

6.874, 6.802, 20.390, 20.490, HST.506

Computational Systems Biology

Deep Learning in the Life Sciences

Lecture 8: TF binding

Gene regulation, DNA regulatory code,
3D conformation folding

Prof. Manolis Kellis

Guest lecture: David Kelley

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
- Three-dimensional chromatin structure, Hi-C, ChIA-PET, TADs, Loop Extrusion

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 \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow Motif matches

4. Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures

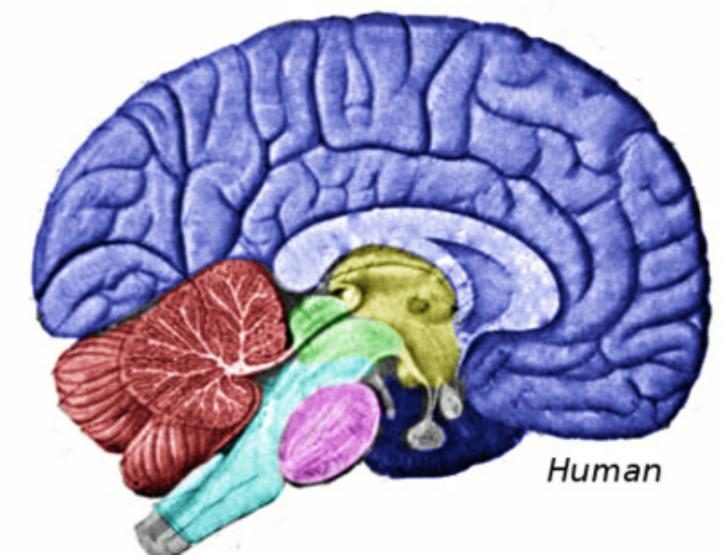
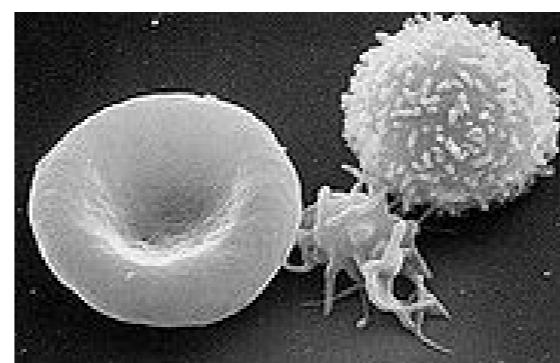
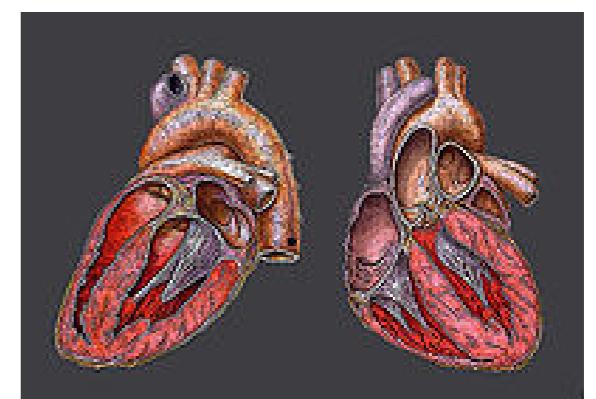
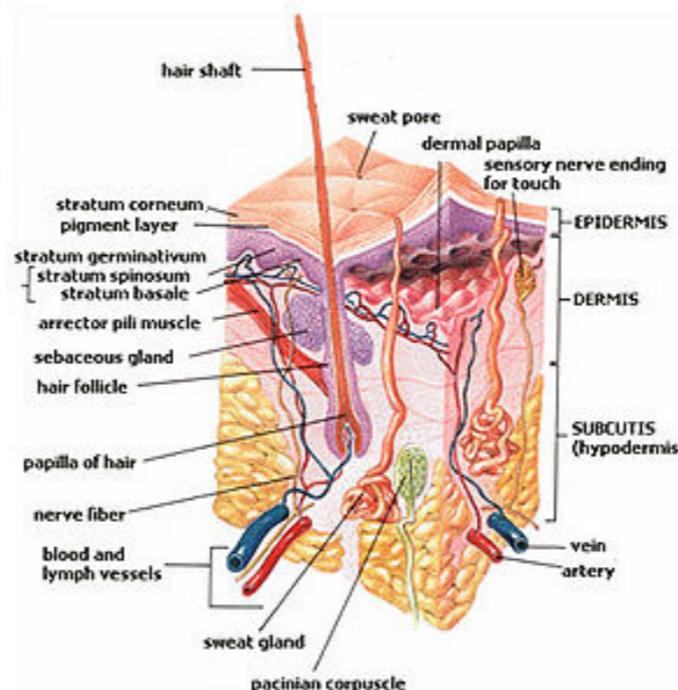
- DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
- DeepSea: Train model directly on mutational impact prediction
- Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
- ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
- DeepLIFT: Model interpretation based on neuron activation properties
- DanQ: Recurrent Neural Network for sequential data analysis

5. Guest Lecture: David Kelley on Basset and Deep Learning for Hi-C looping

1a. Basics of gene regulation

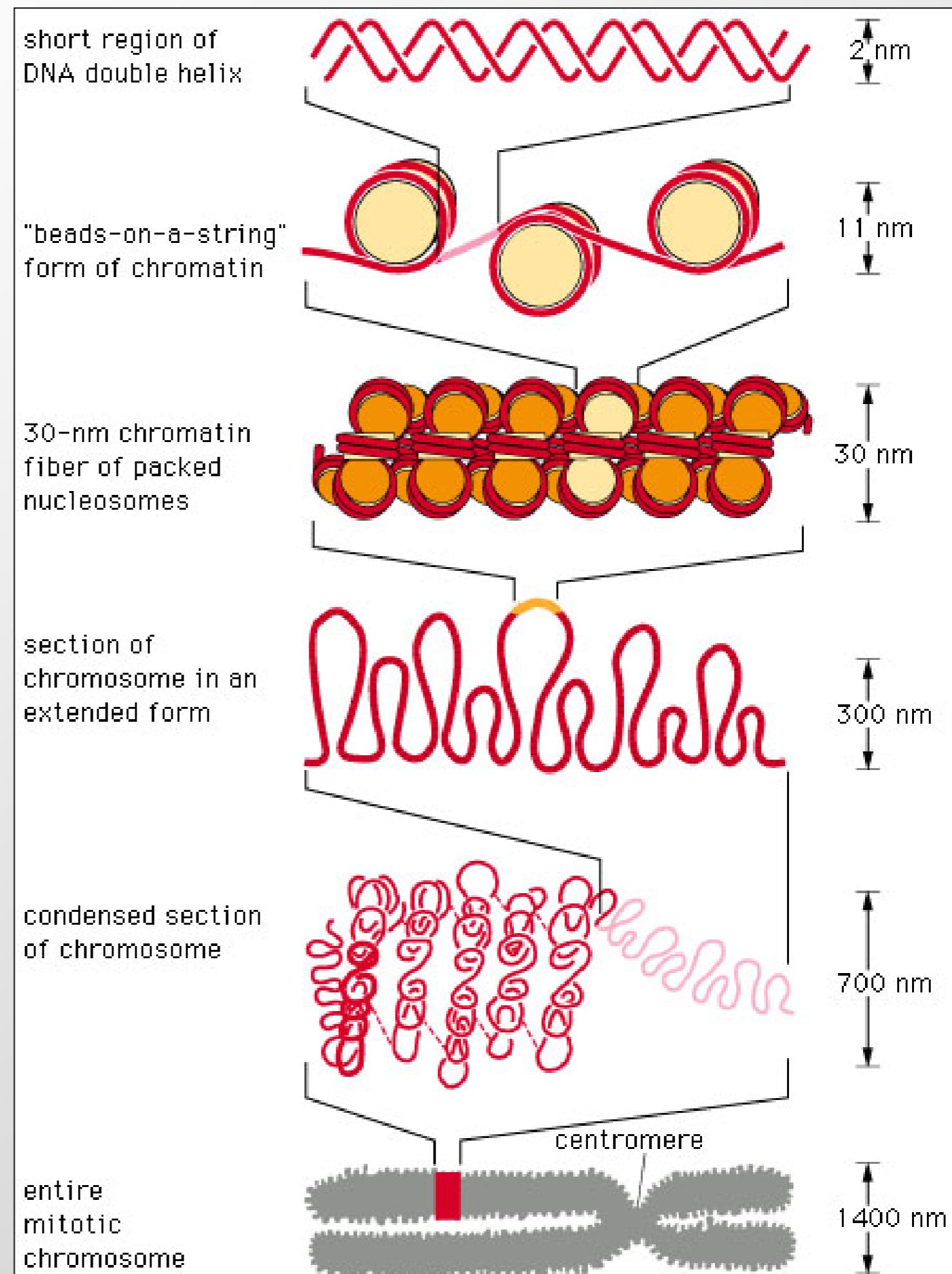
One Genome – Many Cell Types

ACCAAGTTACGACGGTCA
GGGTACTGATAACCCAA
ACCGTTGACCGCATTAA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTGAAAAAAAG
TTTGAGTTGGTTTTTC
ACGGTAGAACGTACCGT
TACCAAGTA

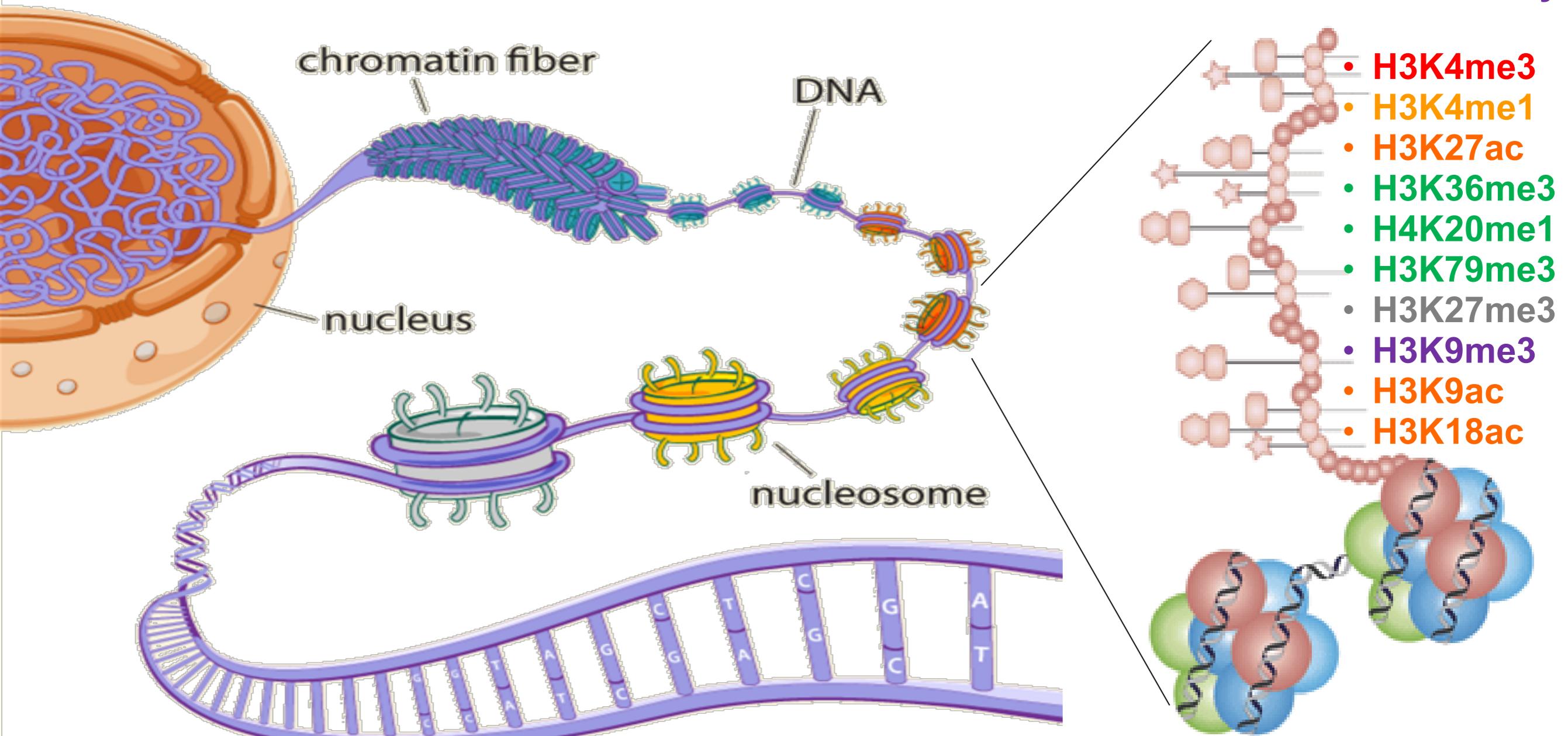
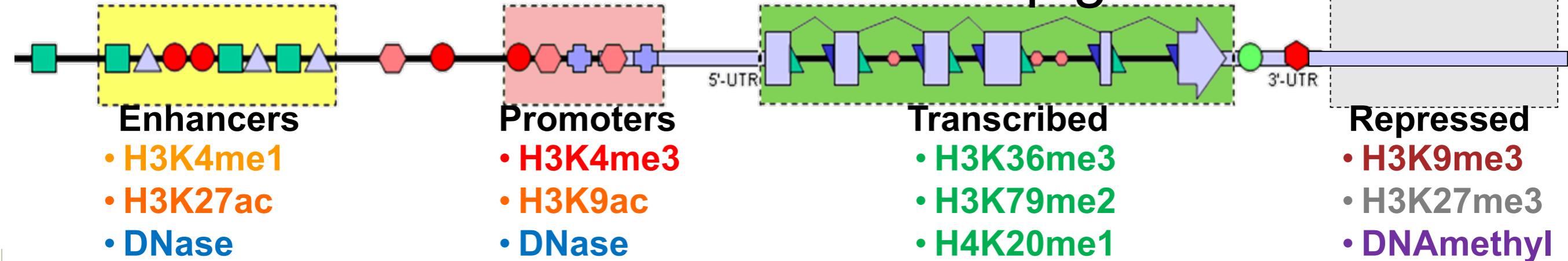


DNA packaging

- 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

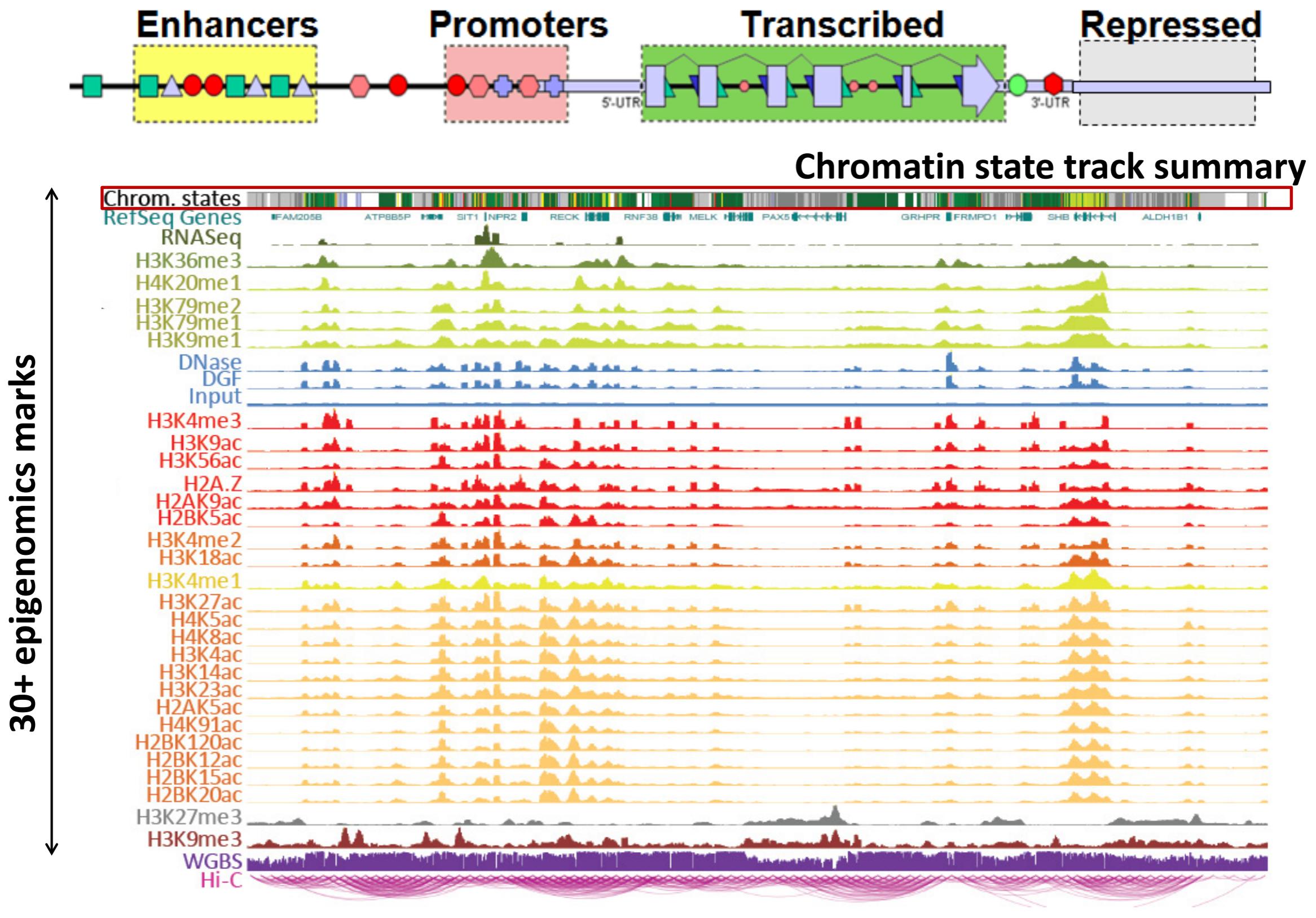


Combinations of marks encode epigenomic state



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

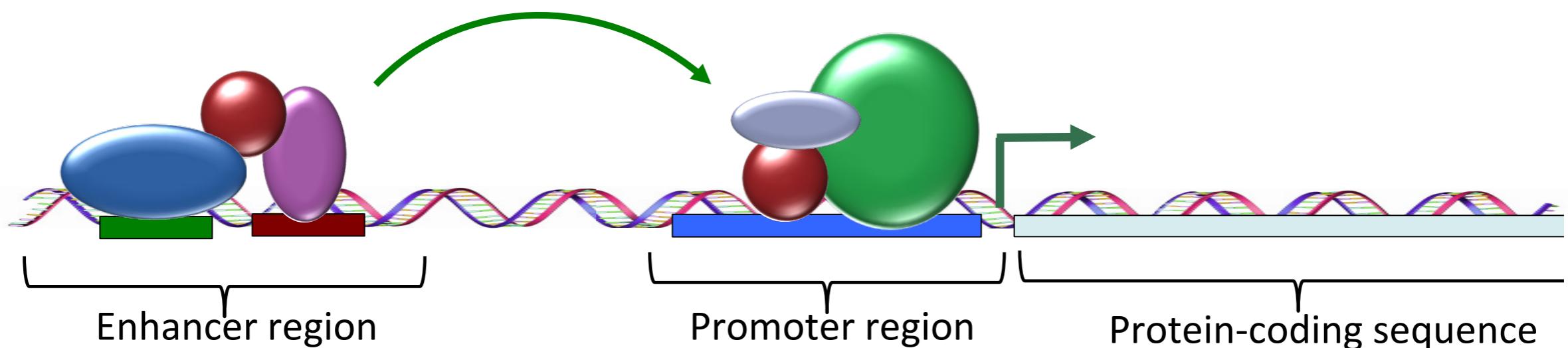
Summarize multiple marks into chromatin states



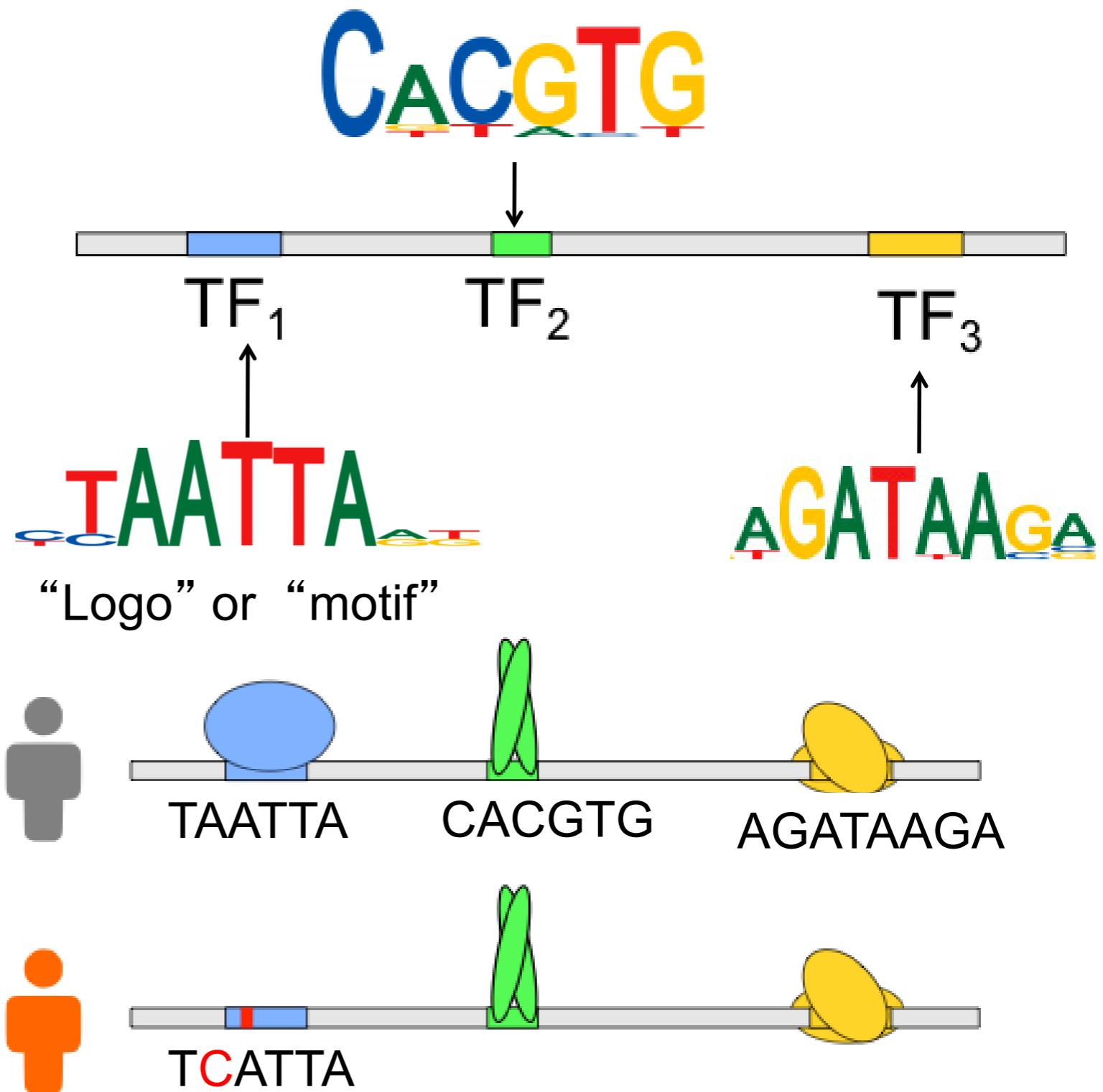
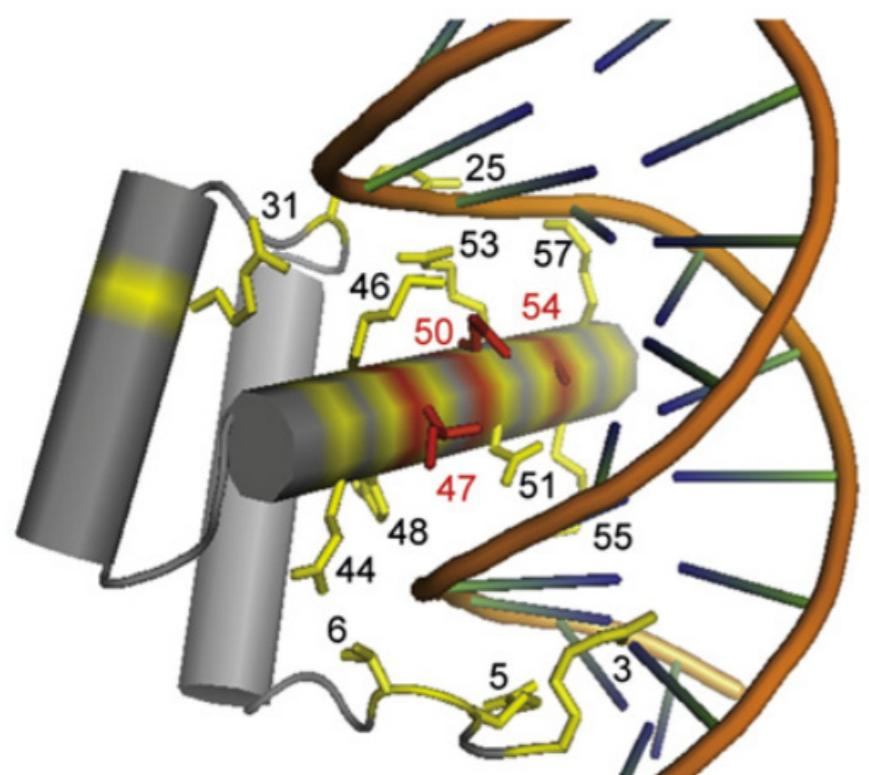
WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

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

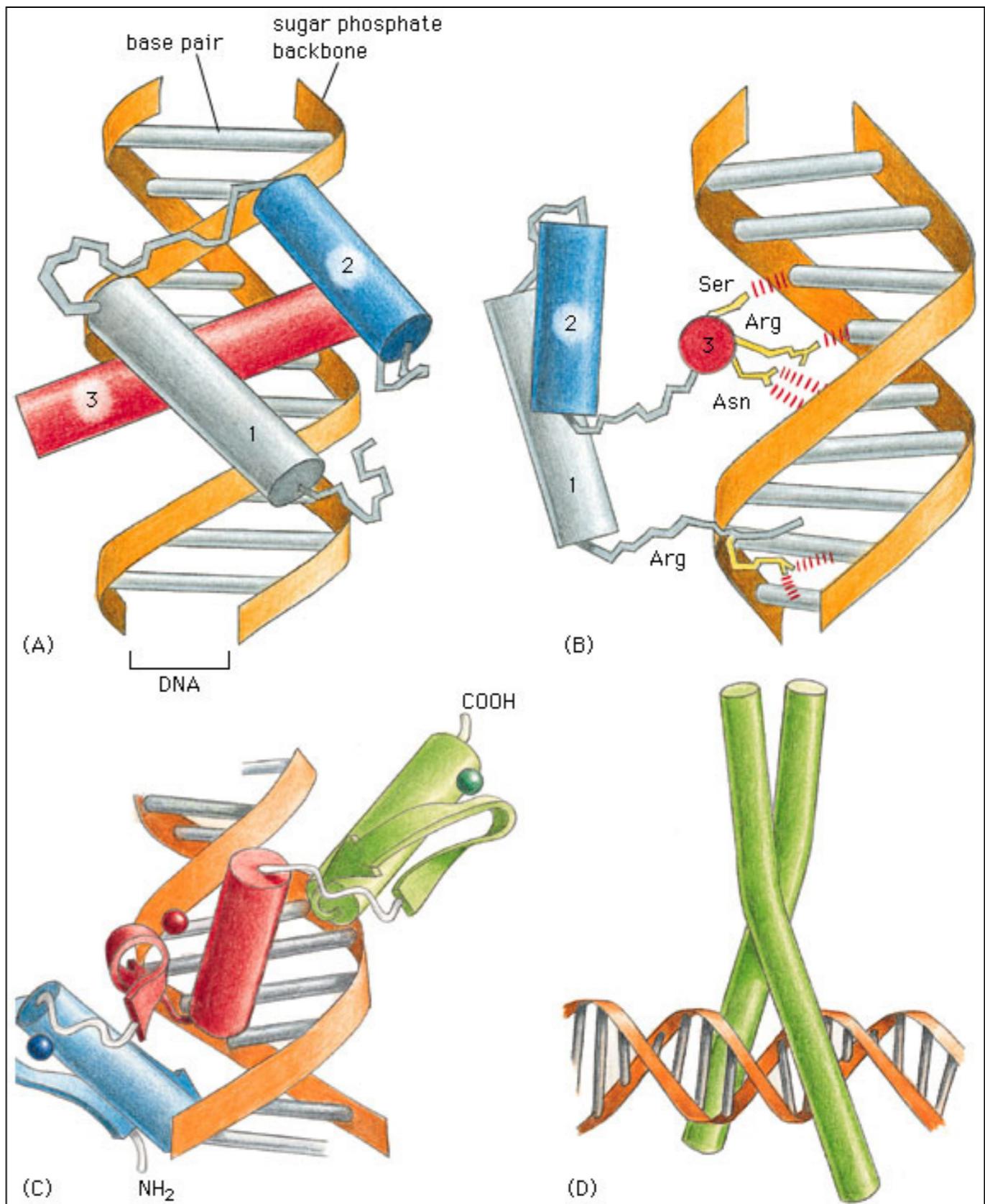


TFs use DNA-binding domains to recognize specific DNA sequences in the genome



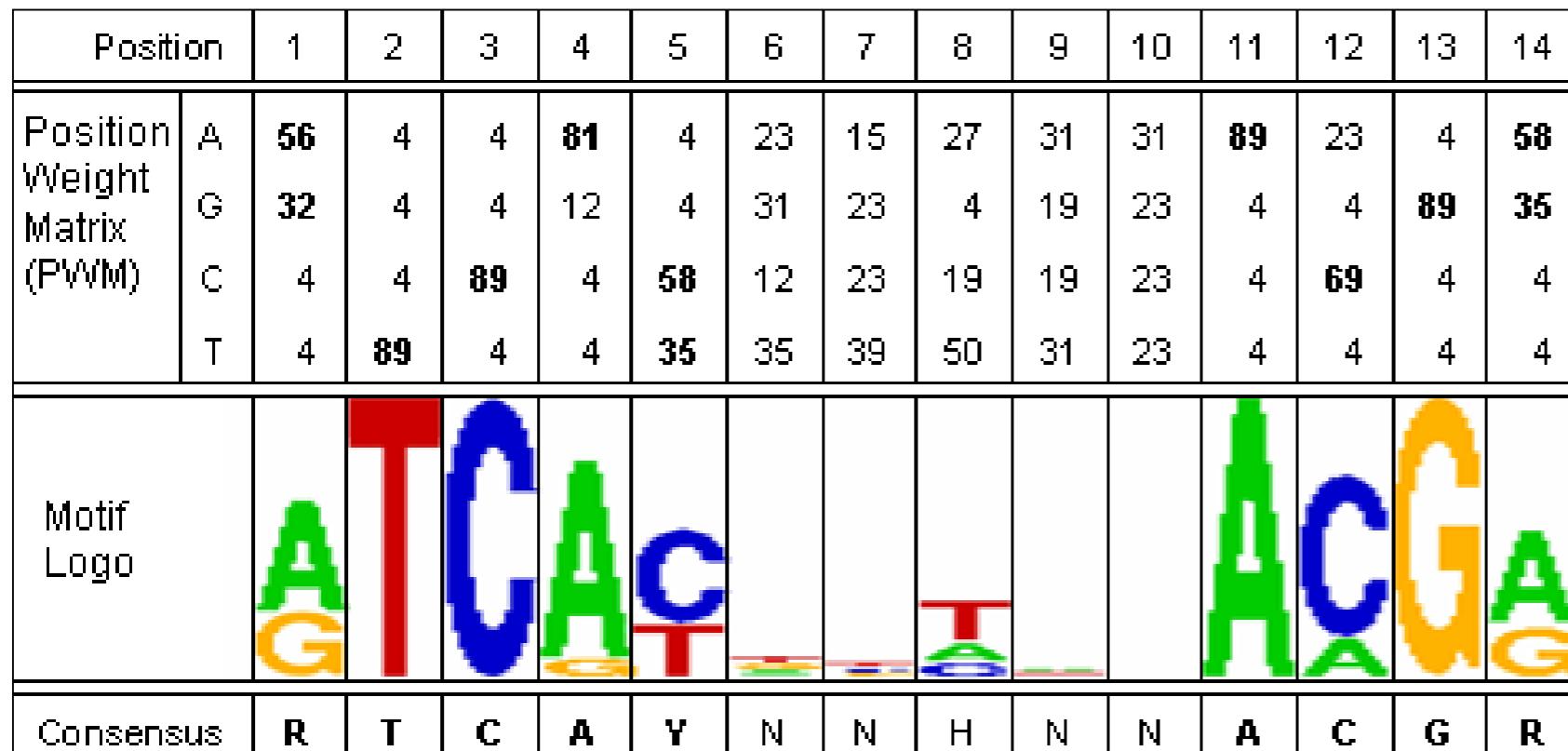
Regulator structure \leftrightarrow recognized motifs

