6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 6: Regulatory genomics

Gene regulation, chromatin accessibility, DNA regulatory code

Prof. Manolis Kellis

Massachusetts Institute of Technology

http://mit6874.github.io

Slides credit: 6.047, Anshul Kundaje, David Gifford

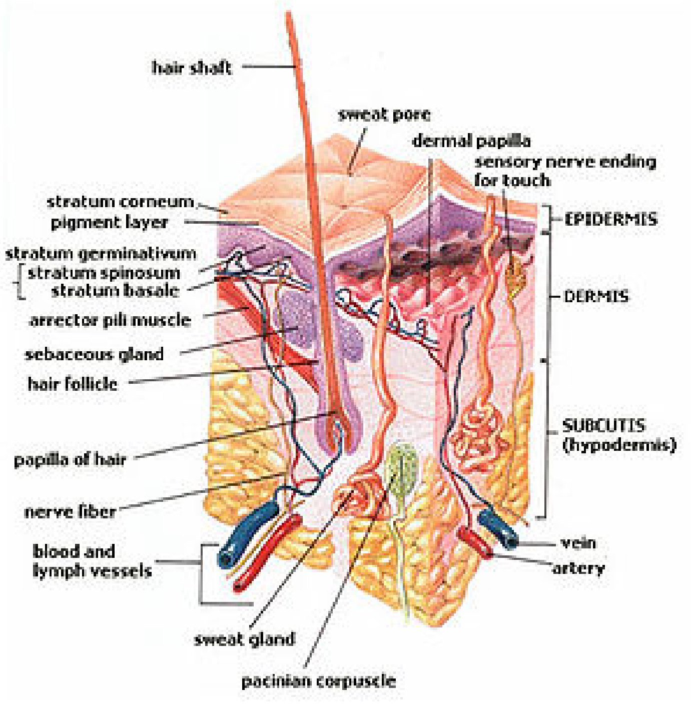
Deep Learning for Regulatory Genomics

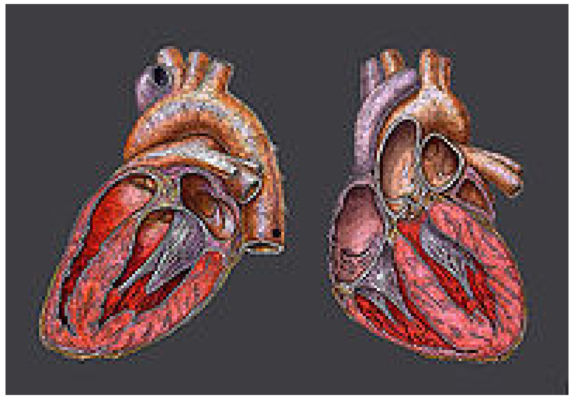
1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⬄ DNA letters. Patches/filters ⬄ Motifs. Higher ⬄ combinations
   - Learning convolutional filters ⬄ Motif discovery. Applying them ⬄ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

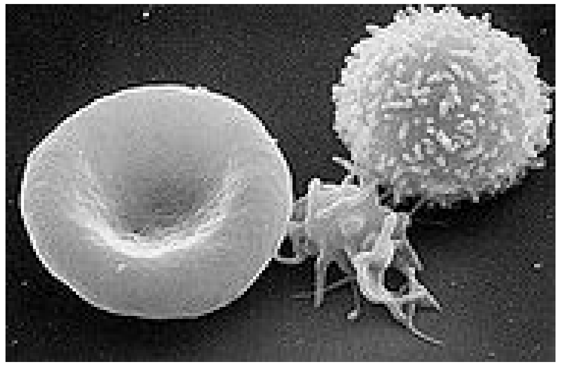
1a. Basics of gene regulation

One Genome – Many Cell Types

```
ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
      TACCAGTA
```







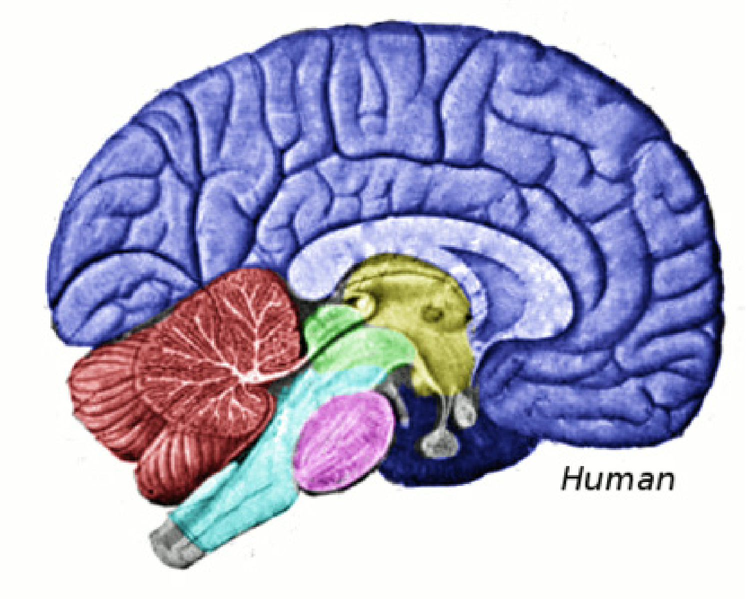
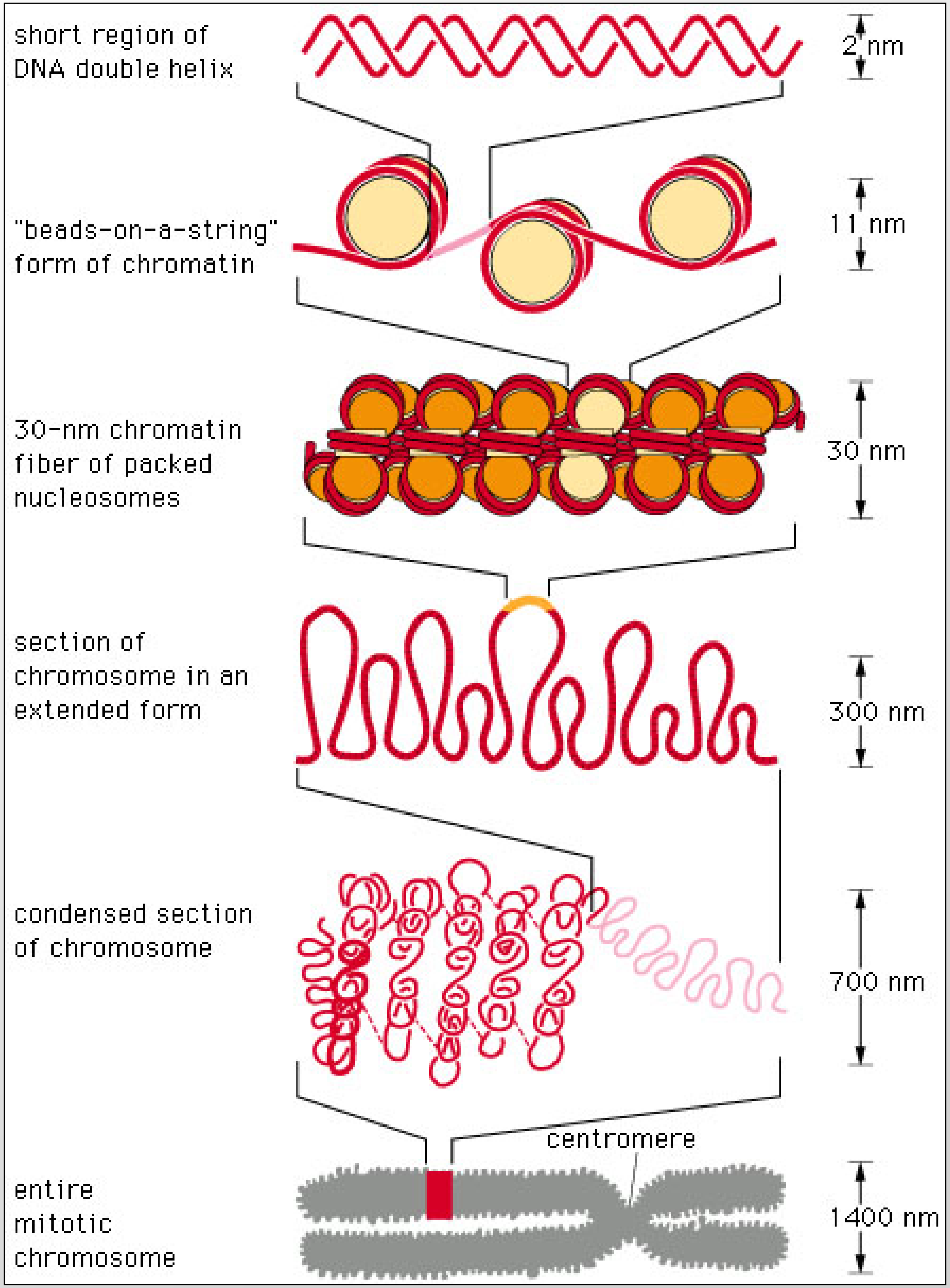


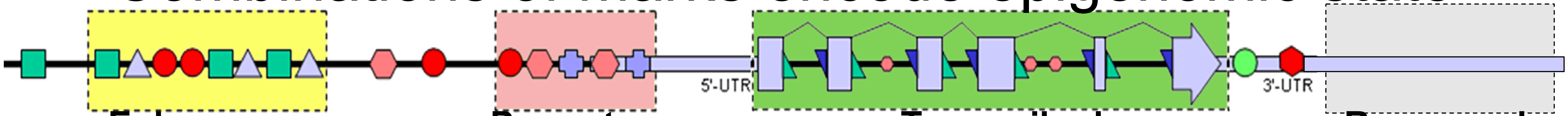
Image Source wikipedia

DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm

Combinations of marks encode epigenomic state

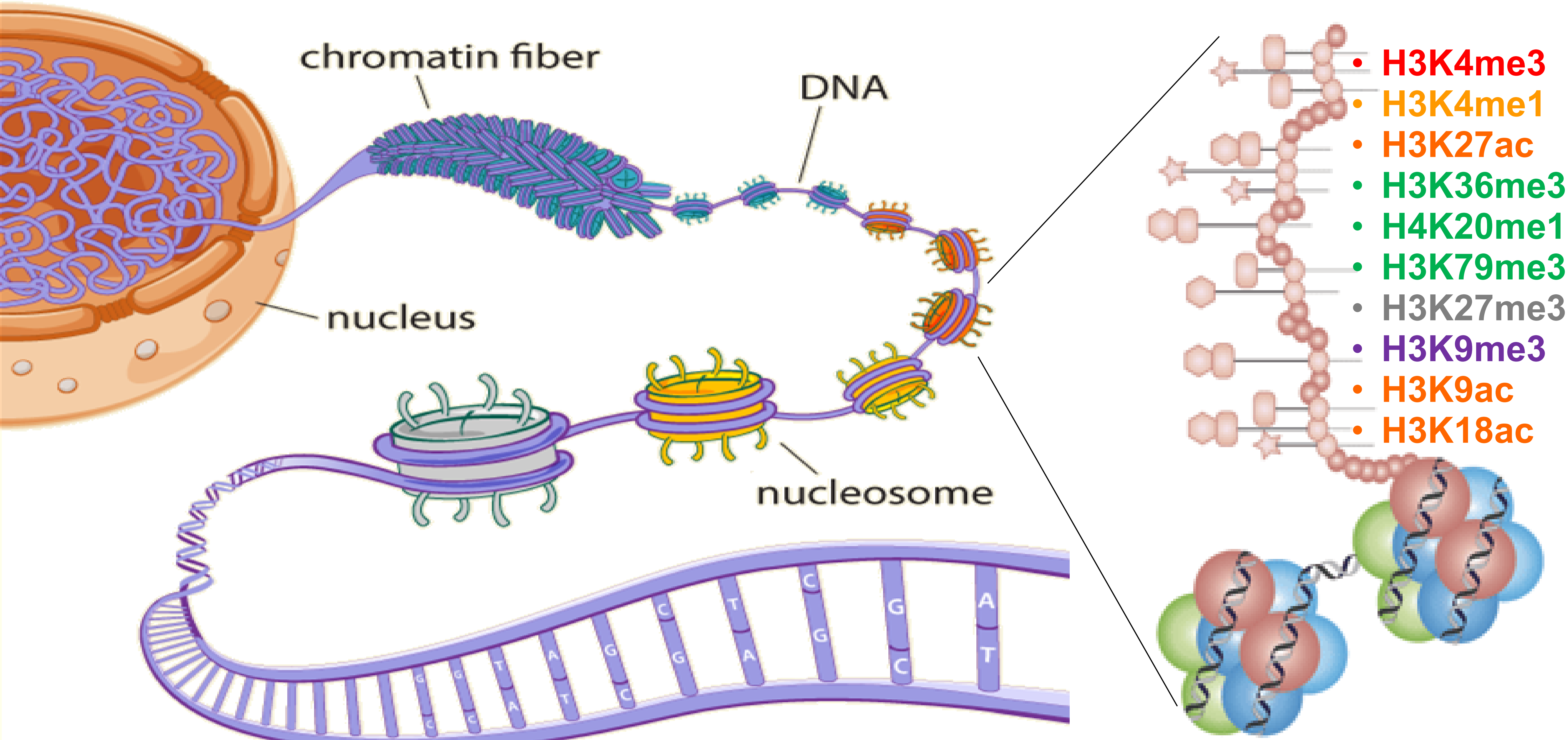


Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

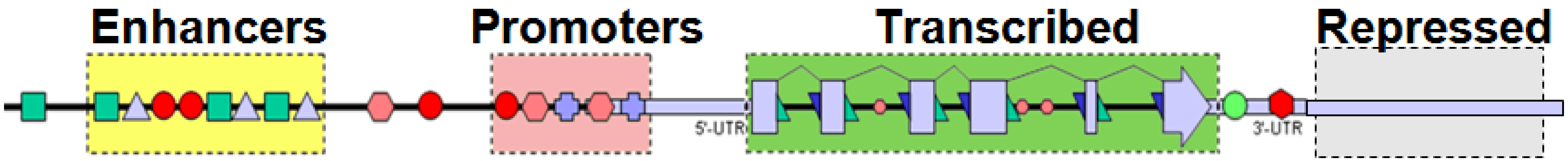
Repressed
- H3K9me3
- H3K27me3
- DNAmethyl

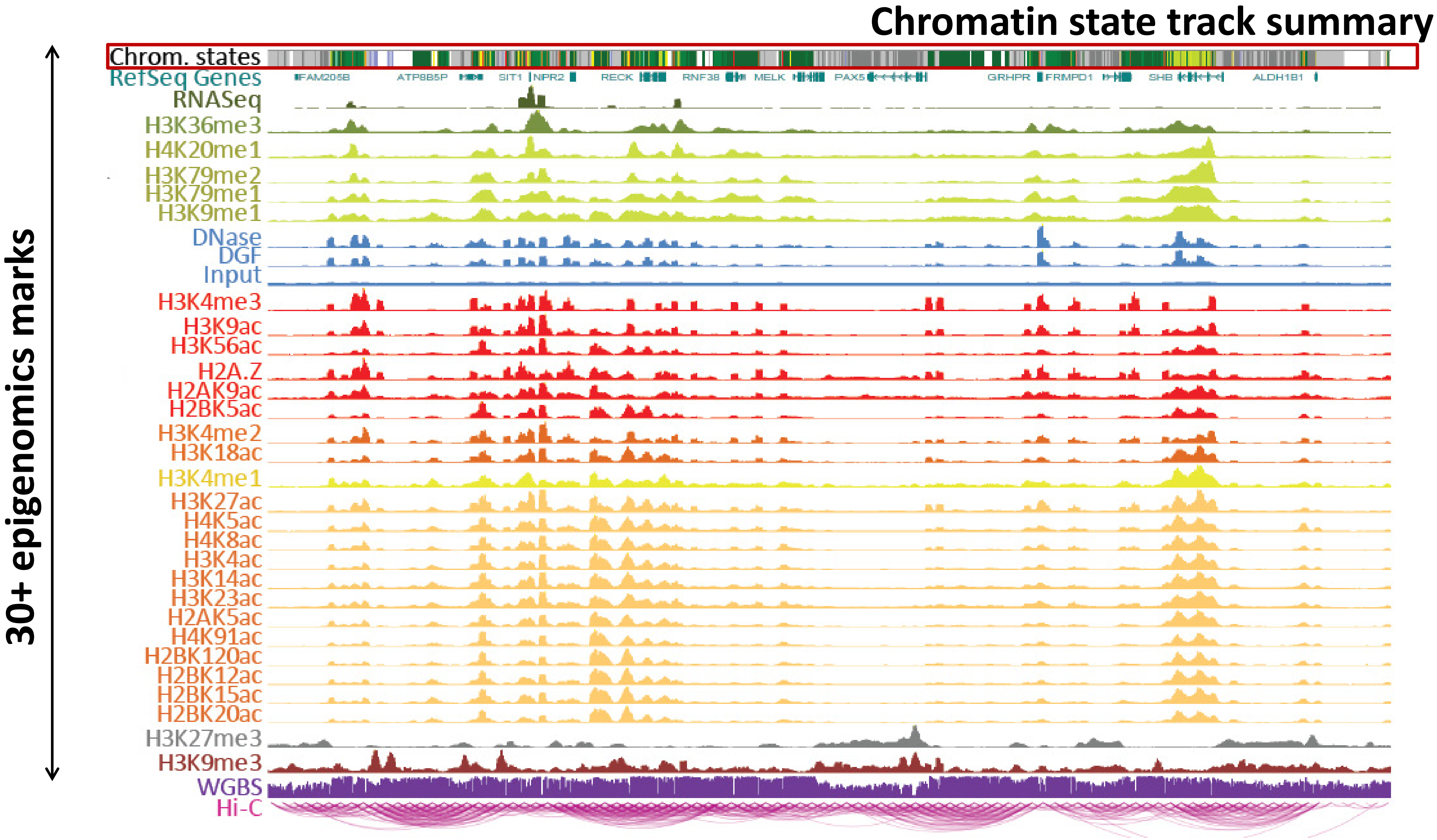


- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me1
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Summarize multiple marks into chromatin states

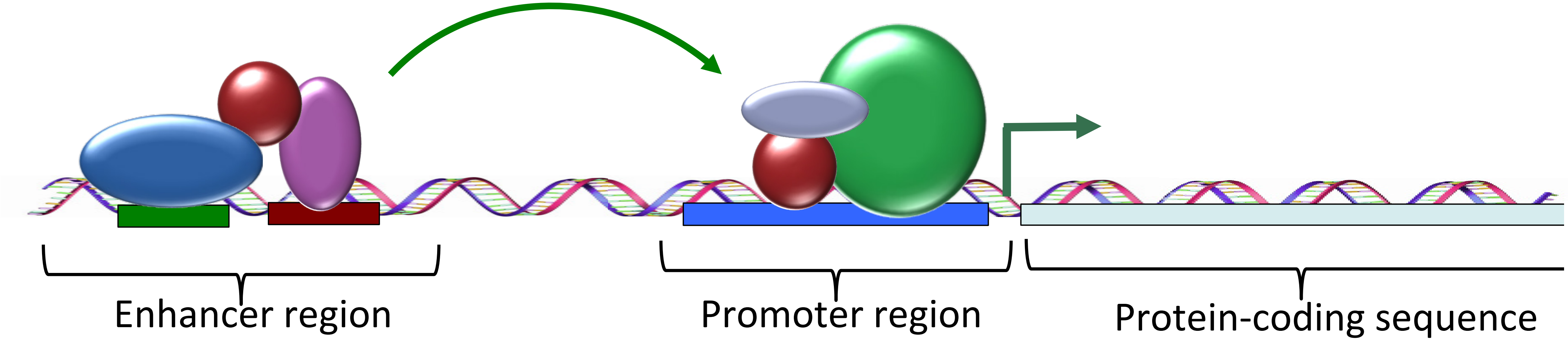




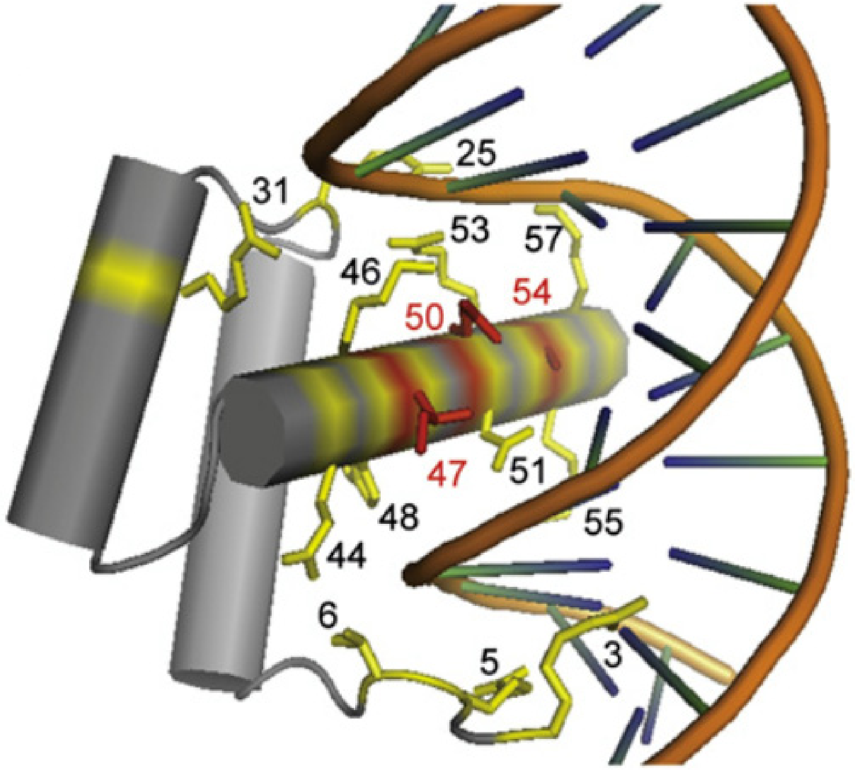
WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

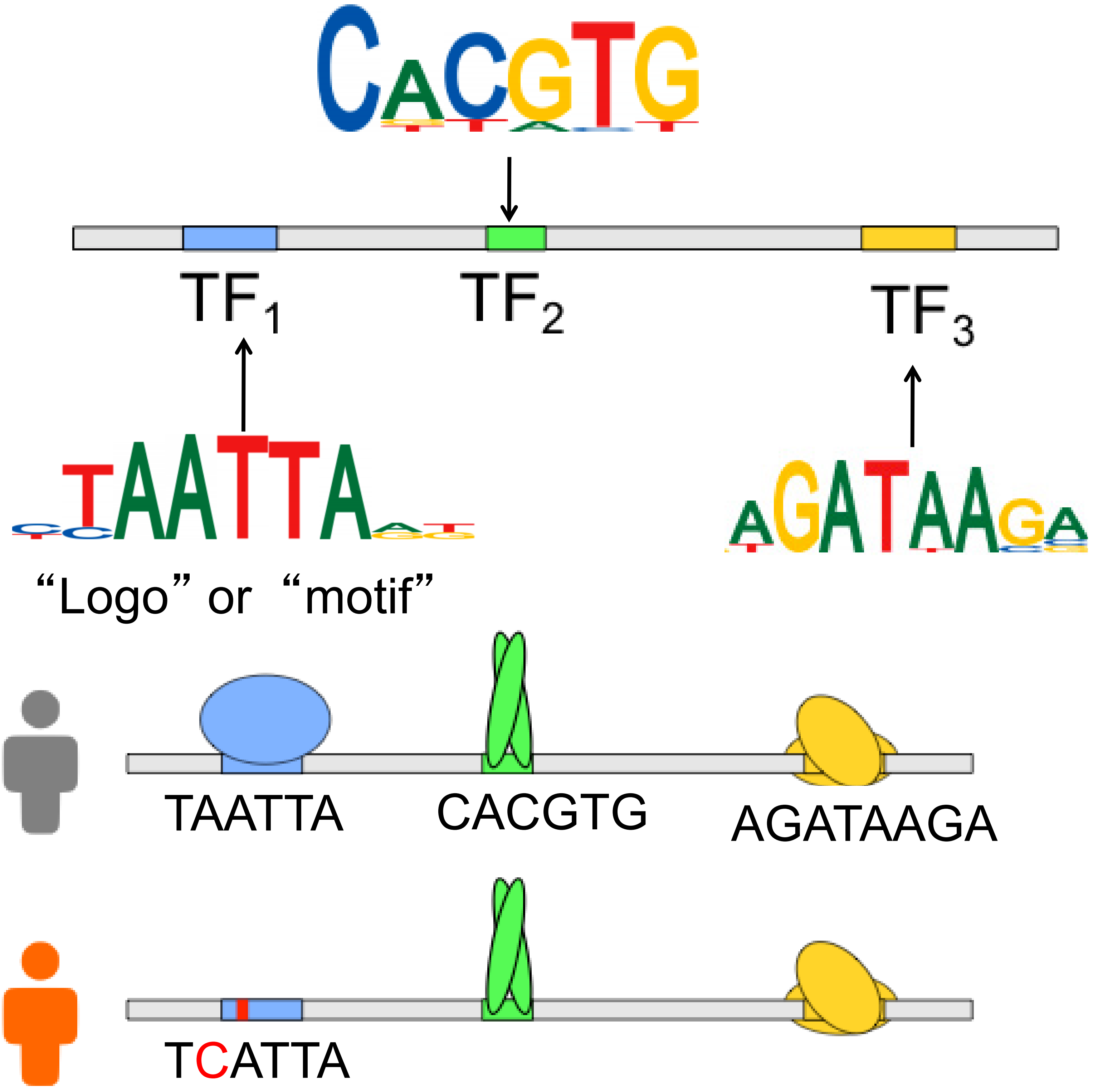
Transcription factors control activation of cell-type-specific promoters and enhancers



TFs use DNA-binding domains to recognize specific DNA sequences in the genome



DNA-binding domain of *Engrailed*



TF_1 TF_2 TF_3

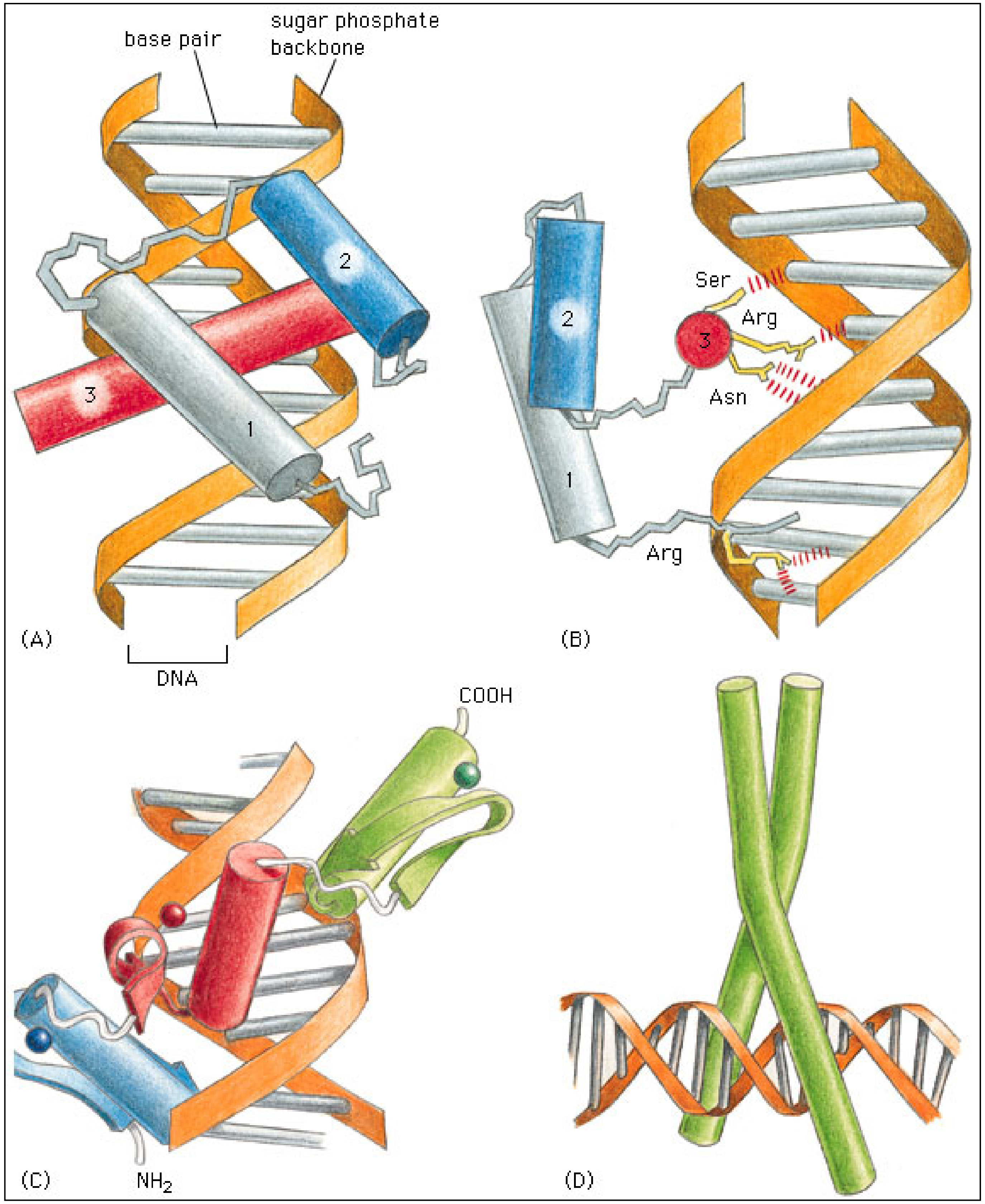
“Logo” or “motif”

TAATTA CACGTG AGATAAGA

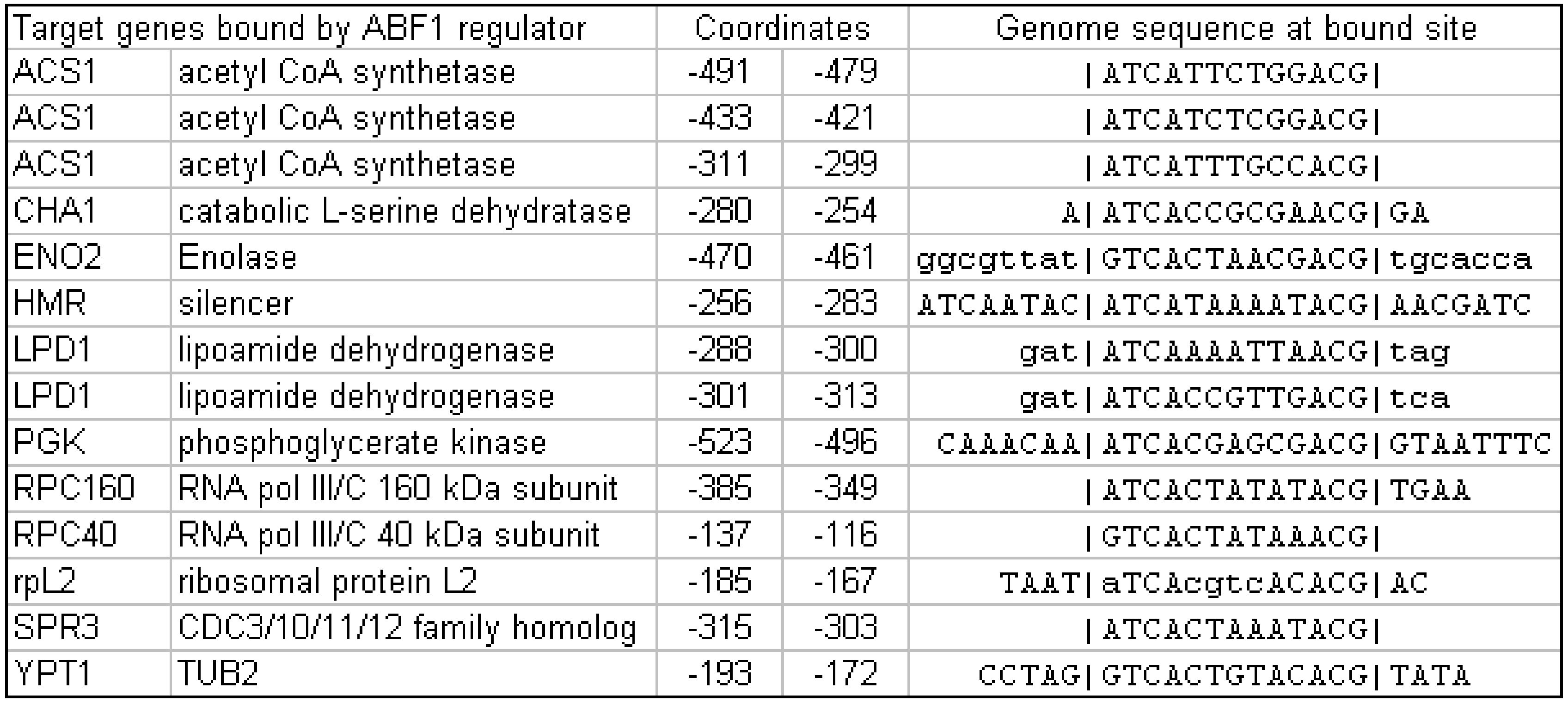
TCATTA

Regulator structure ⬄ recognized motifs

- Proteins ‘feel’ DNA
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)
- 3D Topology dictates specificity
  - Fully constrained positions:
    ➔ every atom matters
  - “Ambiguous / degenerate” positions
    ➔ loosely contacted
- Other types of recognition
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination



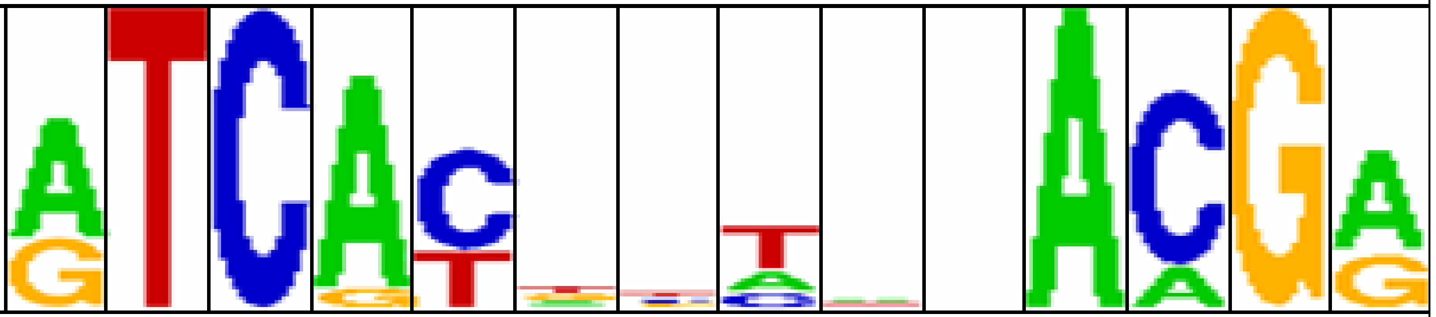
Motifs summarize TF sequence specificity



Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	|ATCATTCTGGACG|
ACS1	acetyl CoA synthetase	-433	-421	|ATCATCTCGGACG|
ACS1	acetyl CoA synthetase	-311	-299	|ATCATTTGCCACG|
CHA1	catabolic L-serine dehydratase	-280	-254	A|ATCACCGCGAACG|GA
ENO2	Enolase	-470	-461	ggcgttat|GTCACTAACGACG|tgcacca
HMR	silencer	-256	-283	ATCAATAC|ATCATAAAATACG|AACGATC
LPD1	lipoamide dehydrogenase	-288	-300	gat|ATCAAAATTAACG|tag
LPD1	lipoamide dehydrogenase	-301	-313	gat|ATCACCGTTGACG|tca
PGK	phosphoglycerate kinase	-523	-496	CAAACAA|ATCACGAGCGACG|GTAATTTC
RPC160	RNA pol III/C 160 kDa subunit	-385	-349	|ATCACTATATACG|TGAA
RPC40	RNA pol III/C 40 kDa subunit	-137	-116	|GTCACTATAAACG|
rpL2	ribosomal protein L2	-185	-167	TAAT|aTCAcgtcACACG|AC
SPR3	CDC3/10/11/12 family homolog	-315	-303	|ATCACTAAATACG|
YPT1	TUB2	-193	-172	CCTAG|GTCACTGTACACG|TATA

Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4

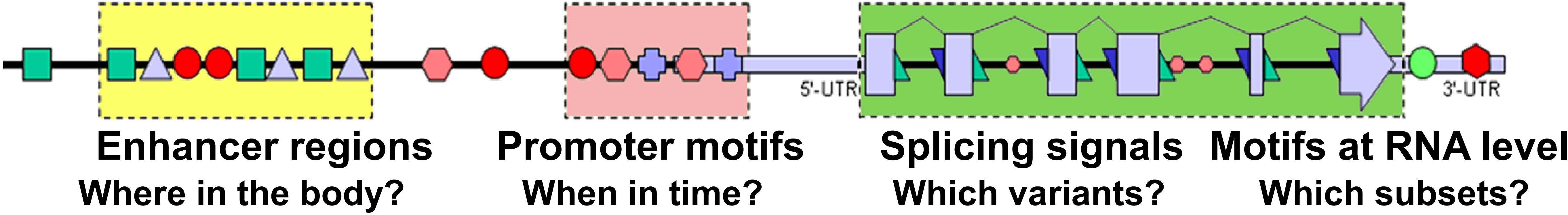
Motif Logo



Consensus	R	T	C	A	Y	N	N	H	N	N	A	C	G	R

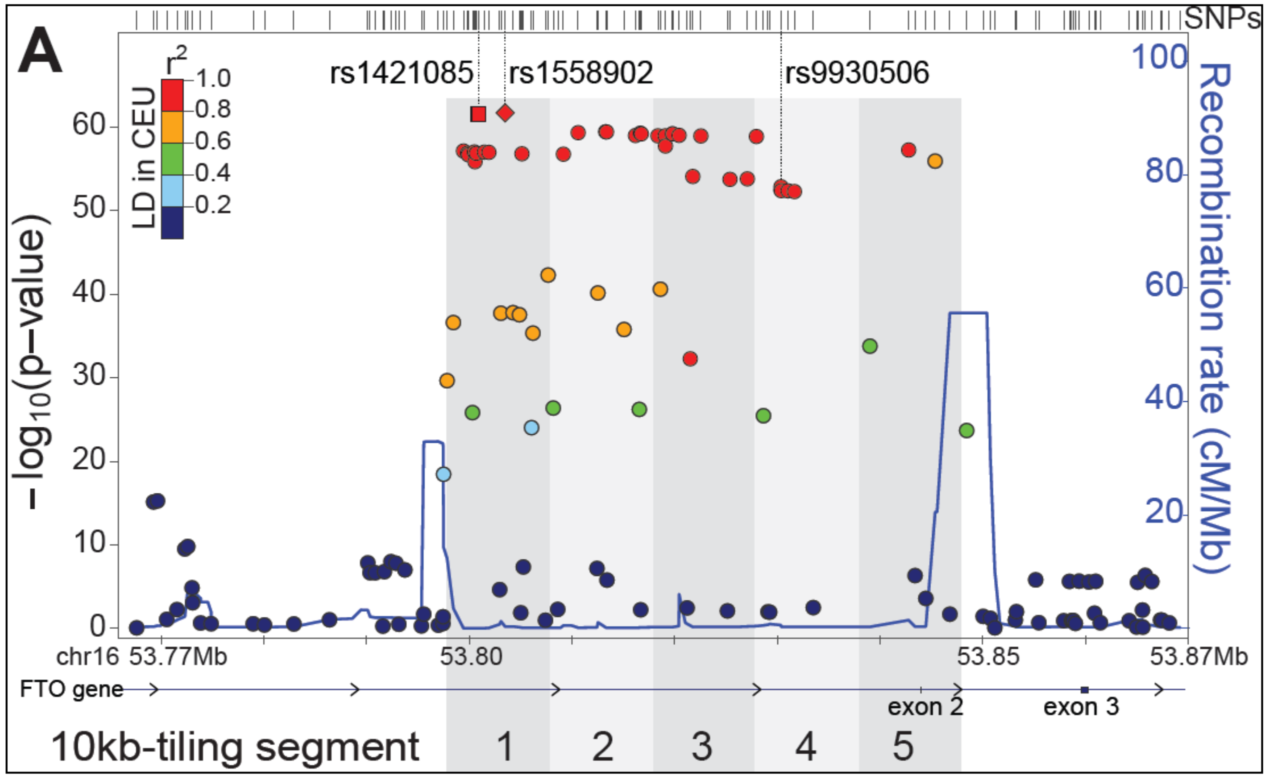
- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

Regulatory motifs at all levels of pre/post-tx regulation

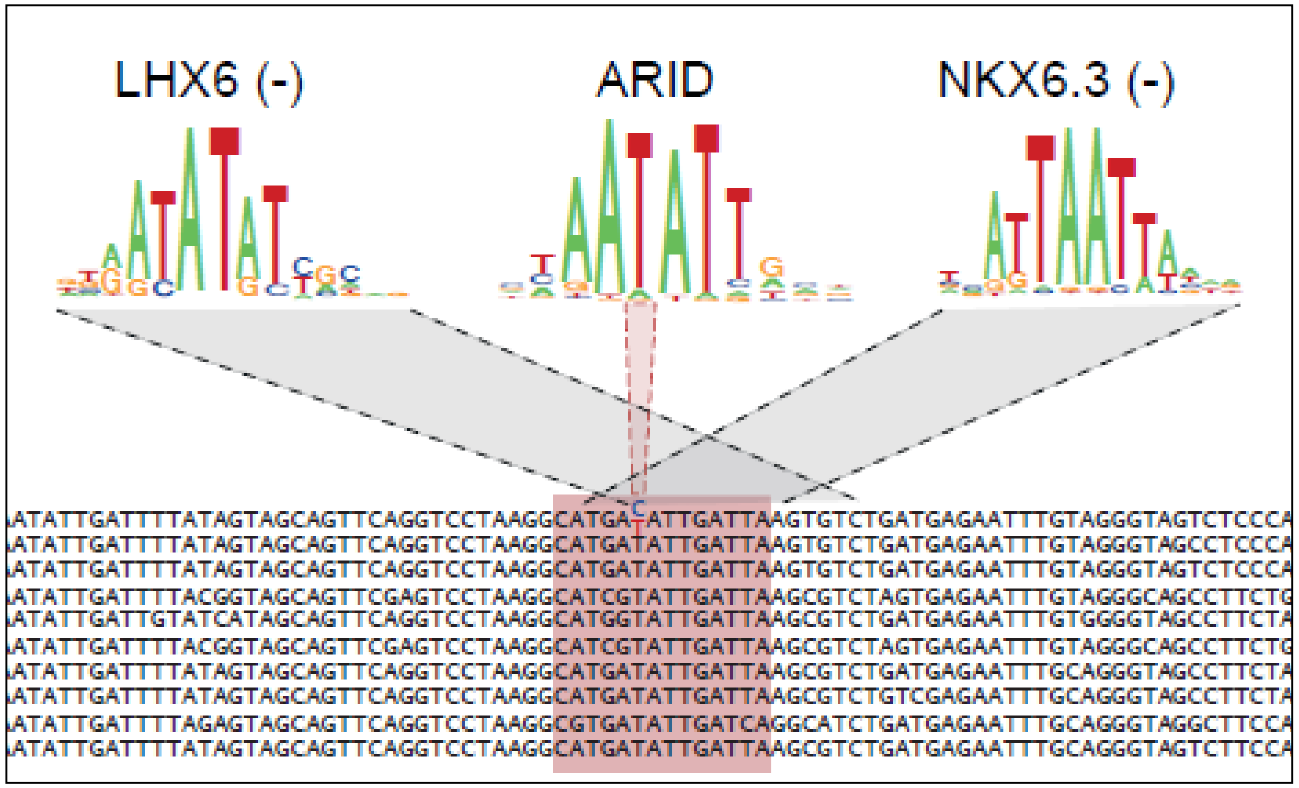


- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
    - Data-centric encoding of addresses
  - Overlaid with 'memory' marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- Today: discovering motifs in co-regulated promoters and *de novo* motif discovery & target identification

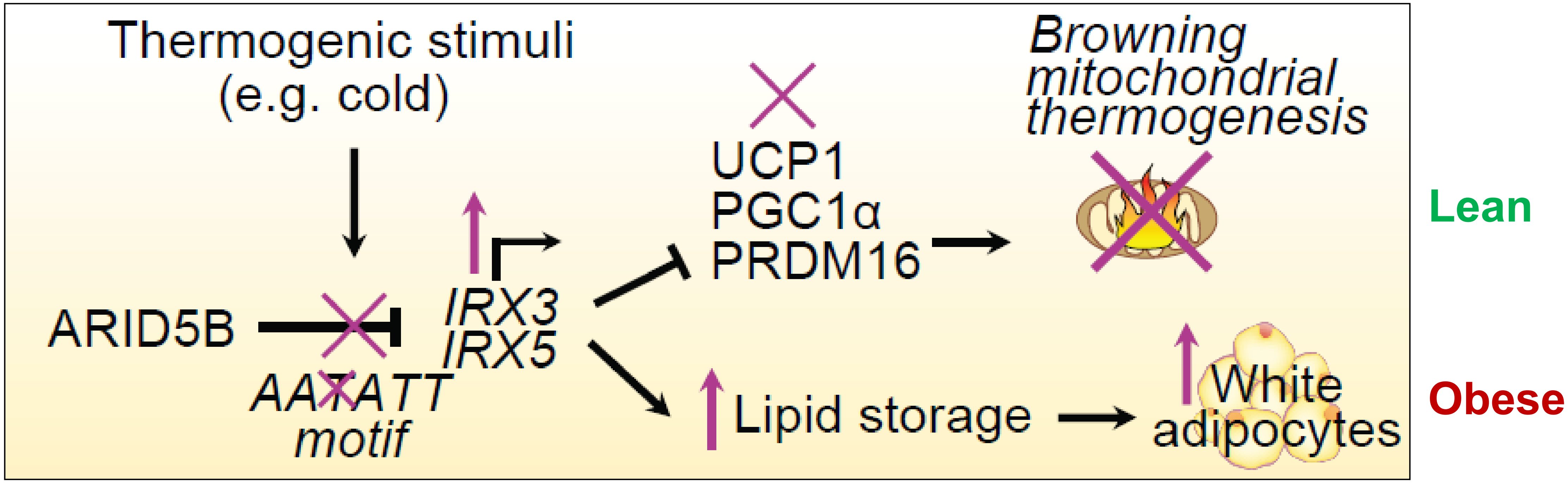
Disrupted motif at the heart of FTO obesity locus



Strongest association with obesity


LHX6 (-)
ARID
NKX6.3 (-)
C
T
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGTGTCTGATGAGAATTTGTAGGGTAGTCTCCCA
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGTGTCTGATGAGAATTTGTAGGGTAGCCTCCCA
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGTGTCTGATGAGAATTTGTAGGGTAGTCTCCCA
ATATTGATTTTACGGTAGCAGTTCGAGTCCTAAGGCATCGTATTGATTAAGCGTCTAGTGAGAATTTGTAGGGCAGCCTTCTG
ATATTGATTGTATCATAGCAGTTCAGGTCCTAAGGCATGGTATTGATTAAGCGTCTGATGAGAATTTGTGGGGTAGCCTTCTA
ATATTGATTTTACGGTAGCAGTTCGAGTCCTAAGGCATCGTATTGATTAAGCGTCTAGTGAGAATTTGTAGGGCAGCCTTCTG
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGCGTCTGATGAGAATTTGCAGGGTAGCCTTCTA
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGCGTCTGTCGAGAATTTGCAGGGTAGCCTTCTA
ATATTGATTTTAGAGTAGCAGTTCAGGTCCTAAGGCGTGATATTGATCAGGCATCTGATGAGAATTTGCAGGGTAGGCTTCCA
ATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGATATTGATTAAGCGTCTGATGAGAATTTGCAGGGTAGTCTTCCA

C-to-T disruption of AT-rich regulatory motif



Thermogenic stimuli (e.g. cold)

ARID5B

AATATT motif

IRX3
IRX5

UCP1
PGC1α
PRDM16

Browning mitochondrial thermogenesis

Lipid storage

White adipocytes

Lean

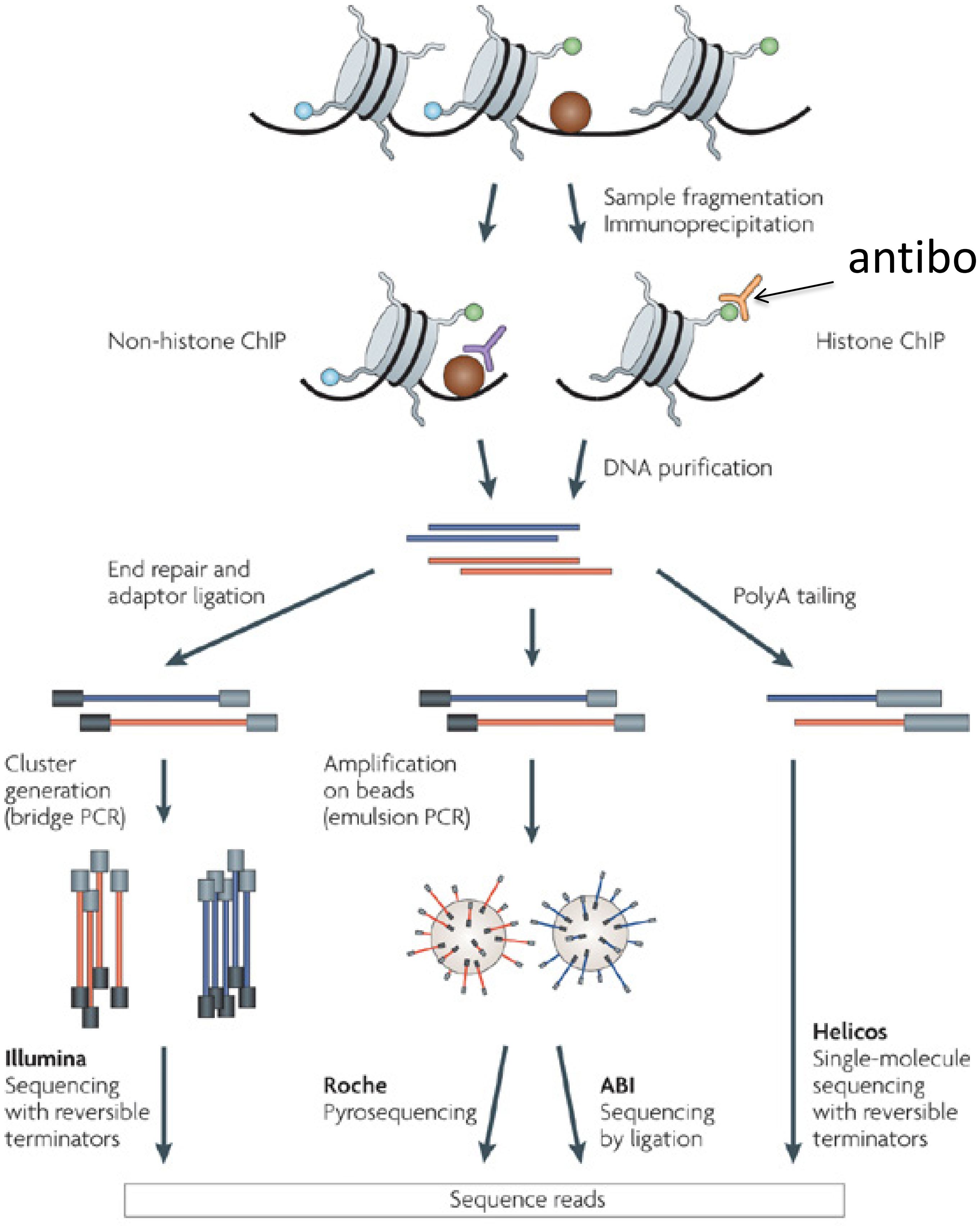
Obese

Restoring motif restores thermogenesis

1b. Technologies for probing gene regulation

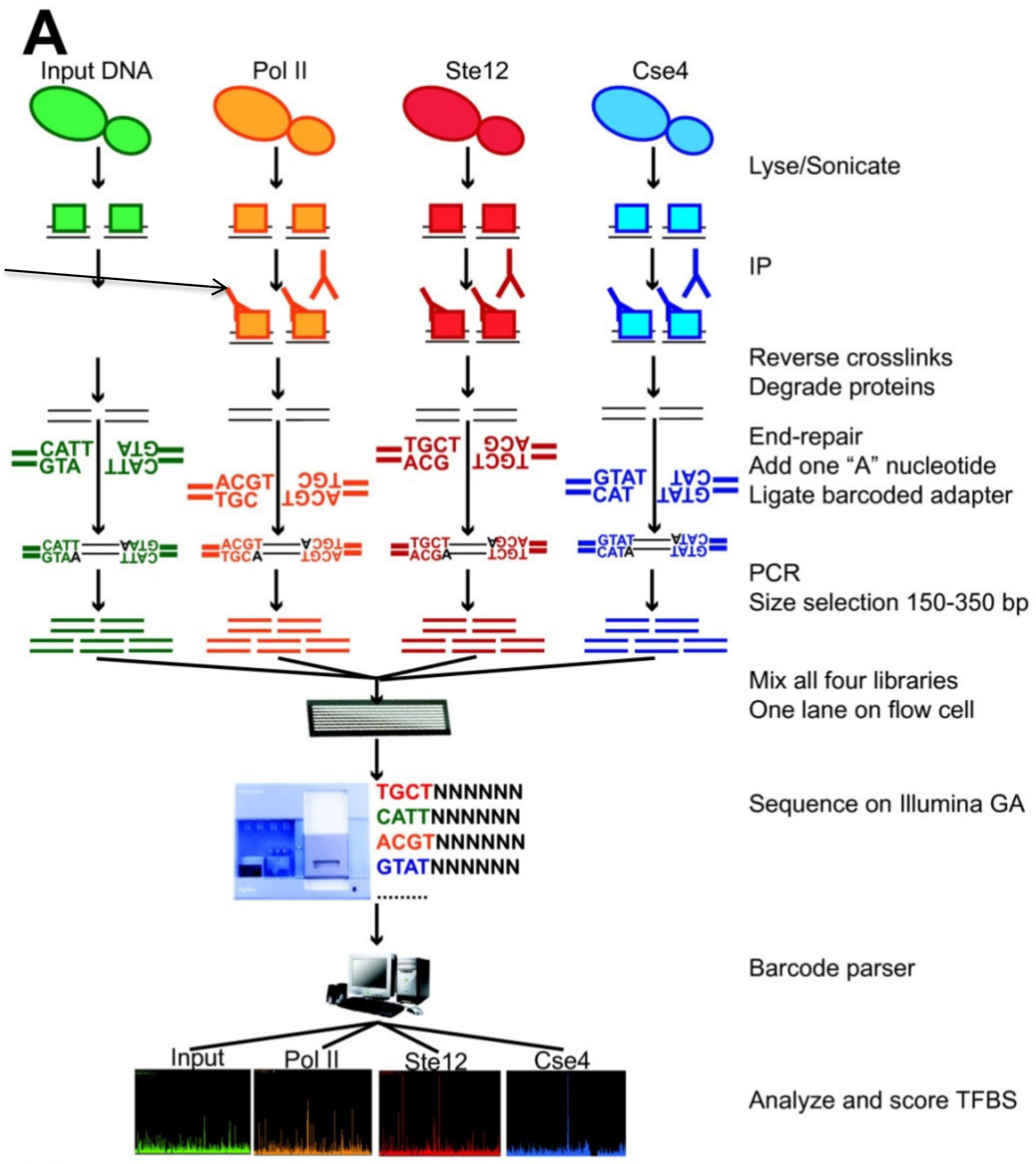
Mapping regulator binding: ChIP-seq

(Chromatin immunoprecipitation followed by sequencing) TF=transcription factor


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads

Nature Reviews | Genetics

antibody



Bar-coded multiplexed sequencing

ChIP-chip and ChIP-Seq technology overview

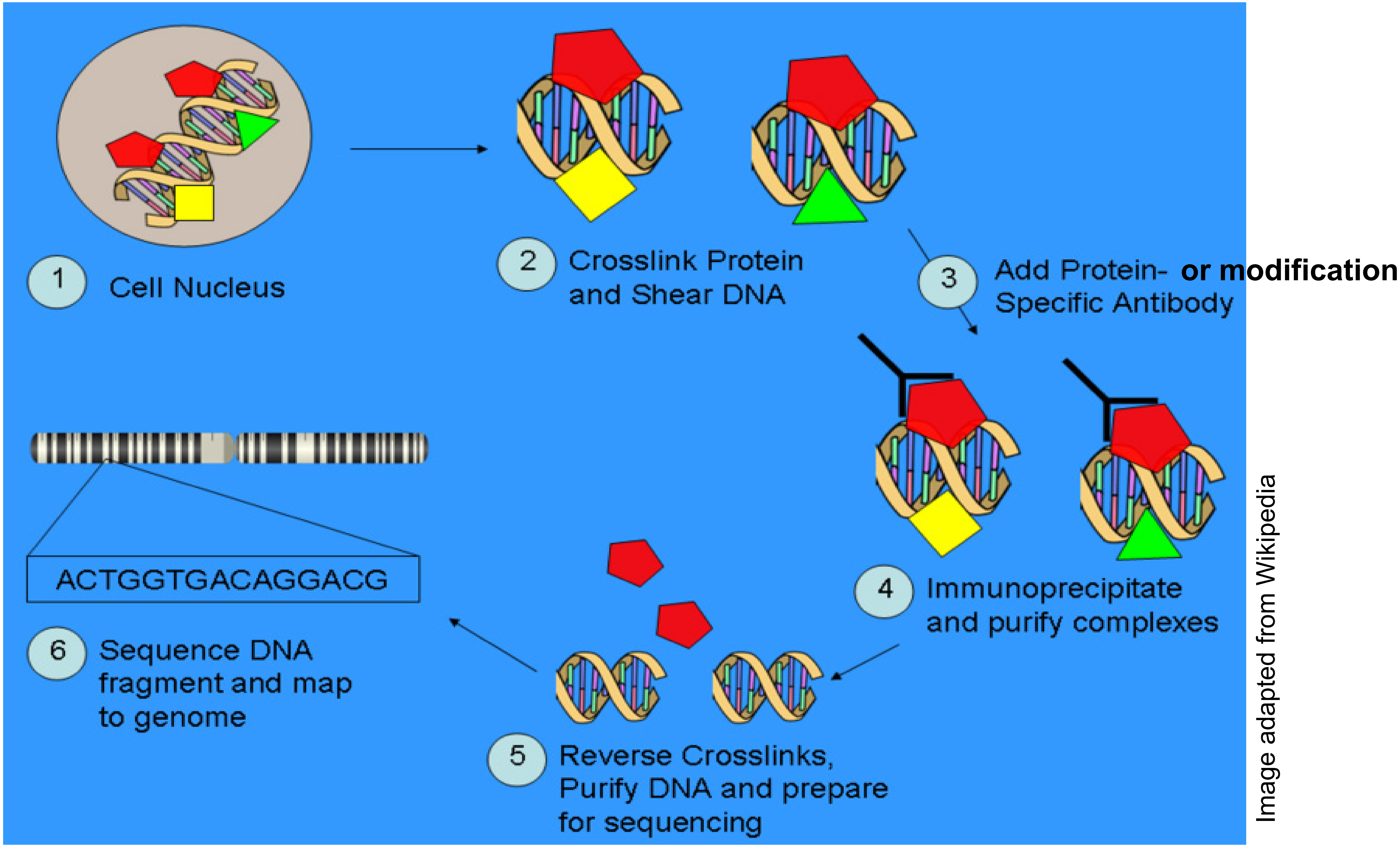

1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein- or modification Specific Antibody
4 Immunoprecipitate and purify complexes
5 Reverse Crosslinks, Purify DNA and prepare for sequencing
6 Sequence DNA fragment and map to genome
ACTGGTGACAGGACG

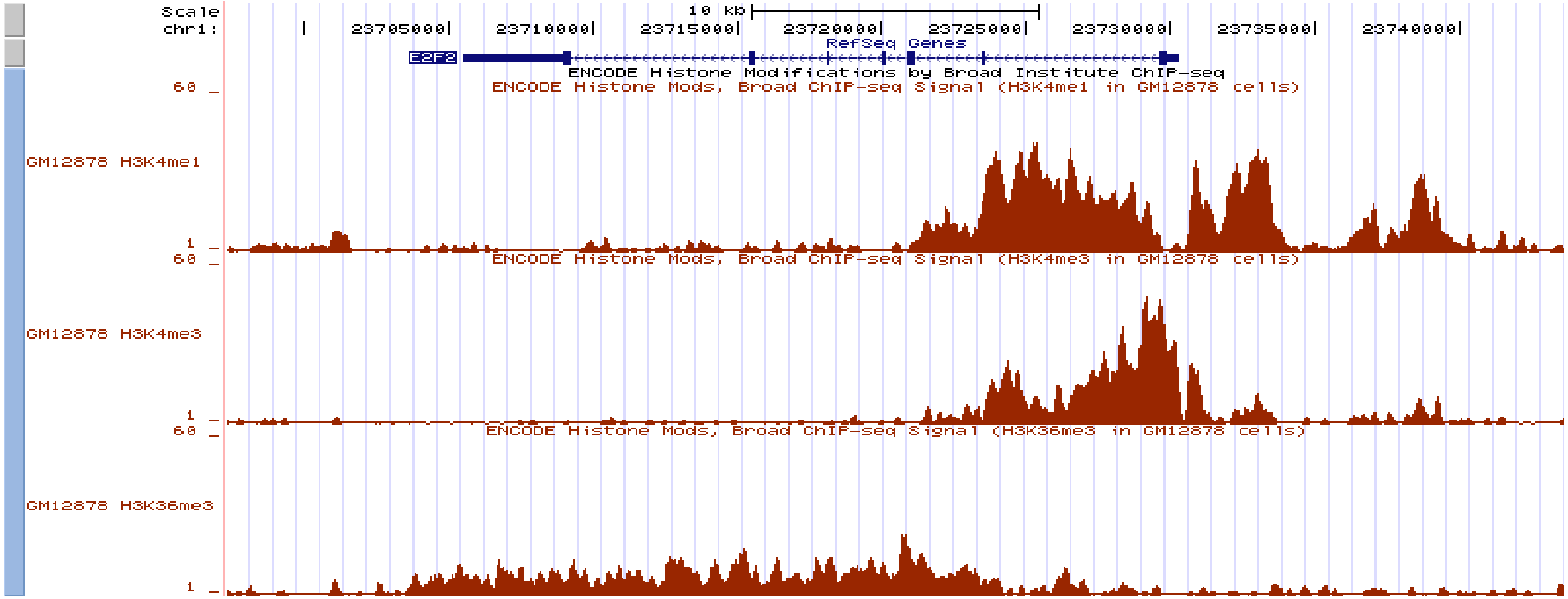
Image adapted from Wikipedia

Modification-specific antibodies → Chromatin Immuno-Precipitation

followed by: ChIP-chip: array hybridization

ChIP-Seq: Massively Parallel Next-gen Sequencing

ChIP-Seq Histone Modifications: What the raw data looks like

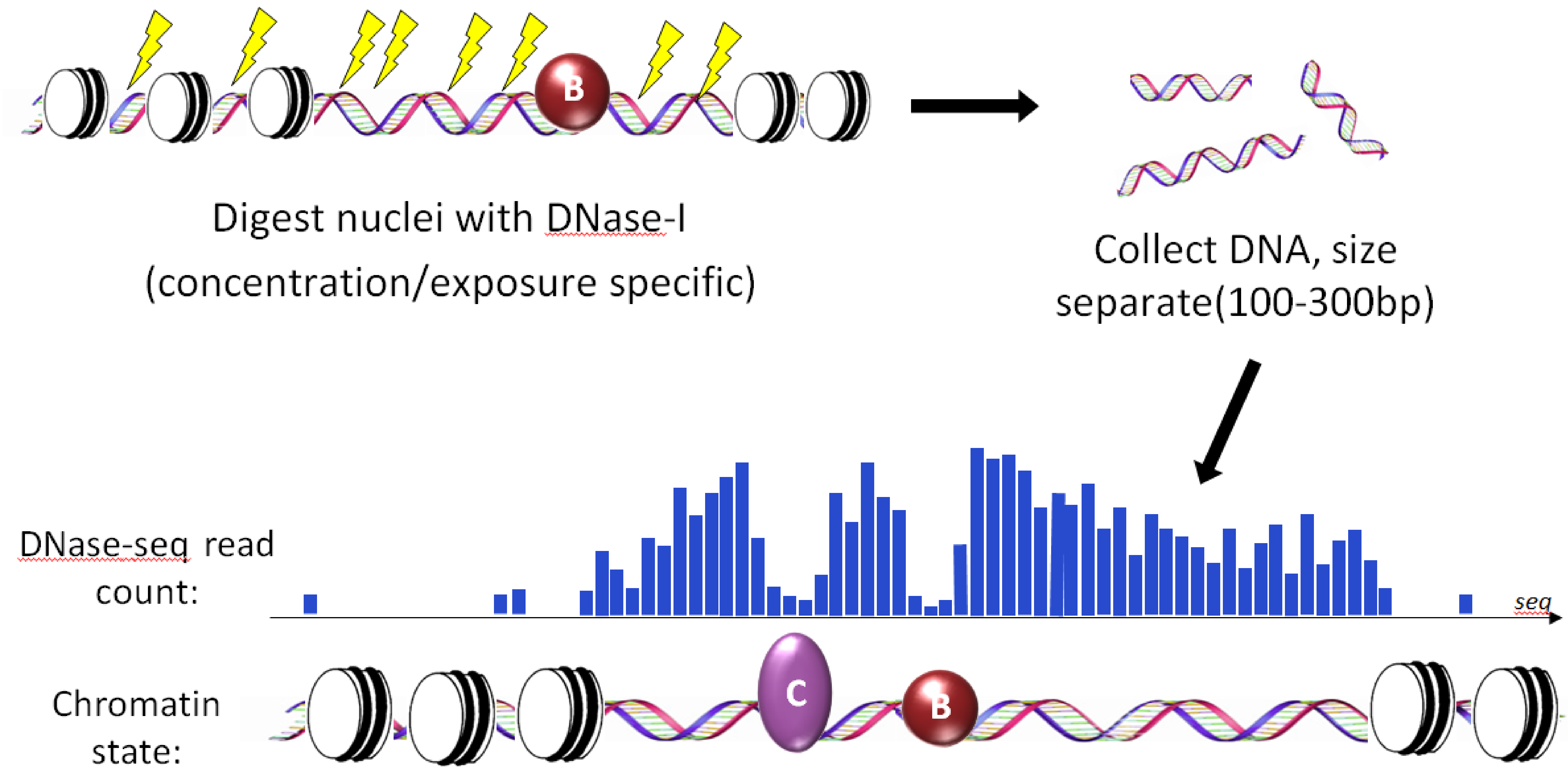

Scale
10 kb
chr1:
23705000
23710000
23715000
23720000
23725000
23730000
23735000
23740000
RefSeq Genes
E2F2
ENCODE Histone Modifications by Broad Institute ChIP-seq
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me1 in GM12878 cells)
60
GM12878 H3K4me1
1
60
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K4me3 in GM12878 cells)
GM12878 H3K4me3
1
60
ENCODE Histone Mods, Broad ChIP-seq Signal (H3K36me3 in GM12878 cells)
GM12878 H3K36me3
1

- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Chromatin accessibility can reveal TF binding

Sherwood, RI, et al. "**Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape**" *Nat. Biotech 2014.*

DNase-seq reveals genome protection profiles



Digest nuclei with DNase-I (concentration/exposure specific)

