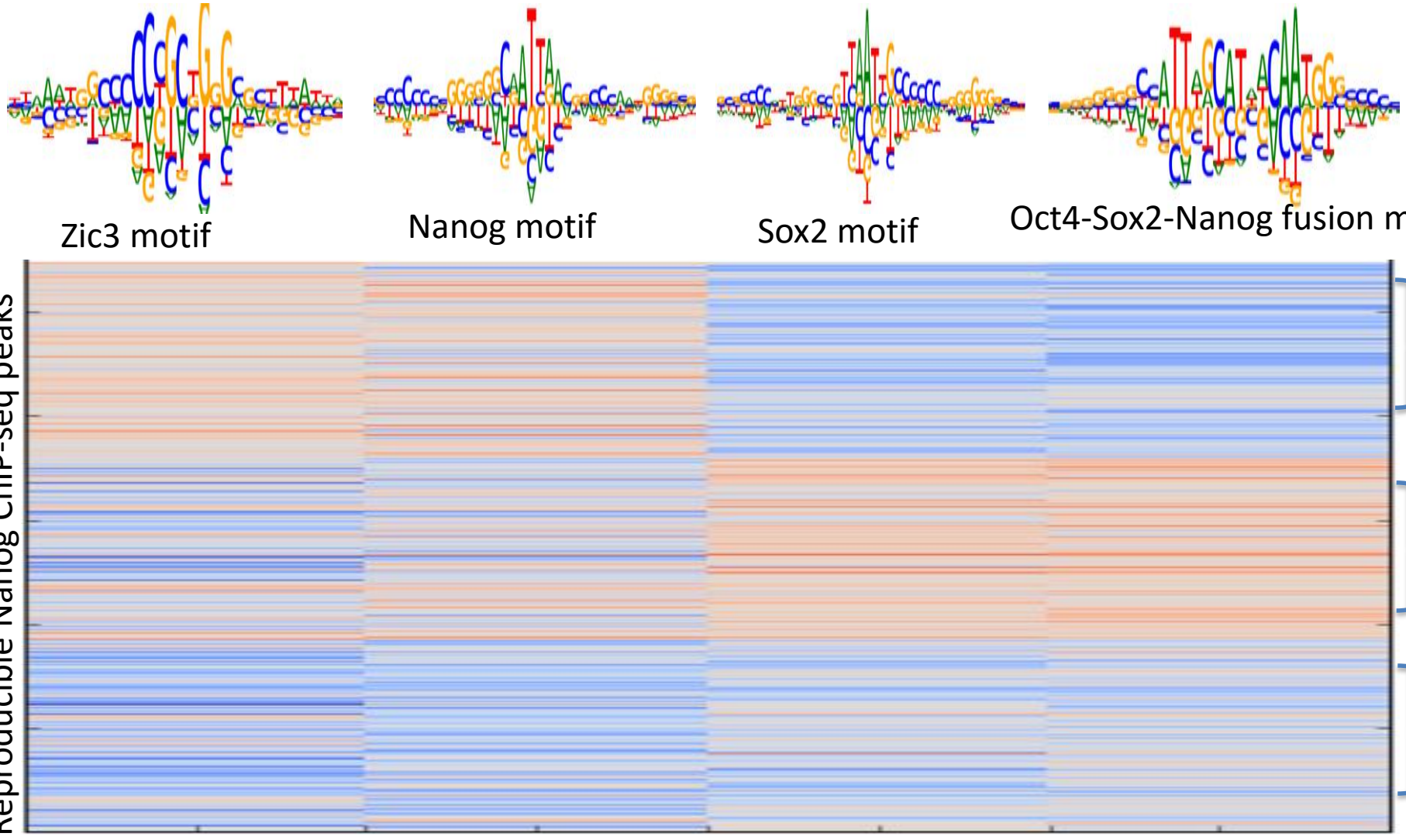
Underlying PWM



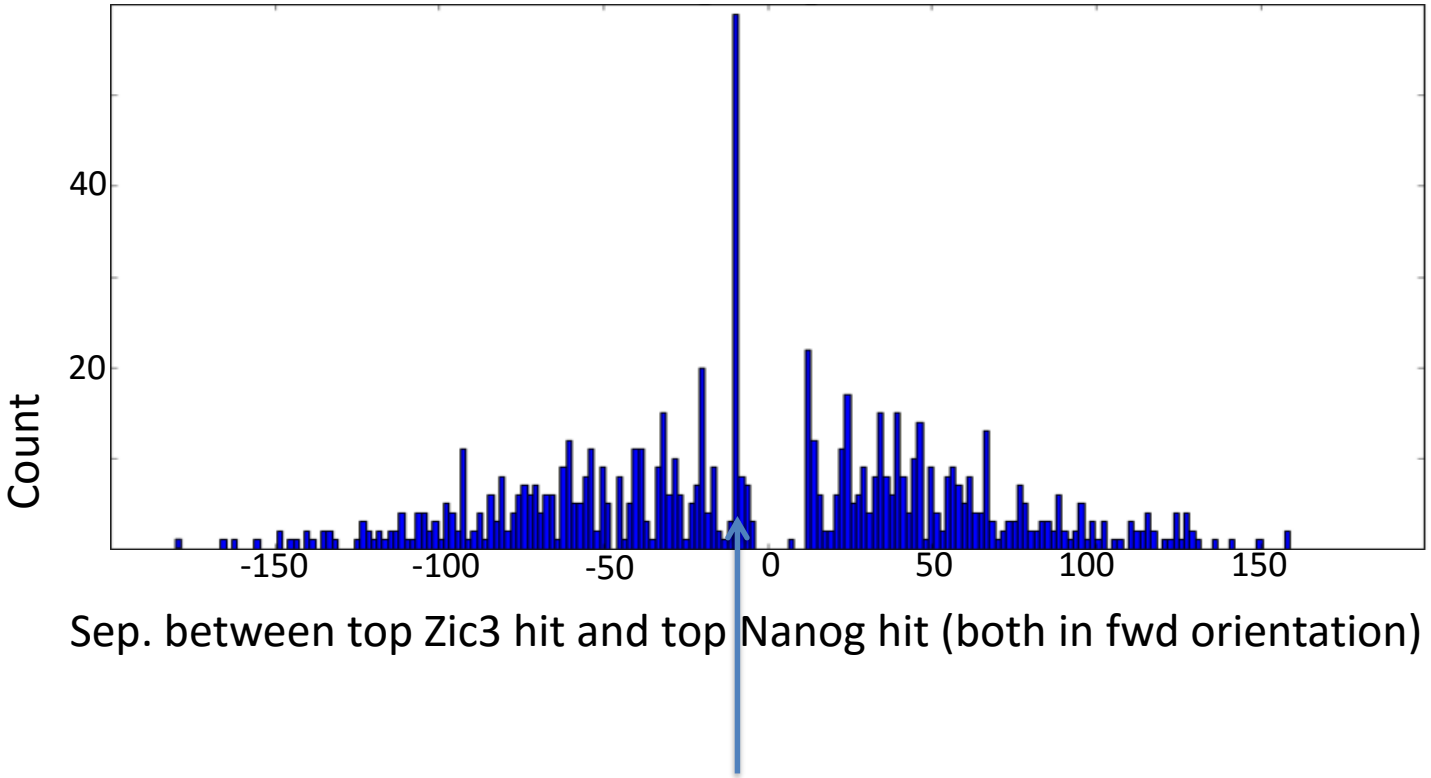
Zic3 (Jaspar; Zhao et al.)