- 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



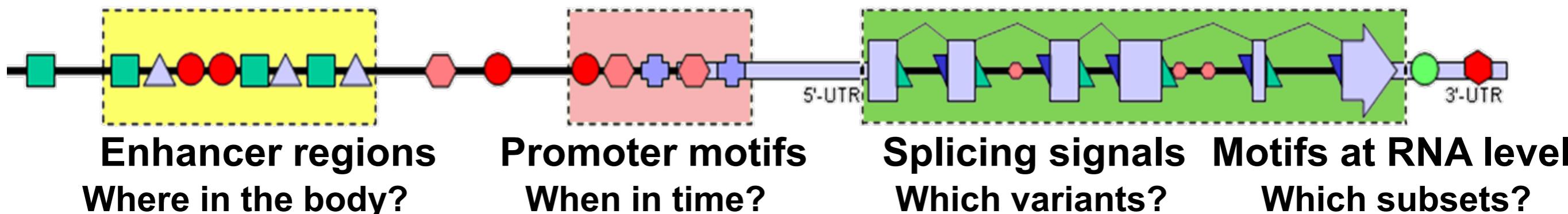
Motifs summarize TF sequence specificity

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	ATCATTGCCACG	
CHA1	catabolic L-serine dehydratase	-280	-254	A ATCACCGCGAACG GA	
ENO2	Enolase	-470	-461	ggcgttat GTCACTAACGACG tgccacca	
HMR	silencer	-256	-283	ATCAATAC ATCATAAAATACG AACGATC	
LPD1	lipoamide dehydrogenase	-288	-300	gat ATCAAAATTAAACG 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	ATCACTATATAACG 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	



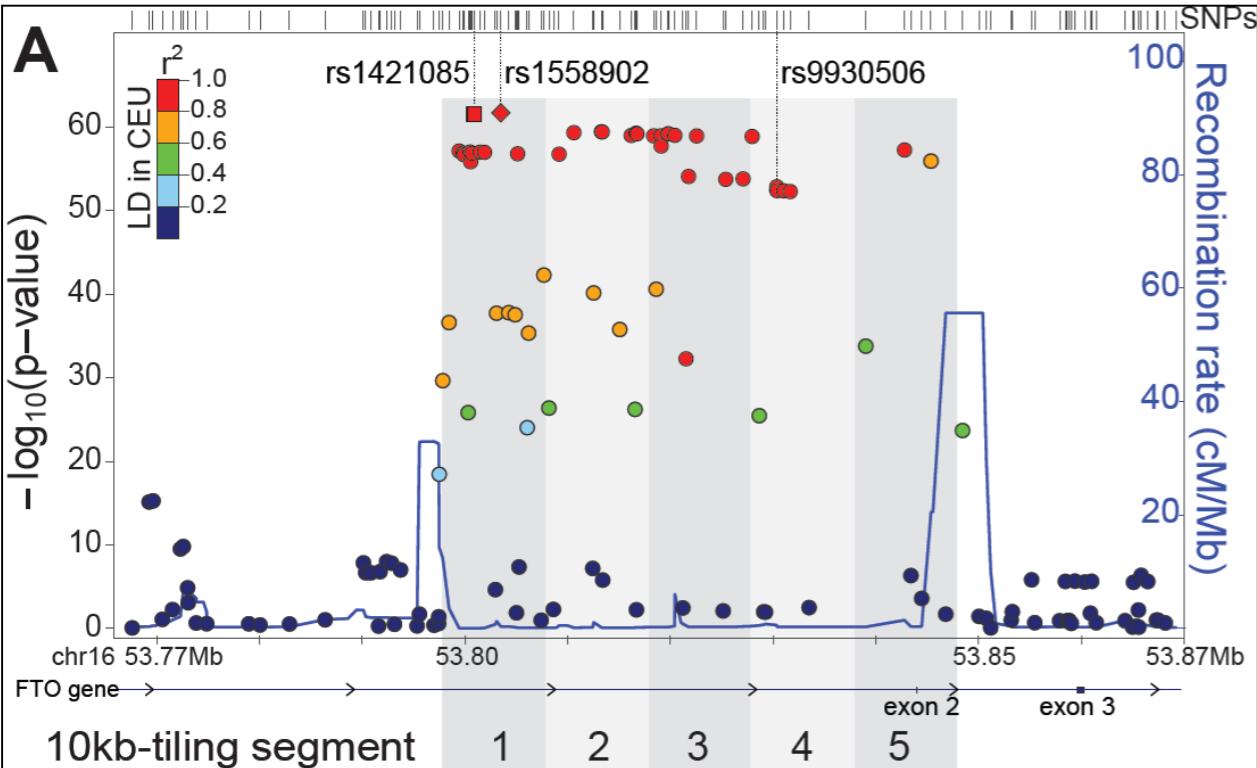
- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
 - Independence
 - Fixed spacing

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

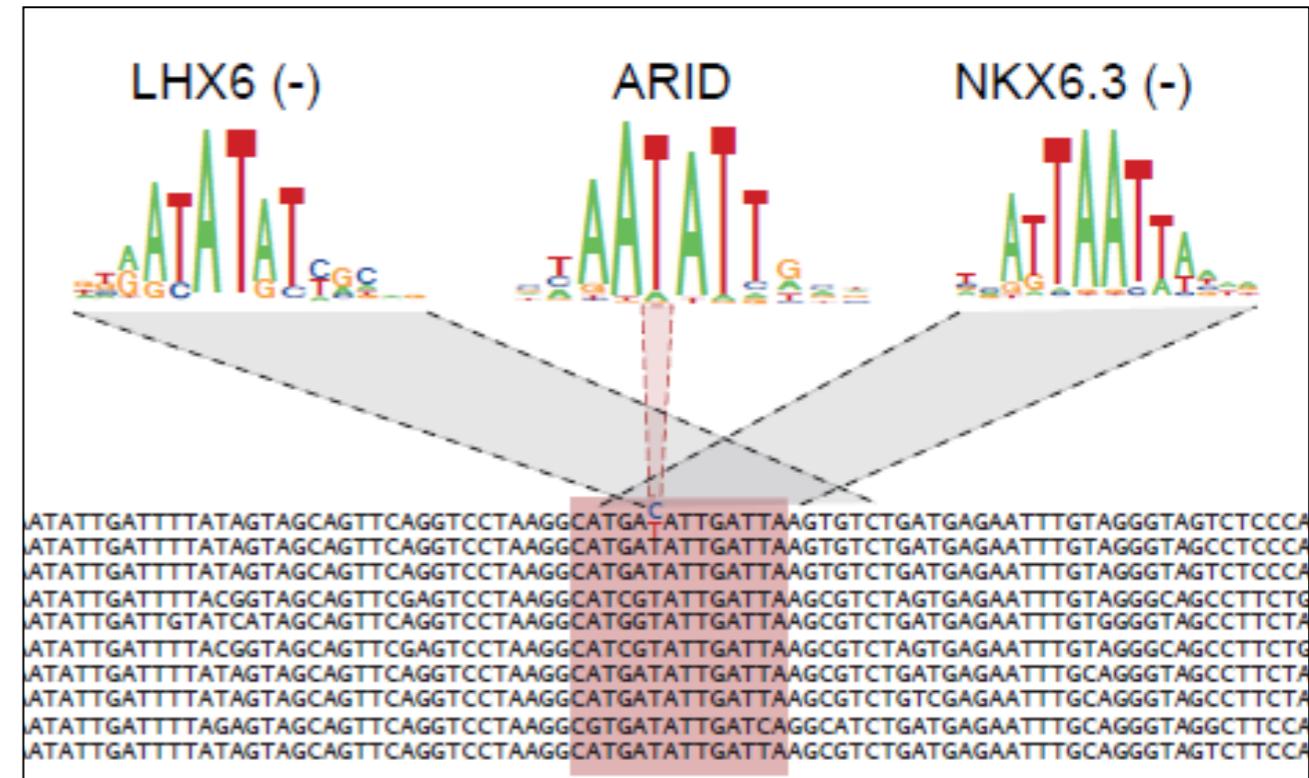


- The parts list: ~20-30k genes
 - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
 - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
 - Combinatorial coding of 'unique tags'
 - Data-centric encoding 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

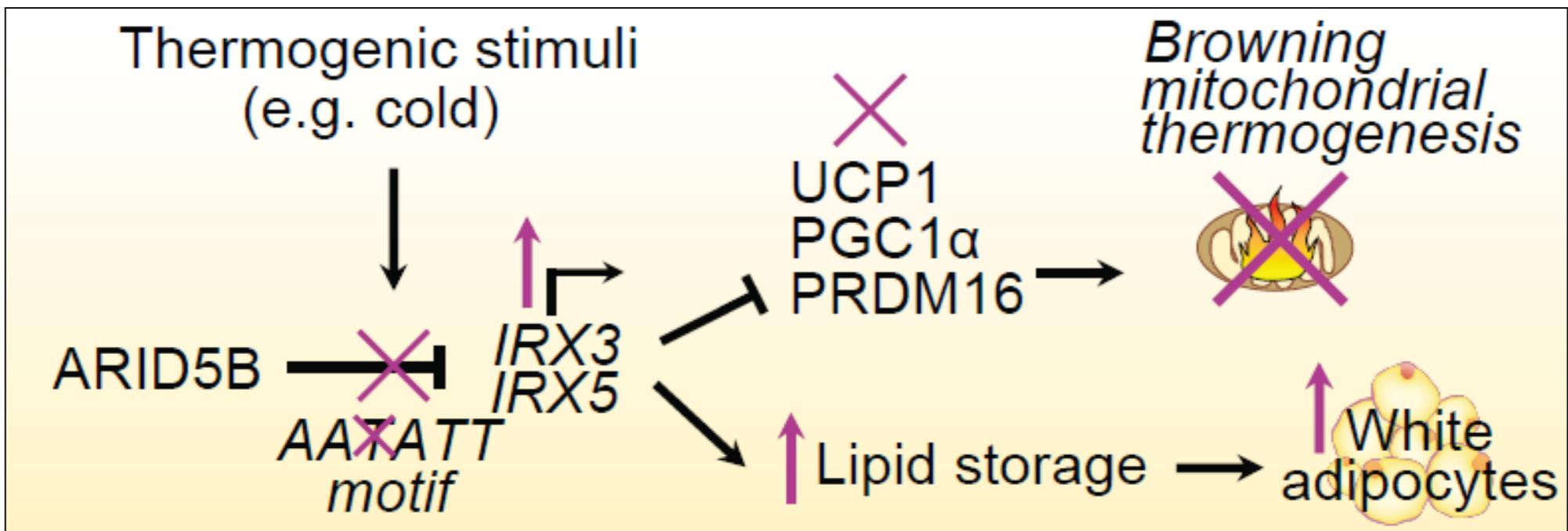
Disrupted motif at the heart of FTO obesity locus



**Strongest association
with obesity**



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

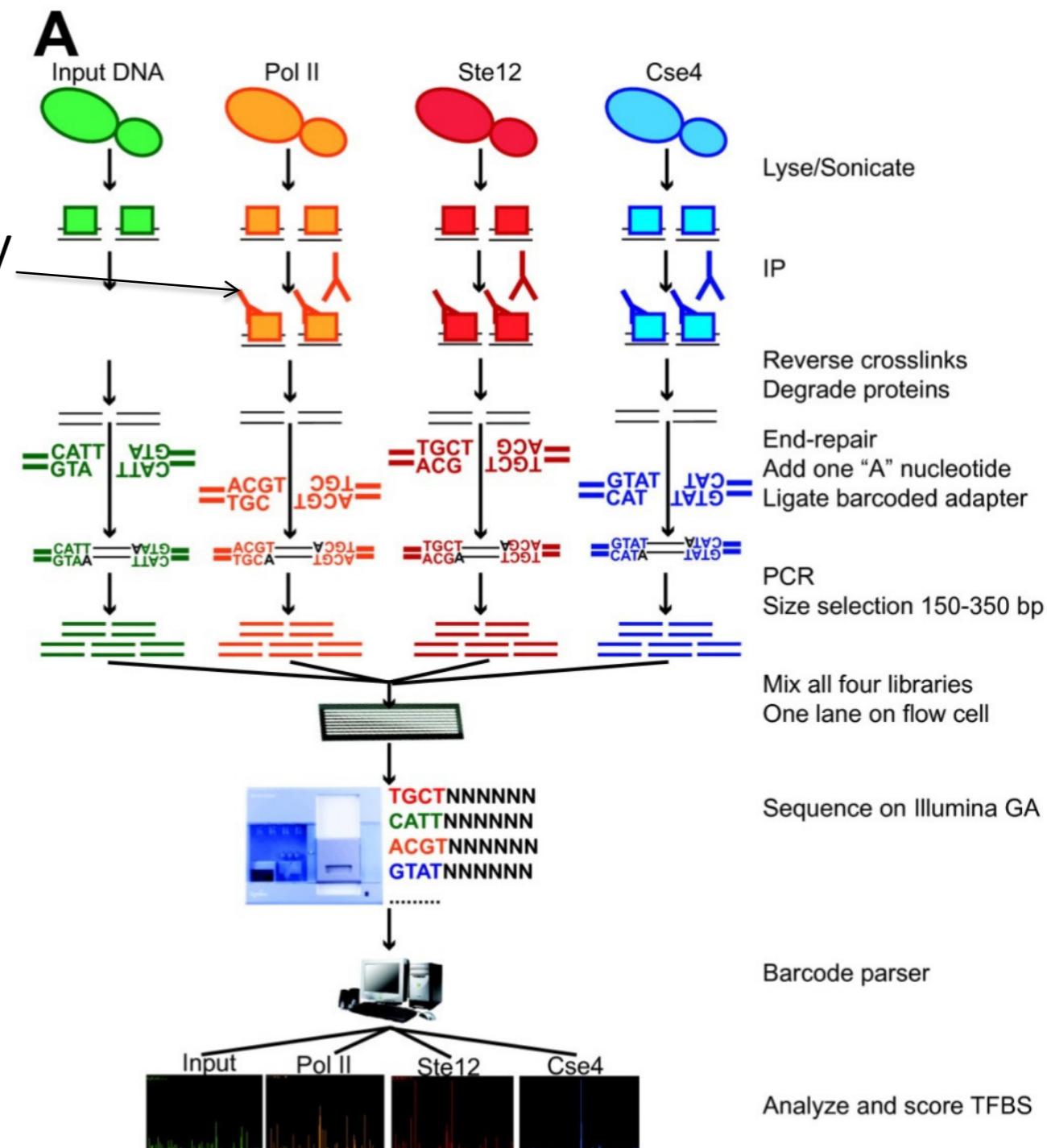
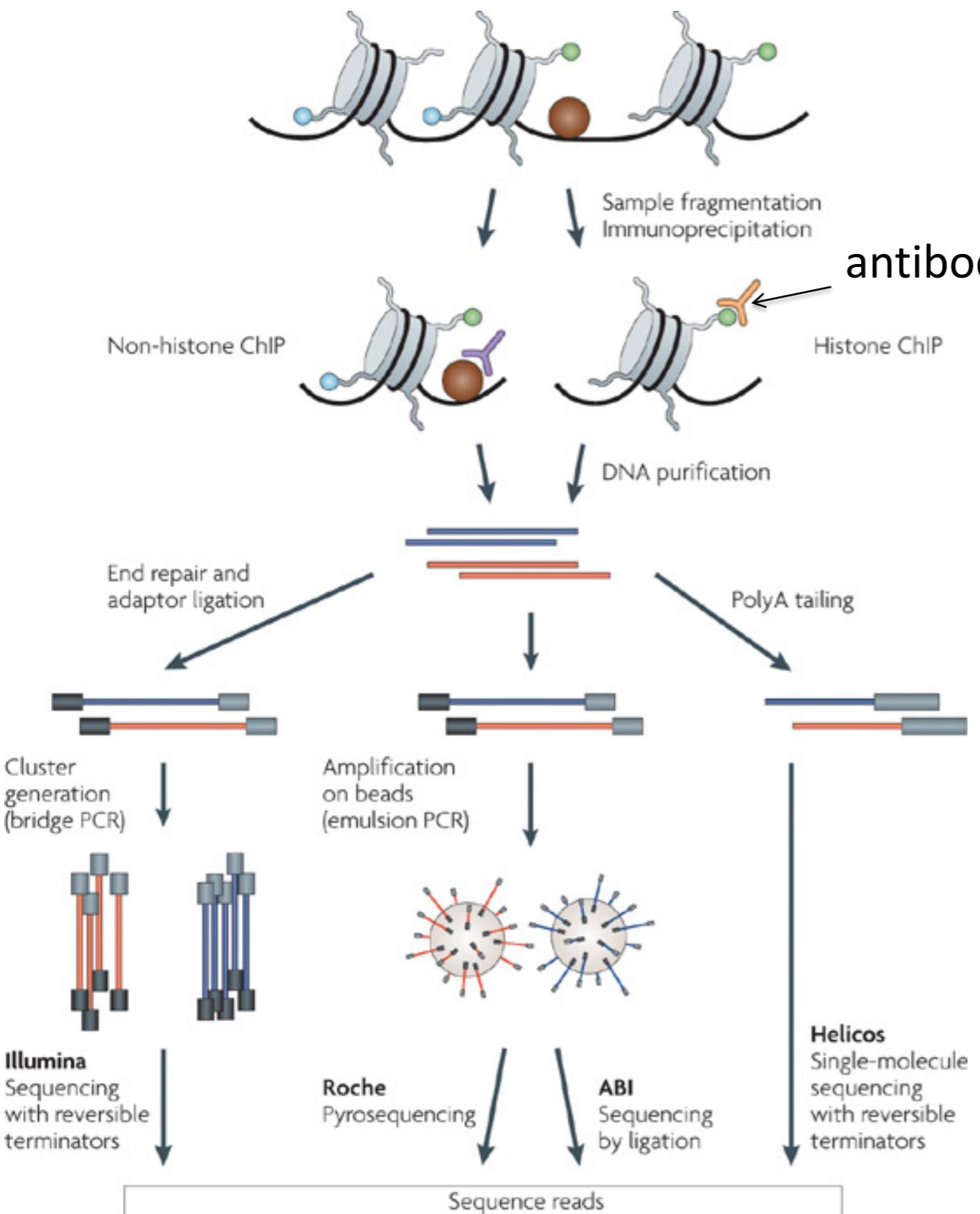


Restoring motif restores thermogenesis

1b. Technologies for probing gene regulation

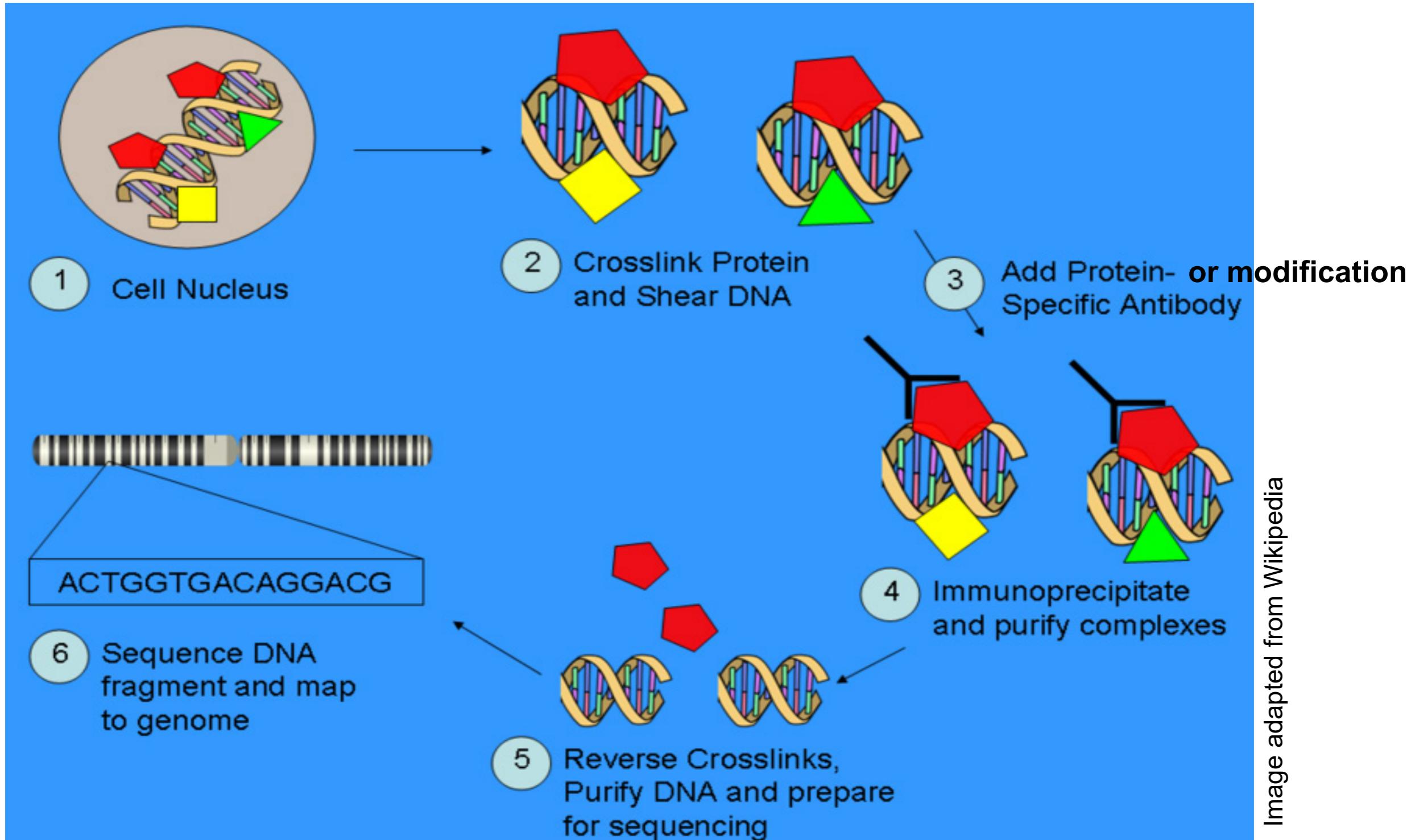
Mapping regulator binding: ChIP-seq

(Chromatin immunoprecipitation followed by sequencing) TF=transcription factor



Bar-coded multiplexed sequencing

ChIP-chip and ChIP-Seq technology overview

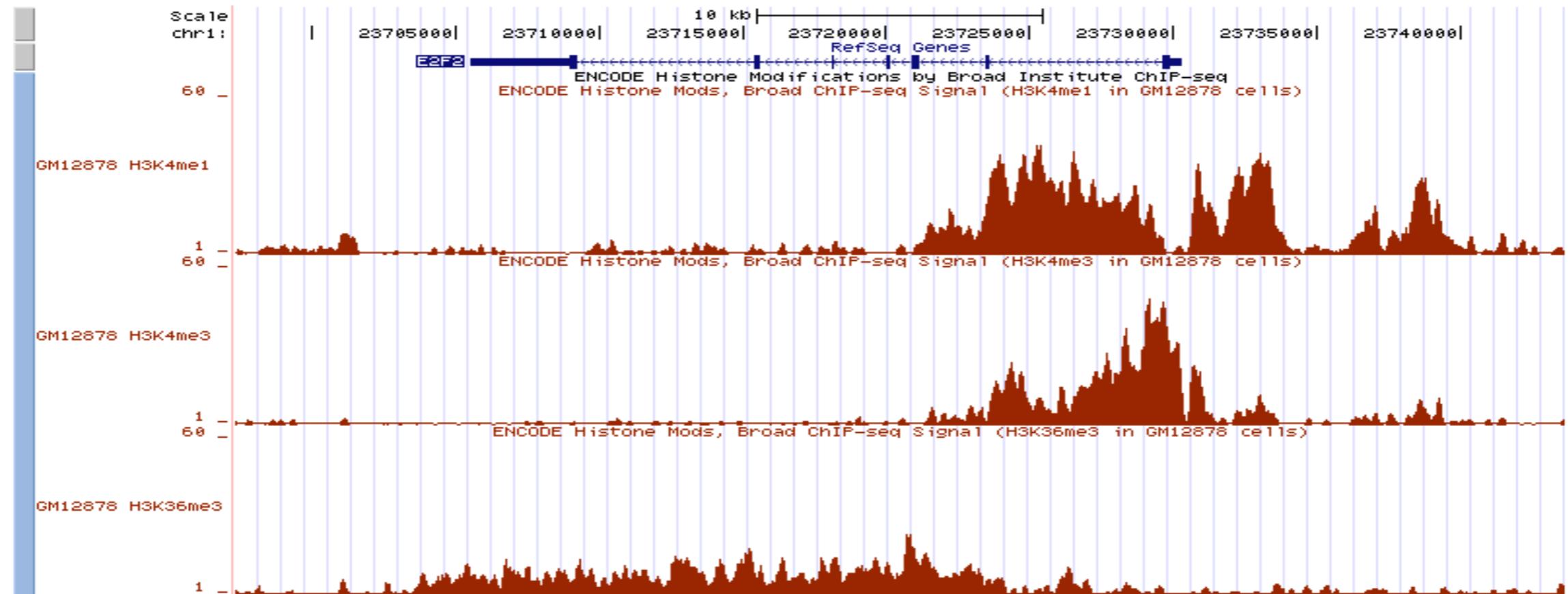


Modification-specific antibodies → Chromatin Immuno-Precipitation

followed by: ChIP-chip: array hybridization

ChIP-Seq: Massively Parallel Next-gen Sequencing

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

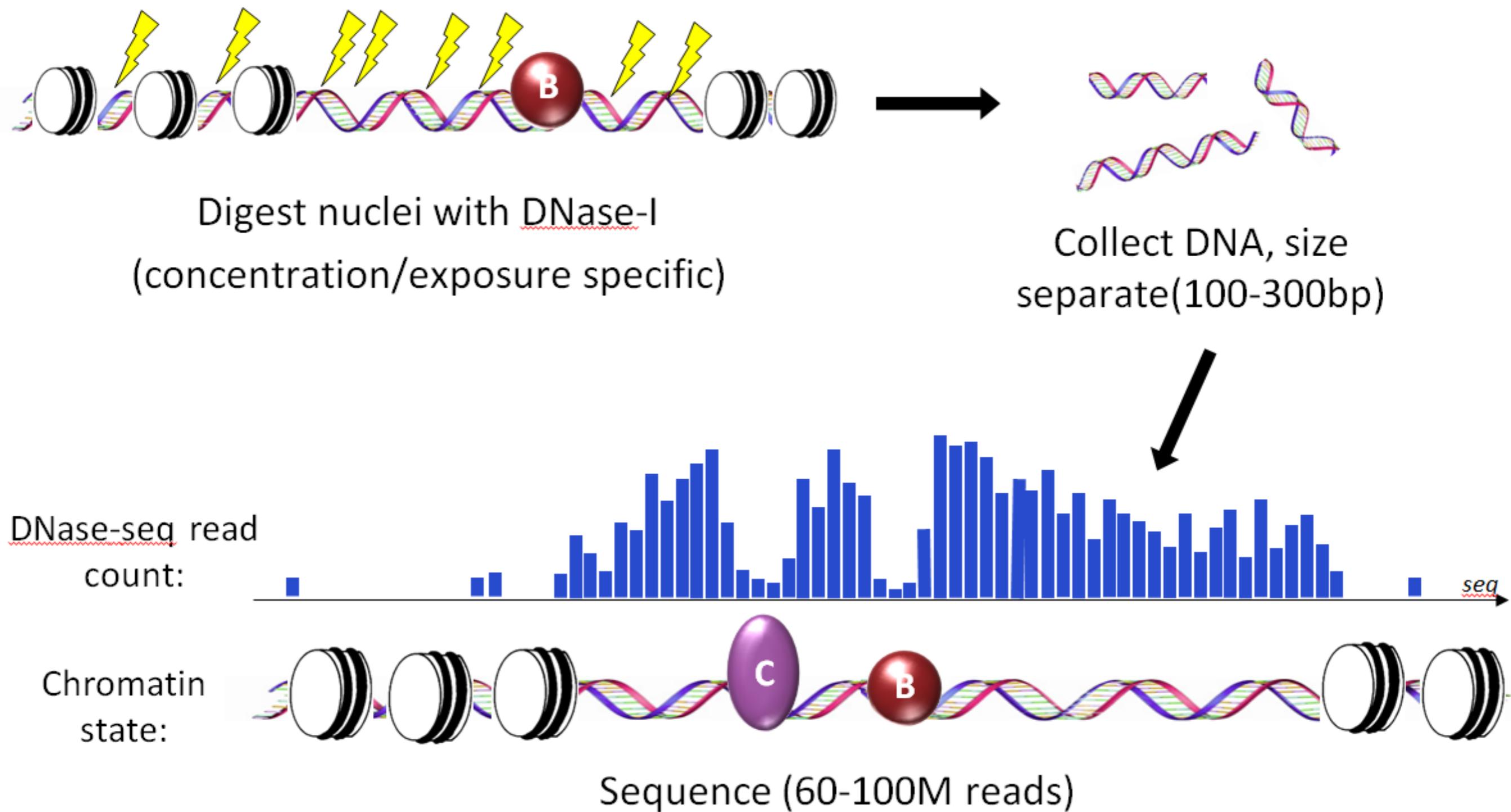


- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth.
Typically on the order of 10 million mapped reads.