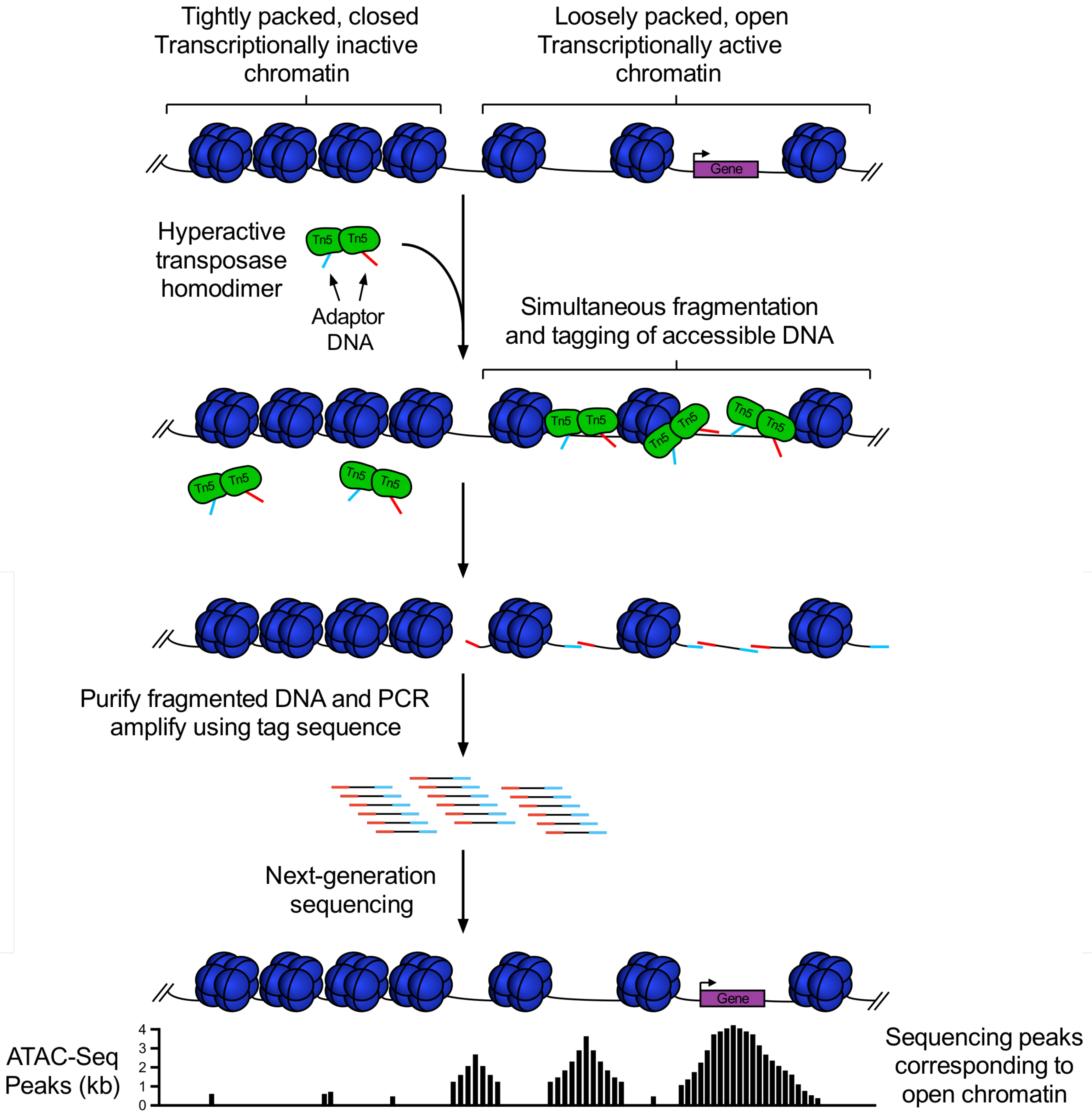
Collect DNA, size separate(100-300bp)

DNase-seq read count:

Chromatin state:

Sequence (60-100M reads)

Assay for Transposase-Accessible Chromatin (ATAC-seq)



Tightly packed, closed
Transcriptionally inactive
chromatin

Loosely packed, open
Transcriptionally active
chromatin

Hyperactive
transposase
homodimer

Adaptor
DNA

Simultaneous fragmentation
and tagging of accessible DNA

Purify fragmented DNA and PCR
amplify using tag sequence

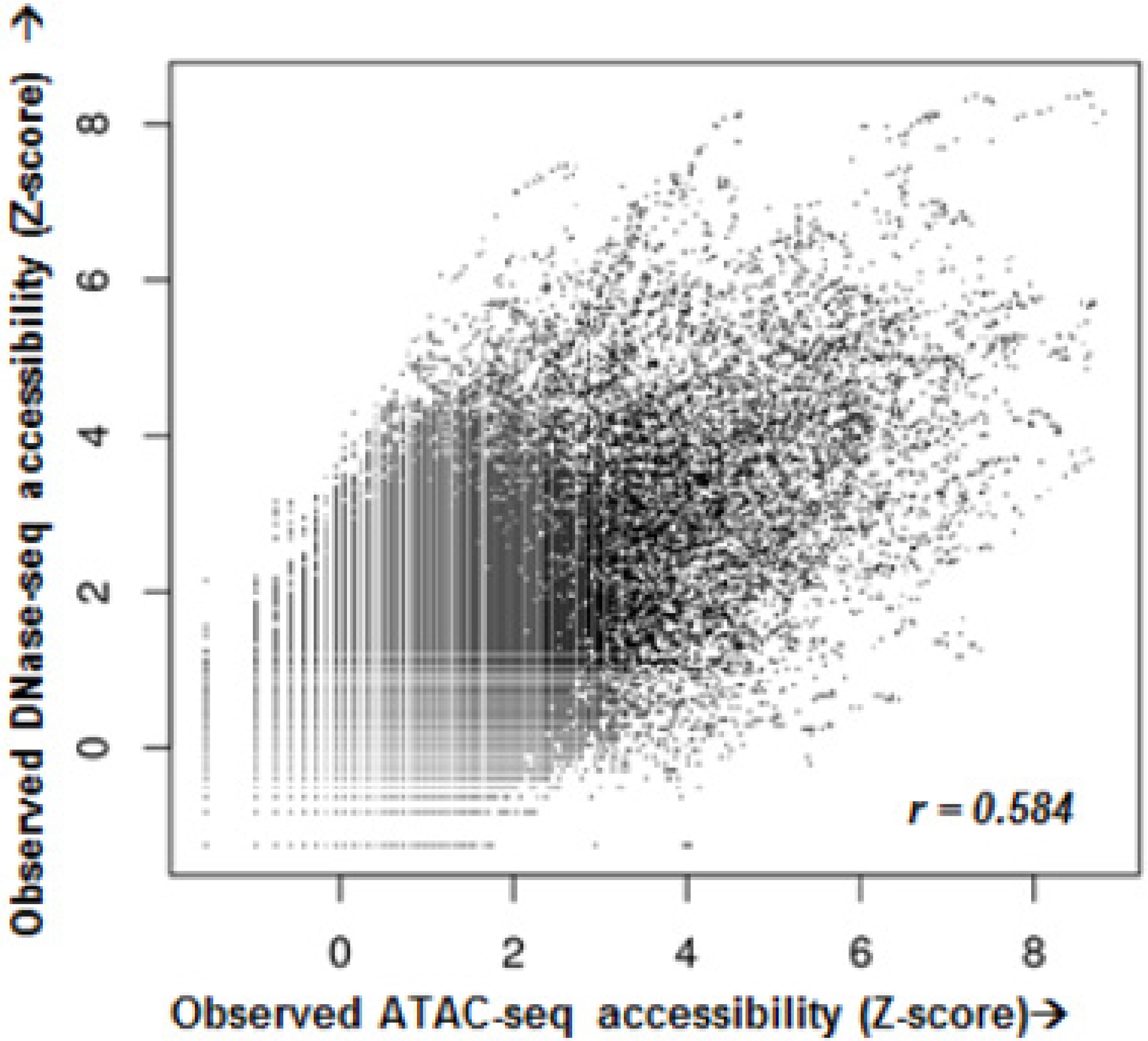
Next-generation
sequencing

ATAC-Seq
Peaks (kb)

Sequencing peaks
corresponding to
open chromatin

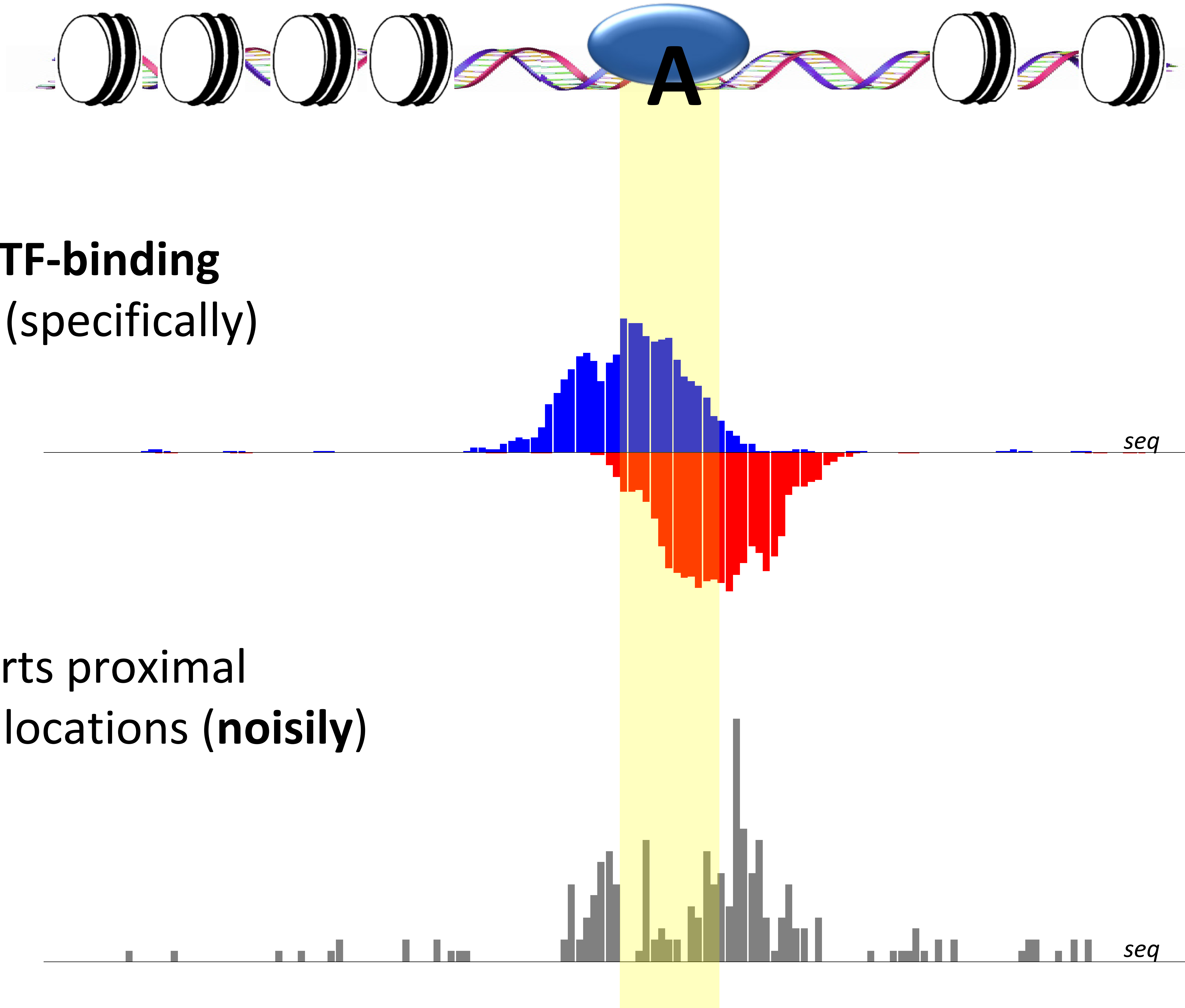
ATAC-seq and DNase-seq are not identical

GM12878, Chr. 14,
Each point is accessibility in a 2 kb window



Hashimoto TB, et al. "A Synergistic DNA Logic Predicts Genome-wide Chromatin Accessibility" *Genome Research* 2016

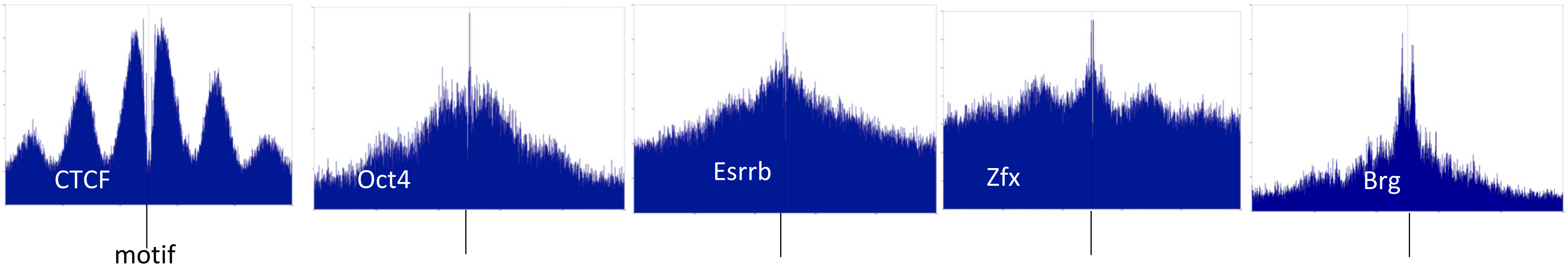
DNase-seq is less defined evidence than ChIP-seq



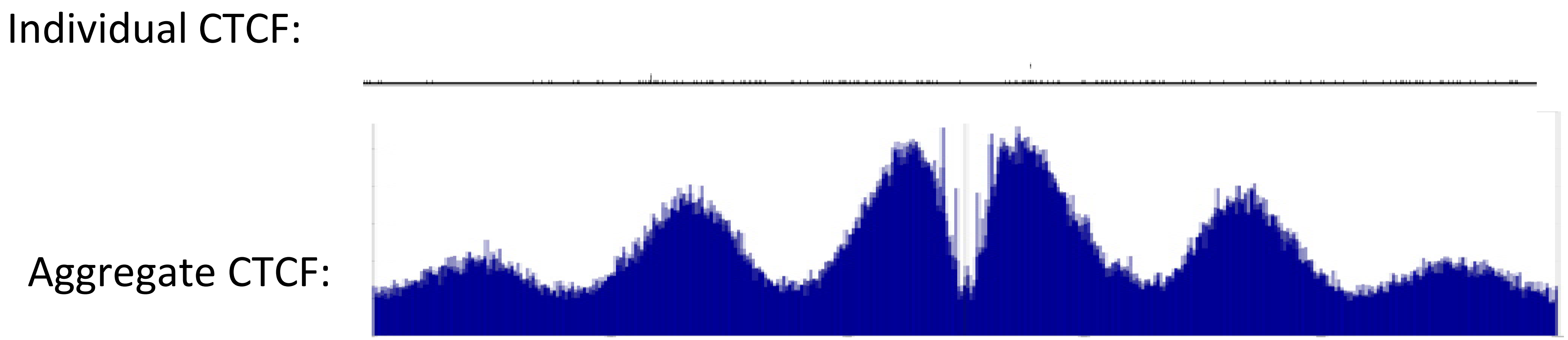
ChIP-seq reports **TF-binding** locations regions (specifically)

DNase-seq reports proximal **TF-non-binding** locations (**noisily**)

Bound factors leave distinct DNase-seq profiles

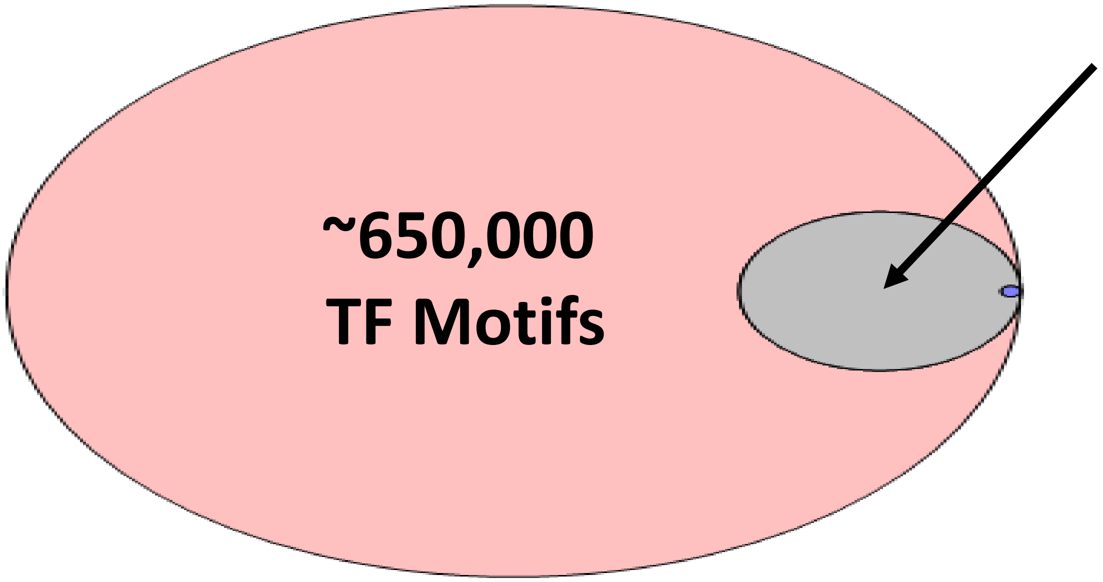


Individual binding site prediction is difficult



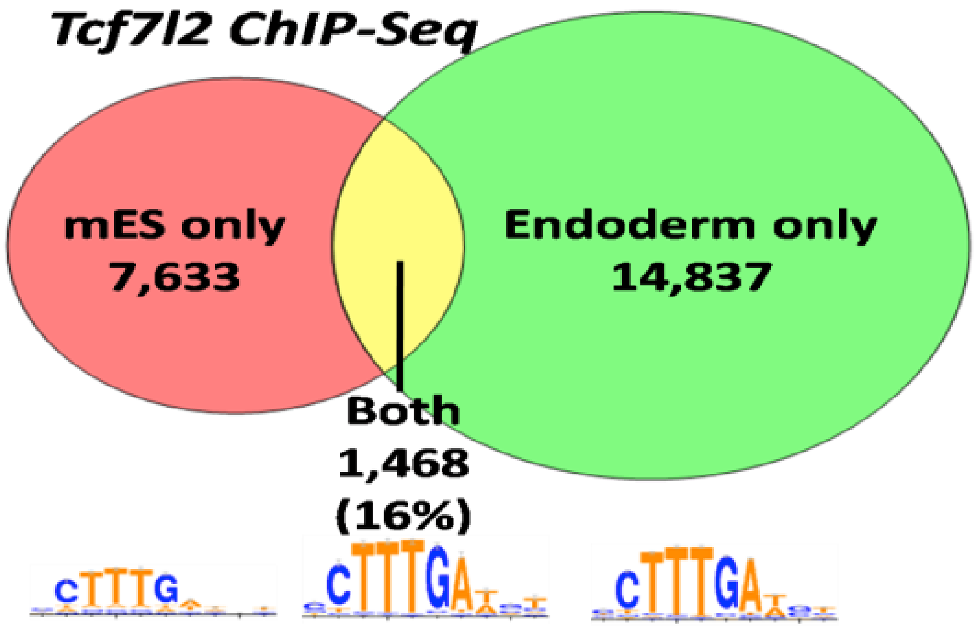
Motifs can predict TF binding

~50,000 binding sites for a typical TF



~650,000 TF Motifs

Binding sites change across time



Tcf7l2 ChIP-Seq

mES only 7,633

Endoderm only 14,837

Both 1,468 (16%)

Chromatin accessibly influences transcription factor binding

- Modeling accessibility profiles yields binding predictions and pioneer factor discovery
- Asymmetric accessibility is induced by *directional pioneers*
- The binding of *settler factors* can be enabled by proximal pioneer factor binding

Sherwood, RI, et al. "**Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape**" *Nat. Biotech 2014.*

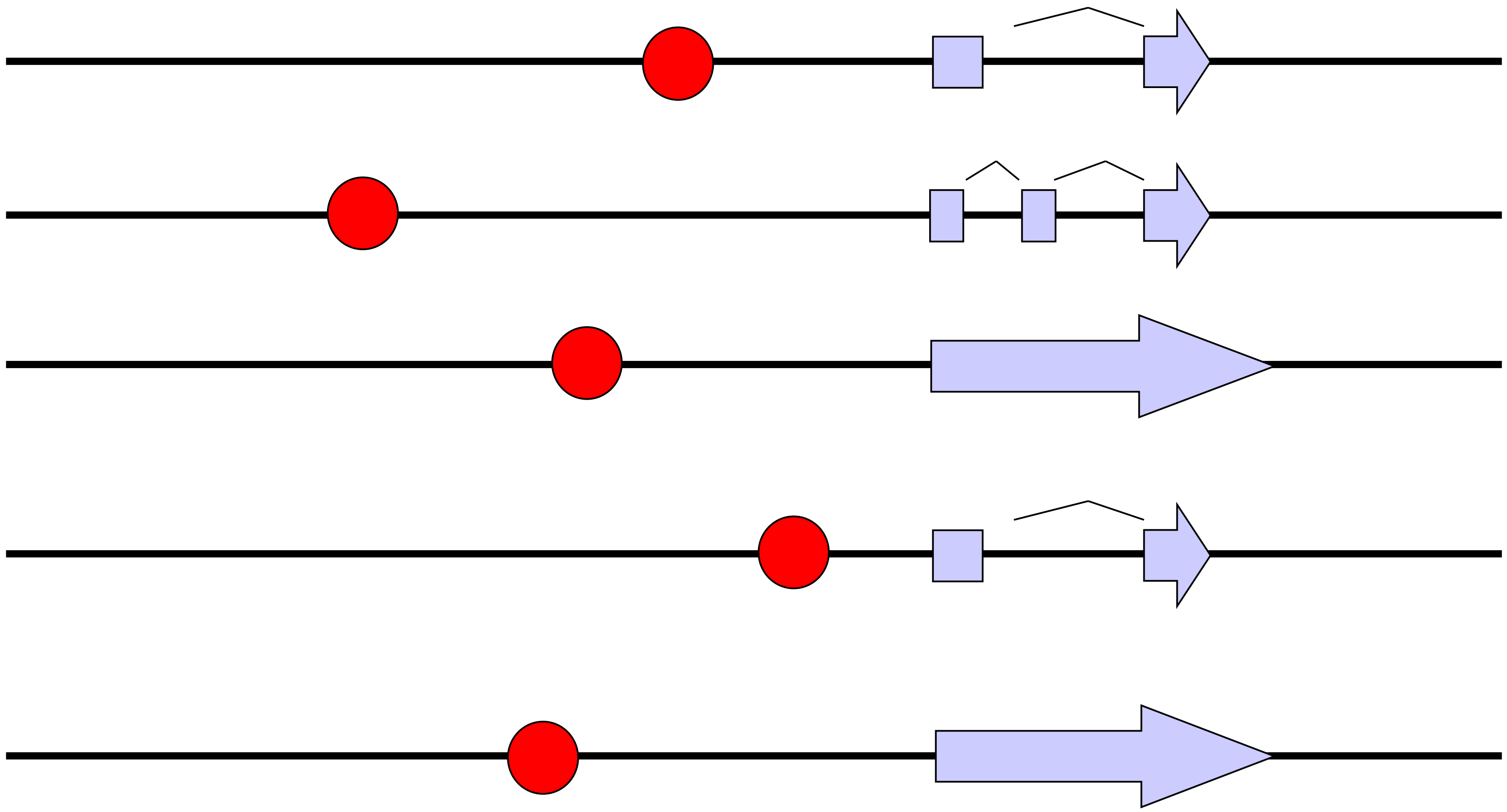
Deep Learning for Regulatory Genomics

1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⇔ DNA letters. Patches/filters ⇔ Motifs. Higher ⇔ combinations
   - Learning convolutional filters ⇔ Motif discovery. Applying them ⇔ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

2. Classical regulatory genomics (before Deep Learning)

Enrichment-based discovery methods

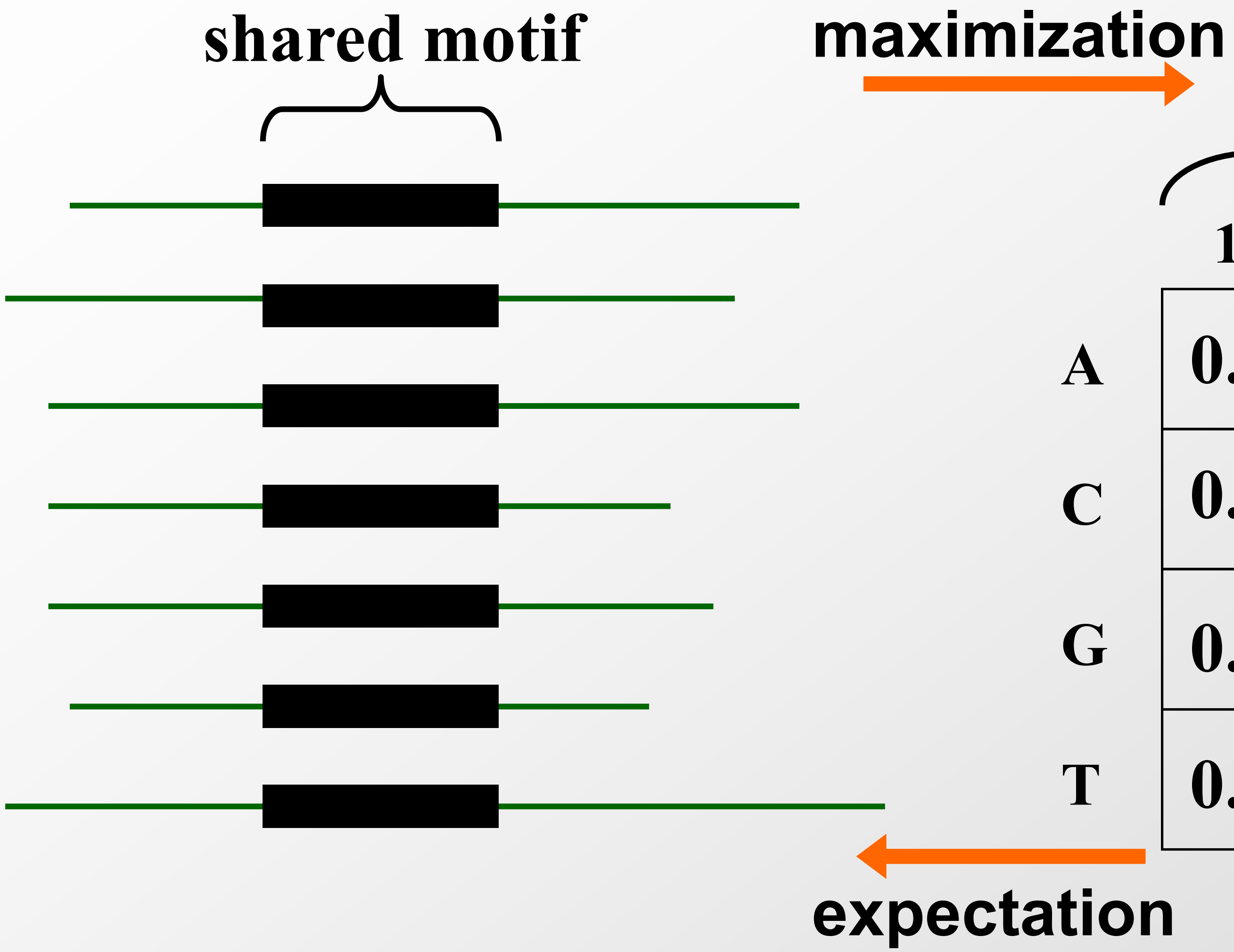
Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions



- **Align the promoters to each other using local alignment**
- **Use expert knowledge for what motifs should look like**
- **Find ‘median’ string by enumeration (motif/sample driven)**
- **Start with conserved blocks in the upstream regions**

Starting positions ⇔ Motif matrix

- given <u>aligned</u> sequences ➔ easy to compute profile matrix



sequence positions

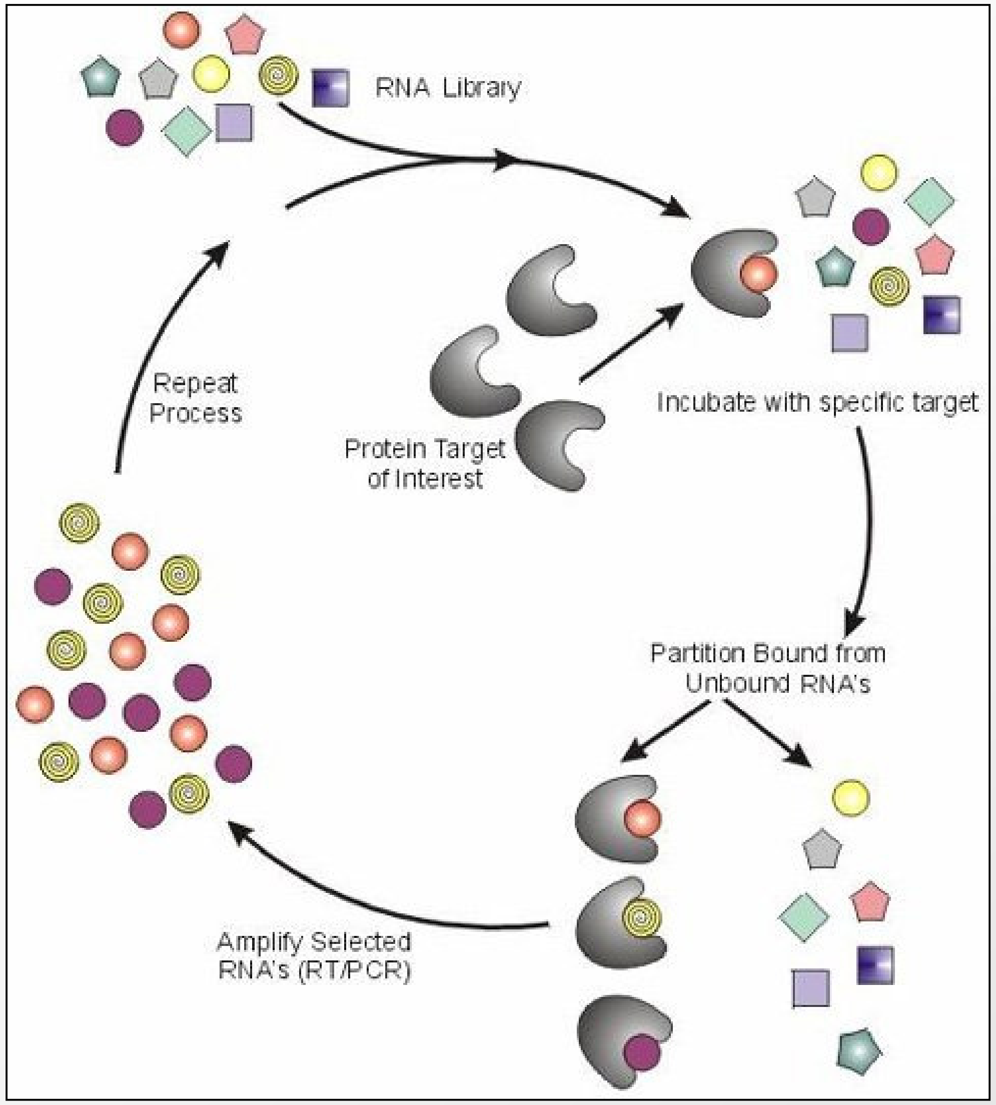
	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

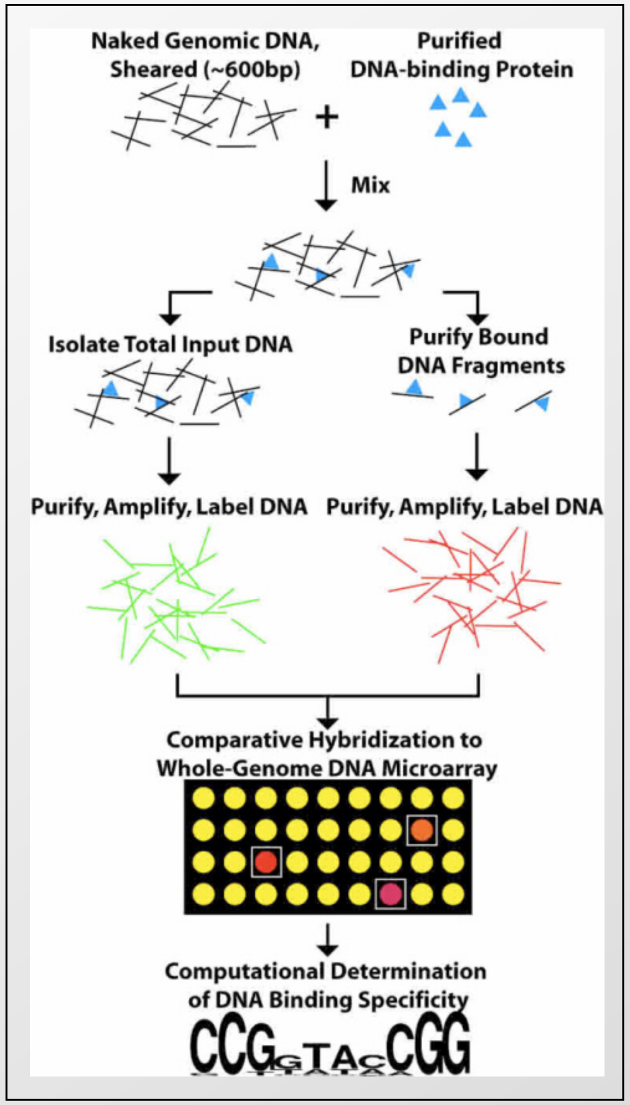
given profile matrix

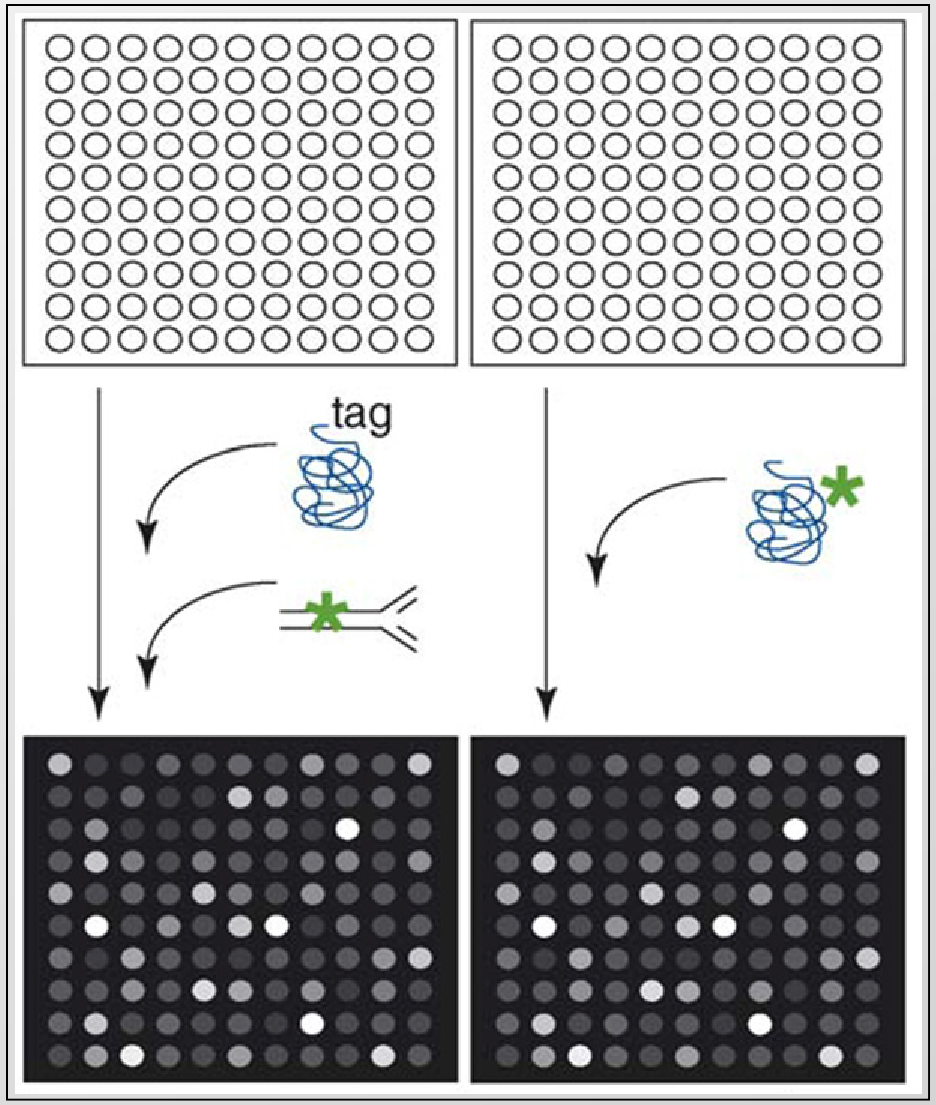
- **easy to find starting position probabilities**

Key idea: Iterative procedure for estimating both, given uncertainty
(learning problem with hidden variables: the starting positions)

Experimental factor-centric discovery of motifs


RNA Library
Incubate with specific target
Protein Target of Interest
Repeat Process
Partition Bound from Unbound RNA's
Amplify Selected RNA's (RT/PCR)


Naked Genomic DNA, Sheared (~600bp)
Purified DNA-binding Protein
+
Mix
Isolate Total Input DNA
Purify Bound DNA Fragments
Purify, Amplify, Label DNA
Purify, Amplify, Label DNA
Comparative Hybridization to Whole-Genome DNA Microarray
Computational Determination of DNA Binding Specificity
CCGGTACCGG



SELEX (Systematic Evolution of Ligands by Exponential Enrichment; Klug & Famulok, 1994).

DIP-Chip (DNA-immunoprecipitation with microarray detection; Liu et al., 2005)

PBMs (Protein binding microarrays; Mukherjee, 2004)
Double stranded DNA arrays

Approaches to regulatory motif discovery

Region-based motif discovery

- Expectation Maximization (e.g. MEME)
  - Iteratively refine positions / motif profile
- Gibbs Sampling (e.g. AlignACE)
  - Iteratively sample positions / motif profile
- Enumeration with wildcards (e.g. Weeder)
  - Allows global enrichment/background score
- Peak-height correlation (e.g. MatrixREDUCE)
  - Alternative to cutoff-based approach

Genome-wide

- Conservation-based discovery (e.g. MCS)
  - Genome-wide score, up-/down-stream bias

In vitro / trans

- Protein Domains (e.g. PBMs, SELEX)
  - In vitro motif identification, seq-/array-based

Deep Learning for Regulatory Genomics

1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⬄ DNA letters. Patches/filters ⬄ Motifs. Higher ⬄ combinations
   - Learning convolutional filters ⬄ Motif discovery. Applying them ⬄ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

Deep convolutional neural network

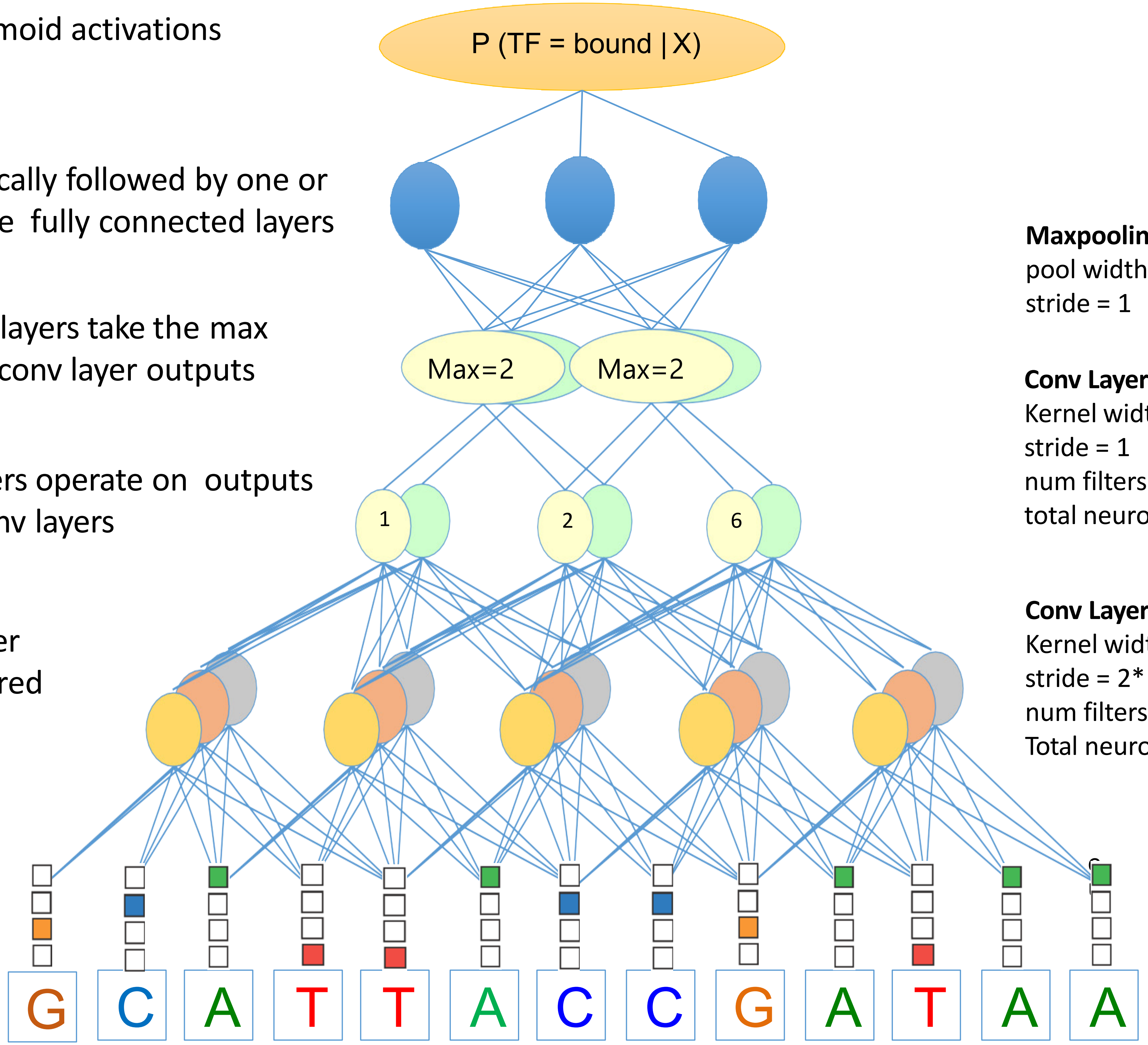
Sigmoid activations

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer (same color = shared weights)



Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num channels = 2
total neurons = 6

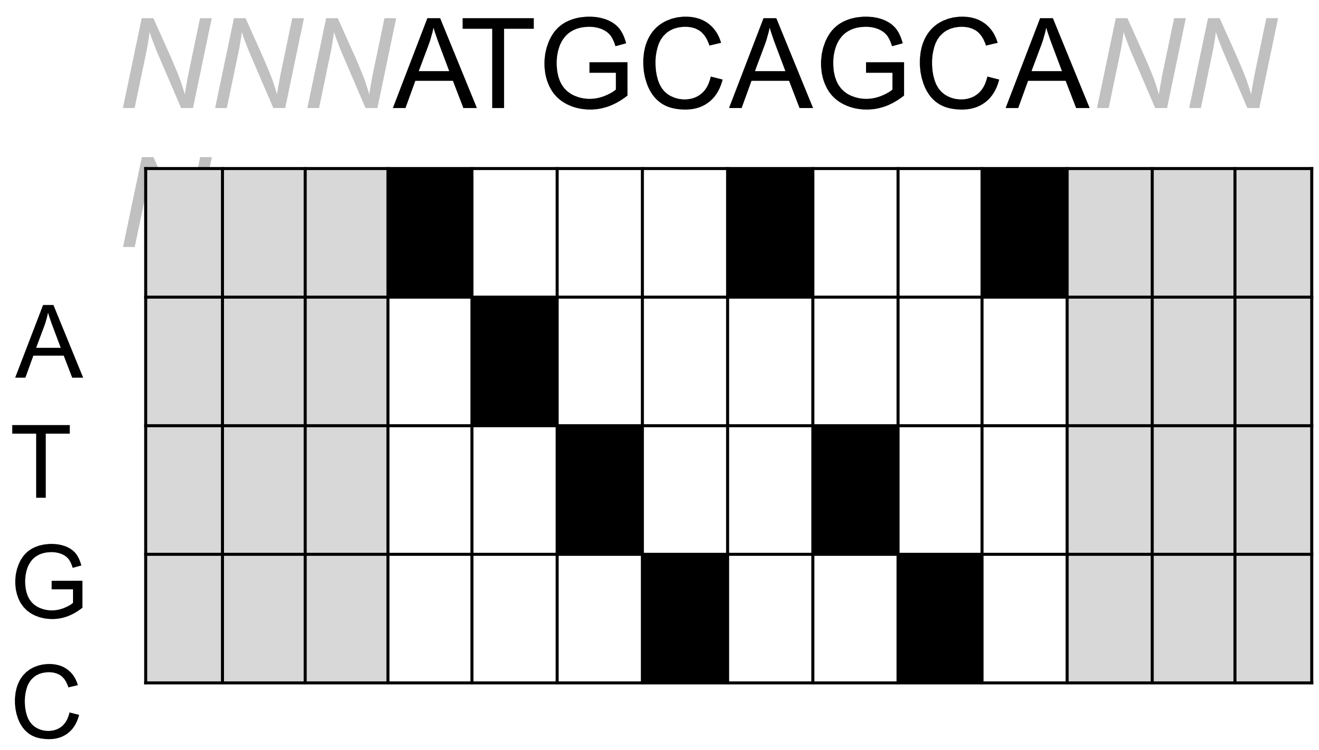
Conv Layer 1
Kernel width = 4
stride = 2*
num filters / num channels = 3
Total neurons = 15

***for genomics, a stride of 1 for conv layers is recommended**

3a. CNNs for Regulatory Genomics Foundations (Low-level features)

An example of using CNN to model DNA sequence

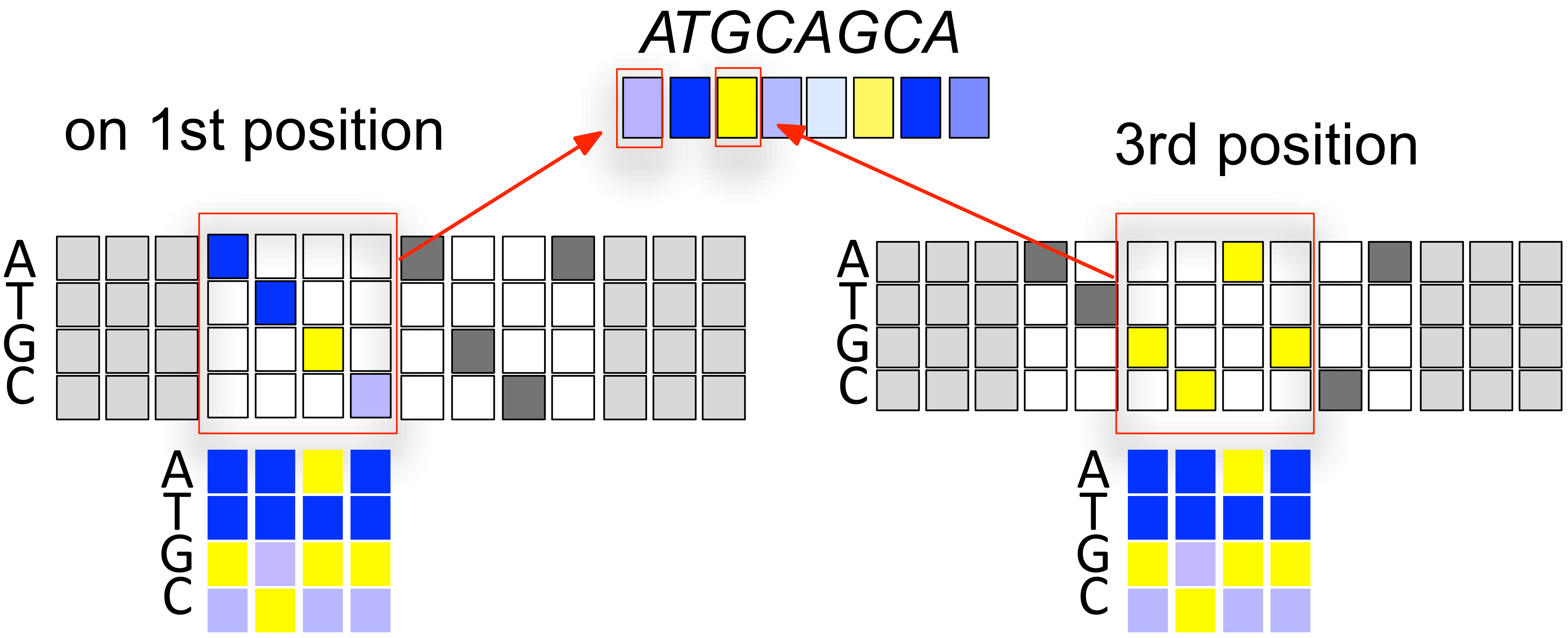
Representing DNA sequence as 2D matrix:



Matrix representation of
DNA sequence
(darker = stronger)

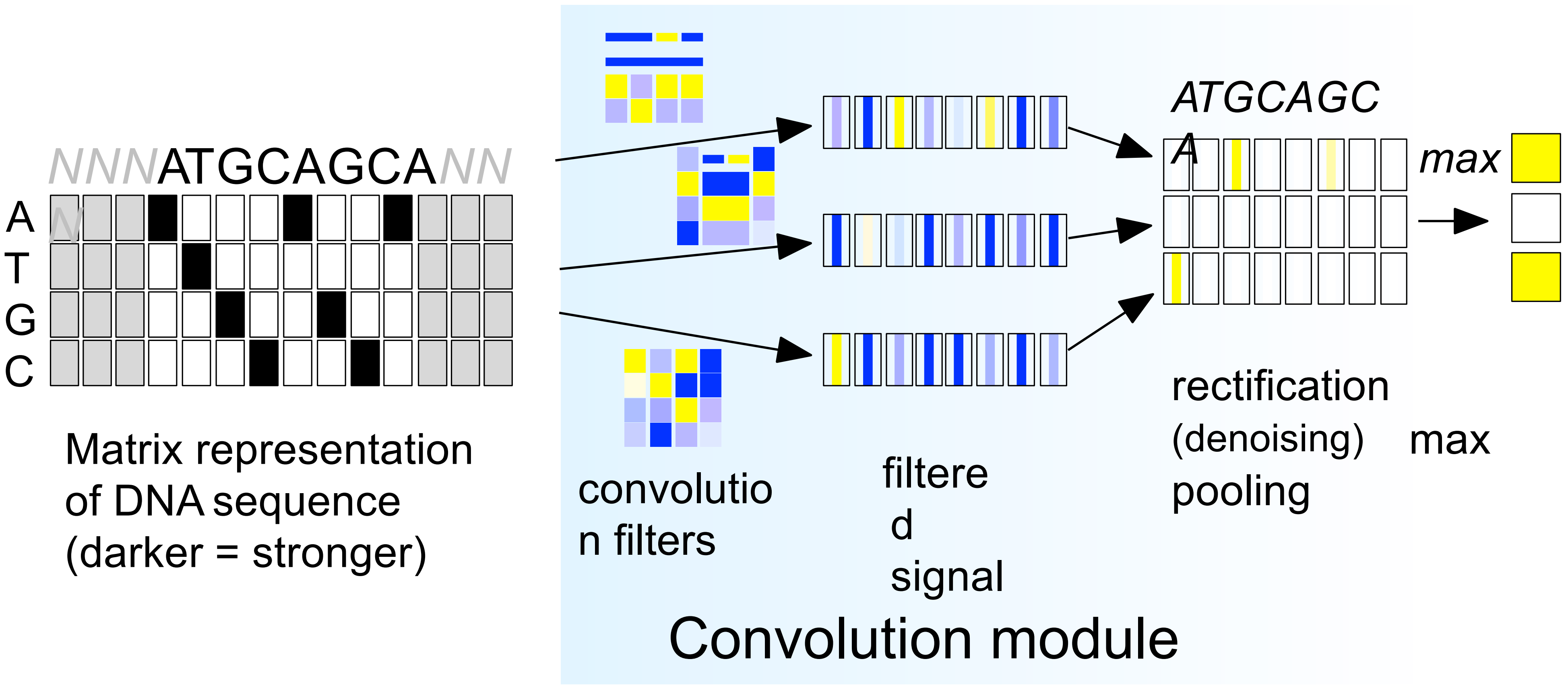
Convolution – extracting invariant feature

Applying 4 bp sequence filter along the DNA matrix:



Yellow = high activity; blue = low activity

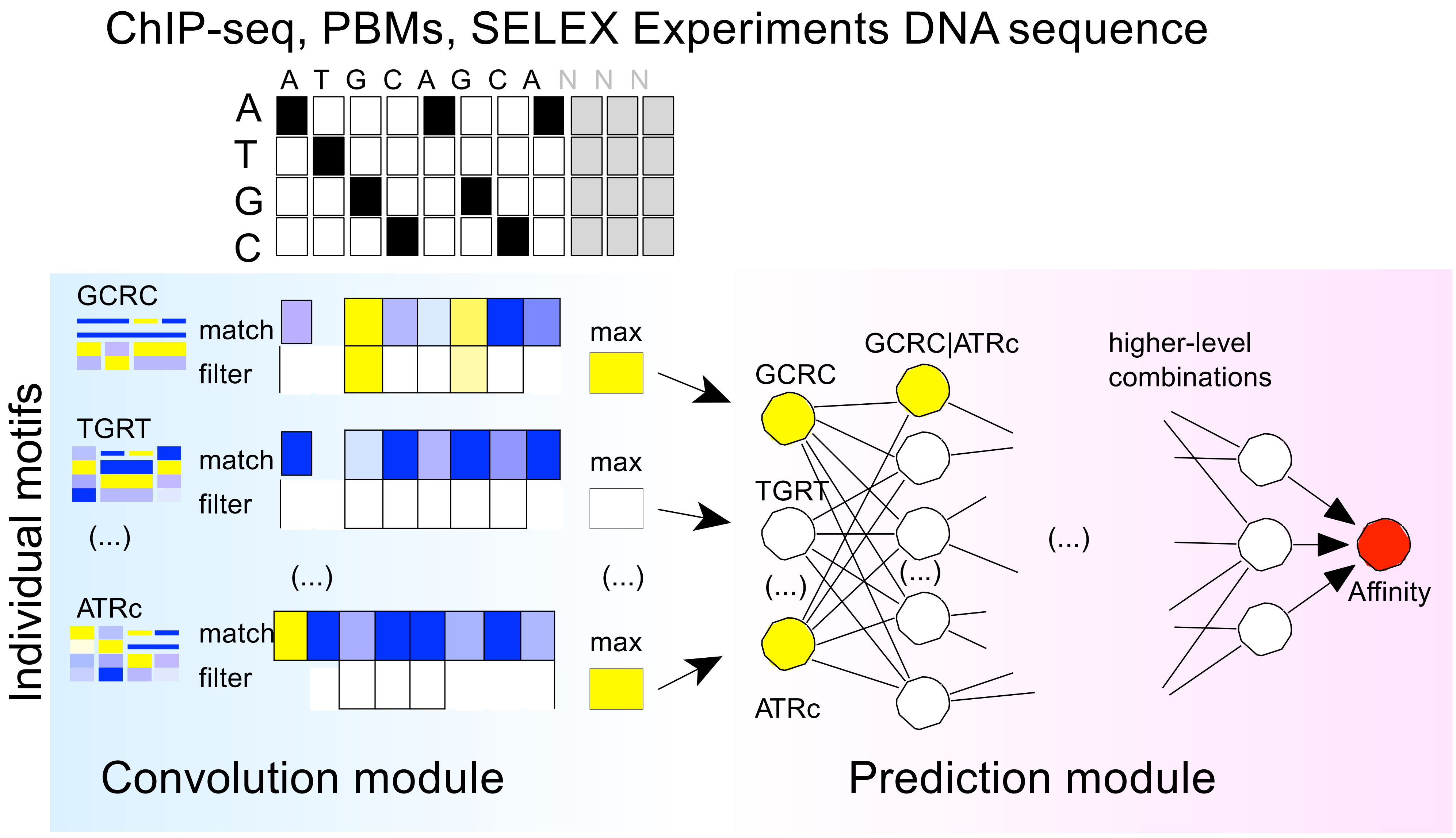
Convolution – extracting invariant feature



Rectification = ignore signals below some threshold.

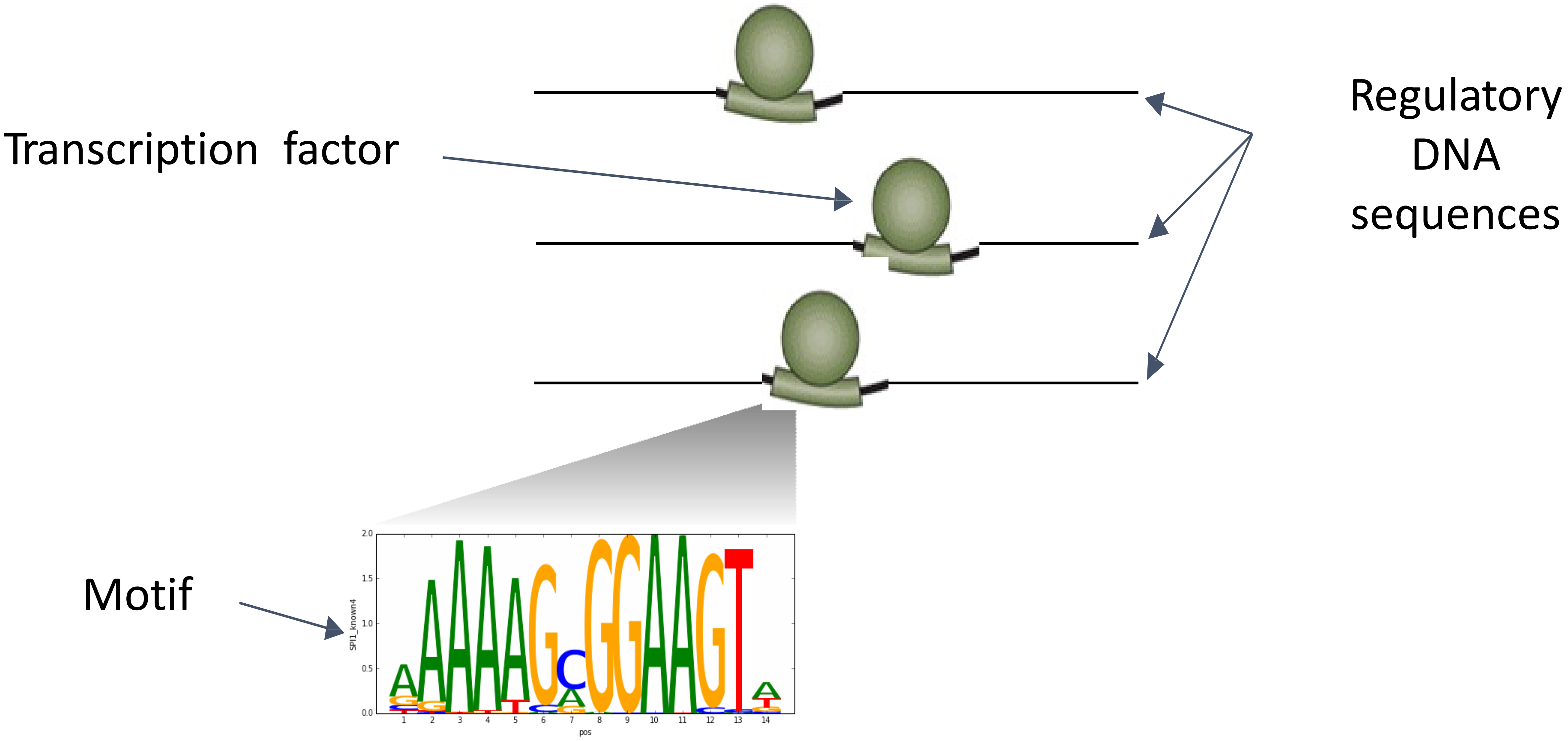
Pooling = summary of each channel by max or average.

Prediction using extracted features map



[Park and Kellis, 2015]

Key properties of regulatory sequence



Transcription factor

Regulatory DNA sequences

Motif

TRANSCRIPTION FACTOR BINDING

Regulatory proteins called **transcription factors (TFs)** bind to high affinity sequence patterns (**motifs**) in regulatory DNA

Sequence motifs: PWM

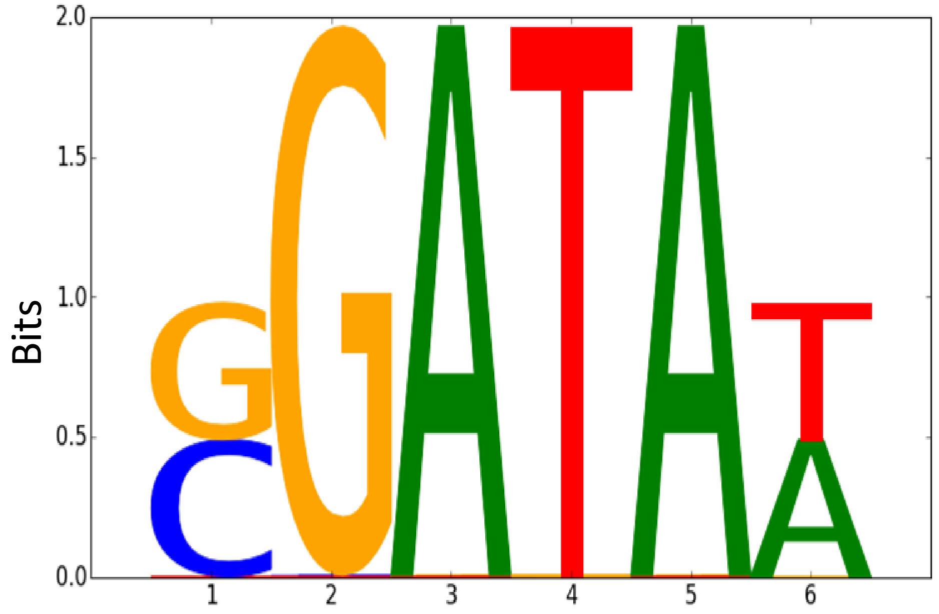
GGATAA
CGATAA
CGATAT
GGATAT

Set of aligned sequences
Bound by TF

$p_i(x_i = a_i)$

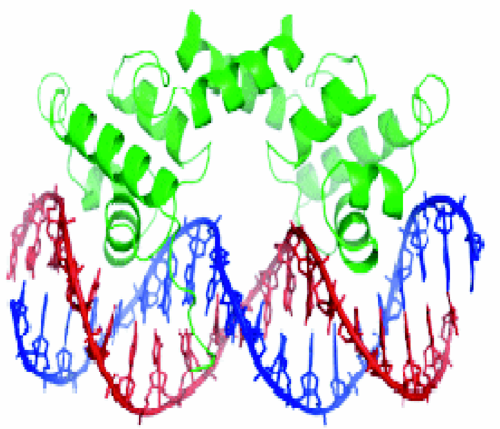
A	0	0	1	0	1	0.5
C	0.5	0	0	0	0	0
G	0.5	1	0	0	0	0
T	0	0	0	1	0	0.5

Position weight matrix
(PWM)



PWM logo

https://en.wikipedia.org/wiki/Sequence_logo



..ATGGATTCCTCC..
..GCATATAGCTAT..
..GTGAACTGGCTG..