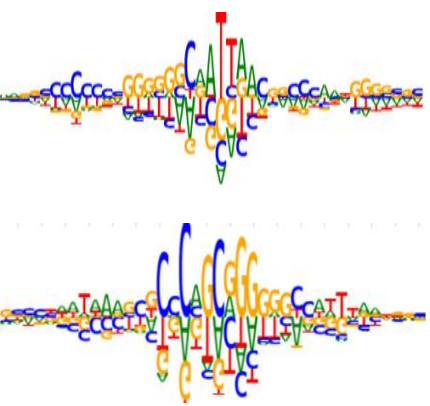
Heterogeneity

At least 3 distinct classes on Nanog sites

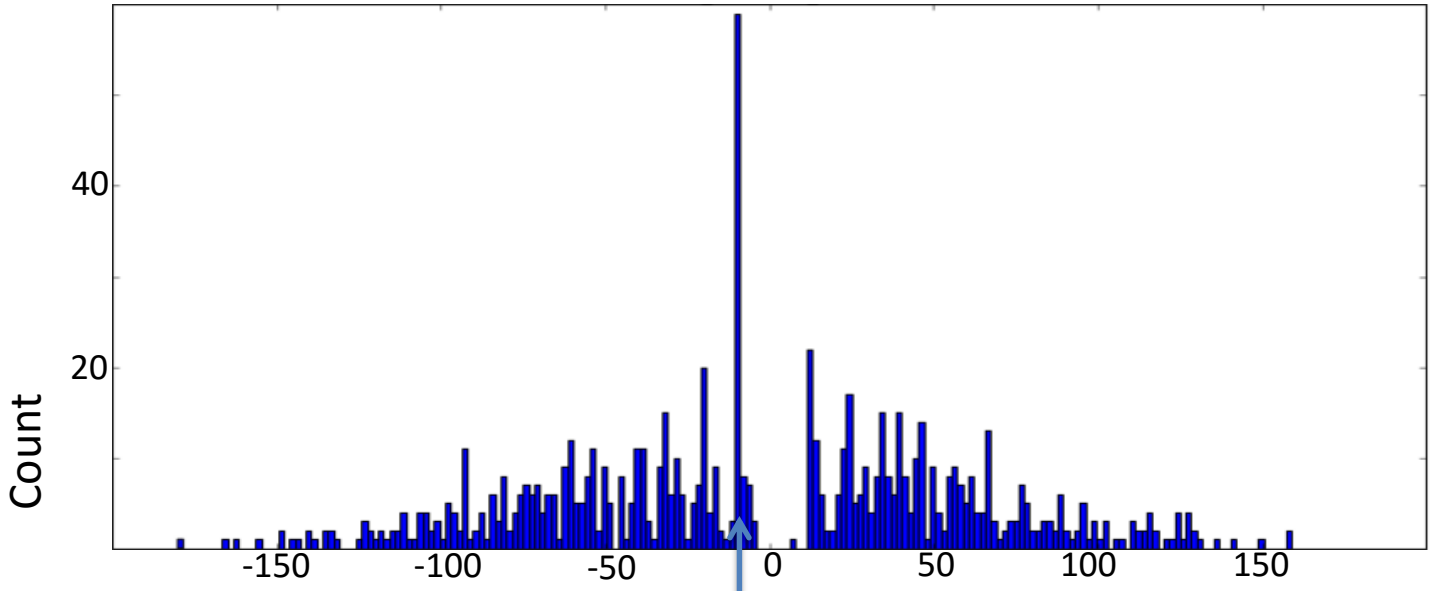


Sequence grammar involving Nanog and Zic3

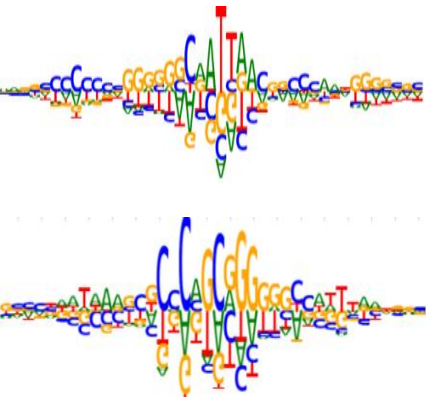


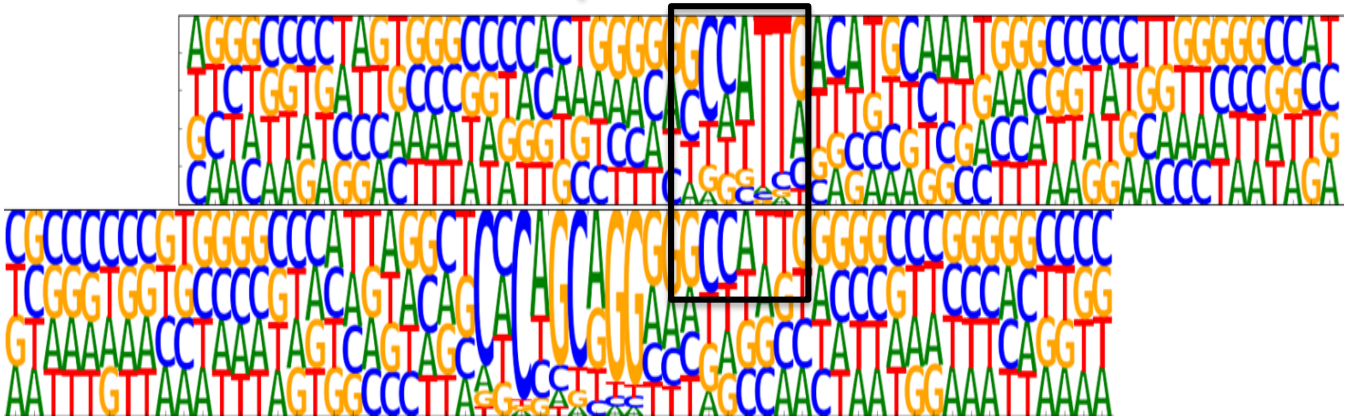


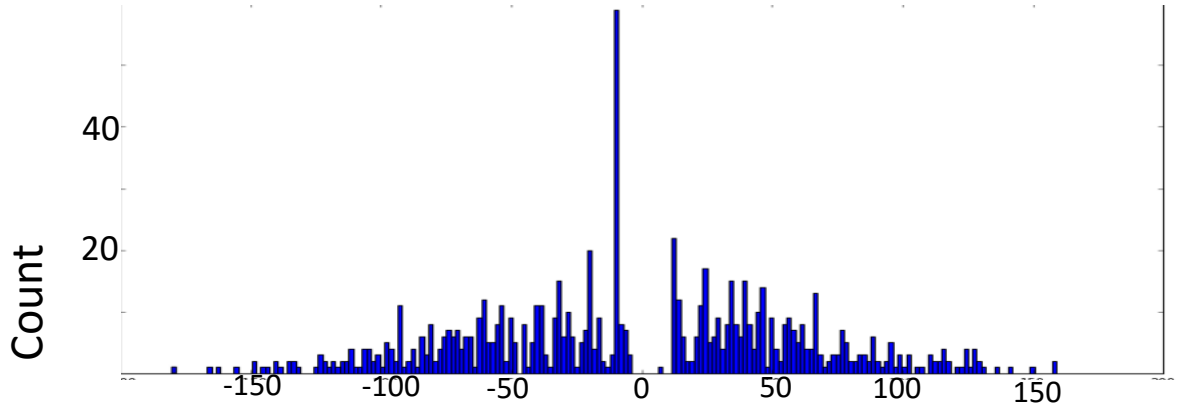
Sequence grammar involving Nanog and Zic3



10bp shift

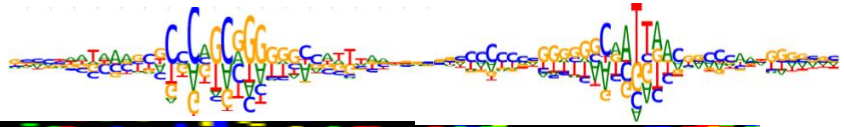


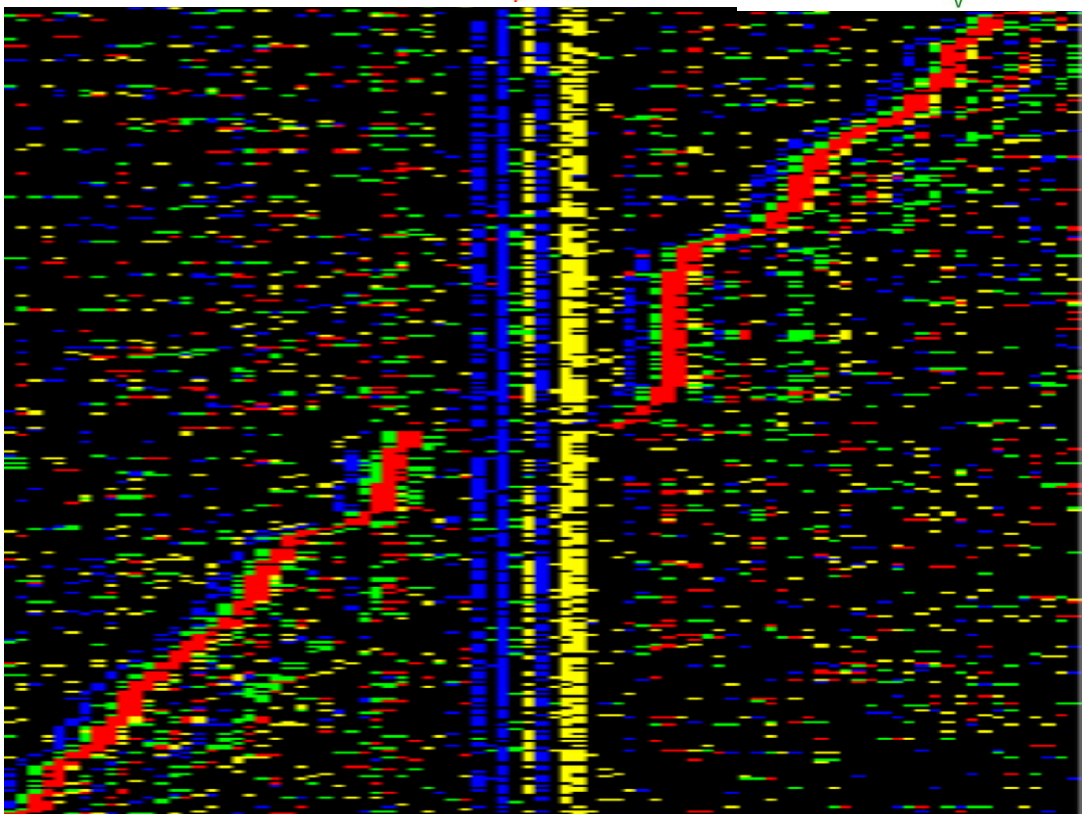




Sep. between top Zic3 hit and top Nanog hit (both in fwd orientation)

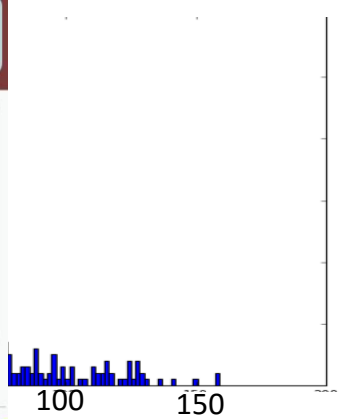
position



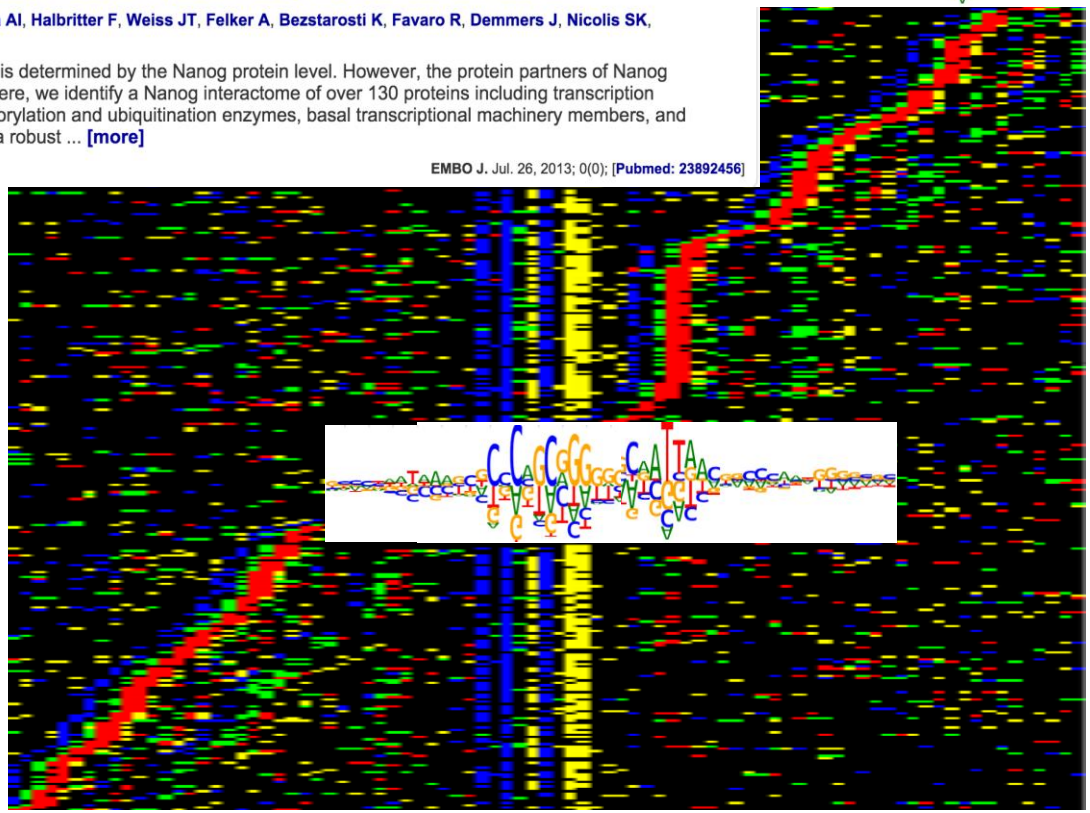


Regions around Zic3 motif sorted by separation

(Colored by base and deepLIFT score)



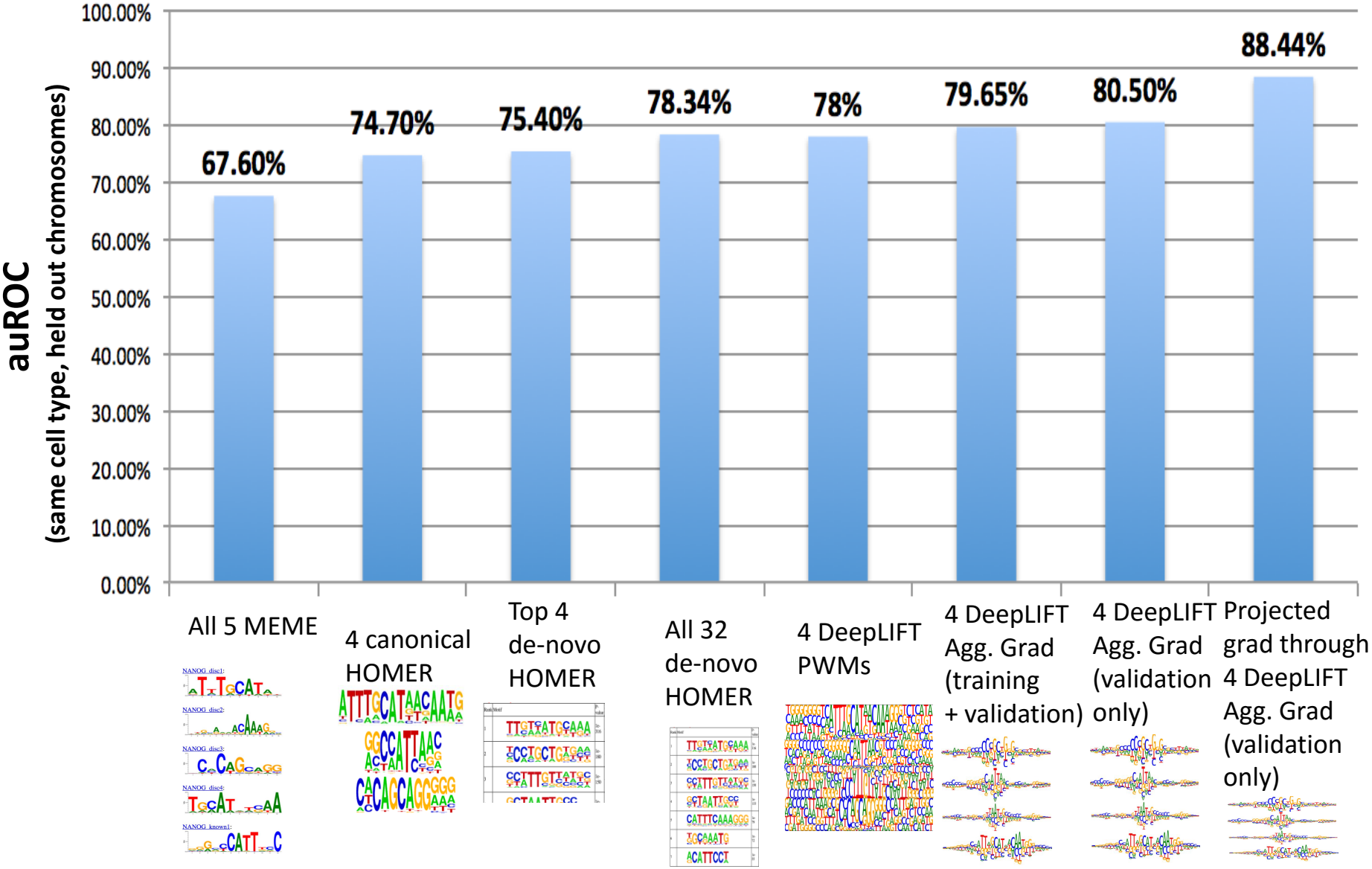
hit (both in fwd orientation)