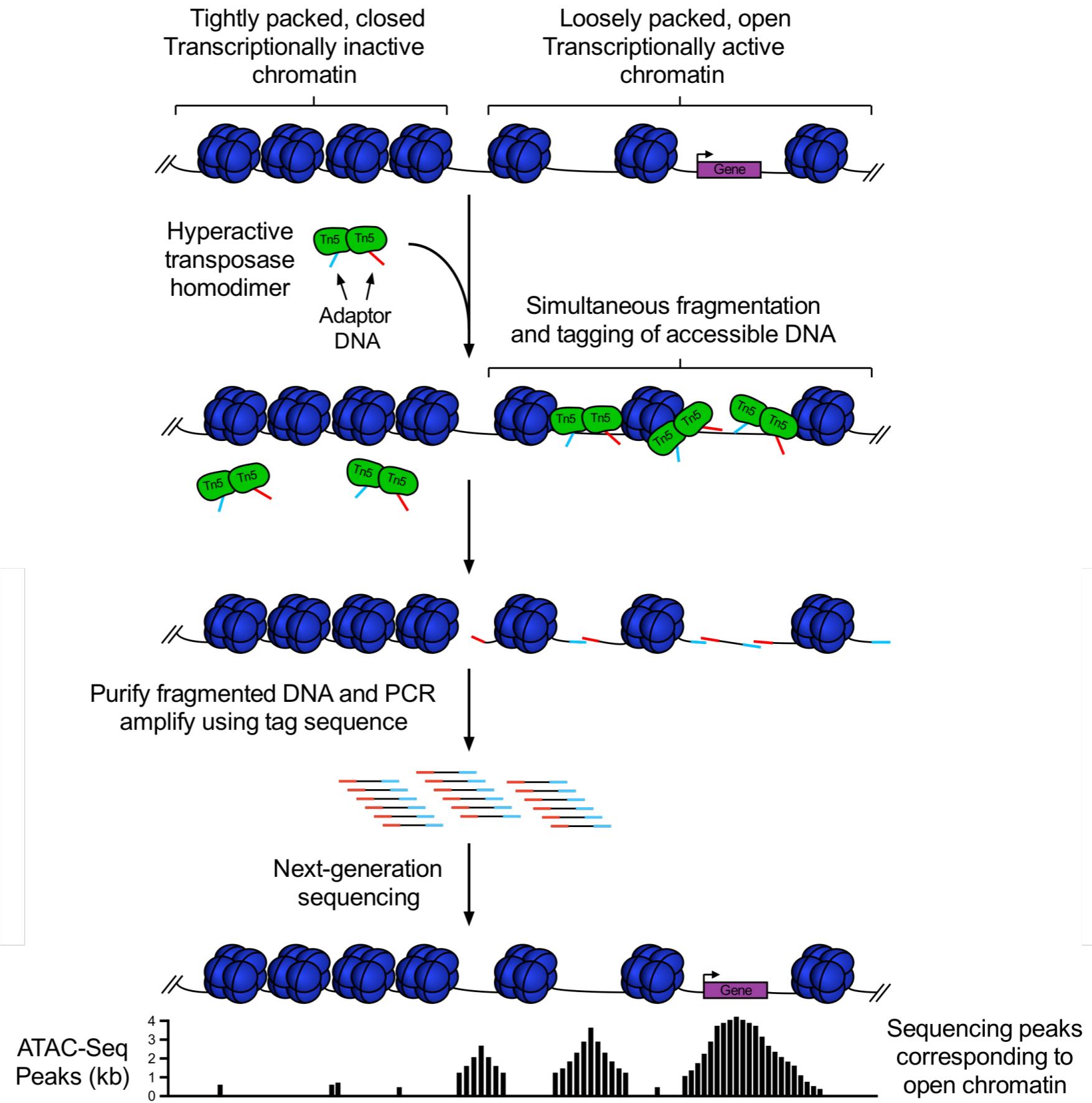
Chromatin accessibility can reveal TF binding

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

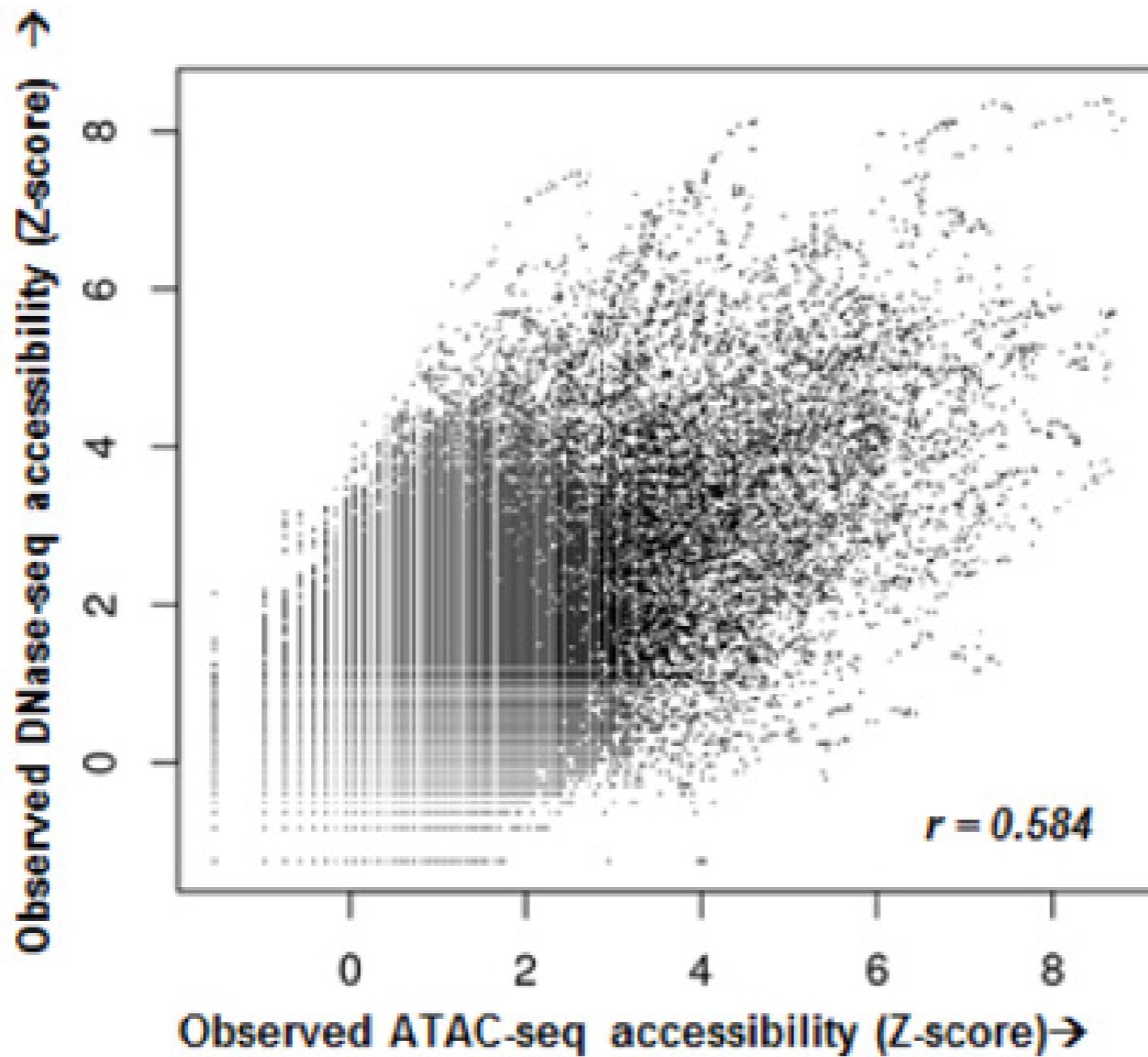


Assay for Transposase-Accessible Chromatin (ATAC-seq)



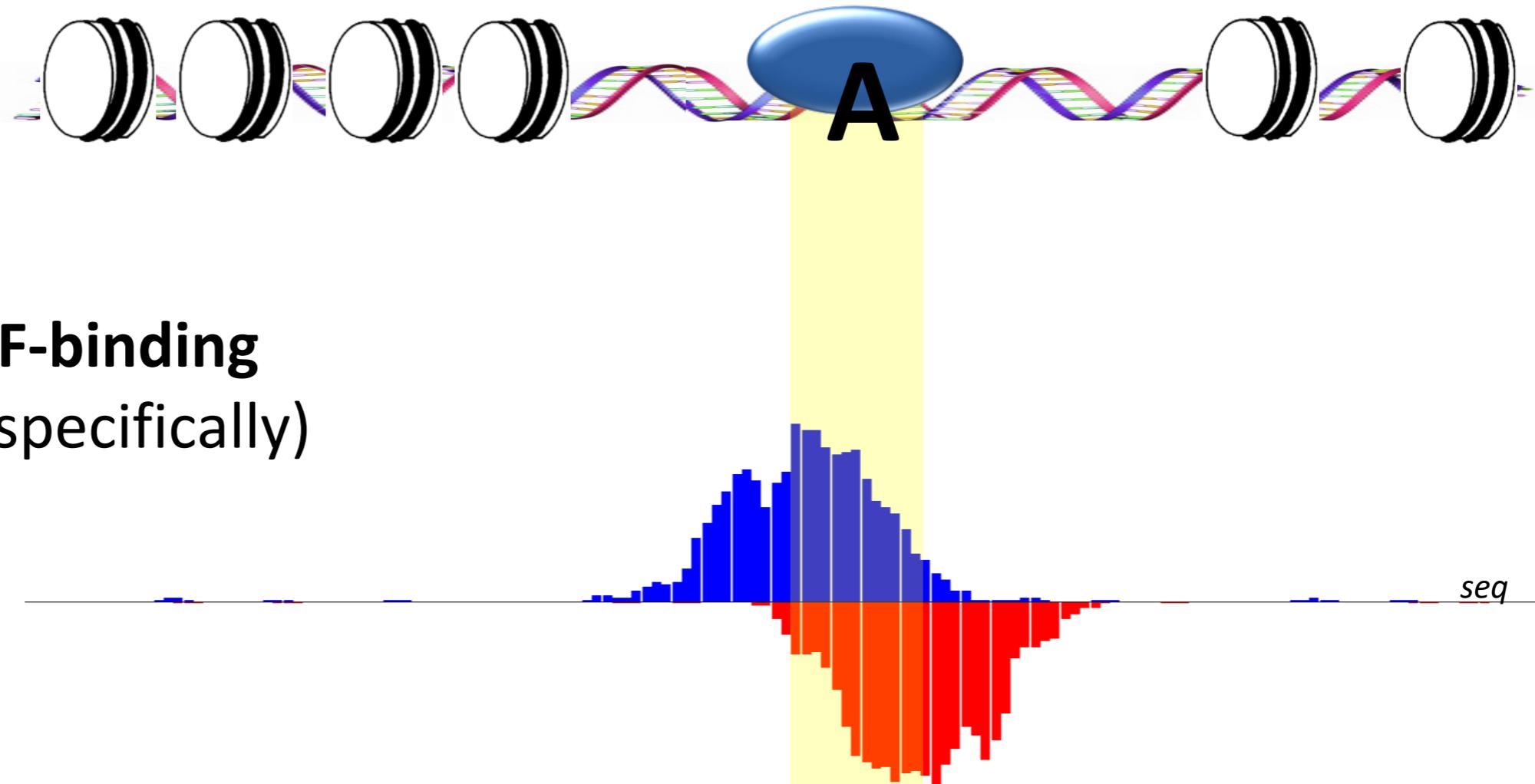
ATAC-seq and DNase-seq are not identical

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

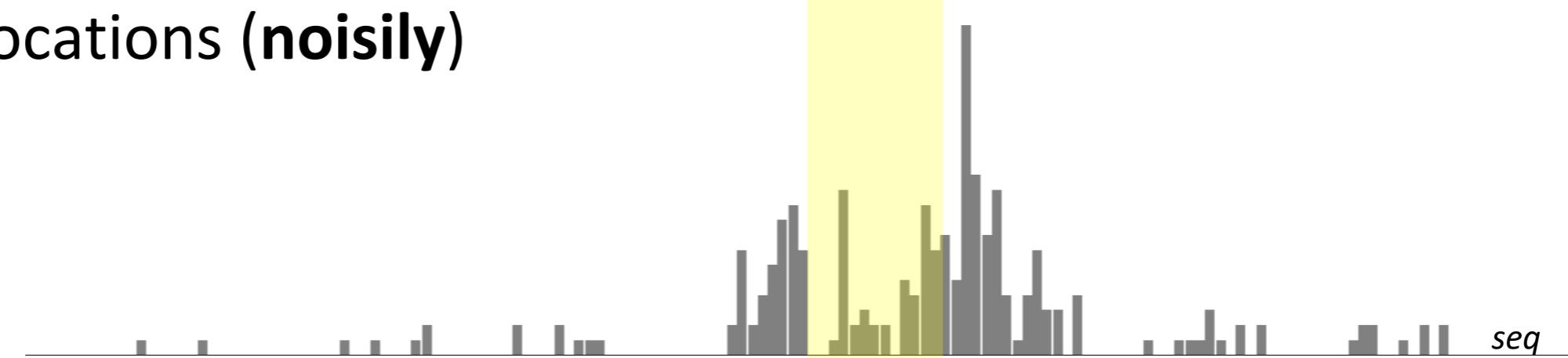


DNase-seq is less defined evidence than ChIP-seq

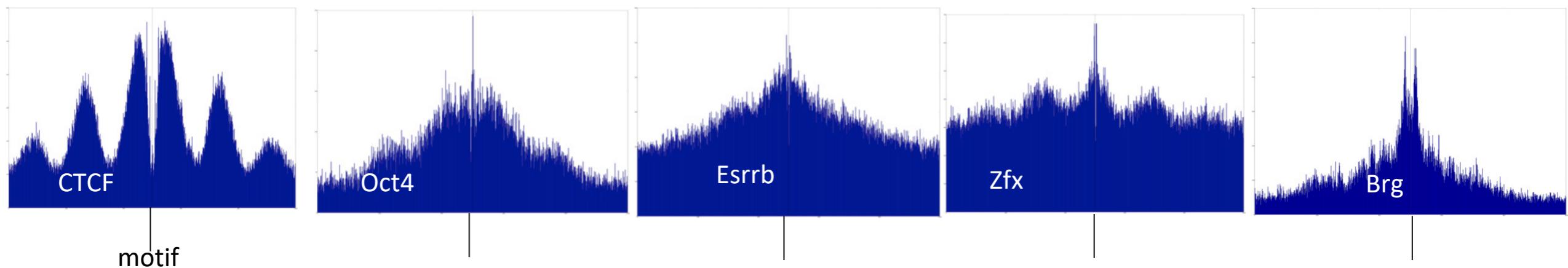
ChIP-seq reports **TF-binding** locations regions (specifically)



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

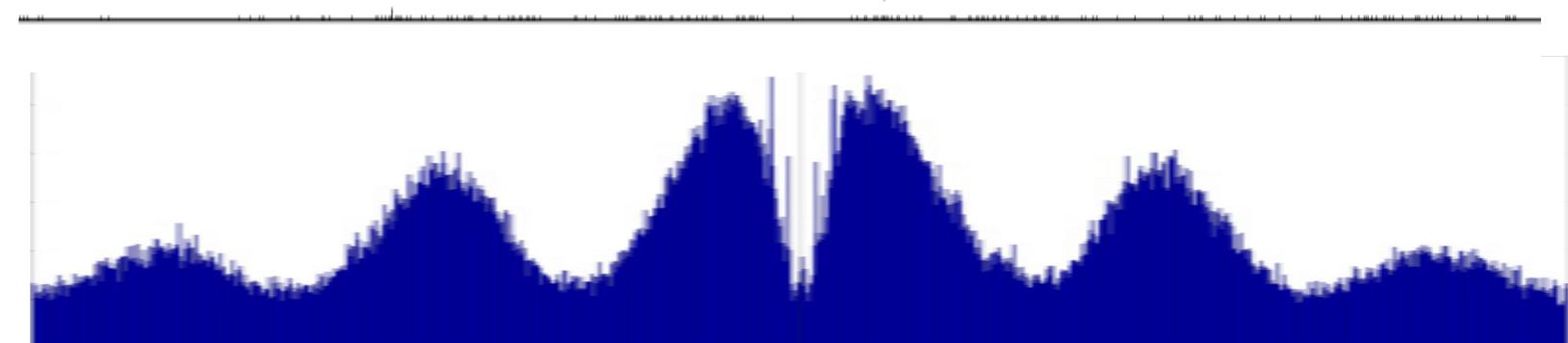


Bound factors leave distinct DNase-seq profiles



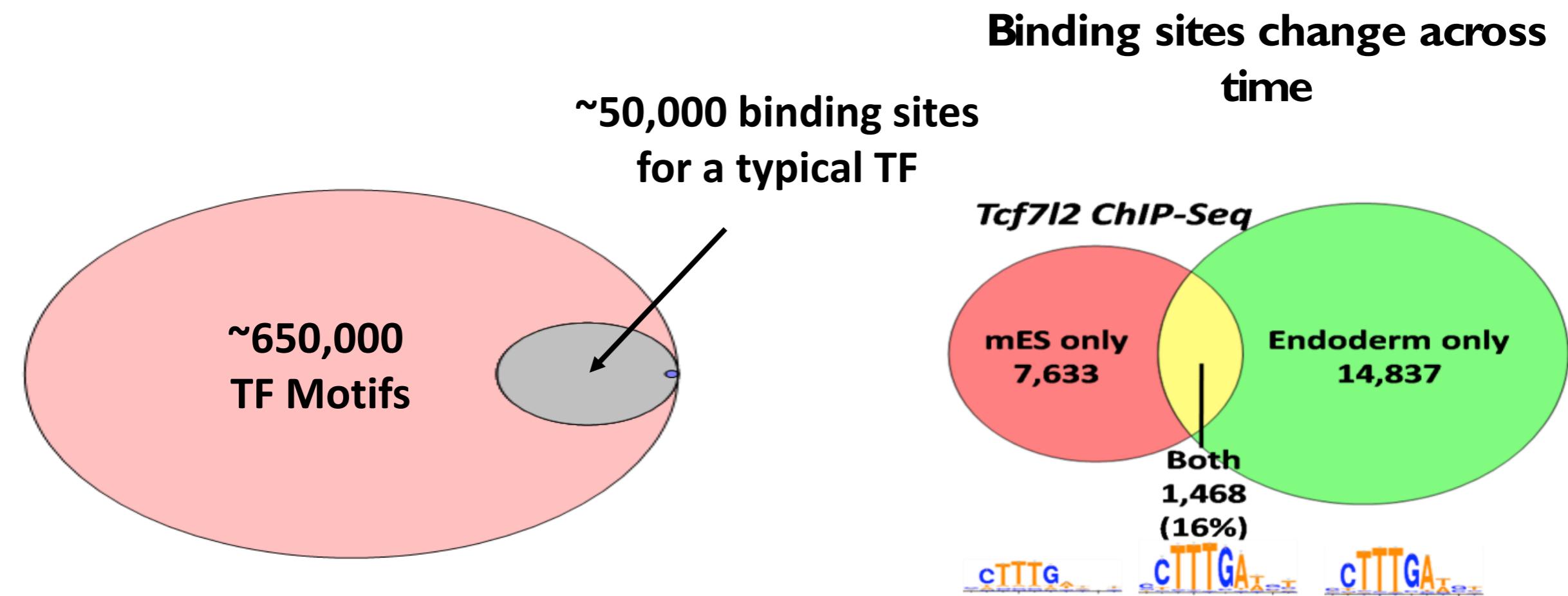
Individual binding site prediction is difficult

Individual CTCF:



Aggregate CTCF:

Motifs can predict TF binding



Chromatin accessibility 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
- Three-dimensional chromatin structure, Hi-C, ChIA-PET, TADs, Loop Extrusion

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 \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow Motif matches

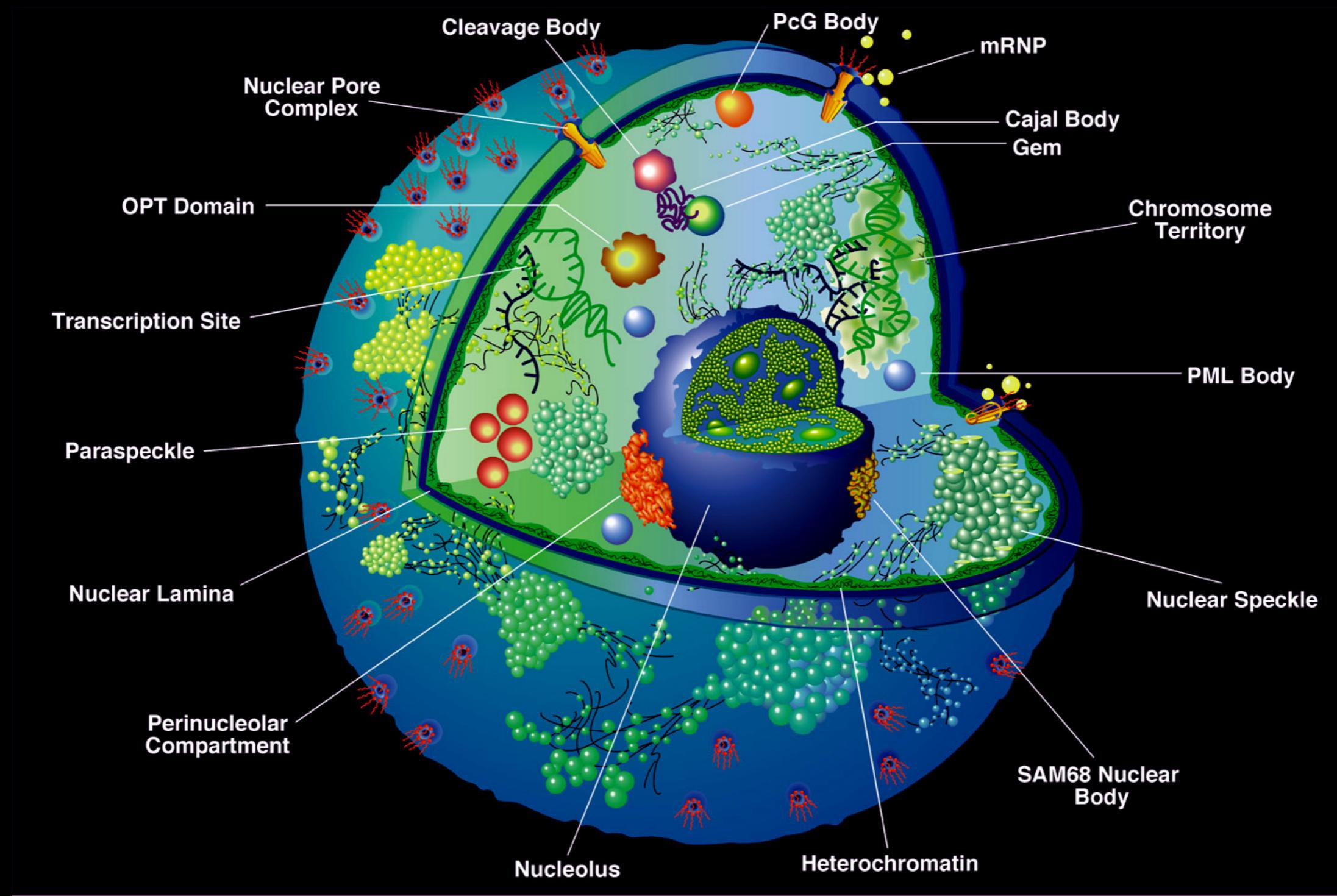
4. Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures

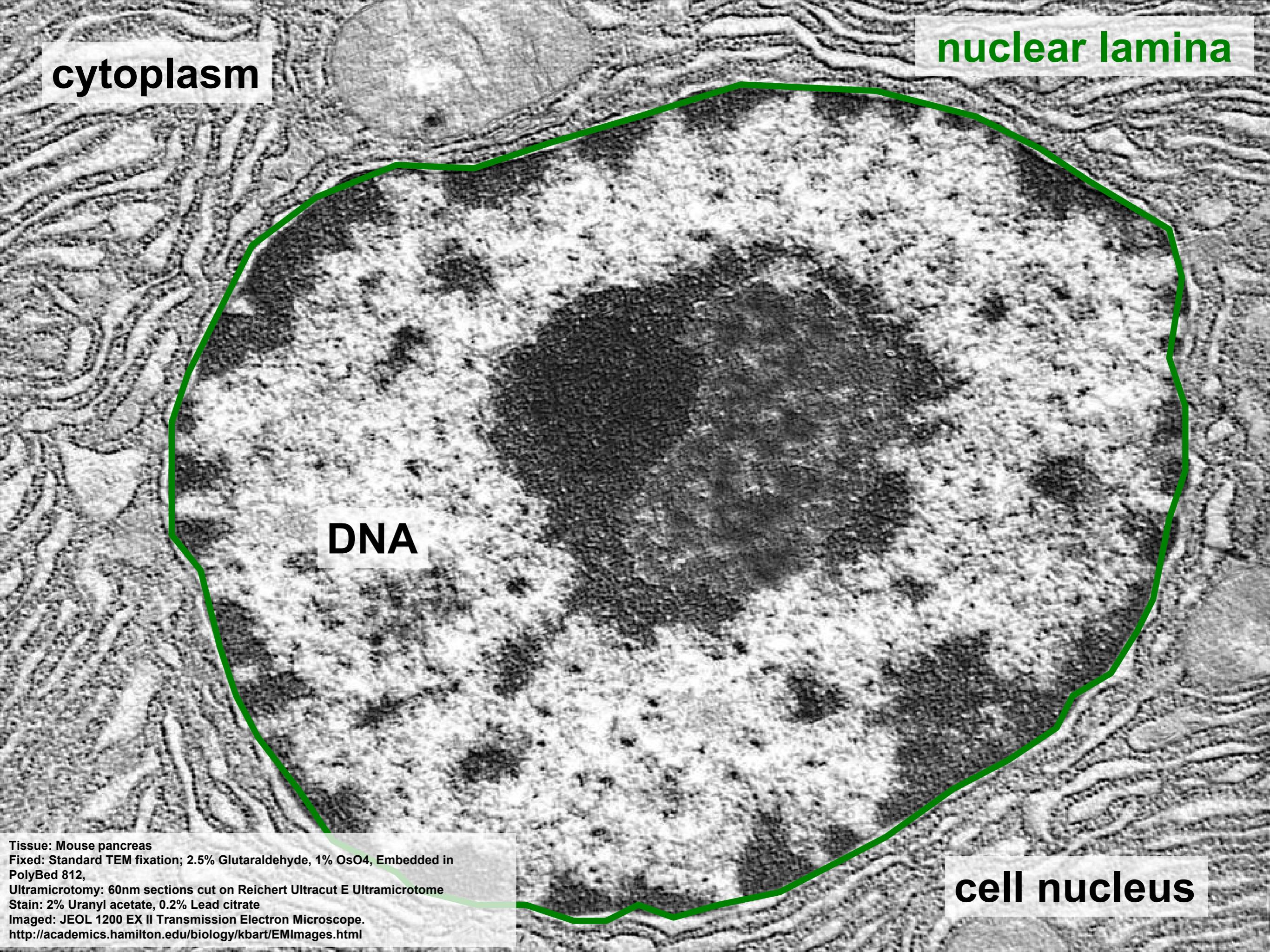
- DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
- DeepSea: Train model directly on mutational impact prediction
- Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
- ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
- DeepLIFT: Model interpretation based on neuron activation properties
- DanQ: Recurrent Neural Network for sequential data analysis

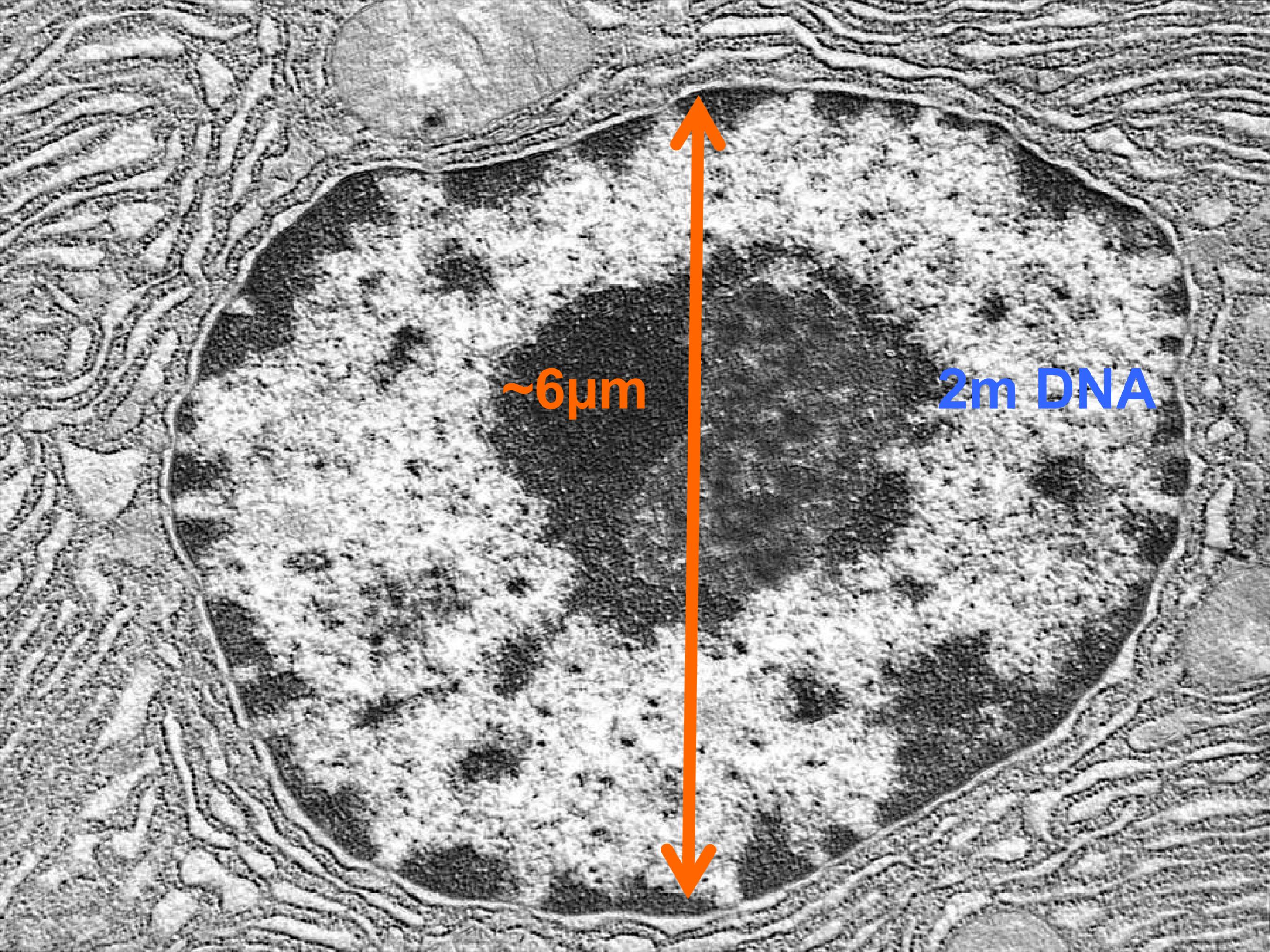
5. Guest Lecture: David Kelley on Basset and Deep Learning for Hi-C looping