The information content (y-axis) of position i is given by:[2]

$$R_i = \log_2(4) - (H_i + e_n)$$

where H_i is the uncertainty (sometimes called the Shannon entropy) of position i

$$H_i = -\sum f_{a,i} \times \log_2 f_{a,i}$$

. The height of letter a in column i is given by

$$\text{height} = f_{a,i} \times R_i$$

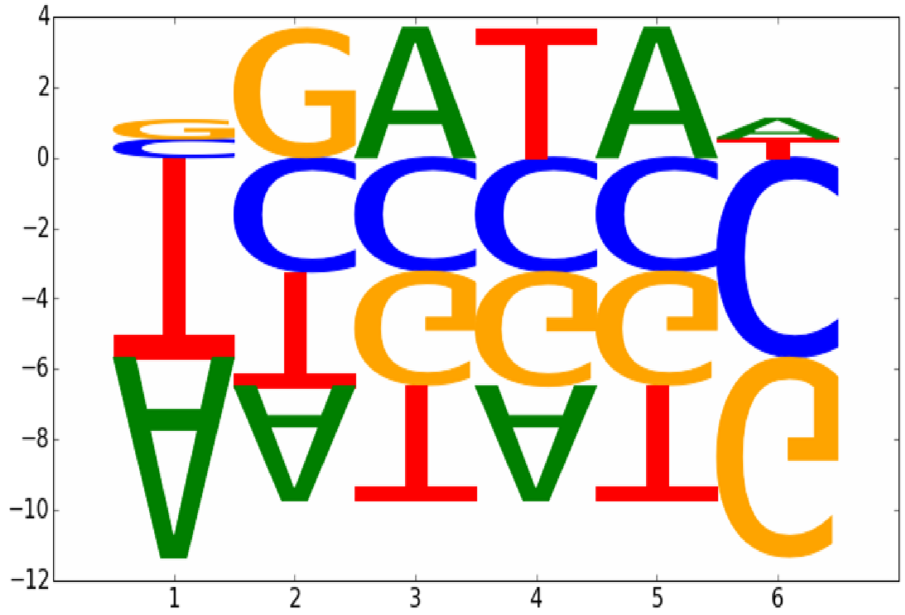
Sequence motifs: PSSM

Accounting for genomic background nucleotide distribution

Position-specific scoring matrix (PSSM)

$$\log_2\left(\frac{p_i(x_i = a_i)}{p_{bg}(x_i = a_i)}\right)$$

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**



PSSM logo

Scoring a sequence with a motif PSSM

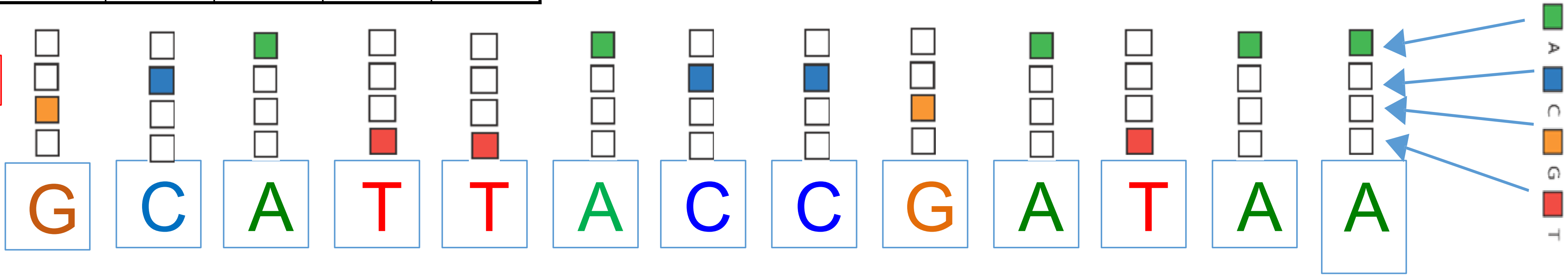
PSSM parameters

Scoring weights **W**

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding **(X)**

Input sequence



Convolution:
Scoring a sequence with a PSSM

Motif match Scores
sum(W * x)

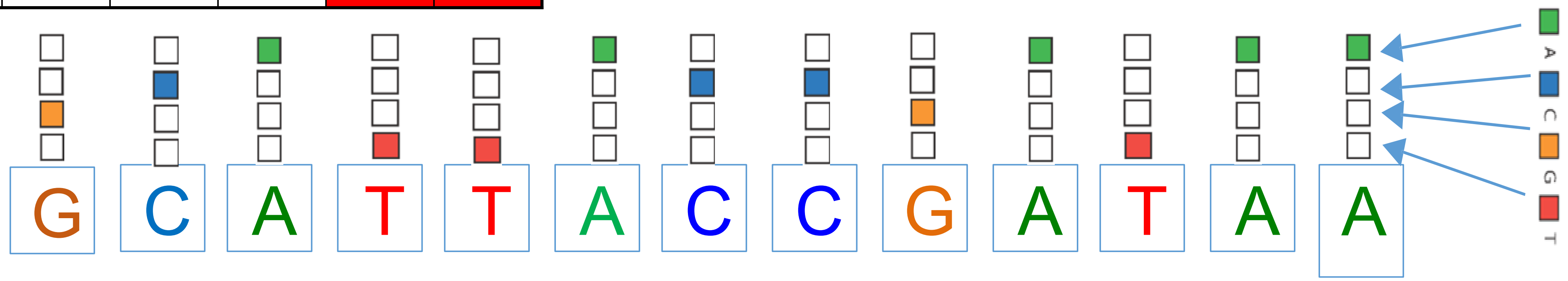
	-5.4											

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding (X)

Input sequence



Convolution

Motif match Scores
sum(W * x)

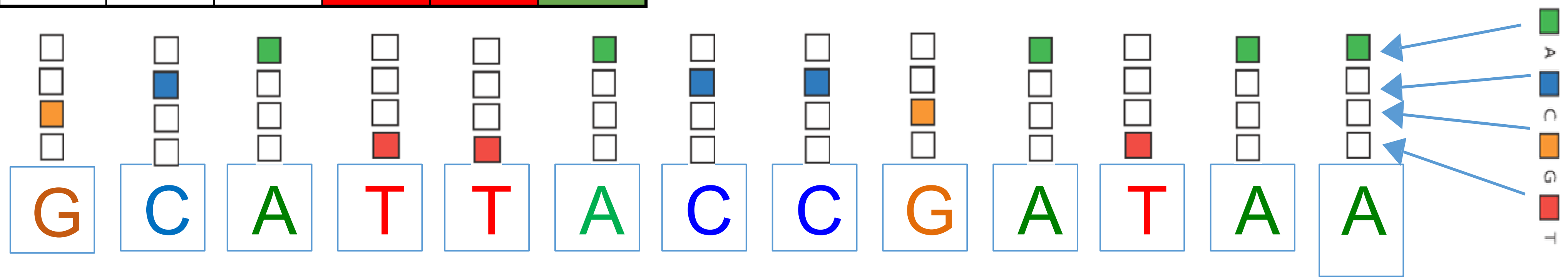
	-5.4	2.0										

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**

One-hot encoding (X)

Input sequence

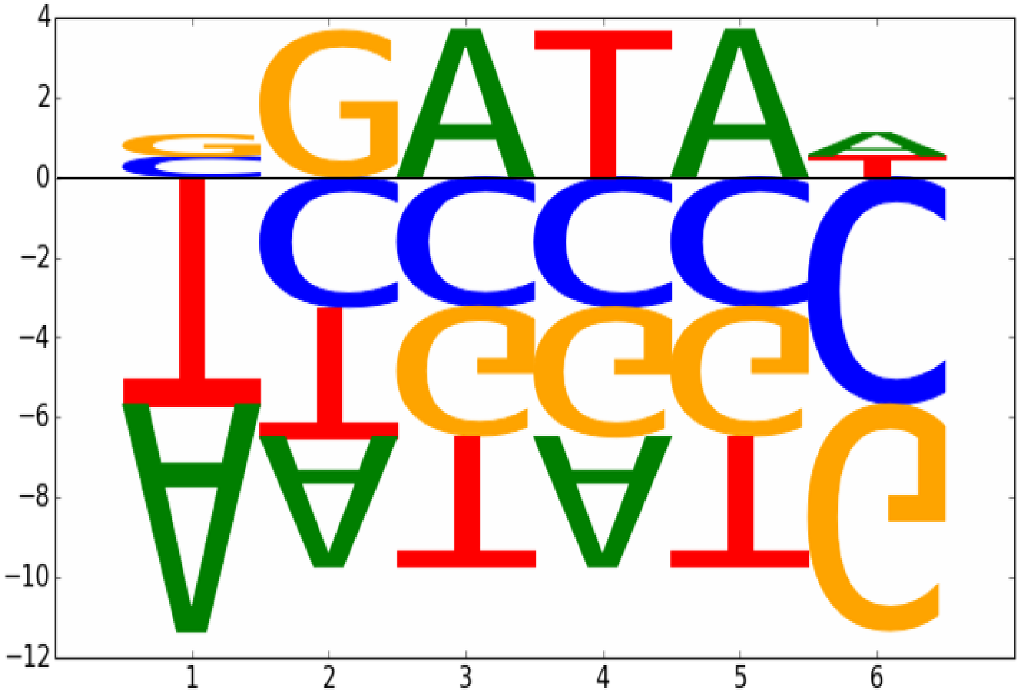


Convolution

Motif match Scores sum(W * x)													
	-2.2	-5.4	2.0	-4.3	-24	-17	-18	-11	-12	16	-5.5	-8.5	-5.2

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**



One-hot encoding (X)	□	□	🟩	□	□	🟩	□	□	□	🟩	□	🟩	🟩
	□	🟦	□	□	□	□	🟦	🟦	□	□	□	□	□
	🟧	□	□	□	□	□	□	□	🟧	□	□	□	□
	□	□	□	🟥	🟥	□	□	□	□	□	🟥	□	□
Input sequence	G	C	A	T	T	A	C	C	G	A	T	A	A

🟩 A
🟦 C
🟧 G
🟥 T

Thresholding scores

Thresholded Motif Scores max(0, W*x)

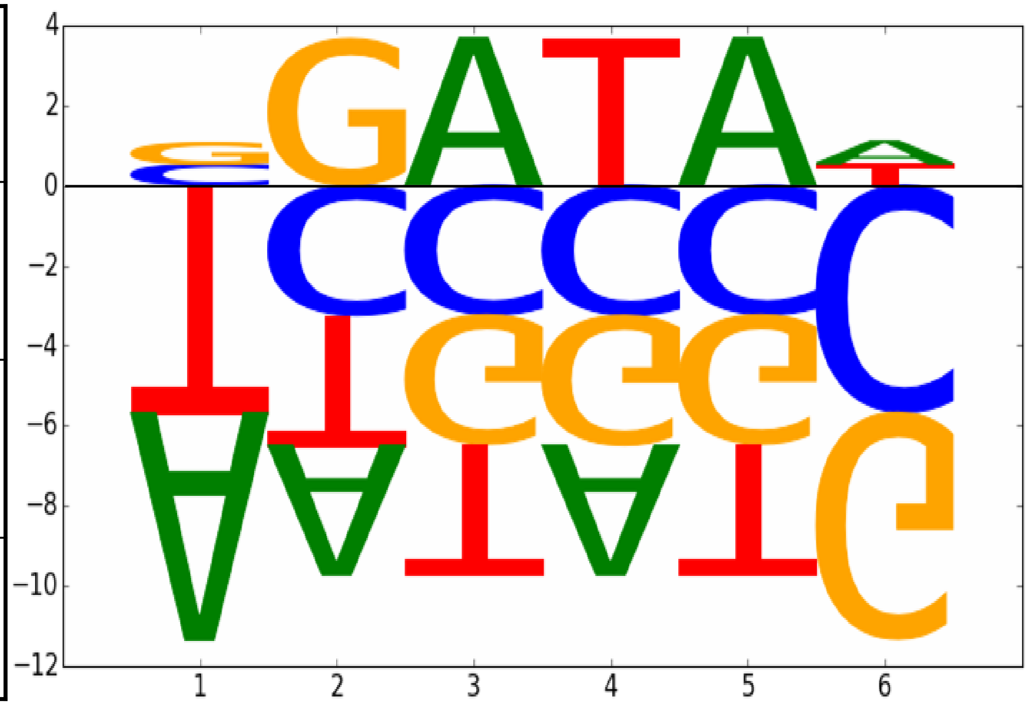
0	0	2.0	0	0	0	0	0	0	16	0	0	0

Motif match Scores W*x

-2.2	-5.4	2.0	-4.3	-24	-17	-18	-11	-12	16	-5.5	-8.5	-5.2

Scoring weights W

A	**-5.7**	**-3.2**	**3.7**	**-3.2**	**3.7**	**0.6**
C	**0.5**	**-3.2**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
G	**0.5**	**3.7**	**-3.2**	**-3.2**	**-3.2**	**-5.7**
T	**-5.7**	**-3.2**	**-3.2**	**3.7**	**-3.2**	**0.5**



One-hot encoding (X)

A	□	□	🟩	□	□	🟩	□	□	□	🟩	□	🟩	🟩
C	□	🟦	□	□	□	□	🟦	🟦	□	□	□	□	□
G	🟧	□	□	□	□	□	□	□	🟧	□	□	□	□
T	□	□	□	🟥	🟥	□	□	□	□	□	🟥	□	□

Input sequence

G	C	A	T	T	A	C	C	G	A	T	A	A

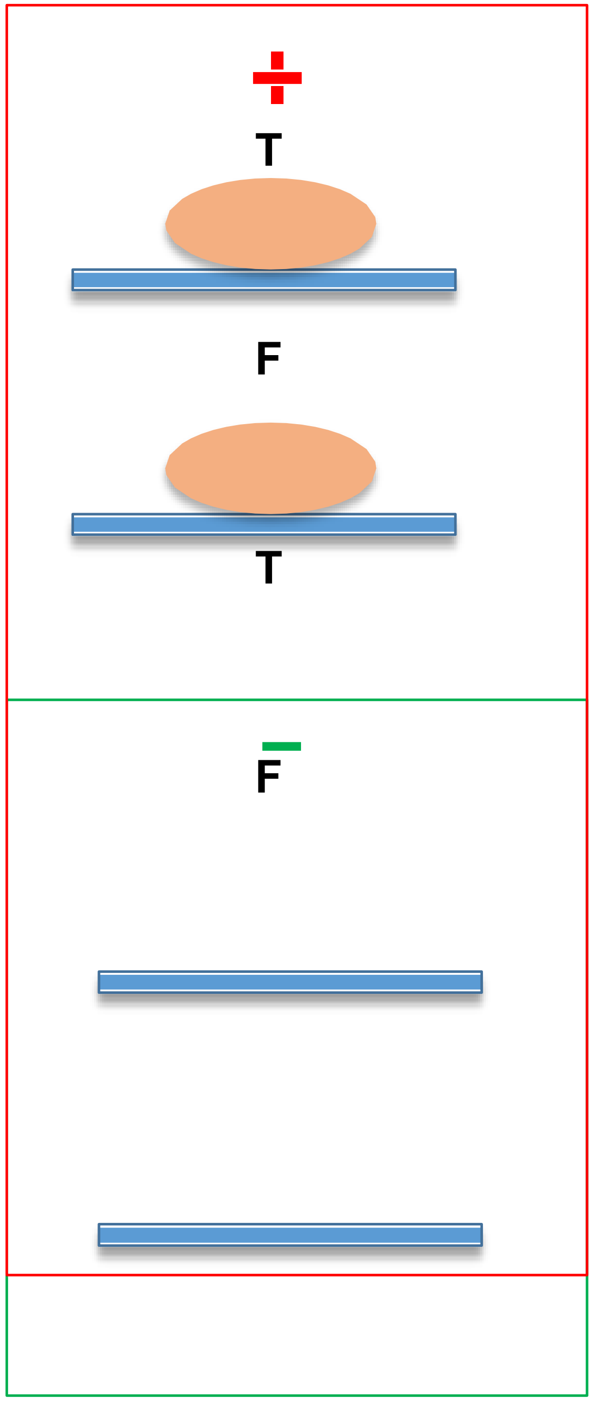
3b. CNNs for Regulatory Genomics Foundations (Higher-level learning)

Learning patterns in regulatory DNA sequence

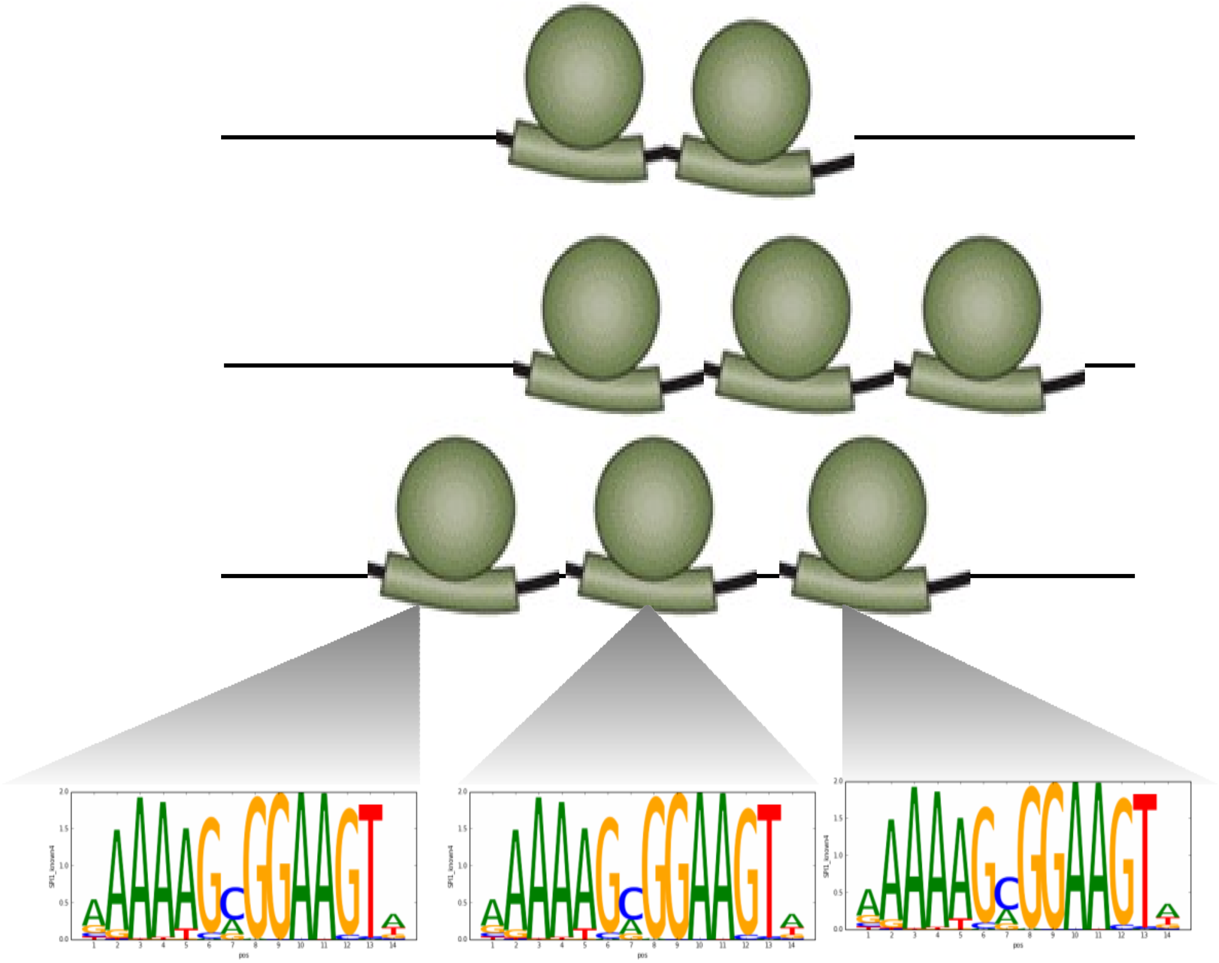
- Positive class of genomic sequences bound a transcription factor of interest

Can we learn patterns in the DNA sequence that distinguish these 2 classes of genomic sequences?

- Negative class of genomic sequences not bound by a transcription factor of interest



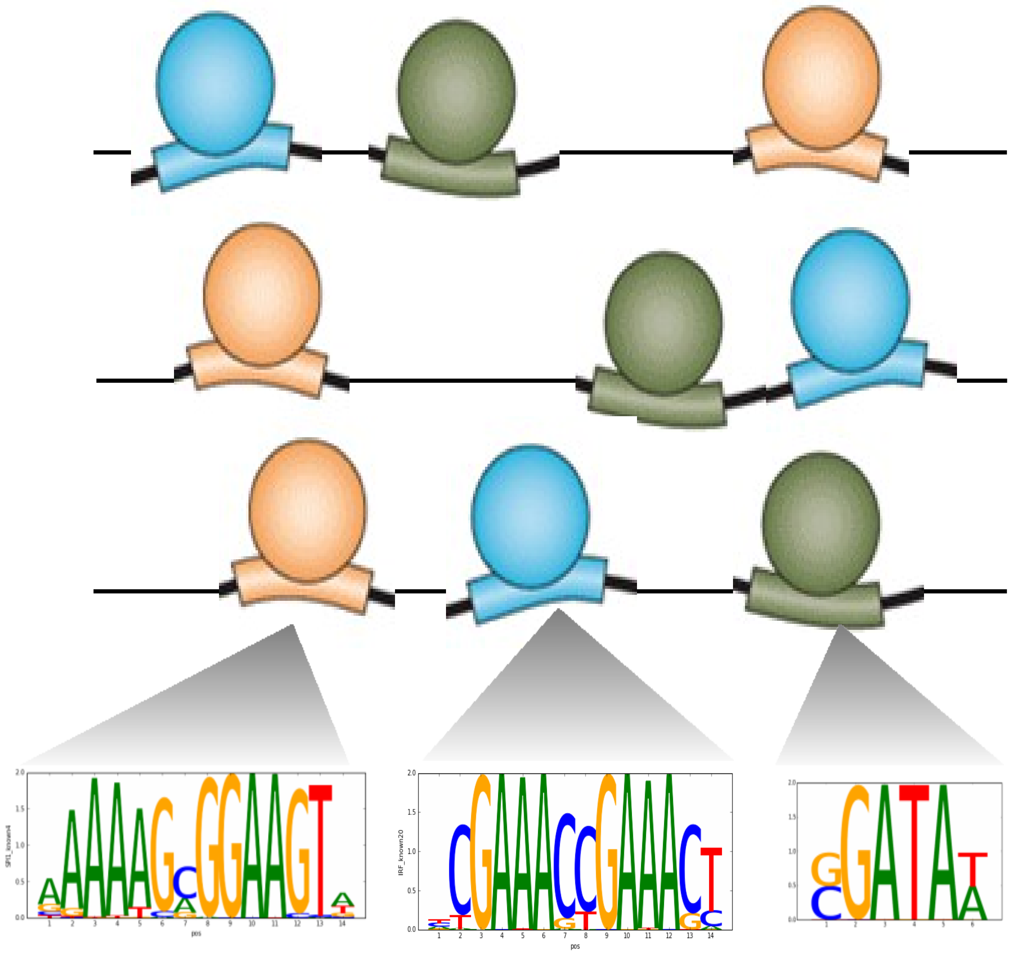
Key properties of regulatory sequence



HOMOTYPIC MOTIF DENSITY

Regulatory sequences often contain **more than one binding instance** of a TF resulting in **homotypic clusters of motifs of the same TF**

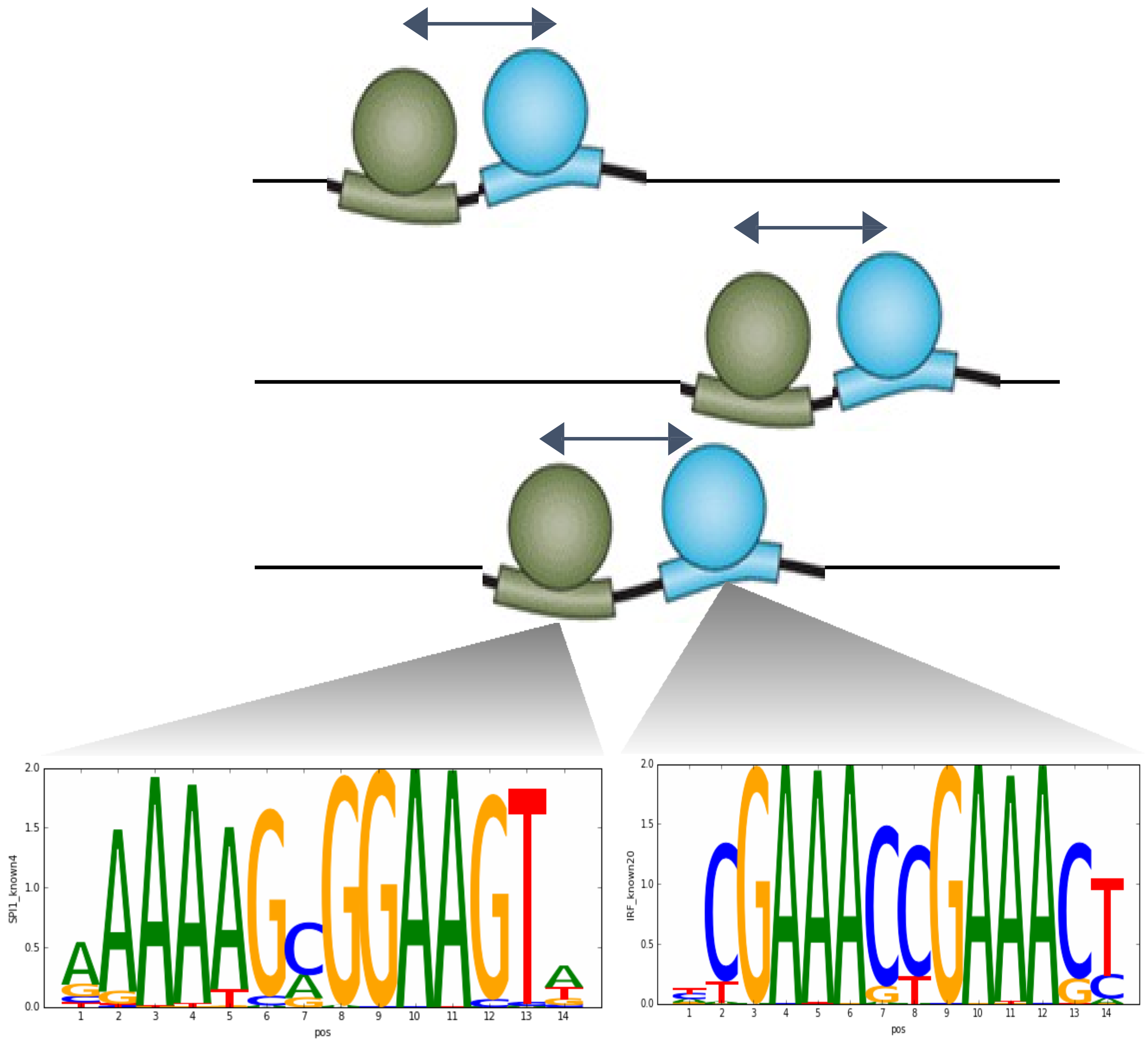
Key properties of regulatory sequence



HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences often bound by **combinations of TFs** resulting in **heterotypic clusters of motifs of different TFs**

Key properties of regulatory sequence



SPATIAL GRAMMARS OF HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences are often bound by **combinations of TFs** with specific **spatial and** **positional constraints** resulting in distinct **motif grammars**

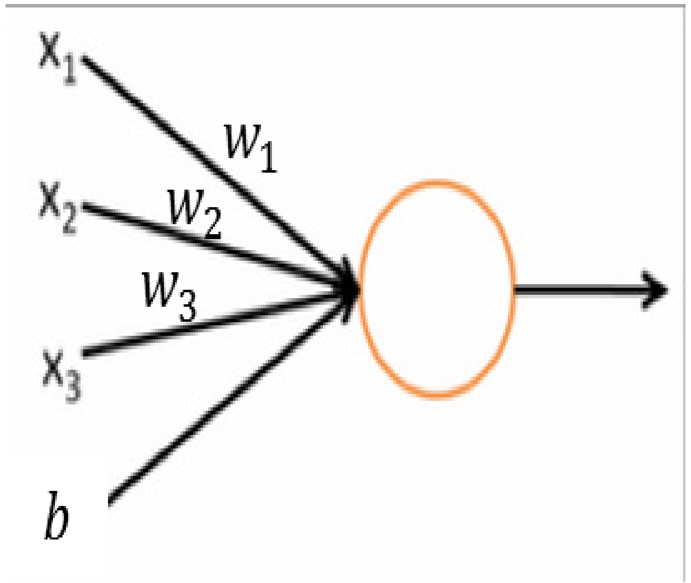
A simple classifier (An artificial neuron)

$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1 . x_1 + w_2 . x_2 + w_3 . x_3 + b$$

Linear function



Z

Training the neuron means learning the optimal w's and b

A simple classifier (An artificial neuron)

Logistic / Sigmoid

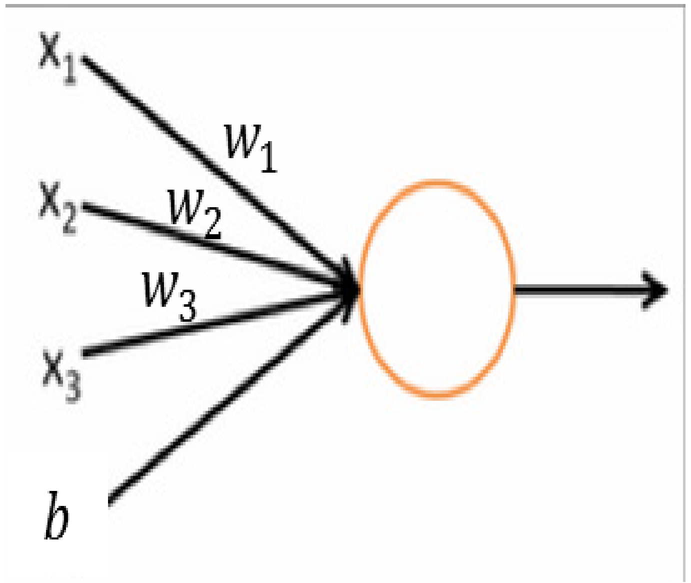
Useful for predicting probabilitie

$$Y = F(x_1, x_2, x_3)$$

parameters

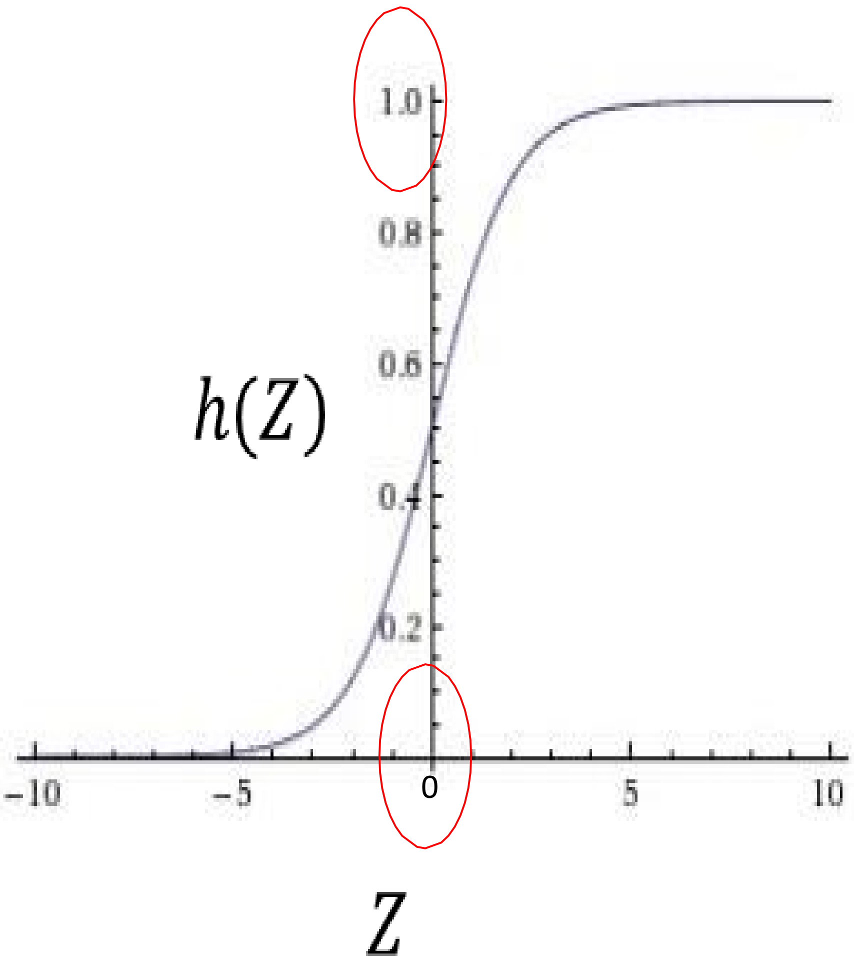
$$Z = w_1.x_1 + w_2.x_2 + w_3.x_3 + b$$

$$Y = h(Z)$$



Y

Non-linear function



Training the neuron means learning the optimal w’s and b

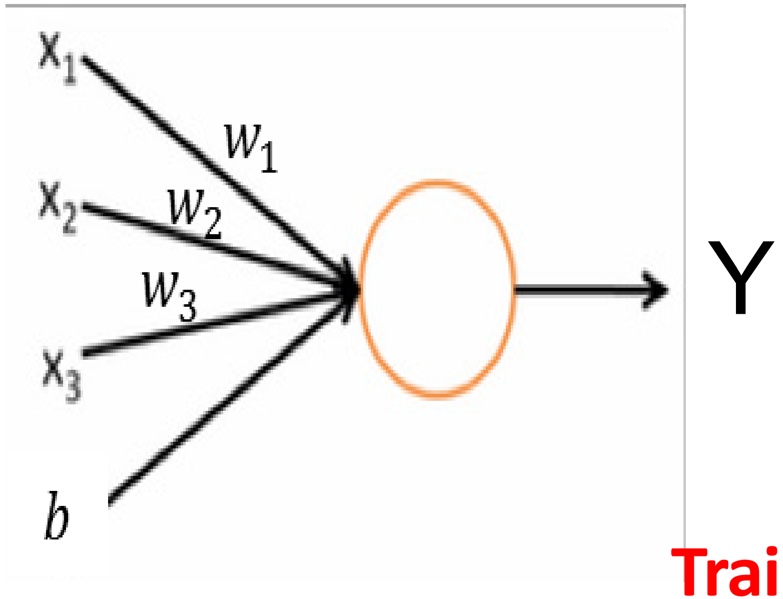
A simple classifier (An artificial neuron)

$$Y = F(x_1, x_2, x_3)$$

parameters

$$Z = w_1.x_1 + w_2.x_2 + w_3.x_3 + b$$

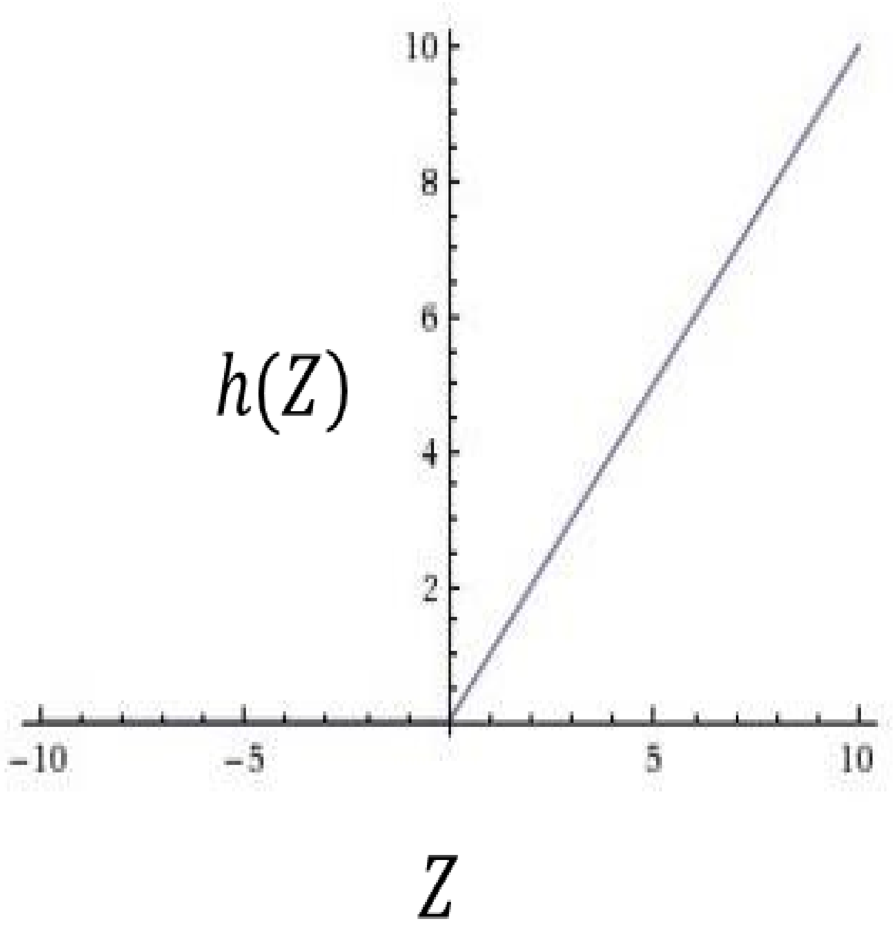
$$Y = h(Z)$$



ReLu (Rectified Linear Unit)

Useful for thresholding

Non-linear function



Training the neuron means learning the optimal w’s and b

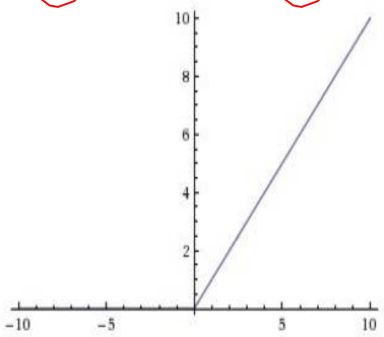
Artificial neuron can represent a motif

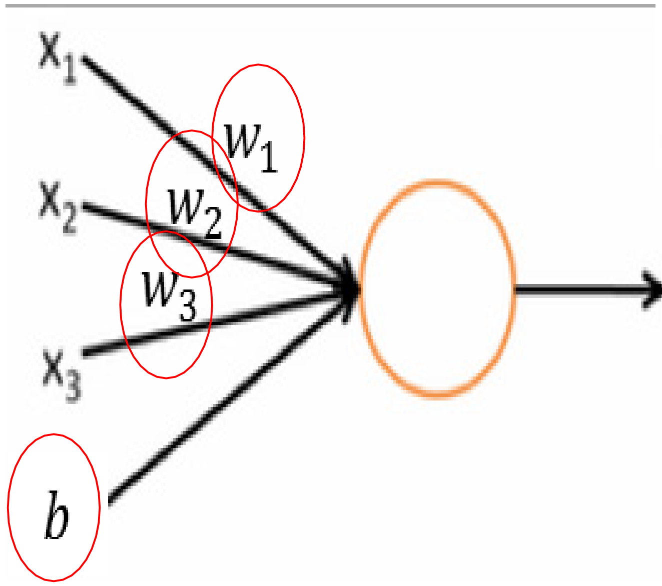
$$Y = F(x_1, x_2, x_3)$$

parameters

$$\boxed{Z} = w_1 . x_1 + w_2 . x_2 + w_3 . x_3 + b$$

$$Y = \boxed{h(Z)}$$





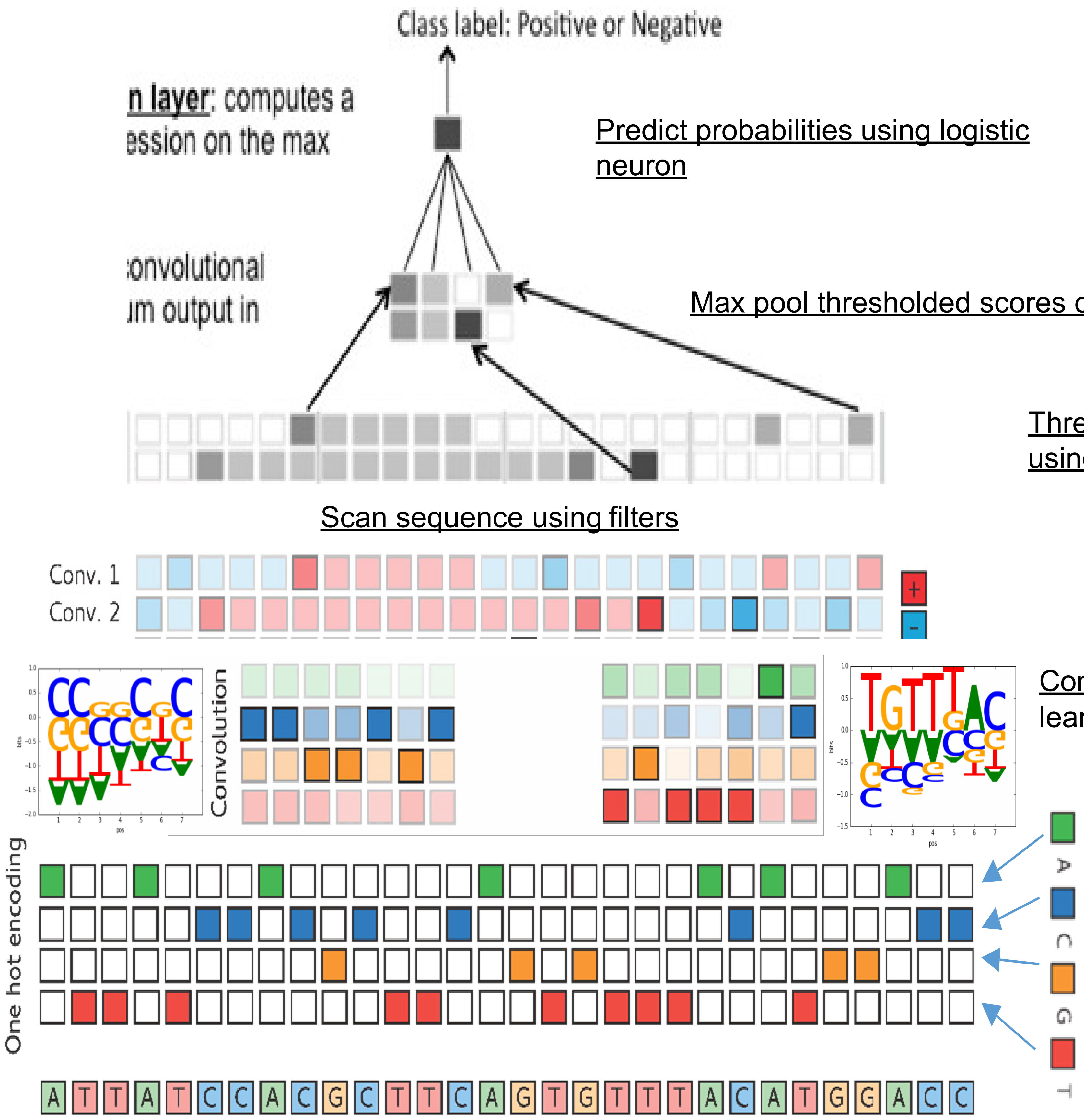
Thresholded Motif Scores max(0, W*x)	0	0	2.0	0	0	0
Motif match Scores sum(W * x)	-2.2	-5.5	2.0	-4.3	-24	-17

Scoring weights W

A	-5.7	-3.2	3.7	-3.2	3.7	0.6
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding (X)	□ □ 🟧 □	□ 🟦 □ □	🟩 □ □ □	□ □ □ 🟥	□ □ □ 🟥	🟩 □ □ □
Input sequence	G	C	A	T	T	A

Biological motivation of Deep CNN



Predict probabilities using logistic neuron

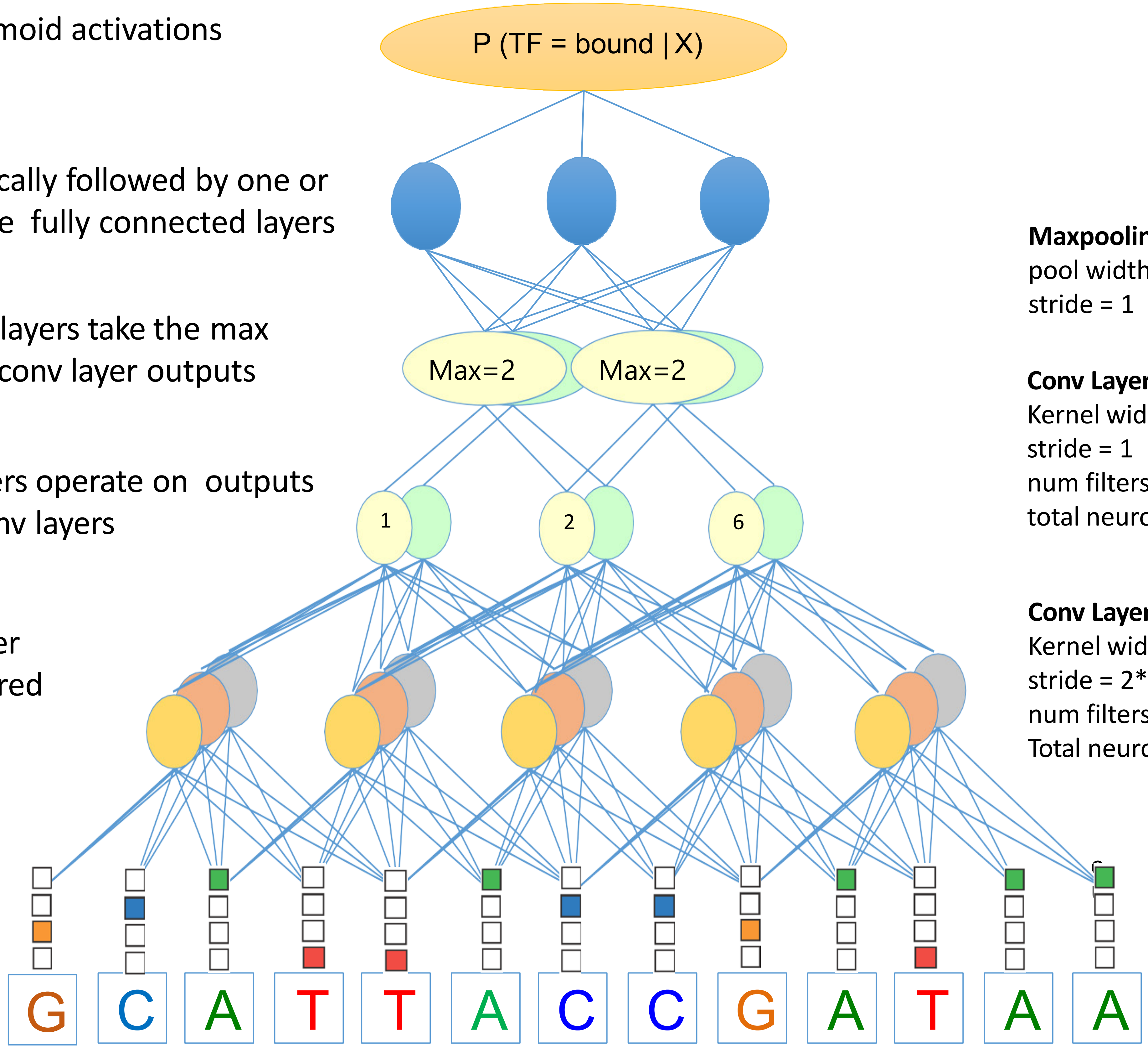
Max pool thresholded scores over windows

Threshold scores using ReLU

Scan sequence using filters

Convolutional filters learn motifs (PSSM)

Deep convolutional neural network



Sigmoid activations

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer (same color = shared weights)

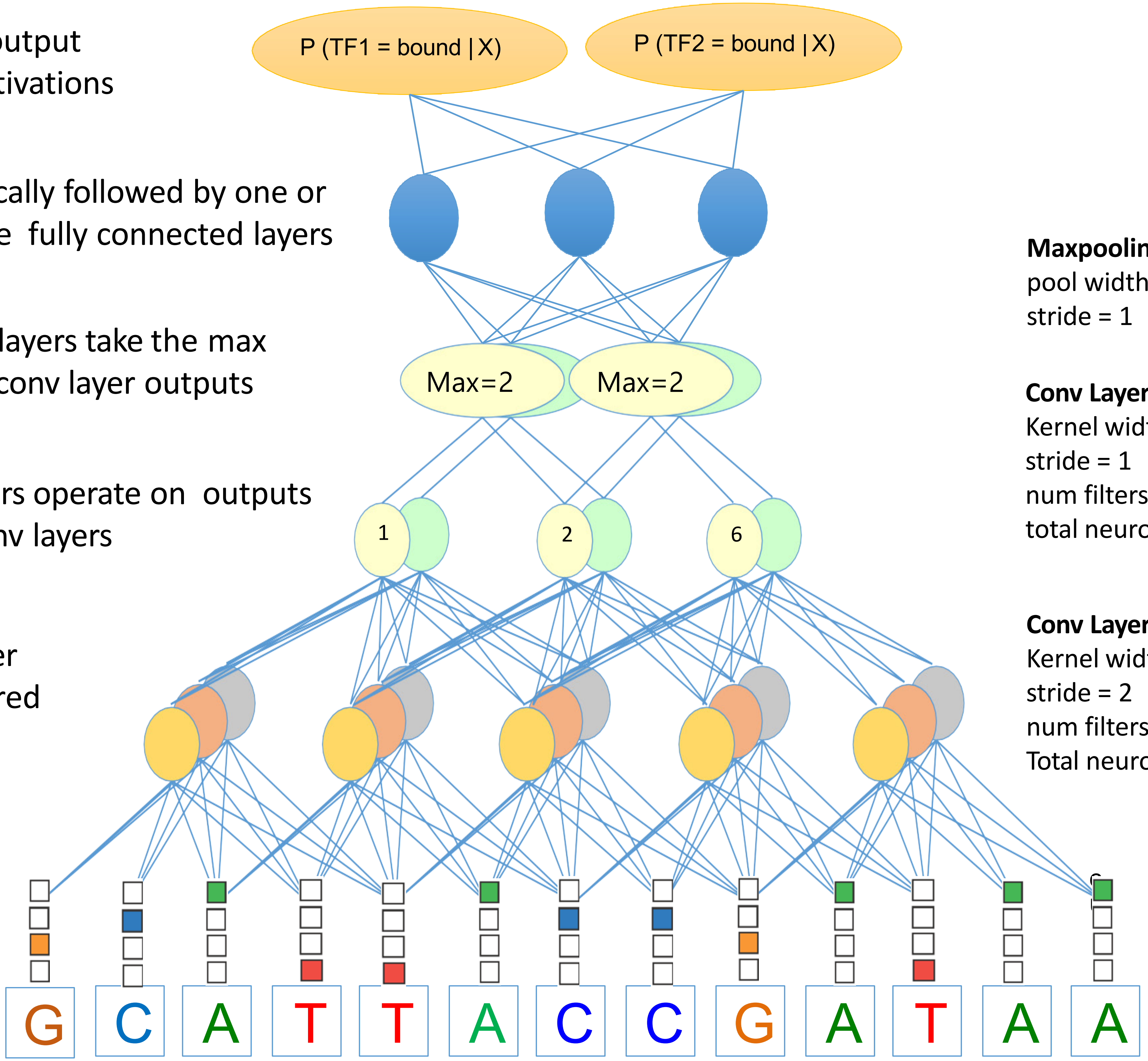
Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num channels = 2
total neurons = 6

Conv Layer 1
Kernel width = 4
stride = 2*
num filters / num channels = 3
Total neurons = 15

***for genomics, a stride of 1 for conv layers is recommended**

Multi-task CNN



Multi-task output (sigmoid activations here)

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer (same color = shared weights)

Maxpooling layer
pool width = 2
stride = 1

Conv Layer 2
Kernel width = 3
stride = 1
num filters / num channels = 2
total neurons = 6

Conv Layer 1
Kernel width = 4
stride = 2
num filters / num channels = 3
Total neurons = 15

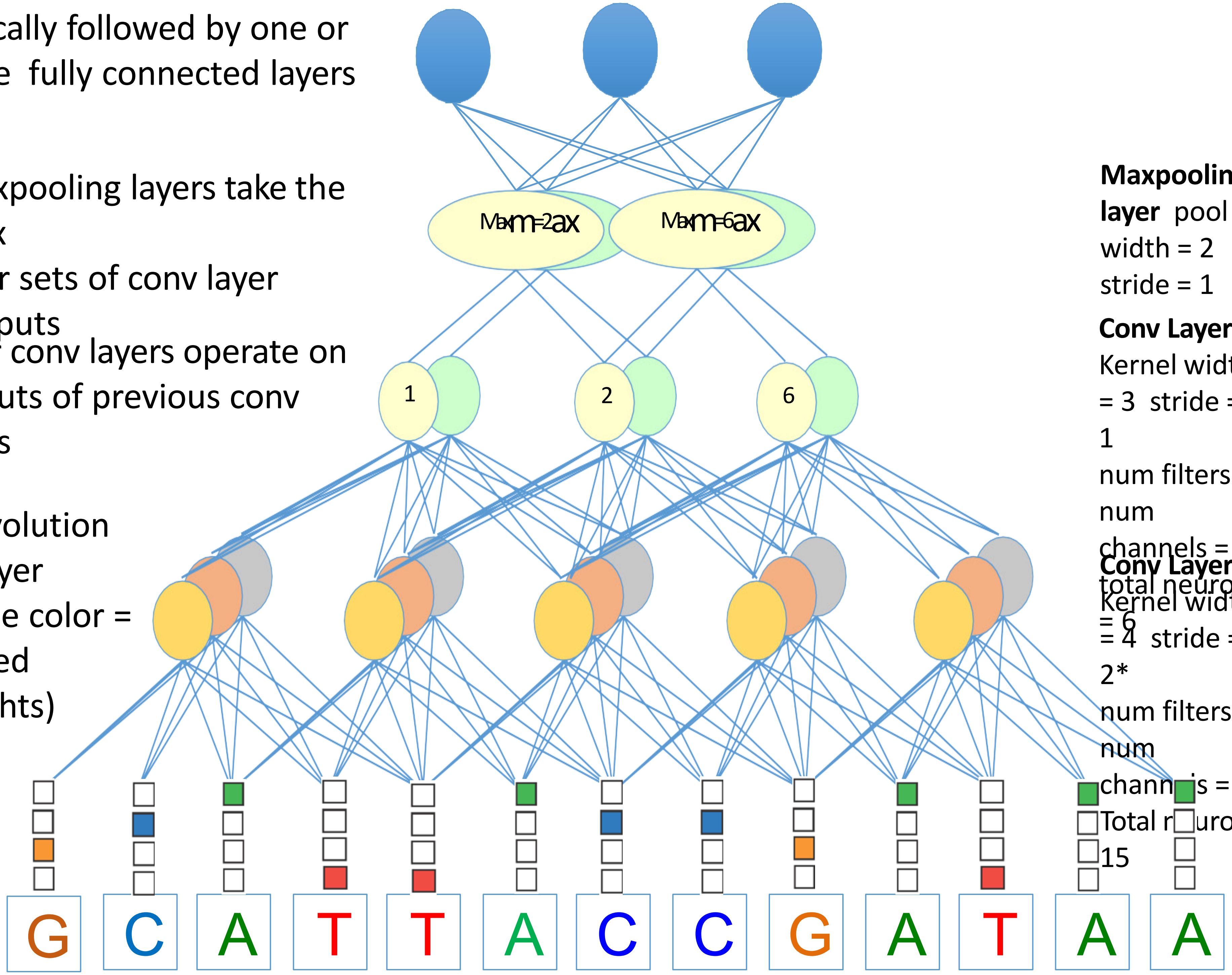
Multi-task CNN

Typically followed by one or more fully connected layers

Maxpooling layers take the max over sets of conv layer outputs

Later conv layers operate on outputs of previous conv layers

Convolutional layer (same color = shared weights)



Maxpooling layer pool width = 2 stride = 1

Conv Layer 2 Kernel width = 3 stride = 1
num filters / num channels = 2
total neurons = 6

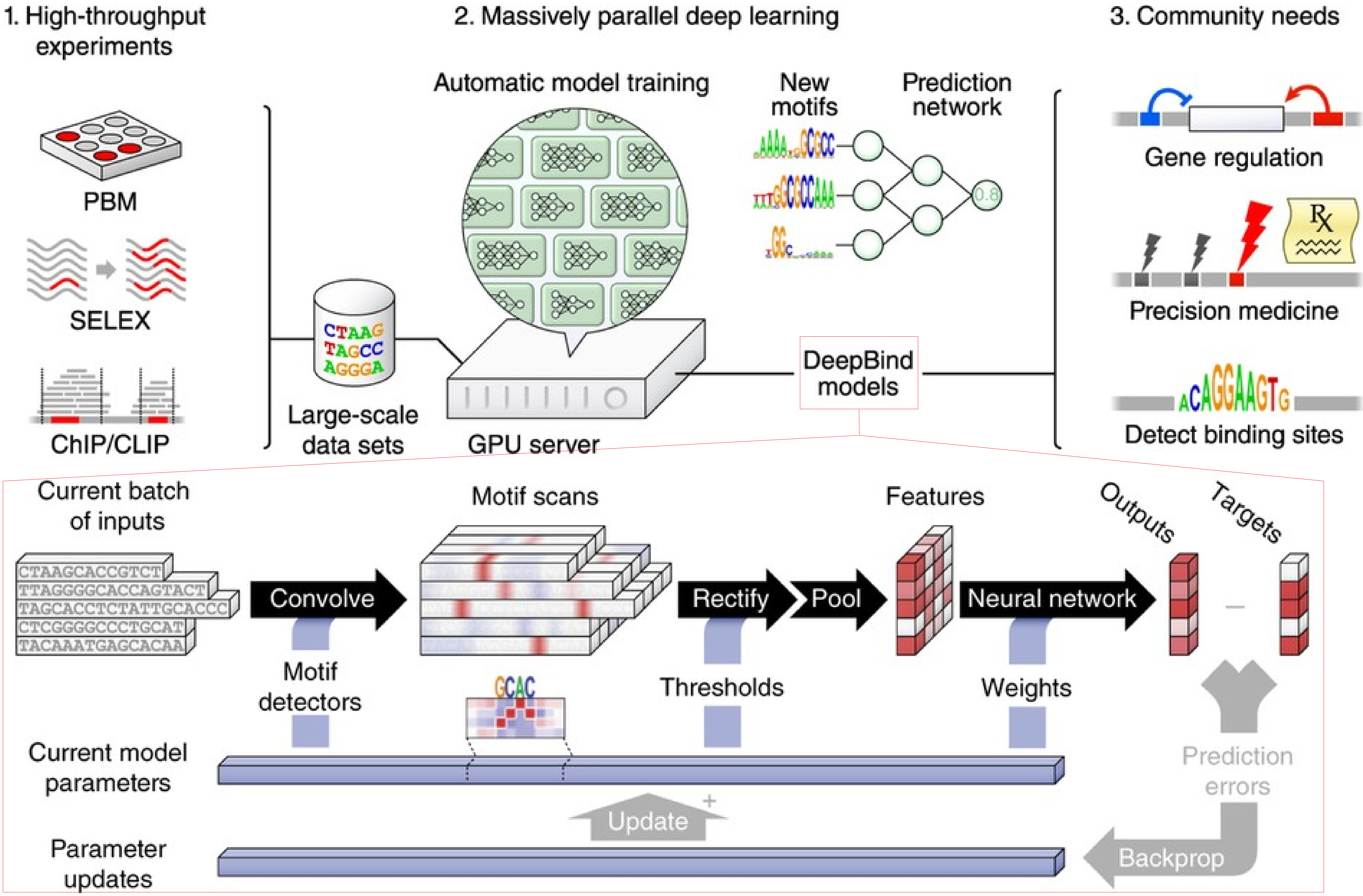
Conv Layer 1 Kernel width = 4 stride = 2*
num filters / num channels = 3
Total neurons = 15

Deep Learning for Regulatory Genomics

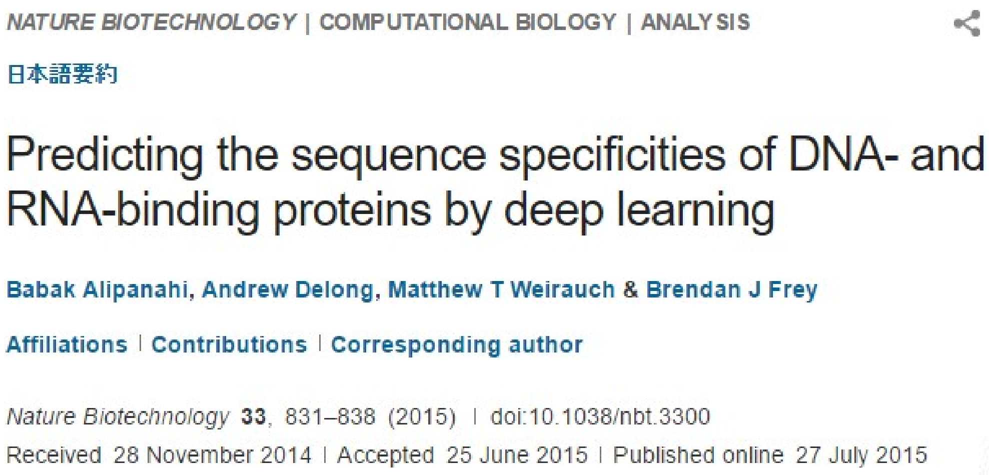
1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⇔ DNA letters. Patches/filters ⇔ Motifs. Higher ⇔ combinations
   - Learning convolutional filters ⇔ Motif discovery. Applying them ⇔ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

4. Regulatory Genomics CNNs in Practice:
(a) DeepBind

DeepBind



[Alipanahi et al., 2015]



NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

日本語要約

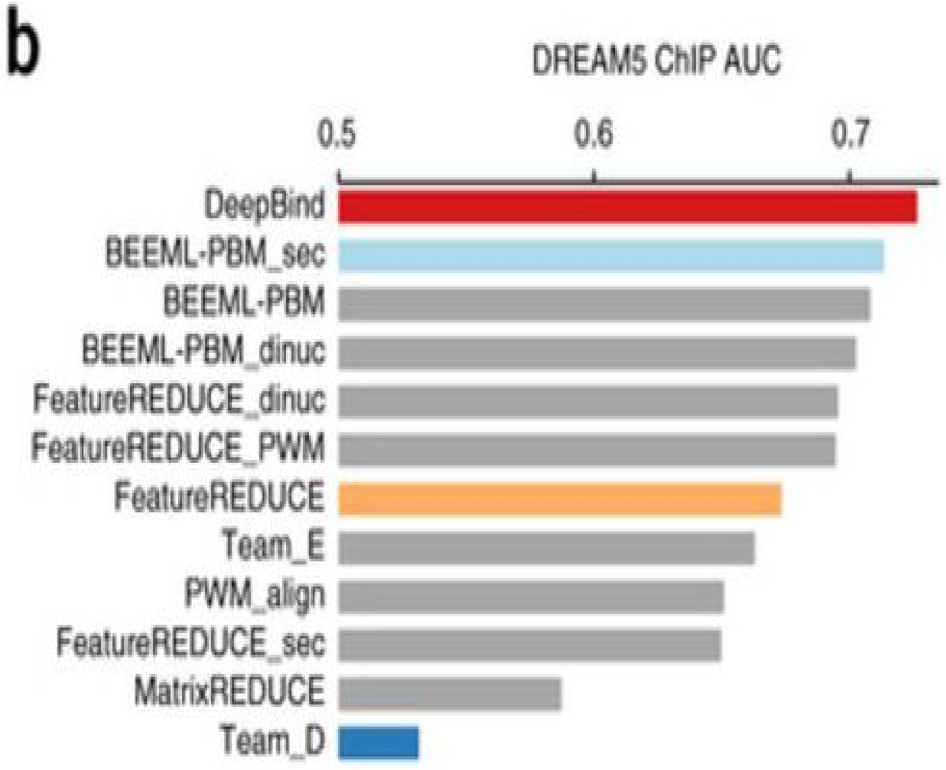
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

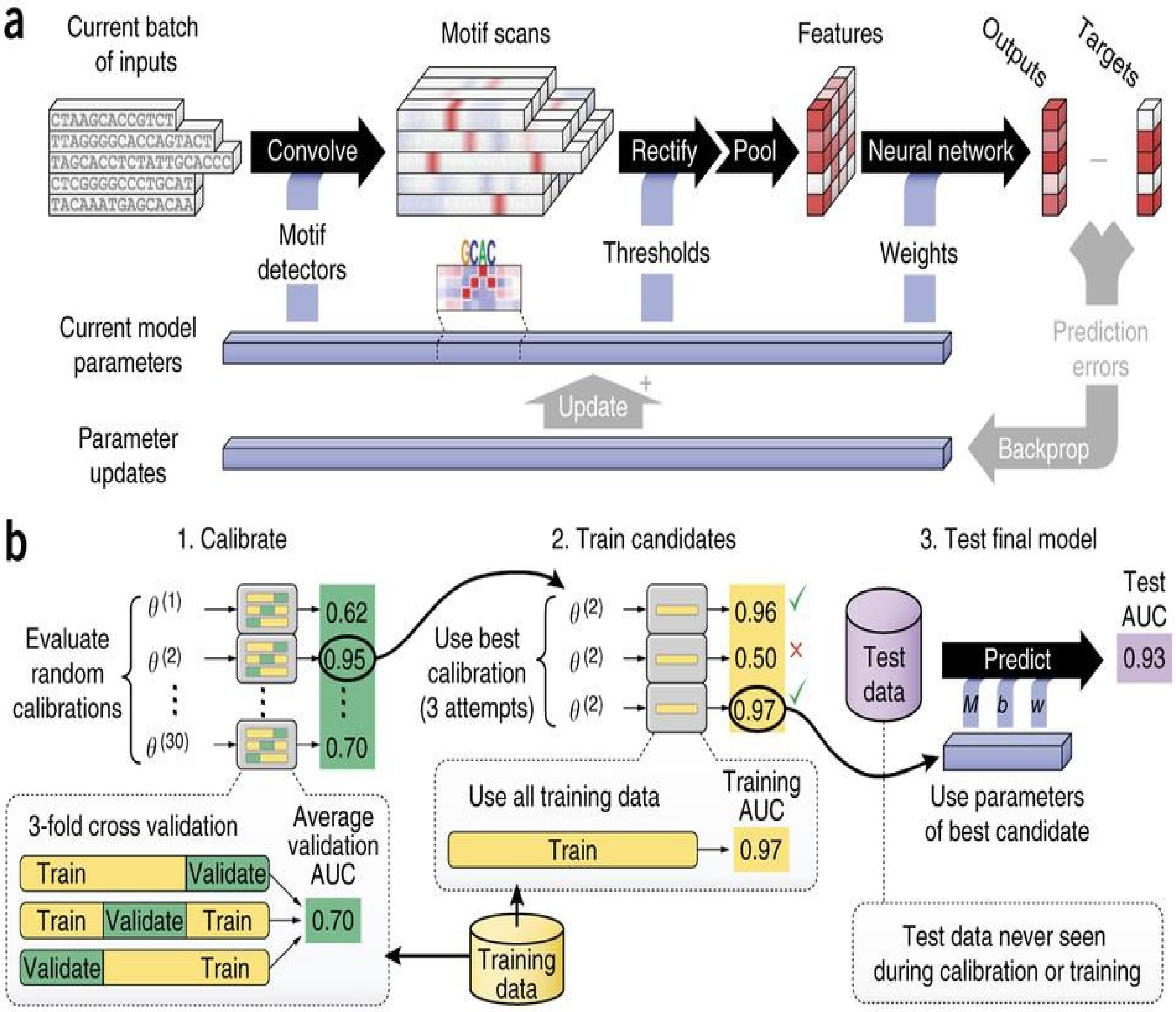
Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Affiliations | Contributions | Corresponding author

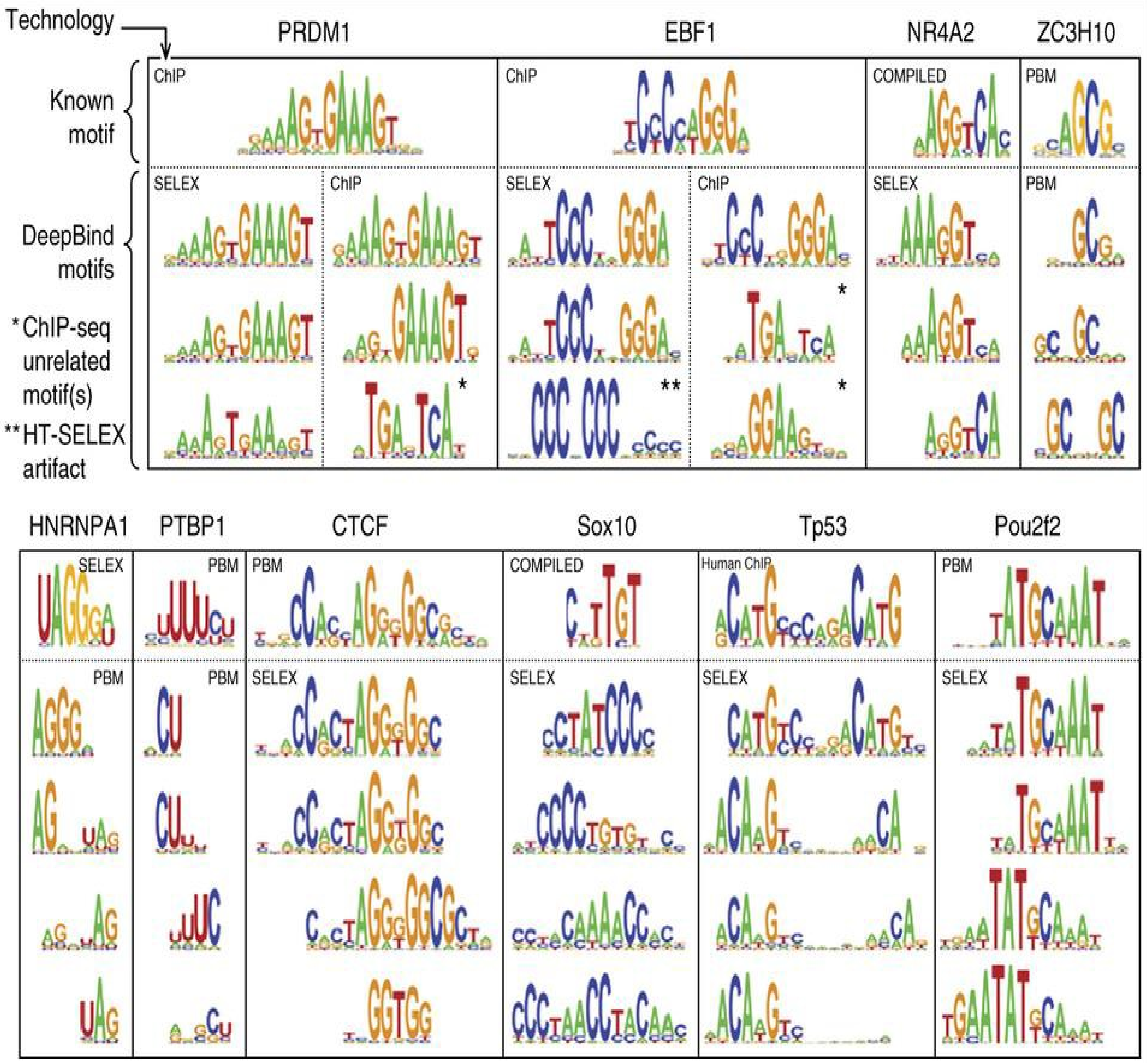
Nature Biotechnology **33**, 831–838 (2015) | doi:10.1038/nbt.3300

Received 28 November 2014 | Accepted 25 June 2015 | Published online 27 July 2015

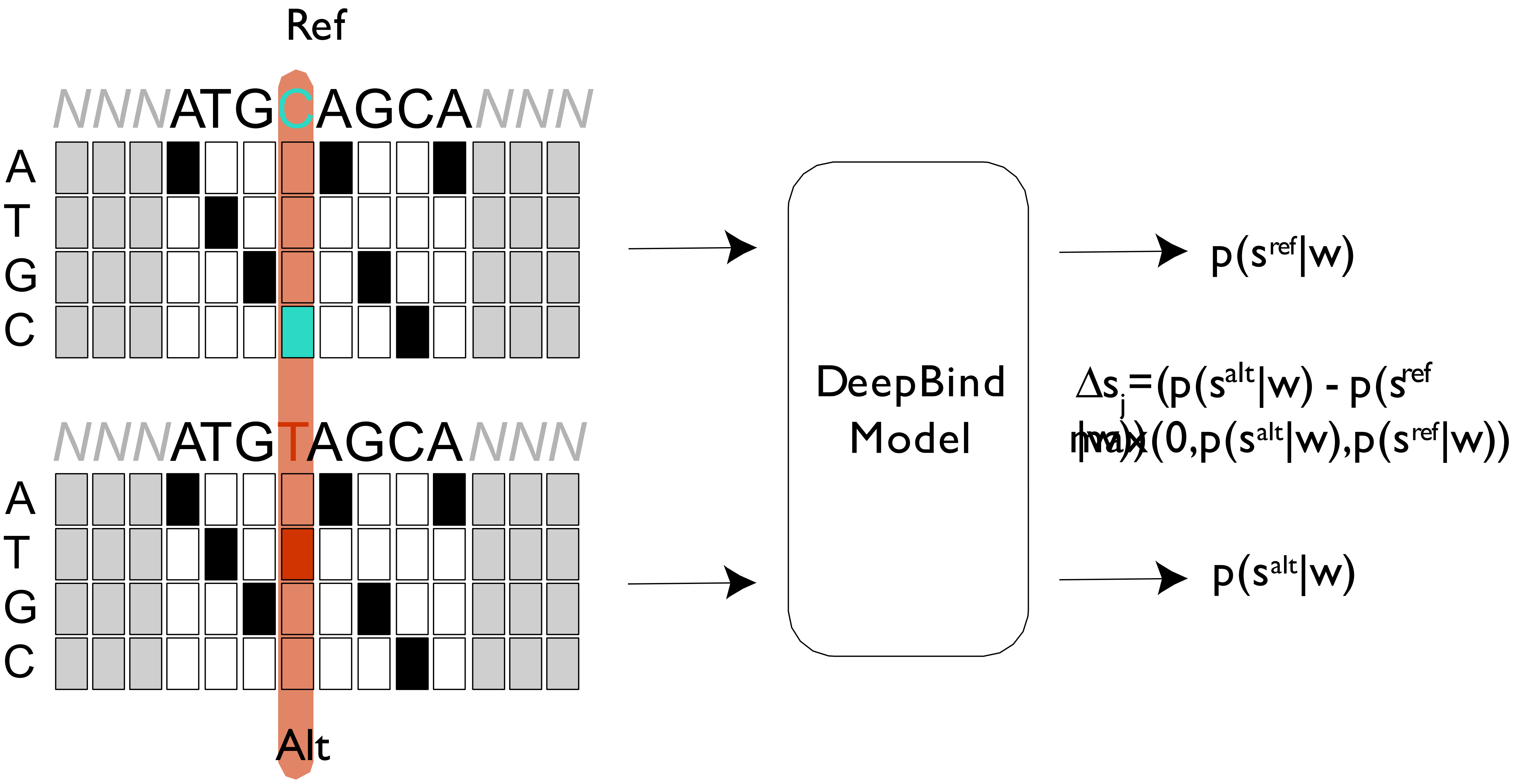




http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html

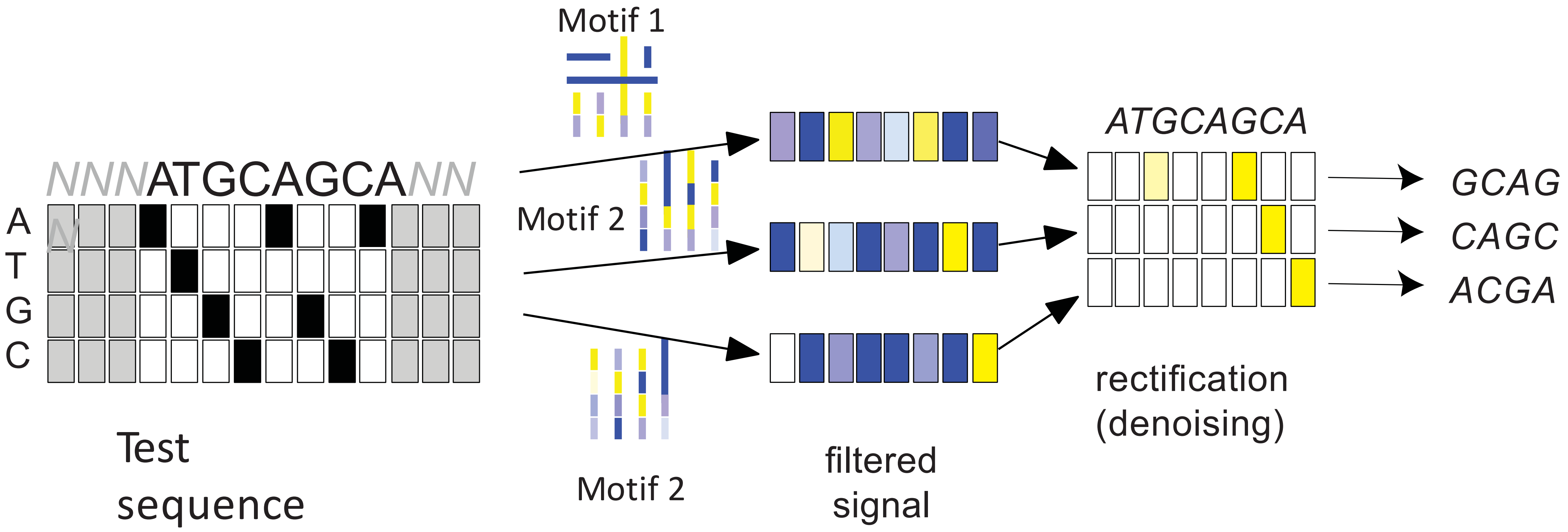

Technology
PRDM1
EBF1
NR4A2
ZC3H10
Known motif
DeepBind motifs
* ChIP-seq unrelated motif(s)
** HT-SELEX artifact
ChIP
ChIP
COMPILED
PBM
SELEX
ChIP
SELEX
ChIP
SELEX
PBM
*
**
*
*
HNRNPA1
PTBP1
CTCF
Sox10
Tp53
Pou2f2
SELEX
PBM
PBM
COMPILED
Human ChIP
PBM
PBM
PBM
SELEX
SELEX
SELEX
SELEX

Constructing mutation map



$\Delta s_j = (p(s^{alt}|w) - p(s^{ref}|w)) \max(0, p(s^{alt}|w), p(s^{ref}|w))$

Constructing sequence logo



Motif 1

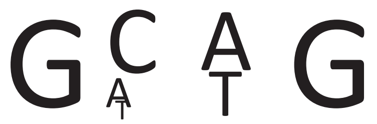
GCAG
GCTG
GATG
.
.
.
GTAG
GCTG

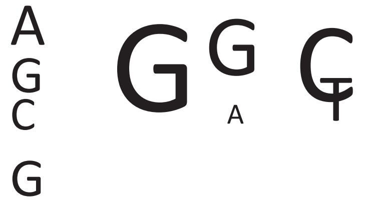
Motif 2

CAGC
GGTC
AGTC
.
.
.
AGGC
GGTG

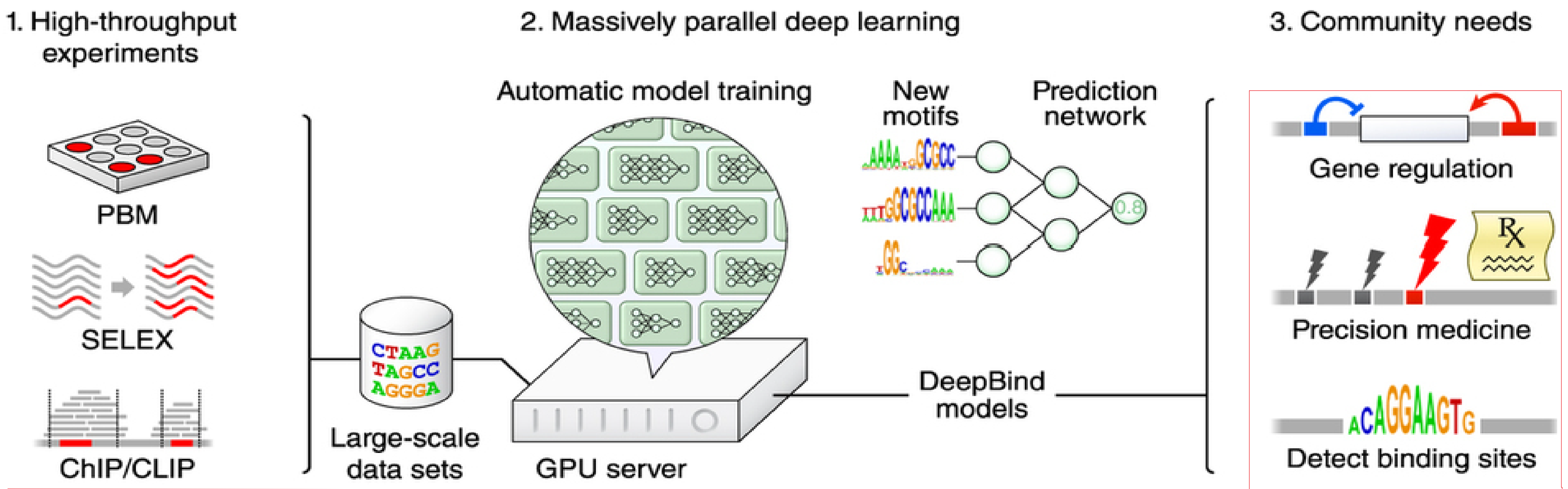
...