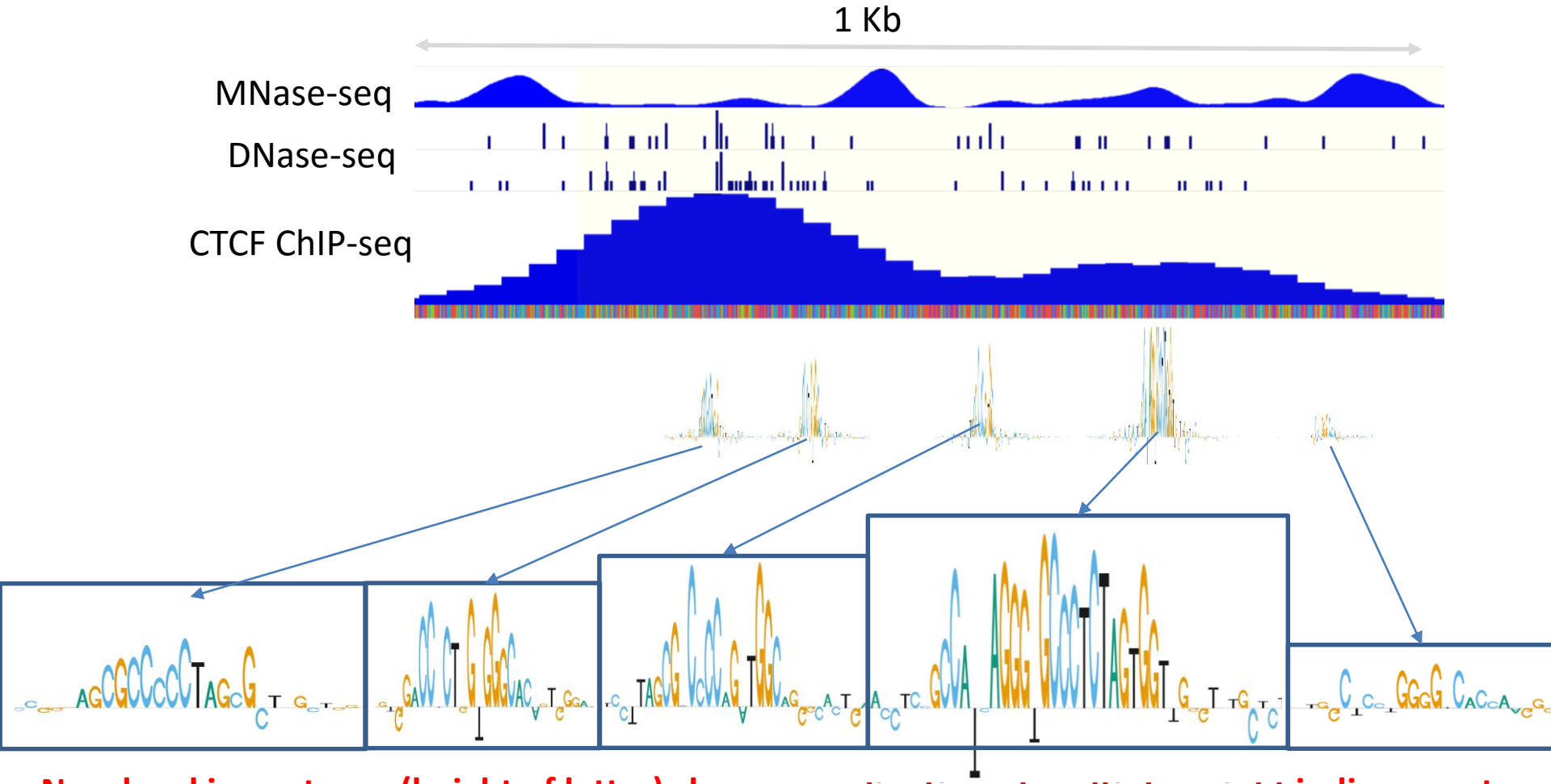
Regions around Zic3 motif sorted by separation

(Colored by base and deepLIFT score)

DeepLIFT motifs are more predictive than classical PWMs even in simple logistic regression models



High resolution point binding events and sequence grammars at a CTCF double peak

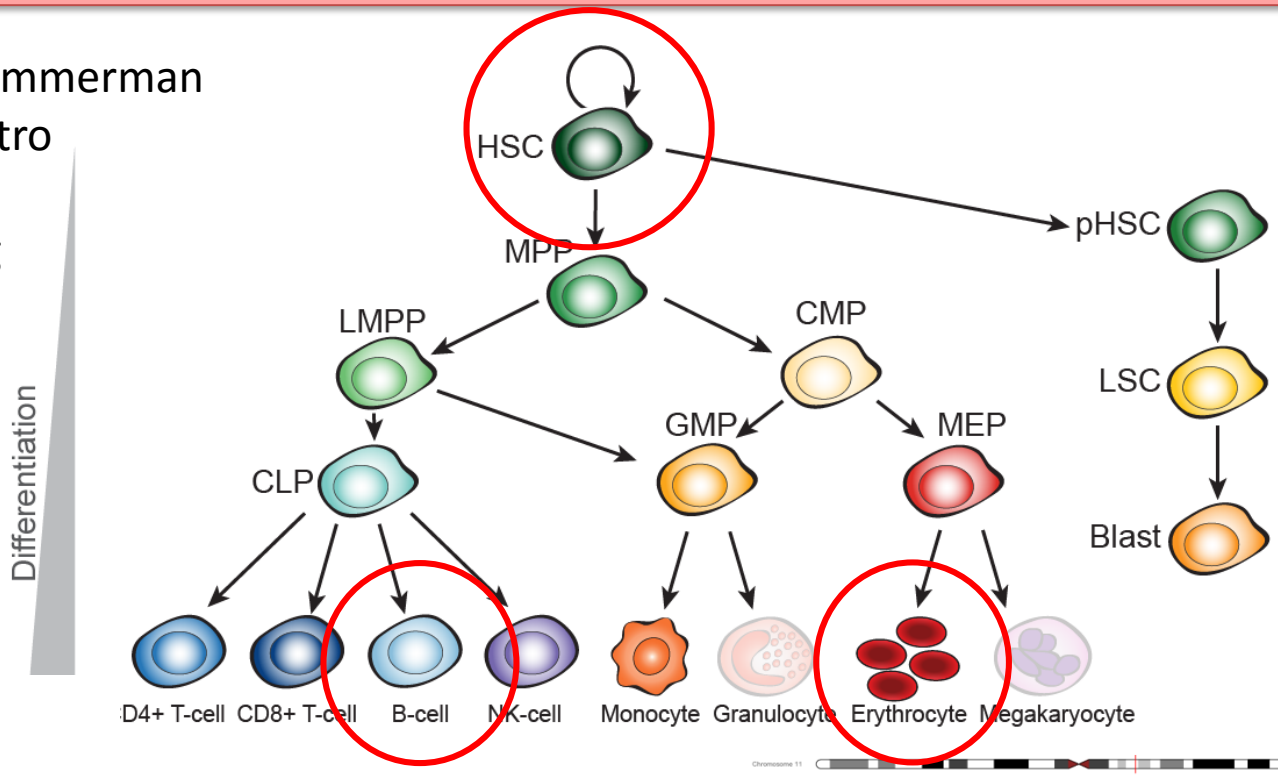


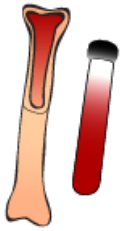
Nuc. level importance (height of letter) shows coordination of multiple point binding events

Deep learning sequence determinants of context-specific chromatin accessibility across hematopoietic cell types

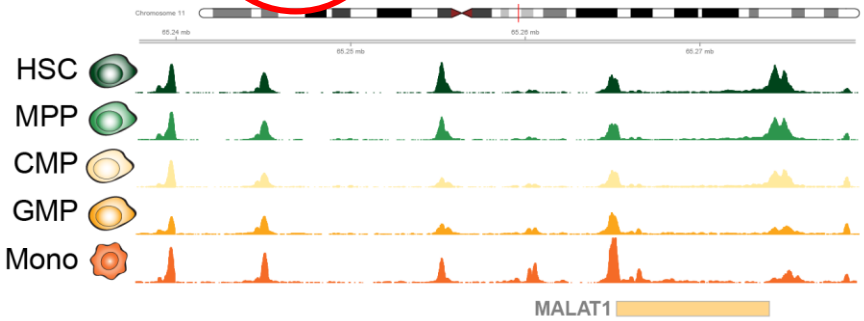
Ryan Corces-Zimmerman
Jason Buenrostro
Will Greenleaf
Howard Chang
Ravi Majeti

Peyton Greenside





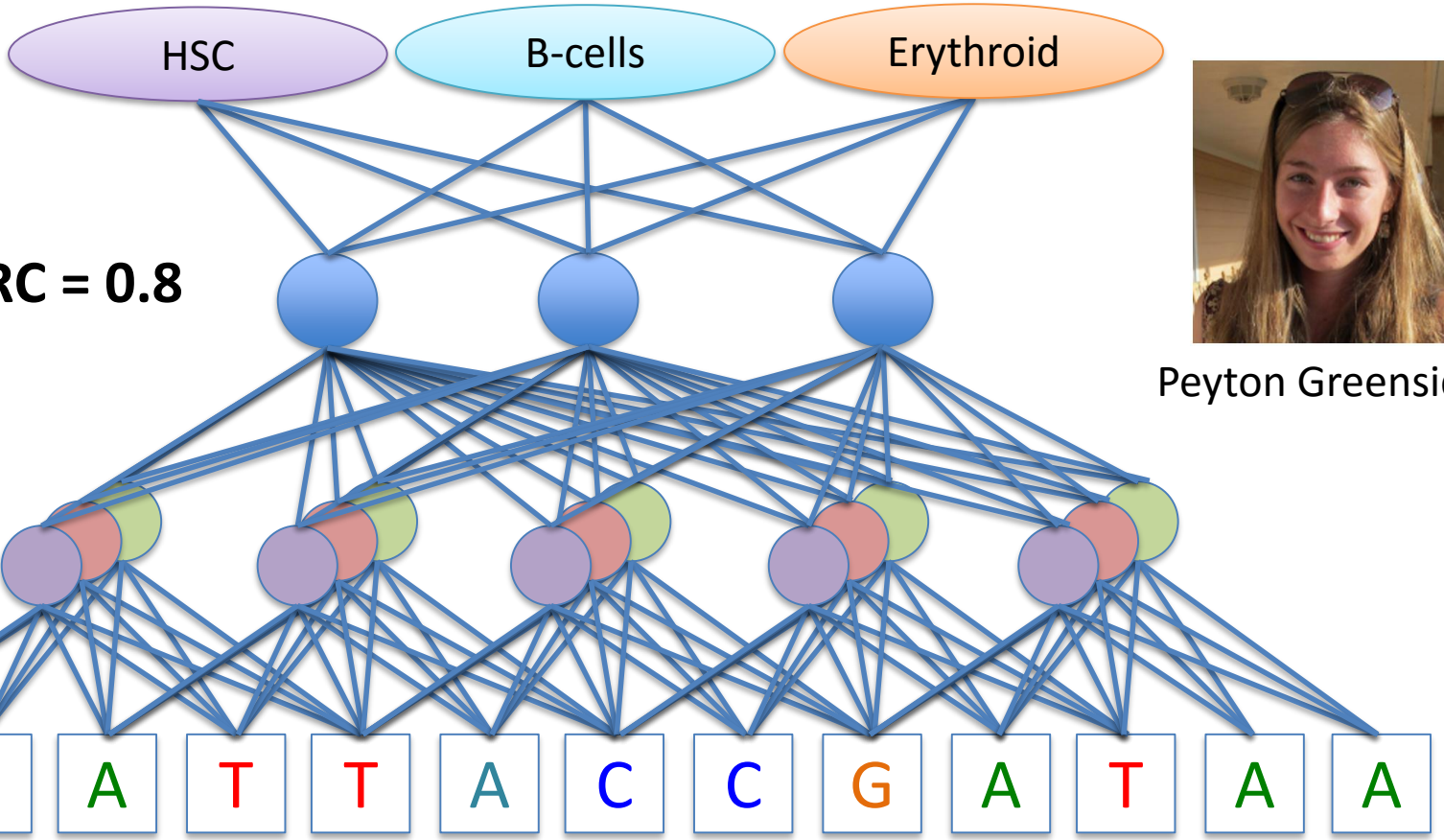
FACS → ATAC-seq



400K ATAC-seq peaks across all cell types

Multi-task deep CNN model of context-specific chromatin accessibility

Output: Accessible (+1) vs. not accessible (0)