1c. 3D chromatin structure

A model of the (mammalian) nucleus



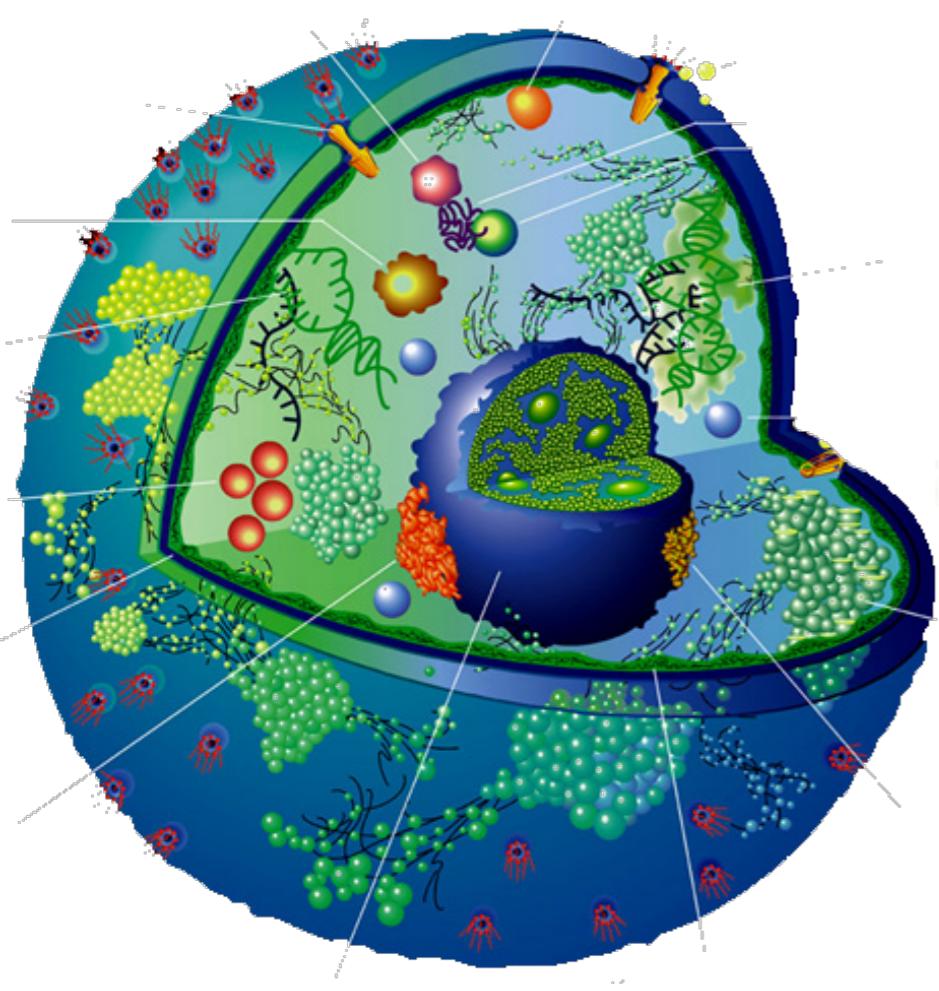




~6 μ m

2m DNA

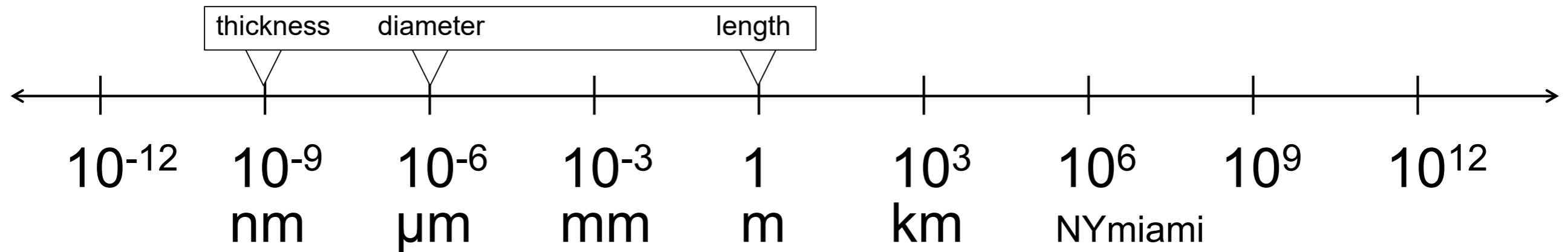
nucleus



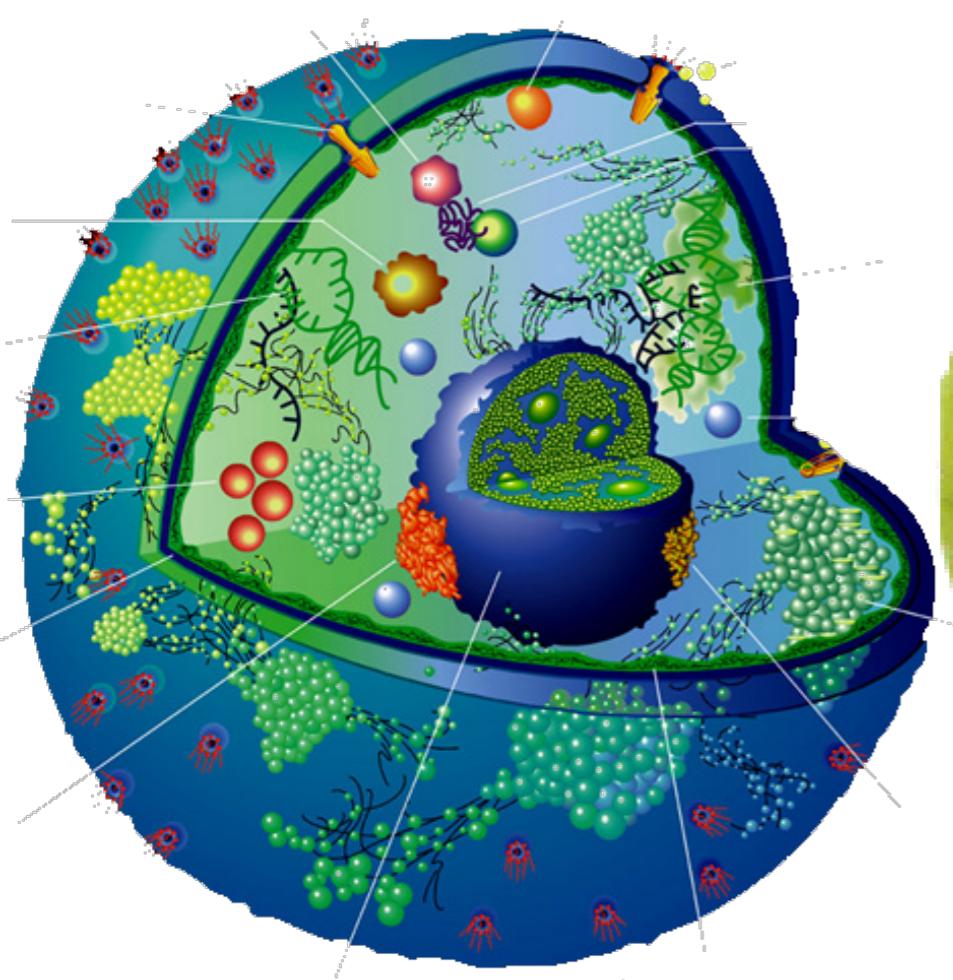
diameter: $6\mu\text{m}$

thread length: 2m

thread thickness: 2.5nm



nucleus

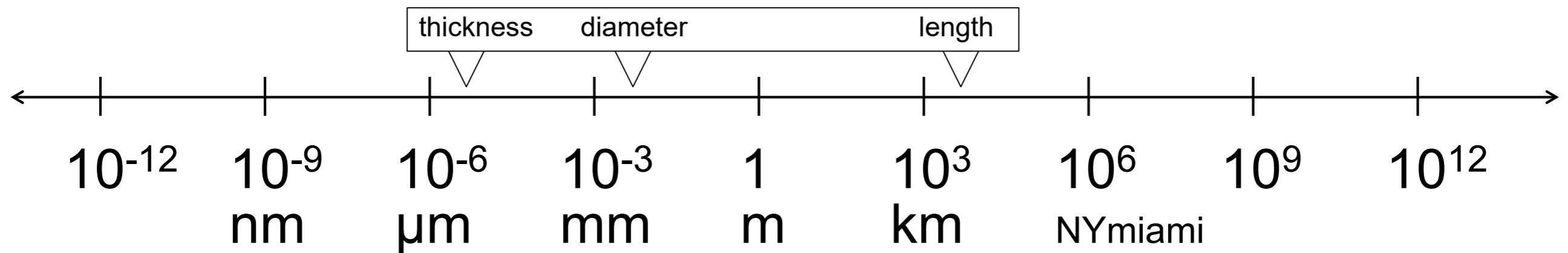


tennis ball



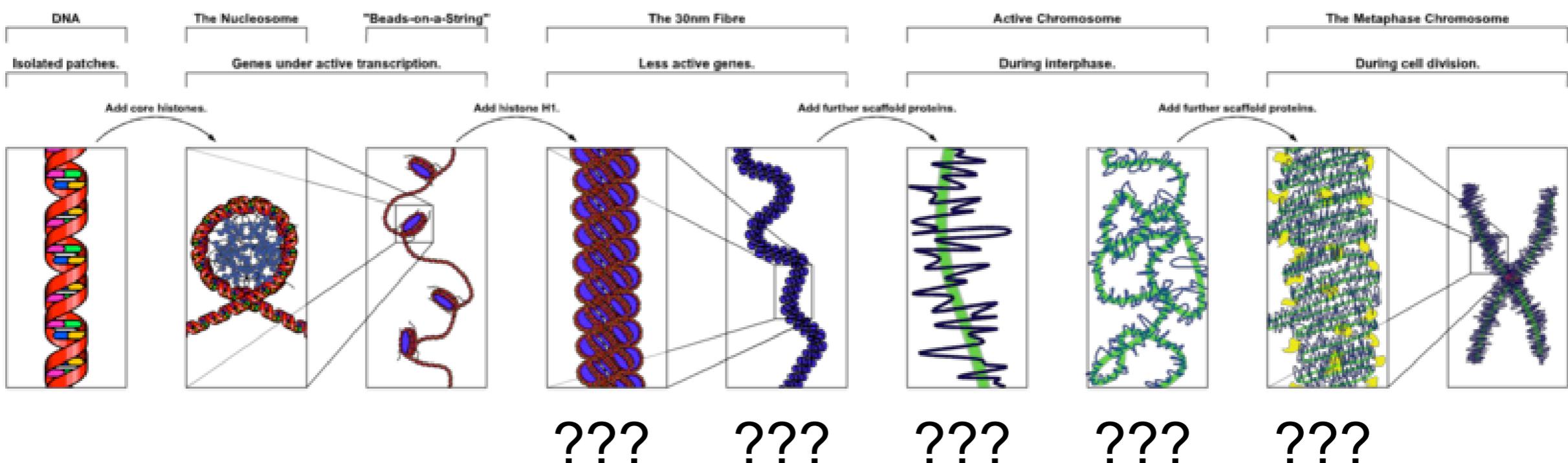
diameter: $6\mu\text{m}$
thread length: 2m
thread thickness: 2.5nm

6.7cm
 $\sim 20\text{km}$ (8 x MIT - Harvard)
 $\sim 20\mu\text{m}$

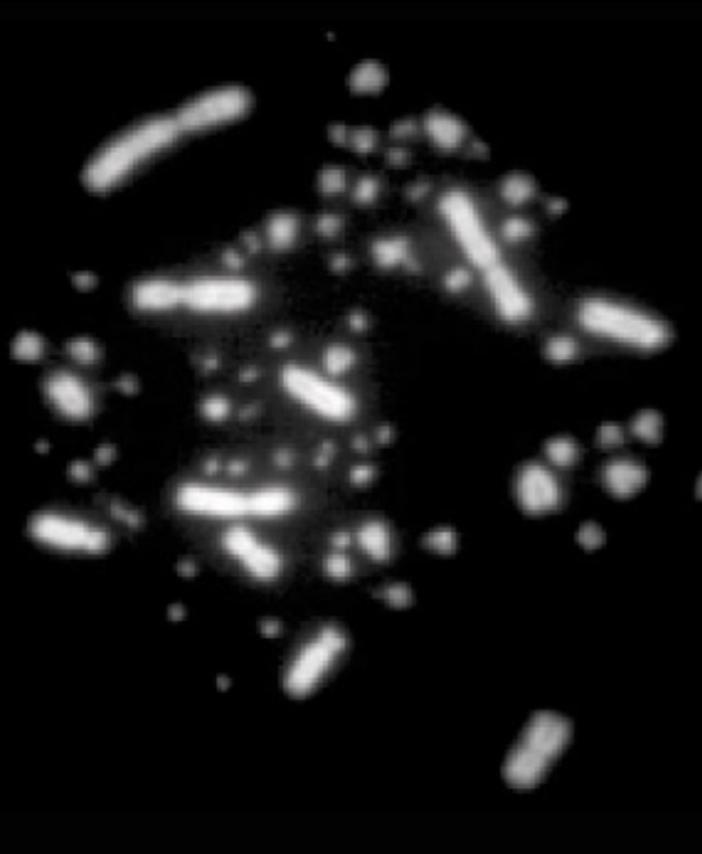
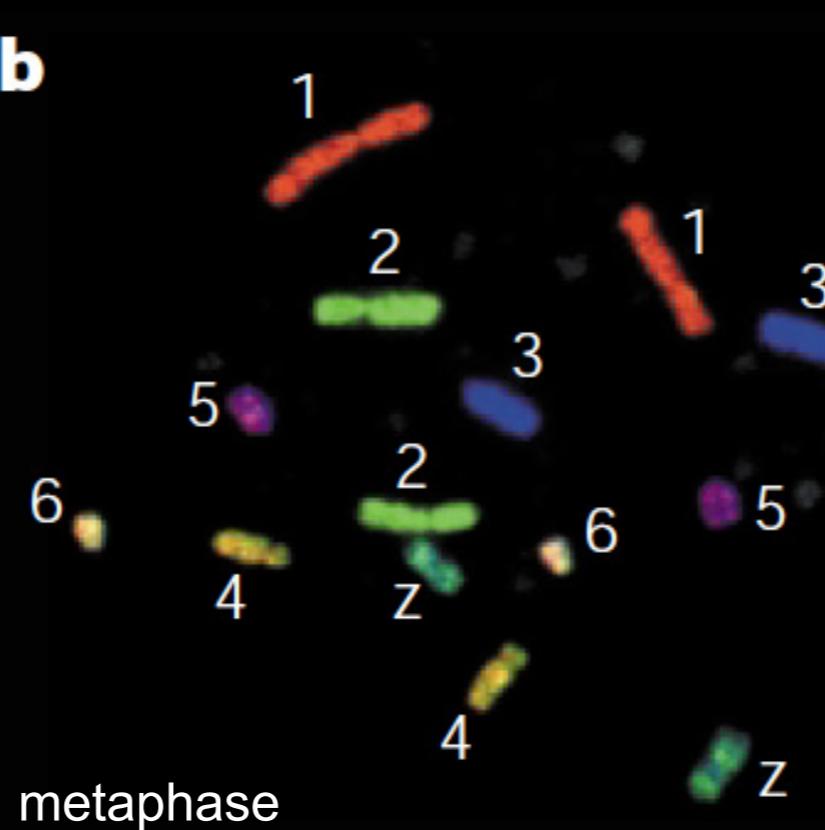


DNA compaction

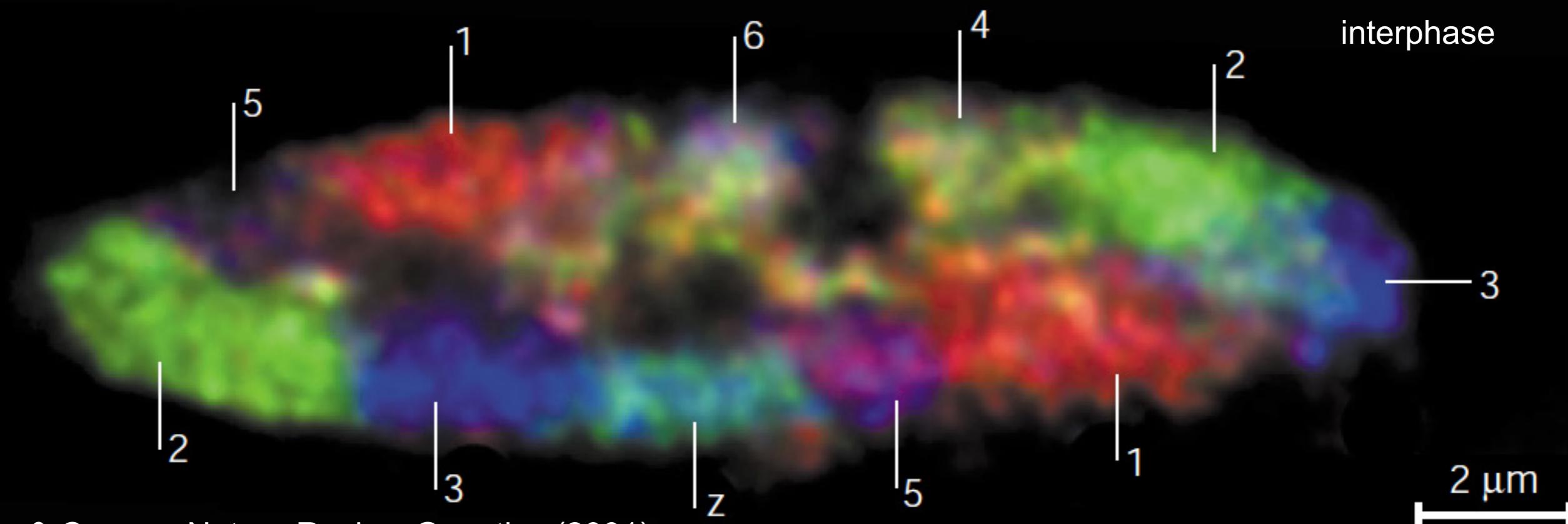
- DNA is **locally** compacted using *histone octamers* to form *nucleosomes*
- DNA is **globally** compacted by way of *chromosomes* (at least, during cell division / mitosis)
- Intermediate packaging mechanisms are subject of heavy speculation



Chromosome territories (CT)

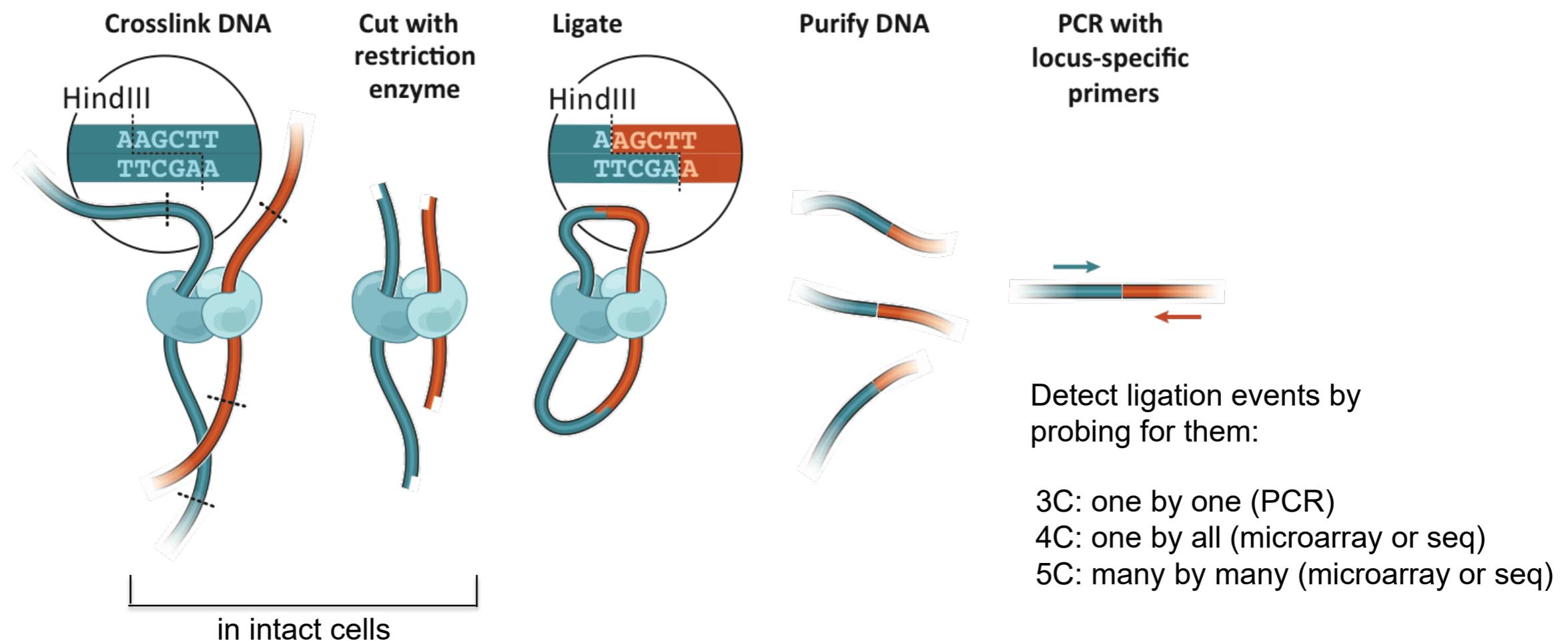
a**b****c**

Chr	1	2	3	4	5	z	6
Cy3	■						■
FITC		■				■	■
Cy5			■	■	■		■

d

3C: Chromosome Conformation Capture

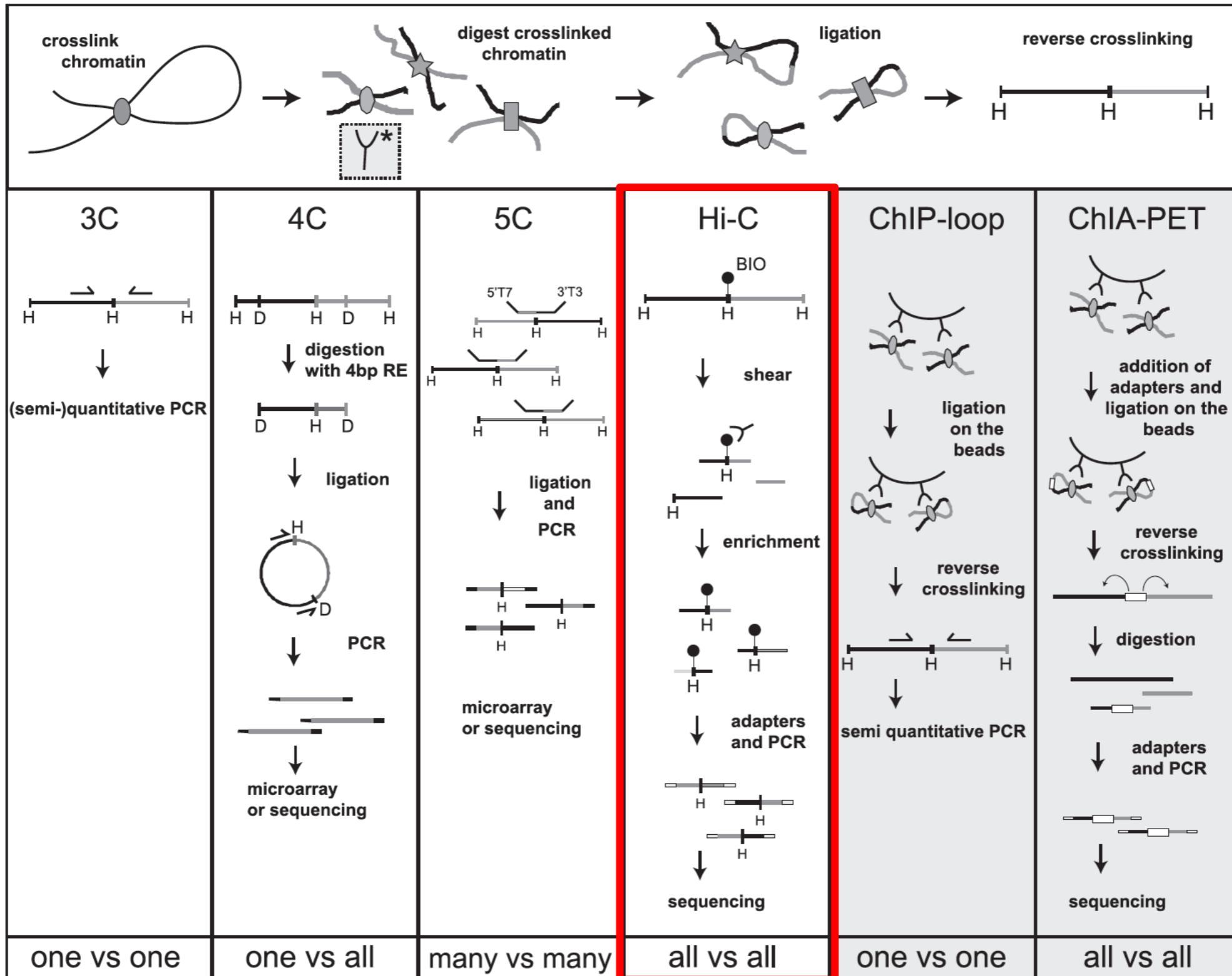
- Detects physical interactions between genomic elements
- Interacting elements are converted into ***ligation products***



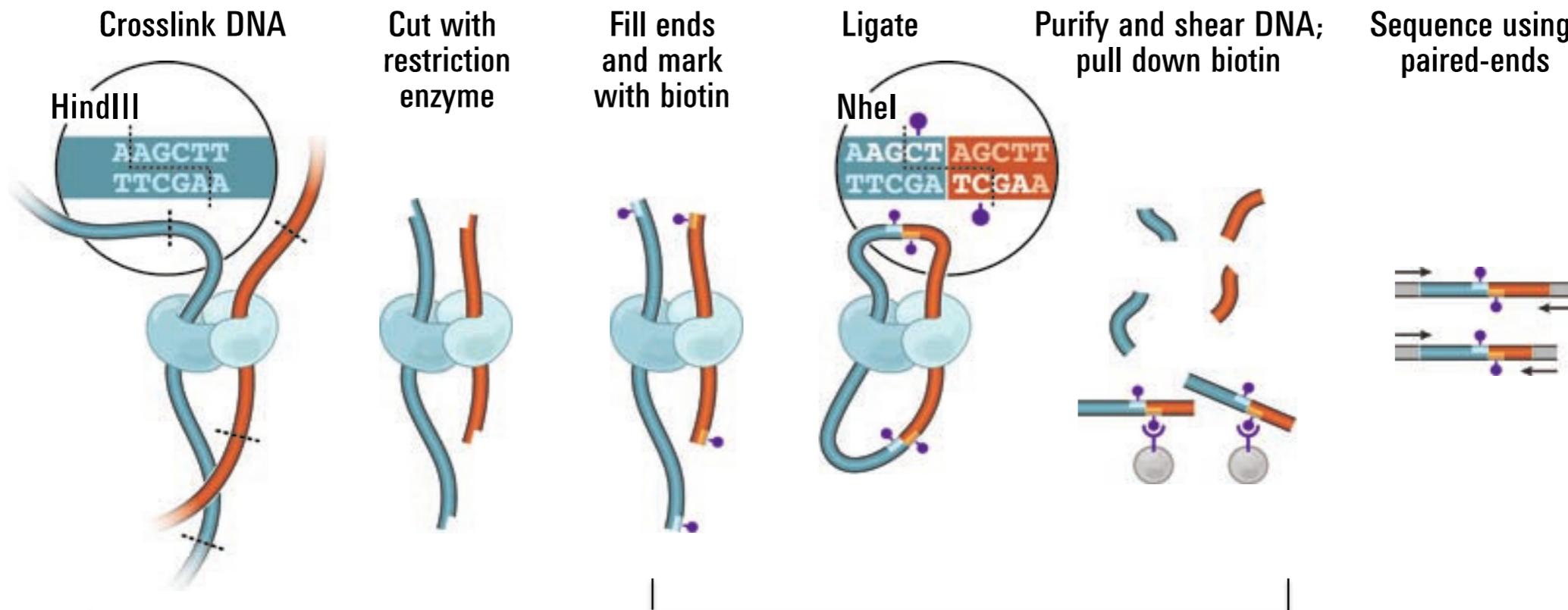
Dekker *et al.* Science 2002

Dostie *et al.* Genome Res. 2006

Chromosome Conformation Capturing (3C) based methods



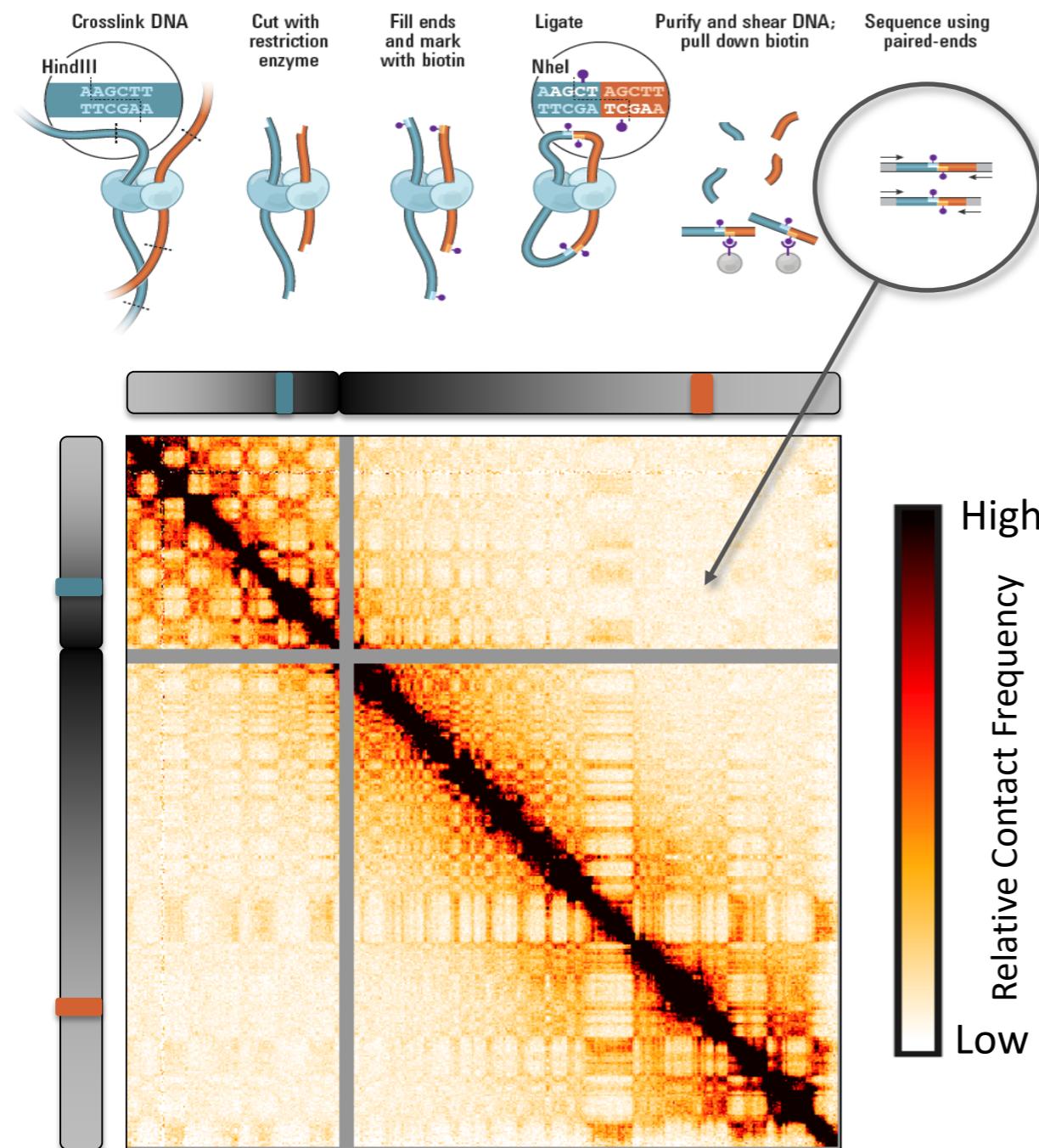
Hi-C: genome-wide 3C



Add adapters and do paired-end sequencing



Hi-C: genome-wide 3C



Unless synchronized, averages over the cell cycle as well!