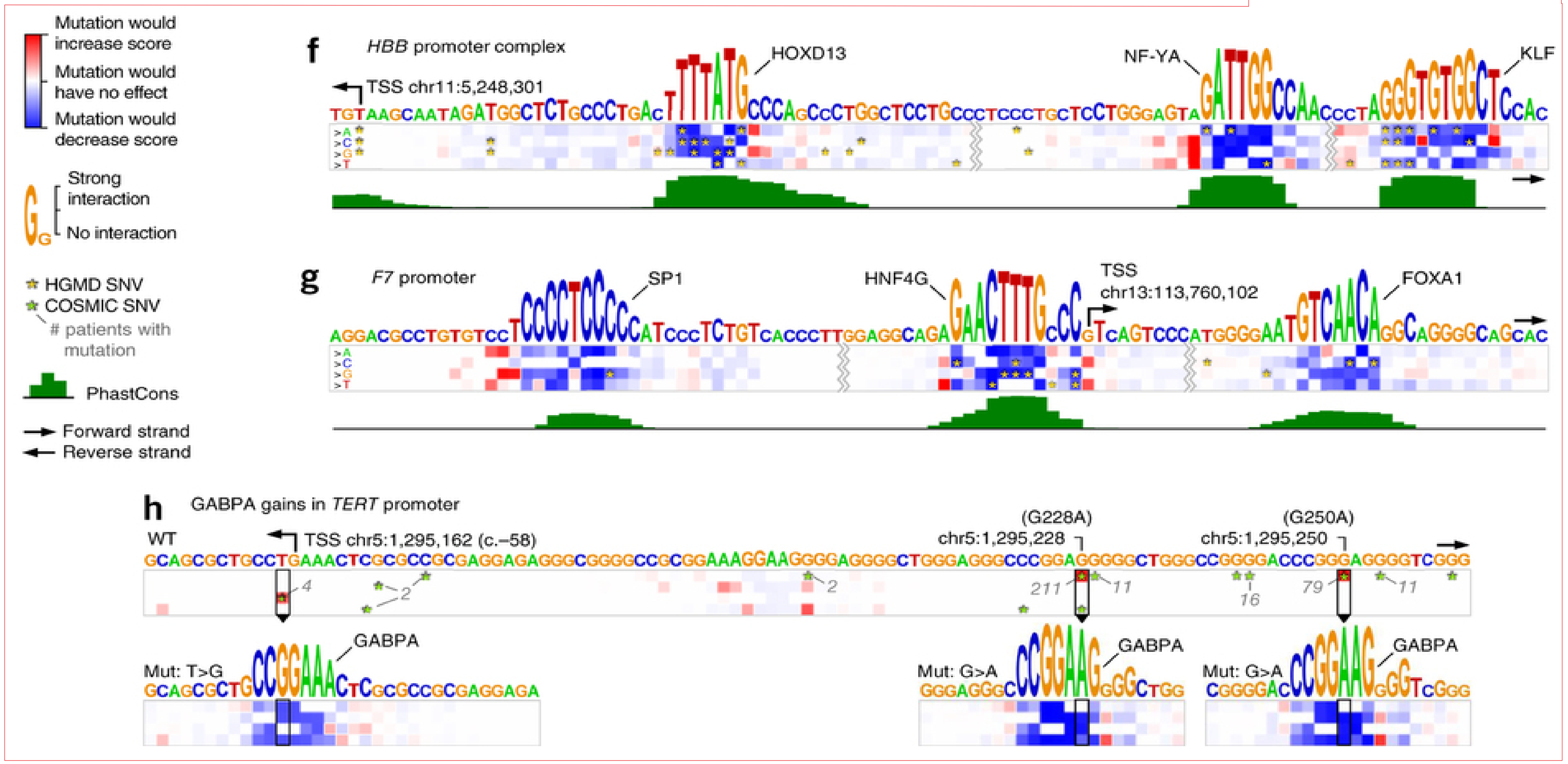
PFM





Predicting disease mutations





[Alipanahi et al., 2015]

DeepBind summary

The key deep learning techniques:

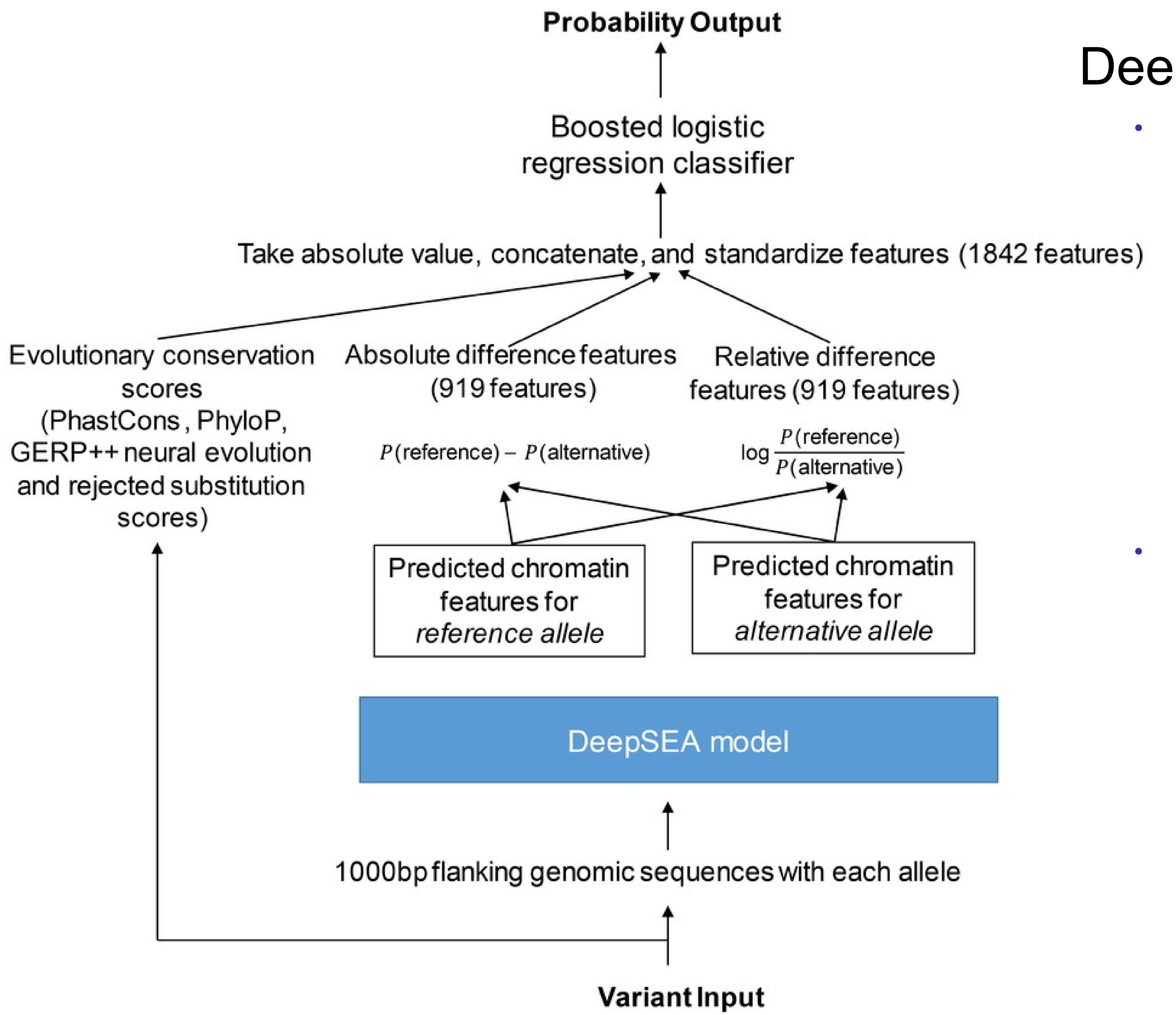
- Convolutional learning
- Representational learning
- Back-propagation and stochastic gradient
- Regularization and dropout
- Parallel GPU computing especially useful for hyperparameter search

Limitations in DeepBind:

- Require defining negative training examples, which is often arbitrary
- Using observed mutation data only as post-hoc evaluation
- Modeling each regulatory dataset separately

Regulatory Genomics CNNs in Practice:
(b) DeepSEA

DeepSea


Probability Output
Boosted logistic regression classifier
Take absolute value, concatenate, and standardize features (1842 features)
Evolutionary conservation scores (PhastCons, PhyloP, GERP++ neural evolution and rejected substitution scores)
Absolute difference features (919 features)
Relative difference features (919 features)
$P(\text{reference}) - P(\text{alternative})$
$\log \frac{P(\text{reference})}{P(\text{alternative})}$
Predicted chromatin features for *reference allele*
Predicted chromatin features for *alternative allele*
DeepSEA model
1000bp flanking genomic sequences with each allele
Variant Input

DeepSea:

- Similar as DeepBind but trained a separate CNN on each of the ENCODE/Roadmap Epigenomic chromatin profiles 919 chromatin features (125 DNase features, 690 TF features, 104 histone features).
- It uses the Δs mutation score as input to train a linear logistic regression to predict GWAS and eQTL SNPs defined from the GRASP database with a P-value cutoff of 1E-10 and GWAS SNPs from the NHGRI GWAS Catalog

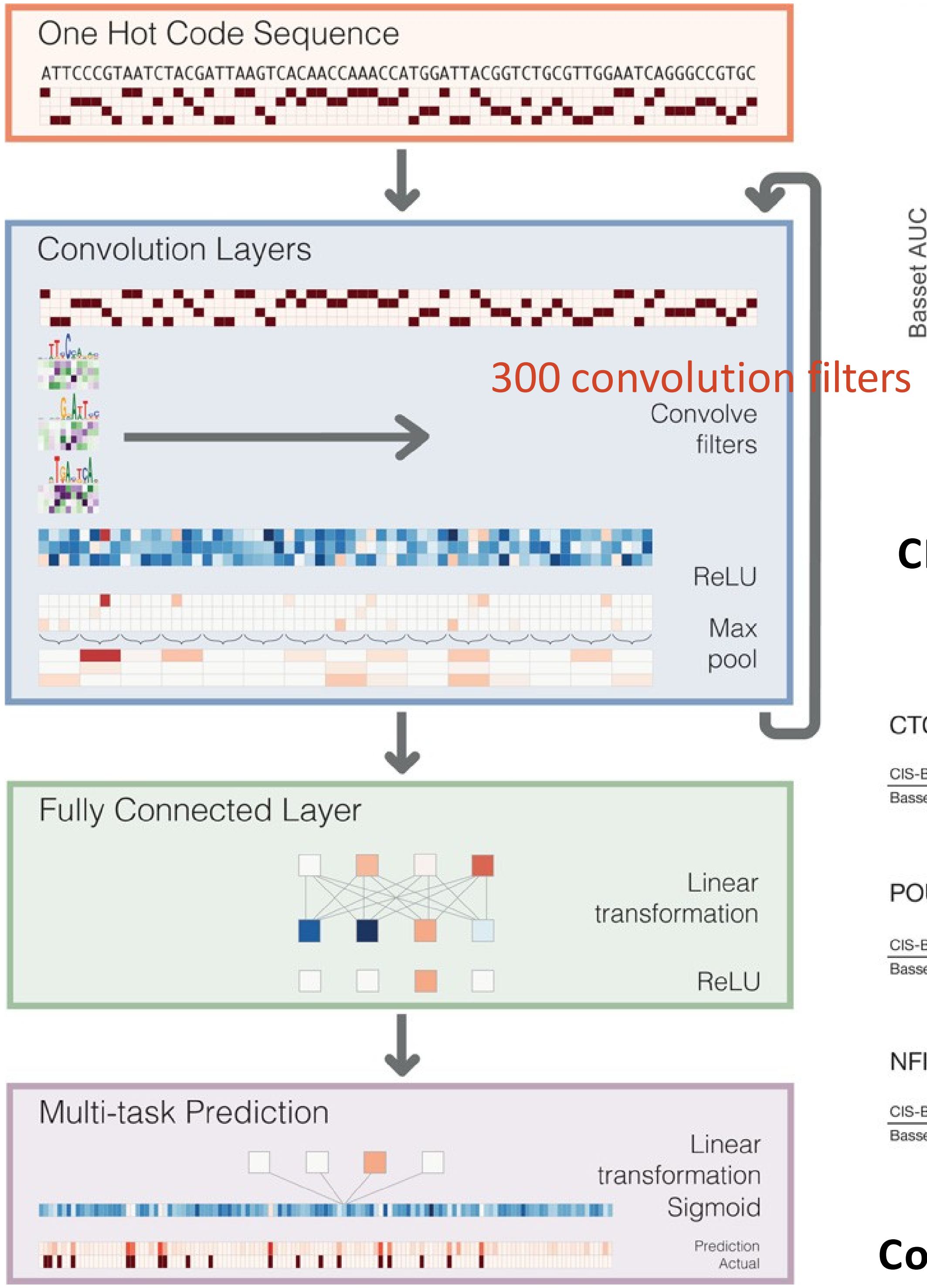
[Zhou and Troyanskaya, 2015]

Regulatory Genomics CNNs in Practice: (c) Basset

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

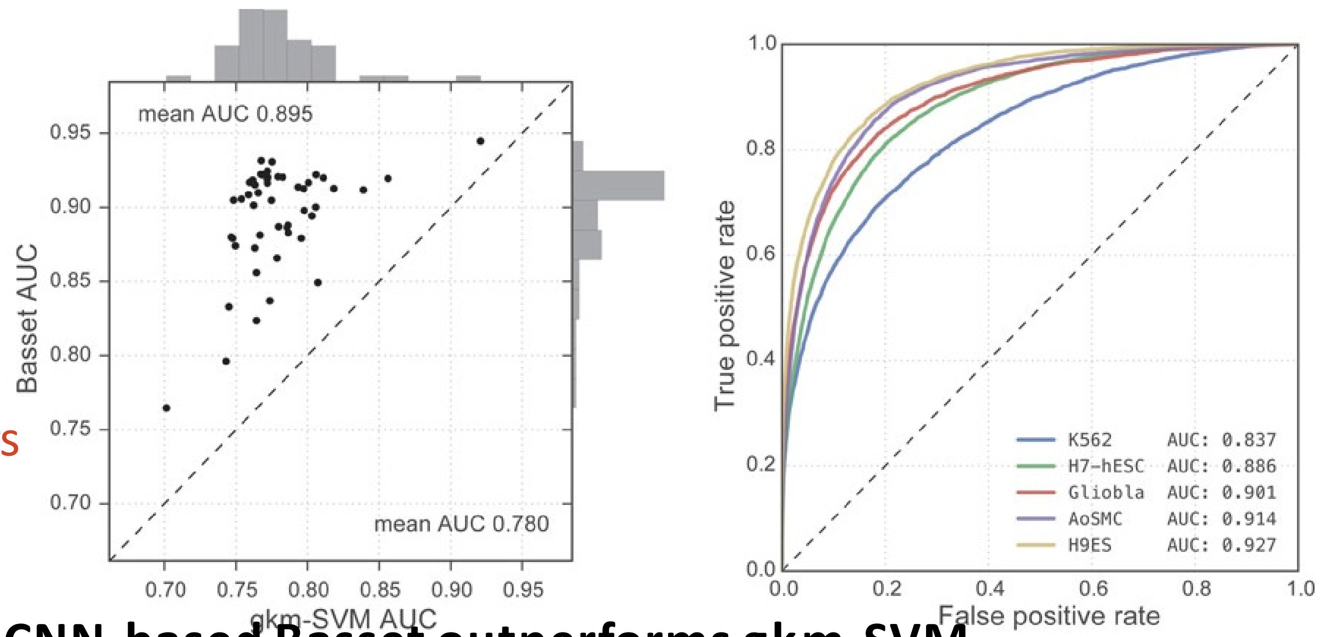
David R. Kelley
Jasper Snoek
John L. Rinn
Genome Research, March 2016

Basset

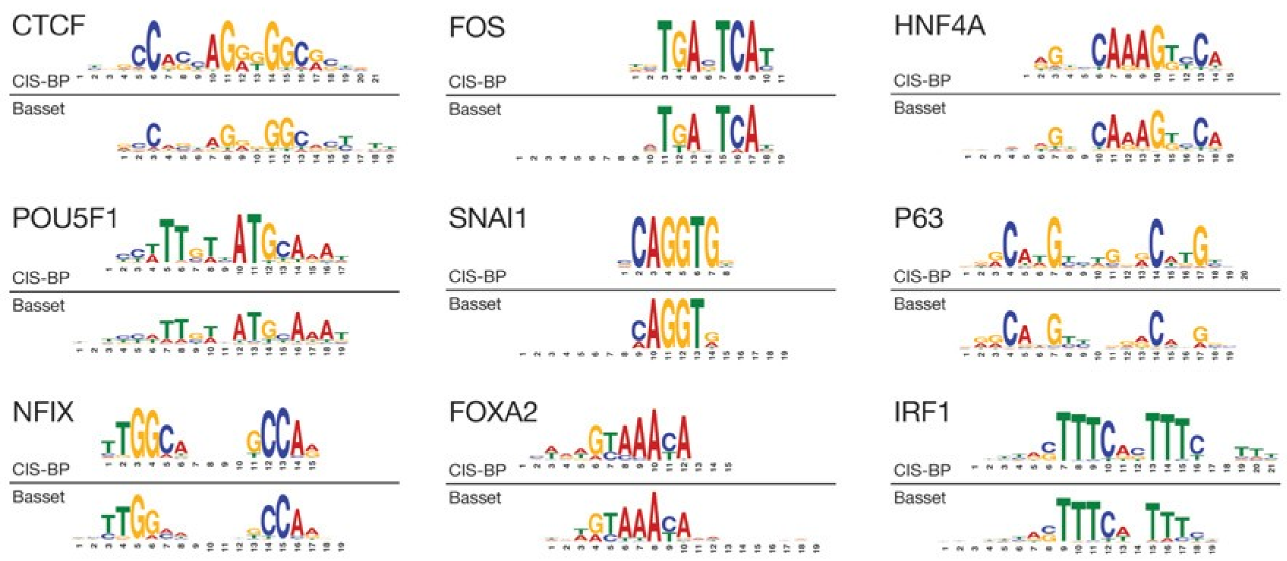


300 convolution filters

Simultaneously predicting DNase sites in 164 cell types



CNN-based Basset outperforms gkm-SVM



Convolutional filters connected to the input sequence recapitulate some known TF motifs

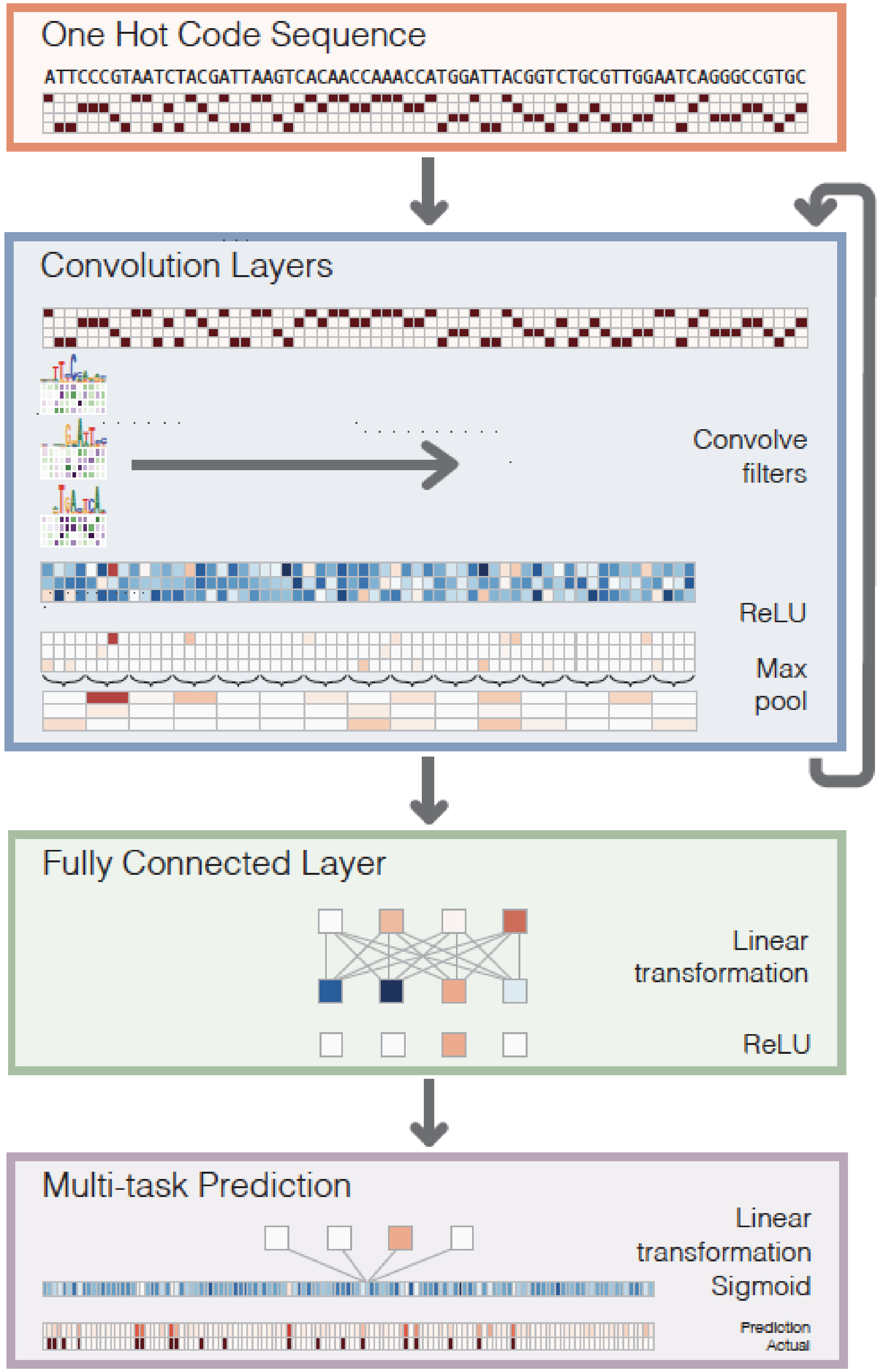
[Kelley et al., 2016]

Bassett architecture for accessibility prediction

Input:
600 bp

1.9 million training examples

Output:
168 bits

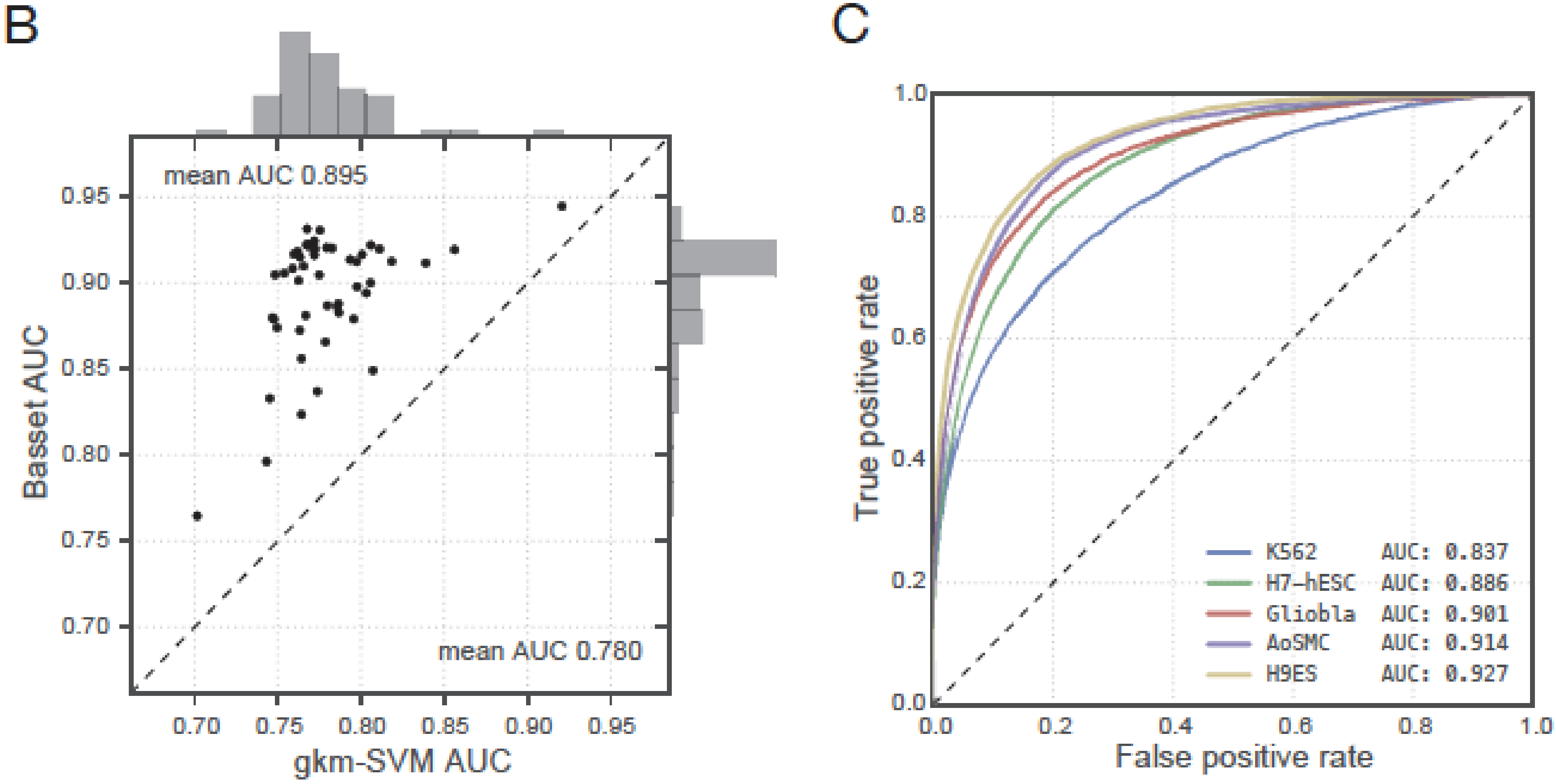


300 filters
3 conv layers
3 FC layers

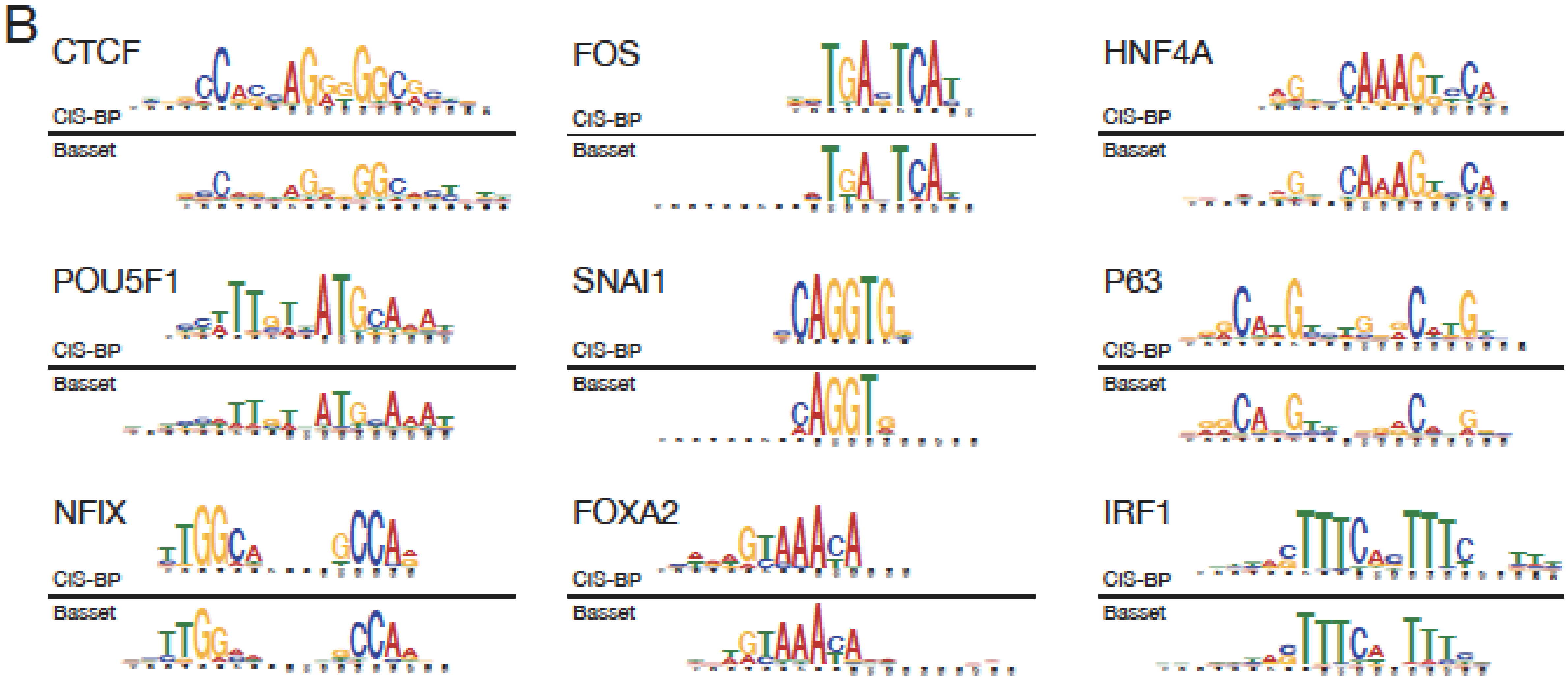
3 fully connected layers

168 outputs
(1 per cell type)

Bassett AUC performance vs. gkm-SVM

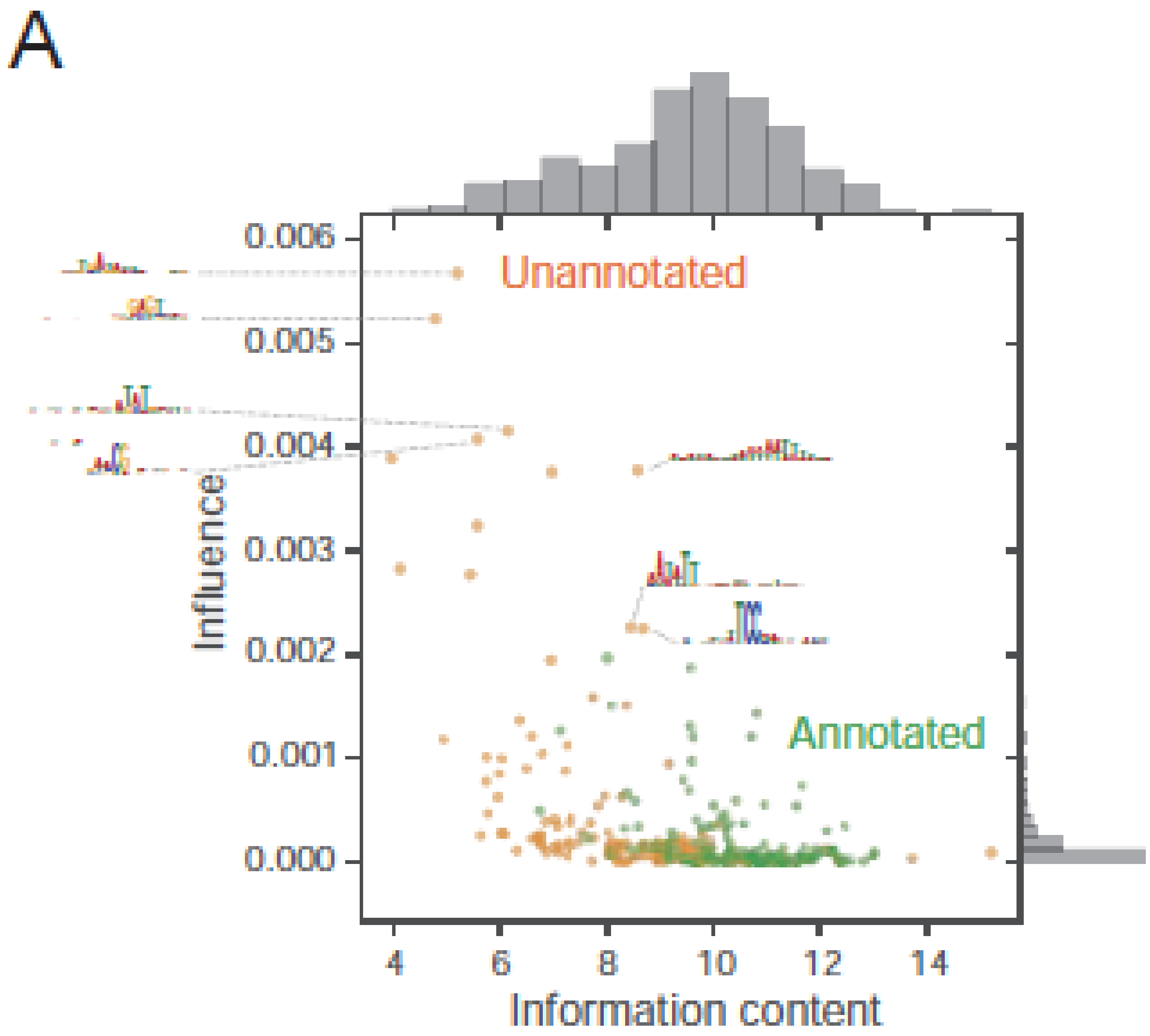


45% of filter derived motifs are found in the CIS-BP database

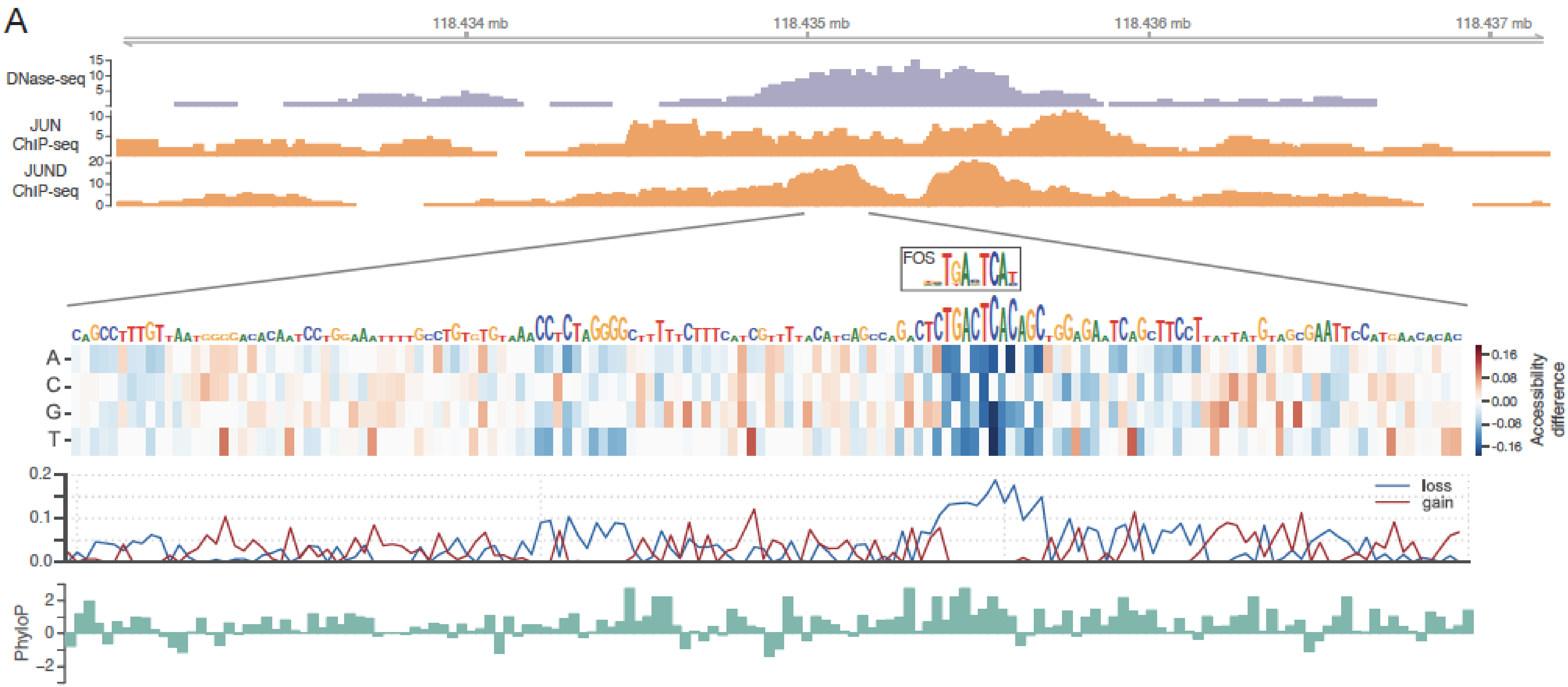


Motifs created by clustering matching input sequences and computing PWM

Motif derived from filters with more information tend to be annotated

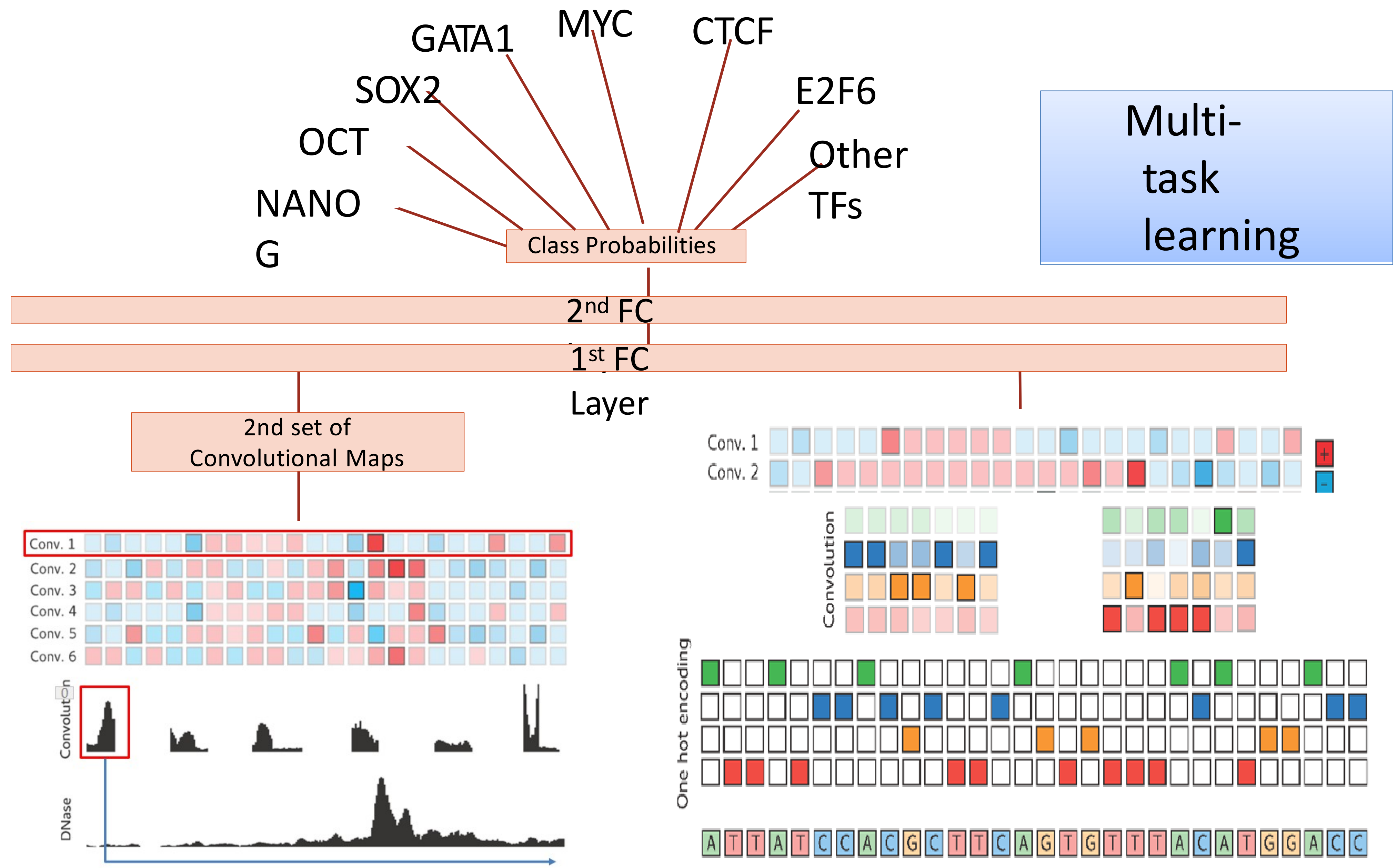


Computational saturation mutagenesis of an AP-1 site reveals loss of accessibility



Regulatory Genomics CNNs in Practice:
(d) Chrompůter

ChromPuter



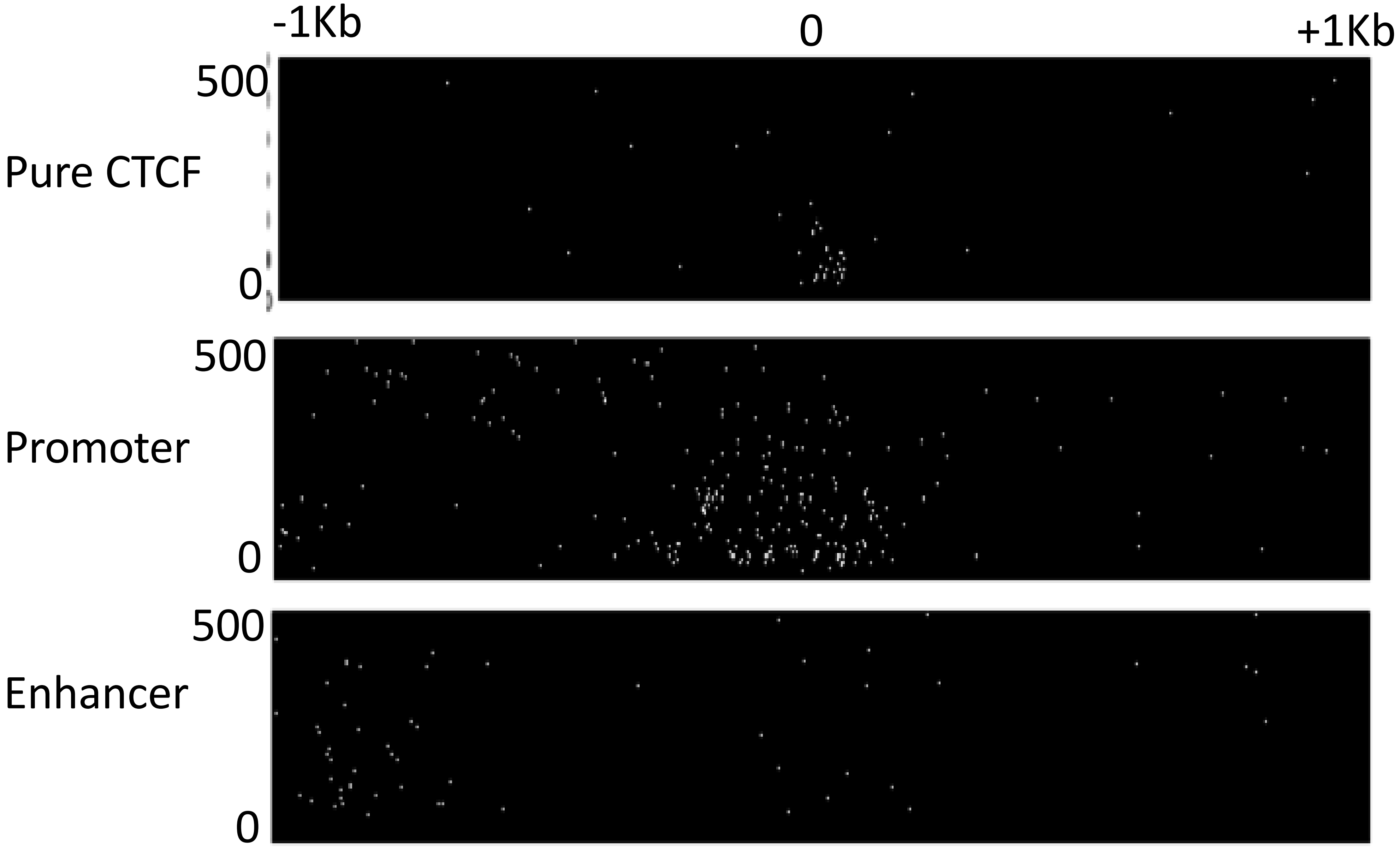
Multi-task learning

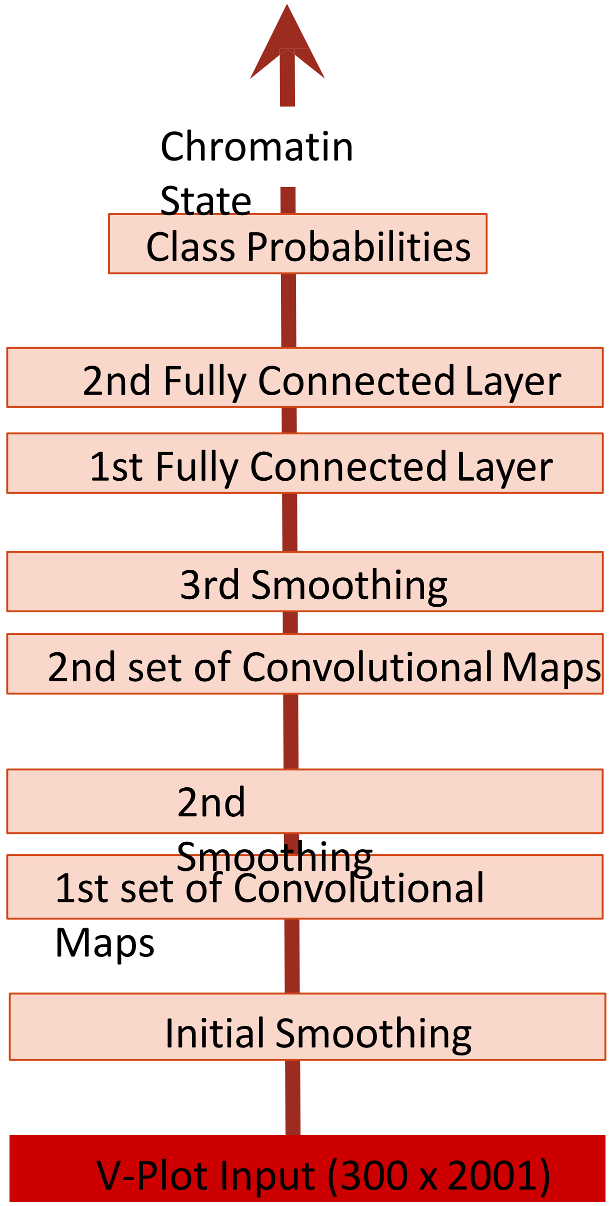
1D DNase-seq/ATAC-seq profile

DNA sequence

(Anshul Kundaje's group from Stanford)

How does a deep conv. neural network transform the raw V-plot input at each layer





Chromatin State

Class Probabilities

2nd Fully Connected Layer

1st Fully Connected Layer

3rd Smoothing

2nd set of Convolutional Maps

2nd Smoothing

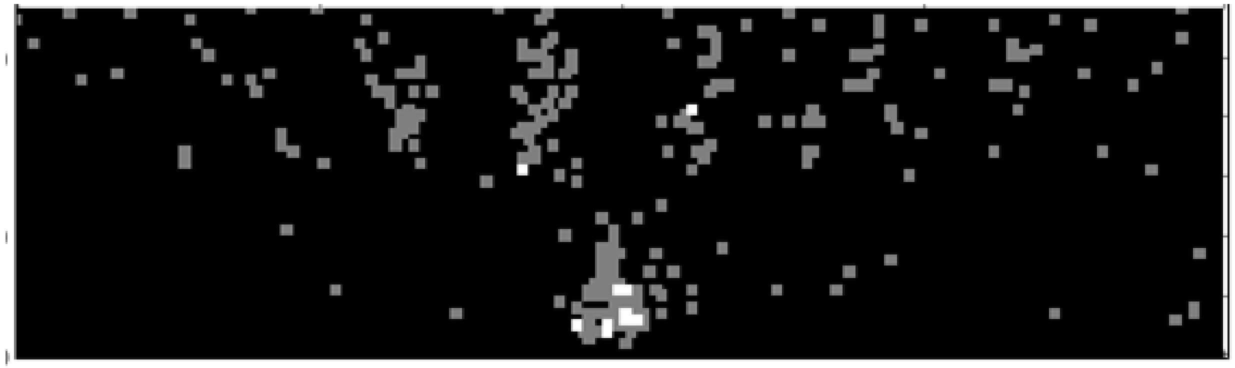
1st set of Convolutional Maps

Initial Smoothing

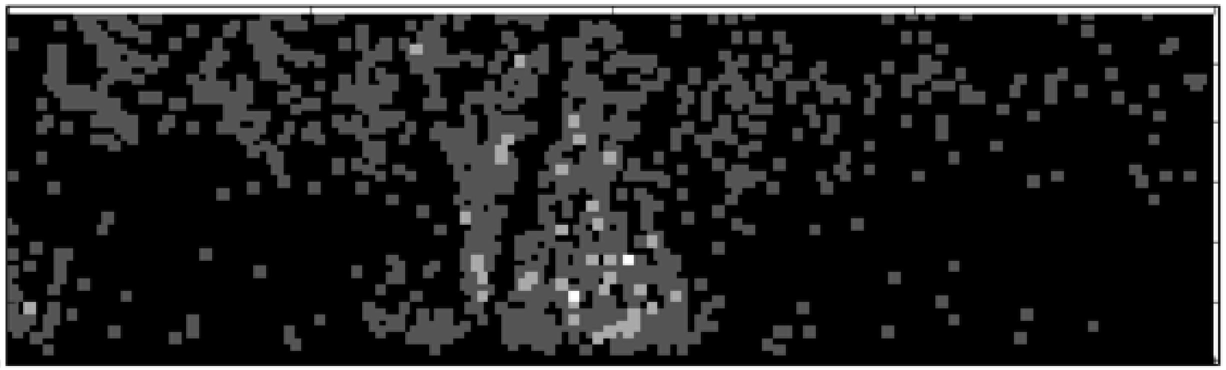
V-Plot Input (300 x 2001)

After initial pooling (smoothing)

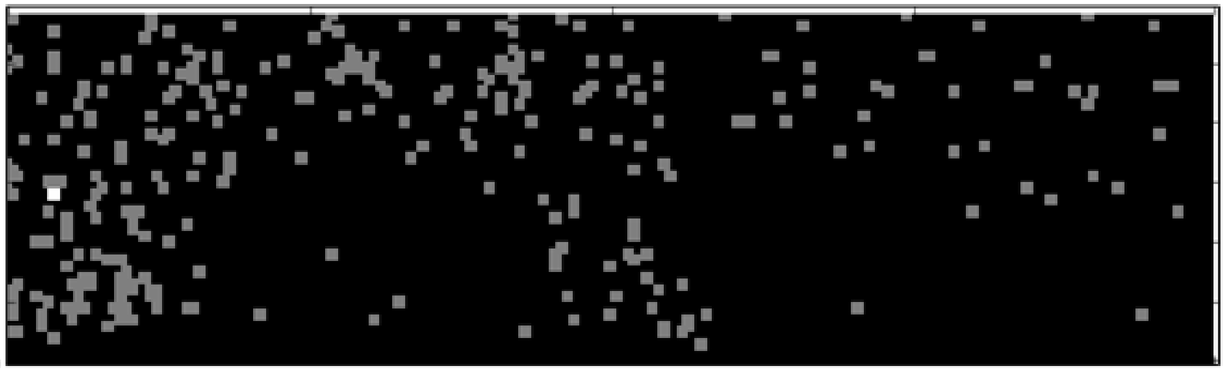
Pure CTCF

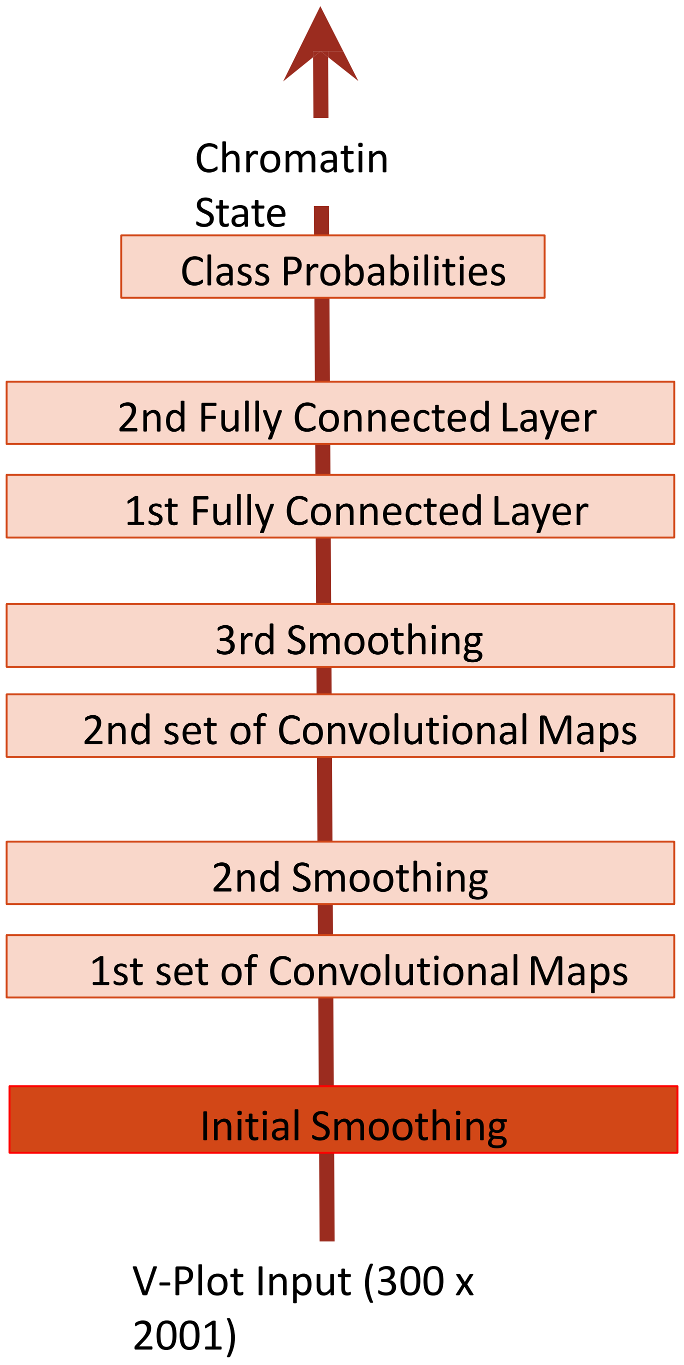


Promoter



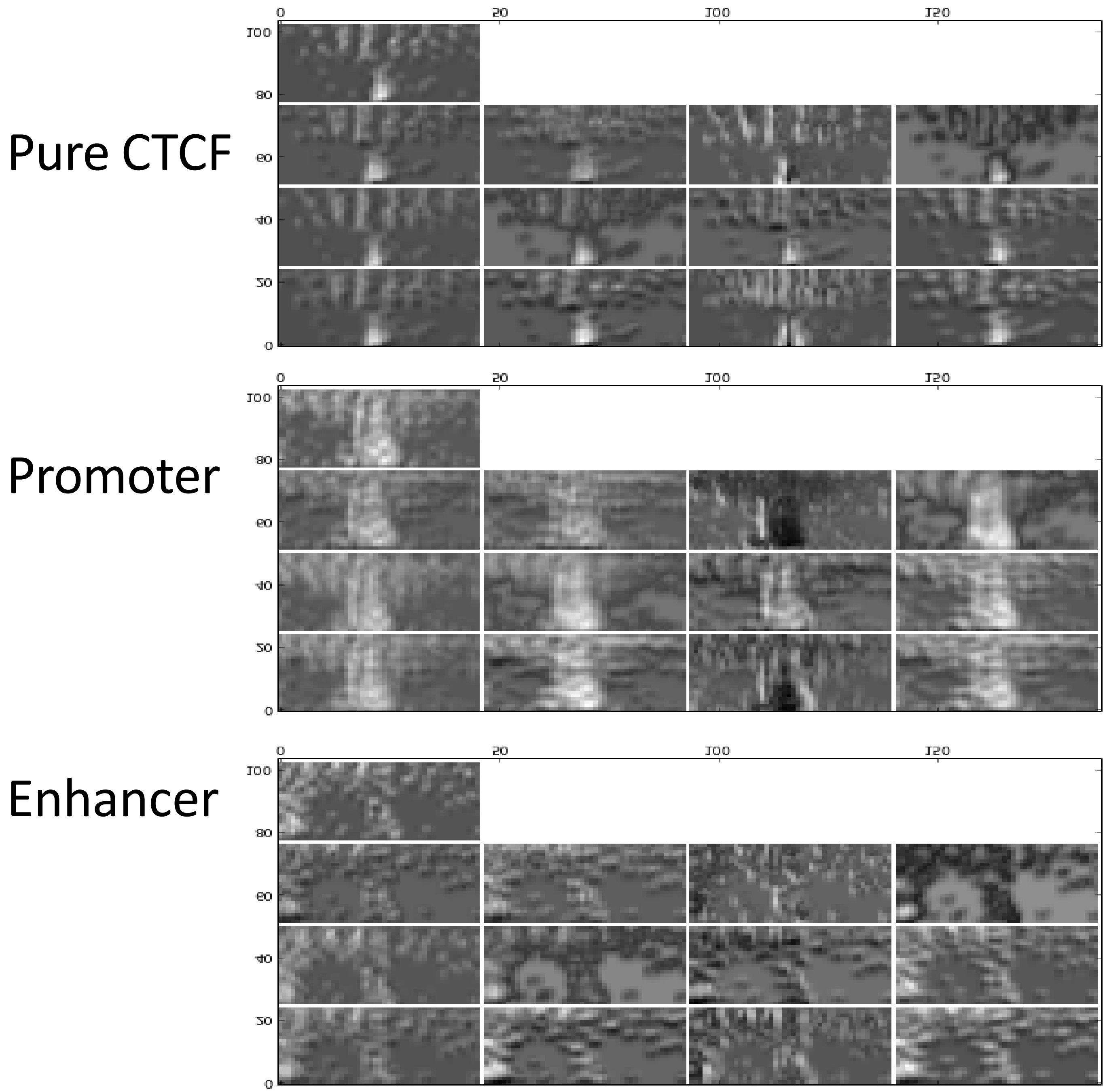
Enhancer

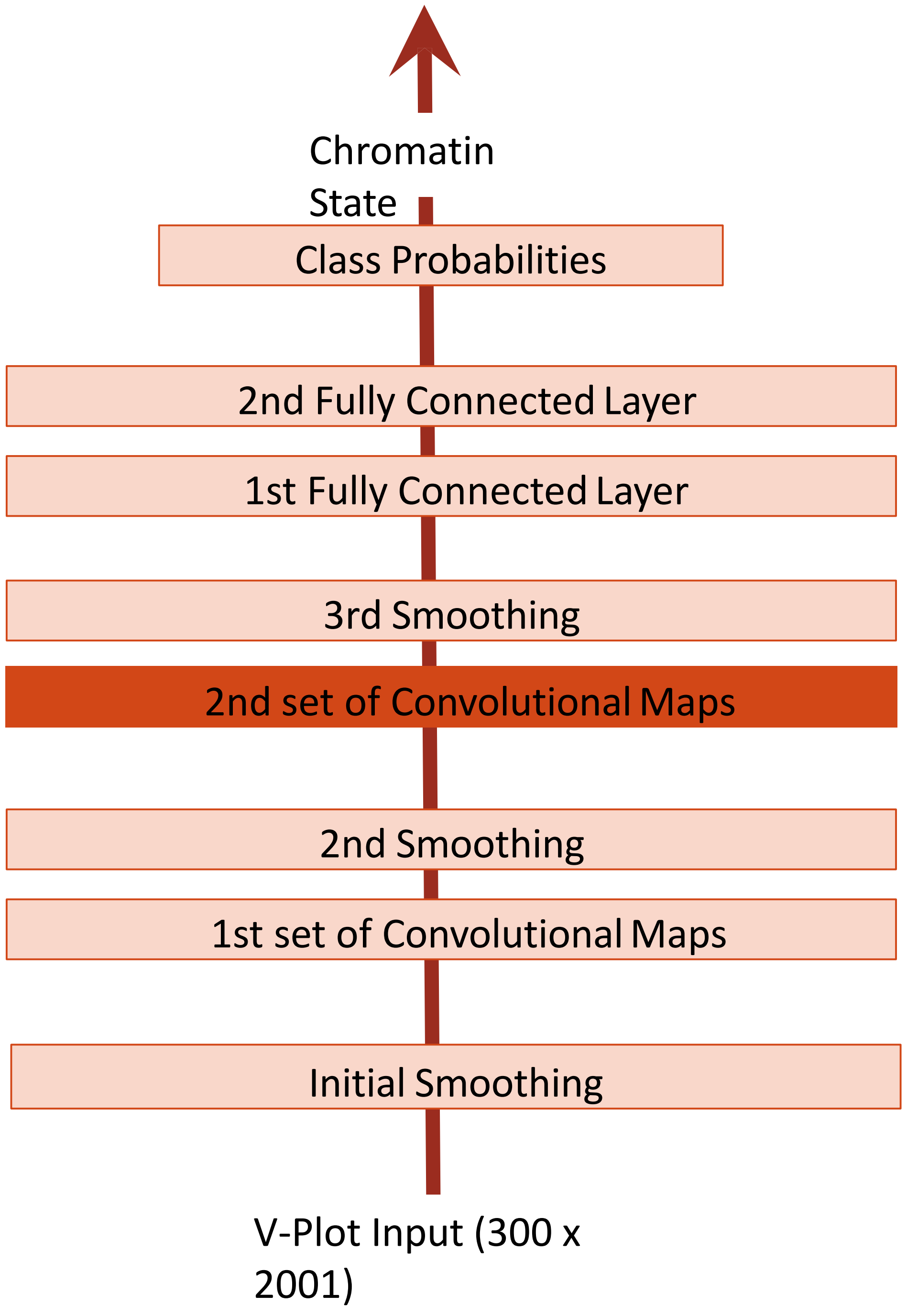




Chromatin State
Class Probabilities
2nd Fully Connected Layer
1st Fully Connected Layer
3rd Smoothing
2nd set of Convolutional Maps
2nd Smoothing
1st set of Convolutional Maps
Initial Smoothing
V-Plot Input (300 x 2001)