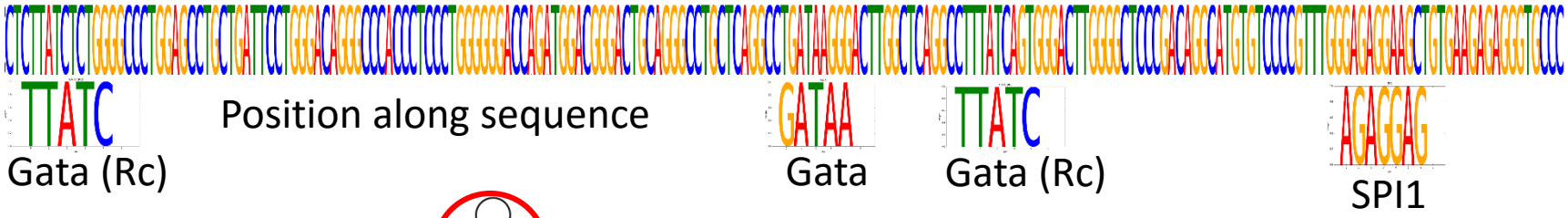


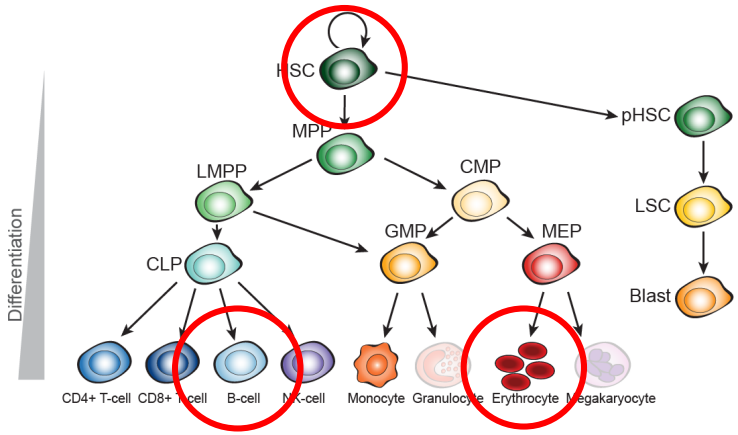
Mean auPRC = 0.8

Peyton Greenside

Input: Raw DNA sequence

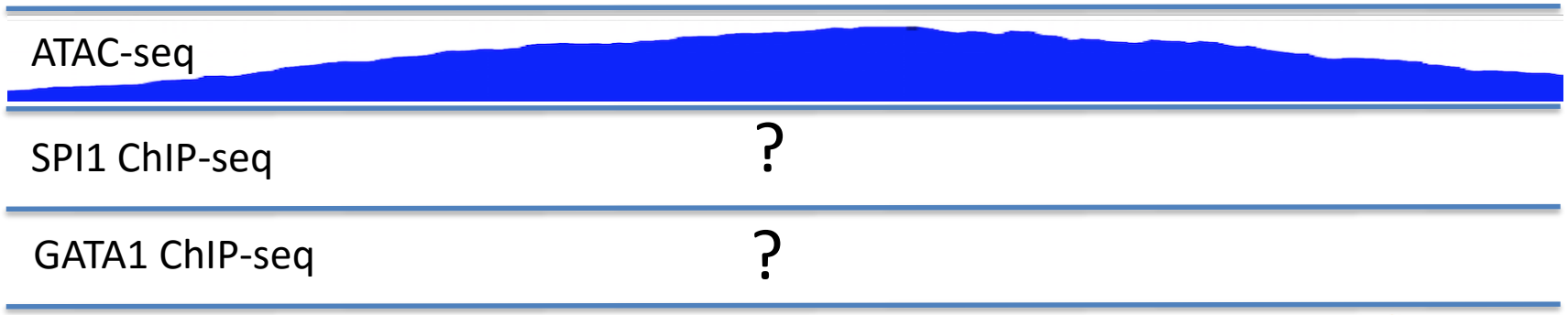
Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts



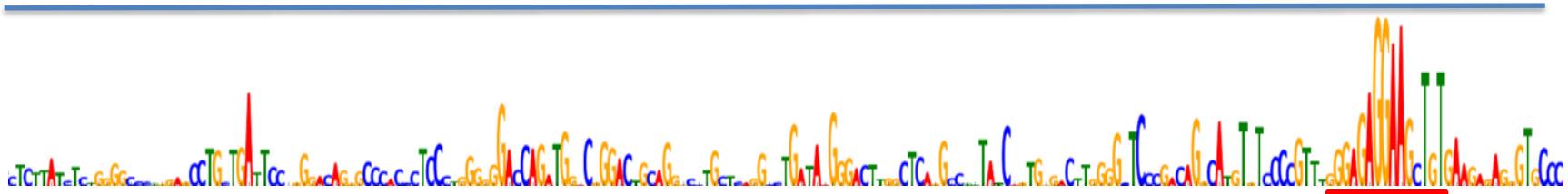


Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

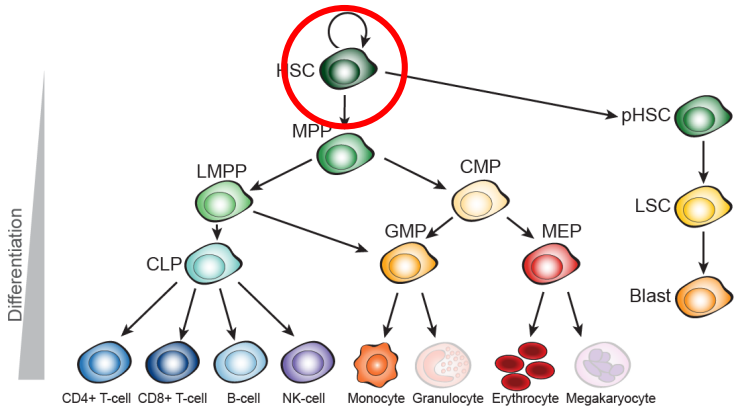


Importance in **HSC's**



Position along sequence

SPI1



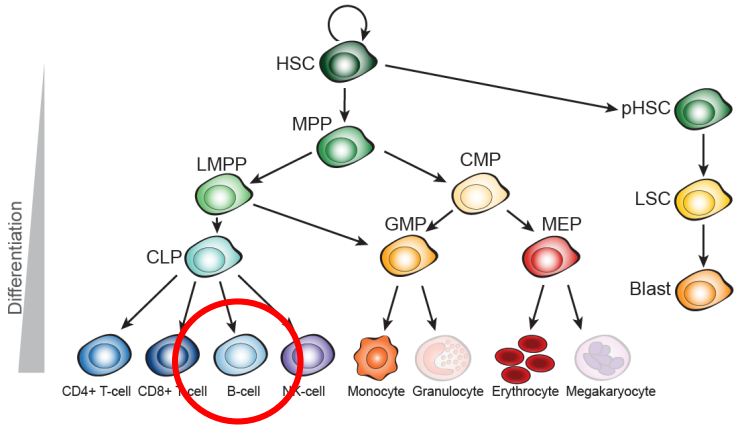
Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

Importance in **B-cells**

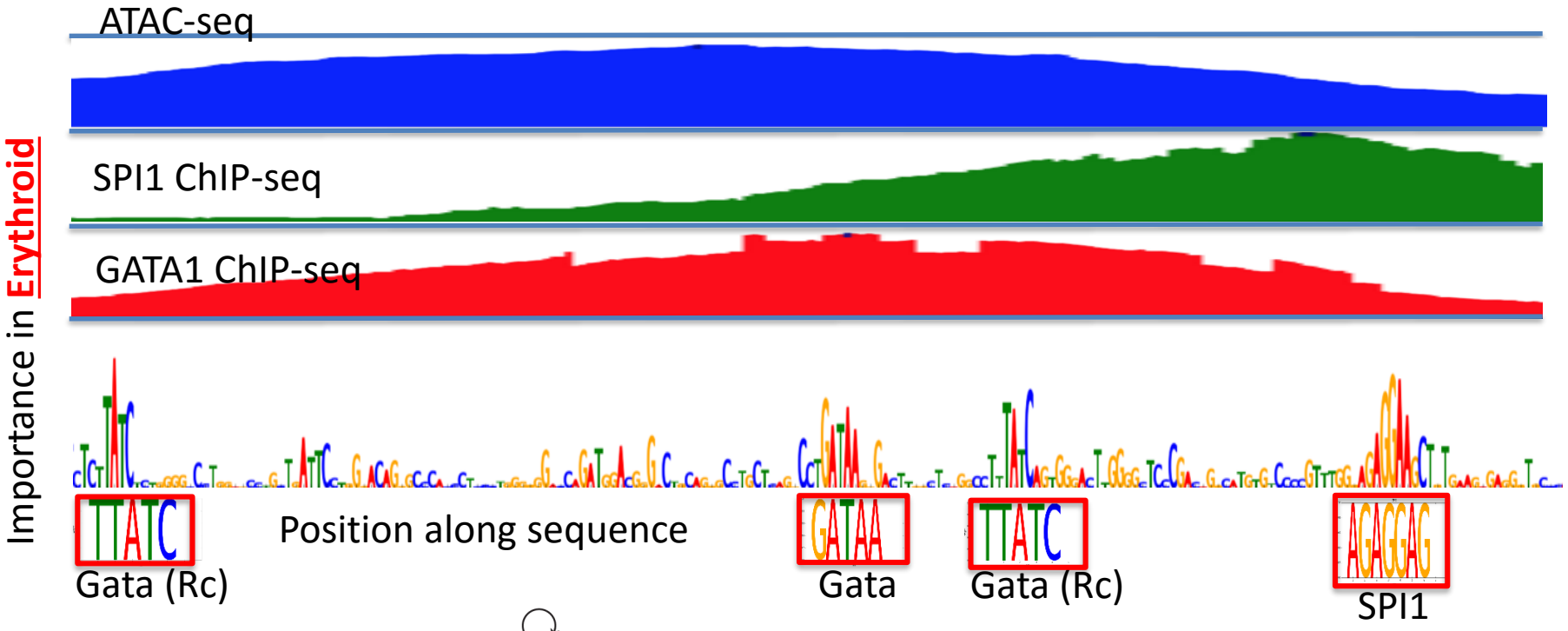
ATAC-seq	No peak
SPI1 ChIP-seq	No peak
GATA1 ChIP-seq	Not expressed