Goal:

Measure direct physical interactions
(ensemble average)

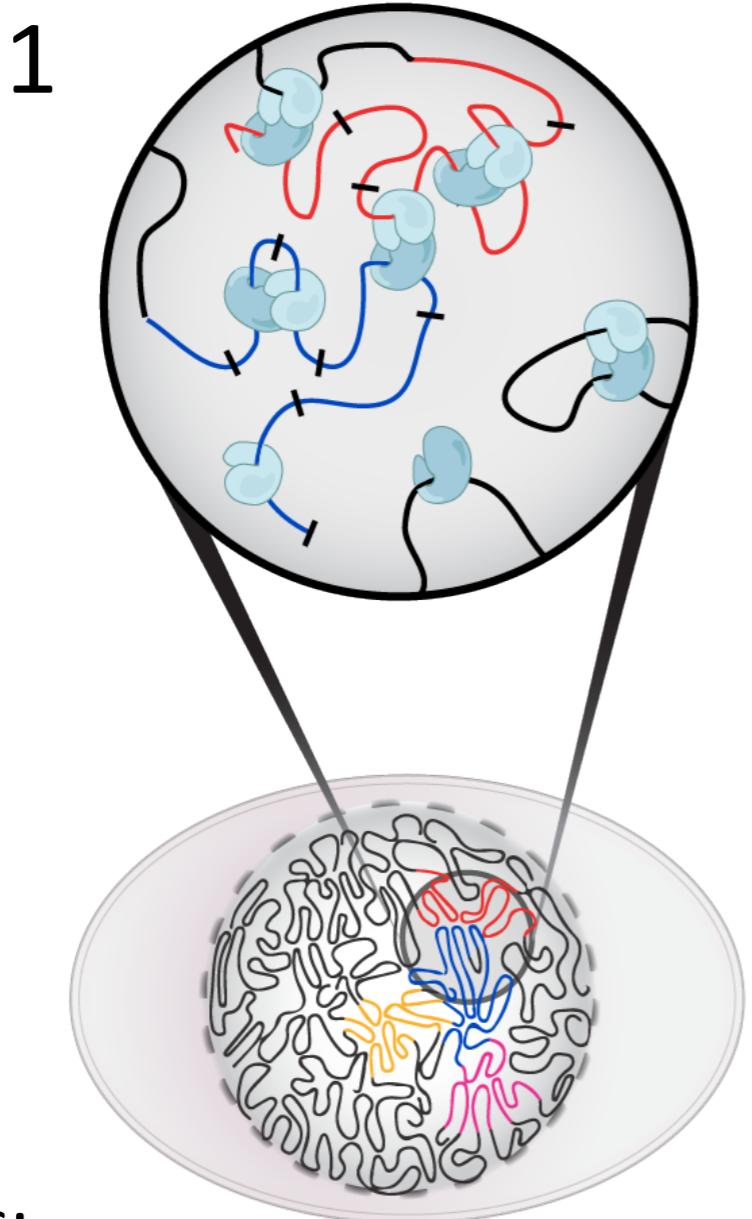
Steps:

1. Crosslink chromatin (freeze contacts)
→ “snapshot”
2. Cut chromatin with restriction enzyme
3. Ligate: captures spatial proximity between fragment.
4. High-throughput sequencing to identify chimeric reads → interactions

Main features:

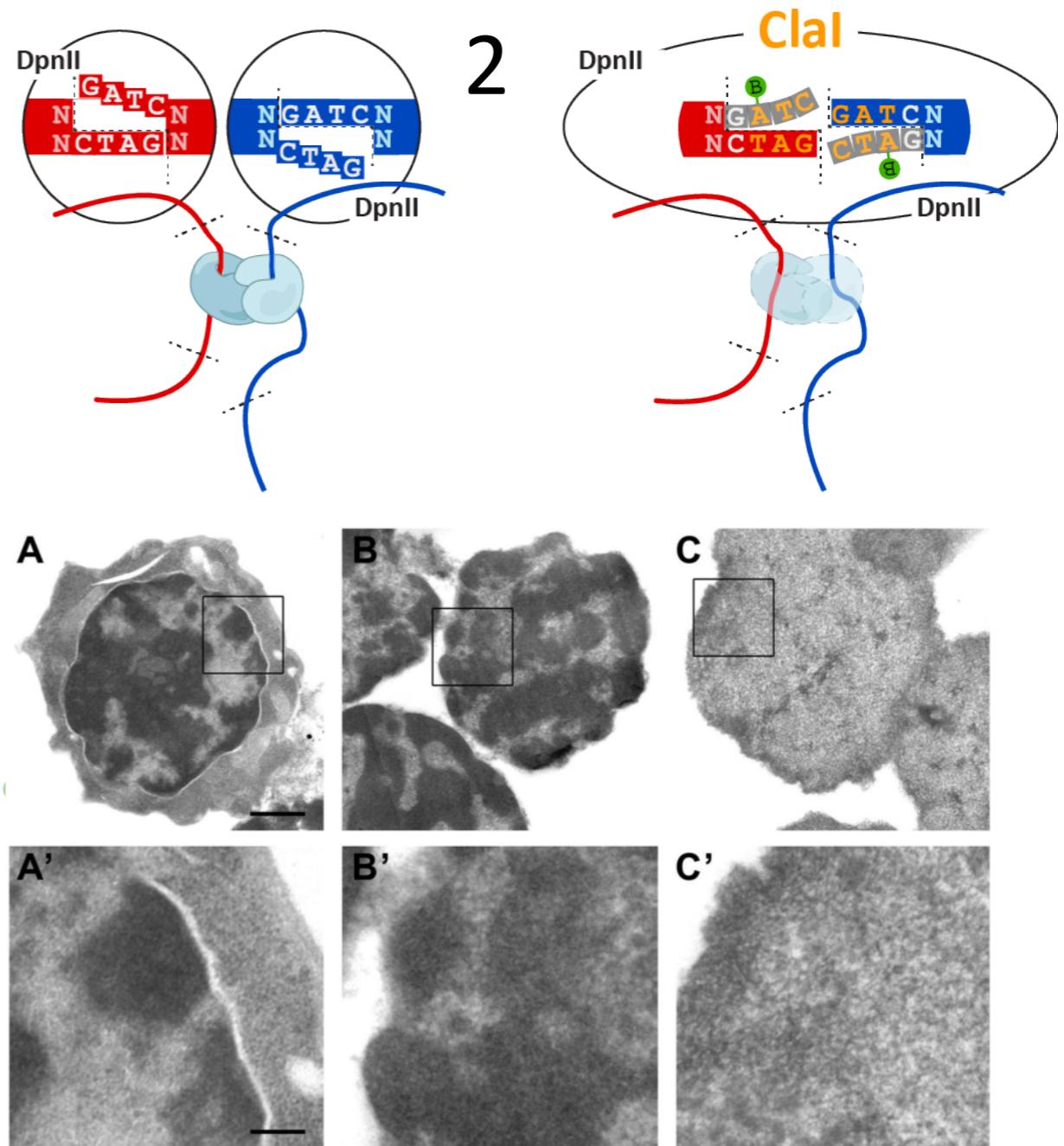
- Bright diagonal
- Decay at increasing distances

Hi-C vs. What-you-See



Critical steps:

1. Crosslinking to fix conformation
2. Digestion and re-ligation
3. Sequencing (biotinylated) junctions

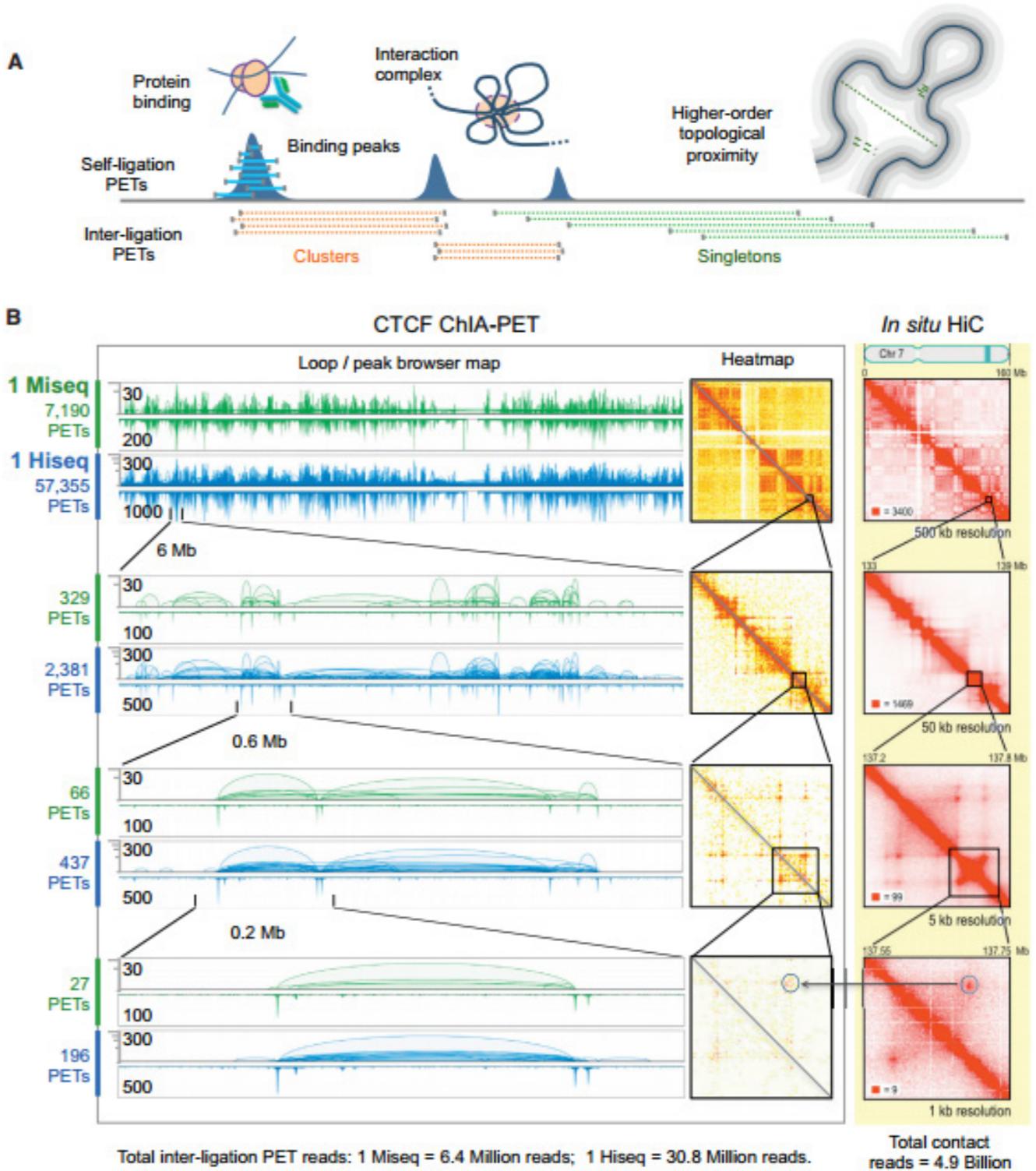


Lieberman-Aiden et al., *Science* 2009

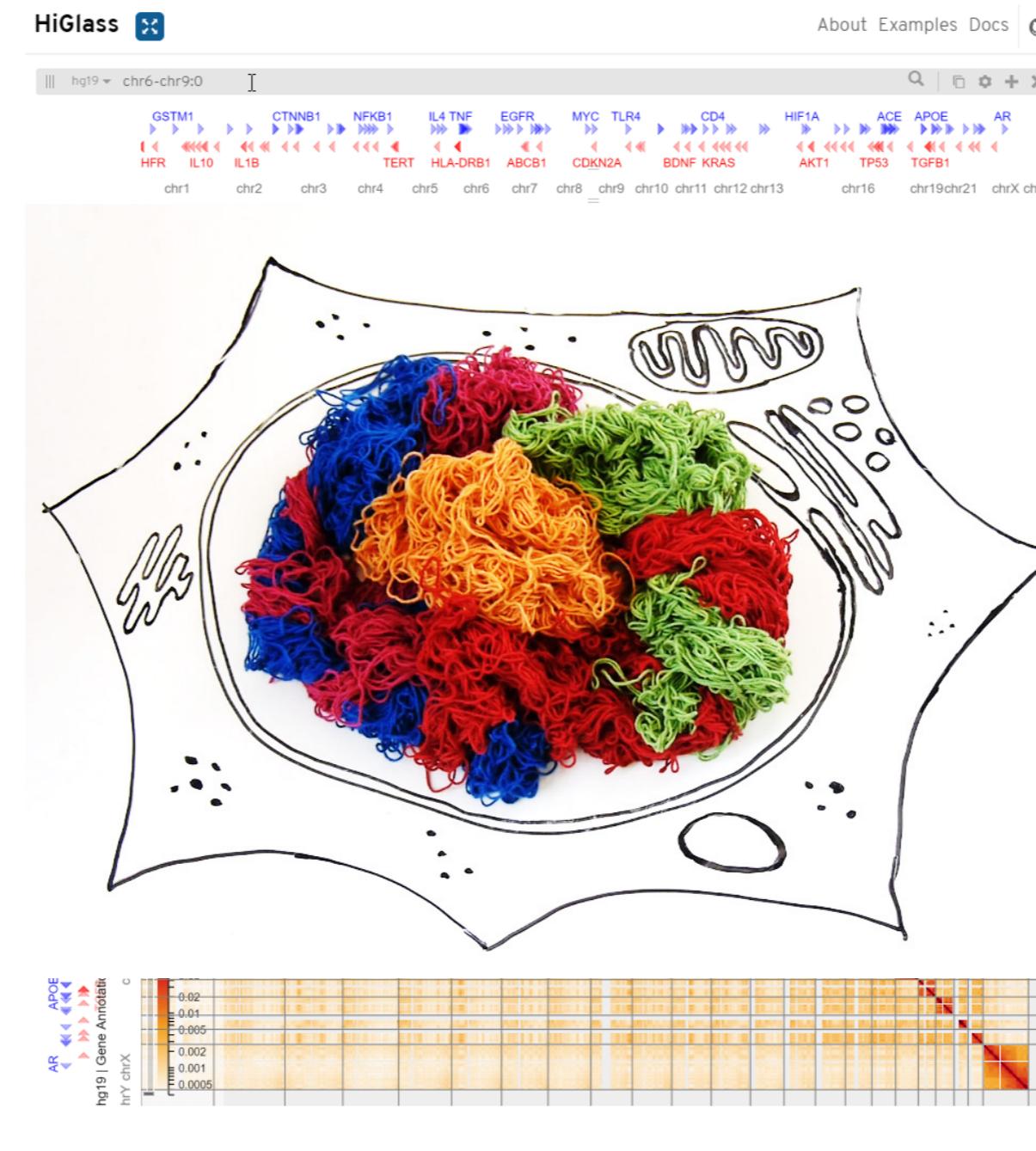
Belaghzal et al., *Genome Methods* 2013

ChIA-PET: Chromatin Interaction Analysis using Paired-End-Tag sequencing

1. self-ligation peaks: binding sites
 2. Inter-ligation: long range interaction
 3. Consistence between CTCF ChIA-PET and Hi-C
 4. ChIA-PET has higher resolution than Hi-C

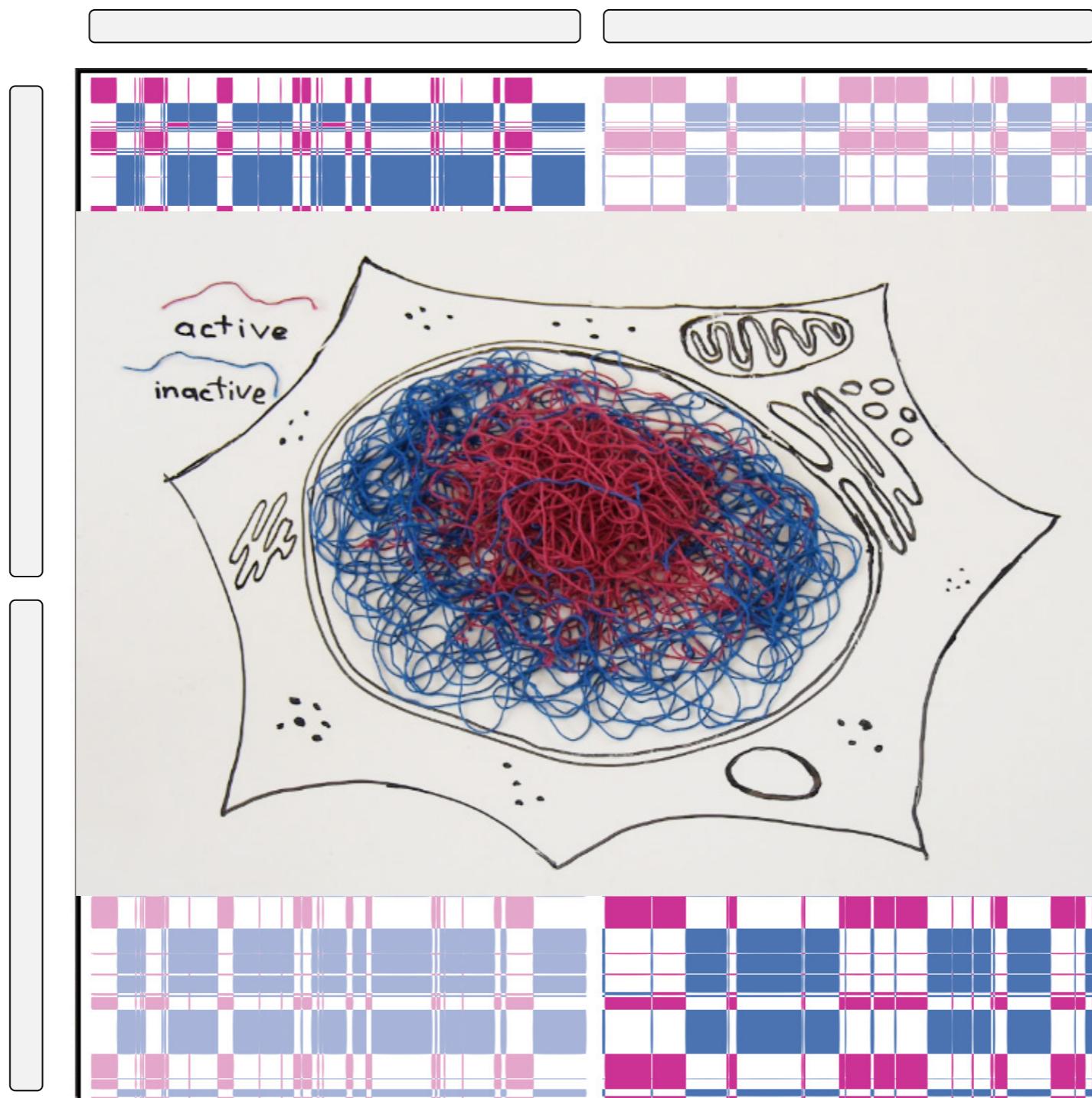


Territoriality

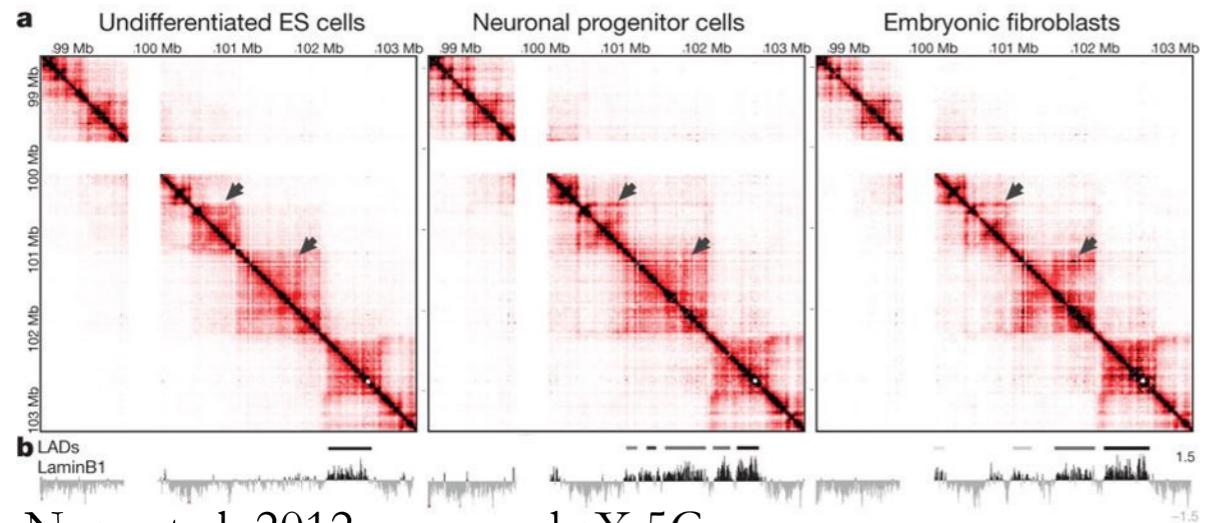


Kerpedjiev P, **Abdennur N**, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, Hwang J, Lee S, Alver BH, Pfister H, Mirny LA, Park PJ, Gehlenborg N. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* 2018 Aug 24;19(1):125.

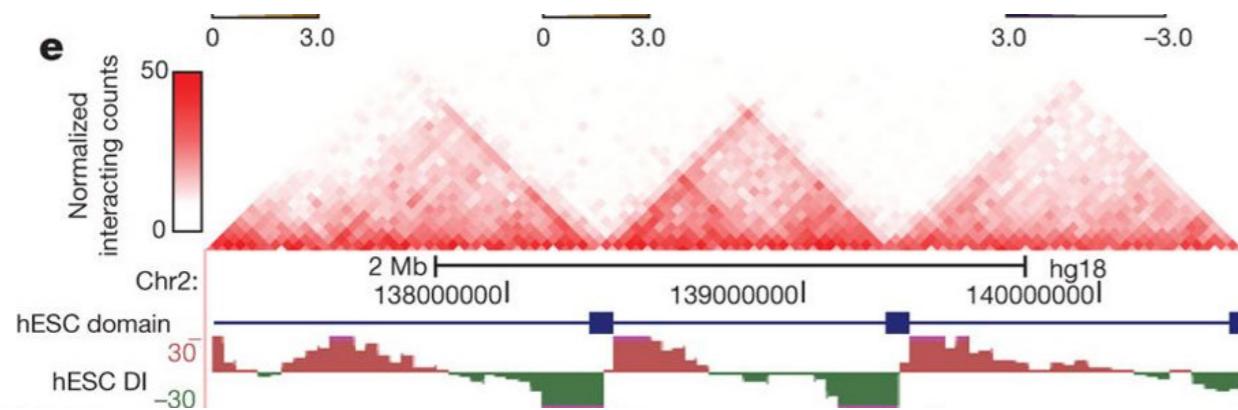
Compartmentalization (segregation)



TADs



Nora et al, 2012: mouse chrX 5C



Dixon et al, 2012: human ESCs and fibroblasts

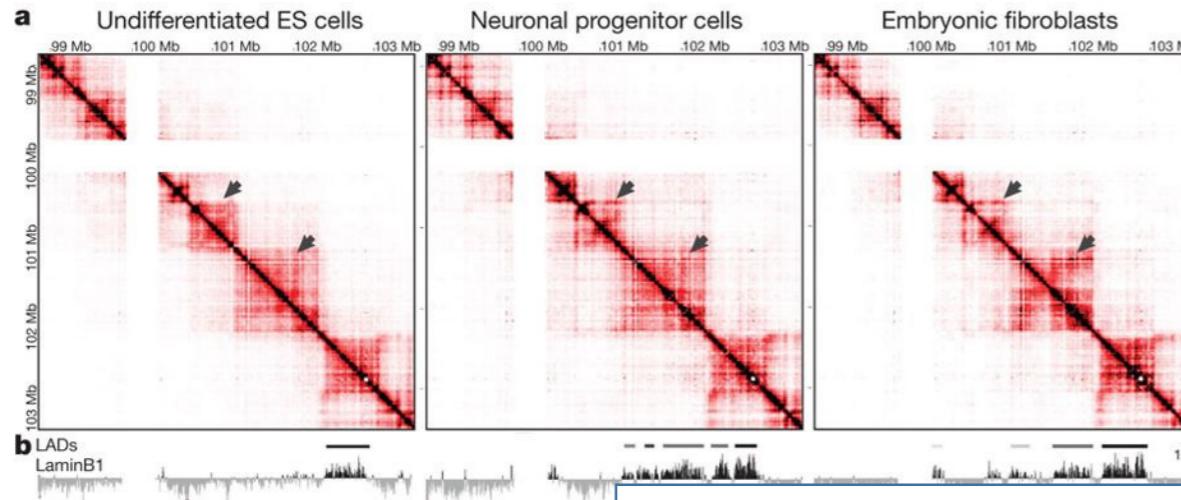
Self-associating intervals, < 1Mb

Sharp boundaries enriched for architectural proteins

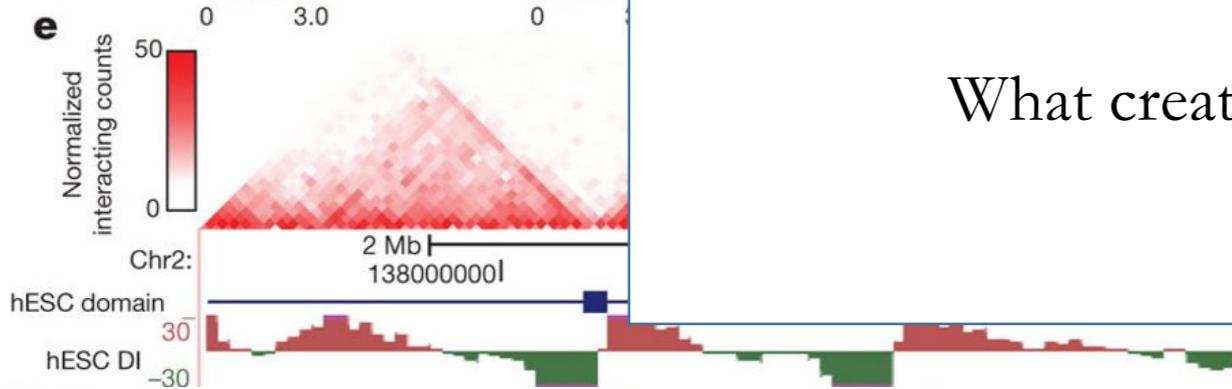
~2-4 fold difference in contact frequency within vs across boundaries

Regulatory neighborhoods

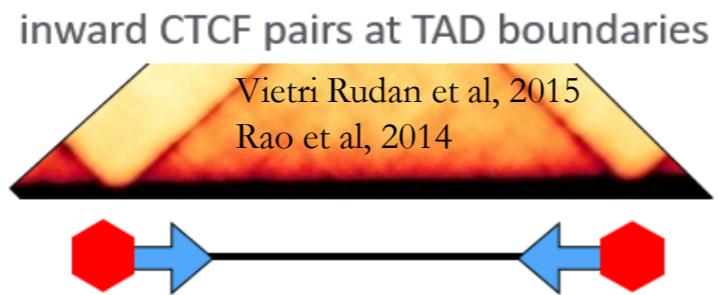
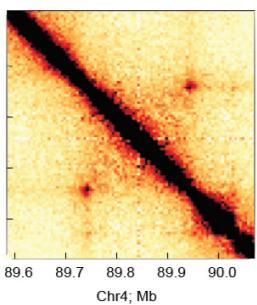
TADs



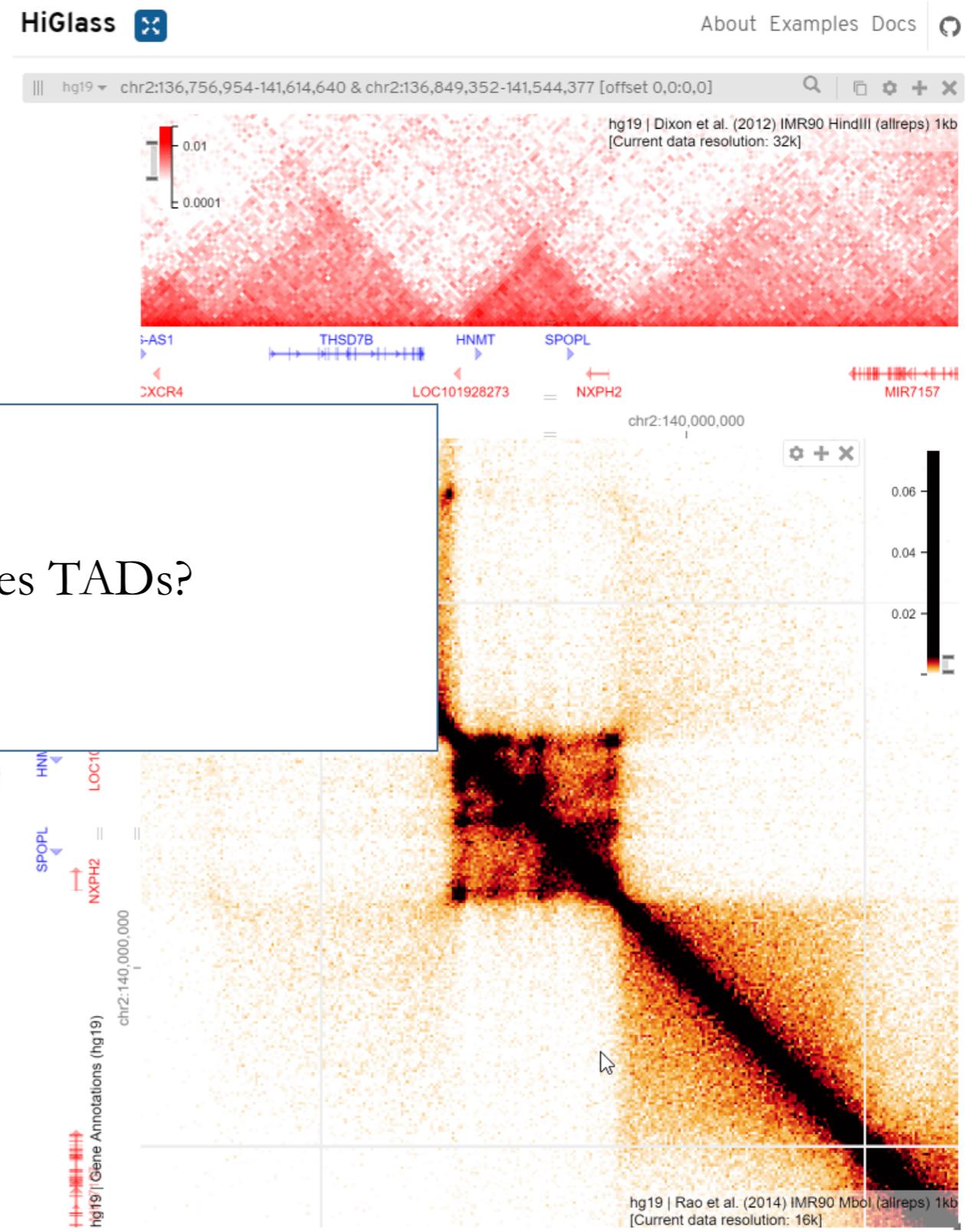
Nora et al, 2012: mouse chrX



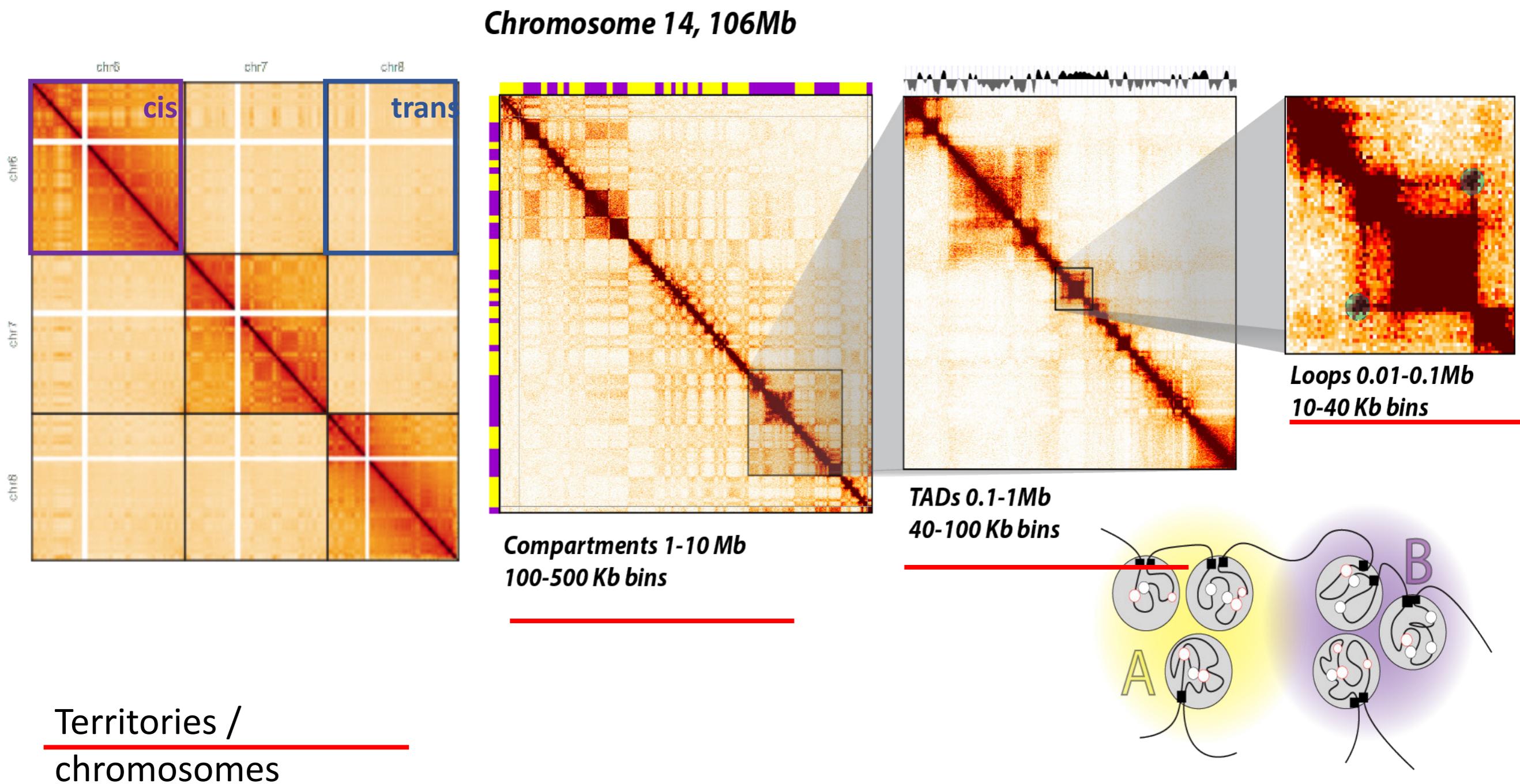
Dixon et al, 2012: human ESCs and fibroblasts