Second set of convolutional maps





Chromatin State

Class Probabilities

2nd Fully Connected Layer

1st Fully Connected Layer

3rd Smoothing

2nd set of Convolutional Maps

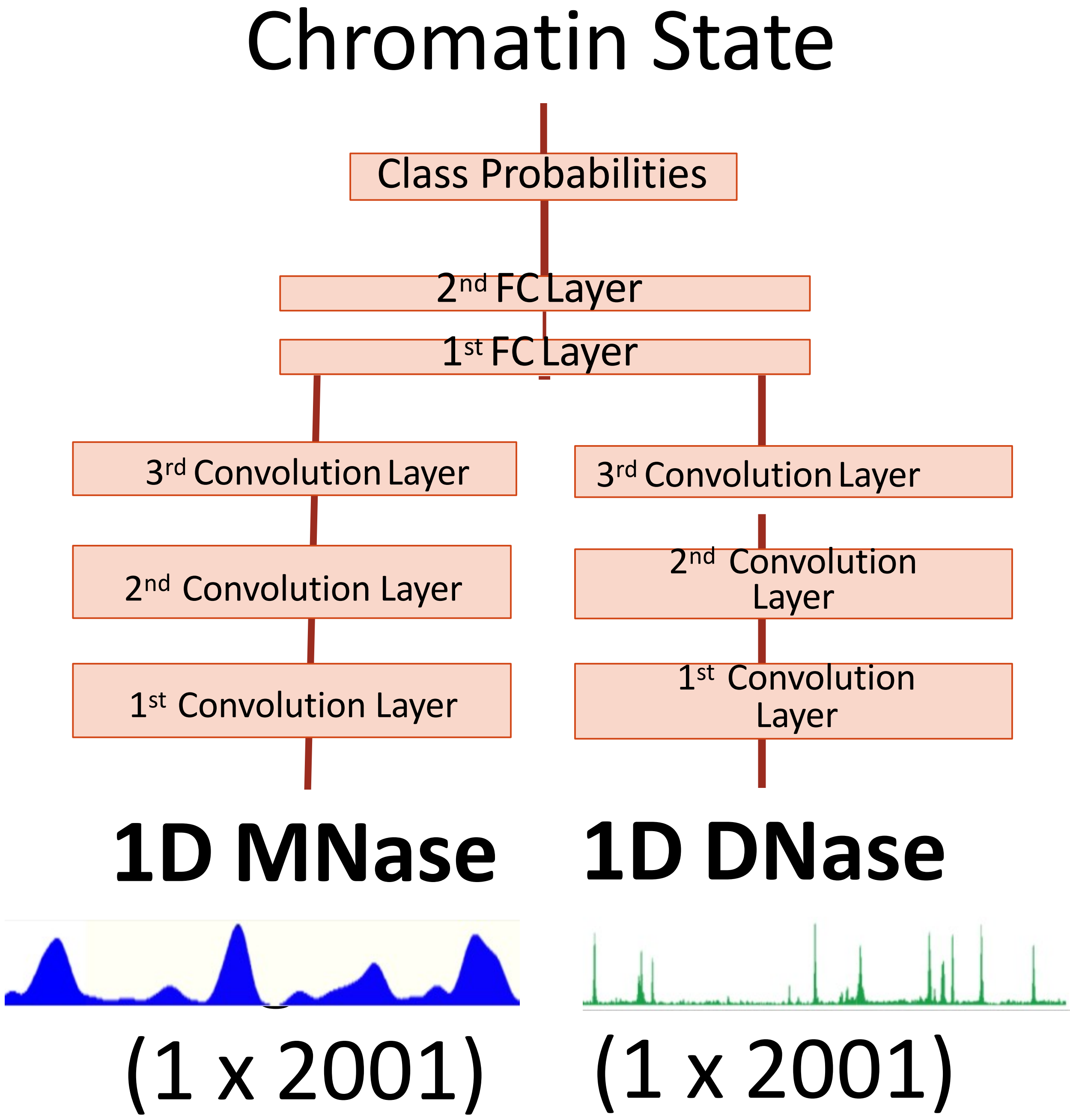
2nd Smoothing

1st set of Convolutional Maps

Initial Smoothing

V-Plot Input (300 x 2001)

Learning from **multiple 1D functional data** (e.g. DNase, MNase)



Chromatin State

Class Probabilities

2nd FC Layer

1st FC Layer

3rd Convolution Layer

2nd Convolution Layer

1st Convolution Layer

3rd Convolution Layer

2nd Convolution Layer

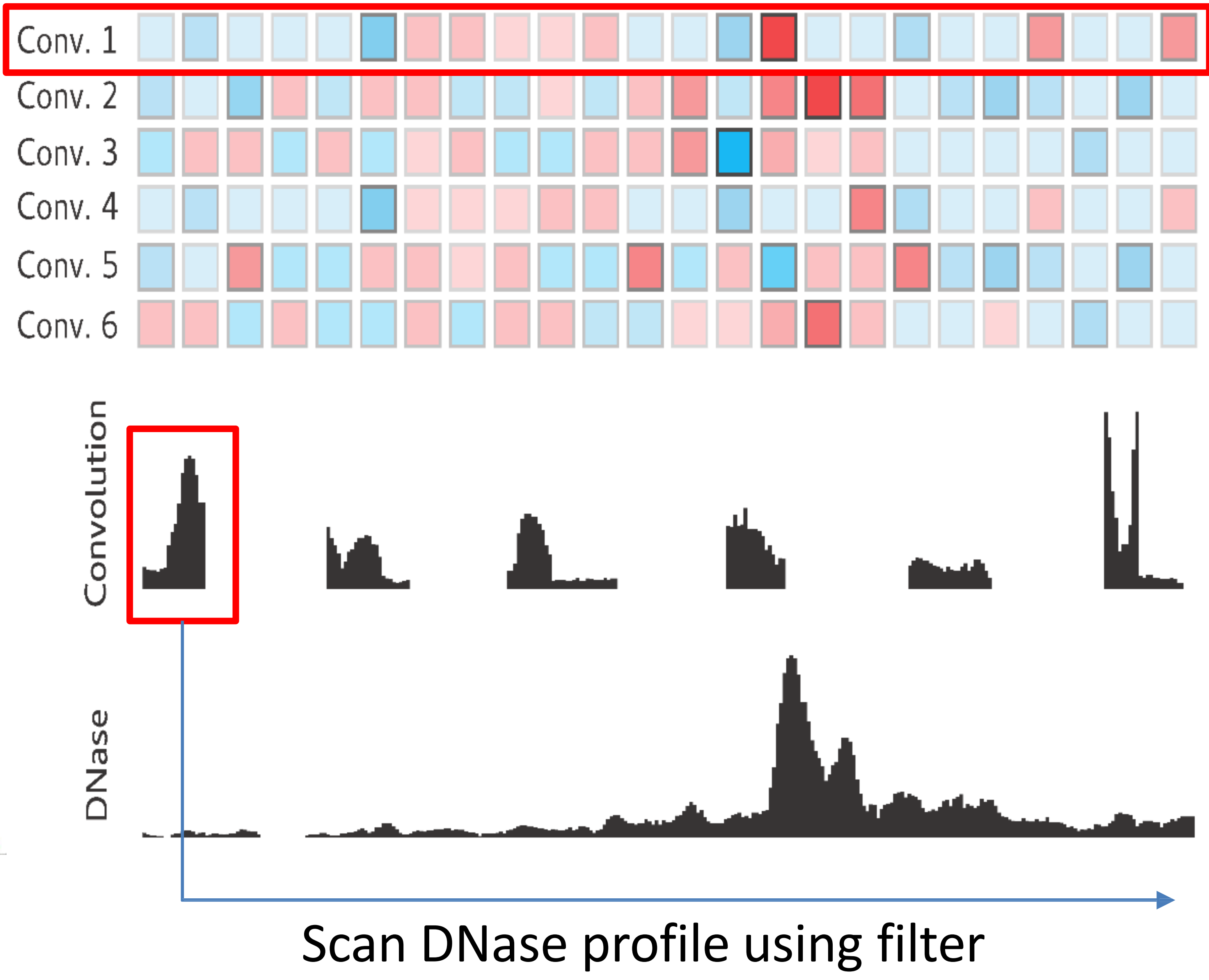
1st Convolution Layer

1D MNase

1D DNase

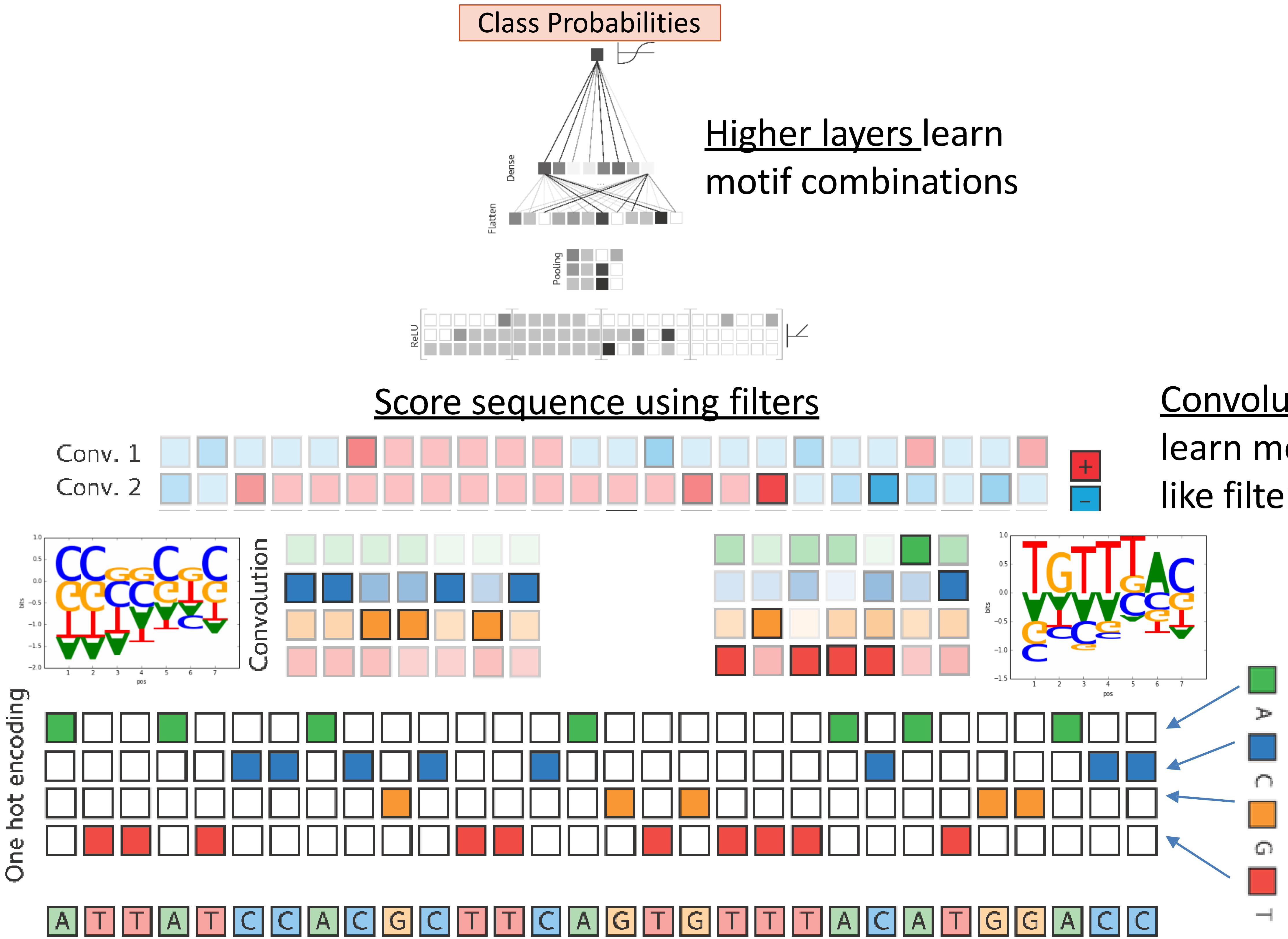
(1 x 2001)

(1 x 2001)



Scan DNase profile using filter

Learning from raw DNA sequence



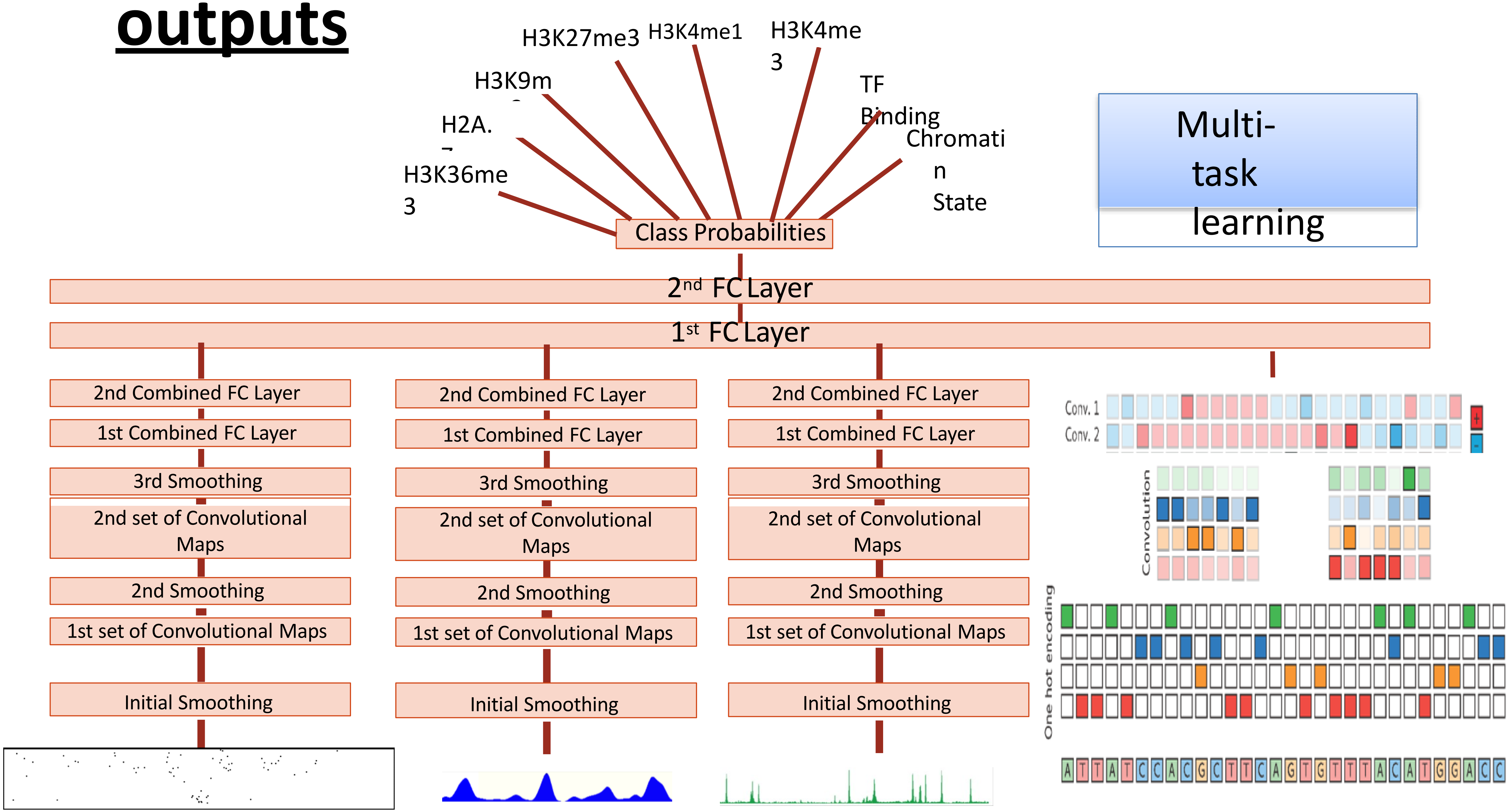
Higher layers learn motif combinations

Score sequence using filters

Convolutional layers learn motif (PWM) like filters

THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence) to simulatenously **predict multiple outputs**



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

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%**

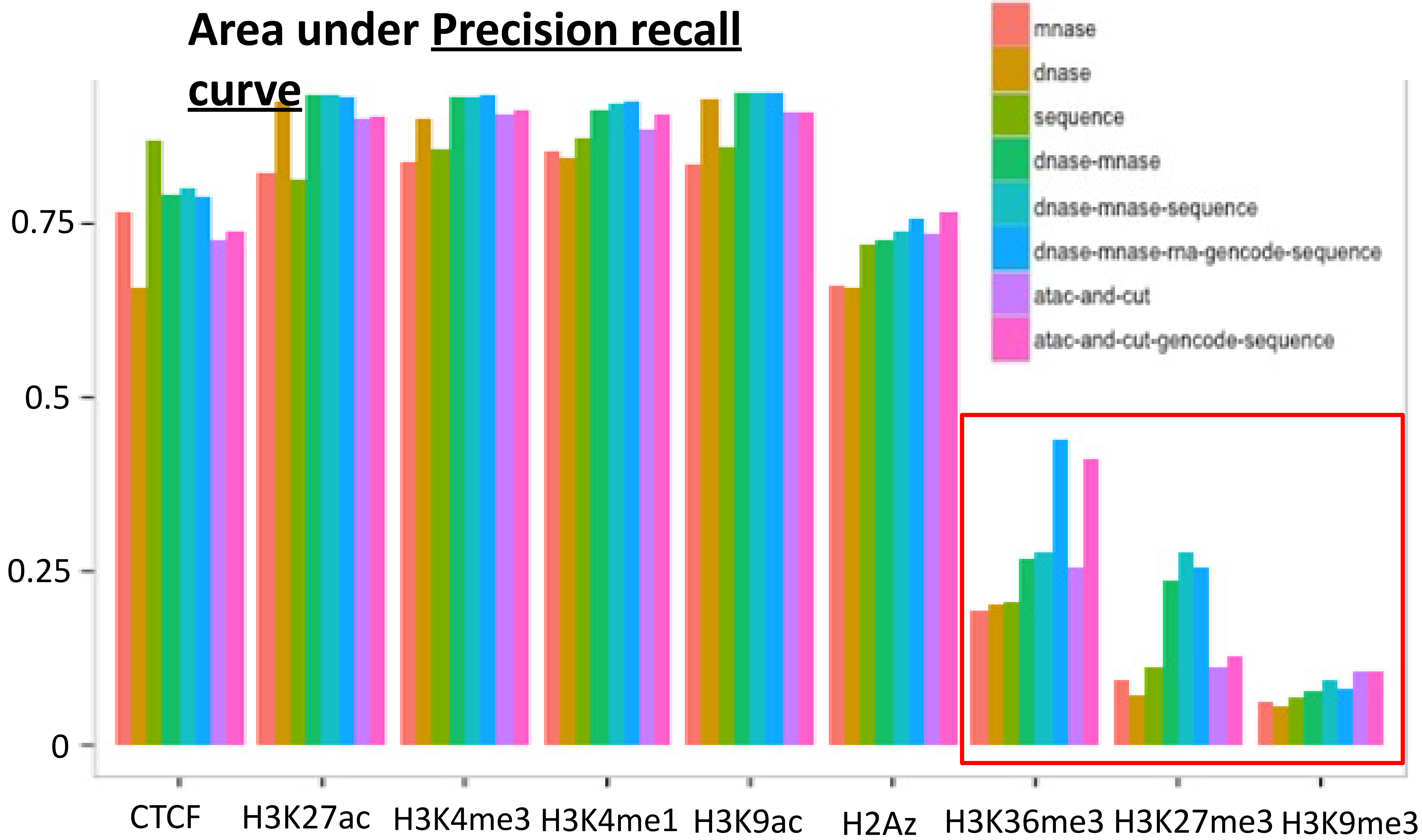
High cross cell-type chromatin state prediction

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

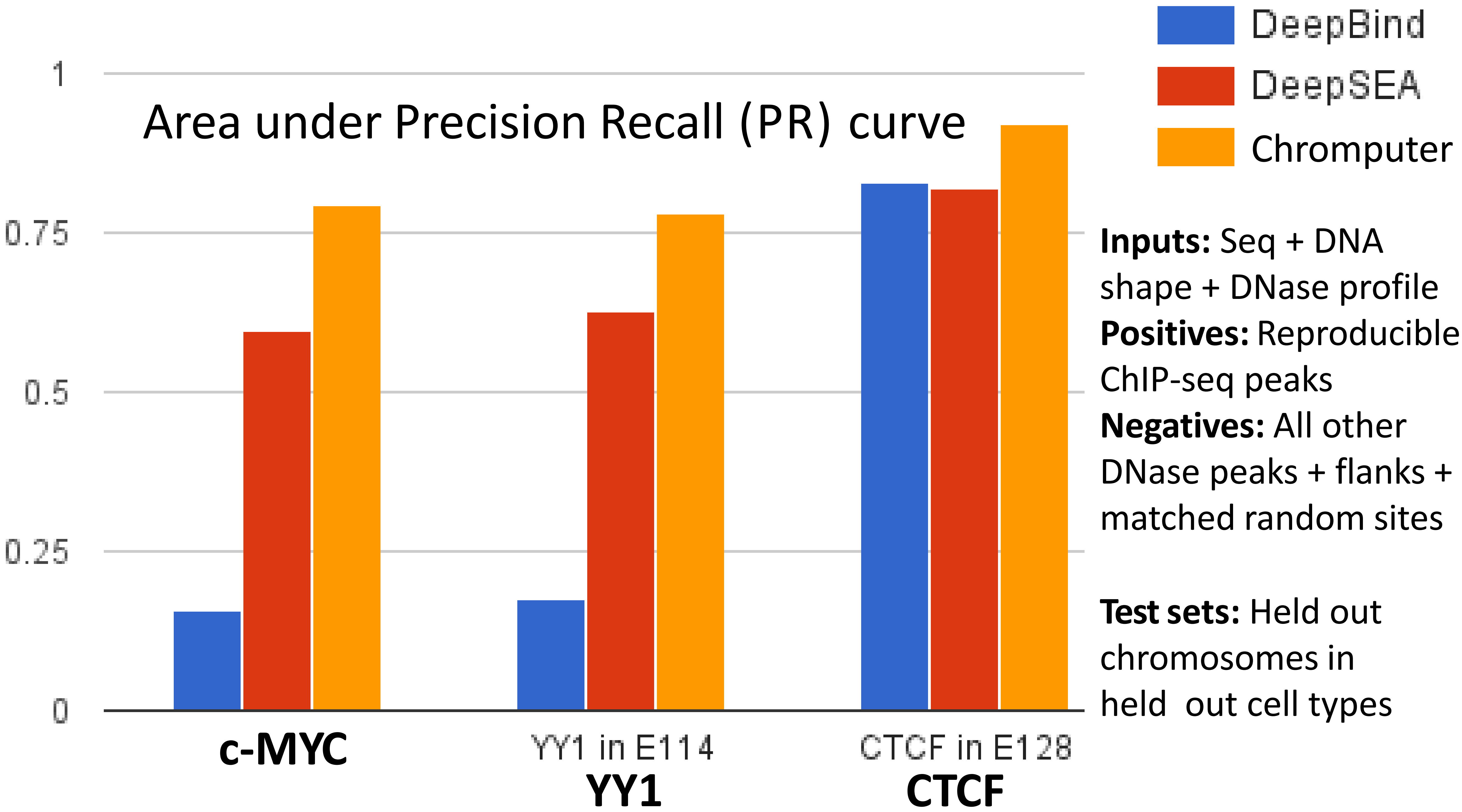
8 class chromatin state accuracy		
Train ↓ / Test →	GM12878	K562
GM12878	0.816	**0.818**
K562	**0.769**	0.844

Predicting individual histone marks from ATAC/DNase/MNase/Sequence

Area under <u>Precision recall curve</u>



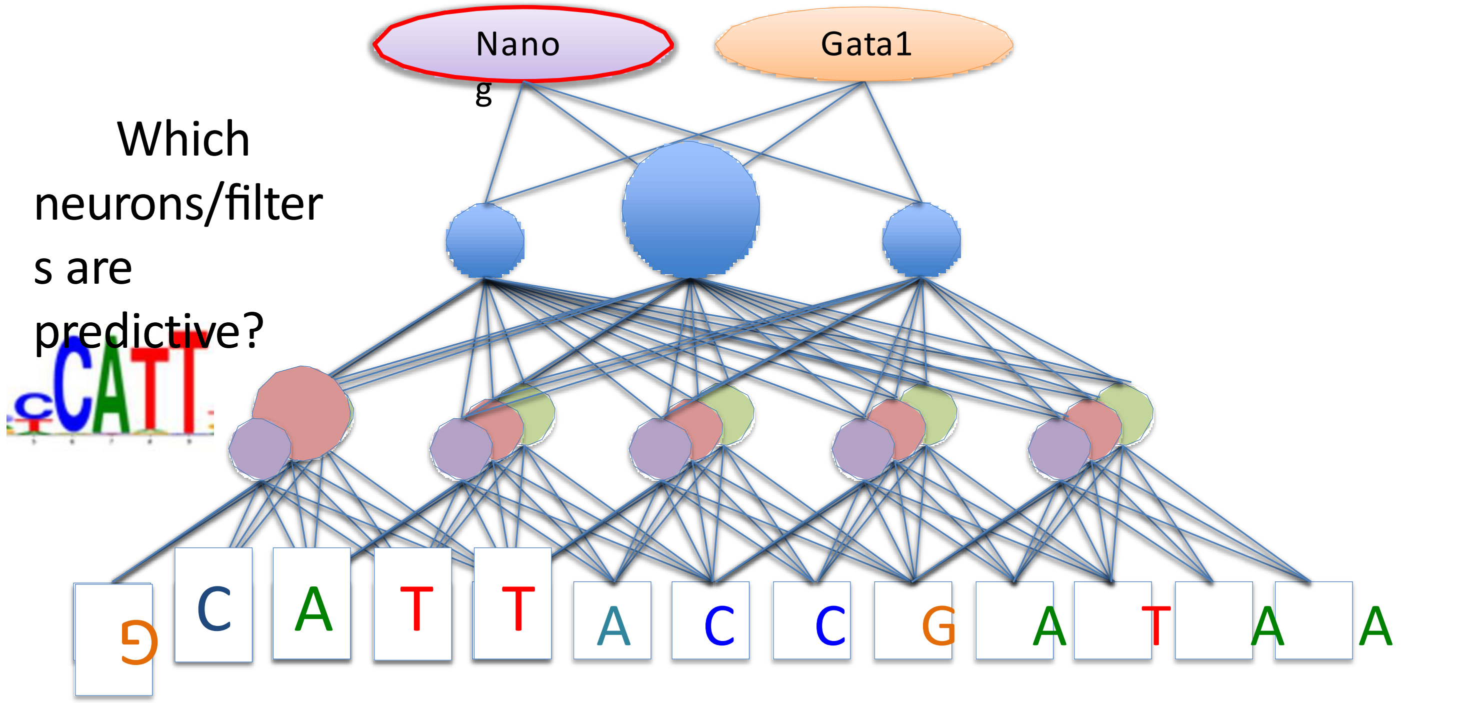
Chromputer trained on TF ChIP-seq predicts cross cell-type in-vivo TF binding with high accuracy



Inputs: Seq + DNA shape + DNase profile
Positives: Reproducible ChIP-seq peaks
Negatives: All other DNase peaks + flanks + matched random sites

Test sets: Held out chromosomes in held out cell types

DeepLift reveals feature importance at the input layer



Which nucleotides in input sequence are contributing to binding

Key idea:

- ReLU is piece-wide linear
- Backpropagation differences of outputs using observed and reference inputs (e.g., inputs of all zeros) to obtain gradient w.r.t. the input
- Importance of any input to any output is the gradients weighted by the input itself

(Anshul Kundaje’s group from Stanford)

Deep Learning for Regulatory Genomics

1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⇔ DNA letters. Patches/filters ⇔ Motifs. Higher ⇔ combinations
   - Learning convolutional filters ⇔ Motif discovery. Applying them ⇔ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

STANFORD
COMPUTER SCIENCE

Stanford
MEDICINE | Department of Genetics

Deep learning at base-resolution reveals cis-regulatory motif syntax

Anshul Kundaje

Twitter:@anshulkundaje

Website: http://anshul.kundaje.net

Acknowledgements

Ziga Avsec

Avanti Shrikumar

Melanie Weilert

Amr Mohamed

Julia Zeitlinger

Funding

1R01HG009674

1U01HG009431

1U24HG009446

ENCODE

R01ES02500902

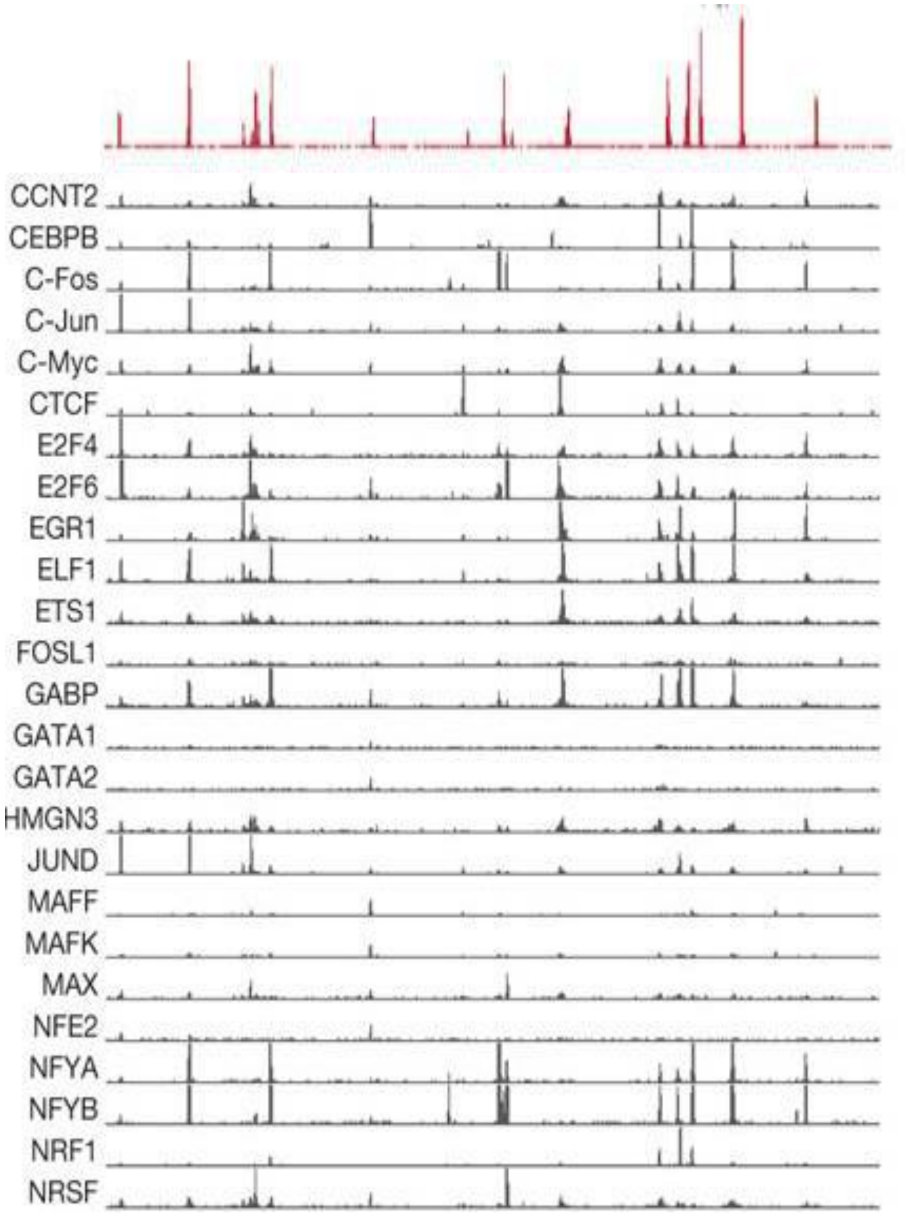
ROADMAP epigenomics PROJECT

1DP2OD022870

NIH DIRECTOR'S NEW INNOVATOR AWARD

- Khyati Dalal
- Sabrina Kruger
- Robin Fropf
- Charles McAnany
- Julien Gagneur

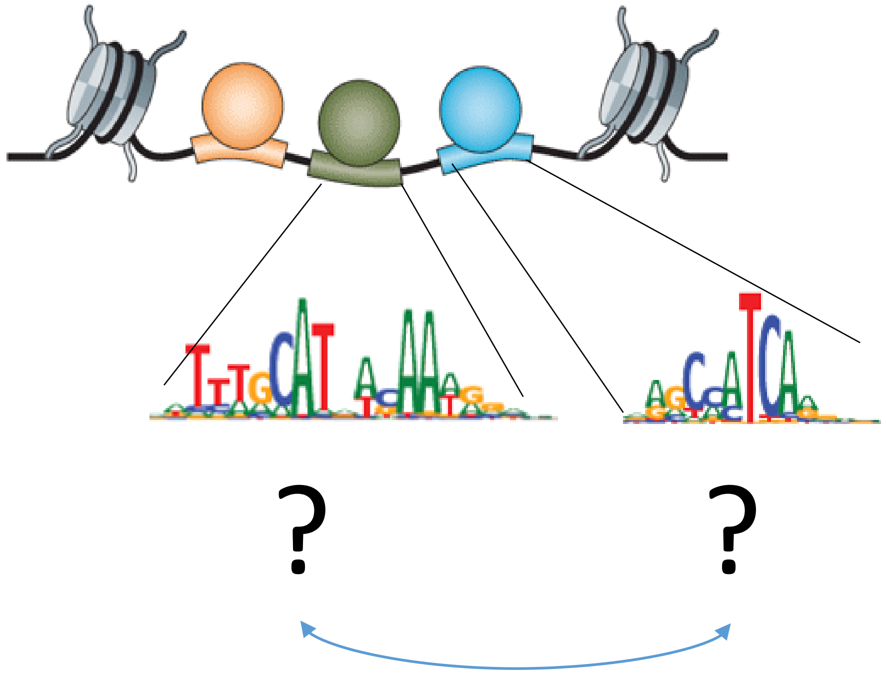
Deciphering syntax of regulatory DNA



chromatin accessibility
(ATAC-seq / DNase-seq)

Transcription factor ChIP-seq experiments

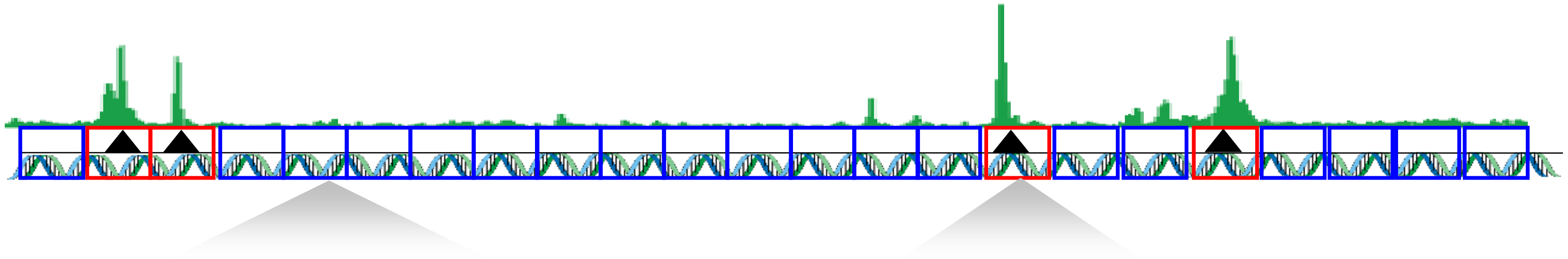
Adapted from Thurman et al 2012



Motif syntax: rules of arrangement, preferred spacing, orientation => cooperativity

Predictive model of regulatory DNA

Transcription factor ChIP-seq data OR chromatin accessibility (DNase-seq / ATAC-seq data)

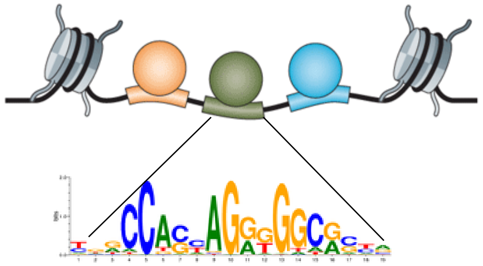


...GACTTGAAACGGCATTG...

Inactive (0) (0.3)

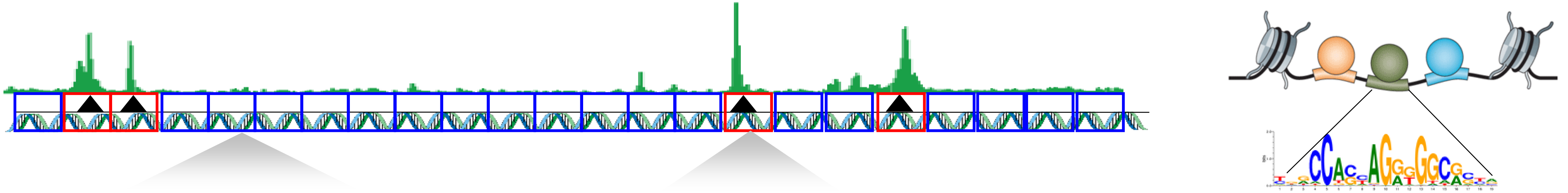
...GACAGATAATGCATTGA...

Active (+1) (20.2)



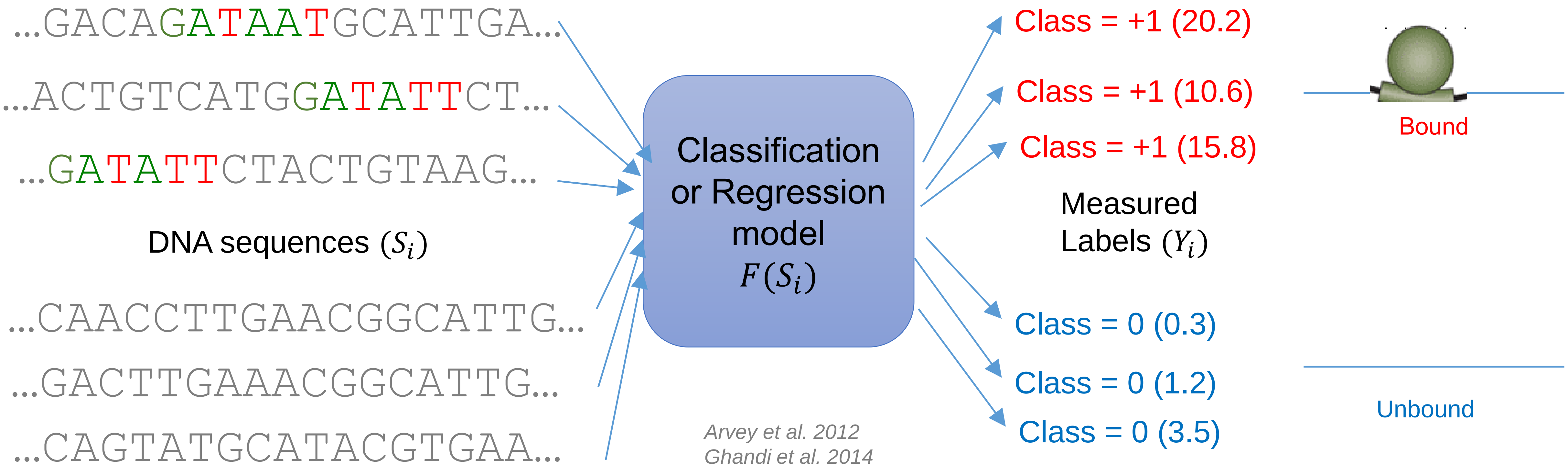
Predictive model of regulatory DNA

Transcription factor ChIP-seq data OR chromatin accessibility (DNase-seq / ATAC-seq data)



...GACTTGAAACGGCATTG...
Inactive (0) (0.3)

...GACAGATAATGCATTGA...
Active (+1) (20.2)



...GACAGATAATGCATTGA...

...ACTGTCATGGATATTCT...

...GATATTCTACTGTAAG...

DNA sequences (S_i)

...CAACCTTGAACGGCATTG...

...GACTTGAAACGGCATTG...

...CAGTATGCATACGTGAA...

Classification or Regression model $F(S_i)$

Class = +1 (20.2)

Class = +1 (10.6)

Class = +1 (15.8)

Measured Labels (Y_i)

Class = 0 (0.3)

Class = 0 (1.2)

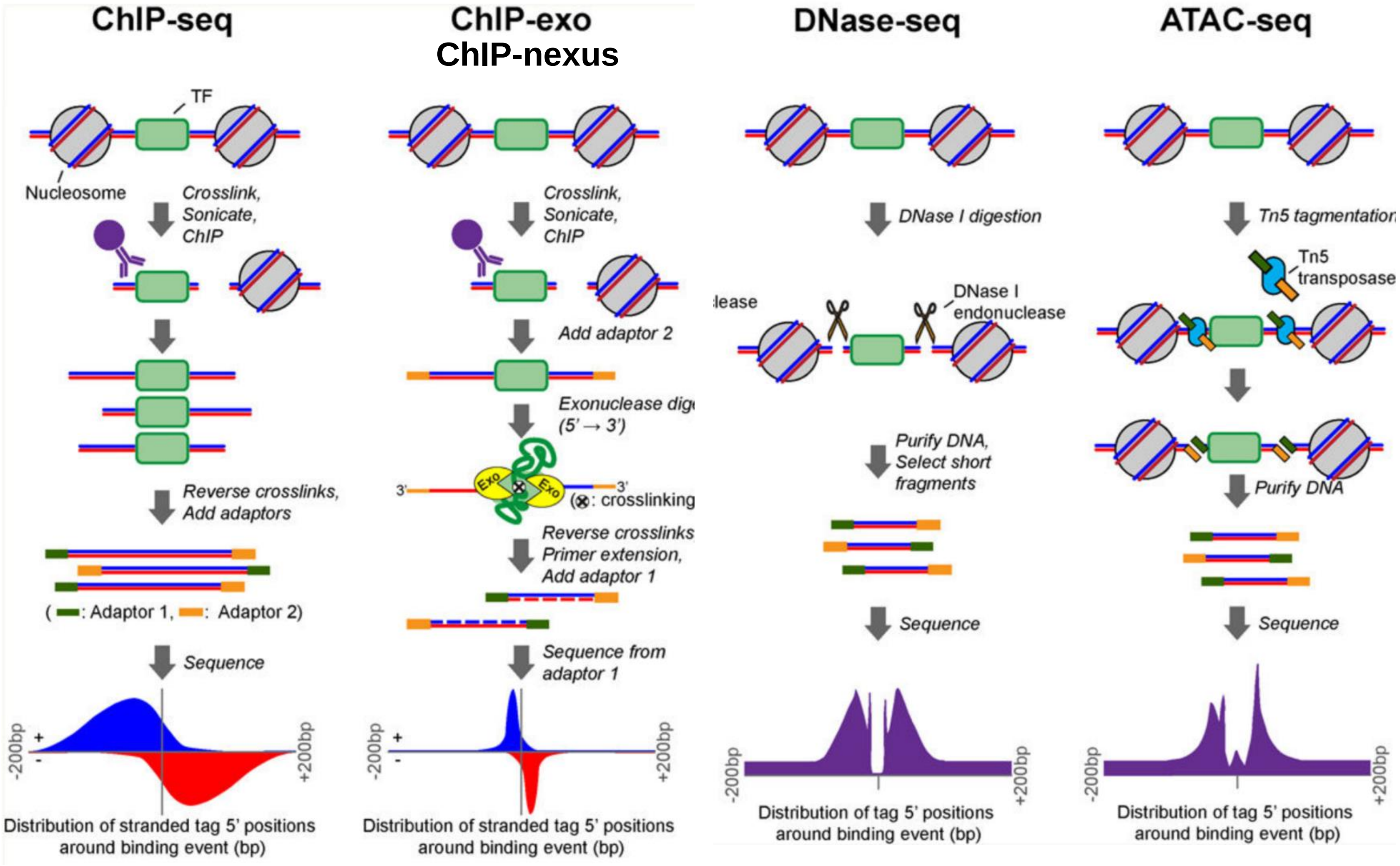
Class = 0 (3.5)

Bound

Unbound

Arvey et al. 2012
Ghandi et al. 2014
Setty et al. 2015

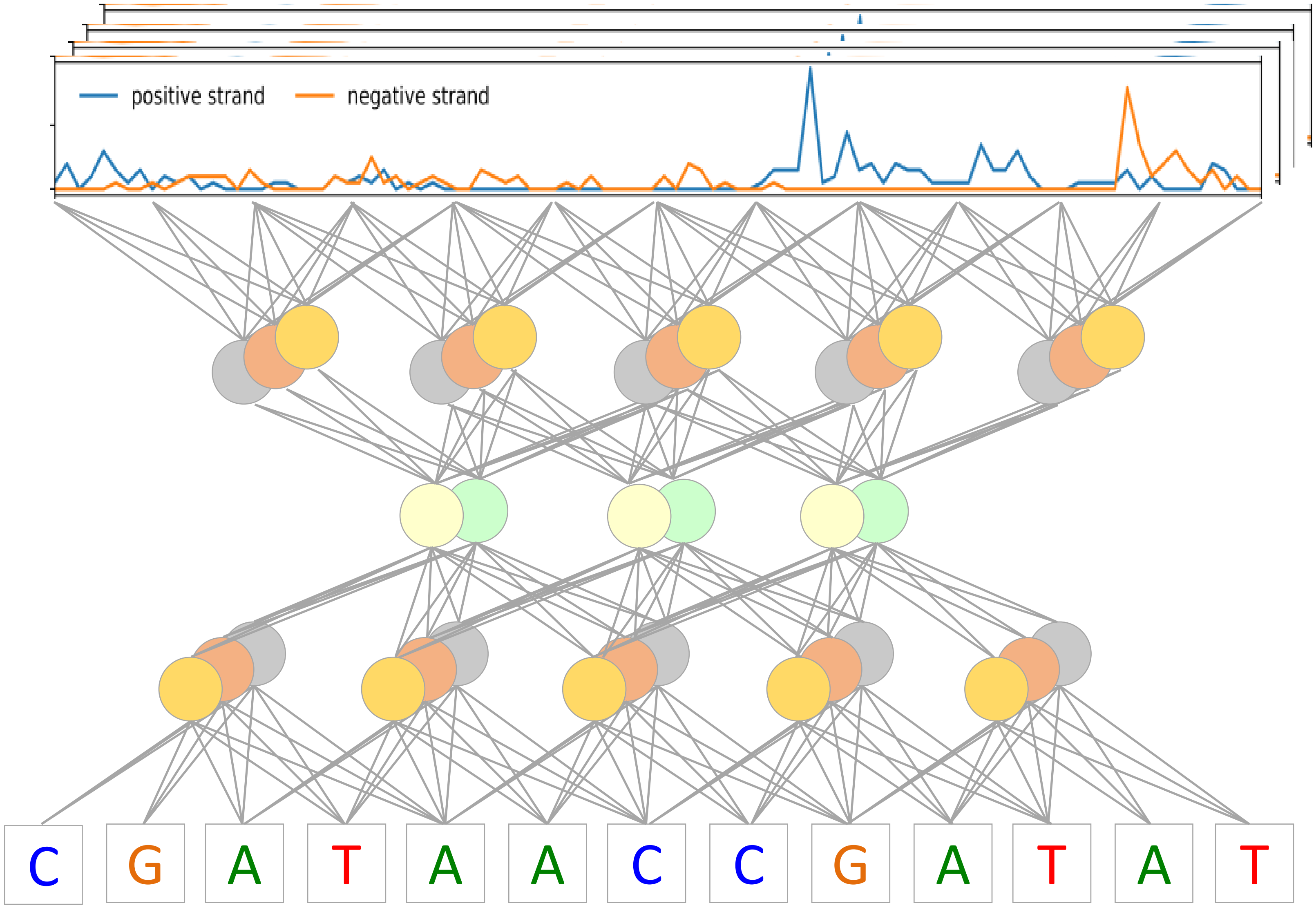
High-resolution 'shapes' of regulatory profiles capture exquisite information about protein-DNA contacts


ChIP-seq
ChIP-exo
ChIP-nexus
DNase-seq
ATAC-seq
TF
Nucleosome
Crosslink, Sonicate, ChIP
Reverse crosslinks, Add adaptors
(: Adaptor 1, : Adaptor 2)
Sequence
+
-
-200bp
+200bp
Distribution of stranded tag 5' positions around binding event (bp)
Crosslink, Sonicate, ChIP
Add adaptor 2
Exonuclease dig
(5' → 3')
3'
Exo
Exo
3'
(⊗: crosslinking
Reverse crosslinks
Primer extension,
Add adaptor 1
Sequence from adaptor 1
+
-
-200bp
+200bp
Distribution of stranded tag 5' positions around binding event (bp)
DNase I digestion
lease
DNase I endonuclease
Purify DNA, Select short fragments
Sequence
-200bp
+200bp
Distribution of tag 5' positions around binding event (bp)
Tn5 tagmentation
Tn5 transposase
Purify DNA
Sequence
-200bp
+200bp
Distribution of tag 5' positions around binding event (bp)

https://doi.org/10.3109/10409238.2015.1051505

BPNet: DNA sequence to base-pair resolution profile regression

stranded base-resolution probability profiles + total read count



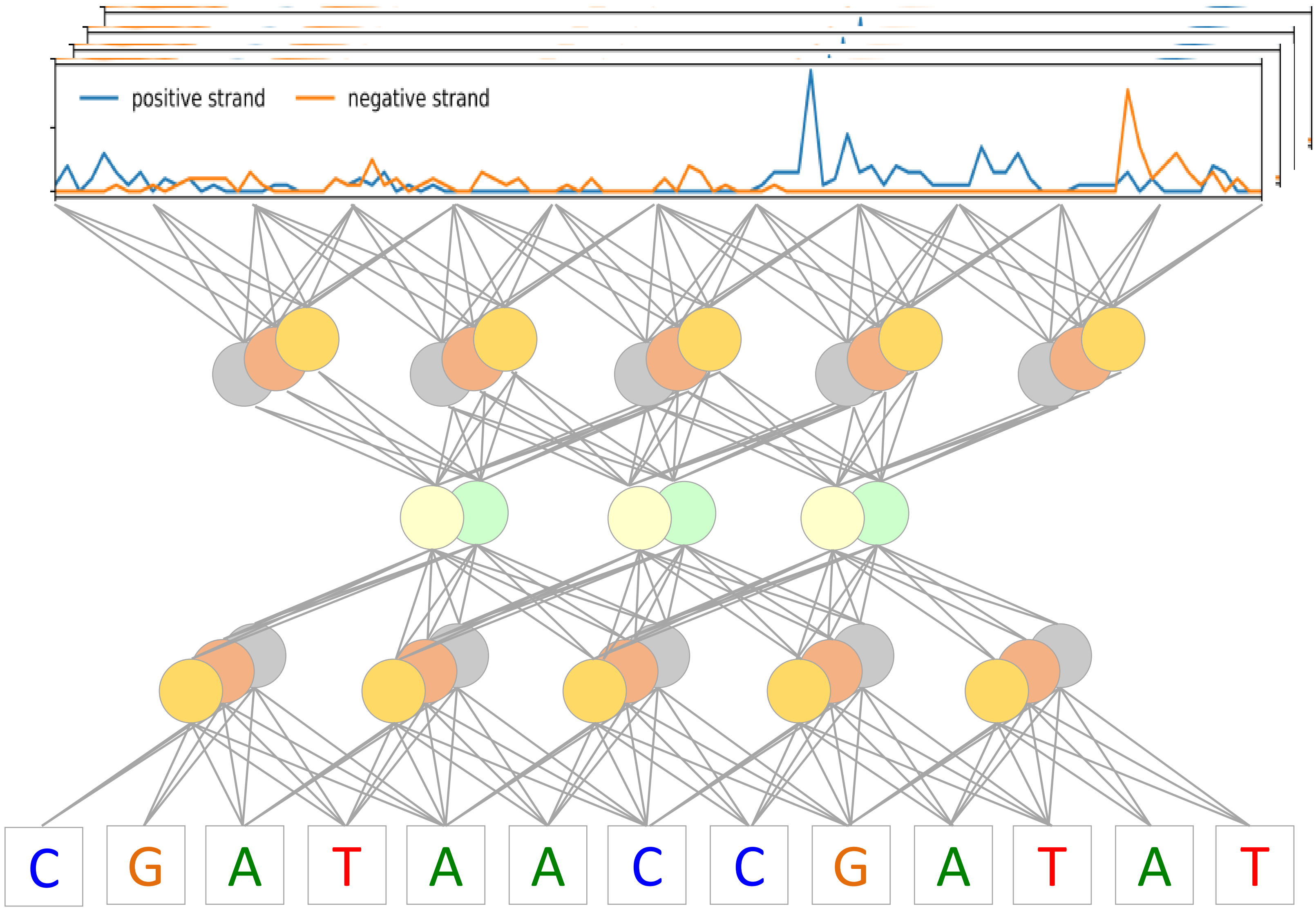
Multi-task training on multiple readouts

1 Kb sequence around all peaks

Ziga Avsec

BPNet: DNA sequence to base-pair resolution profile regression

stranded base-resolution probability profiles + total read count



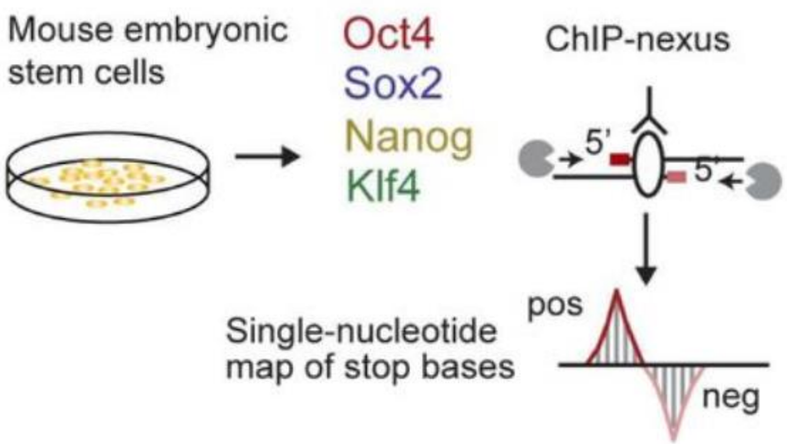
1 Kb sequence around all peaks

Multi-task training on multiple readouts

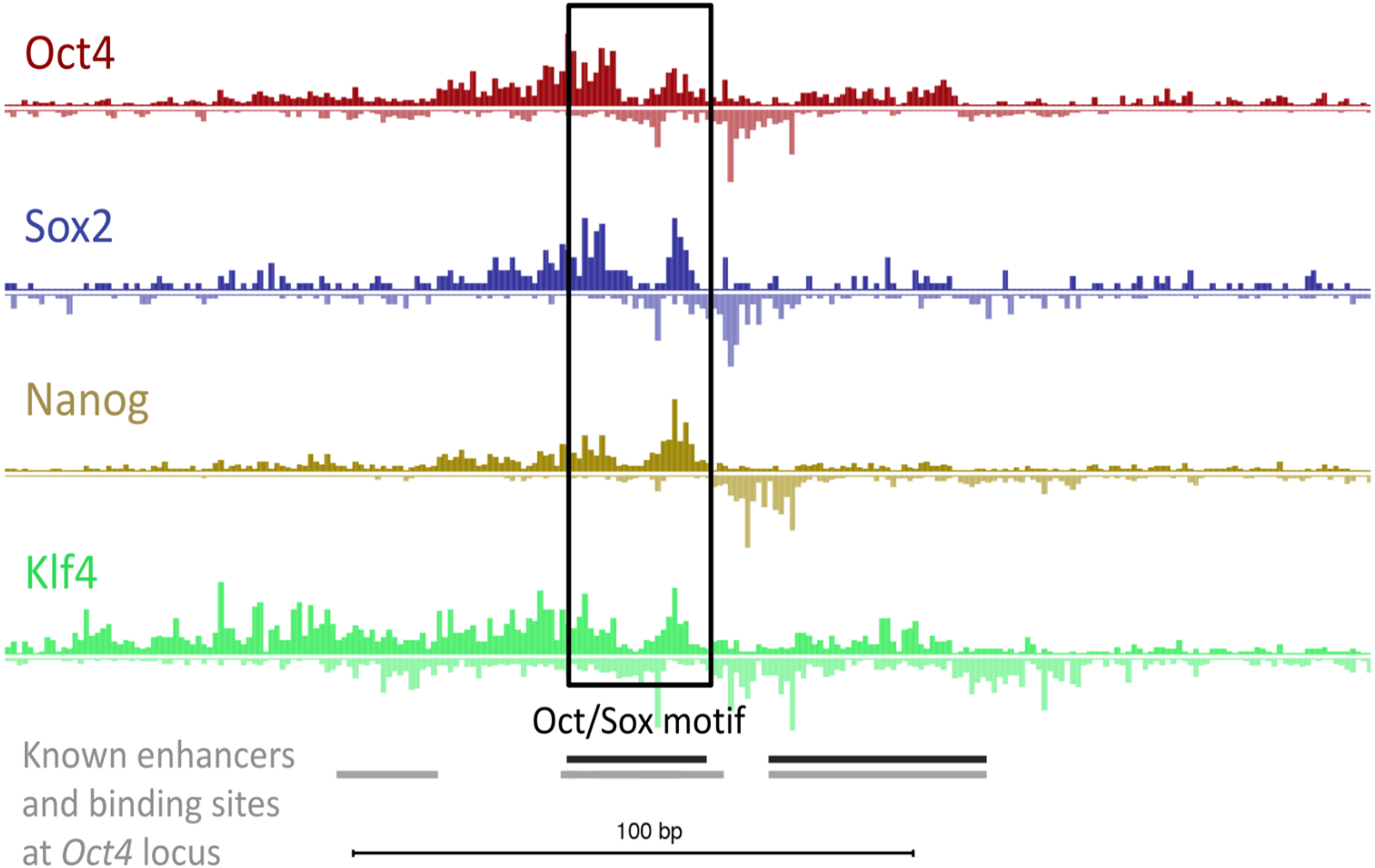
- Novel loss function
  - MSE for log(total counts)
  - Multinomial NLL for profile distribution
- Automatic assay bias correction
- Fully conv. architecture
  - Dilated convolutions
  - Residual connections

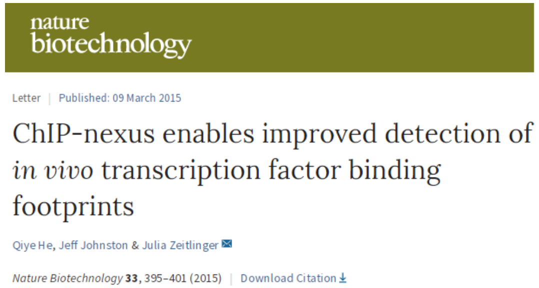
Ziga Avsec

ChIP-exo/nexus: High resolution TF binding footprints


Mouse embryonic stem cells
Oct4
Sox2
Nanog
Klf4
ChIP-nexus
5'
5'
pos
Single-nucleotide map of stop bases
neg

ChIP-nexus data for key transcription factors in mouse embryonic stem (ES) cells





nature biotechnology

Letter | Published: 09 March 2015

ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints

Qiye He, Jeff Johnston & Julia Zeitlinger

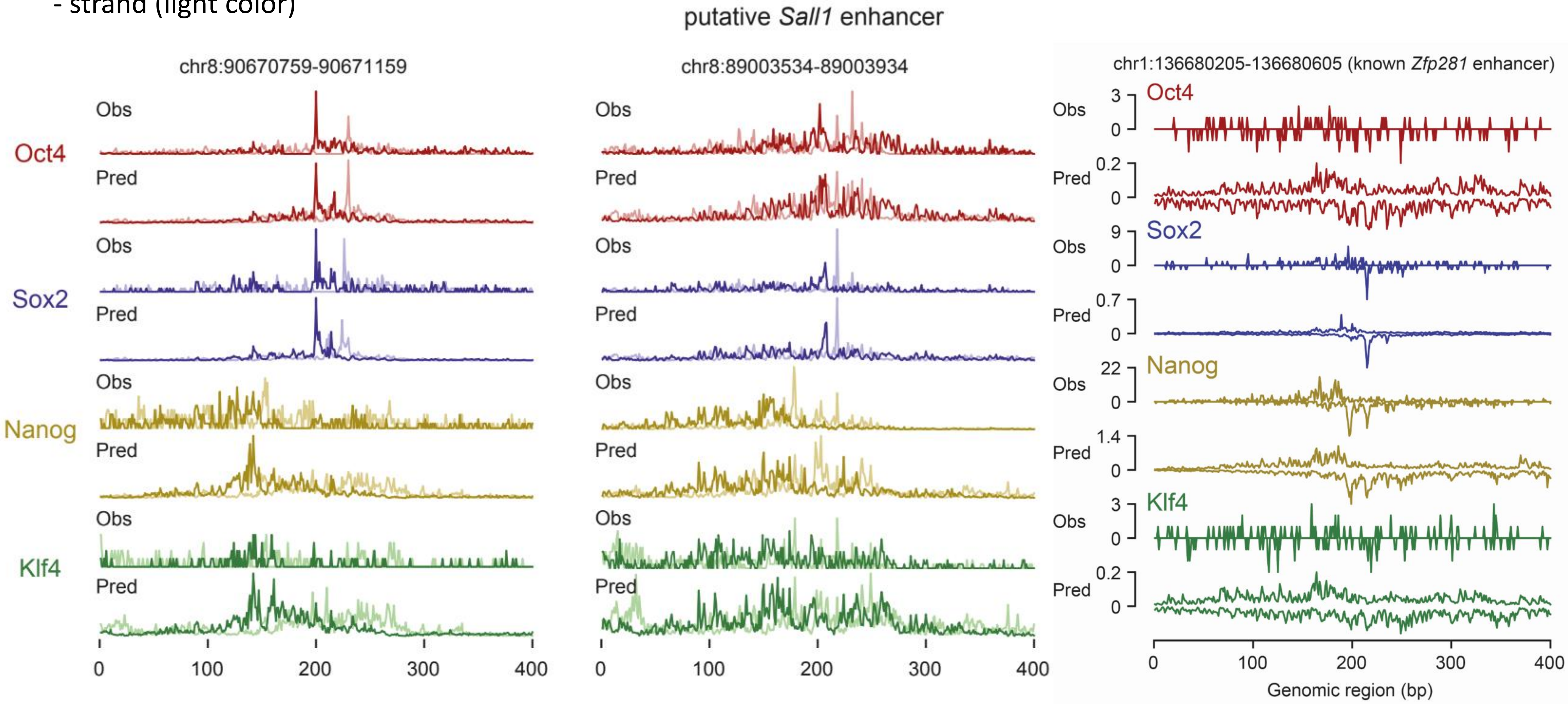
Nature Biotechnology 33, 395–401 (2015) | Download Citation