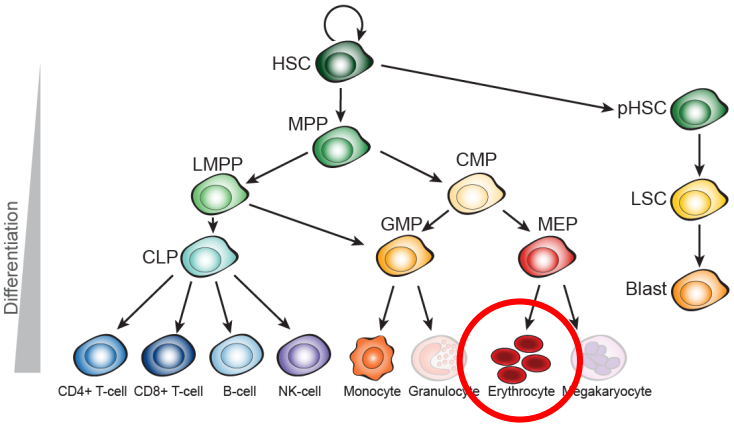
Position along sequence



Peyton Greenside

Context-specific re-use of regulatory sequence in HSC, B-cells and Erythroblasts

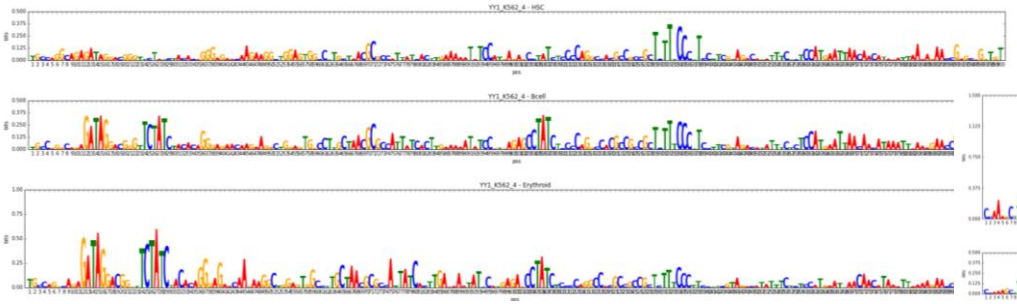




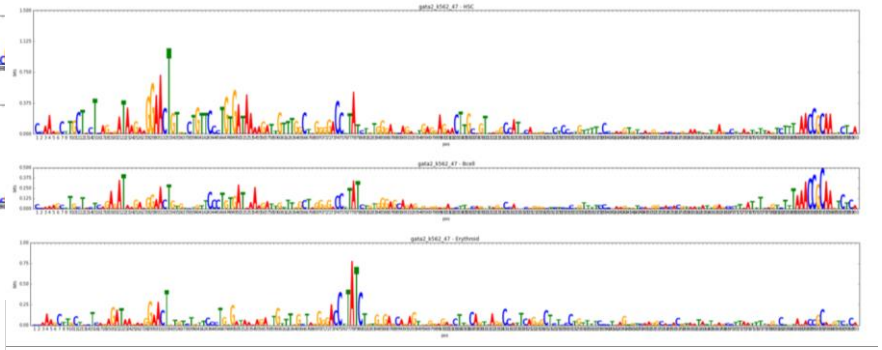
Peyton Greenside

…and much, much more

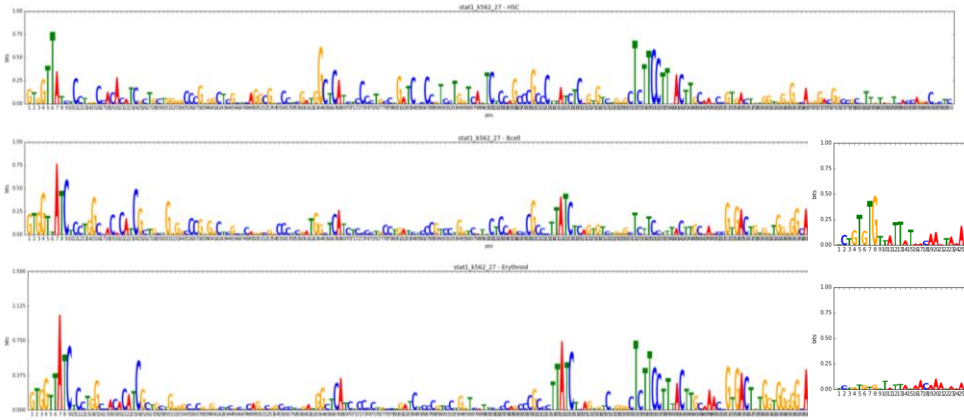
YY1 & GATA



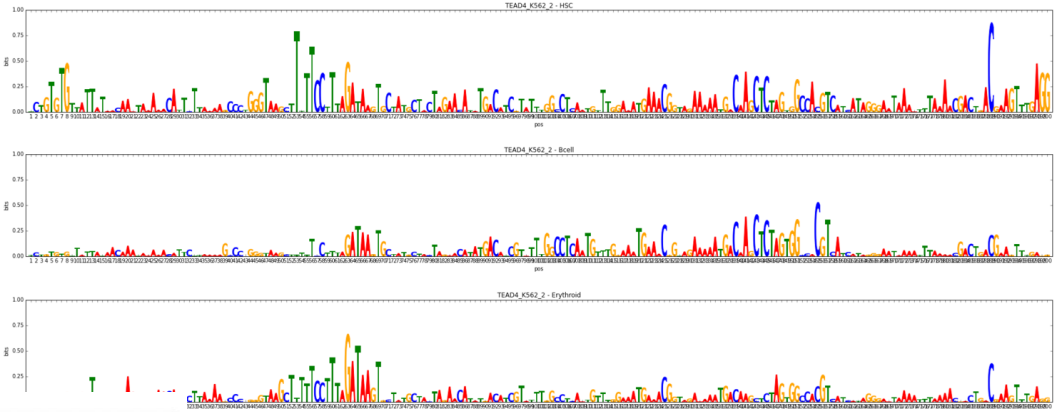
GATA, SPI1, RUNX2



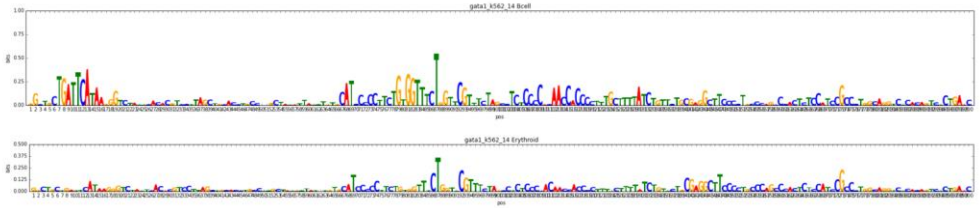
STAT1 & GATA



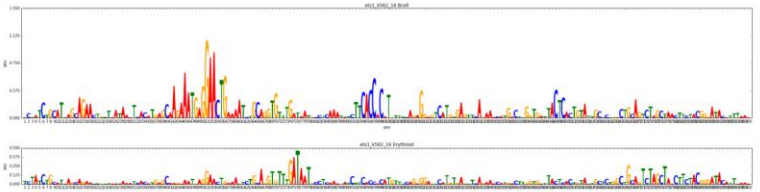
TEAD4 & GATA



AP1 in B-cells only

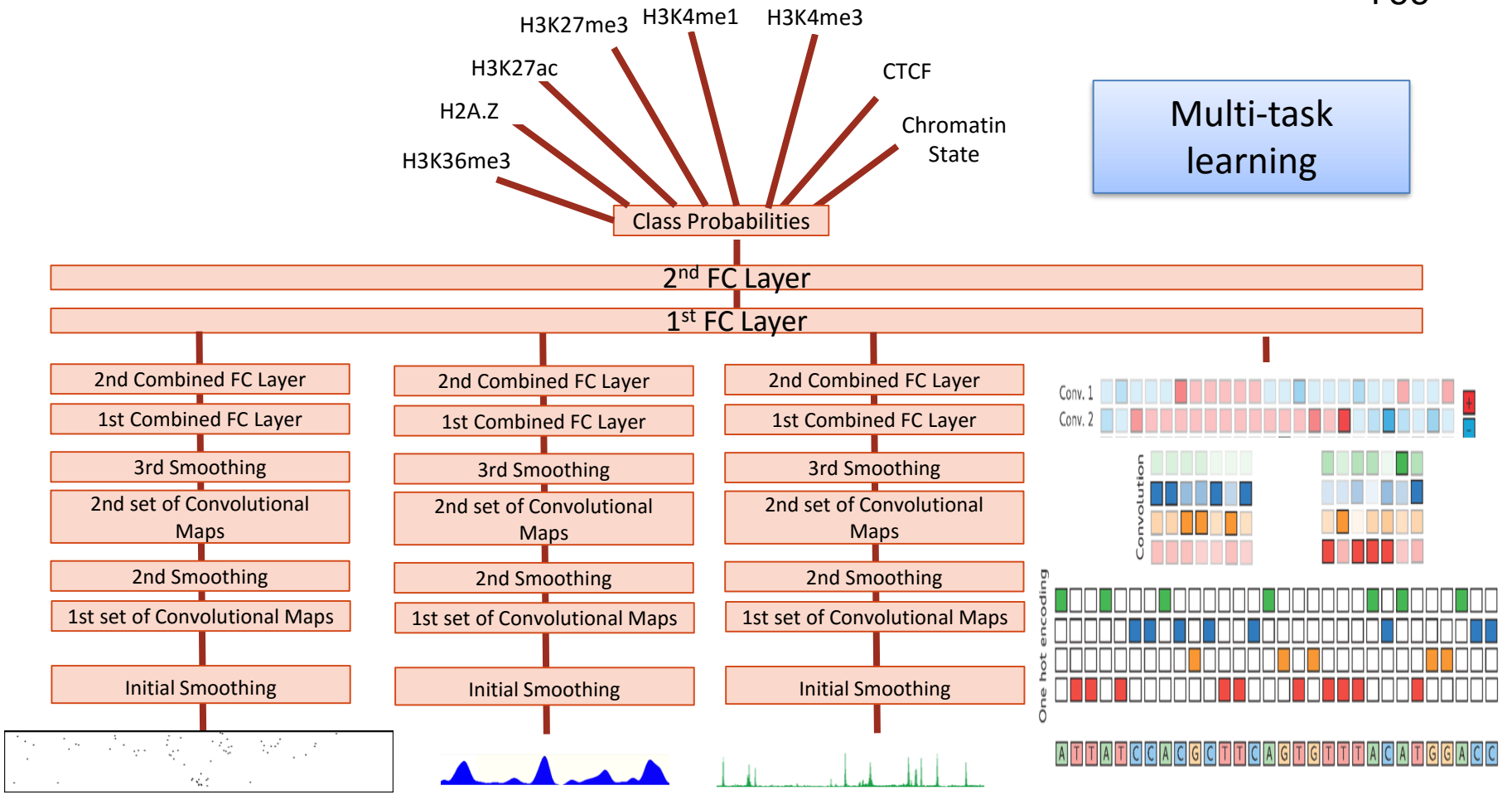


ETS & GATA

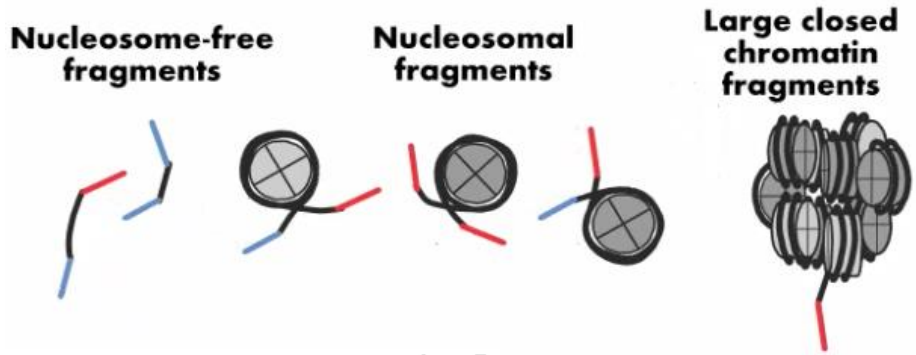


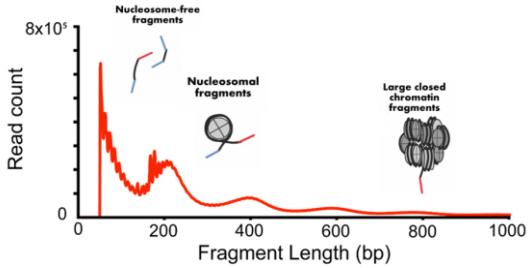
Chromputer predicts multiple histone marks from ATAC-seq or DNase-seq/MNase-seq + sequence with high accuracy