Rao et al, 2014

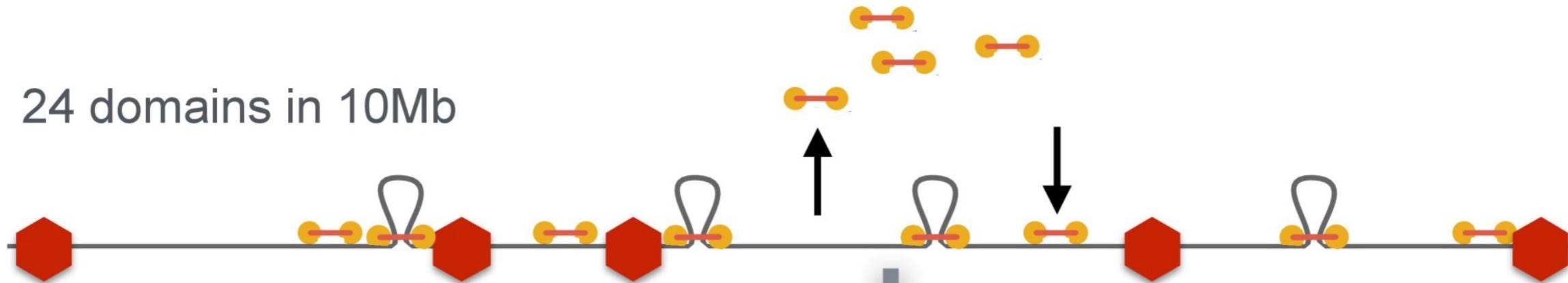


Scales of organization



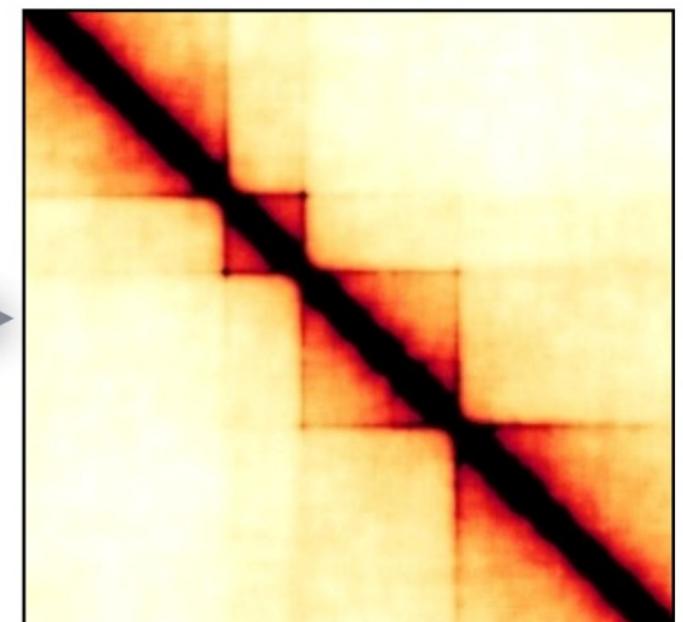
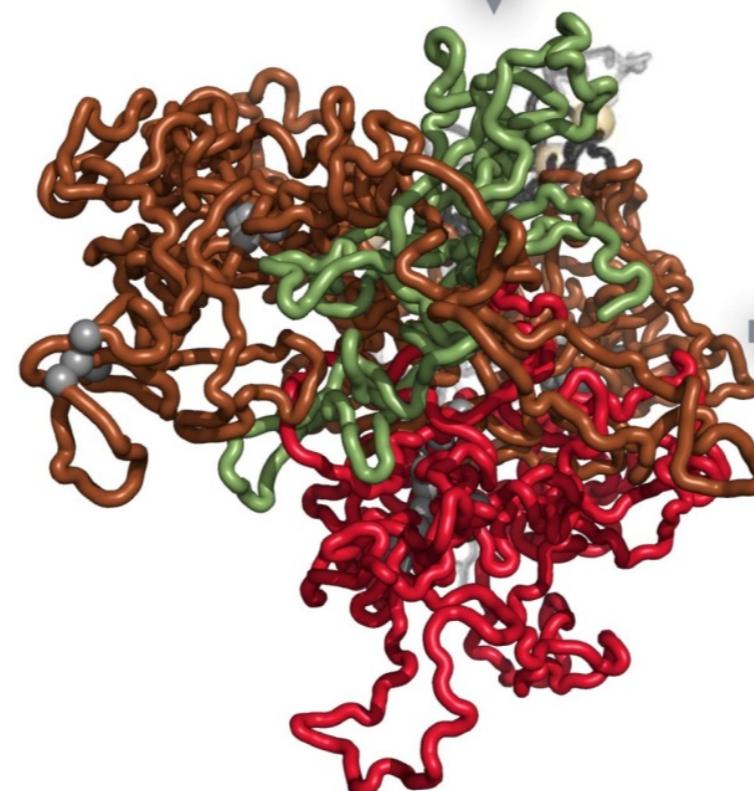
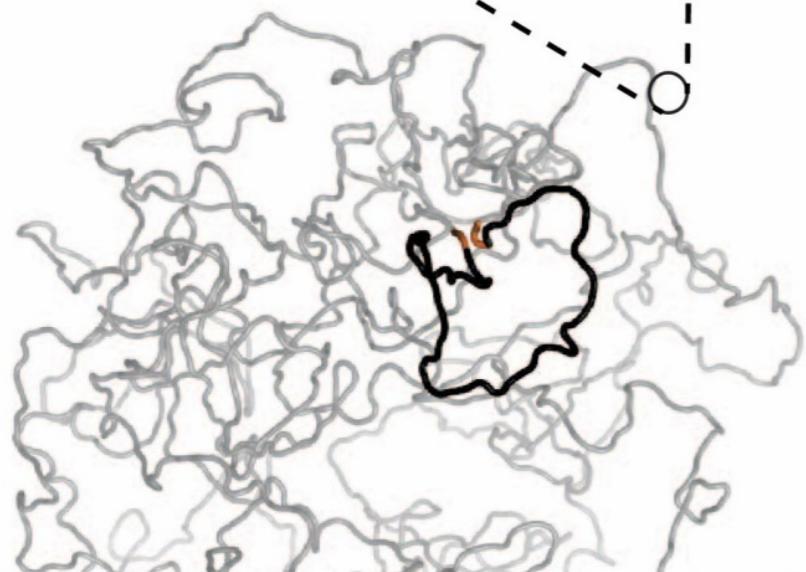
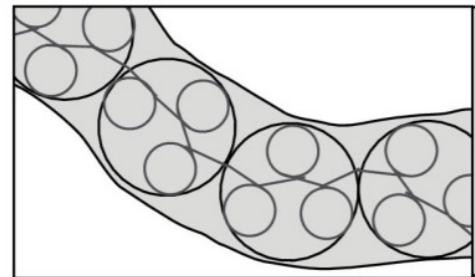
Loop extrusion + polymer model

24 domains in 10Mb



3D simulations of
polymer dynamics

1 monomer = 600bp



16,000 monomers

Formation of chromosomal domains by loop extrusion
<http://biorxiv.org/content/early/2015/08/14/024620>

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
- Three-dimensional chromatin structure, Hi-C, ChIA-PET, TADs, Loop Extrusion

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 \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow Motif matches

4. Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures

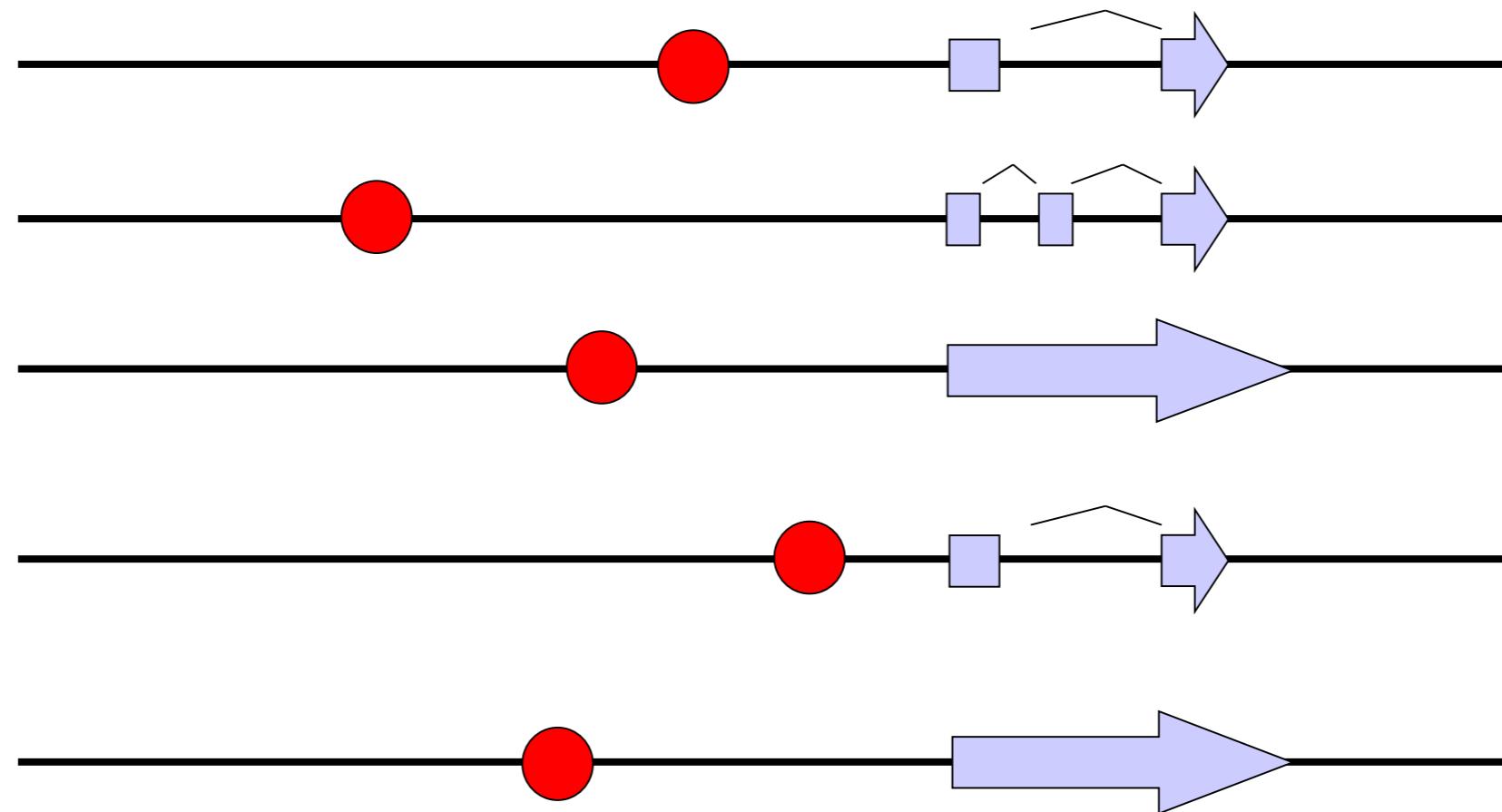
- DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
- DeepSea: Train model directly on mutational impact prediction
- Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
- ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
- DeepLIFT: Model interpretation based on neuron activation properties
- DanQ: Recurrent Neural Network for sequential data analysis

5. Guest Lecture: David Kelley on Basset and Deep Learning for Hi-C looping

2. Classical regulatory genomics (before Deep Learning)

Enrichment-based discovery methods

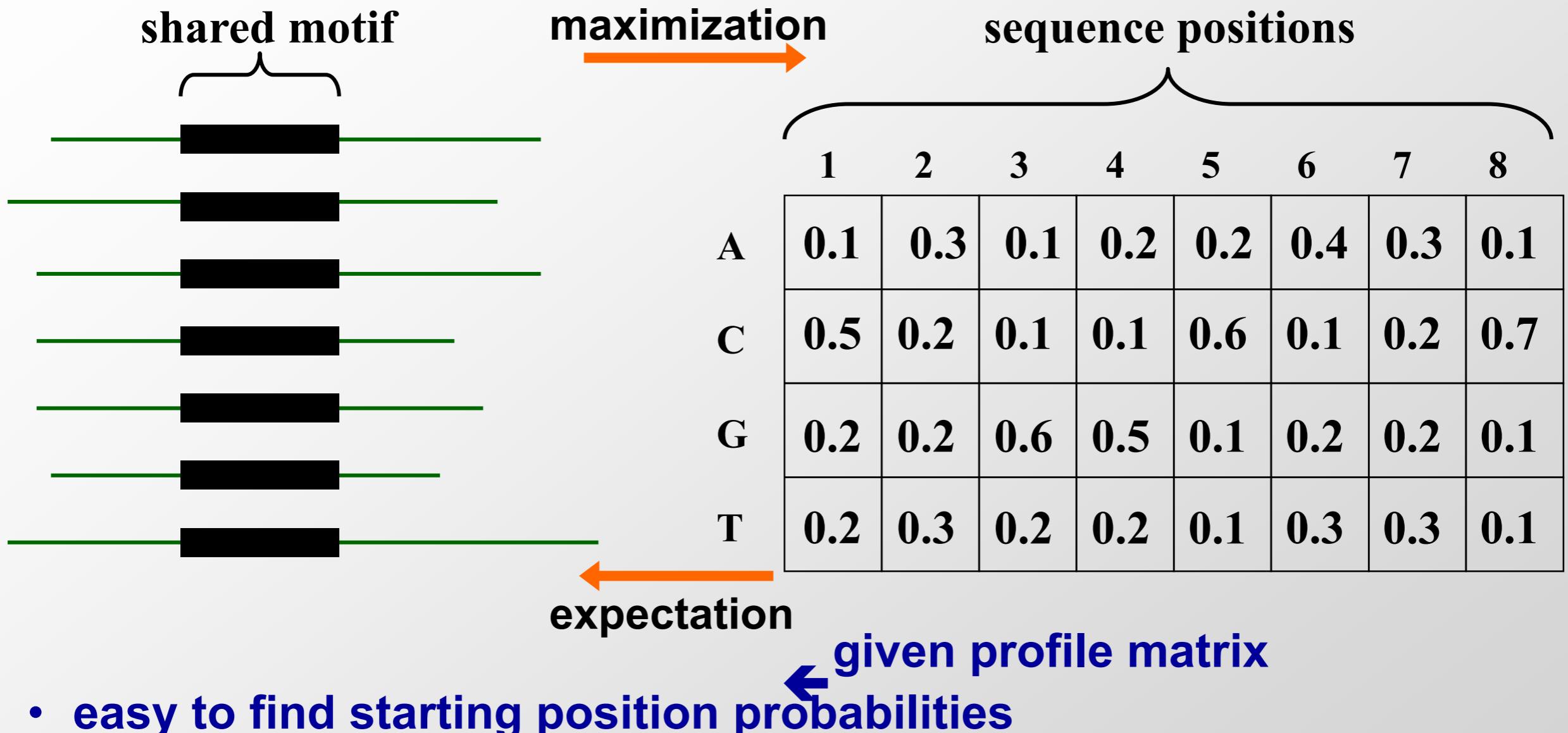
Given a set of **co-regulated/functionally related genes**,
find common motifs in their promoter regions



- 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

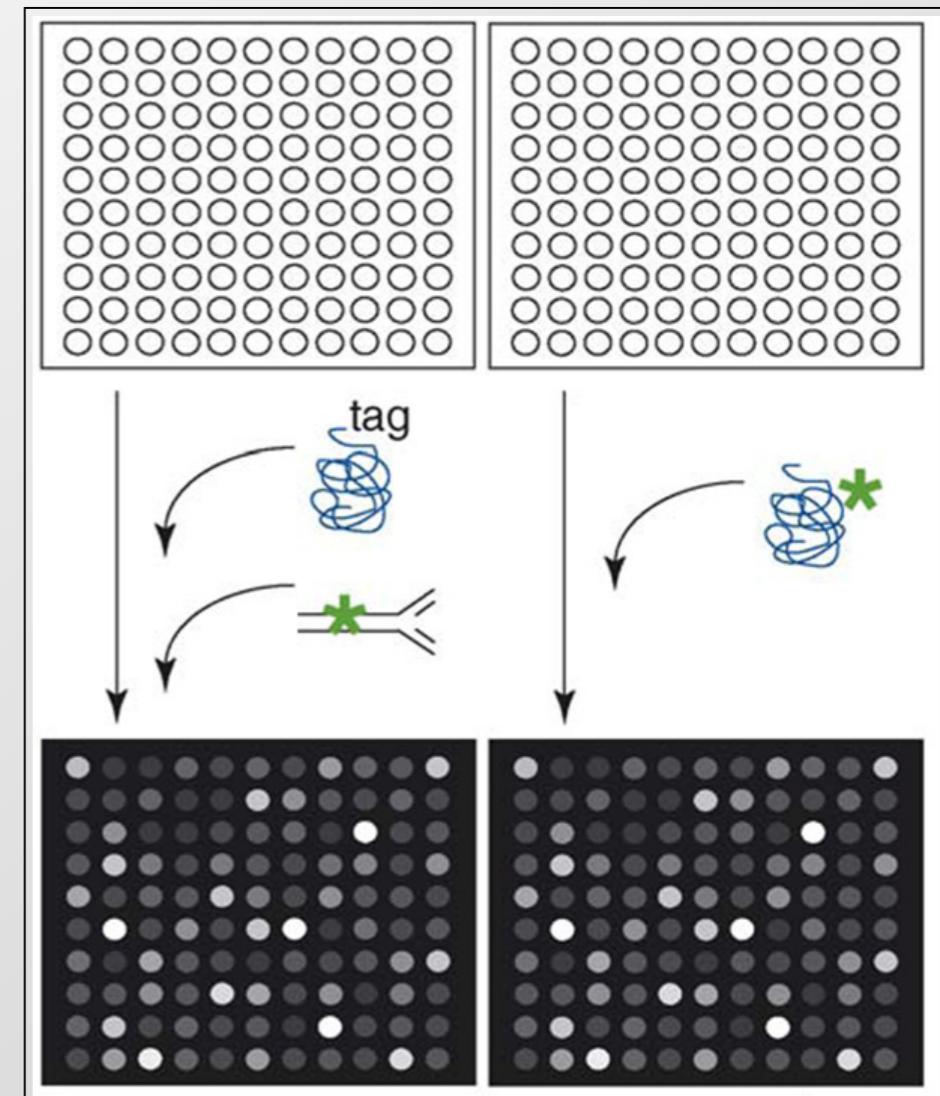
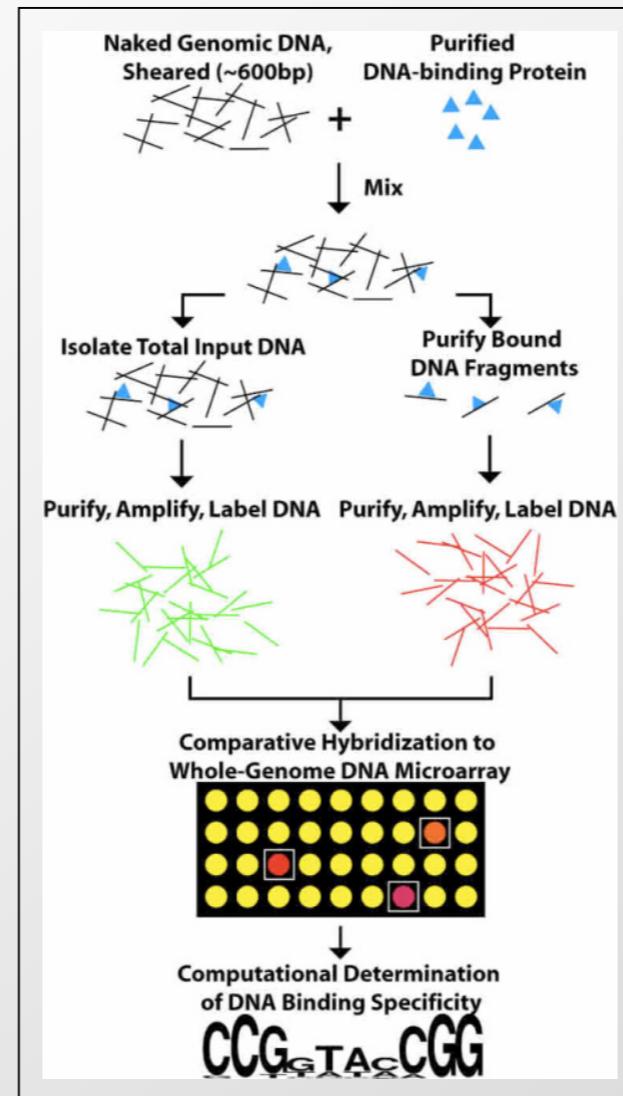
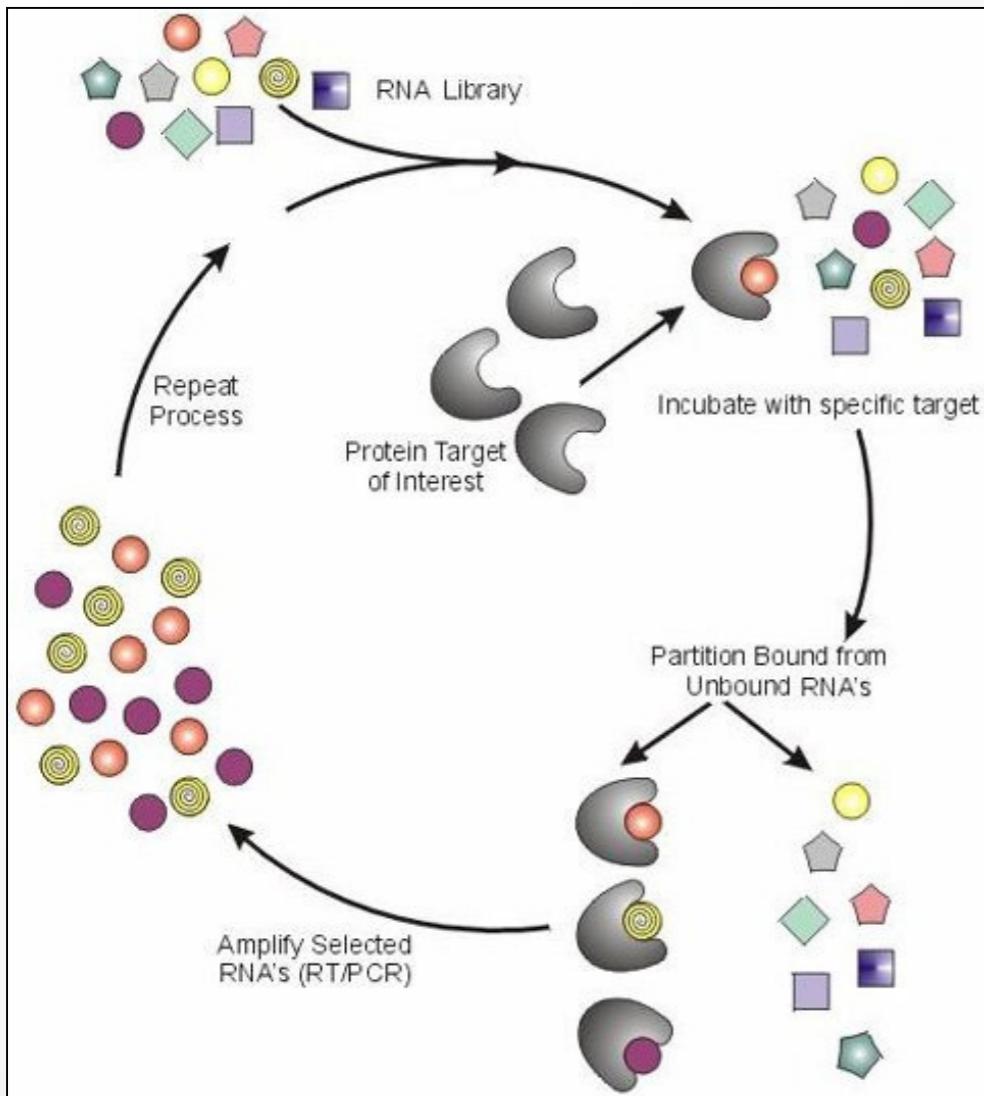
Starting positions \leftrightarrow Motif matrix

- given aligned sequences \rightarrow easy to compute profile matrix



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

Experimental factor-centric discovery of motifs



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
- Three-dimensional chromatin structure, Hi-C, ChIA-PET, TADs, Loop Extrusion

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

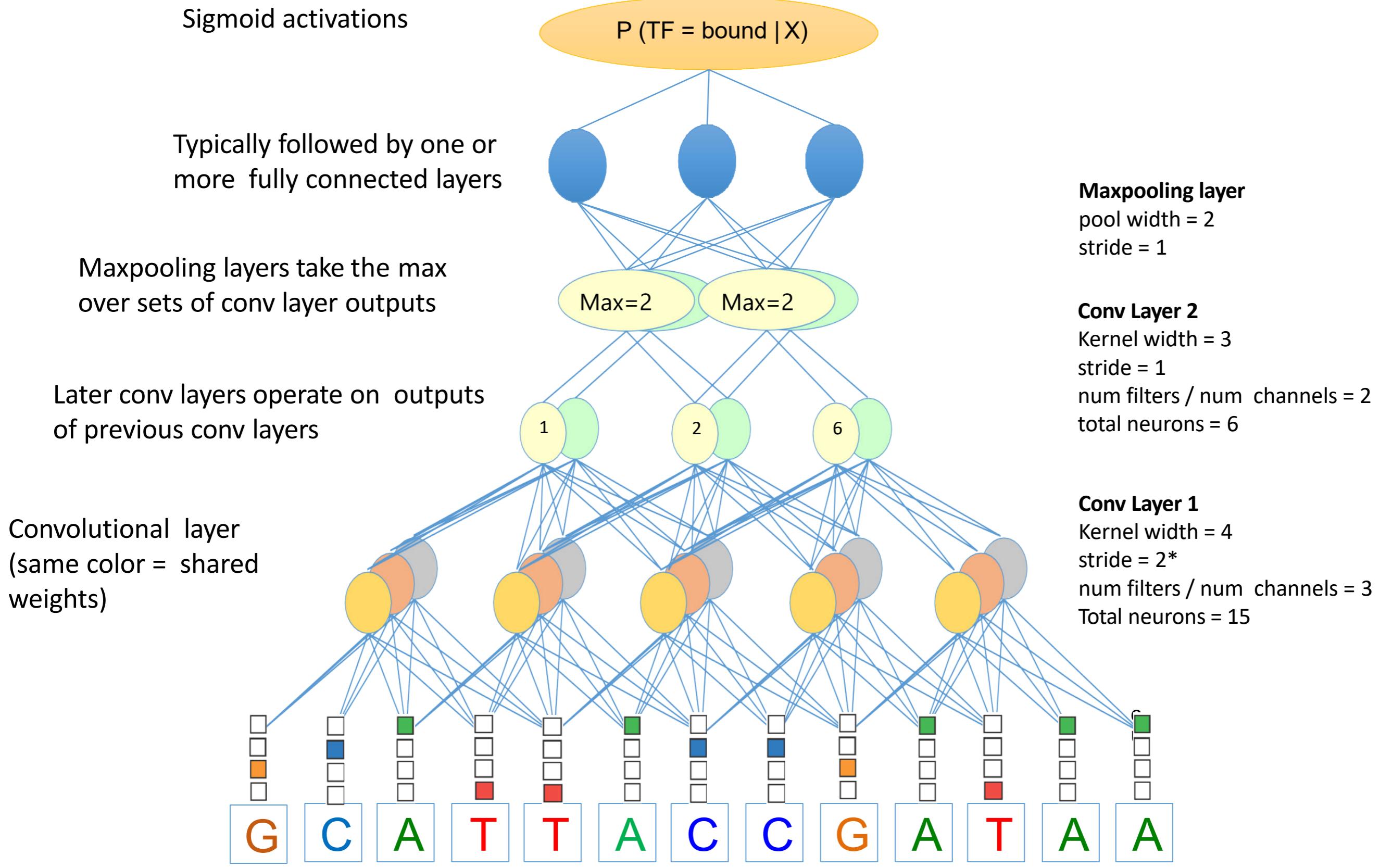
- Key idea: pixels \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow Motif matches

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: David Kelley on Basset and Deep Learning for Hi-C looping