Julia Zeitlinger lab

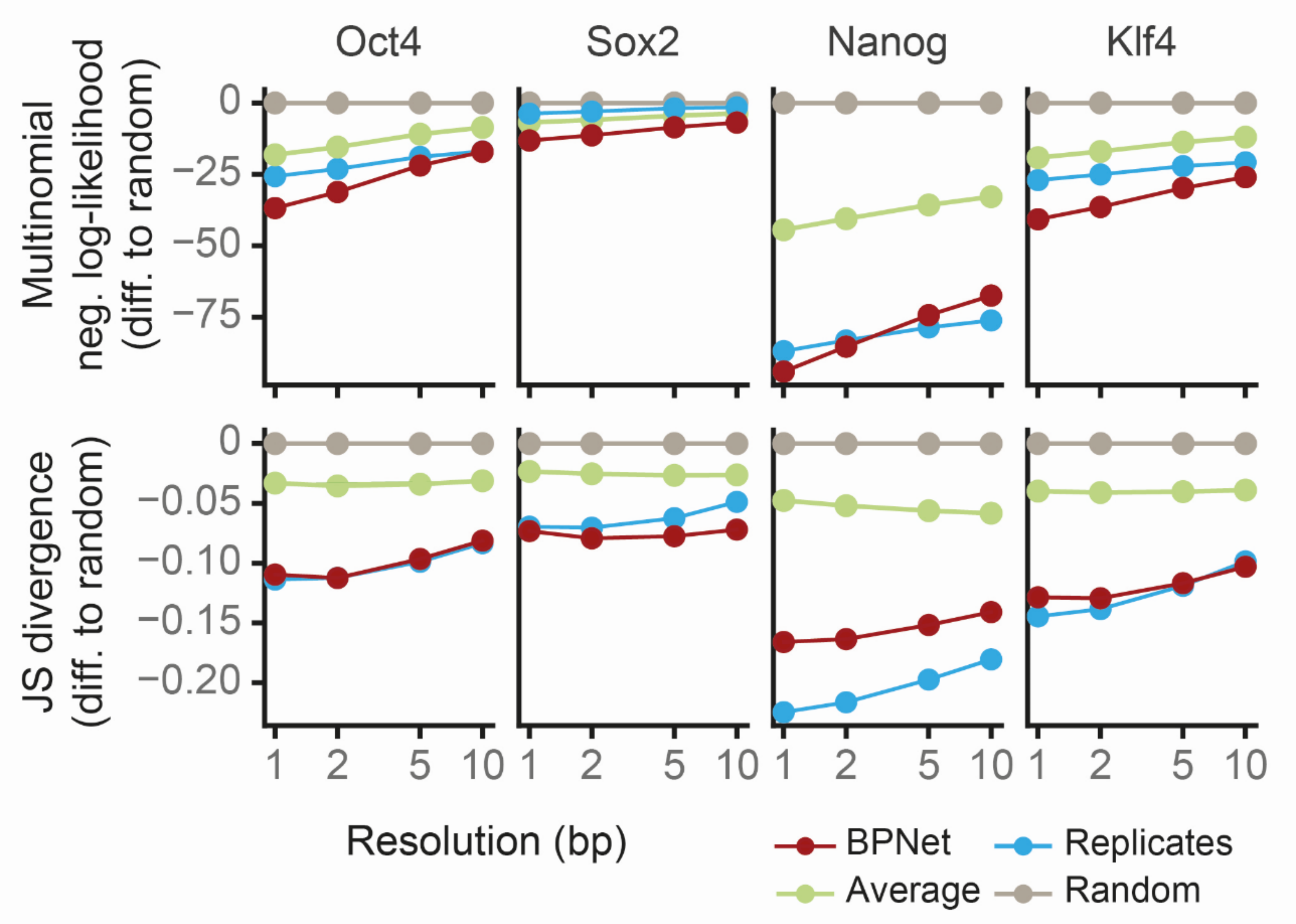
BPNet predicts base resolution binding footprints with unprecedented accuracy

+ strand (dark color)

- strand (light color)



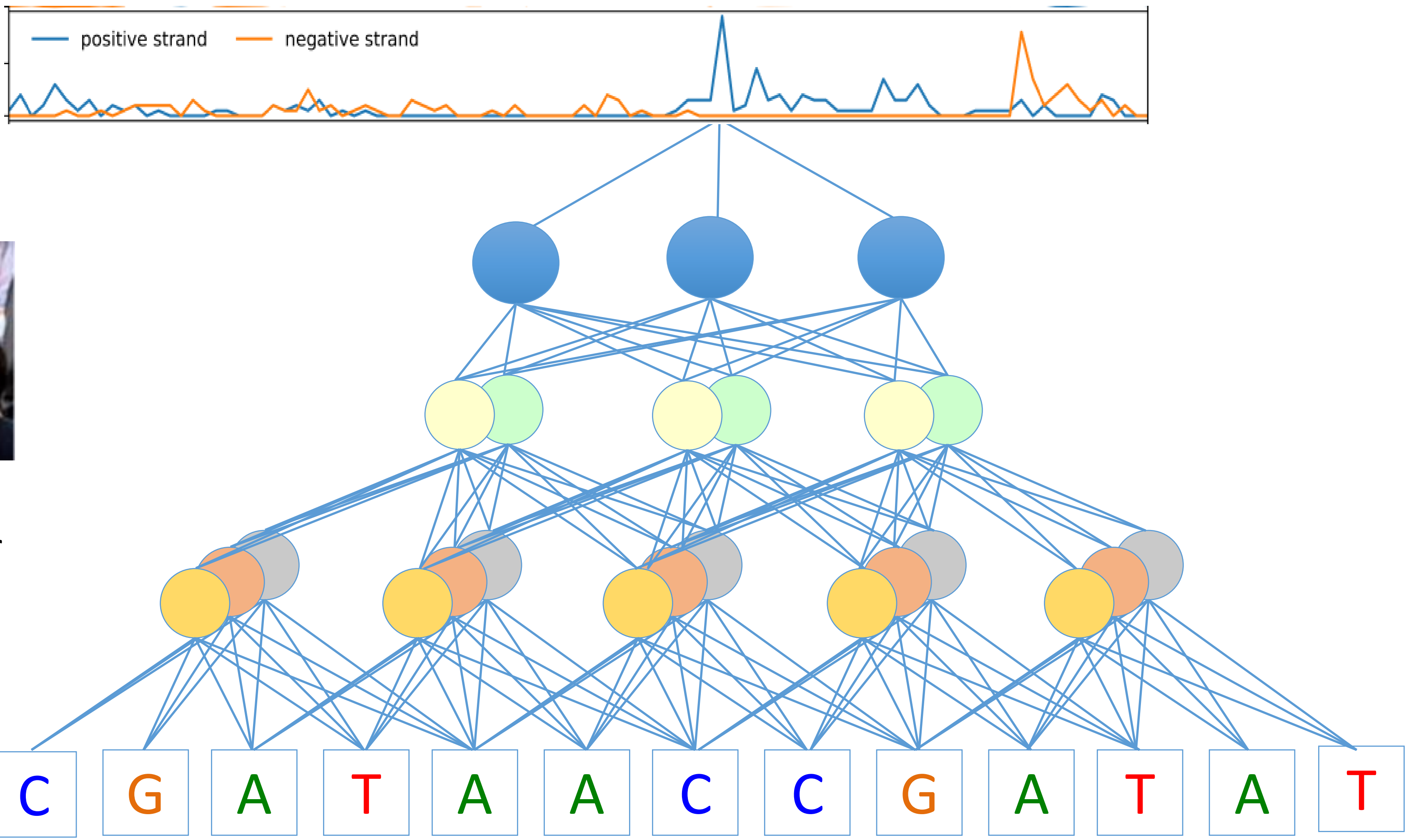
Profile prediction is on par with concordance from replicate experiments



Deciphering predictive motifs and motif instances

DeepLIFT: Inferring predictive nucleotides at individual binding events

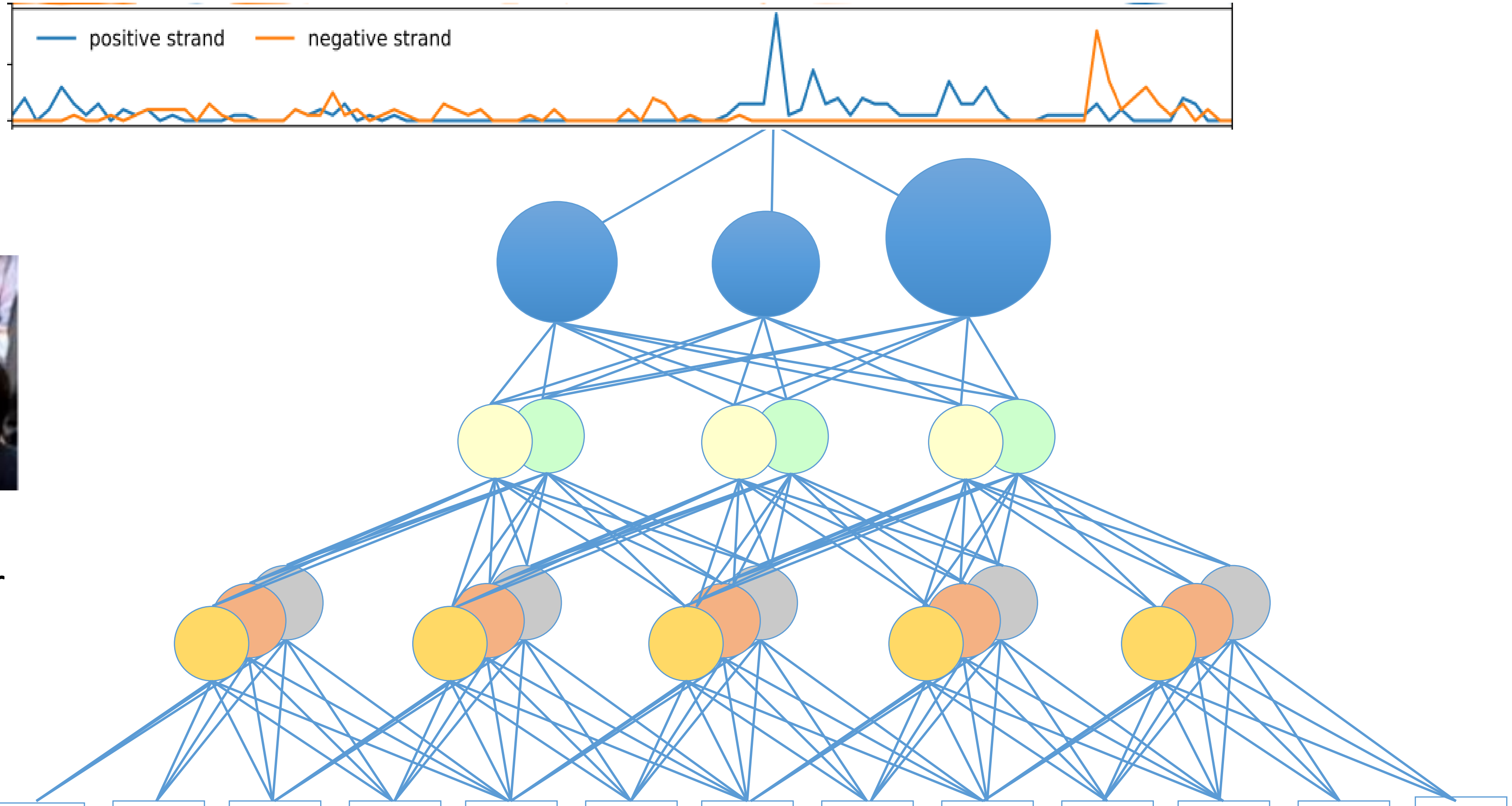
Avanti Shrikumar



A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

DeepLIFT: Inferring predictive nucleotides at individual binding events

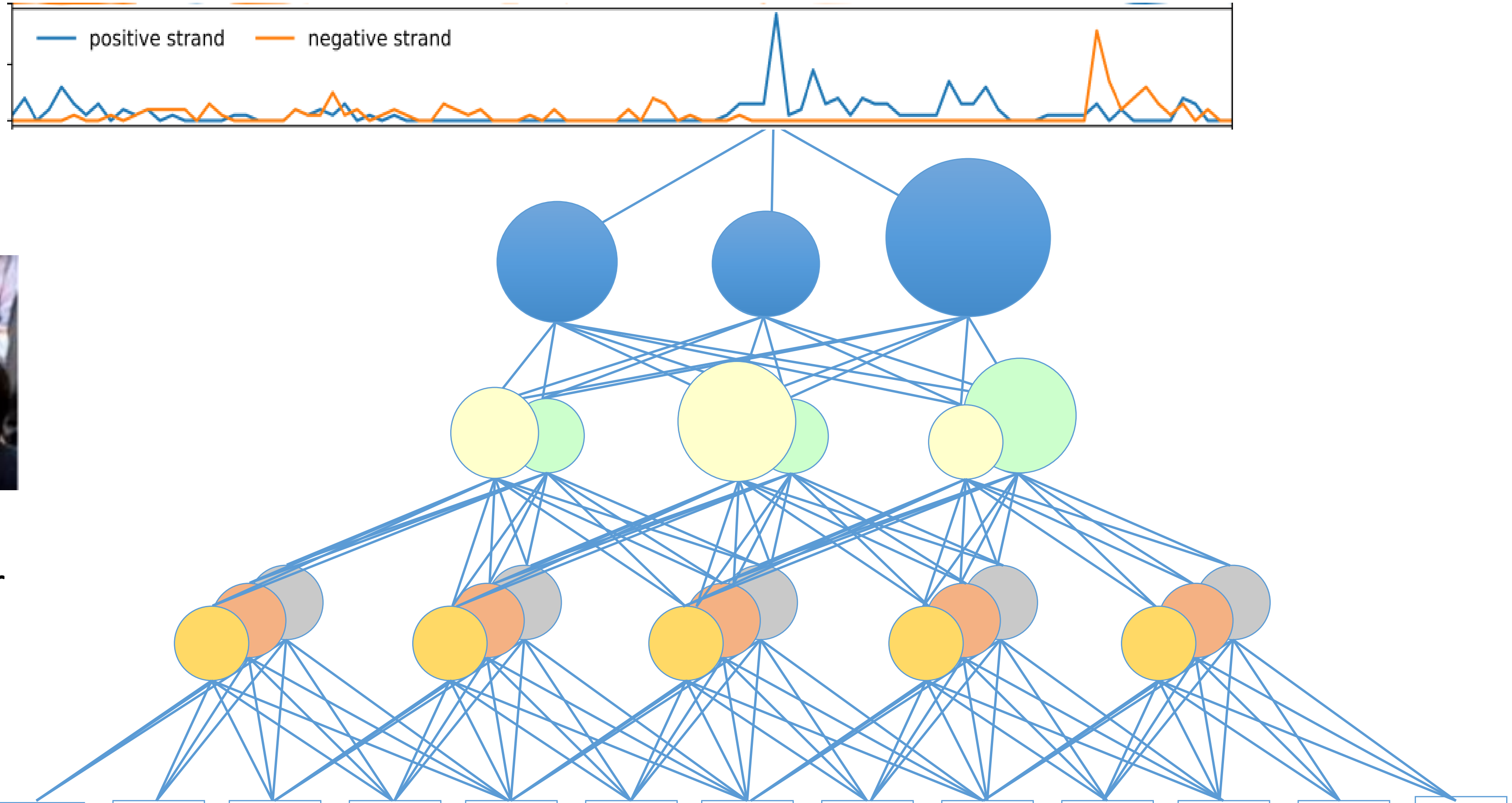
Avanti Shrikumar



	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

DeepLIFT: Inferring predictive nucleotides at individual binding events

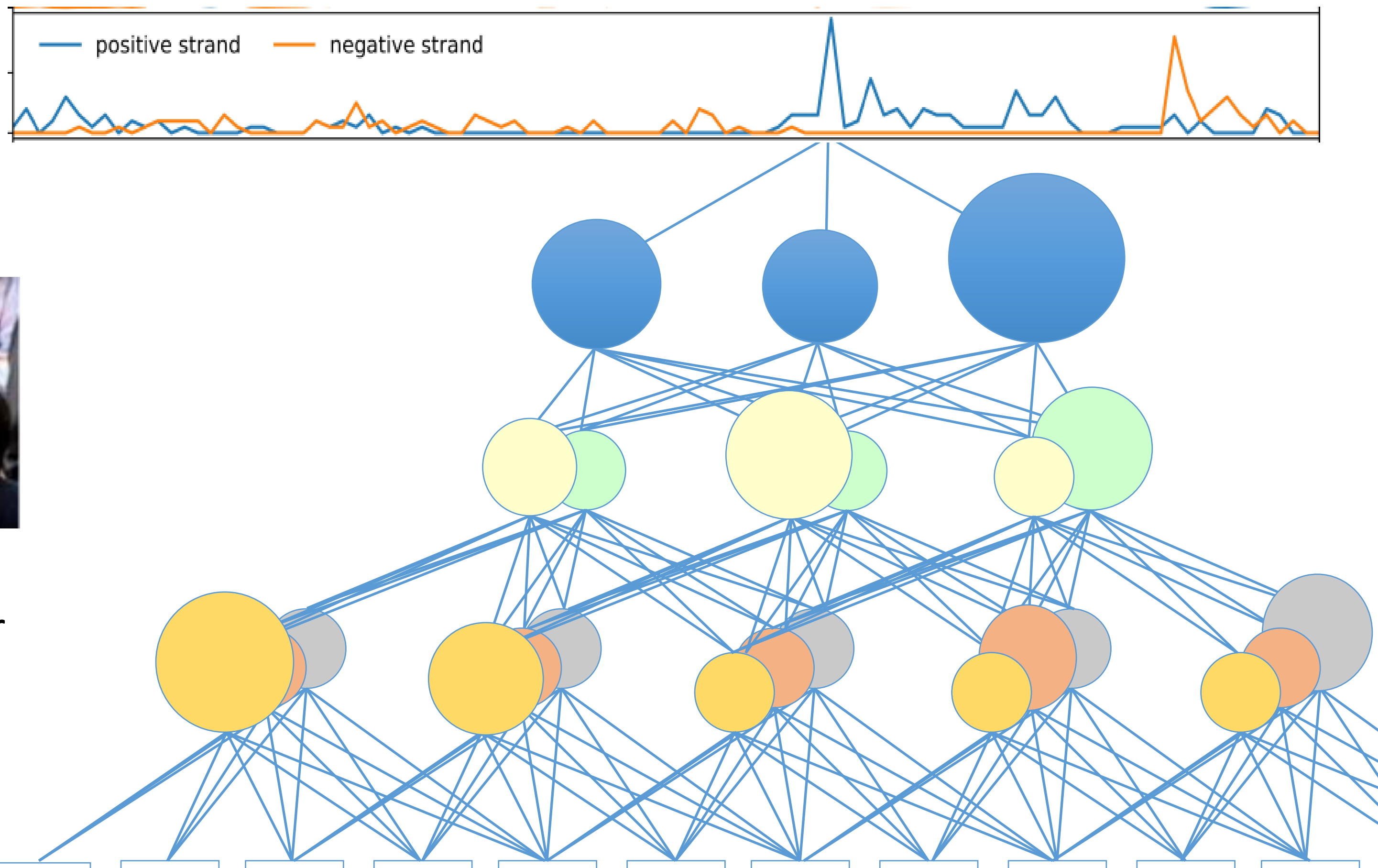
Avanti Shrikumar



	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

DeepLIFT: Inferring predictive nucleotides at individual binding events

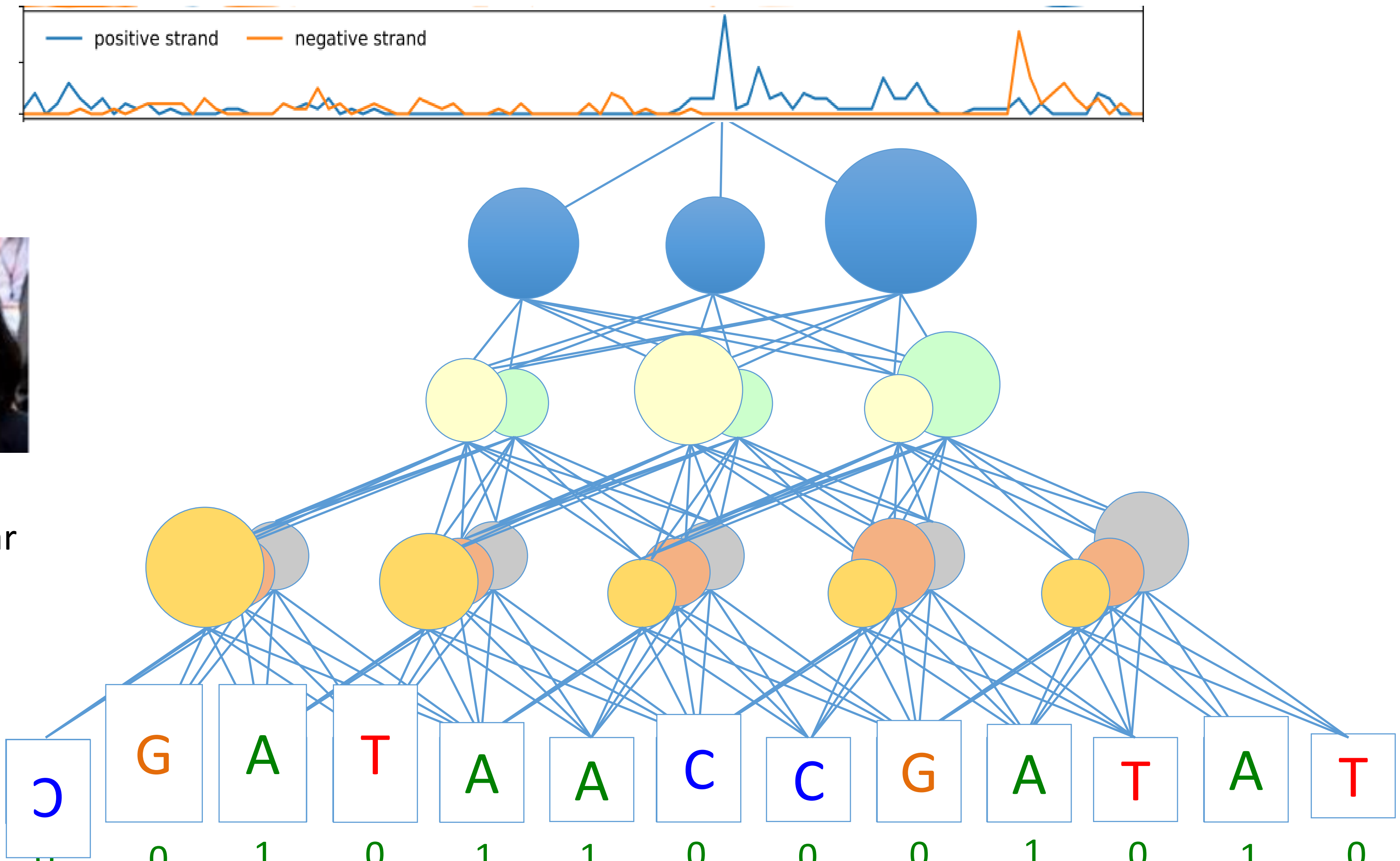
Avanti Shrikumar



	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

DeepLIFT: Inferring predictive nucleotides at individual binding events

Avanti Shrikumar

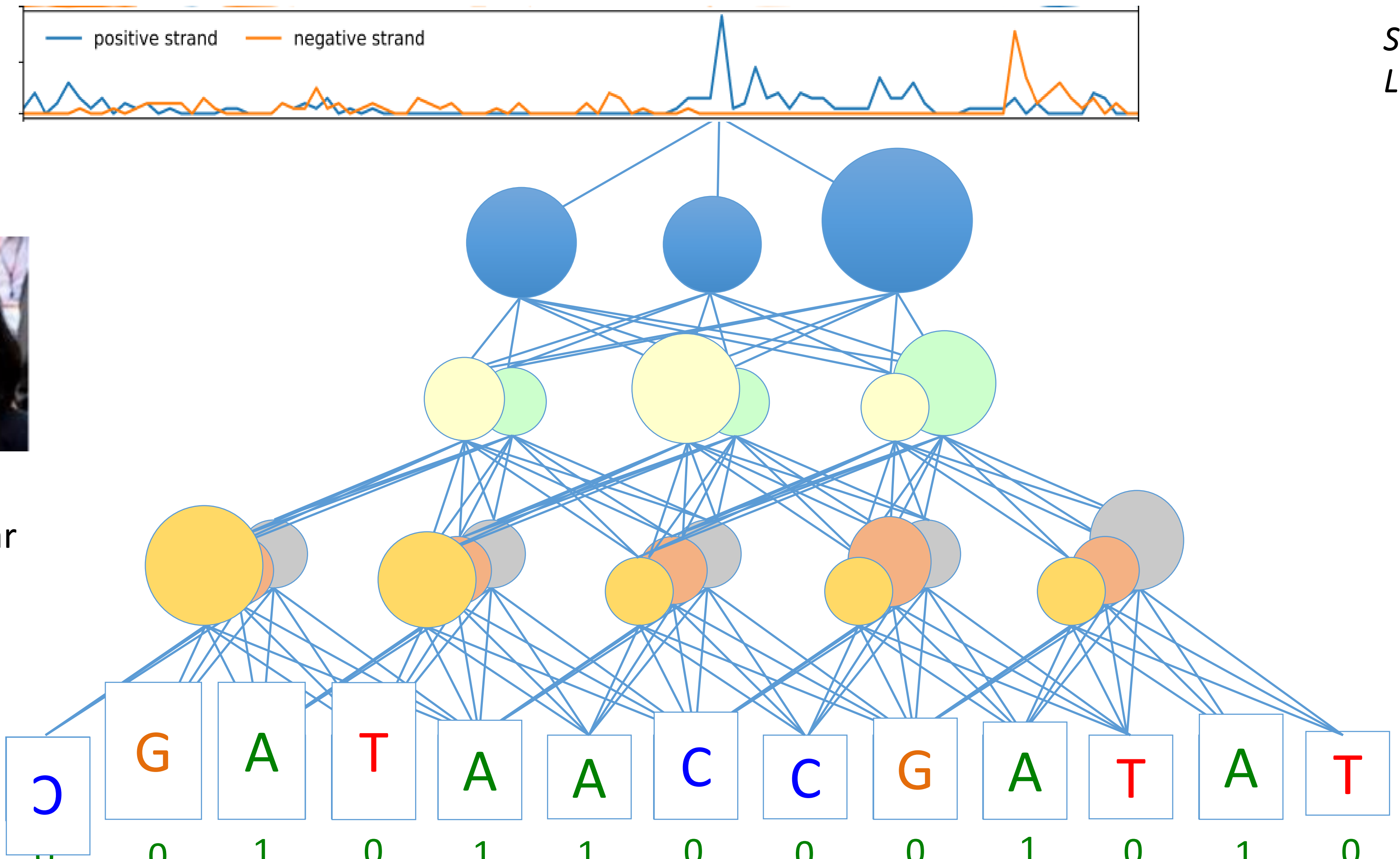


A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

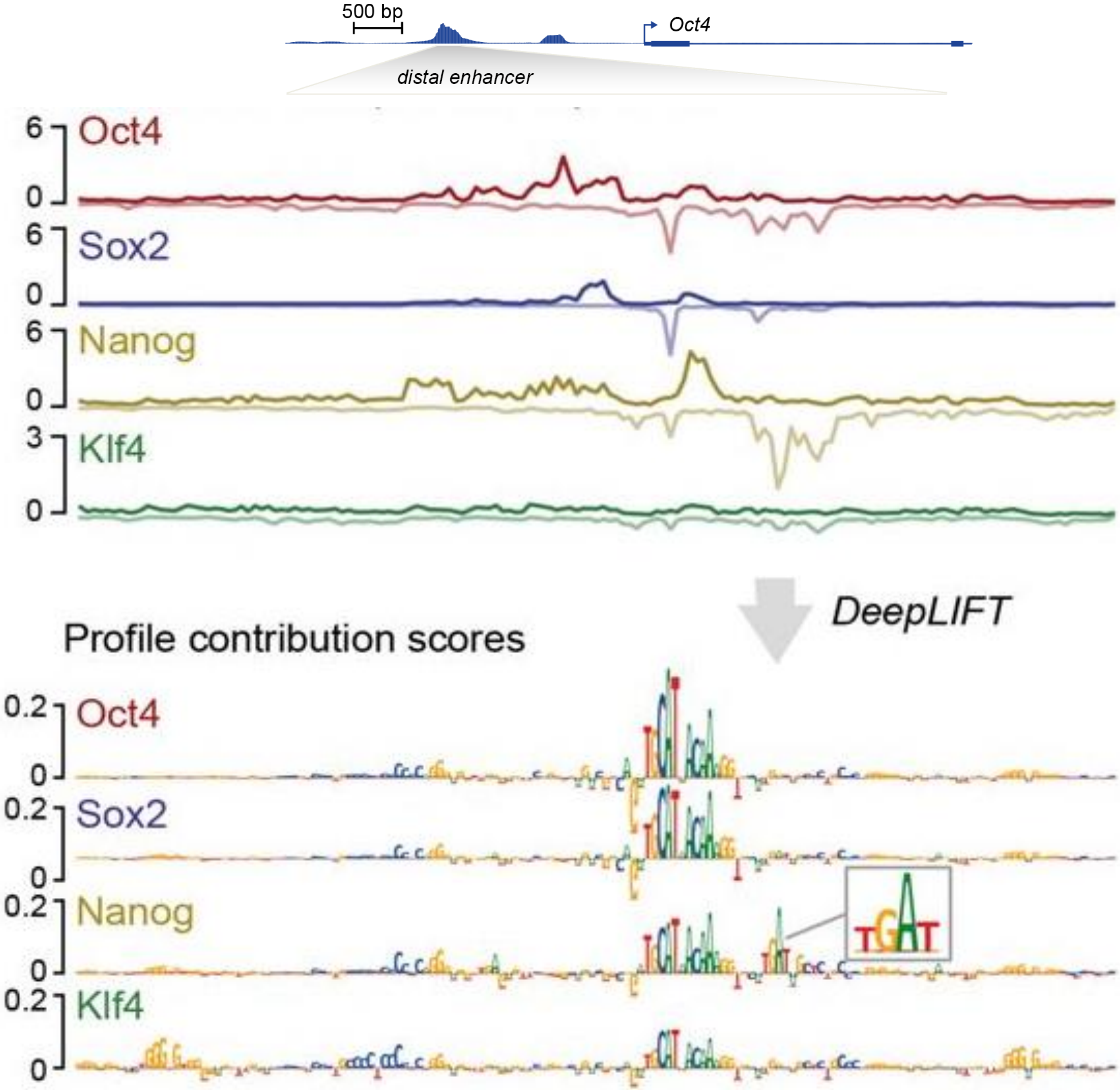
DeepLIFT: Inferring predictive nucleotides at individual binding events

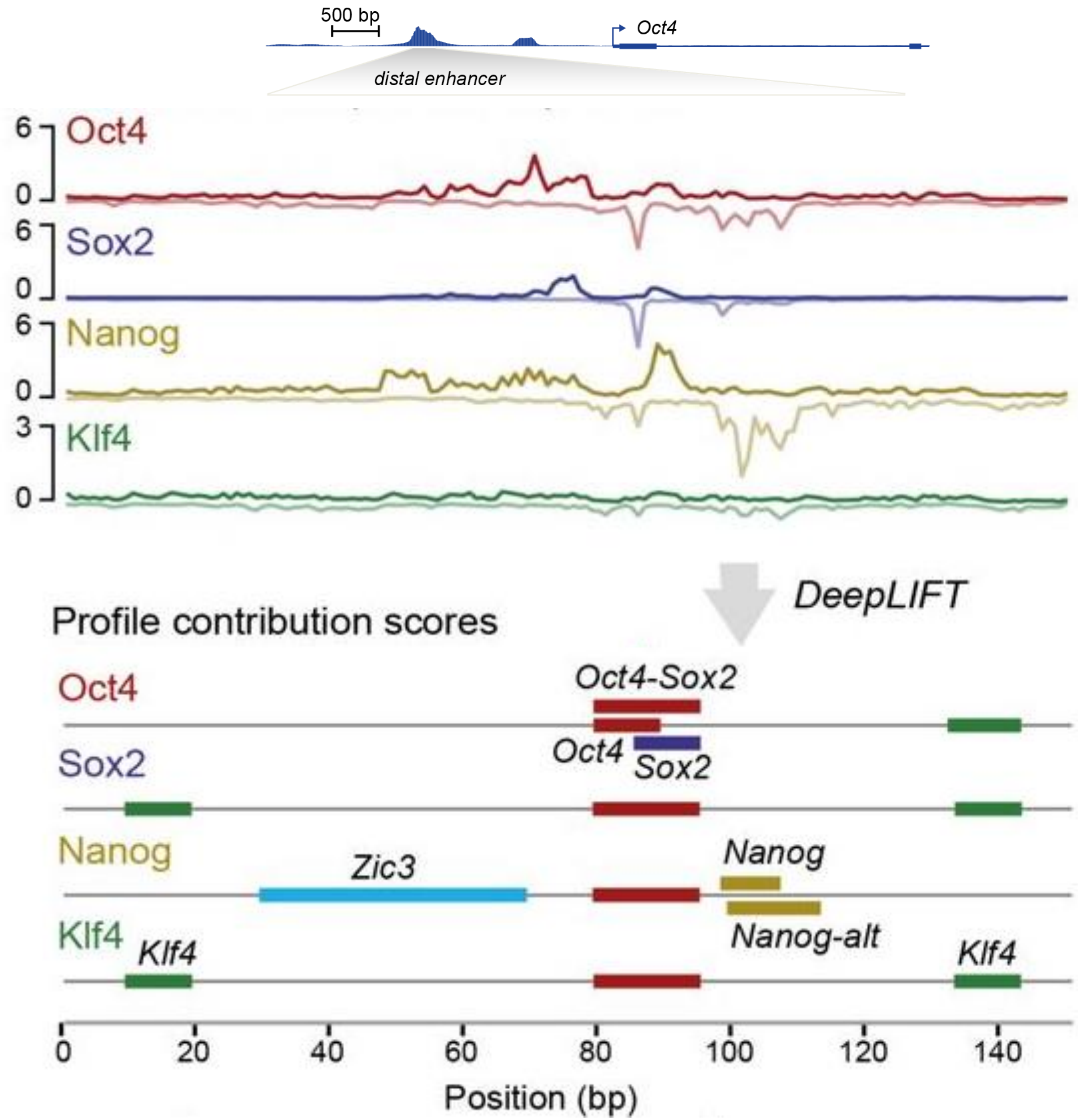
Shrikumar et al. ICML 2017
Lundberg et al. NeurIPS 2017

Avanti Shrikumar



A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1


500 bp
Oct4
distal enhancer
6
0
Oct4
6
0
Sox2
6
0
Nanog
3
0
Klf4
DeepLIFT
Profile contribution scores
0.2
0
Oct4
0.2
0
Sox2
0.2
0
Nanog
TGAT
0.2
0
Klf4


500 bp
Oct4
distal enhancer
6
0
Oct4
6
0
Sox2
6
0
Nanog
3
0
Klf4
DeepLIFT
Profile contribution scores
Oct4
Oct4-Sox2
Oct4
Sox2
Sox2
Nanog
Zic3
Nanog
Nanog-alt
Klf4
Klf4
Klf4
0
20
40
60
80
100
120
140
Position (bp)

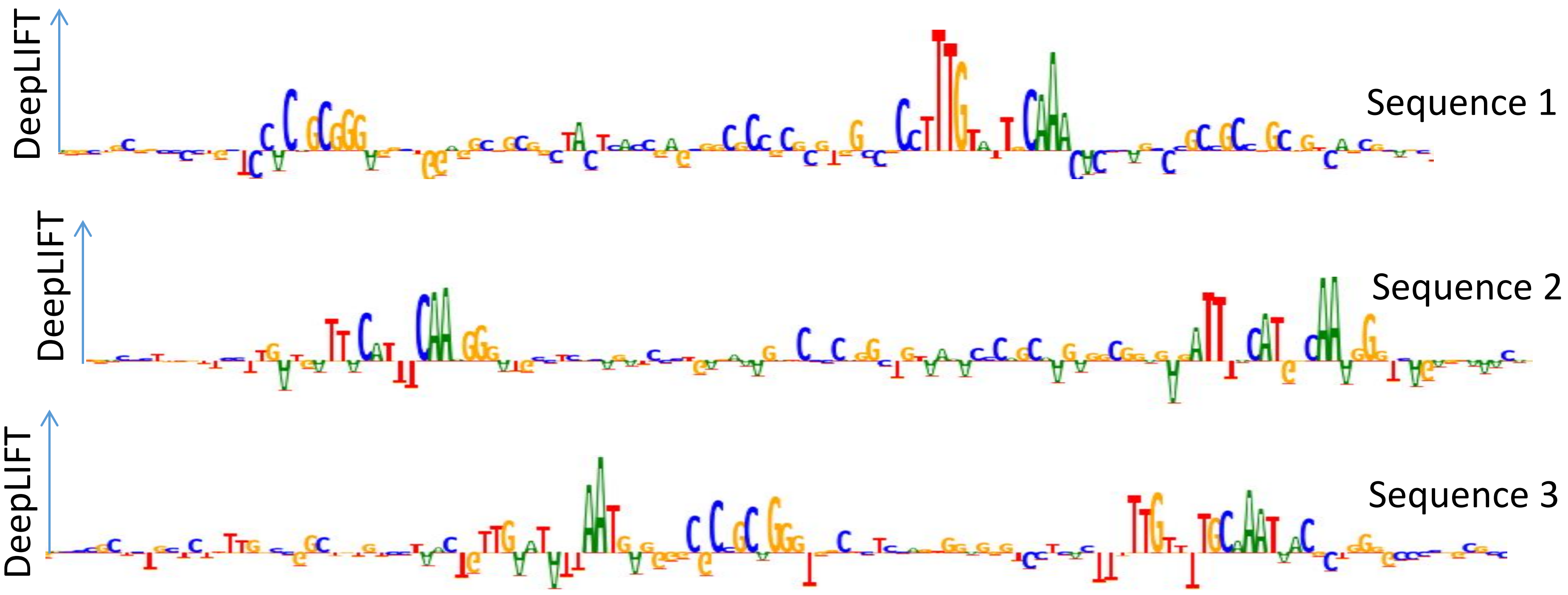
TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Shrikumar et al. 2018, arxiv

CODE: *https://github.com/kundajelab/tfmodisco*

TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Insight: conv. filter contributions are integrated at the nucleotide level

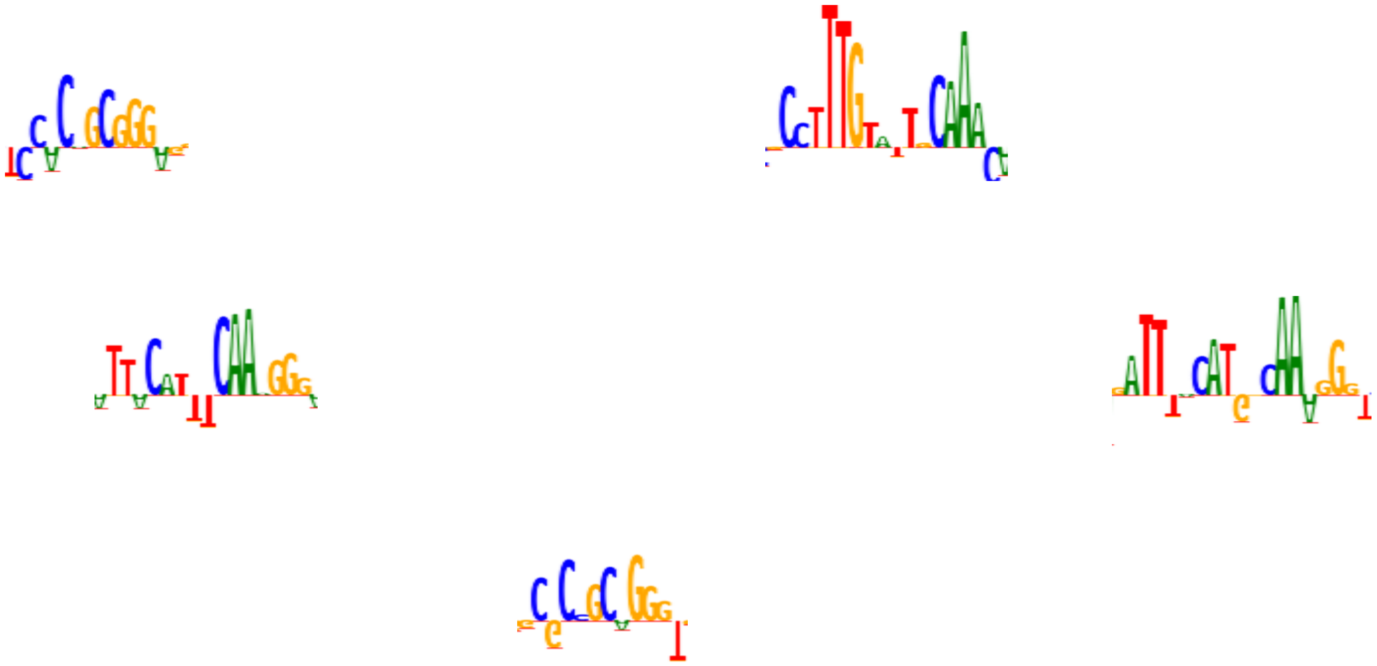


Shrikumar et al. 2018, arxiv

CODE: https://github.com/kundajelab/tfmodisco

TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Insight: conv. filter contributions are integrated at the nucleotide level

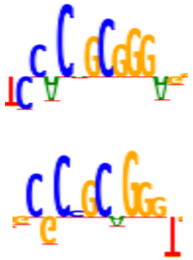


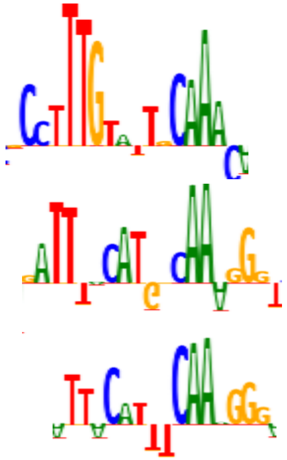
Shrikumar et al. 2018, arxiv

CODE: *https://github.com/kundajelab/tfmodisco*

TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Insight: conv. filter contributions are integrated at the nucleotide level



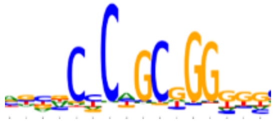


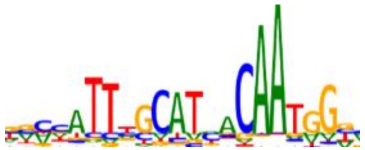
Shrikumar et al. 2018, arxiv

CODE: https://github.com/kundajelab/tfmodisco

TF-MoDISCO: Cluster and consolidate predictive subsequences into contribution weight matrix (CWM) motifs

Insight: conv. filter contributions are integrated at the nucleotide level

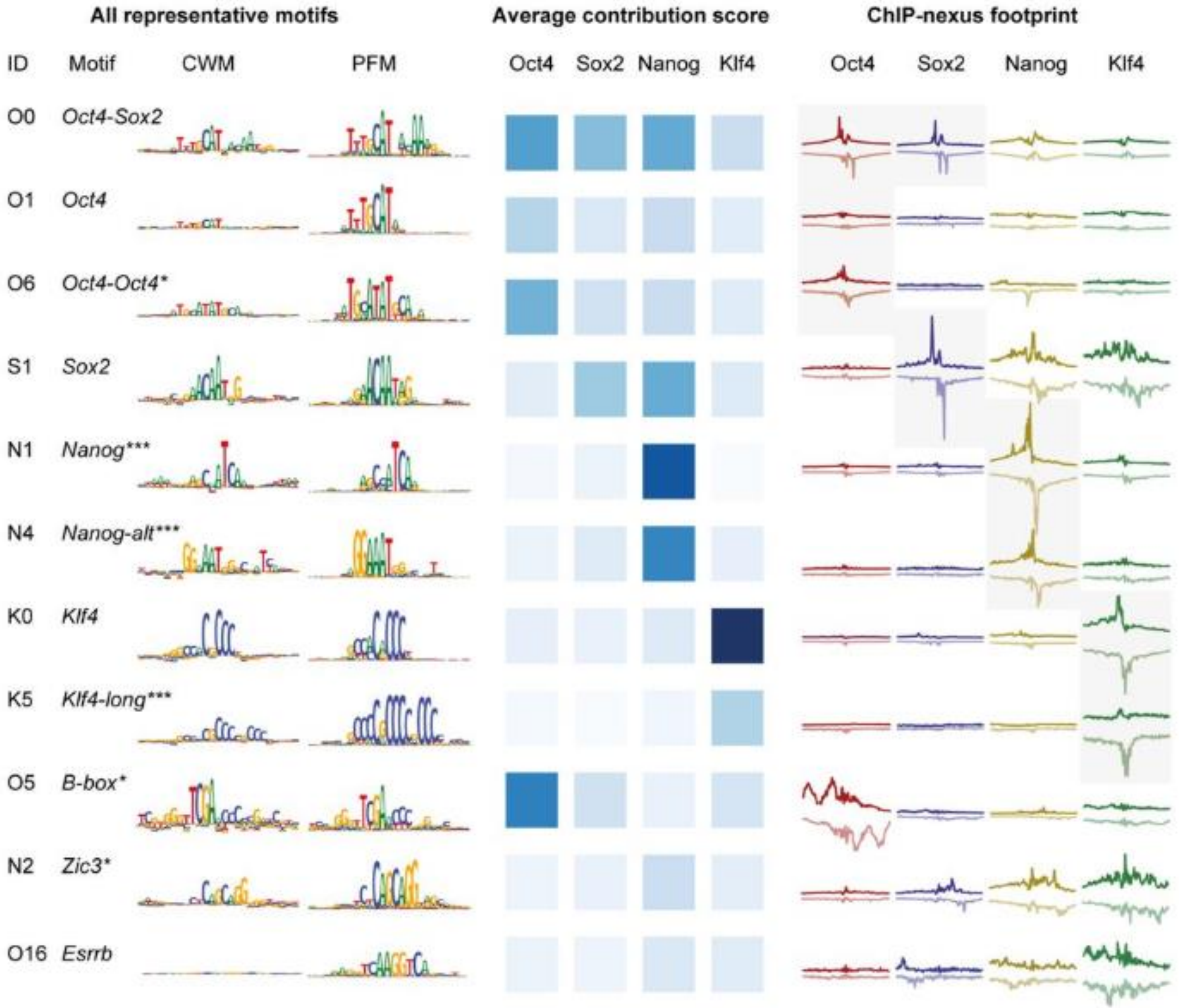




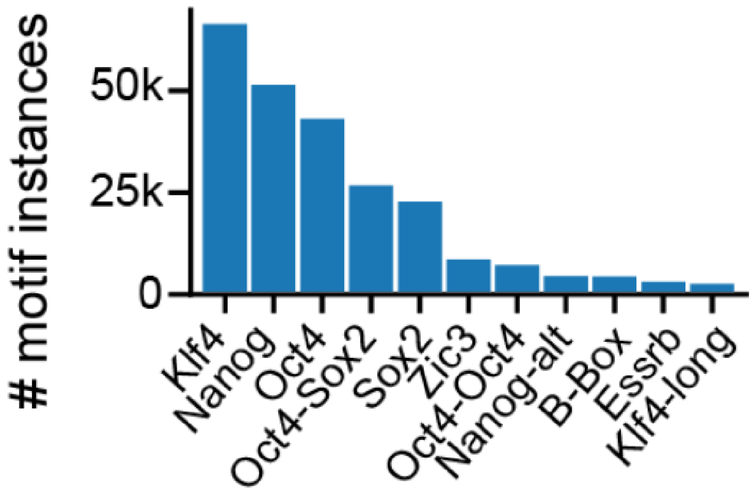
Shrikumar et al. 2018, arxiv

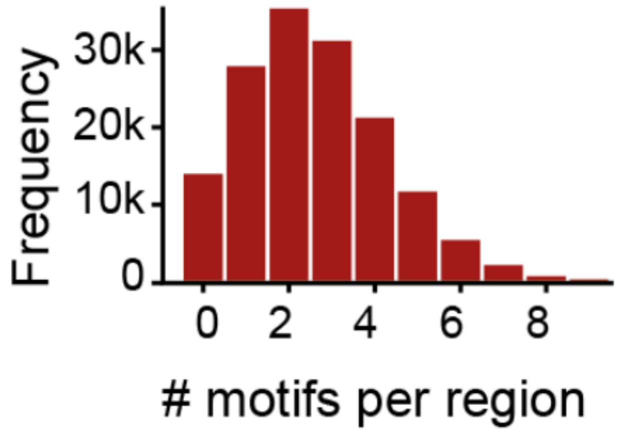
CODE: *https://github.com/kundajelab/tfmodisco*

Consolidated motifs with combinatorial footprints

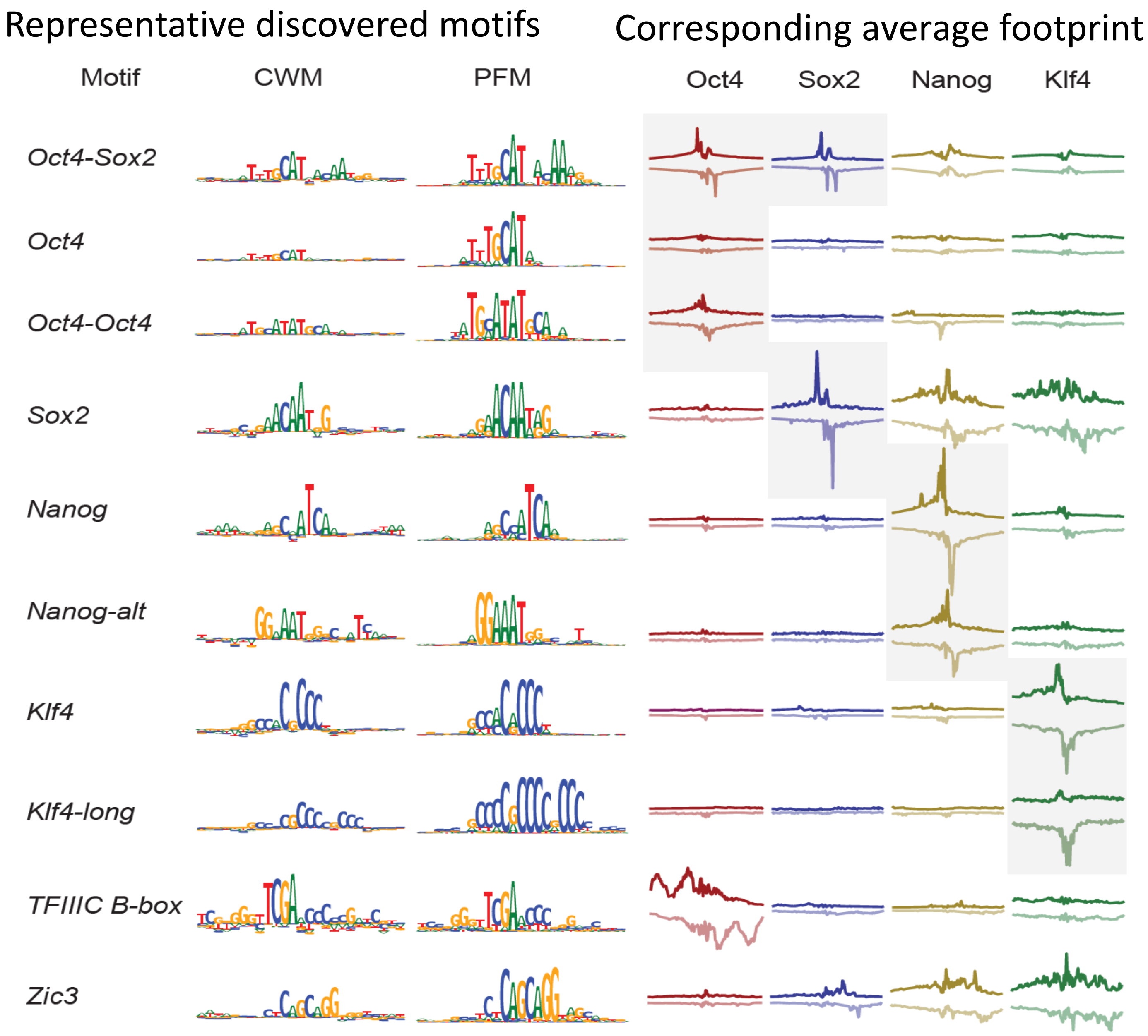


50 motifs for 4 TFs



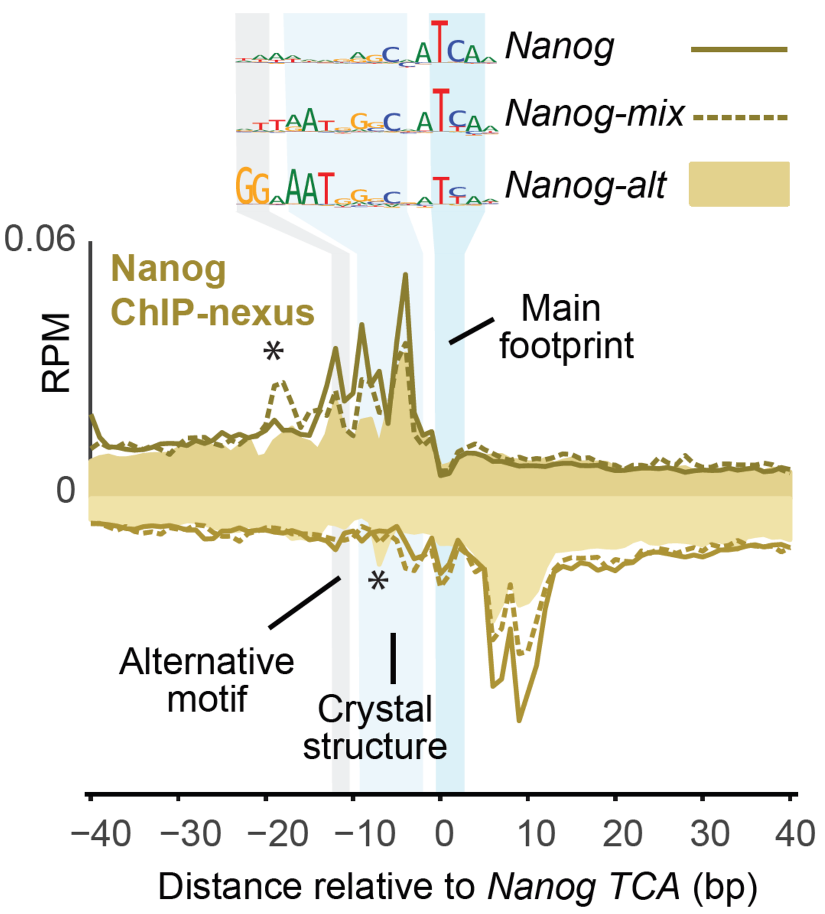


Multiple binding motifs for Nanog



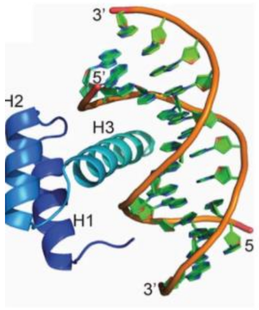
50 motifs for 4 TFs

Subtle differences in Nanog motifs

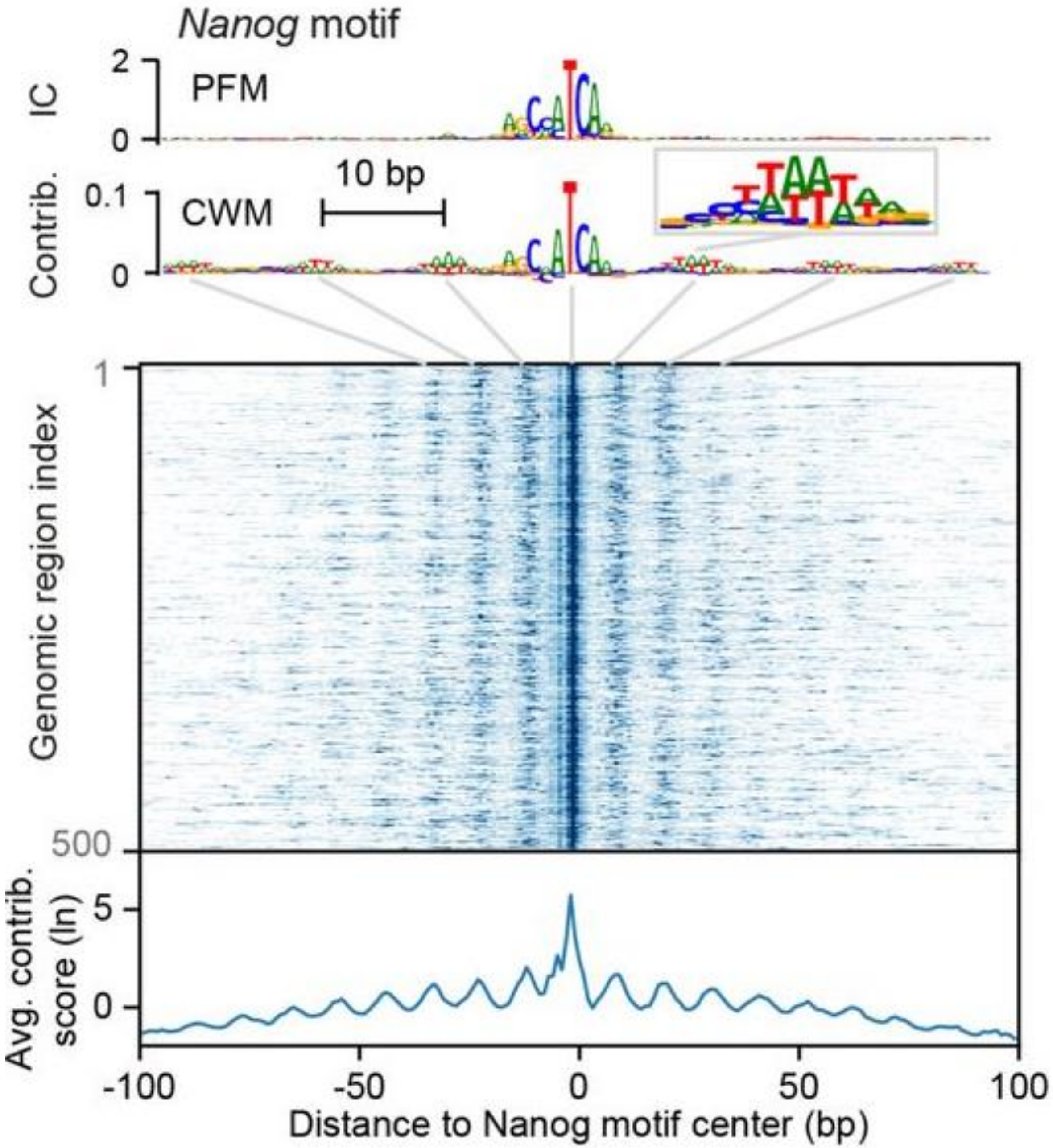


Deciphering motif syntax derived TF cooperativity

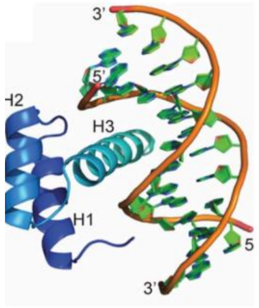
10.5 bp helical periodic flanking pattern for Nanog



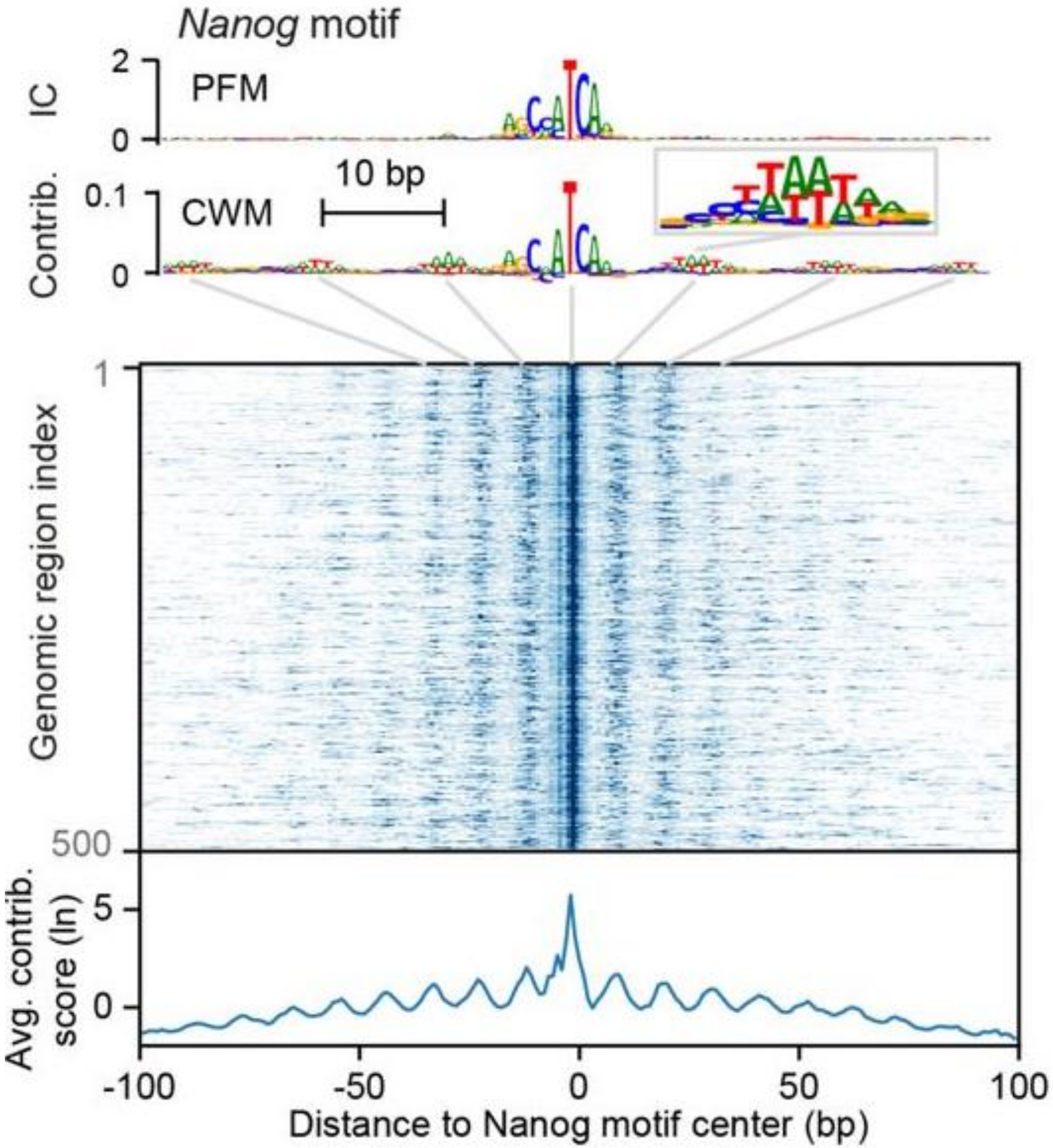
Nanog homeodomain
Hayakshi et al. PNAS 2015

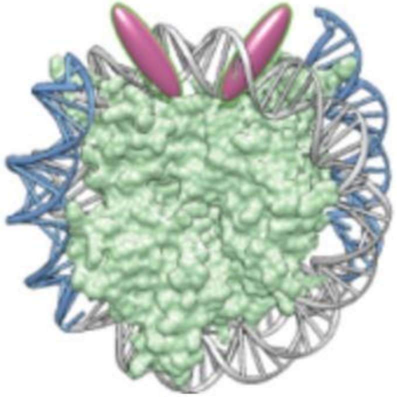


10.5 bp helical periodic flanking pattern for Nanog



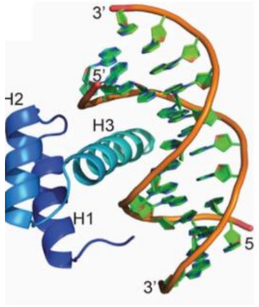
Nanog homeodomain
Hayakshi et al. PNAS 2015



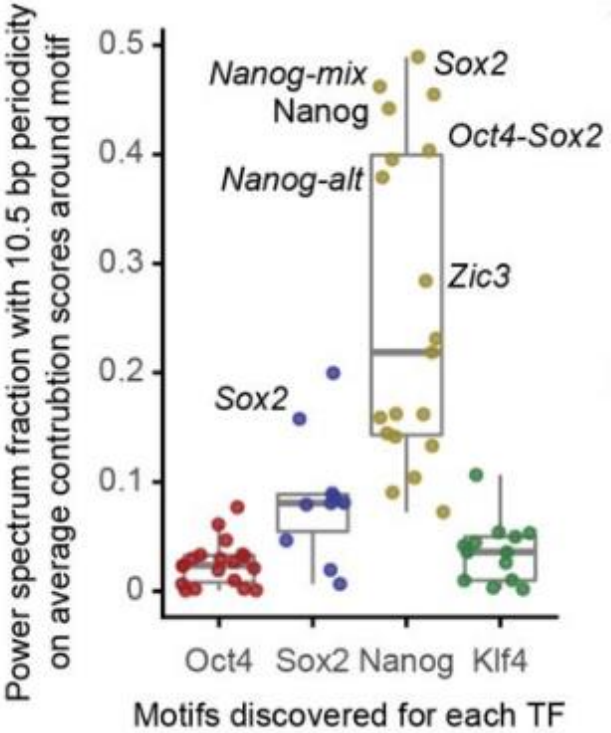


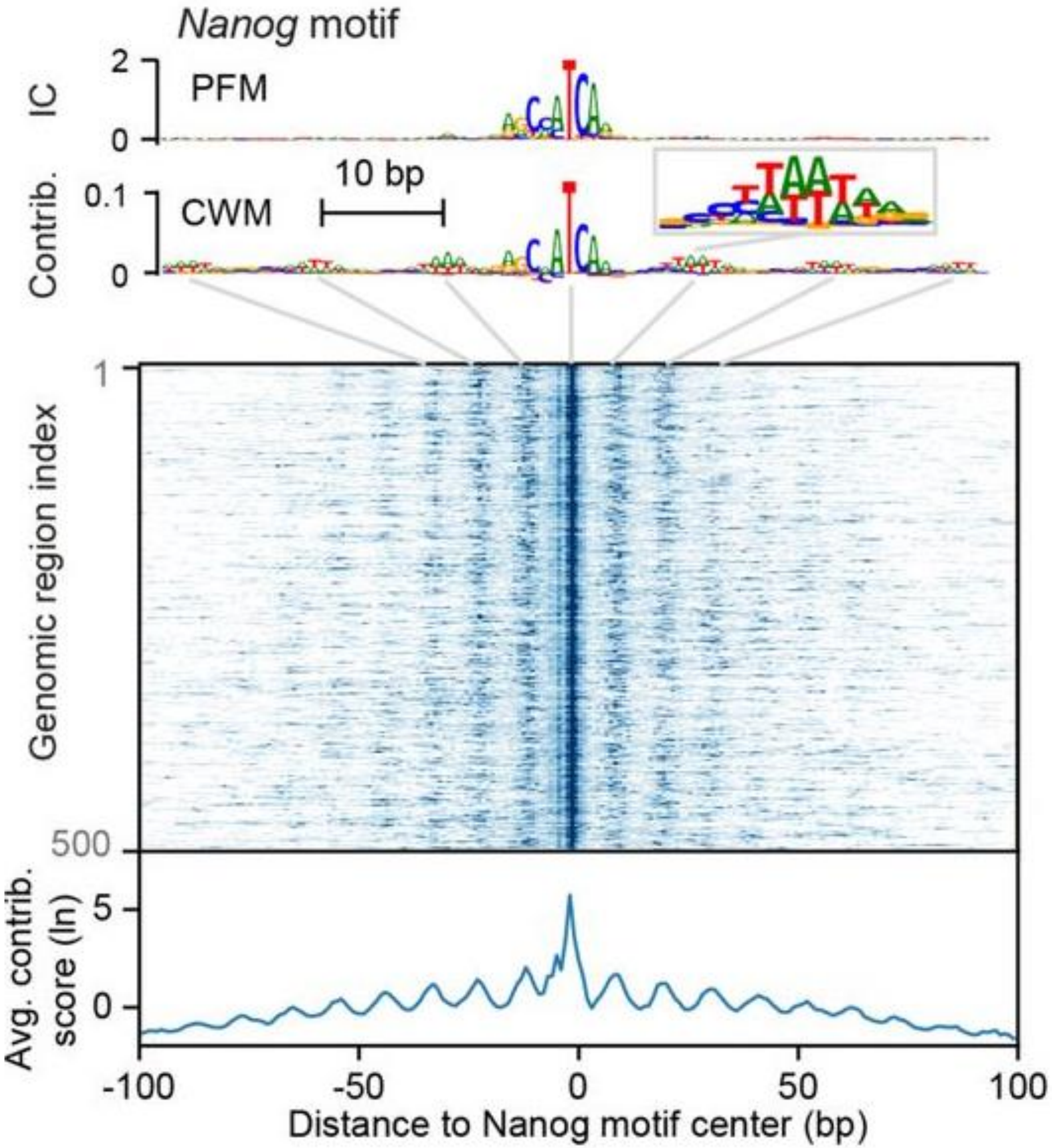
10 bp periodic binding of homeobox TFs to nucleosome DNA
from recent *in vitro* NCAP-SELEX data
(Zhu et al. Nature 2018)