Chuan Sheng Foo



ATAC-seq generates variable length fragments reflecting different aspects of chromatin architecture



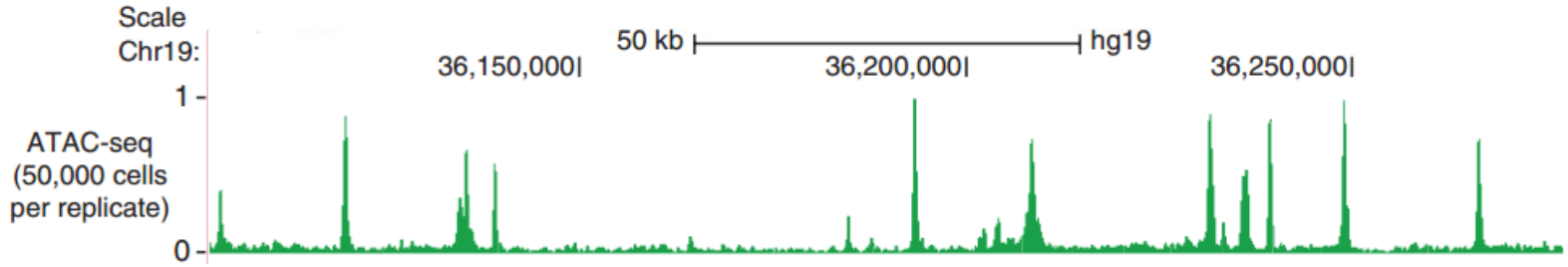


NATURE METHODS | ARTICLE

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

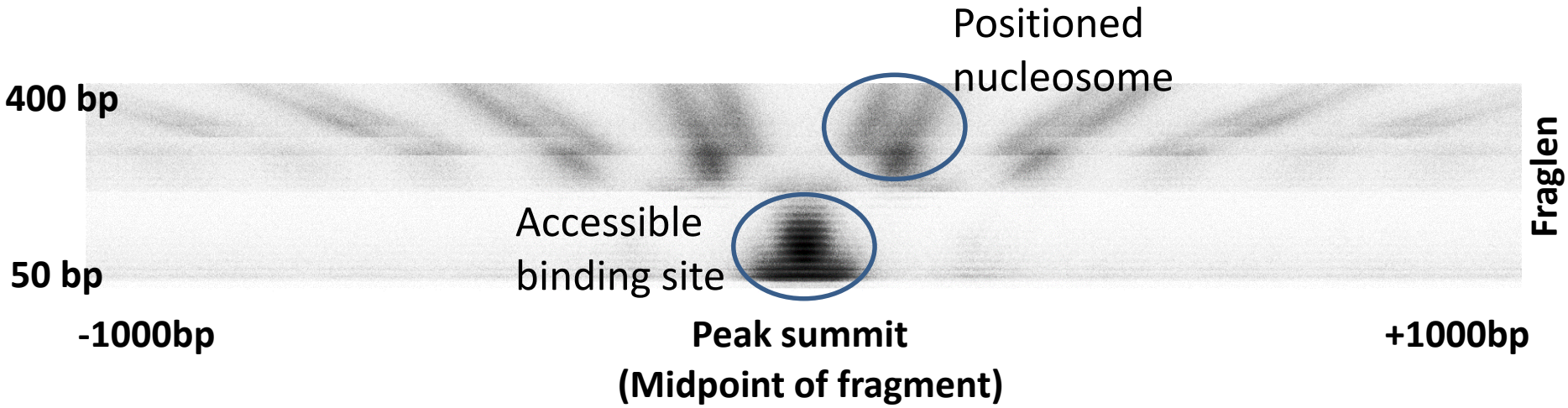
Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang & William J Greenleaf

Paired-end sequencing



ATAC-seq peaks identify chromatin accessible regulatory elements

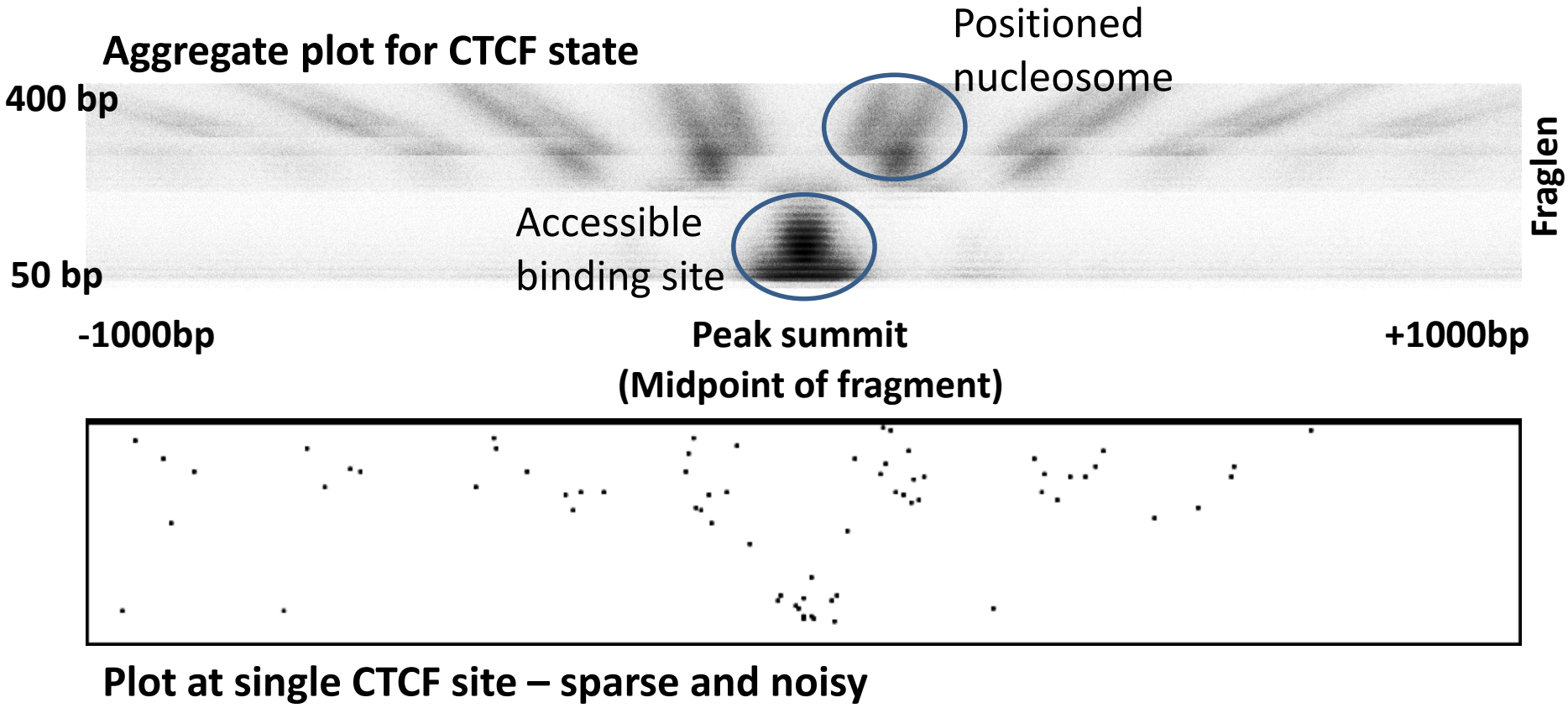
Position-aware 2D fragment length distributions (V-plots)



Aggregate plot for all ATAC-seq peaks in CTCF state

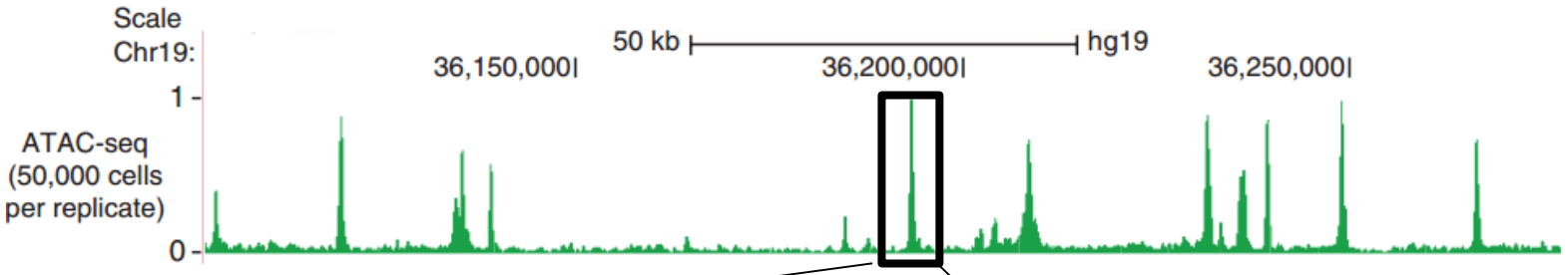
V-plots were first introduced by Henikoff et al. 2011, PNAS

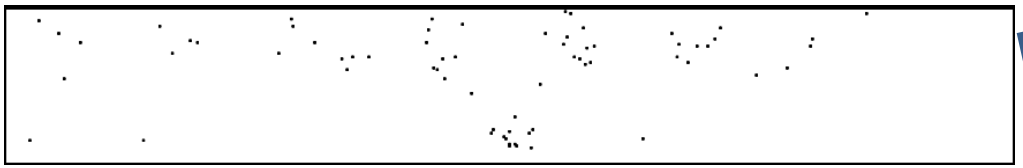
Position-aware 2D fragment length distributions (V-plots)



V-plots were first introduced by Henikoff et al. 2011, PNAS

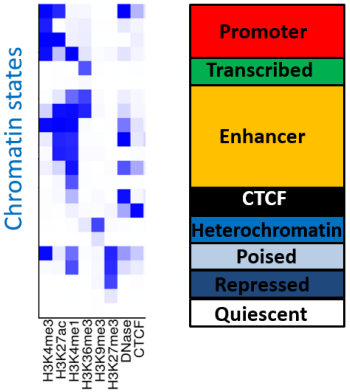
Can we predict chromatin states/histone marks at ATAC-peaks?





2kb around ATAC-peak

Which of 8 chromatin states?



Which histone mark is present?

Image classification task!