Deep convolutional neural network

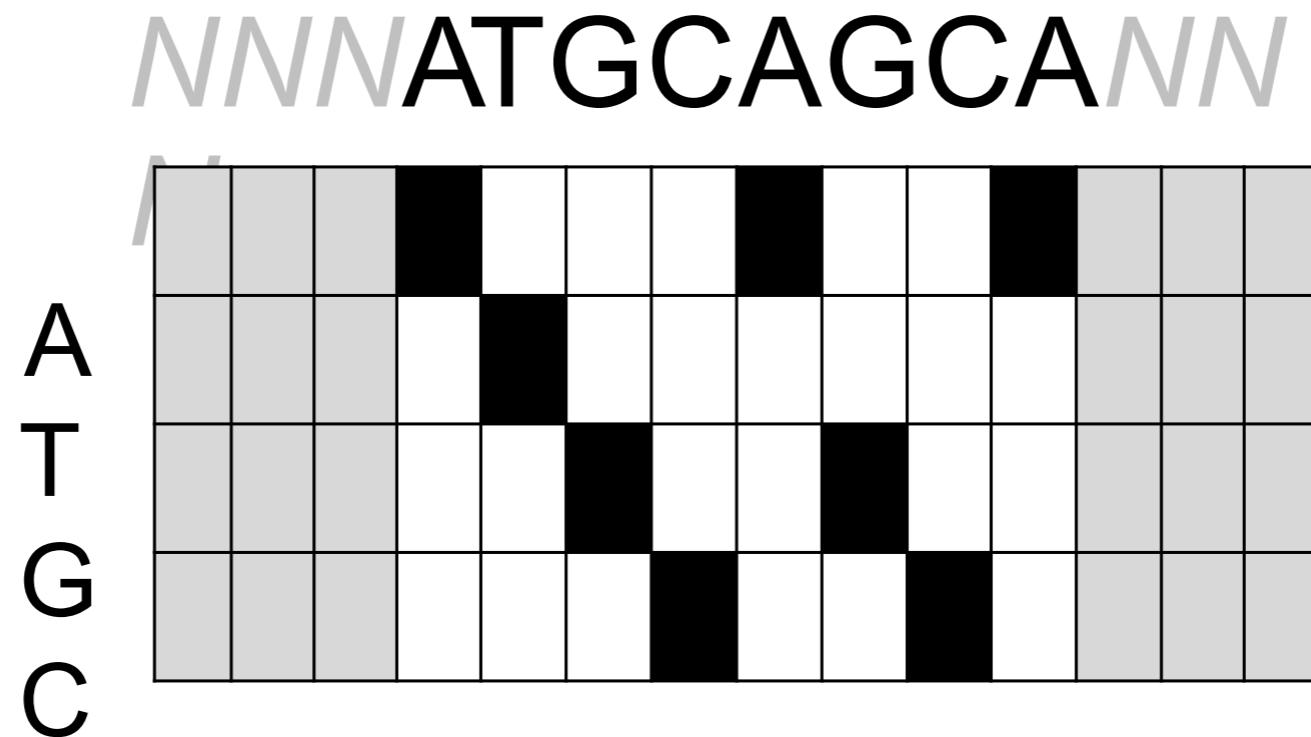


*for genomics, a stride of 1 for conv layers is recommended

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

An example of using CNN to model DNA sequence

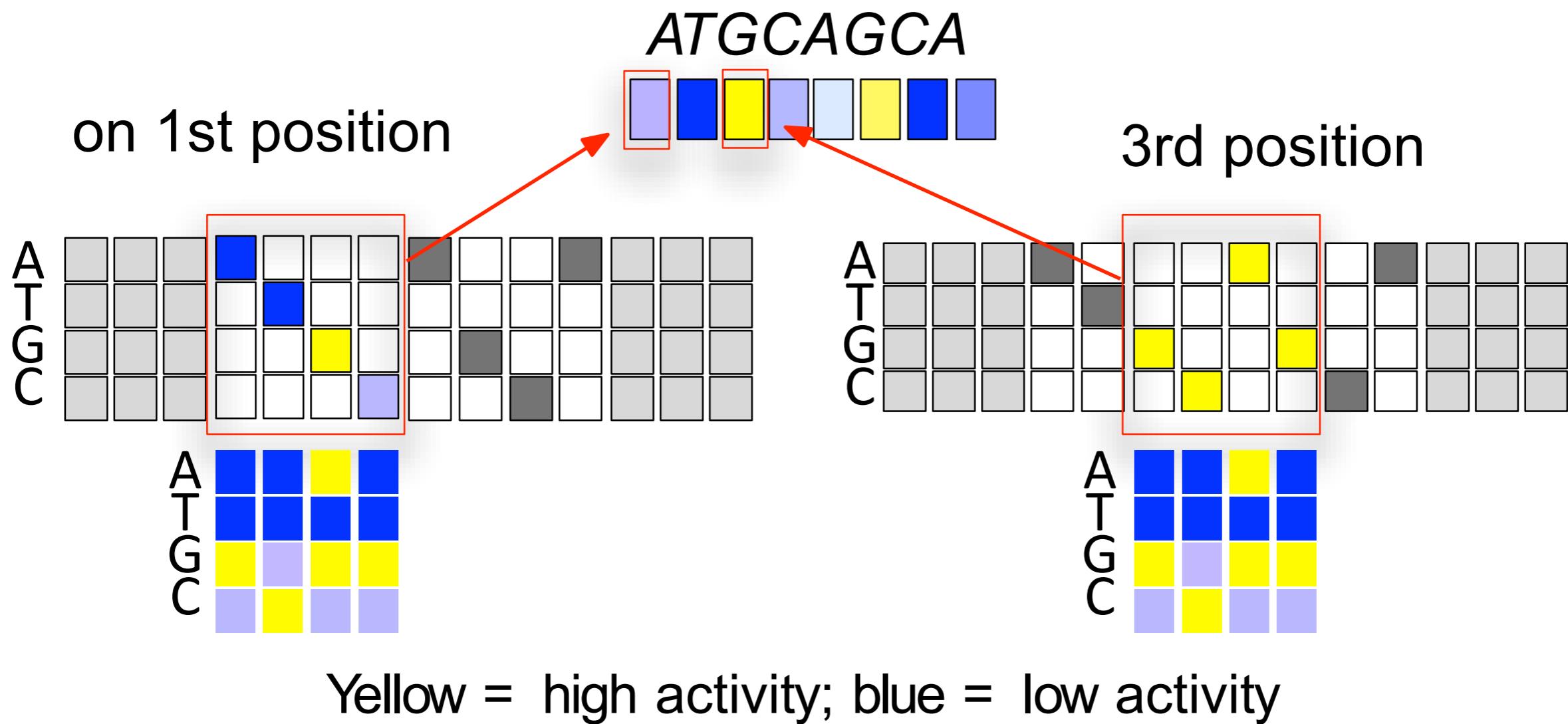
Representing DNA sequence as 2D matrix:



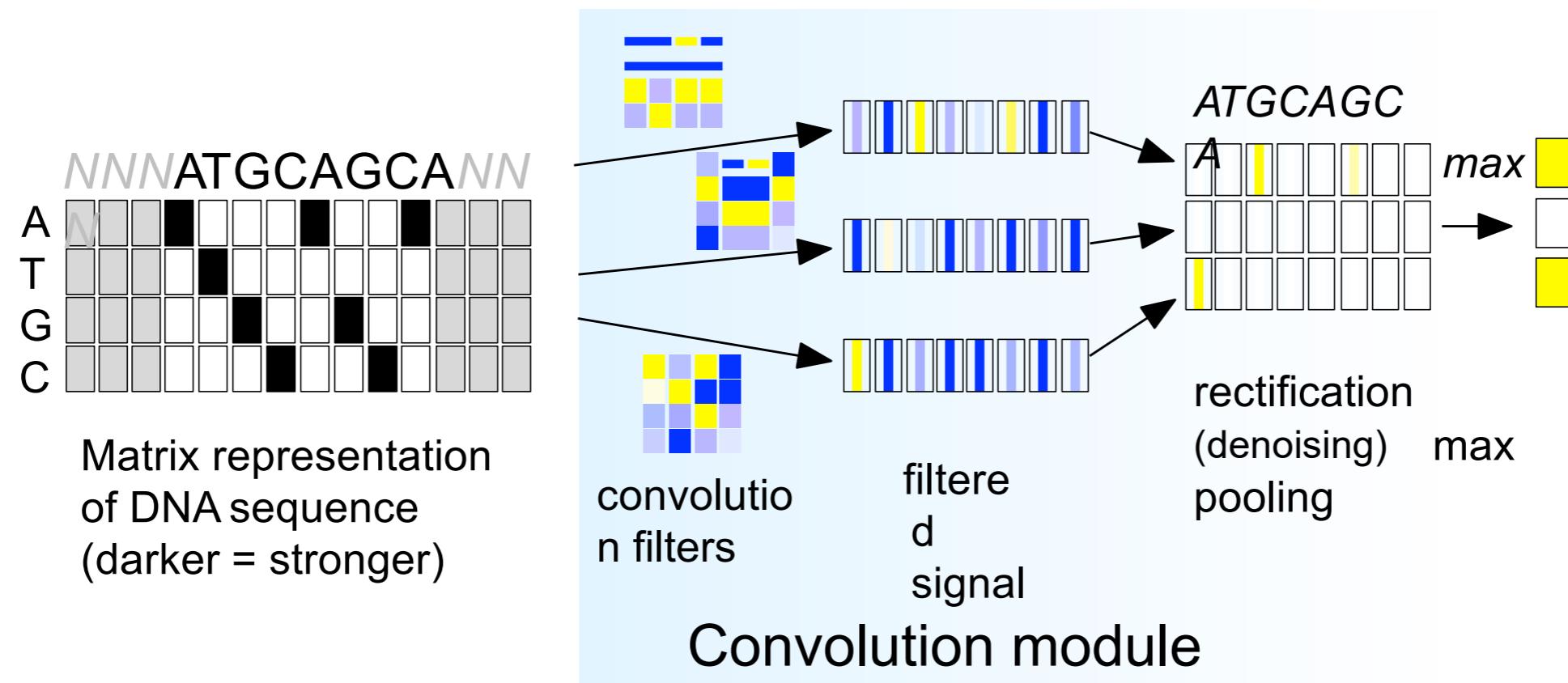
Matrix representation of
DNA sequence
(darker = stronger)

Convolution – extracting invariant feature

Applying 4 bp sequence filter along the DNA matrix:



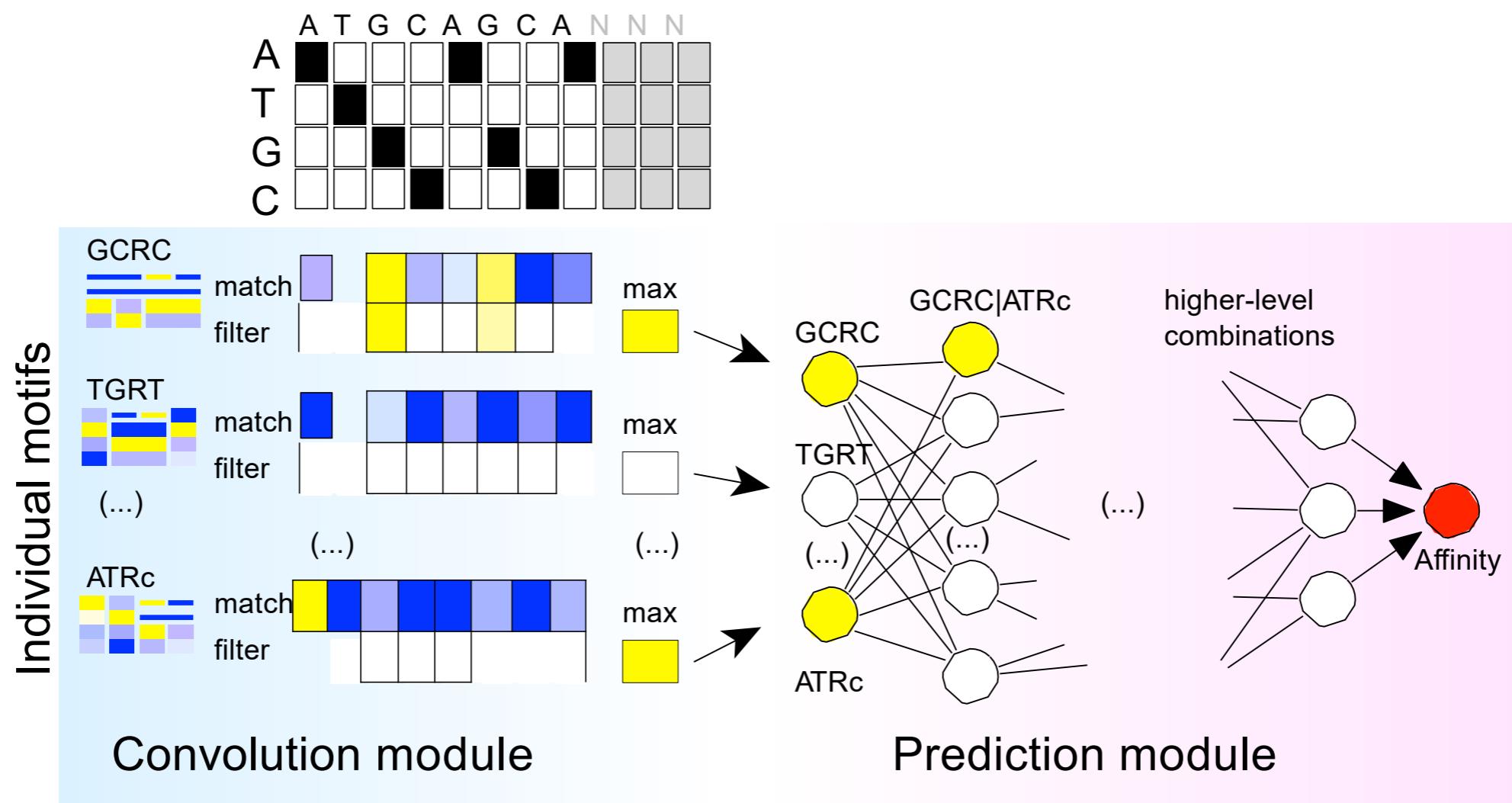
Convolution – extracting invariant feature



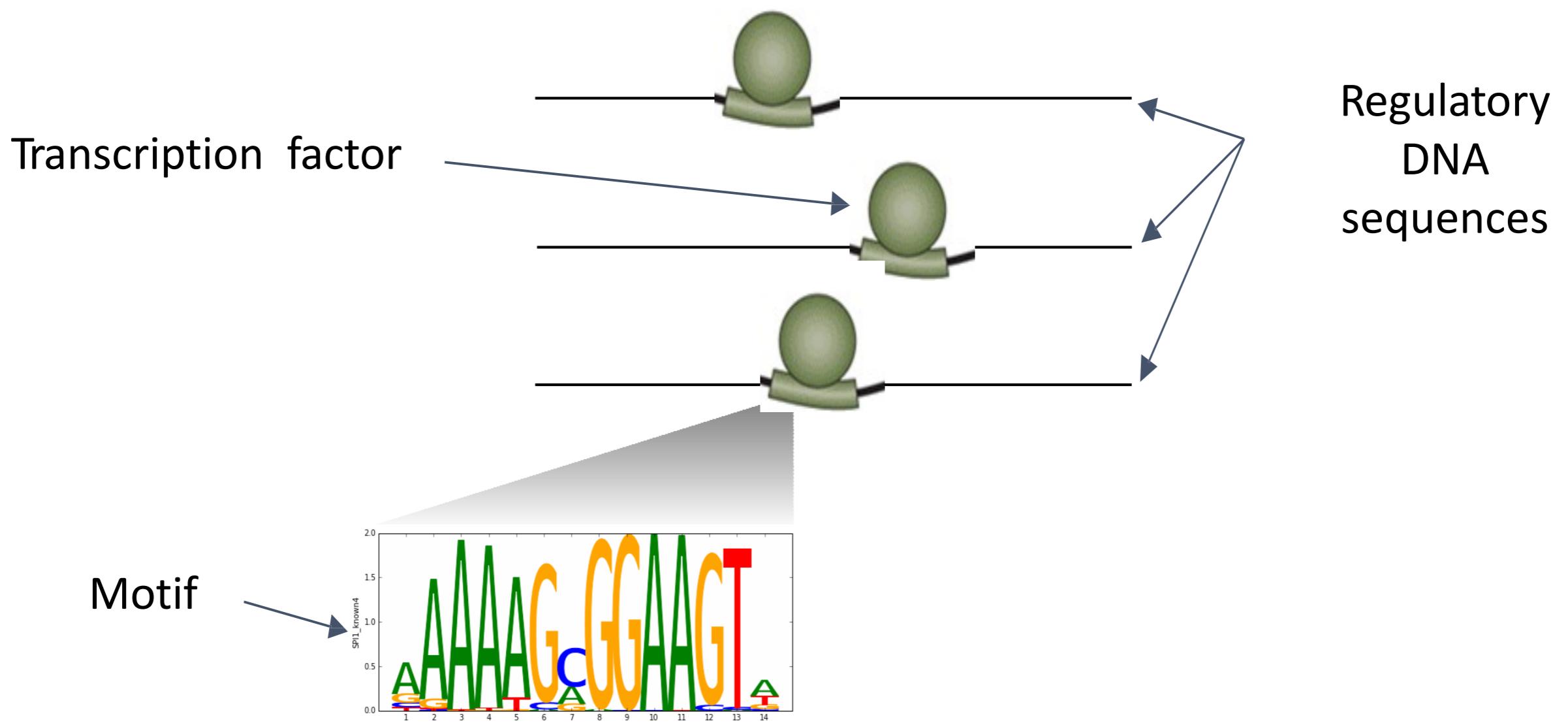
Rectification = ignore signals below some threshold.
Pooling = summary of each channel by max or average.

Prediction using extracted features map

ChIP-seq, PBMs, SELEX Experiments DNA sequence



Key properties of regulatory sequence



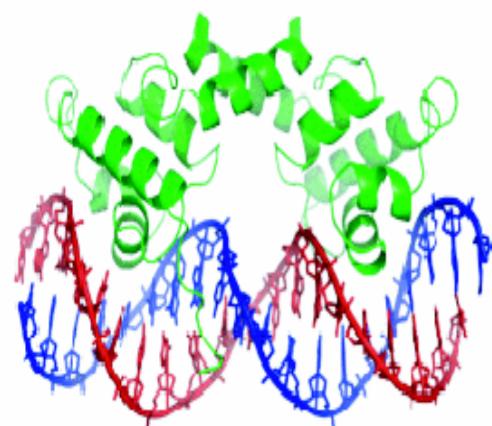
TRANSCRIPTION FACTOR BINDING

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

Sequence motifs: PWM

GGATAA
CGATAA
CGATAT
GGATAT

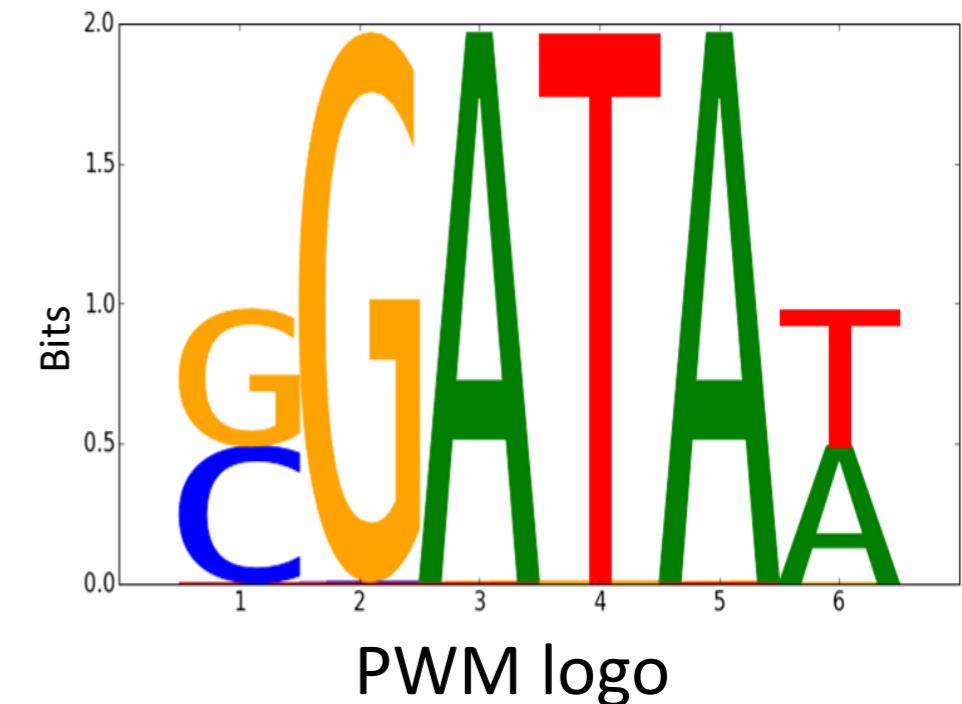
Set of aligned sequences
Bound by TF



..ATGGATTCCCTCC..
..GCATATAGCTAT..
..GTGAACTGGCTG..

	$p_i(x_i = a_i)$					
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)



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

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

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

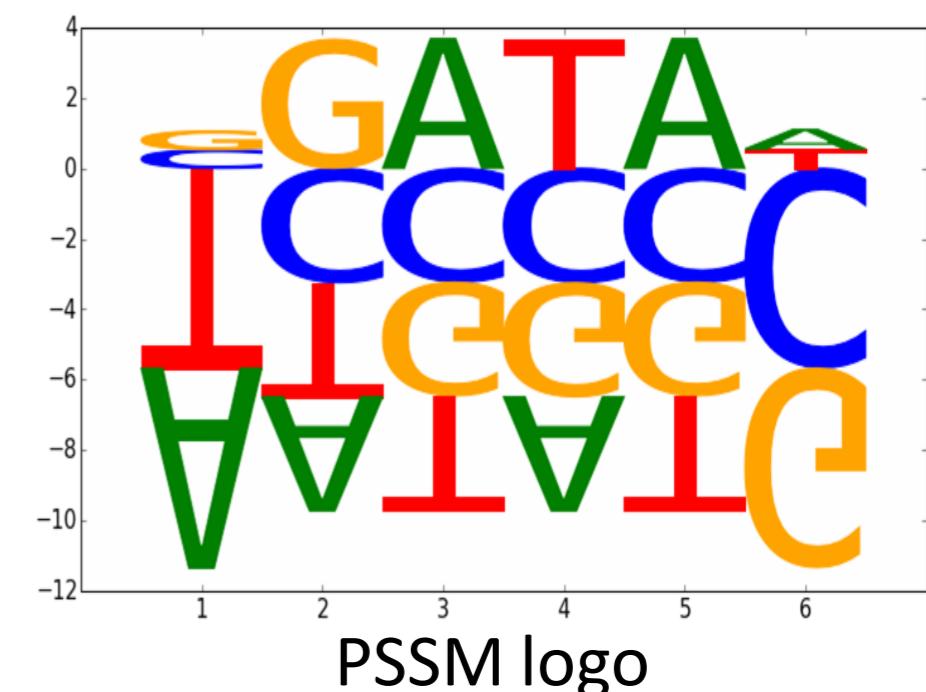
Sequence motifs: PSSM

Accounting for genomic background nucleotide distribution

Position-specific
scoring matrix
(PSSM)

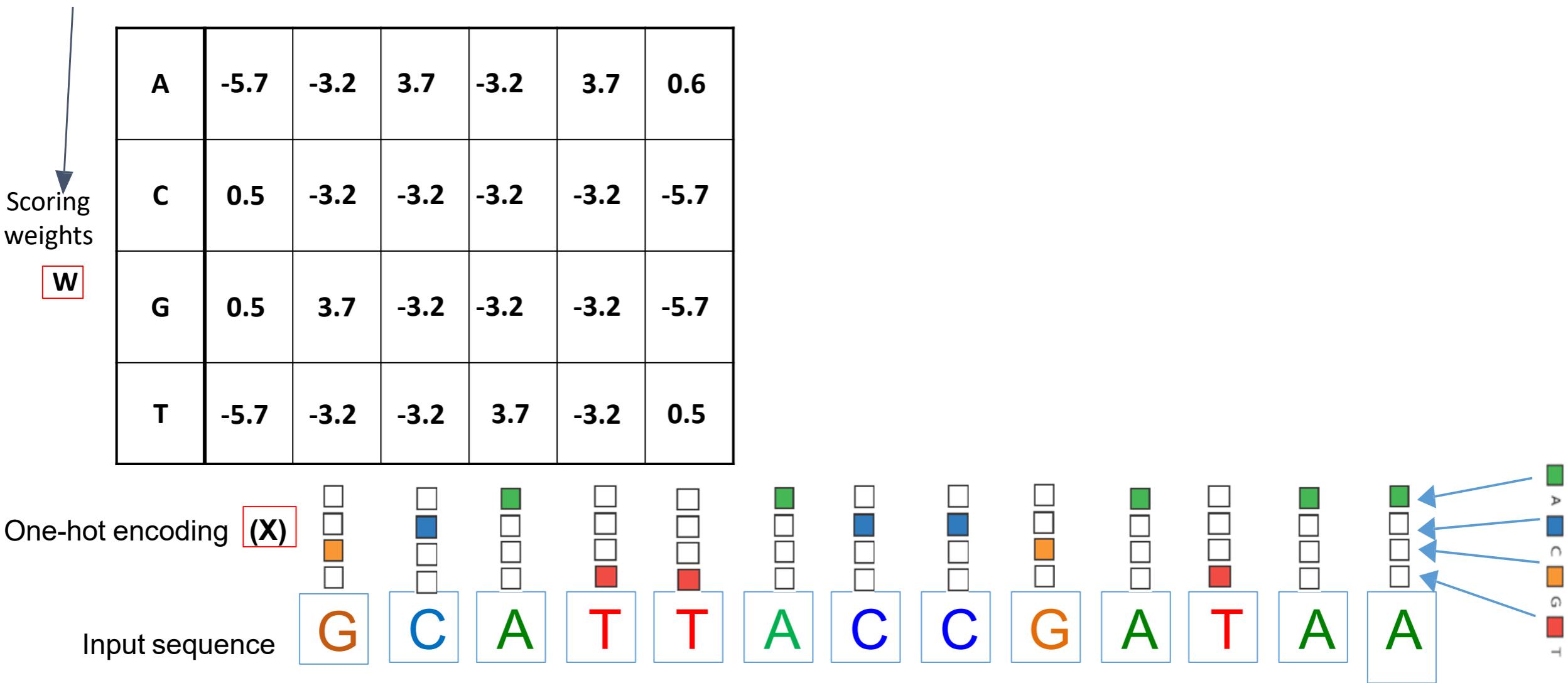
$$\log_2 \left(\frac{p_i(x_i = a_i)}{p_{bg}(x_i = a_i)} \right)$$

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



Scoring a sequence with a motif PSSM

PSSM parameters



Convolution:

Scoring a sequence with a PSSM

Motif match Scores

sum(W * x)

-5.4

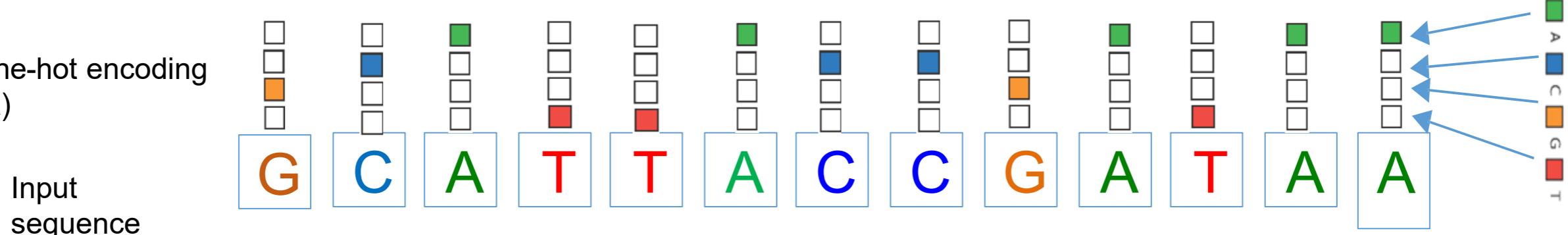
A	-5.7	-3.2	3.7	-3.2	3.7	0.6
---	------	------	-----	------	-----	-----

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)



Convolution

Motif match Scores

sum(w * x)