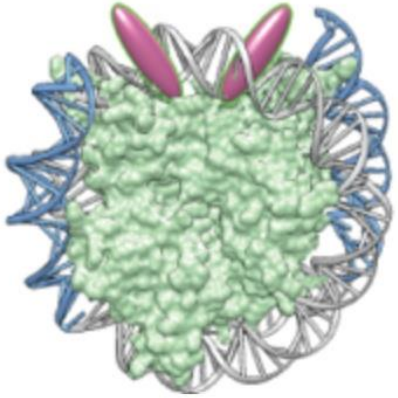
10.5 bp helical periodic flanking pattern for Nanog



Nanog homeodomain
Hayakshi et al. PNAS 2015

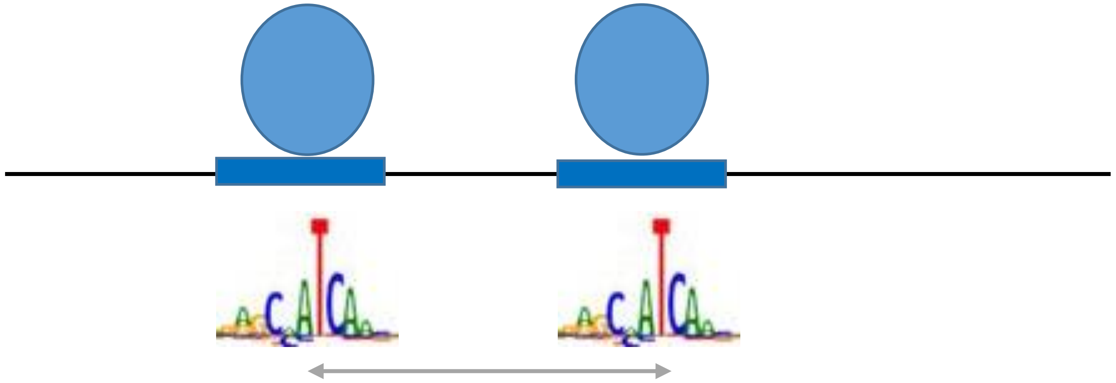


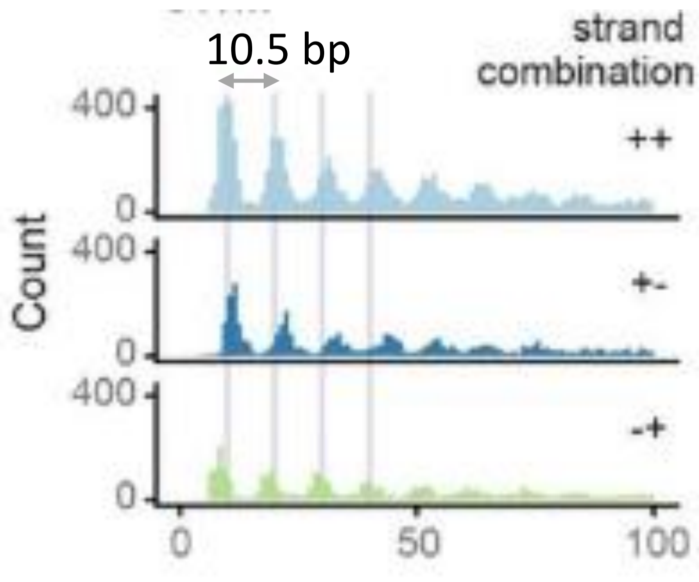




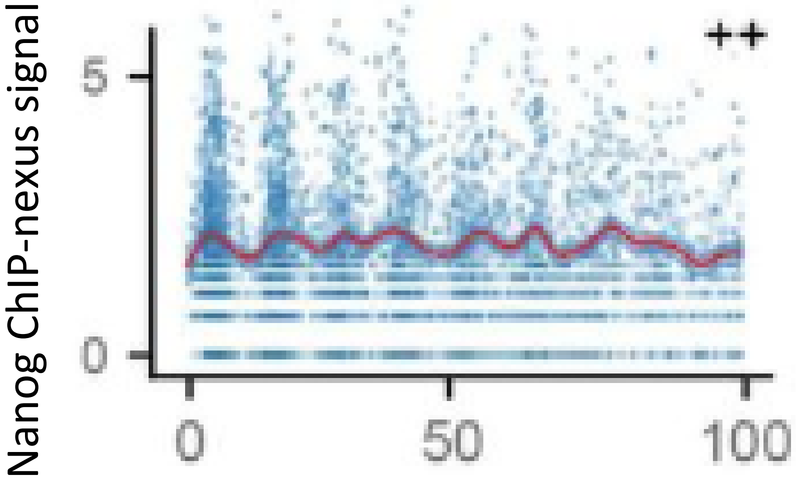
10 bp periodic binding of homeobox TFs to nucleosome DNA
from recent *in vitro* NCAP-SELEX data
(Zhu et al. Nature 2018)

Soft syntax: helical spacing preference between Nanog motifs across all control elements





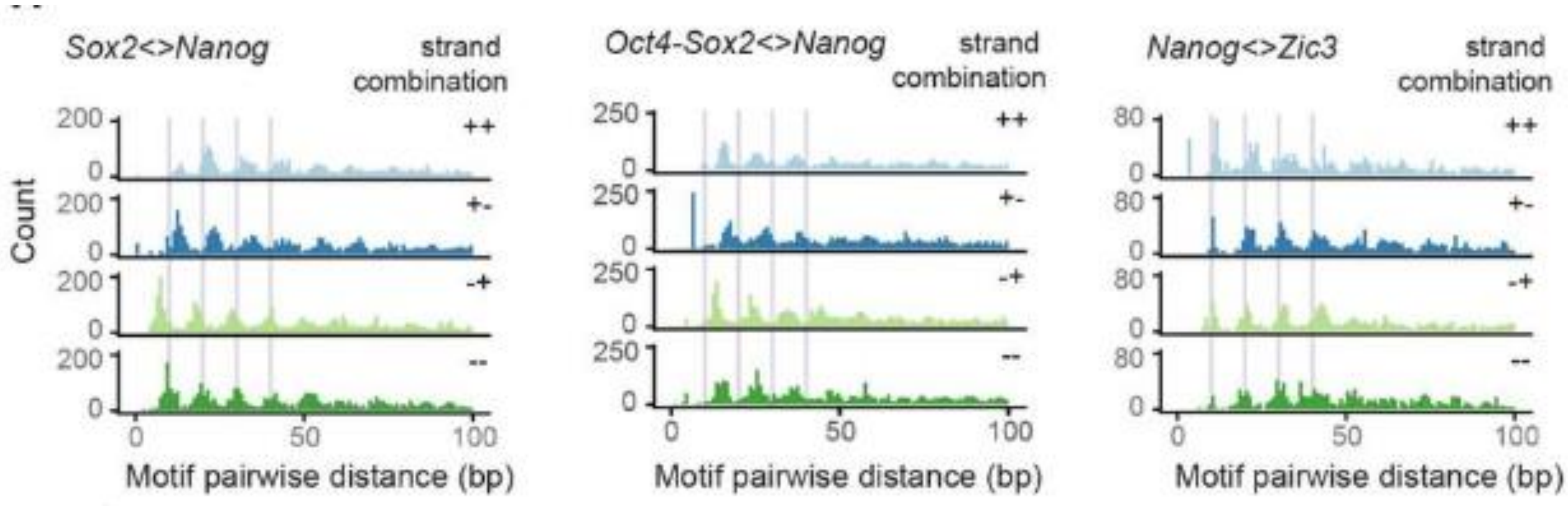
Motif pairwise distance



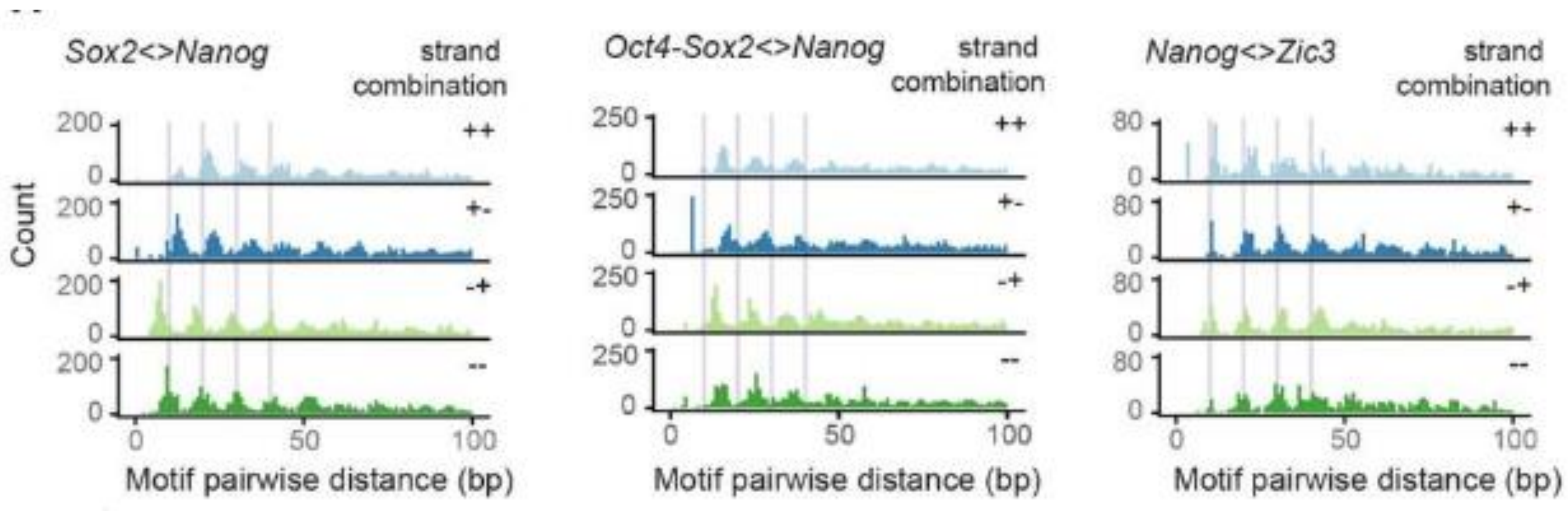
Motif pairwise distance

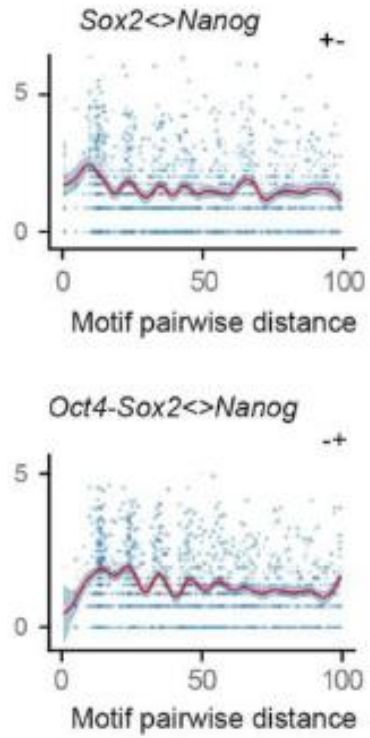
Preferred soft helical spacing preferences between Nanog <> other

Preferred soft helical spacing preferences between Nanog <> other



Preferred soft helical spacing preferences between Nanog <> other





Can we infer “causal” directional cooperative influence of different proteins via motif syntax?

Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

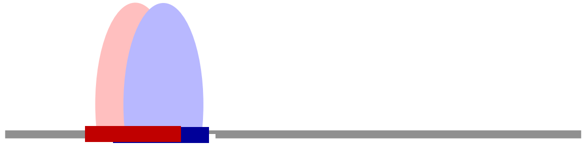
Can we infer “causal” directional cooperative influence of different proteins via motif syntax?

Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

Can we infer “causal” directional cooperative influence of different proteins via motif syntax?

Use BPNet model as in-silico oracle to perform perturbation experiments



1) On synthetic sequences

2) By mutating motifs in genomic regions

Can we infer “causal” directional cooperative influence of different proteins via motif syntax?

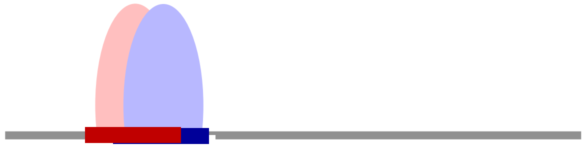
Use BPNet model as in-silico oracle to perform perturbation experiments

1) On synthetic sequences

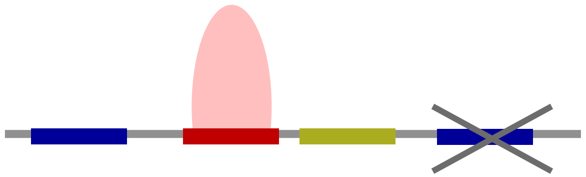
2) By mutating motifs in genomic regions

Can we infer "causal" directional cooperative influence of different proteins via motif syntax?

Use BPNet model as in-silico oracle to perform perturbation experiments



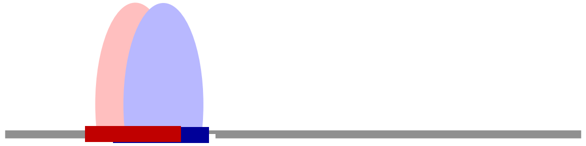
1) On synthetic sequences

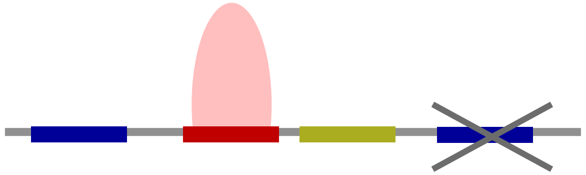


2) By mutating motifs in genomic regions

Can we infer "causal" directional cooperative influence of different proteins via motif syntax?

Use BPNet model as in-silico oracle to perform perturbation experiments





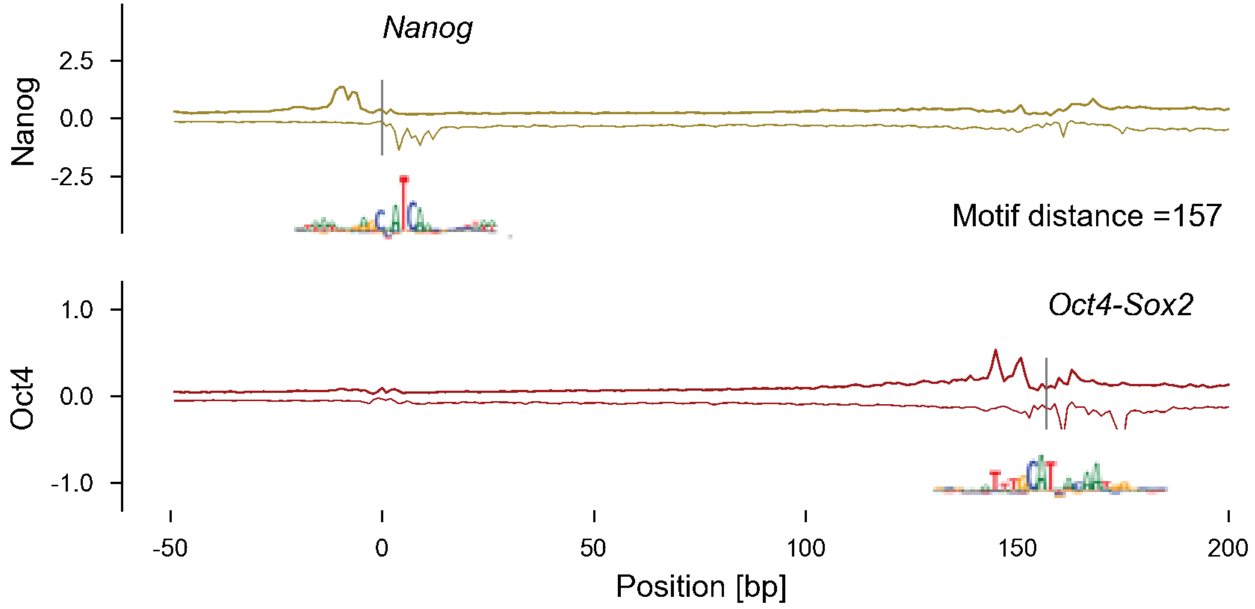
1) On synthetic sequences

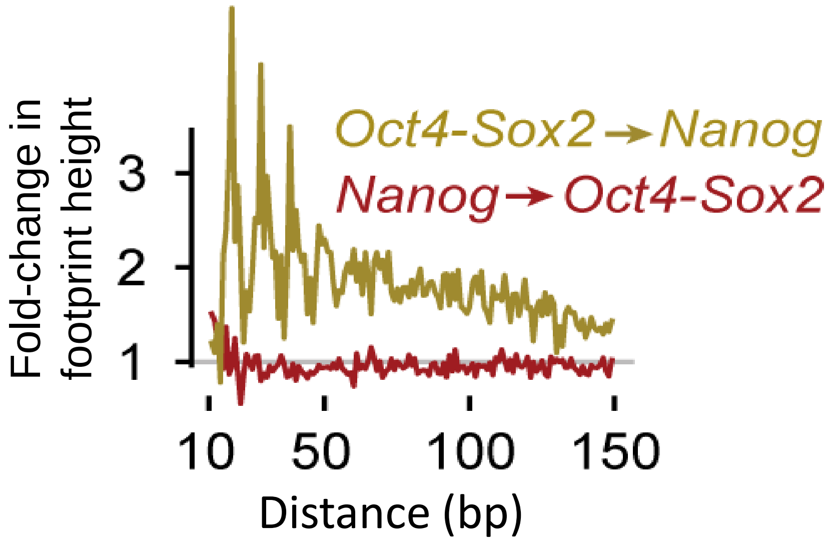
2) By mutating motifs in genomic regions

***In silico* biochemistry**

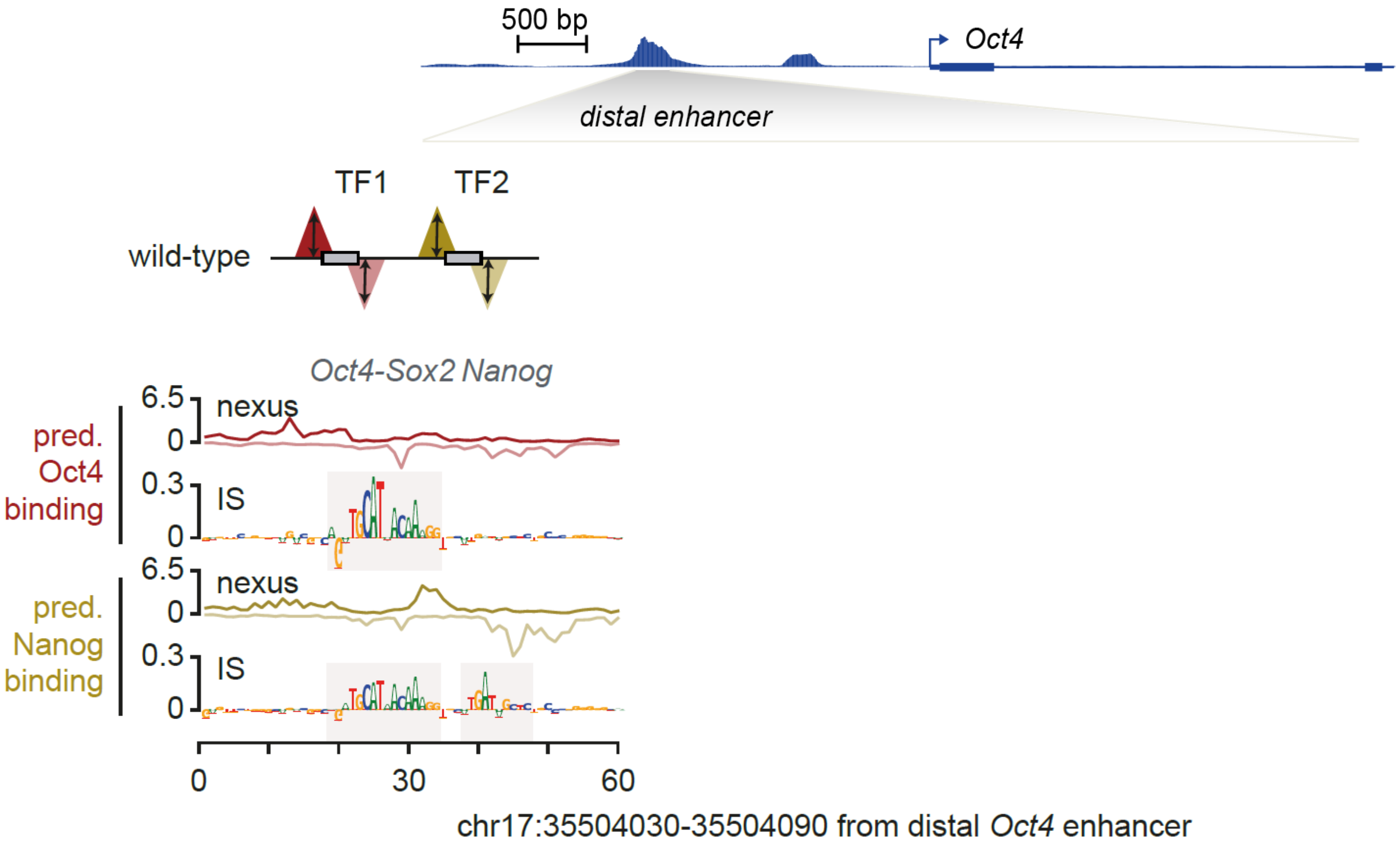
***In silico* genetics**

Cooperative interactions between Oct4 and Nanog as a function of motif spacing using synthetic sequences

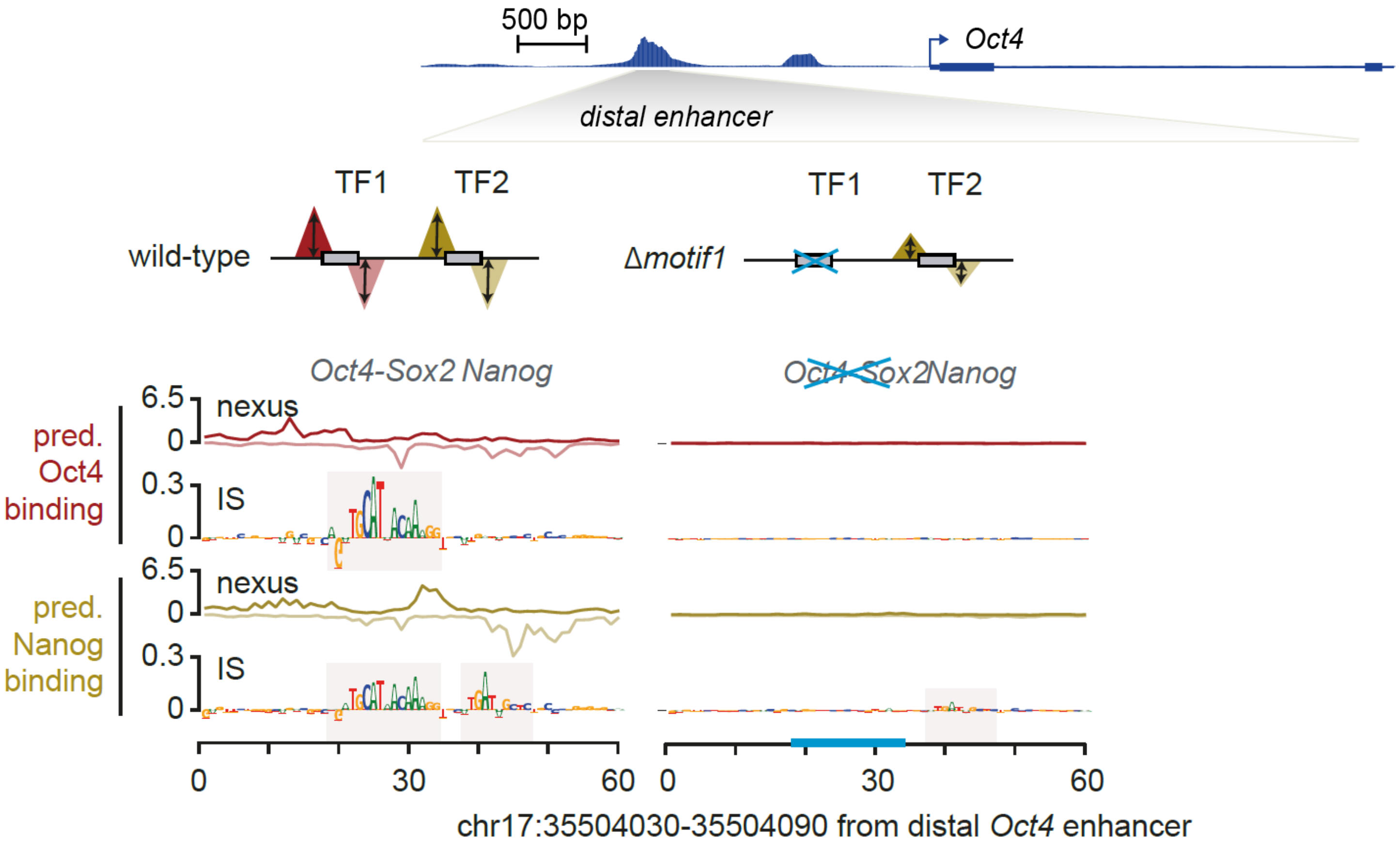




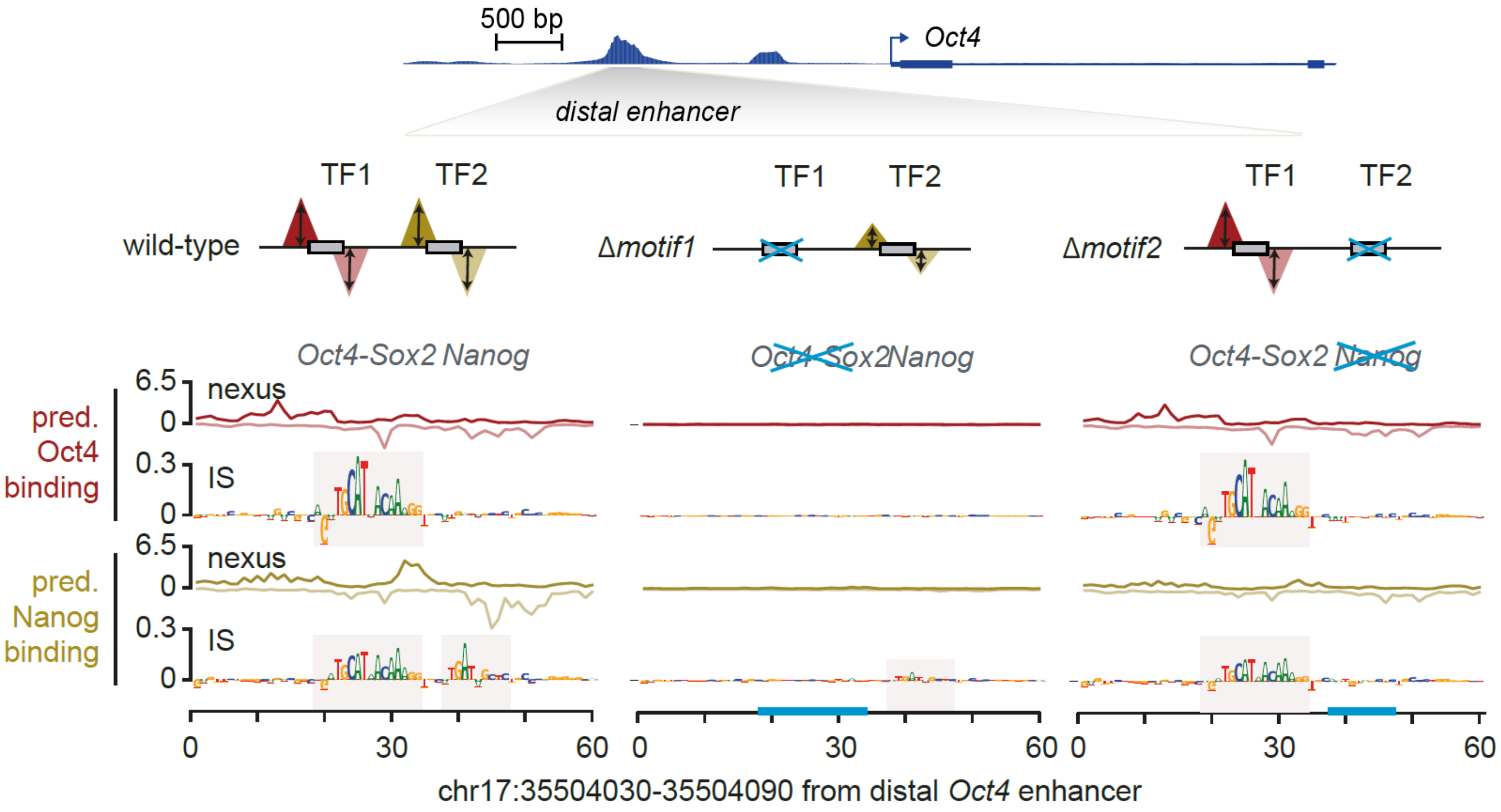
Motif syntax: cooperative TF interactions in genomic enhancers (*in-silico* CRISPR)



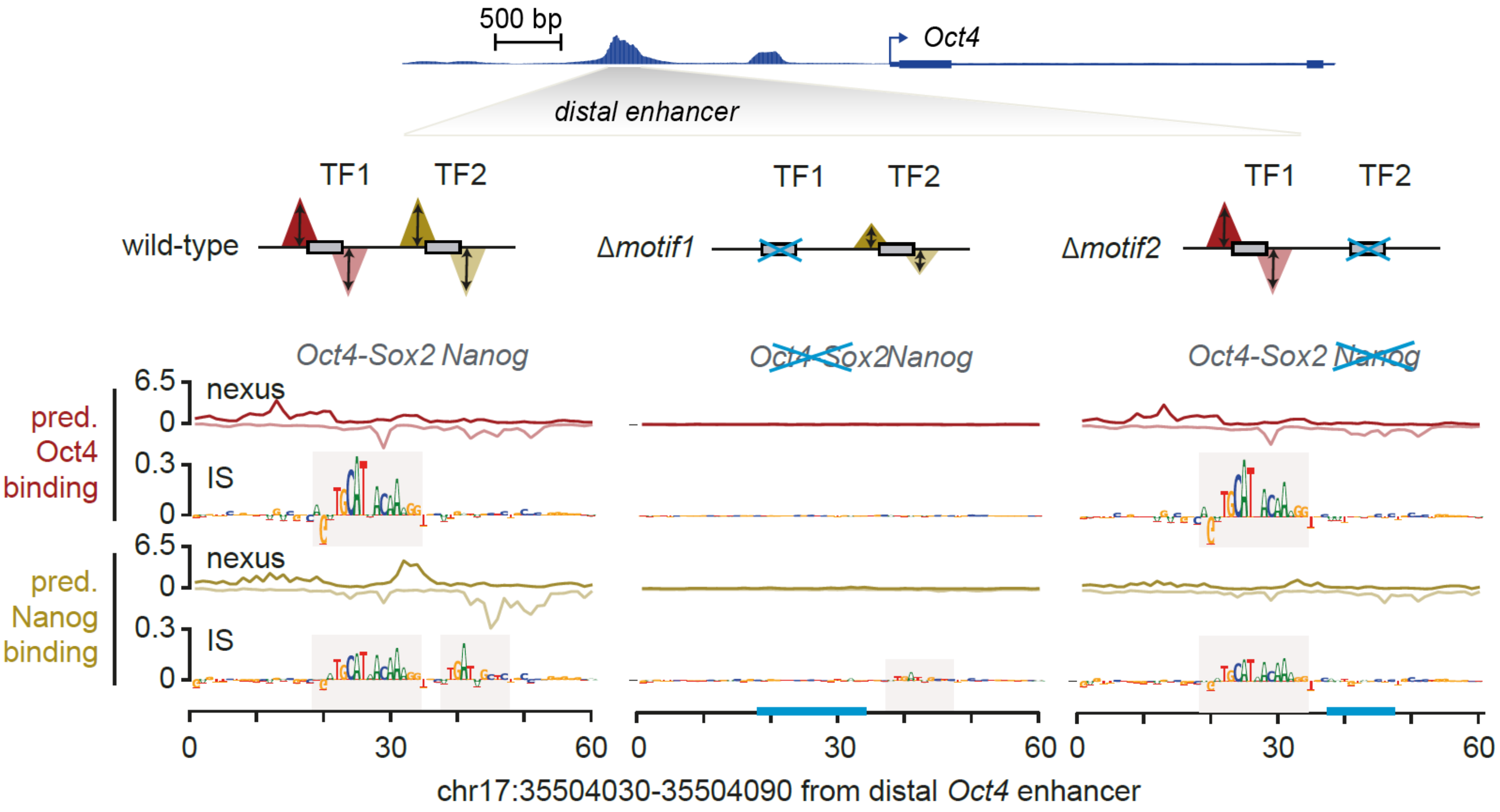
Motif syntax: cooperative TF interactions in genomic enhancers (*in-silico* CRISPR)

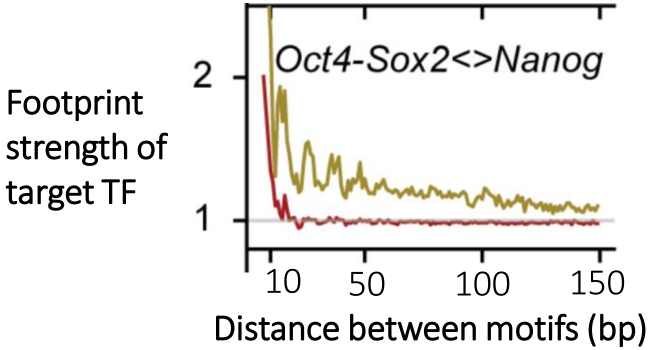


Motif syntax: cooperative TF interactions in genomic enhancers (*in-silico* CRISPR)

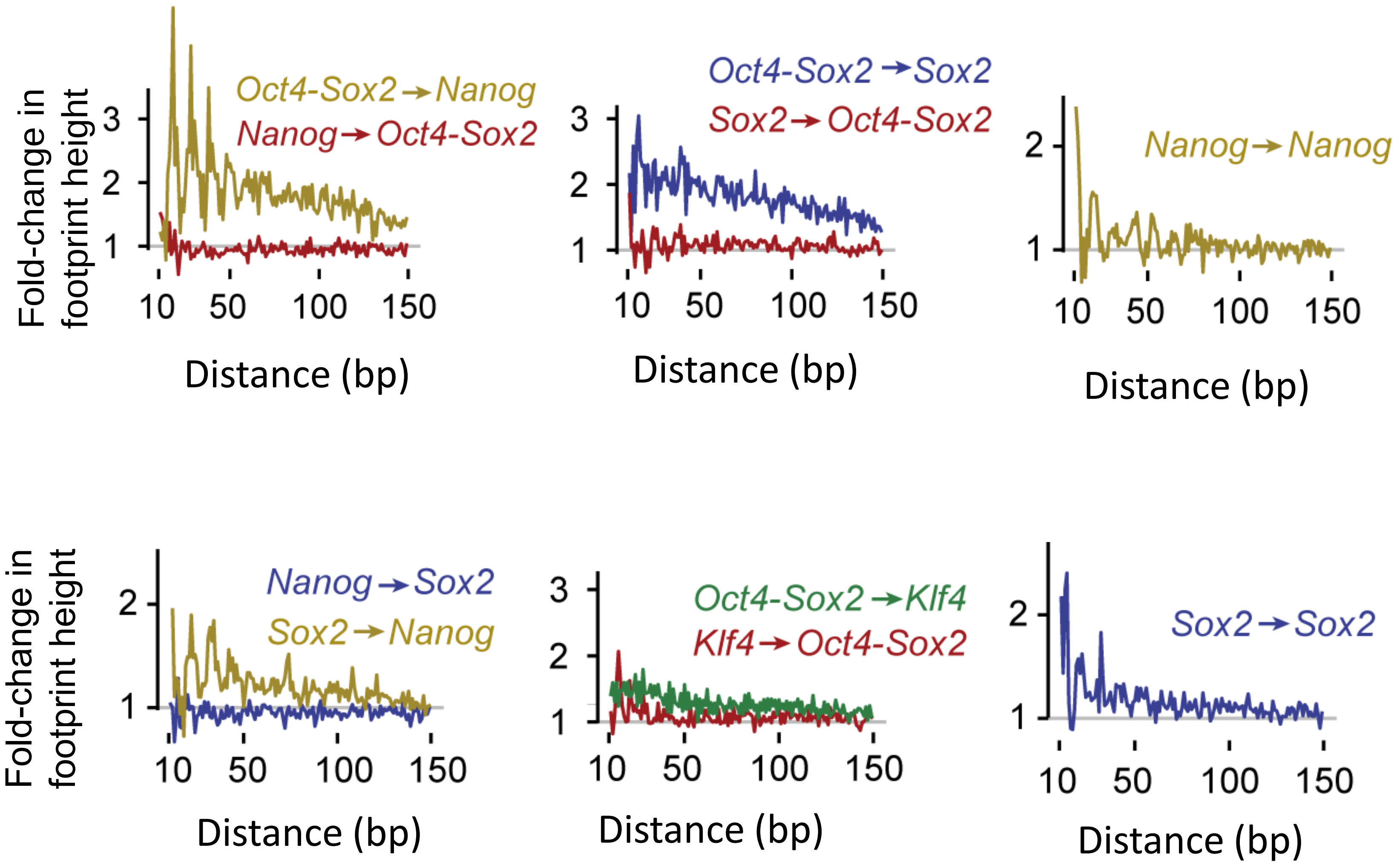


Motif syntax: cooperative TF interactions in genomic enhancers (*in-silico* CRISPR)

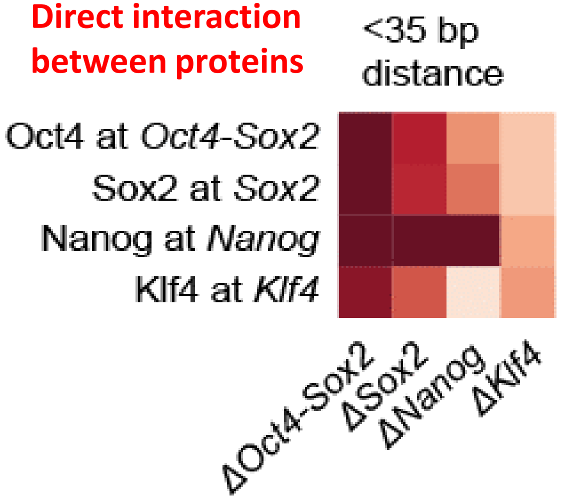




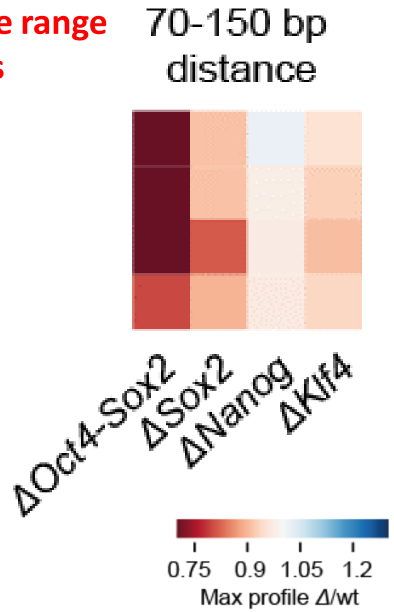
TF cooperativity is often directional & dependent on syntax with different distance ranges



Direct interaction between proteins



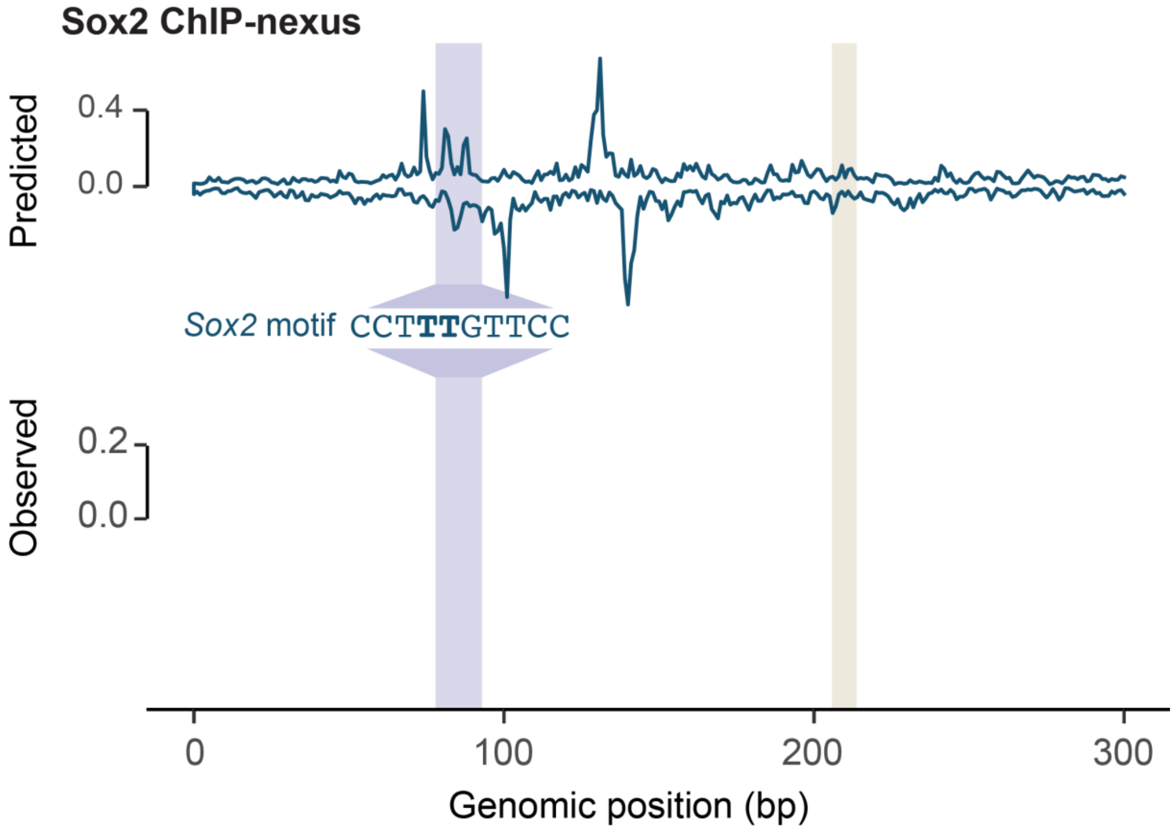
Nucleosome range interactions



CRISPR mutations validate motif syntax Nanog <> Sox2

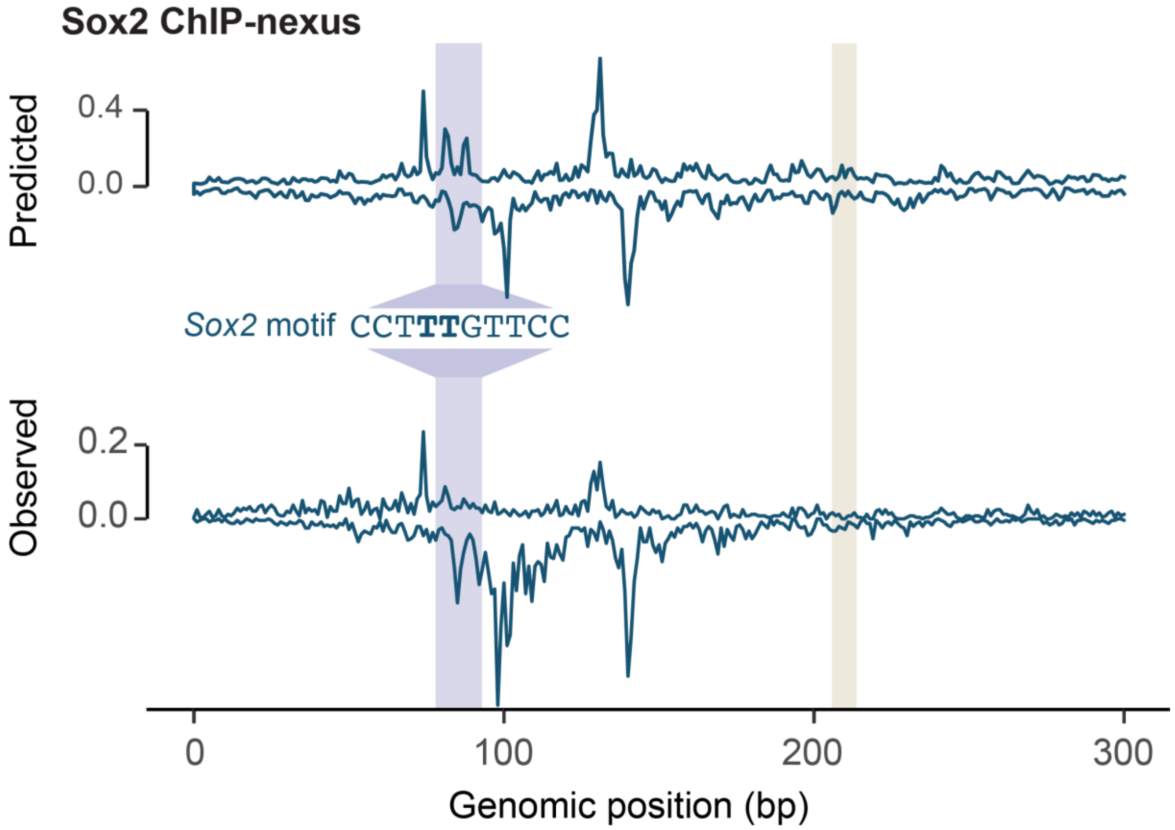
Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



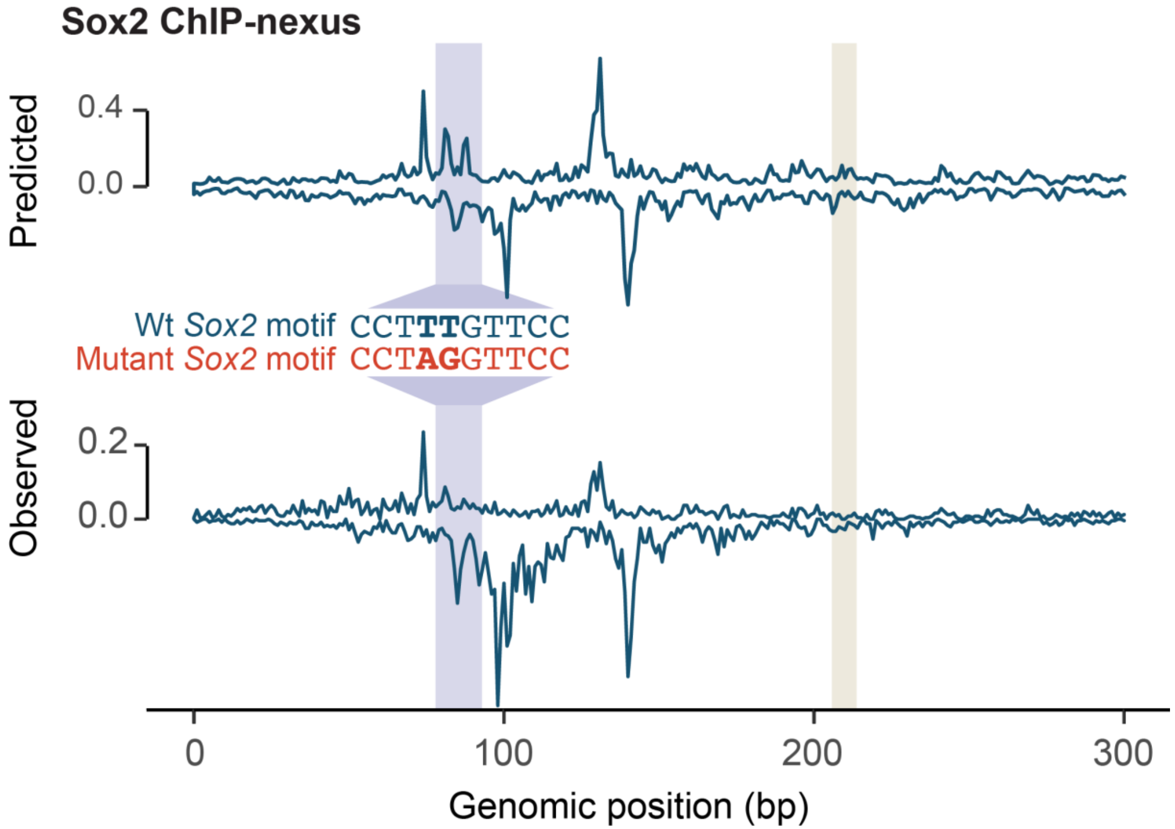
Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



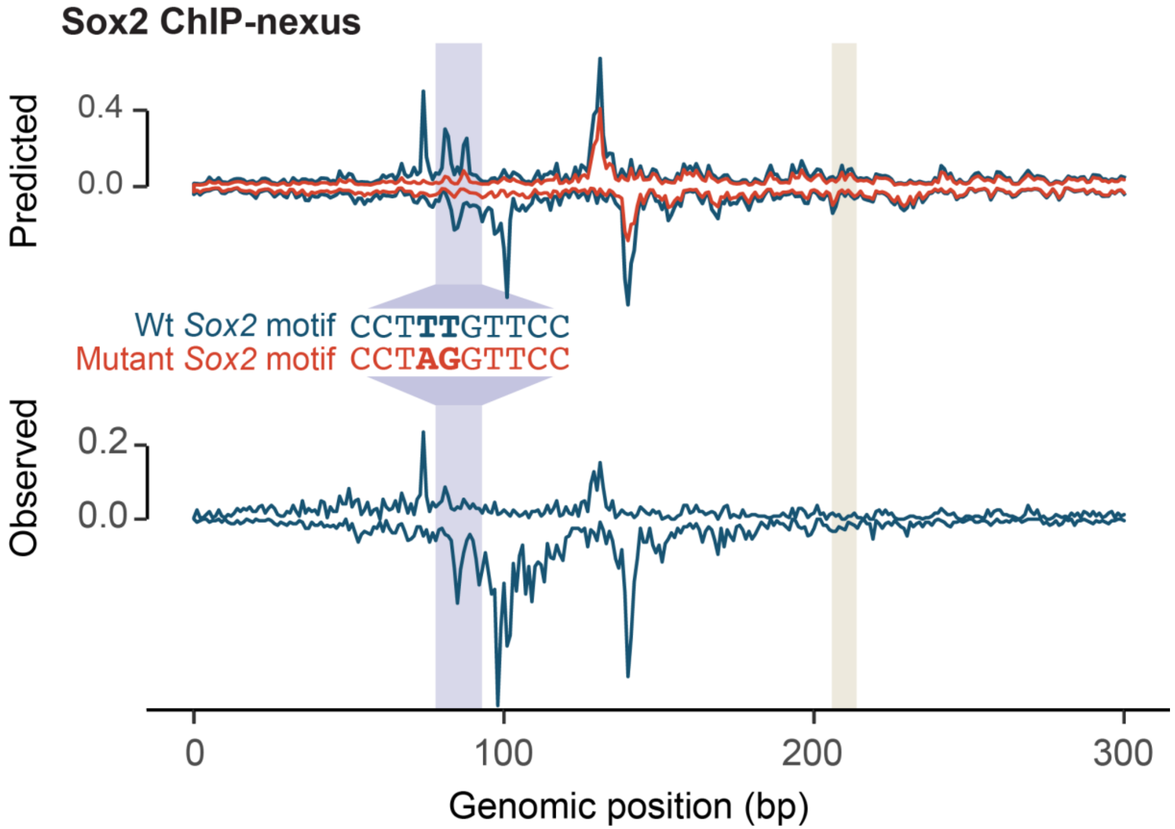
Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



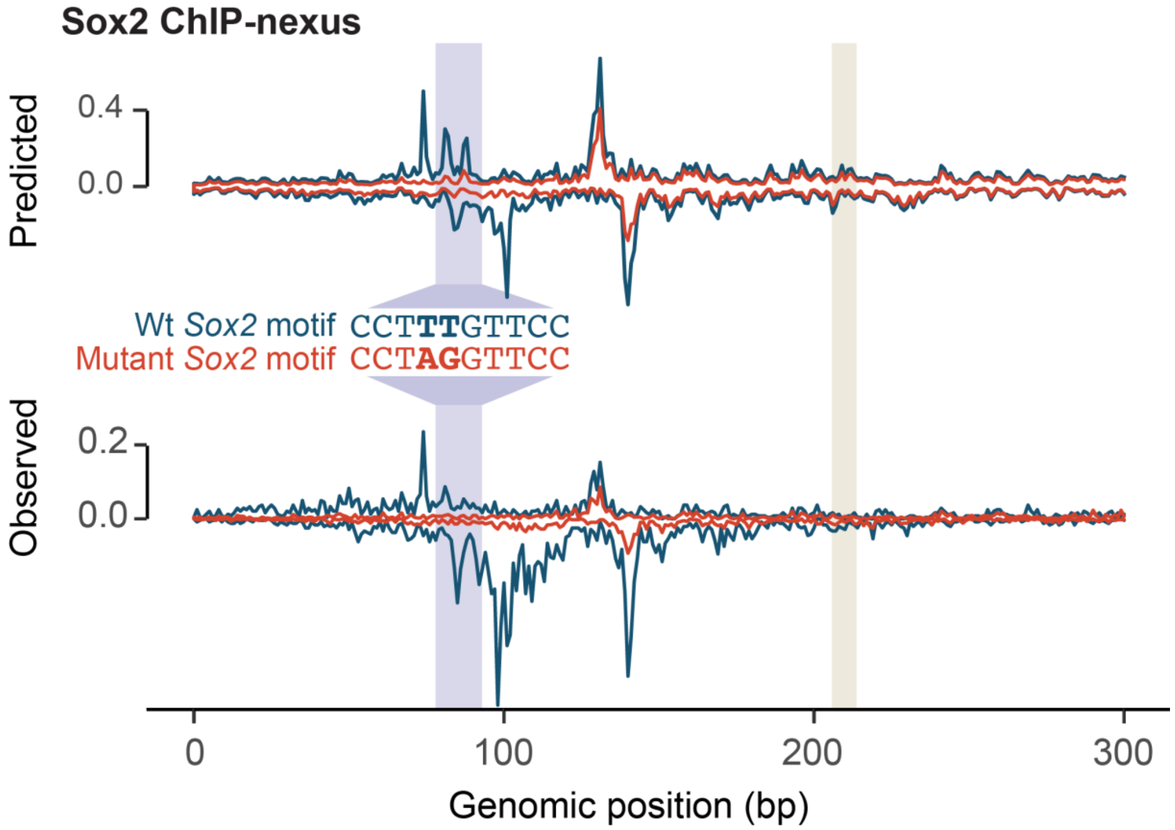
Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



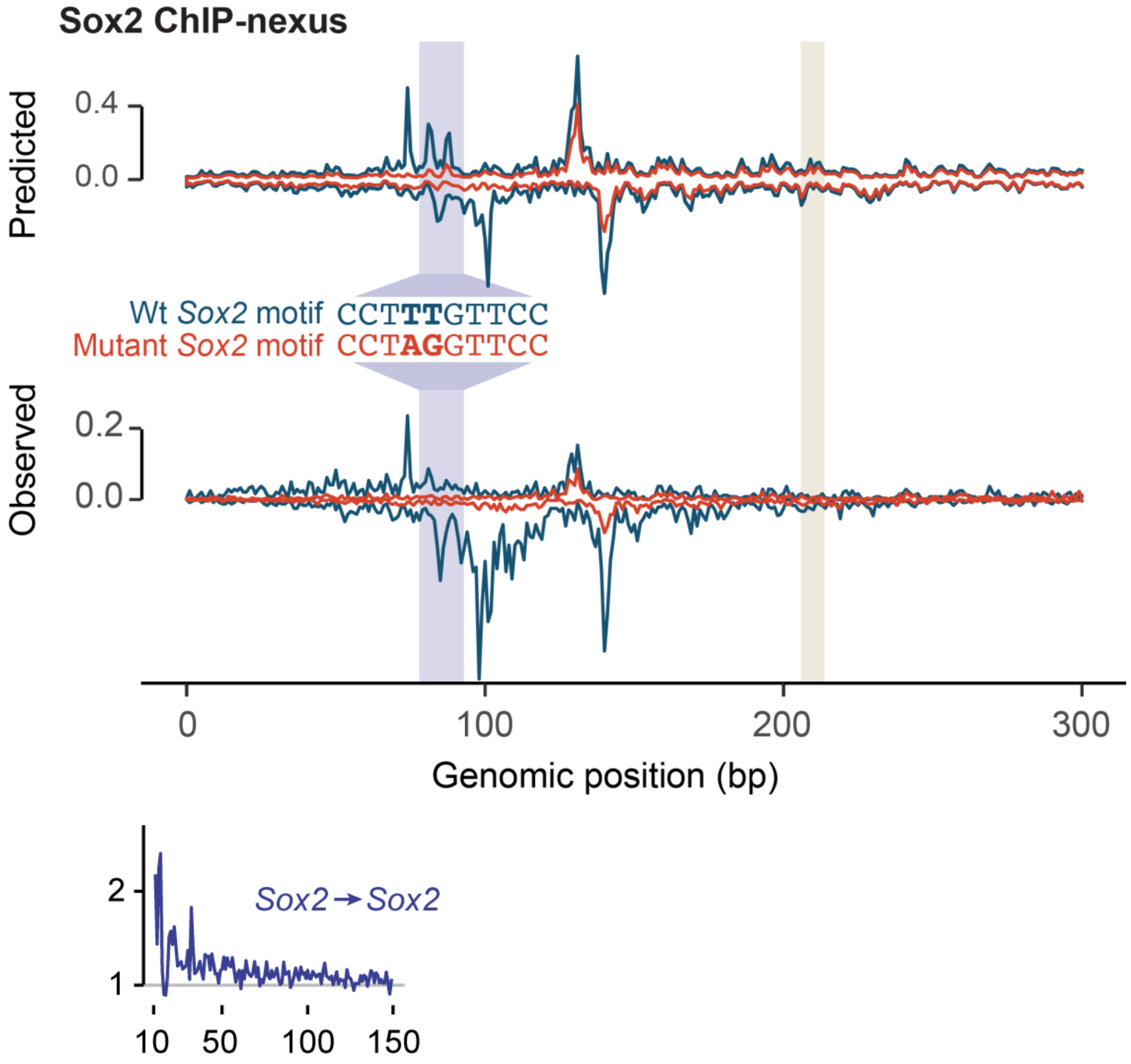
Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2



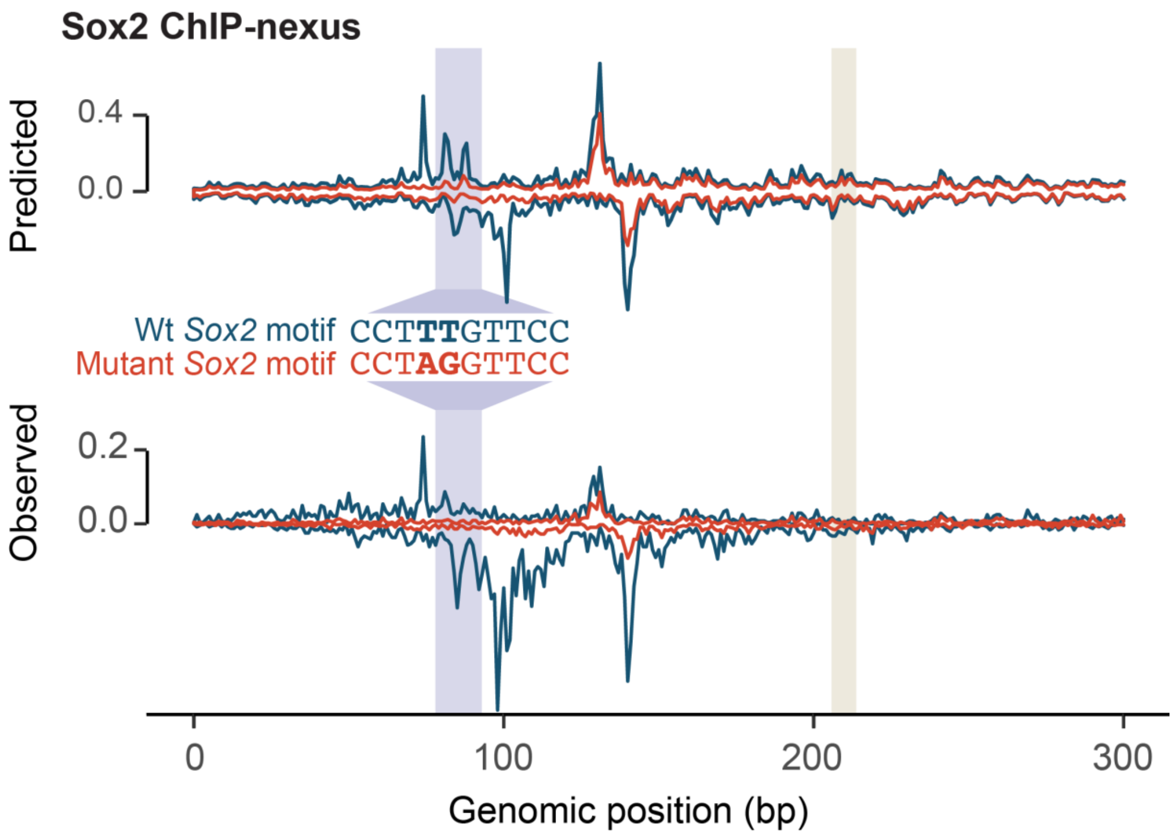
Sabrina Krueger, Melanie Weilert

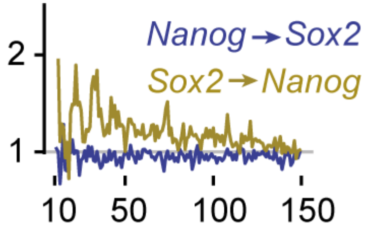
CRISPR mutations validate motif syntax Nanog <> Sox2

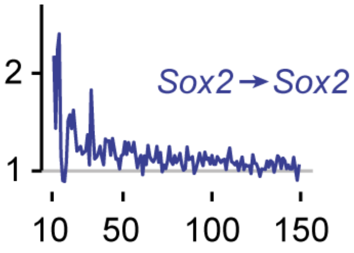


Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2

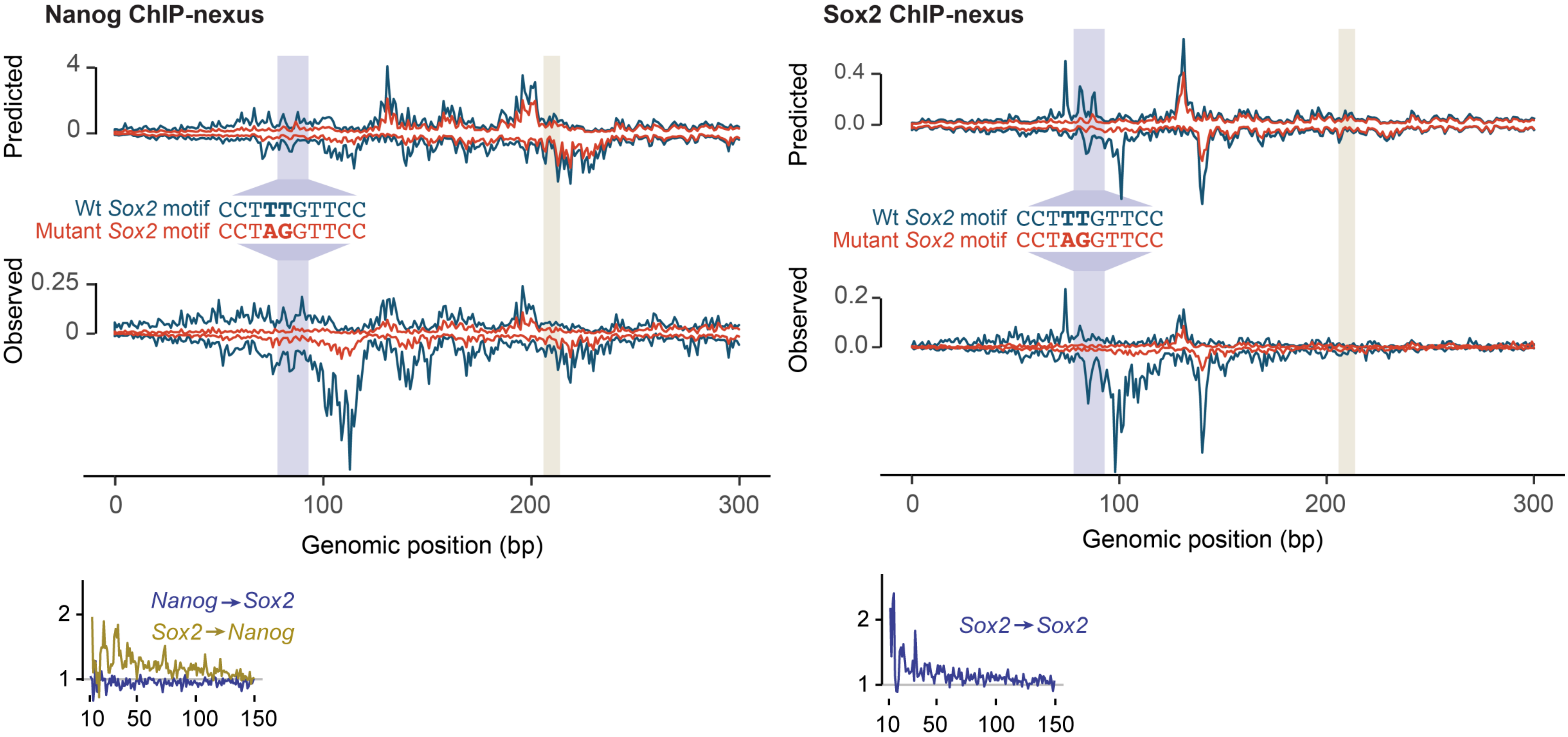






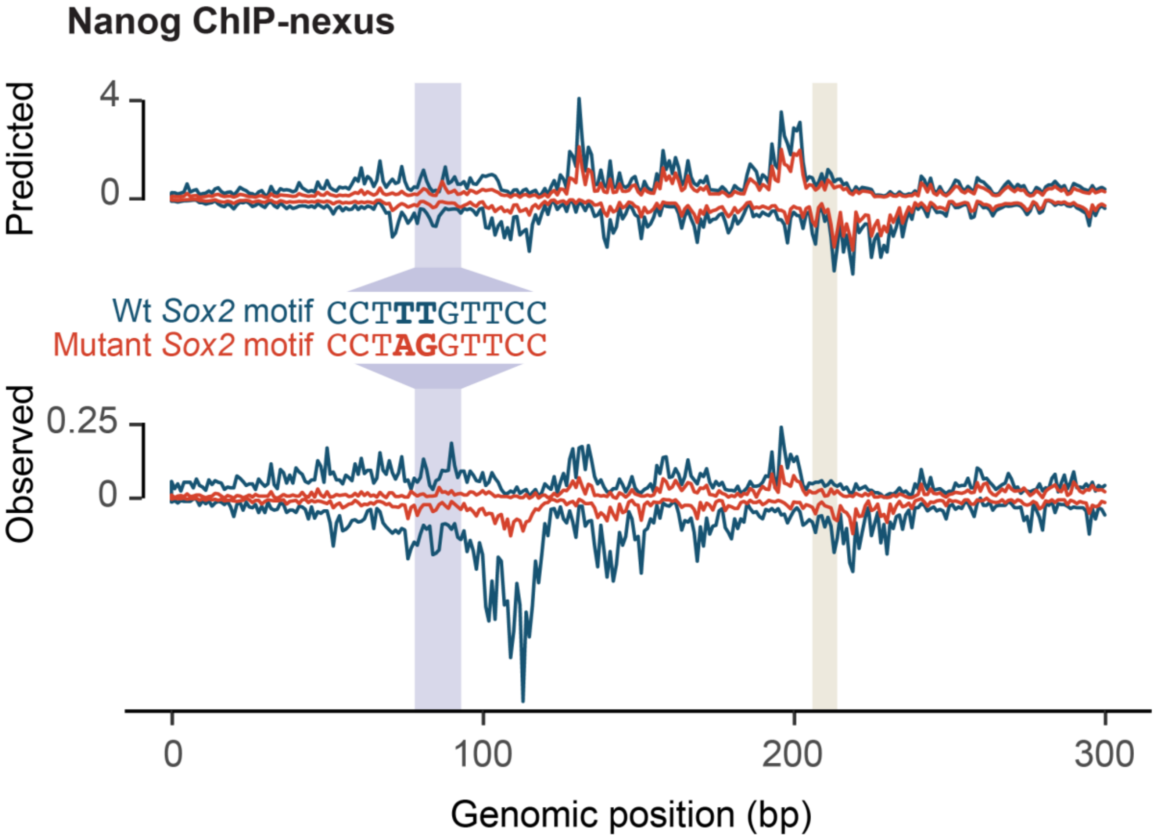
Sabrina Krueger, Melanie Weilert

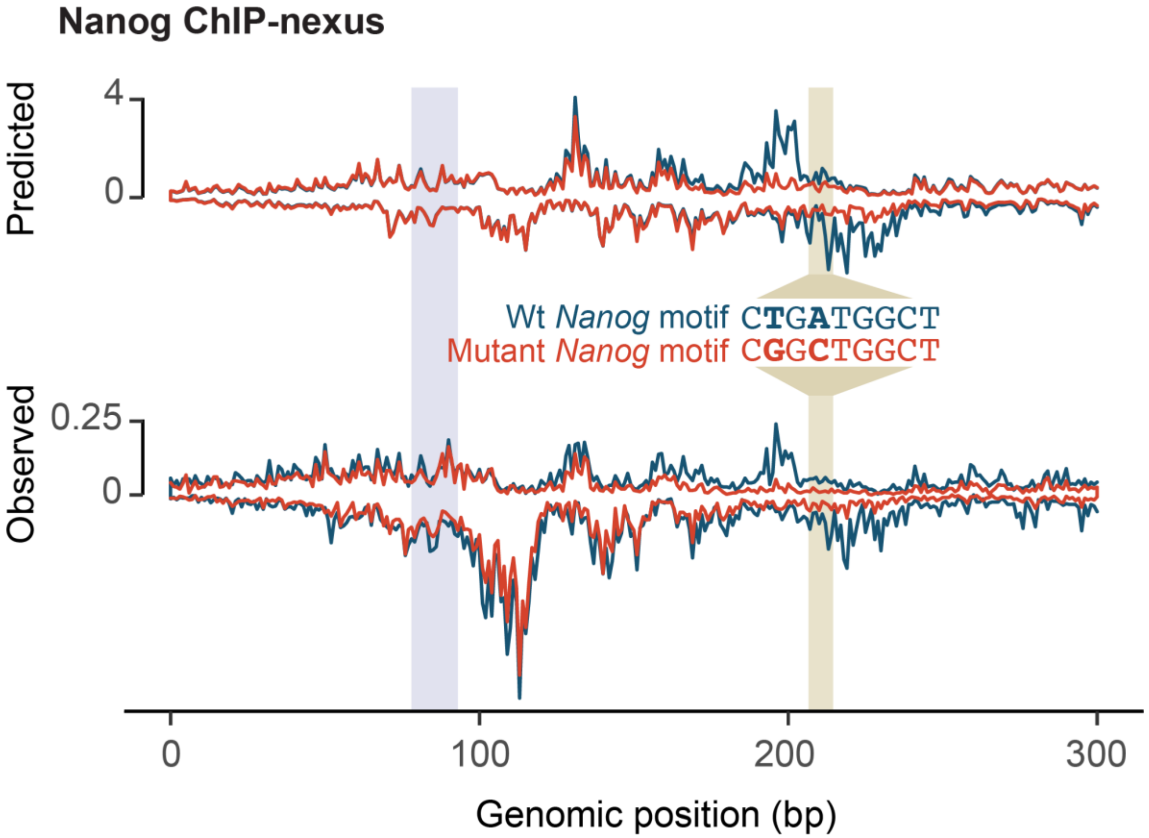
CRISPR mutations validate motif syntax Nanog <> Sox2