Chromatin architecture can predict **chromatin state** in held out chromosome (same cell type GM12878)

Model + Input data types	8-class chromatin state accuracy (%)
Majority class (baseline)	**42%**
Gene proximity	59%
Random Forest: **ATAC-seq** (150M reads)	**61%**
Chromputer: **DNase** (60M reads)	68.1%
Chromputer: **Mnase** (1.5B reads)	69.3%
Chromputer: **ATAC-seq** (150M reads)	**75.9%**
Chromputer: **DNase** + **MNase**	**81.6%**
Chromputer: **ATAC-seq** + **sequence**	83.5%
Chromputer: **DNase** + **MNase** + **sequence**	**86.2%**
Label accuracy across replicates (upper bound)	**88%**

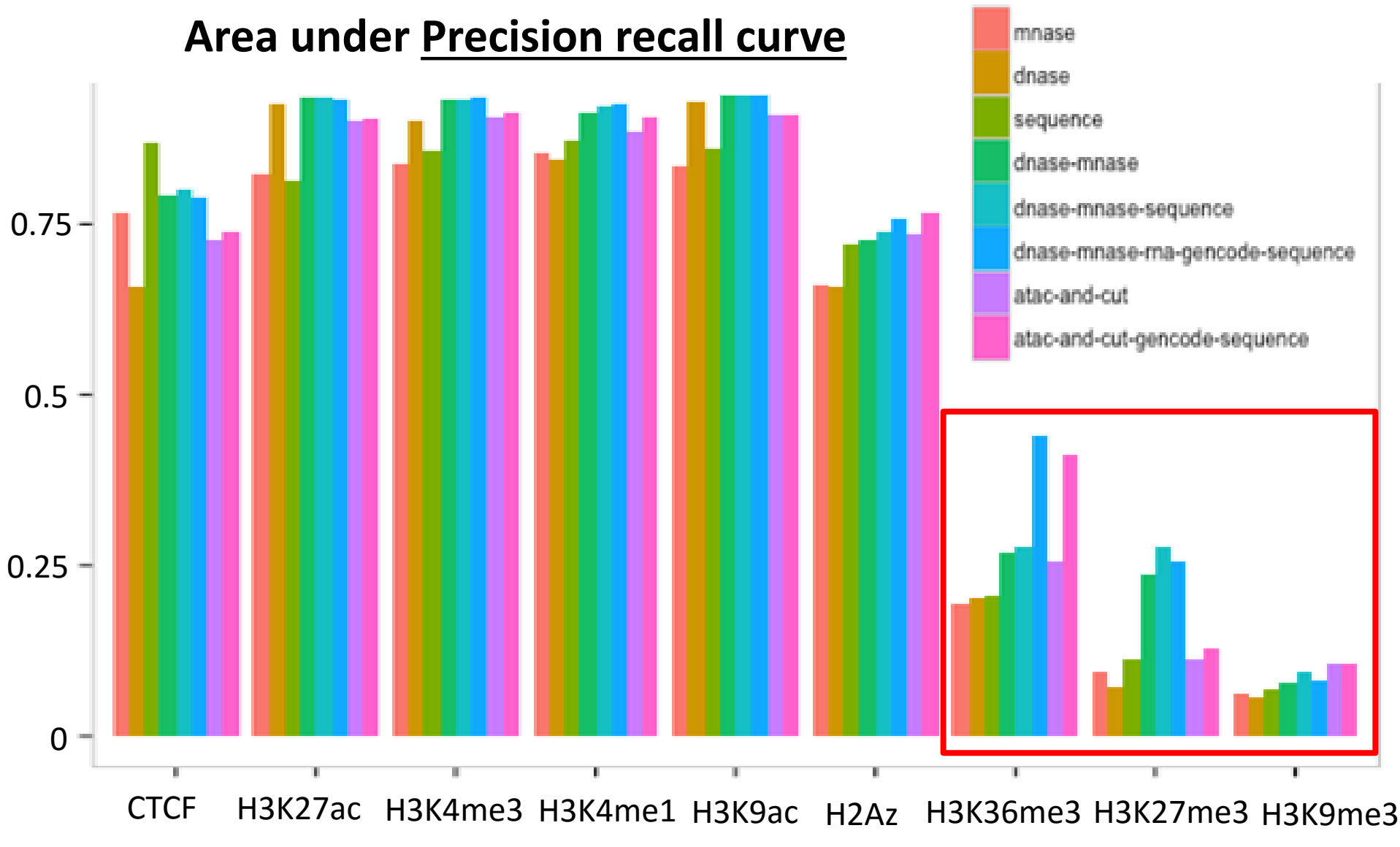
High cross cell-type chromatin state prediction

- Learn model on **DNase and MNase only**
- **Learn on GM12878, predict on K562 (and vice versa)**
- **Requires local normalization** to make signal comparable

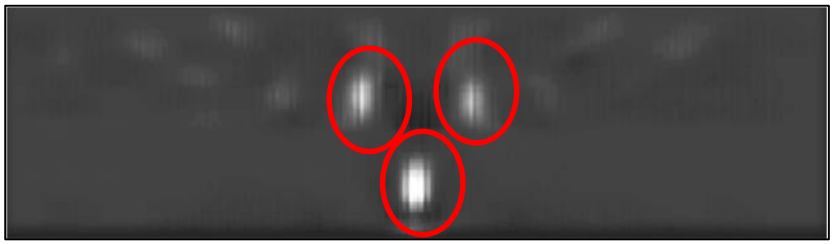
8 class chromatin state accuracy		
Train ↓ / Test →	GM12878	K562
GM12878	0.816	**0.818**
K562	**0.769**	0.844

Predicting individual histone marks

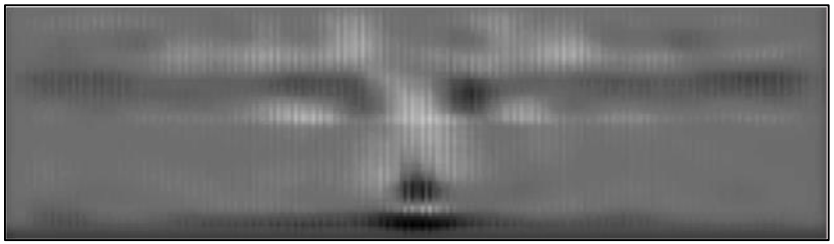
from ATAC/DNase/MNase/Sequence



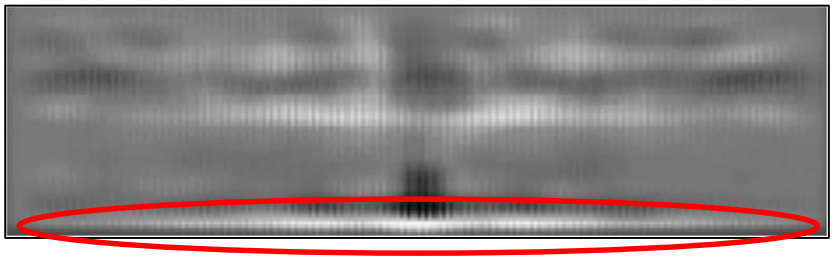
What architecture properties of the ATAC-seq V-plots predict different chromatin states?



CTCF state: centered binding, symmetric phased nucleosomes



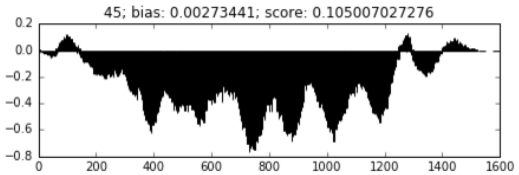
Enhancer state: localized signal, heterogeneity

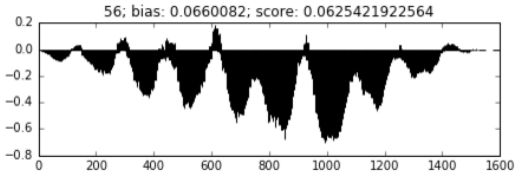


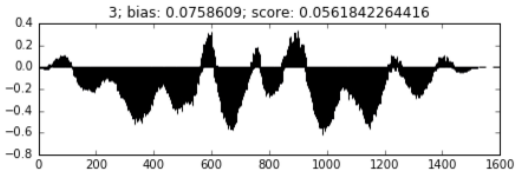
Promoter state: broad regions of accessible chromatin

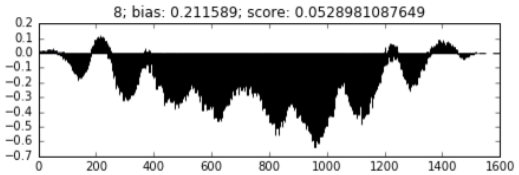
what is the change in classification probability relative to an unbiased classifier if we ***only*** consider the contributions from each pixel

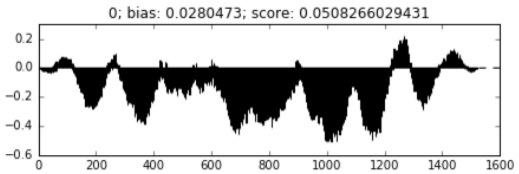
Top scoring MNase filters and activating input patterns for CTCF state

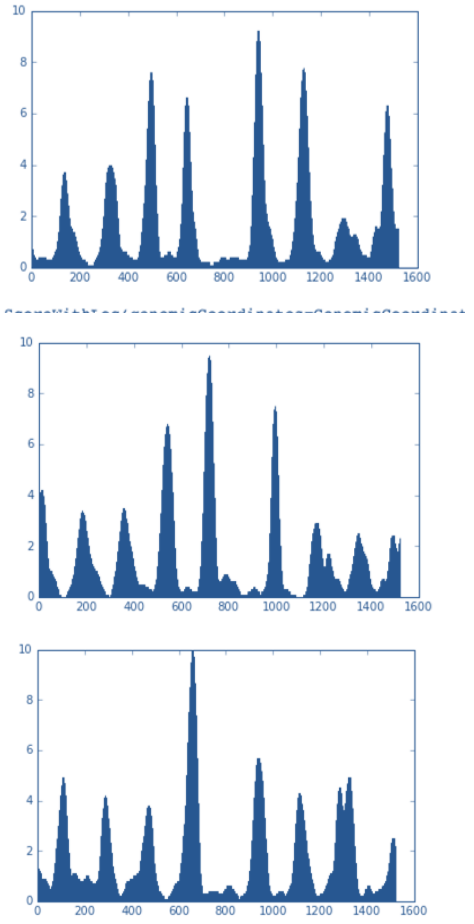








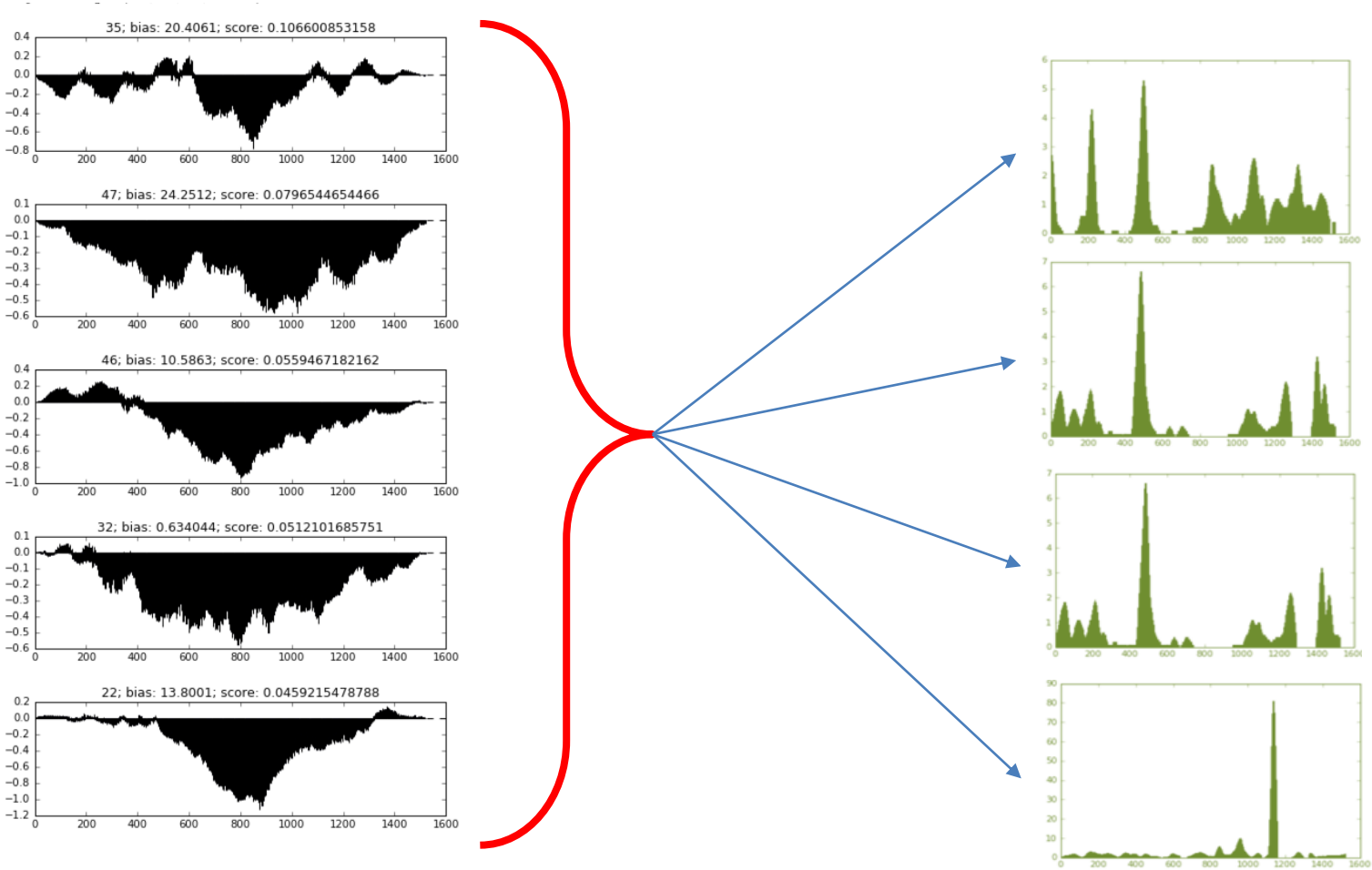




Top scoring MNase filters

Maximally activating input MNase profiles

Top scoring MNase filters and activating input patterns for promoter state



Top scoring MNase filters

Maximally activating input MNase profiles

Summary

- **Chromputer:** Powerful multi-input, multi-output integrative deep learning framework for regulatory genomics
  - Beware of negative set/background selection
  - Beware of performance measures (most prediction tasks are highly unbalanced)
- **DeepLIFT:** efficient method for scoring importance of raw input and intermediate induced features in deep neural networks
  - DNNs learn distributed representations. Caution in over-interpreting individual filters
  - Propagate and integrate multiple filter effects on raw input of individual examples.
  - Cluster 'important' local patterns across examples to learn non-redundant global patterns
- Extensive evidence of **differential usage of sequence grammars** at regulatory elements in different contexts (To be validated with experiments!)

Acknowledgements

Kundaje Lab members

Avanti Shrikumar

Peyton Greenside

Nicholas Sinnott-Armstrong

Johnny Israeli

Rahul Mohan

Irene Kaplow

Chuan Sheng Foo

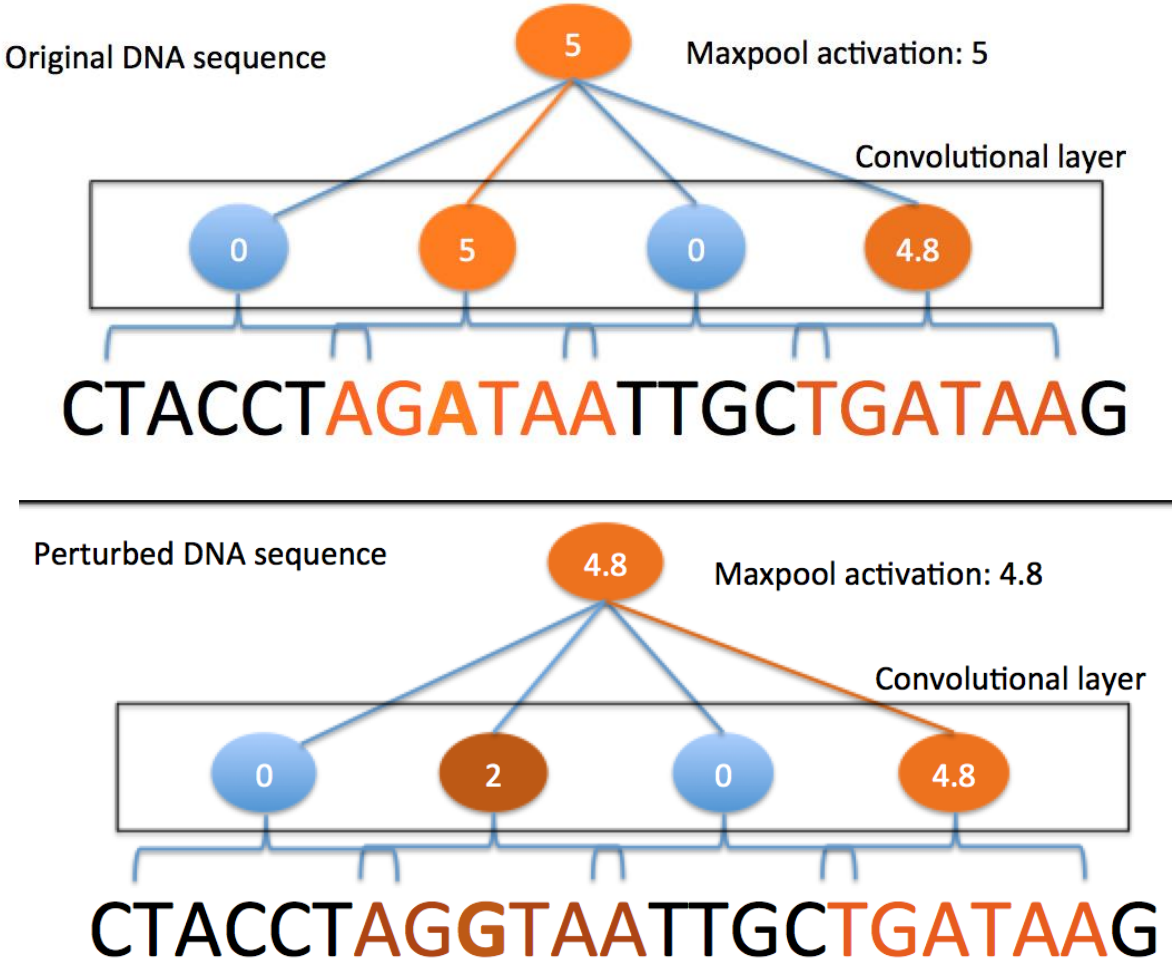
Nathan Boley

Funding

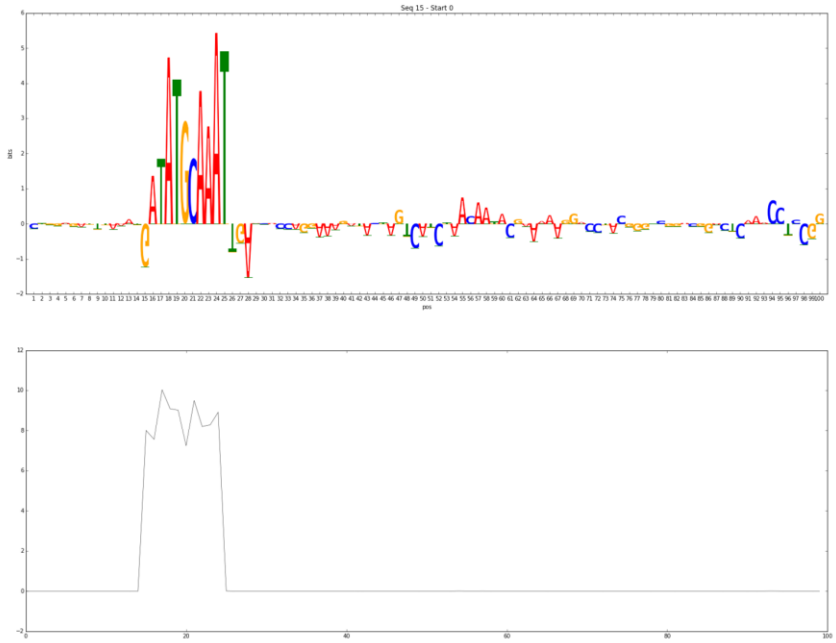
U01HG007919-02 (GGR)
U41-HG007000-04S1
R01ES02500902

Conflict of Interest: Deep Genomics (SAB), Epinomics (SAB)

Buffering and in-silico mutagenesis



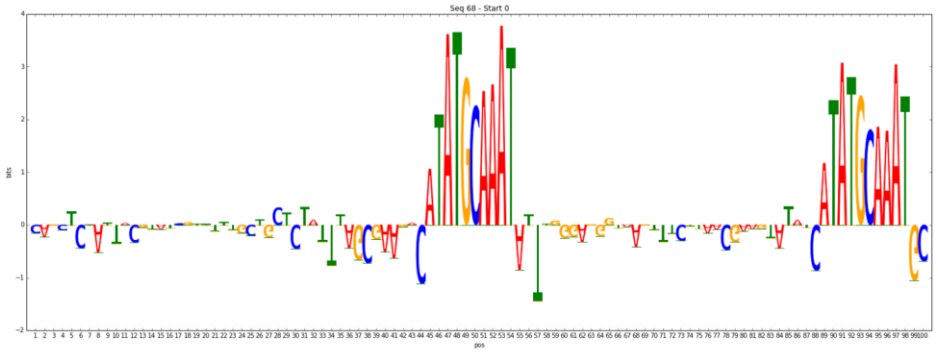
Buffering and in-silico mutagenesis



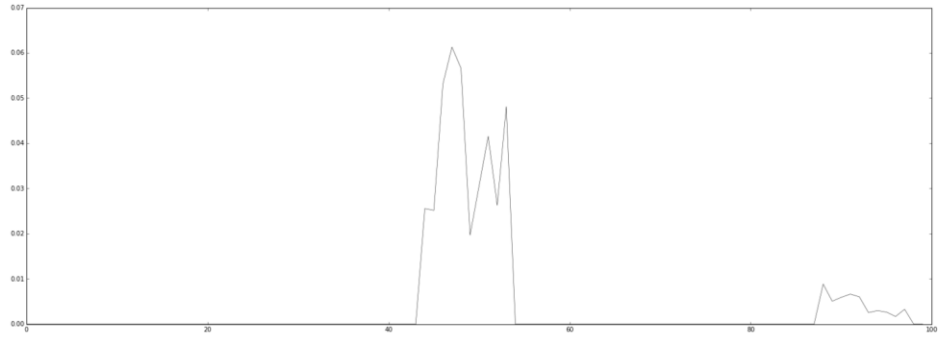
DeepLIFT scores

In-silico mutagenesis scores

Buffering and in-silico mutagenesis

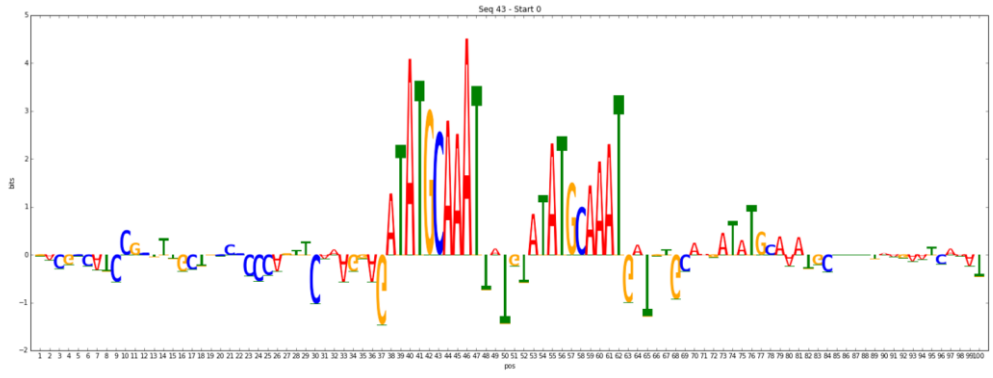


DeepLIFT scores

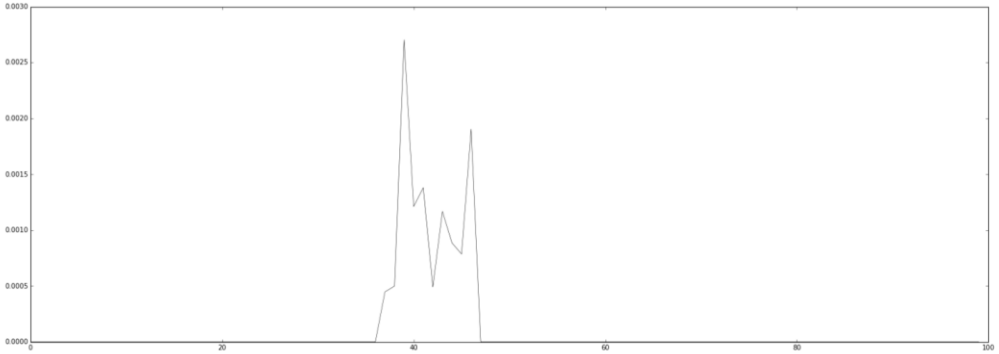


In-silico mutagenesis scores arbitrarily scores left motif stronger than right motif

Buffering and in-silico mutagenesis



DeepLIFT scores



In-silico mutagenesis is unable to detect the second motif to the right