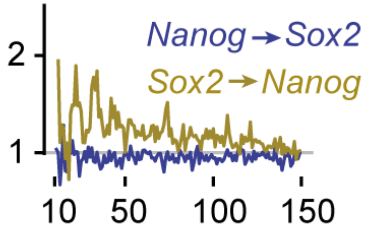


Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2

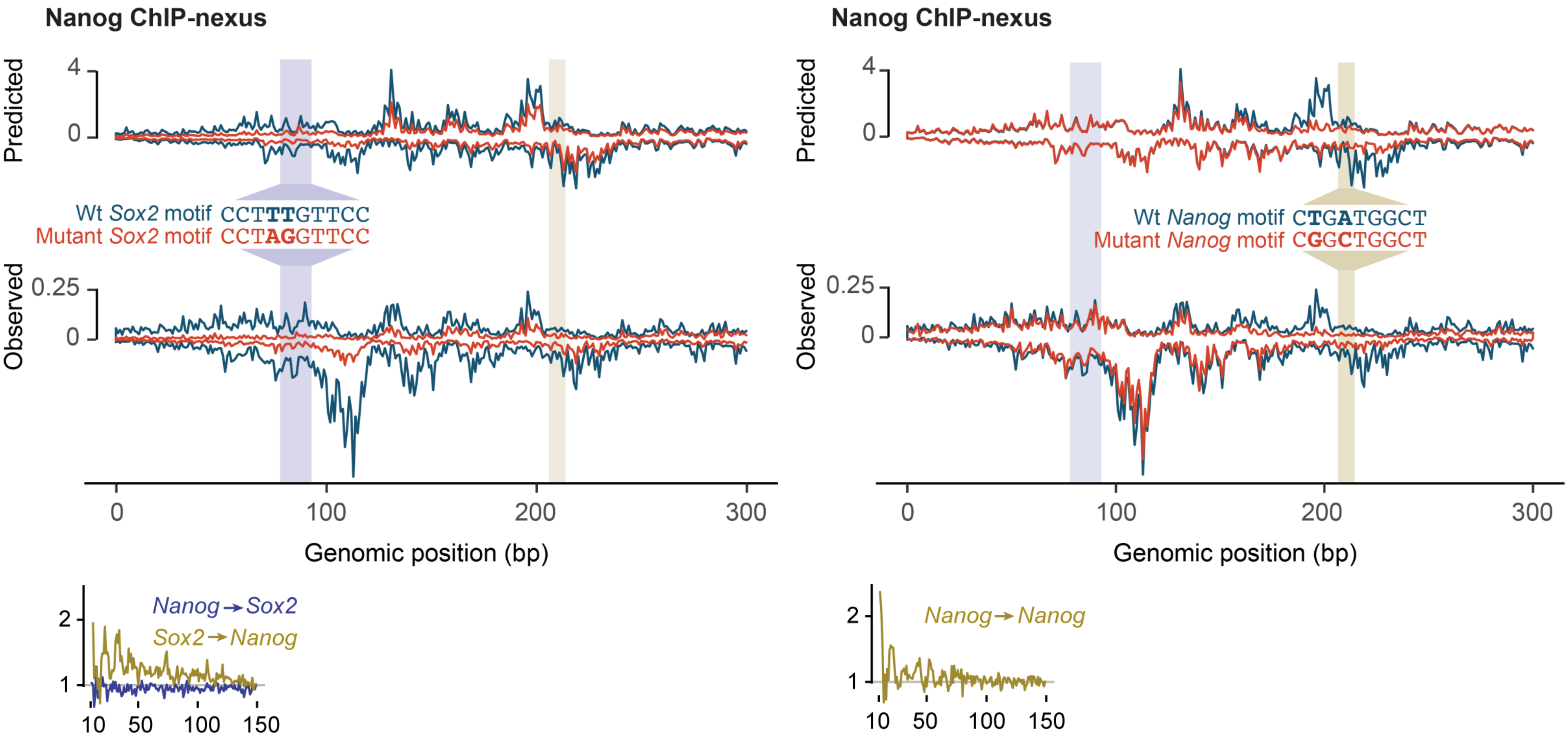






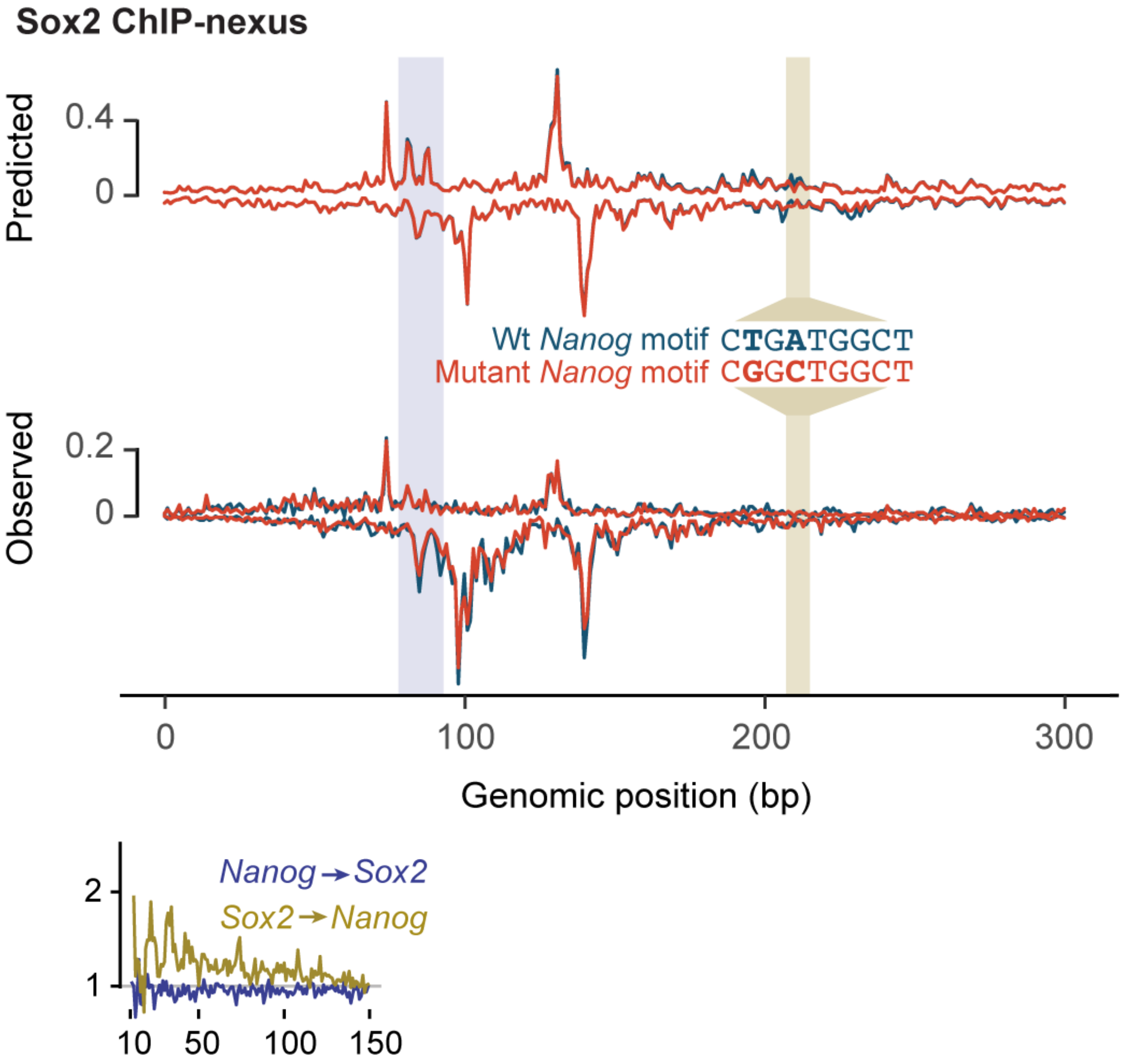
Sabrina Krueger, Melanie Weilert

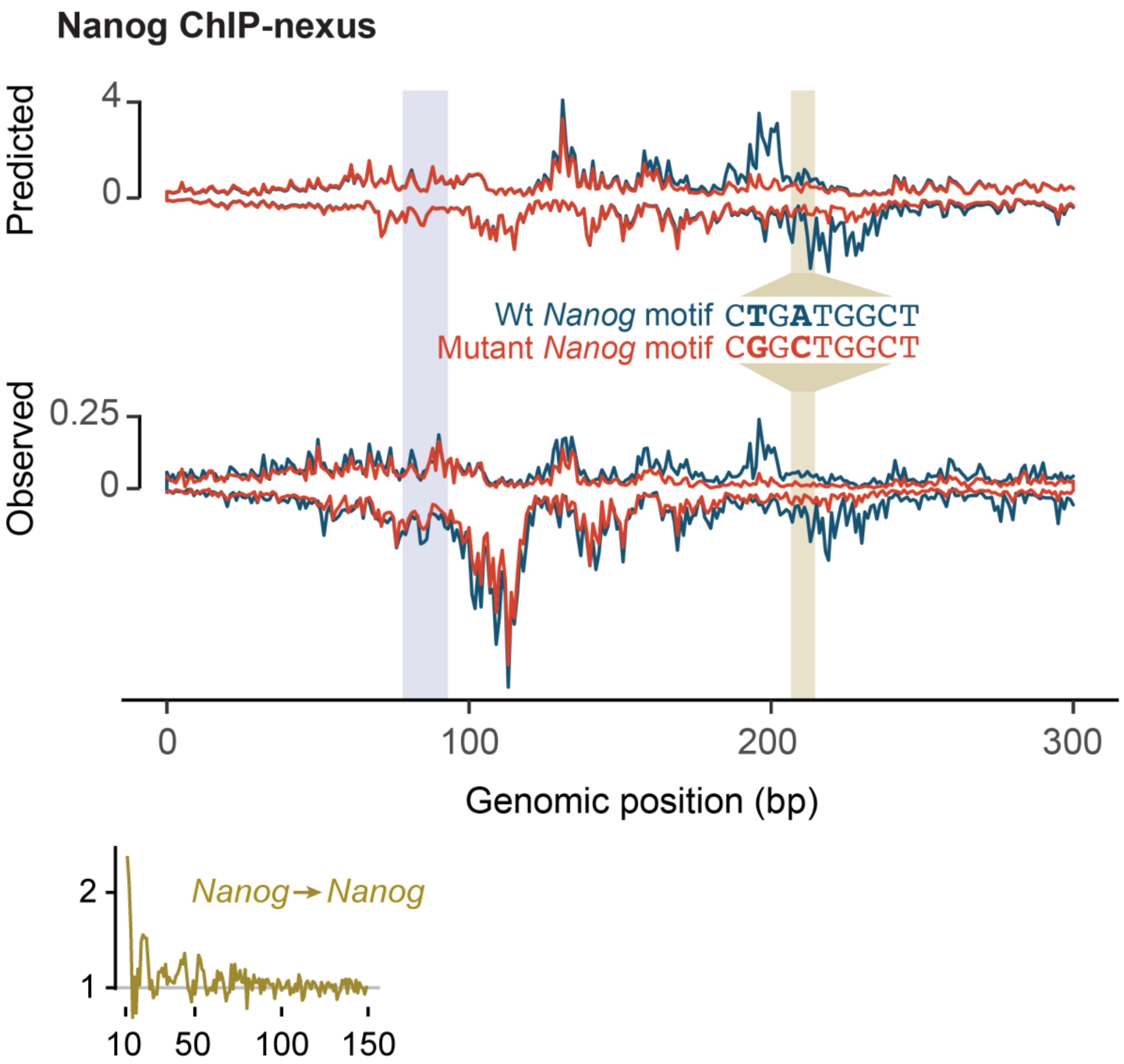
CRISPR mutations validate motif syntax Nanog <> Sox2



Sabrina Krueger, Melanie Weilert

CRISPR mutations validate motif syntax Nanog <> Sox2

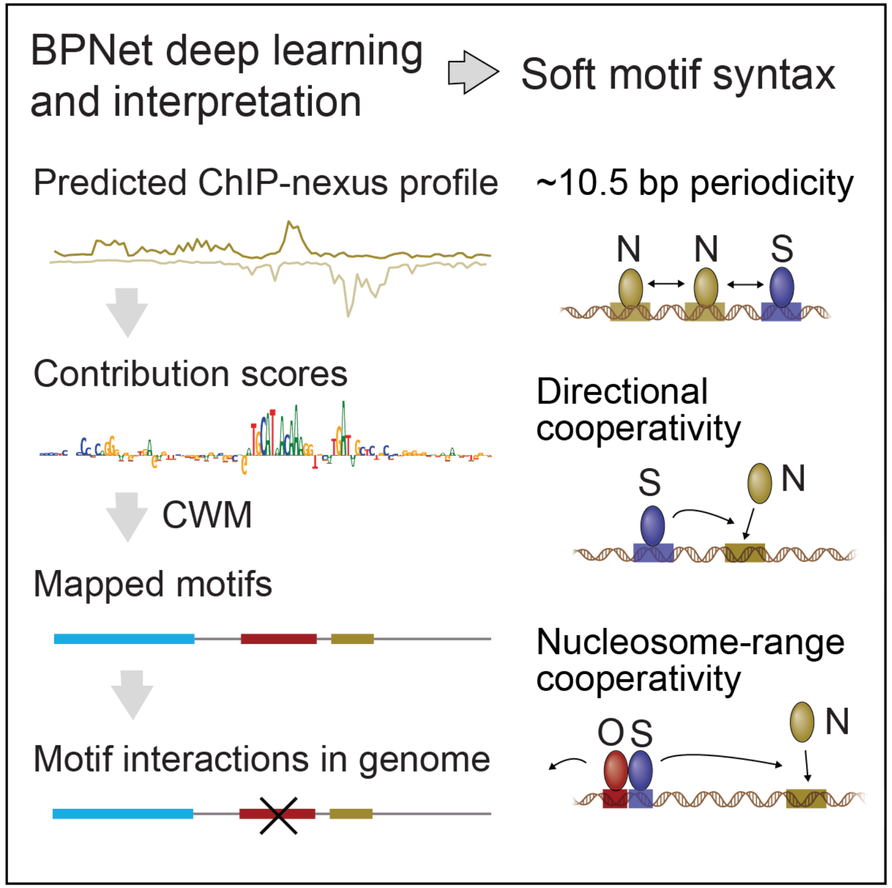




Sabrina Krueger, Melanie Weilert

Summary

- BPNet can map raw DNA sequence to base-resolution regulatory profiles with unprecedented accuracy
  - TF ChIP-exo/nexus, ChIP-seq, CUT&RUN
  - DNase-seq, ATAC-seq, scATAC-seq
  - Histone ChIP-seq / CUT&RUN
  - PRO-seq, RAMPAGE/CAGE
- Interpretation frameworks enable discovery of soft syntax mediated directional TF cooperativity
- Syntax of TF binding is predictive of
  - CRISPR motif perturbation experiments
  - Differential chromatin accessibility after TF knockdown
  - Reporter expression activity


BPNet deep learning and interpretation
Soft motif syntax
Predicted ChIP-nexus profile
Contribution scores
CWM
Mapped motifs
Motif interactions in genome
~10.5 bp periodicity
N
N
S
Directional cooperativity
S
N
Nucleosome-range cooperativity
O S
N

Acknowledgements

Ziga Avsec

Avanti Shrikumar

Melanie Weilert

Amr Mohamed

Julia Zeitlinger

Funding

ALFRED P. SLOAN FOUNDATION 1934

1R01HG009674

1U01HG009431

1U24HG009446

ENCODE

R01ES02500902

ROADMAP epigenomics PROJECT

1DP2OD022870

NIH DIRECTOR'S NEW INNOVATOR AWARD

- Khyati Dalal
- Sabrina Kruger
- Robin Fropf
- Charles McAnany
- Julien Gagneur

Deep Learning for Regulatory Genomics

1. **Biological foundations: Building blocks of Gene Regulation**
   - Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   - Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   - Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   - Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   - Key idea: pixels ⇔ DNA letters. Patches/filters ⇔ Motifs. Higher ⇔ combinations
   - Learning convolutional filters ⇔ Motif discovery. Applying them ⇔ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   - DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   - DeepSea: Train model directly on mutational impact prediction
   - Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   - ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   - DeepLIFT: Model interpretation based on neuron activation properties
   - DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**

NVIDIA

Enhancing epigenomic data with deep learning

Avantika Lal, 3/11/2021

GENOMICS AT NVIDIA

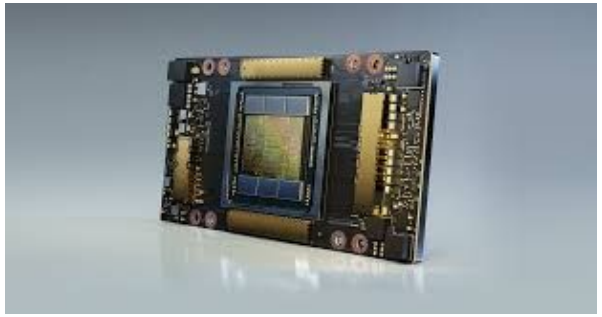
We are a team of scientists and engineers developing software to solve some of the most difficult problems in genomics.

We collaborate with academic institutes and companies across the world.

We apply machine learning, deep learning, and accelerated computing to build faster and more accurate tools - enabling new biological discoveries.

Some areas we work in:

- Single-cell genomics
- Epigenomics
- Cancer
- Variant calling
- Genome assembly
- Long-read sequencing



nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-21765-5 OPEN

Deep learning-based enhancement of epigenomics data with AtacWorks

Avantika Lal[1,3], Zachary D. Chiang[2,3], Nikolai Yakovenko[1], Fabiana M. Duarte[2], Johnny Israeli[1]✉ & Jason D. Buenrostro[2]✉

https://www.nature.com/articles/s41467-021-21765-5

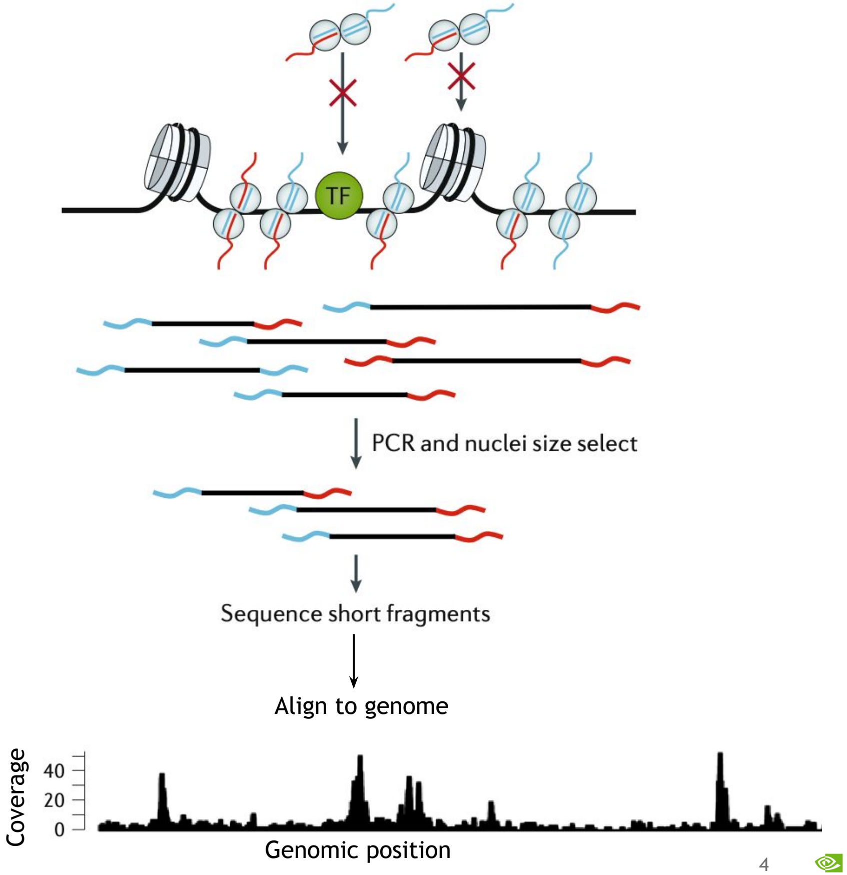
ATAC-SEQ

Chromatin accessibility mapping with DNA sequencing

ATAC-seq measures chromatin accessibility using DNA sequencing.

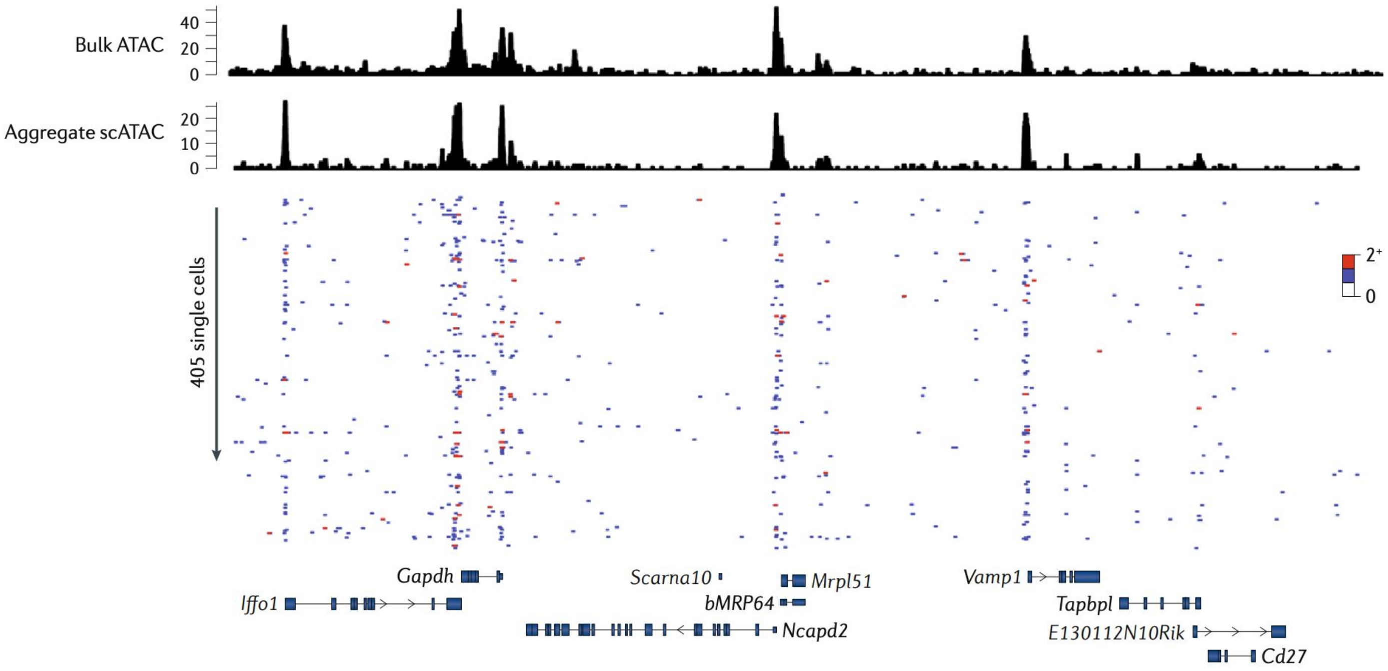
'Peaks' of high sequencing read coverage correspond to regions of open chromatin in the genome.

ATAC-seq helps identify active regulatory elements, build regulatory networks, and study the effect of non-coding variation.

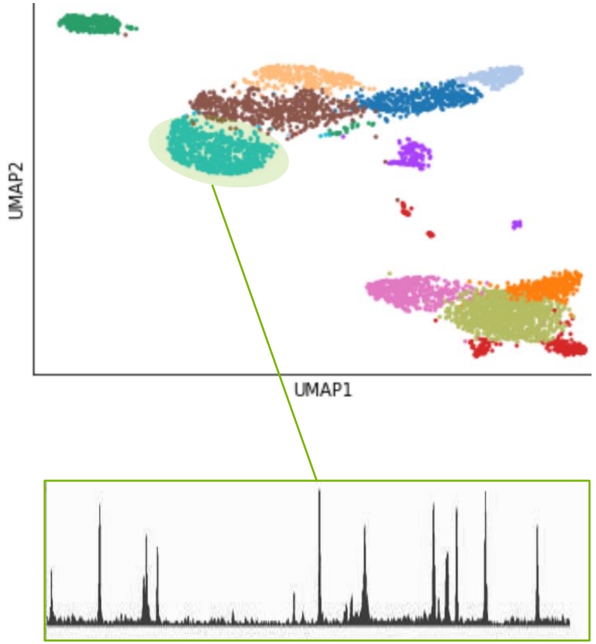


Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20, 207–220 (2019). https://doi.org/10.1038/s41576-018-0089-8

SINGLE-CELL ATAC-SEQ



Cluster cells and identify accessible sites at the cell type level

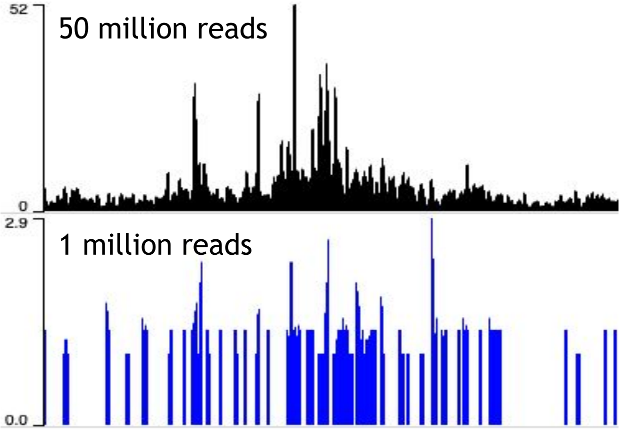


Biological tissues are heterogeneous mixtures of different types of cells. Single-cell sequencing shows us this heterogeneity, but each cell provides only a noisy, sparse signal.

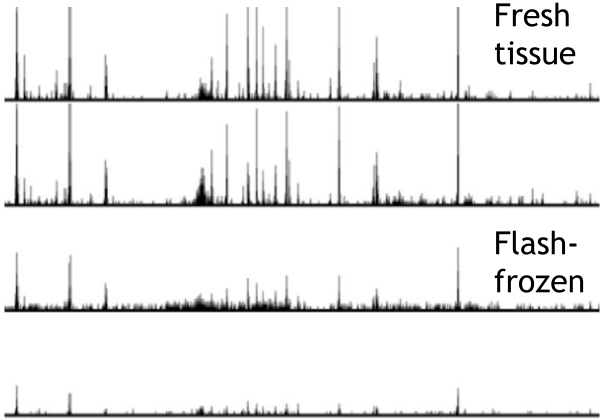
Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 20, 207–220 (2019). https://doi.org/10.1038/s41576-018-0089-8

DATA QUALITY IN ATAC-SEQ

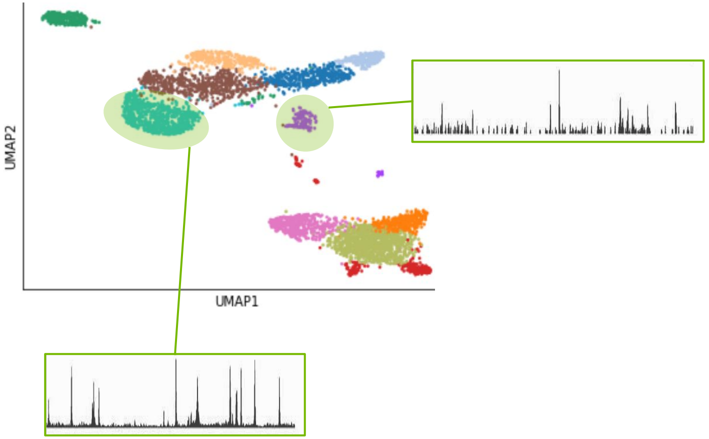
1 Low sequencing depth



2 Sample/experimental factors



3 Low aggregate cell count



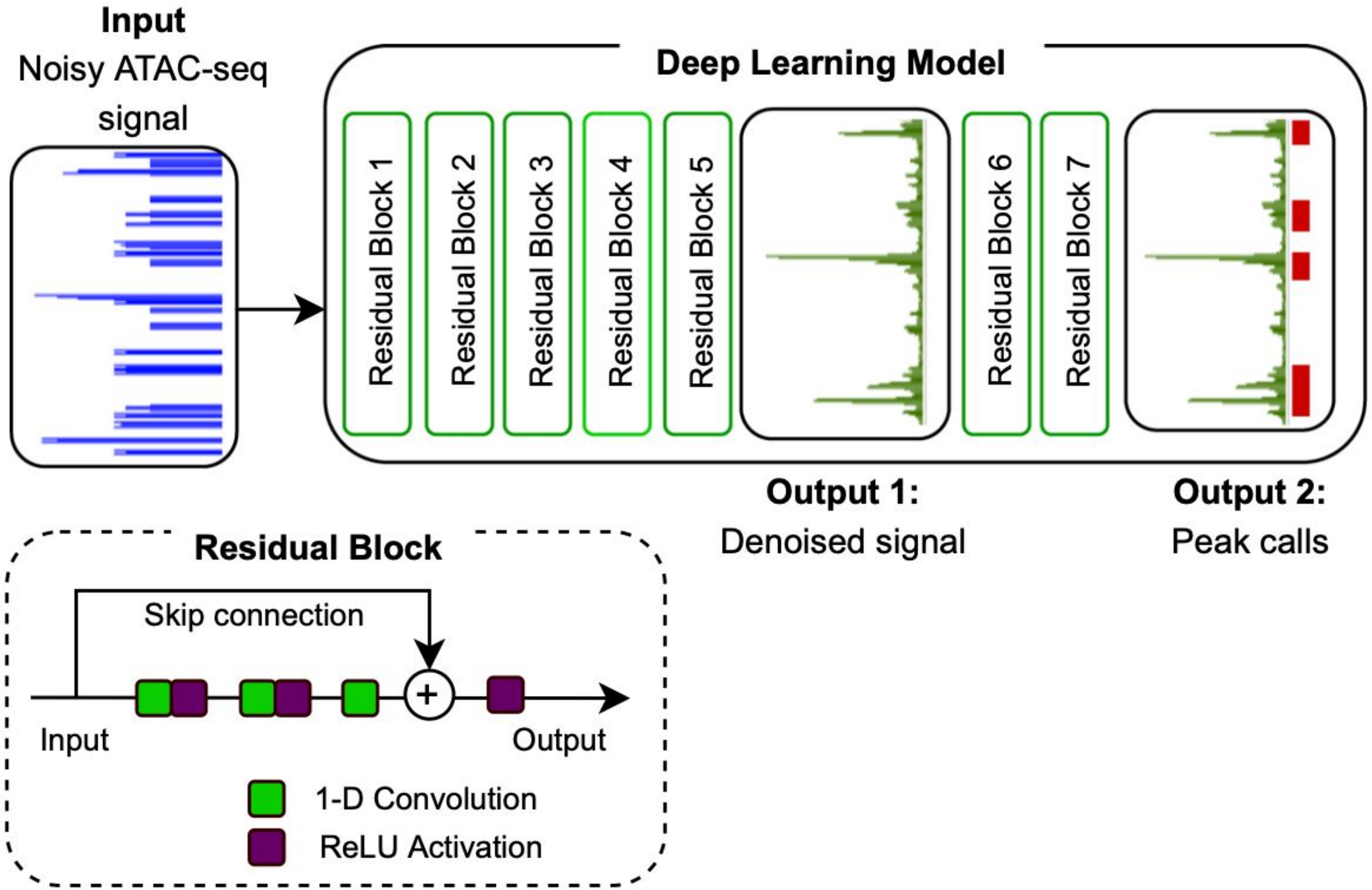
ATACWORKS

AtacWorks takes as input the coverage track from an ATAC-seq experiment, and improves its accuracy.

AtacWorks also identifies the peaks, or open chromatin regions.

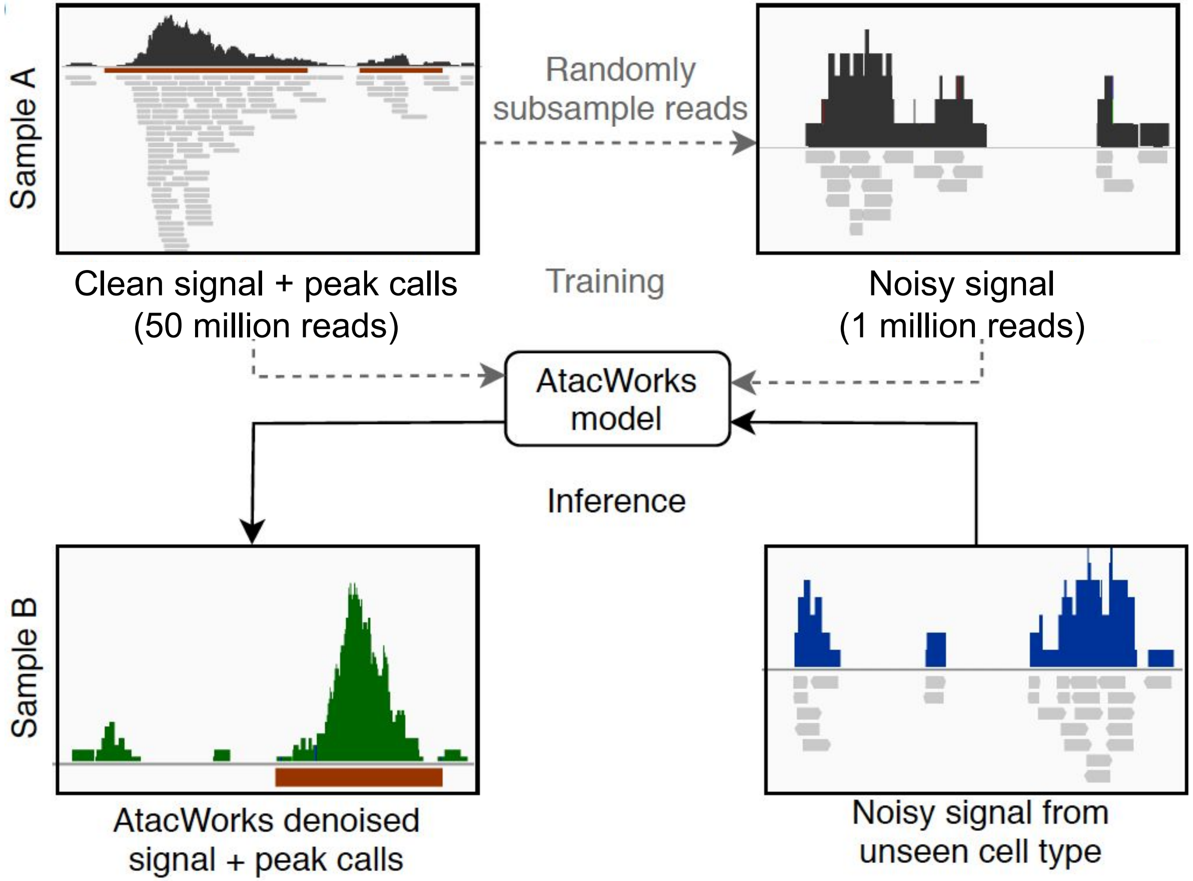
It uses a ResNet (Residual Neural Network) architecture, a convolutional architecture originally used in computer vision.

However, it uses 1-D convolutional layers instead of the 2-D layers used in image analysis.



Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

TRAINING ATACWORKS TO ENHANCE LOW-COVERAGE ATAC-SEQ DATA



Sample A

Randomly subsample reads

Clean signal + peak calls (50 million reads)

Training

Noisy signal (1 million reads)

AtacWorks model

Inference

Sample B

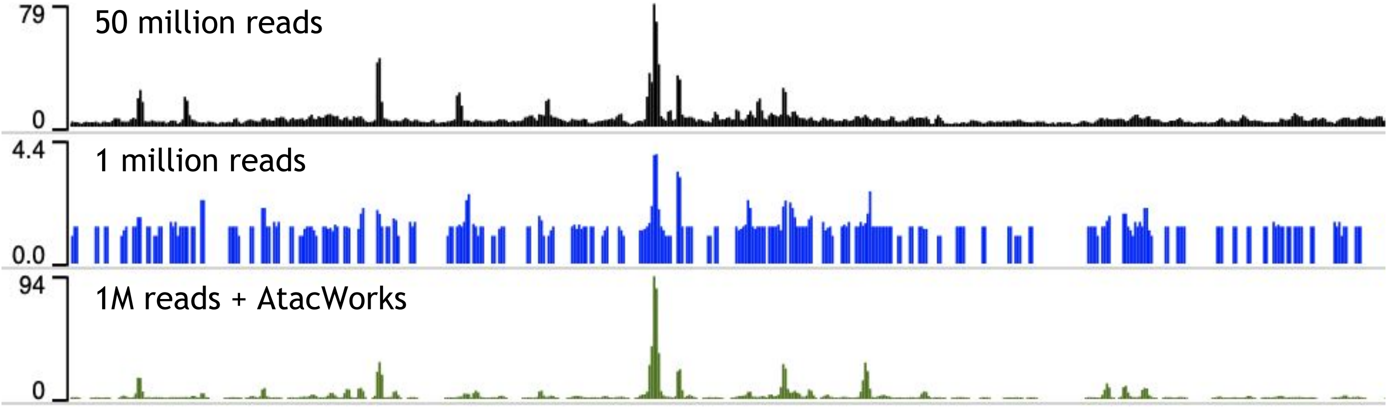
AtacWorks denoised signal + peak calls

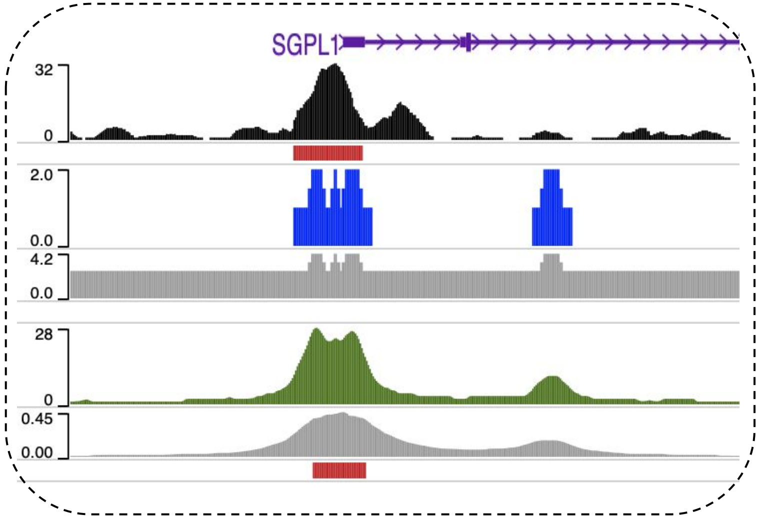
Noisy signal from unseen cell type

ATACWORKS DENOISES AND CALL PEAKS FROM LOW-COVERAGE ATAC-SEQ

Bulk ATAC-seq data from human Erythroblasts

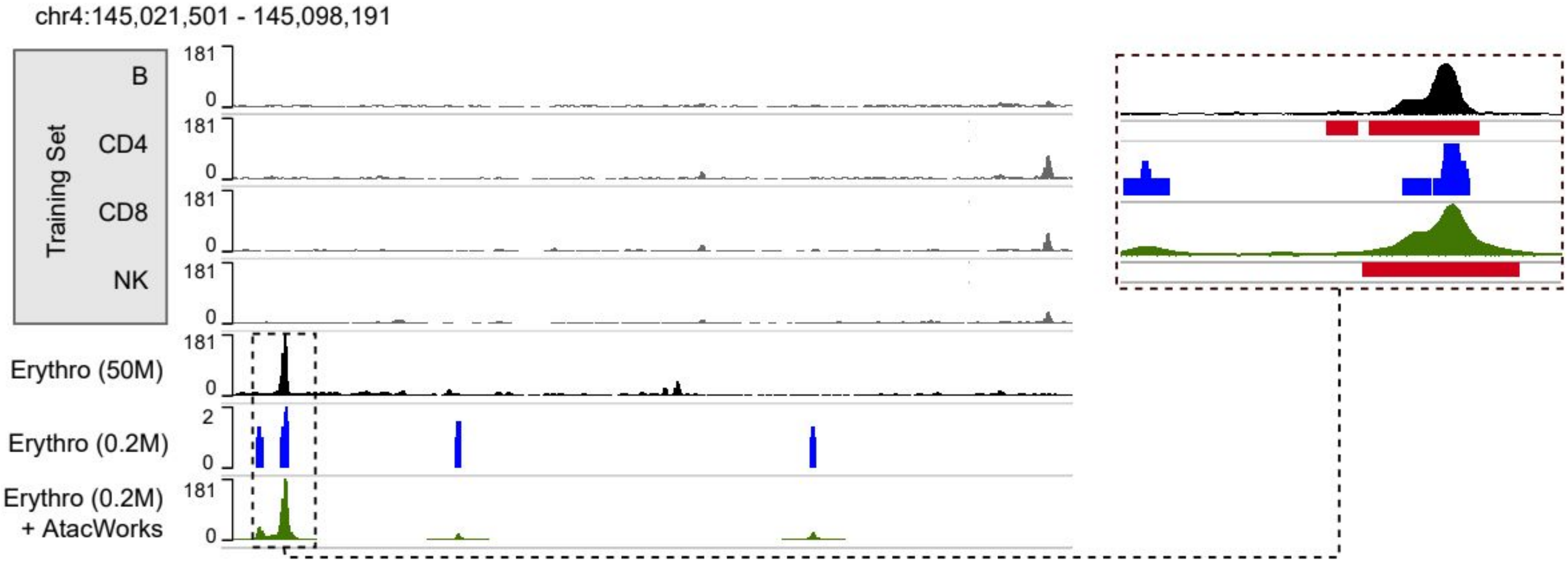
Chr10: 70,400,000-71,450,000





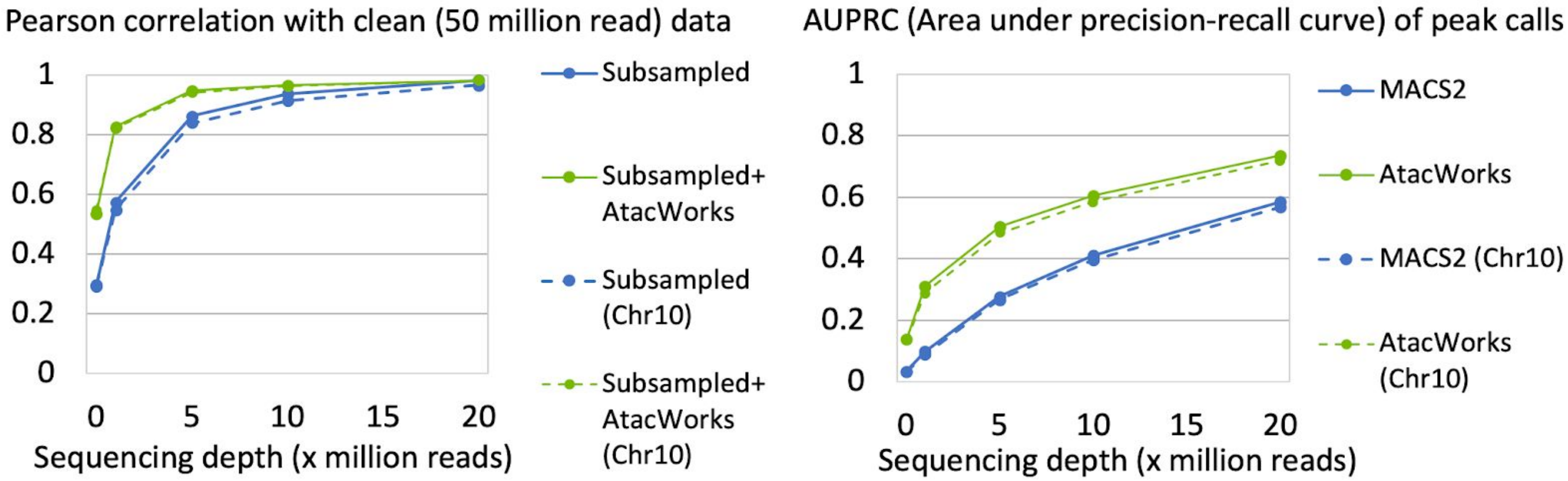
AtacWorks distinguishes real peaks and identifies peaks missed by MACS2.

ATACWORKS GENERALIZES ACROSS CELL TYPES



Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

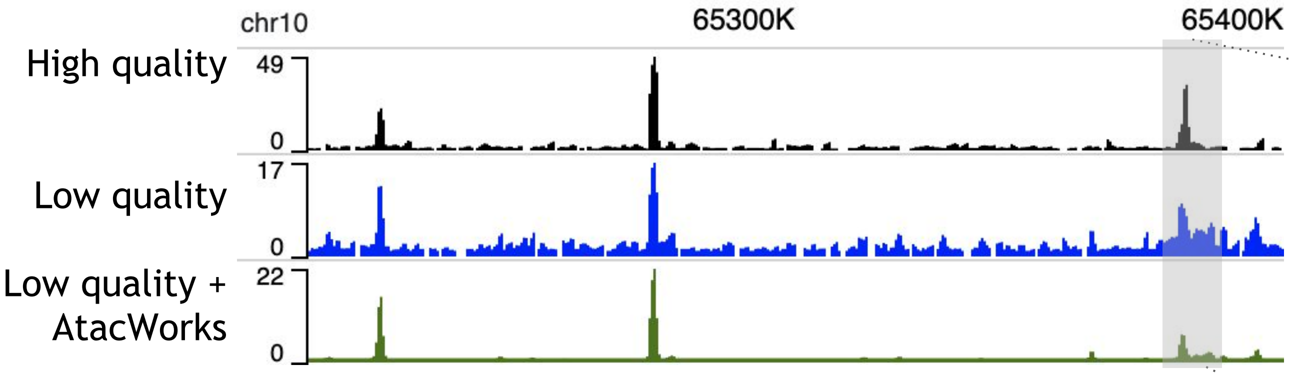
GENOME-WIDE PERFORMANCE METRICS

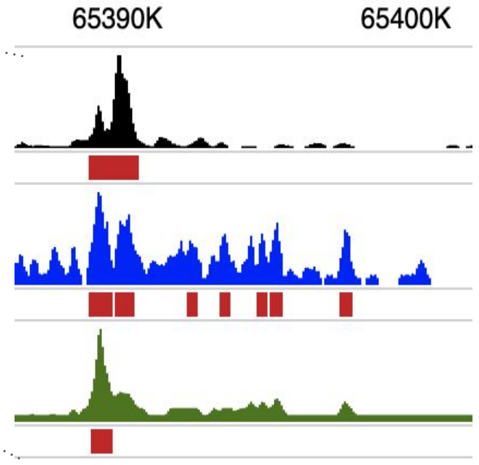


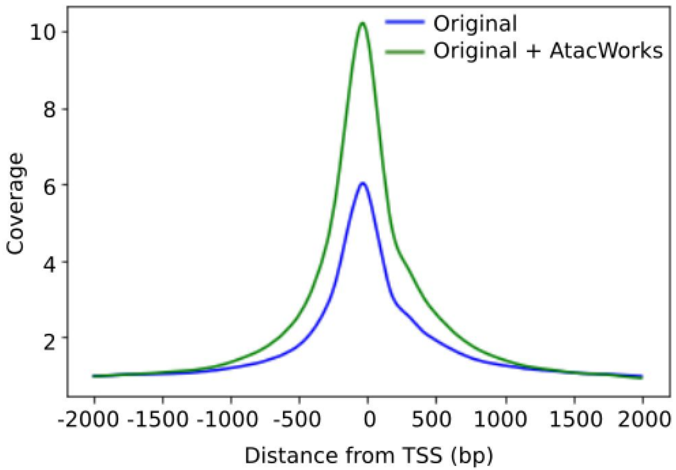
AtacWorks returns equivalent results at 2-5x lower sequencing depth.

ATACWORKS ENHANCES LOW-QUALITY ATAC-SEQ

Bulk ATAC-seq data from human Erythroblasts



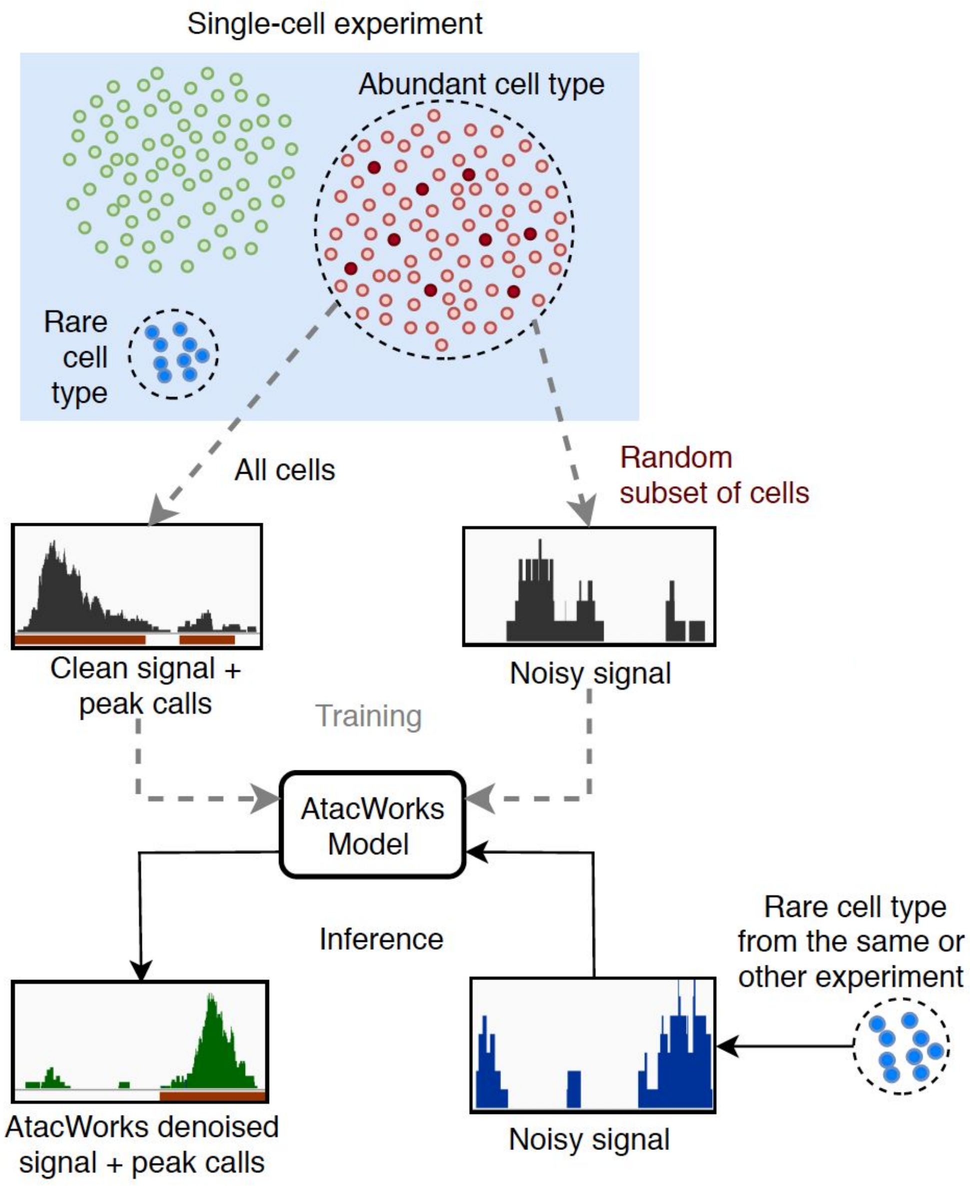




Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

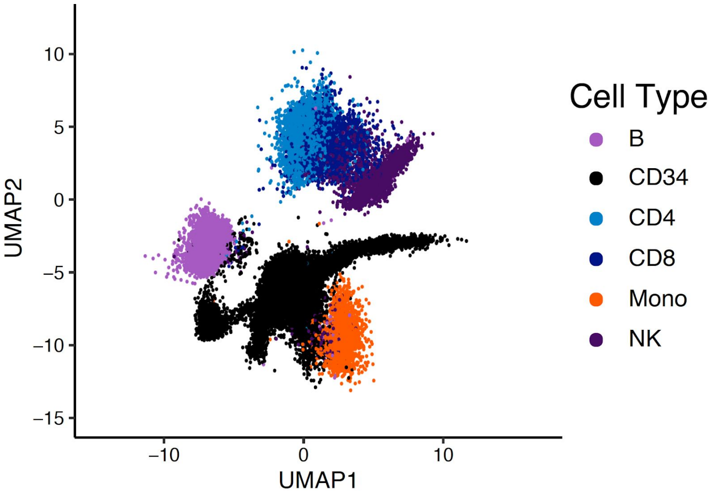
ATACWORKS FOR SINGLE-CELL ATAC-SEQ

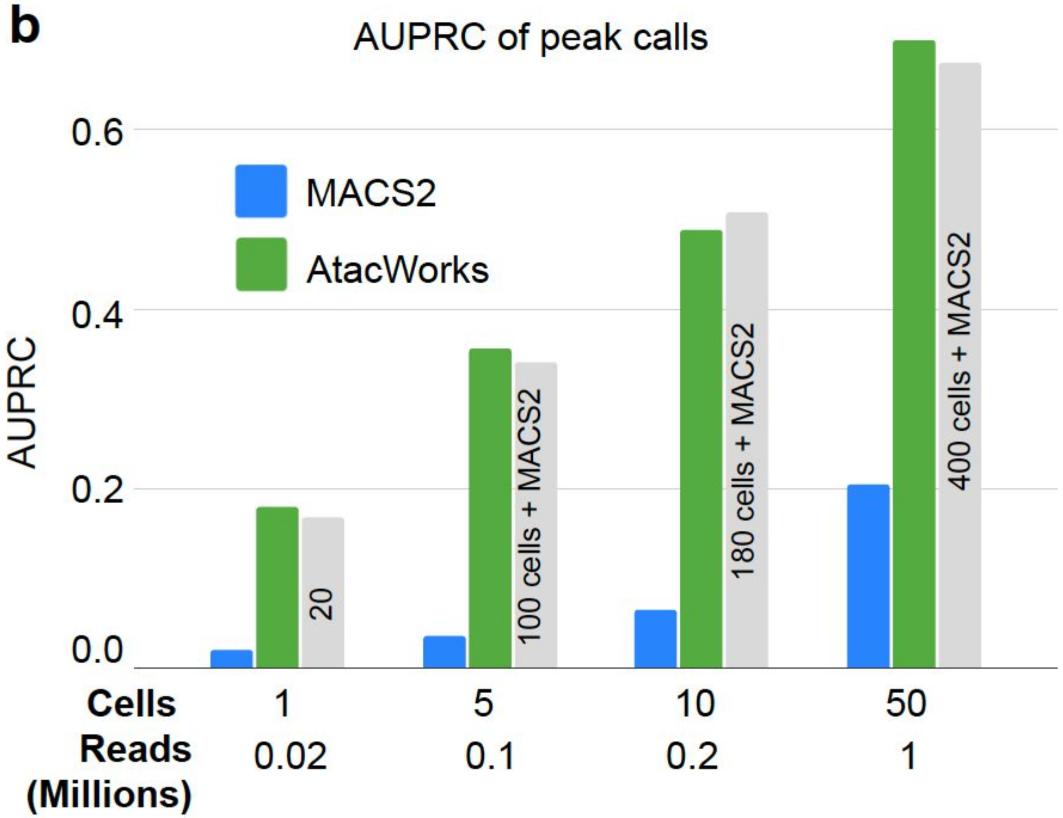
Profiling accessible chromatin in rare cell types



Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

ATACWORKS ENABLES ANALYSIS OF SMALL NUMBERS OF CELLS

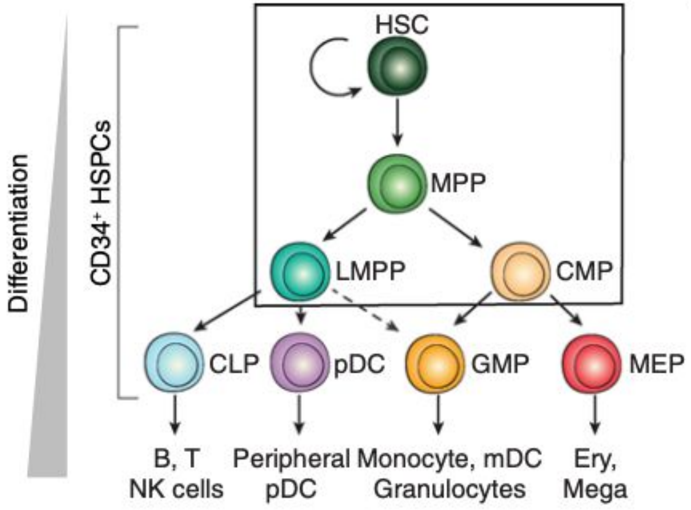


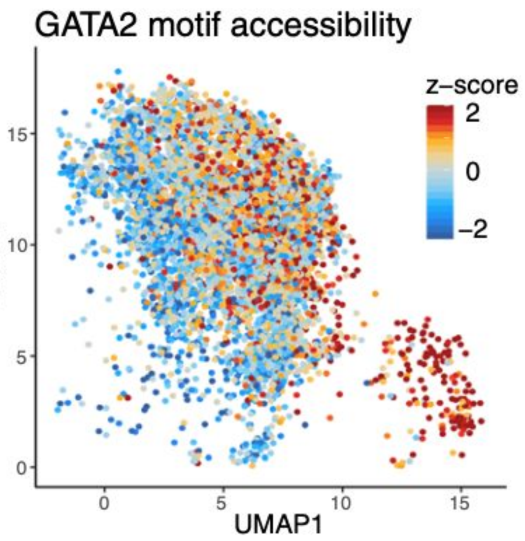


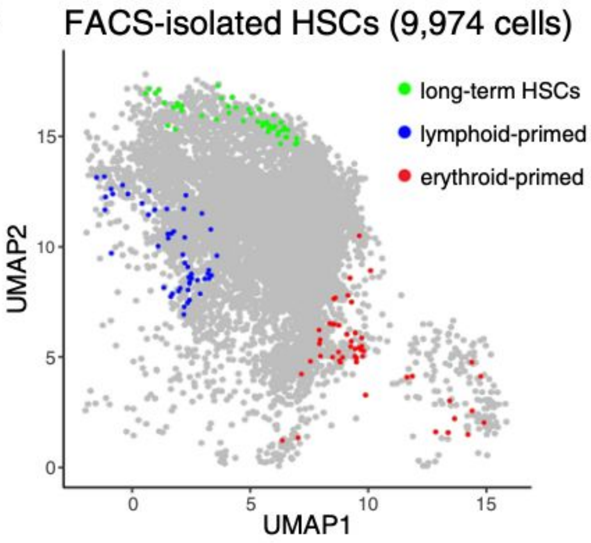
AtacWorks can obtain the same quality from ~10x fewer cells, increasing the resolution of single-cell chromatin accessibility profiling by an order of magnitude.

Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

LINEAGE PRIMING IN HEMATOPOIETIC STEM CELLS

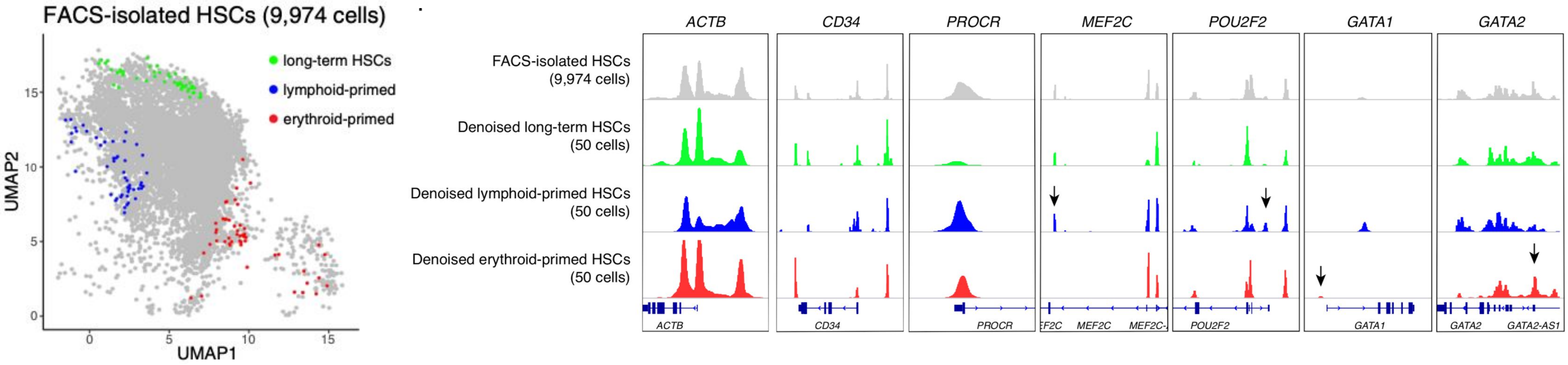






Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

ATACWORKS IDENTIFIES REGULATORY ELEMENTS THAT CONTROL LINEAGE PRIMING



Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

INTERACTIVE EXAMPLE

https://github.com/clara-parabricks/rapids-single-cell-examples/blob/master/notebooks/5k_pbmc_coverage_gpu.ipynb

Built by Raj Movva (MIT CS undergrad)

ACKNOWLEDGMENTS

NVIDIA

Nikolai Yakovenko

Joyjit Daw

Eric Xu

Gary Burnett

Neha Tadimeti

Ohad Mosafi

Rajiv Movva (MIT)

Bryan Catanzaro

Johnny Israeli

Buenrostro Lab, Harvard University

Jason Buenrostro

Zachary Chiang

Fabiana Duarte

CONTACT

Avantika Lal

Senior Scientist
(Deep Learning & Genomics)

NVIDIA

📧 alal@nvidia.com

https://www.linkedin.com/in/avantikalal

@lal_avantika

Internships available!

NVIDIA®

Deep Learning for Regulatory Genomics

1. **Biological foundations: Building blocks of Gene Regulation**
   – Gene regulation: Cell diversity, Epigenomics, Regulators (TFs), Motifs, Disease role
   – Probing gene regulation: TFs/histones: ChIP-seq, Accessibility: DNase/ATAC-seq
2. **Classical methods for Regulatory Genomics and Motif Discovery**
   – Enrichment-based motif discovery: Expectation Maximization, Gibbs Sampling
   – Experimental: PBMs, SELEX. Comparative genomics: Evolutionary conservation.
3. **Regulatory Genomics CNNs (Convolutional Neural Networks): Foundations**
   – Key idea: pixels ⇔ DNA letters. Patches/filters ⇔ Motifs. Higher ⇔ combinations
   – Learning convolutional filters ⇔ Motif discovery. Applying them ⇔ Motif matches
4. **Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures**
   – DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
   – DeepSea: Train model directly on mutational impact prediction
   – Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
   – ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
   – DeepLIFT: Model interpretation based on neuron activation properties
   – DanQ: Recurrent Neural Network for sequential data analysis
5. **Guest Lecture: Anshul Kundaje, Stanford, Deep Learning for Reg. Genomics**
6. **Guest Lecture: Avantika Lal, Nvidia, Deep Learning for ATAC/scATAC**