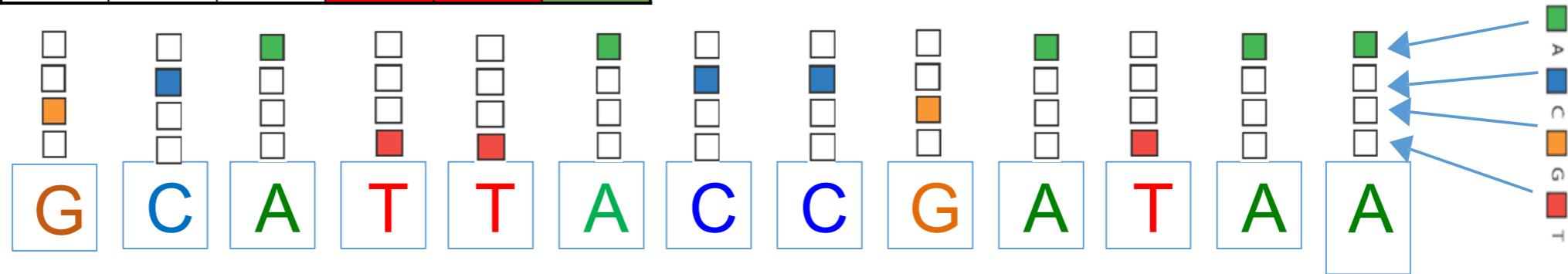
-5.4 2.0

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

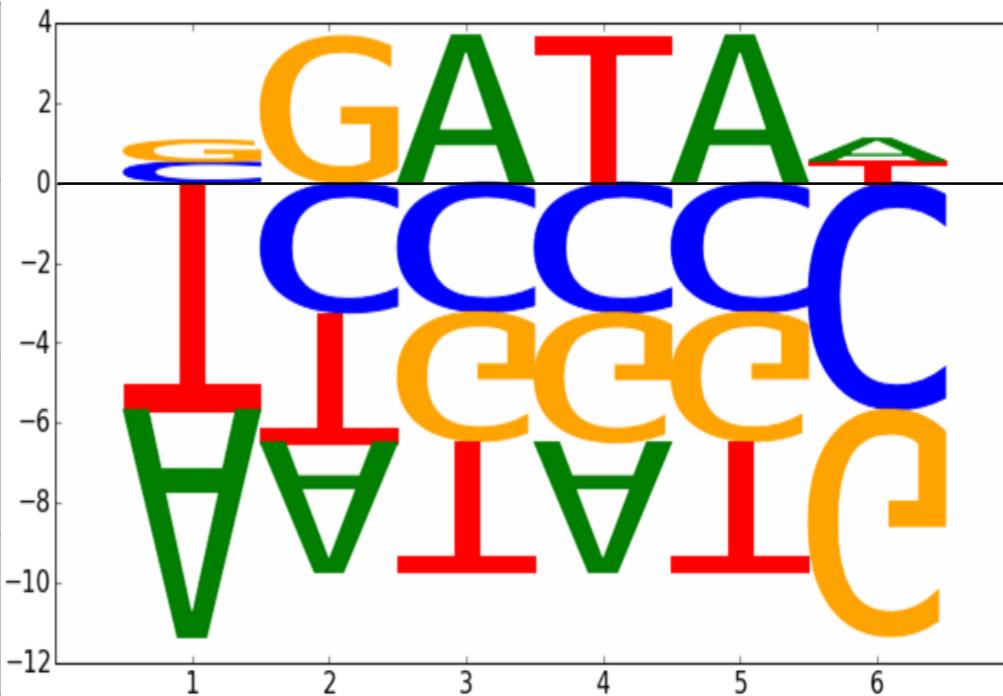


Convolution

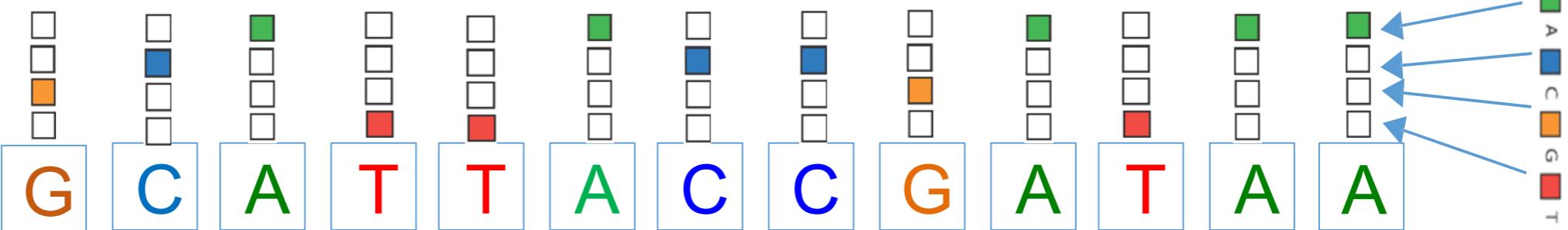
Motif match Scores
 $\sum(W * x)$

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

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



Scoring weights
W



Thresholding scores

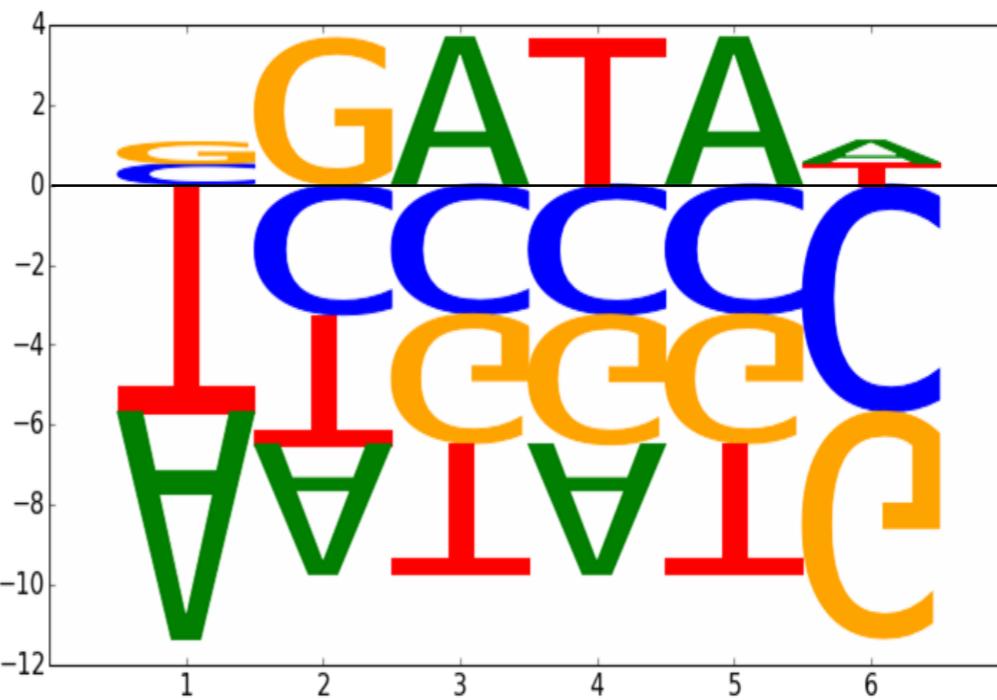
Thresholded
Motif Scores
 $\max(0, W^*x)$

0	0	2.0	0	0	0	0	0	16	0	0	0
---	---	-----	---	---	---	---	---	----	---	---	---

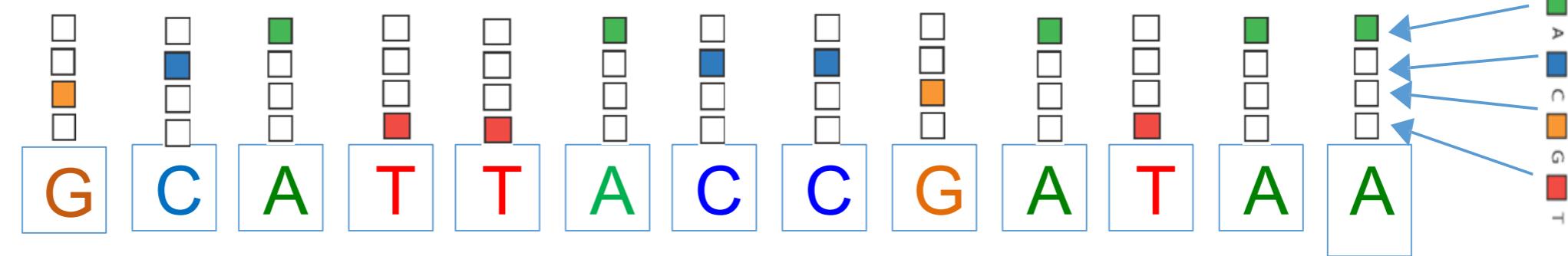
Motif match
Scores
 W^*x

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

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



Scoring
weights
 W



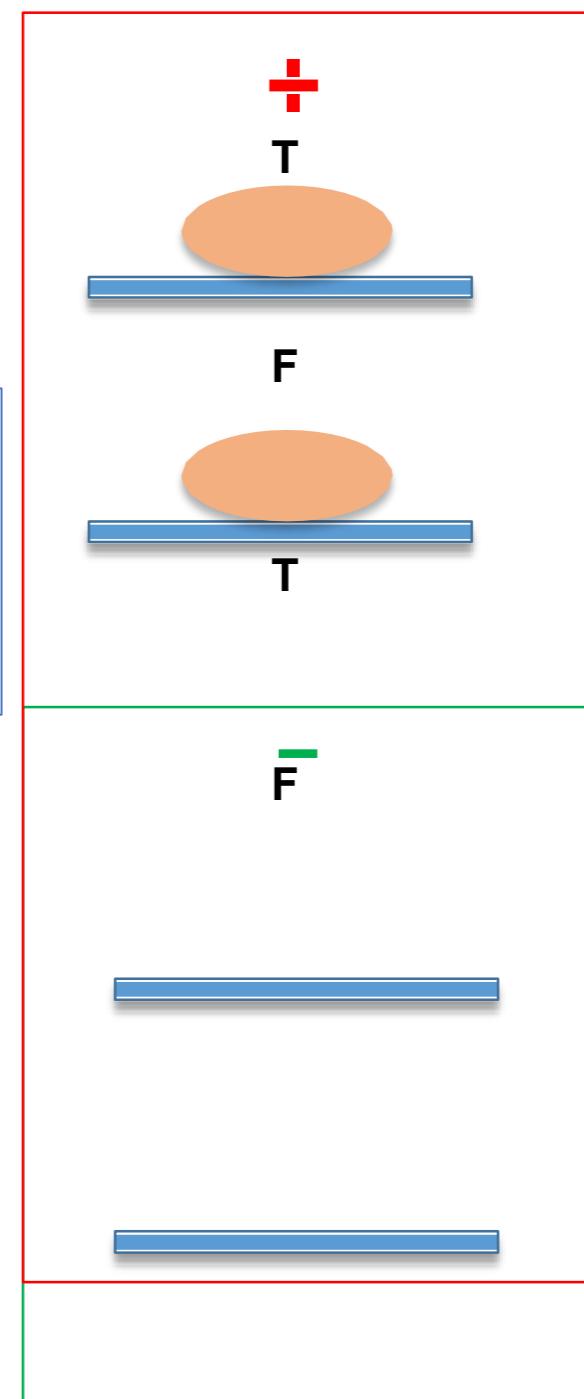
3b. CNNs for Regulatory Genomics Foundations (Higher-level learning)

Learning patterns in regulatory DNA sequence

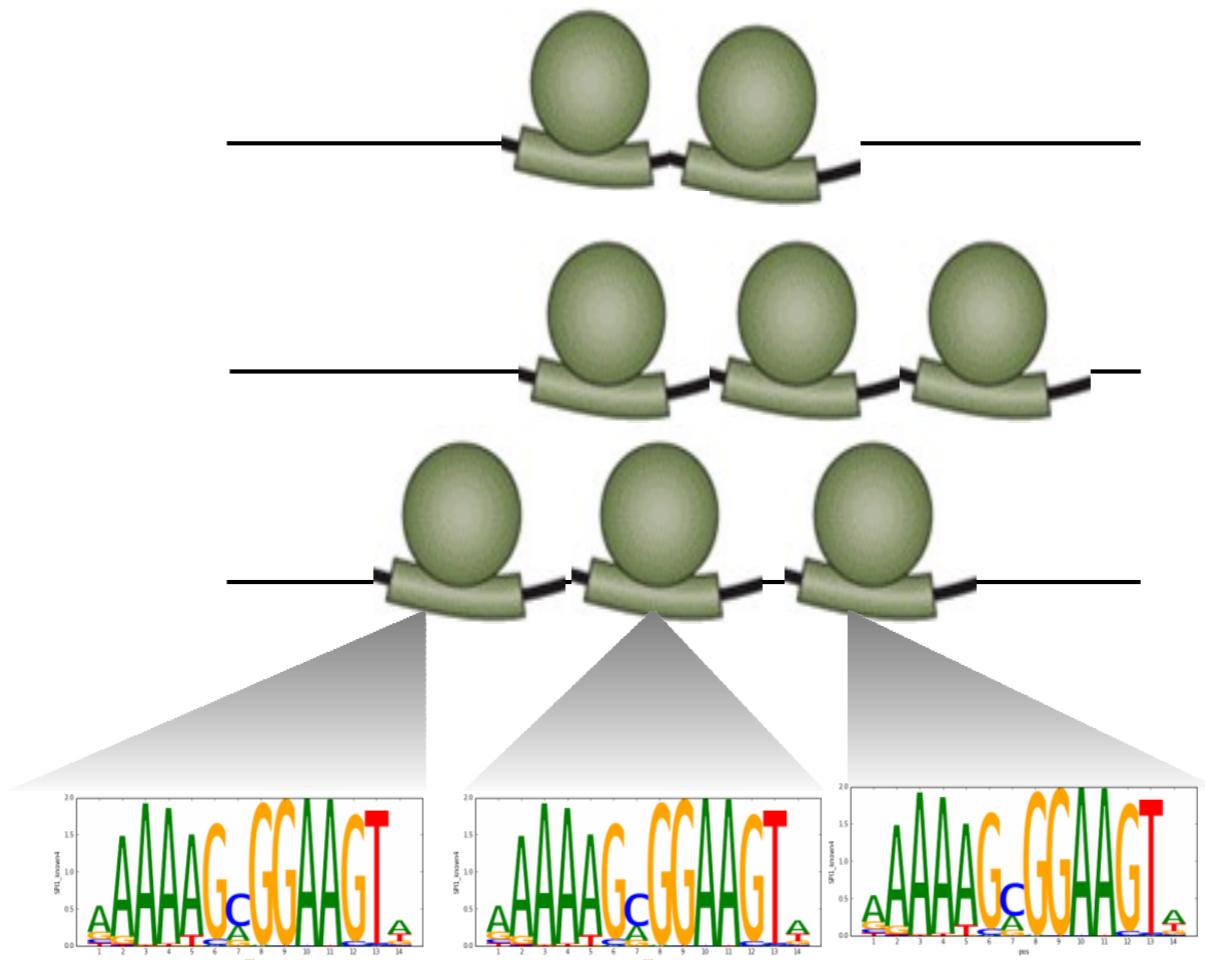
- Positive class of genomic sequences bound a transcription factor of interest

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



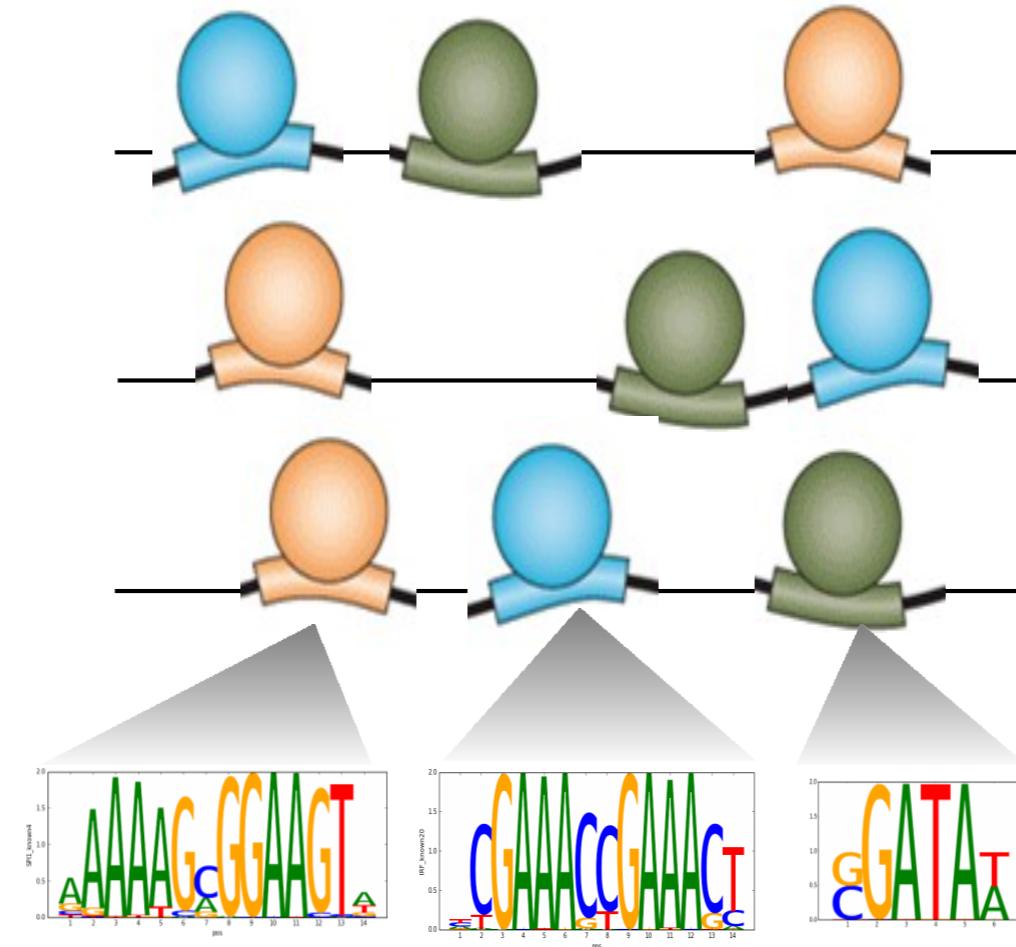
Key properties of regulatory sequence



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

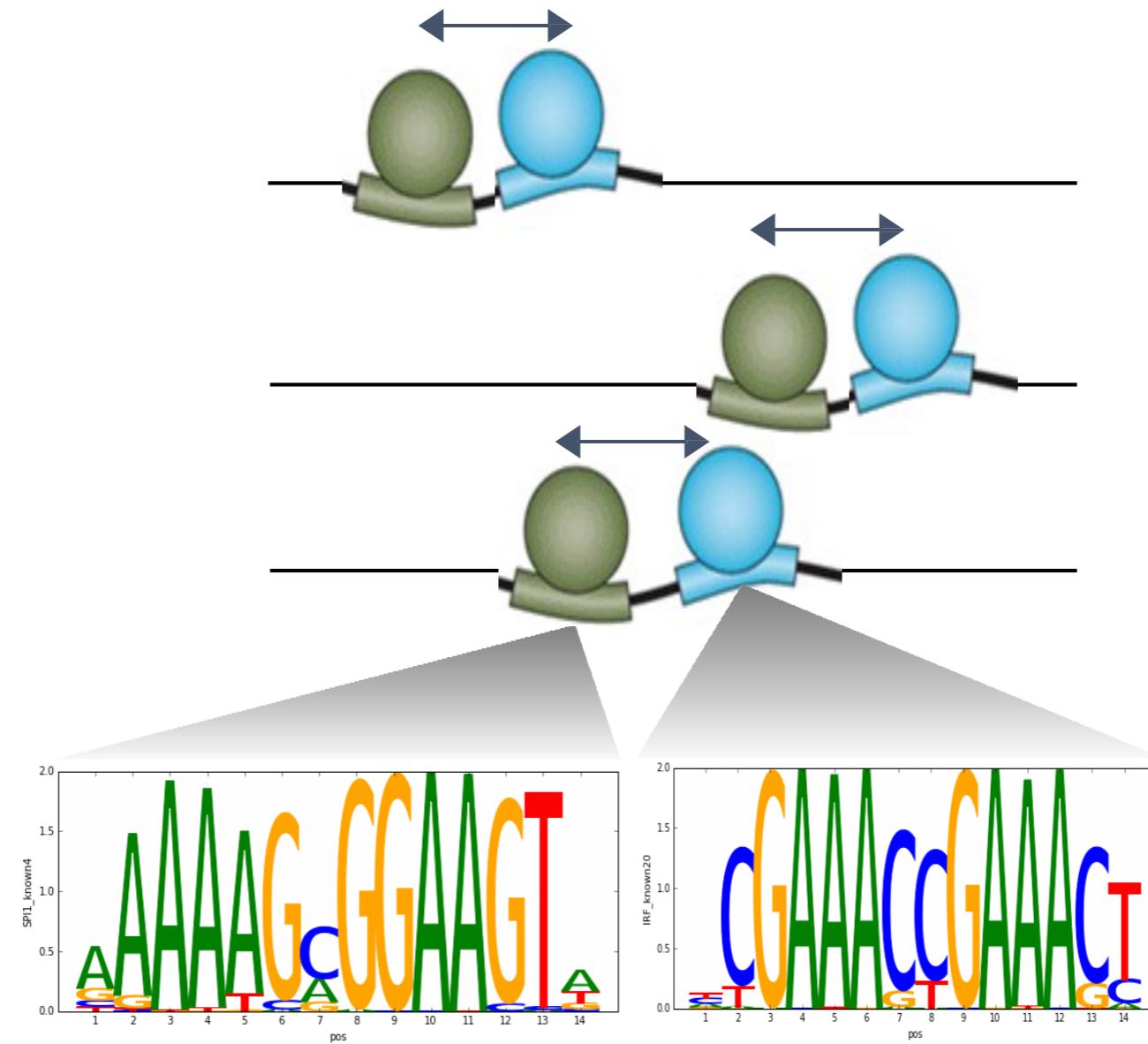
Key properties of regulatory sequence



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 HETEROGENEOUS MOTIF COMBINATIONS

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

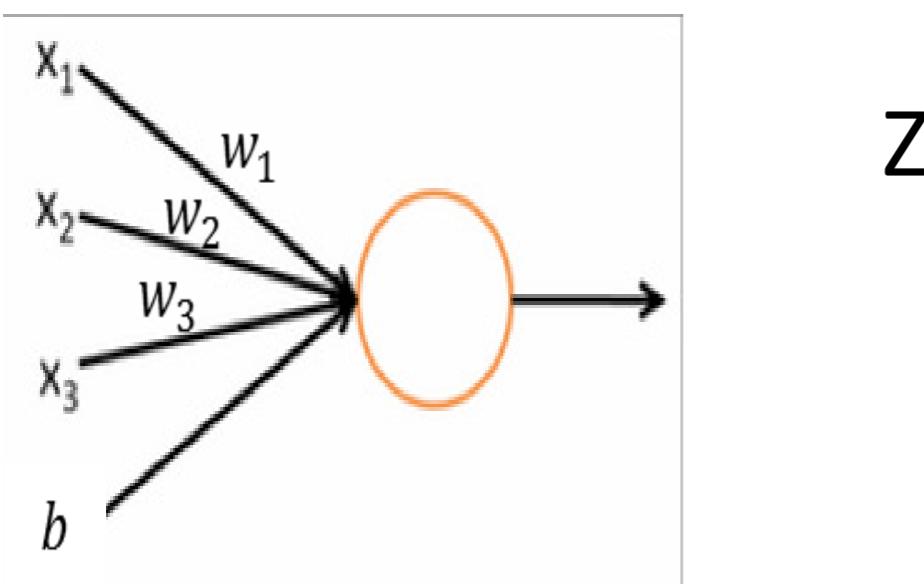
A simple classifier (An artificial neuron)

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

$$Z = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b$$

parameters

Linear function



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

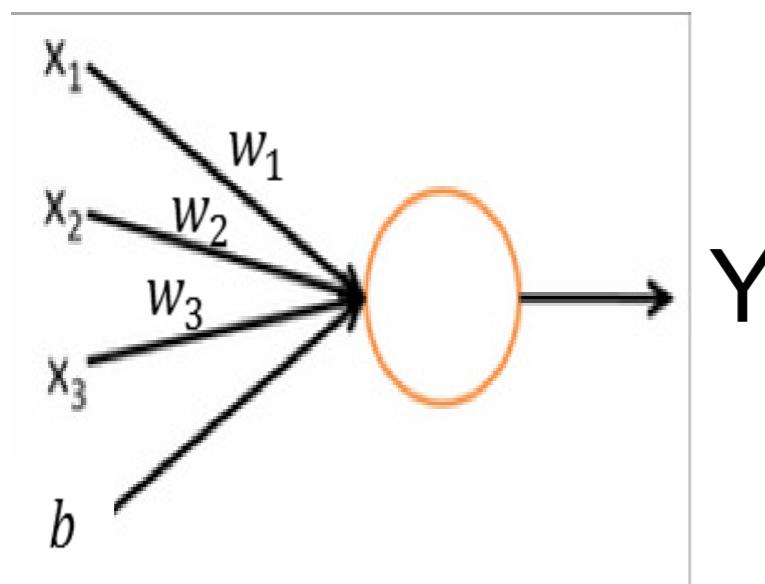
A simple classifier (An artificial neuron)

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

parameters

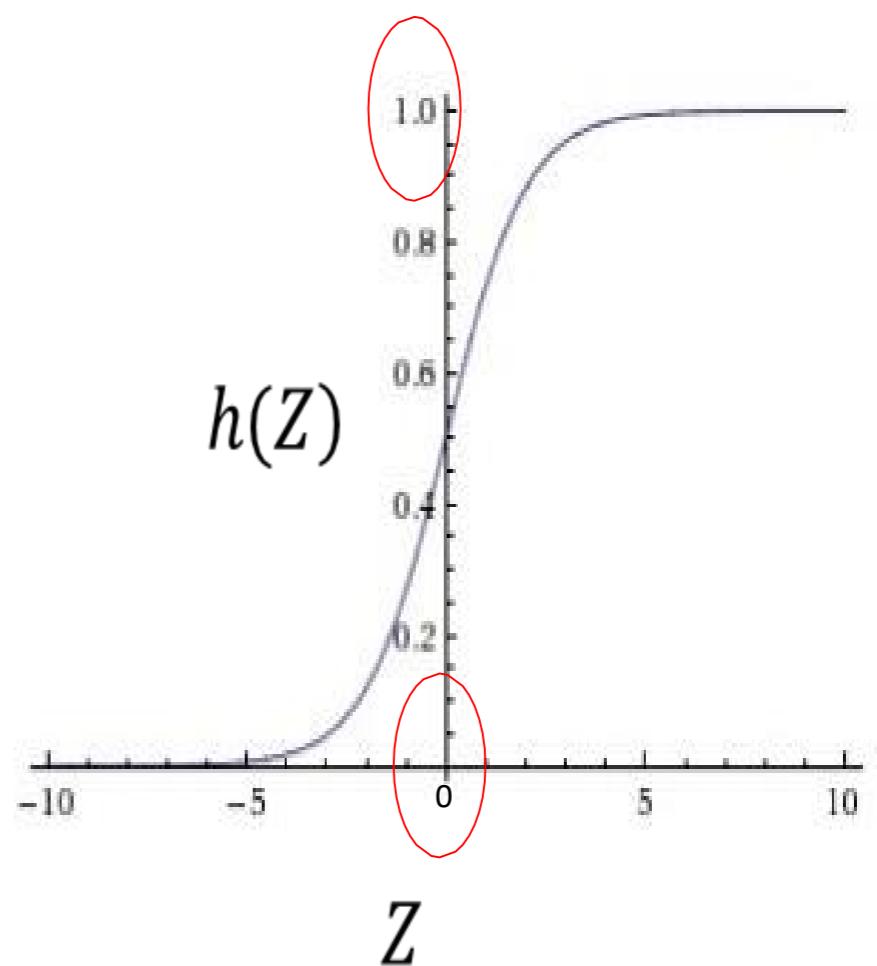
$$Z = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b$$

$$Y = h(Z)$$



Non-linear function

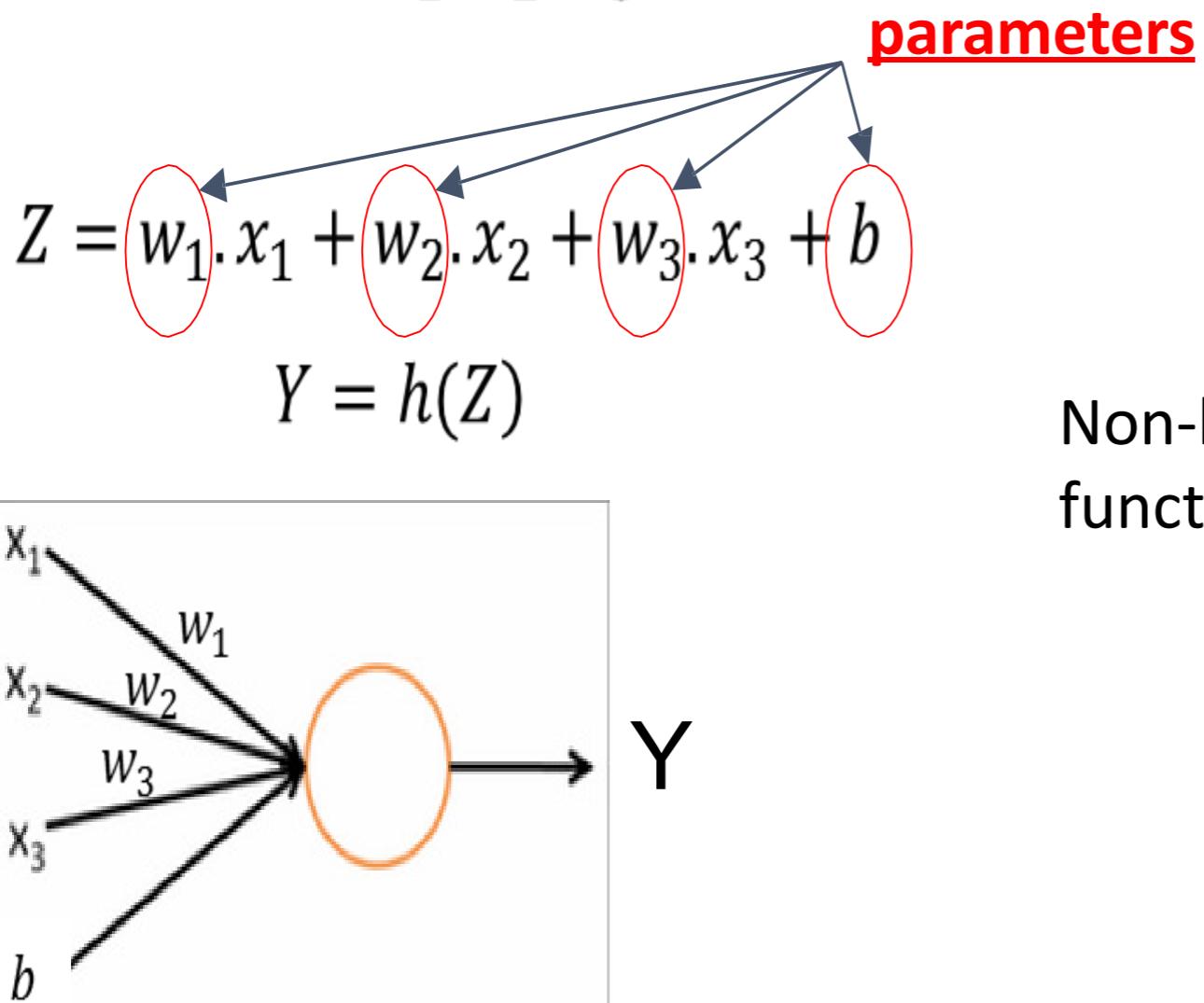
Logistic / Sigmoid
Useful for predicting probabilities



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

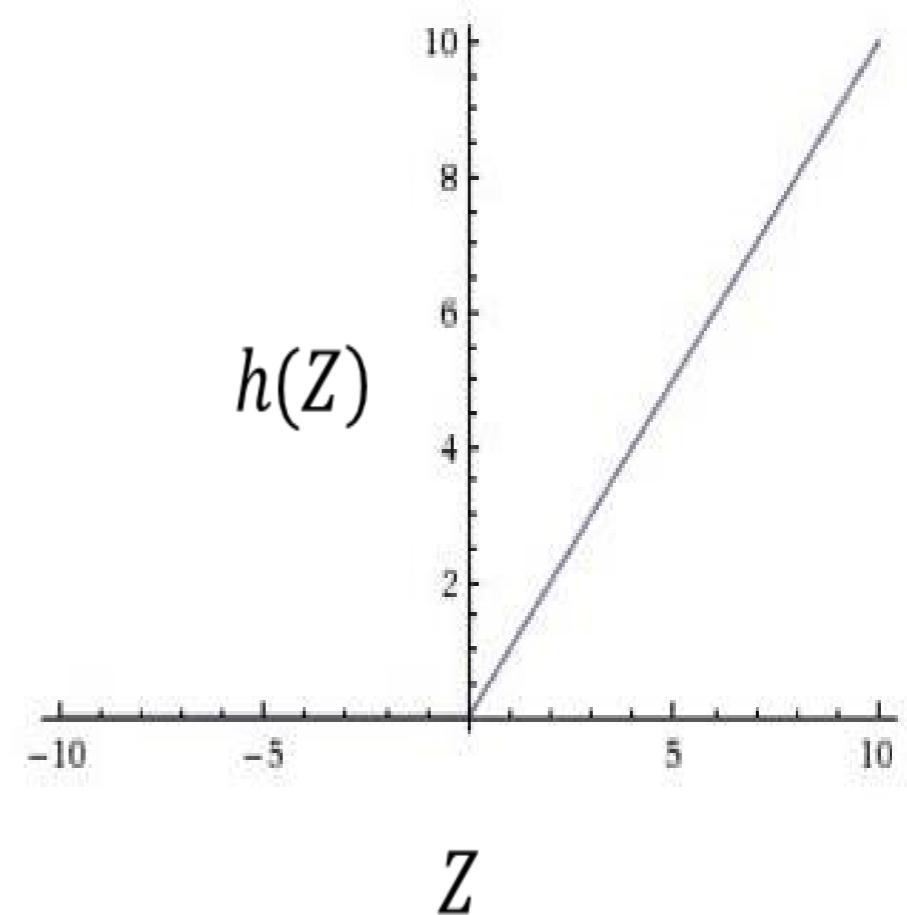
A simple classifier (An artificial neuron)

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



Non-linear function

ReLU (Rectified Linear Unit)
Useful for thresholding



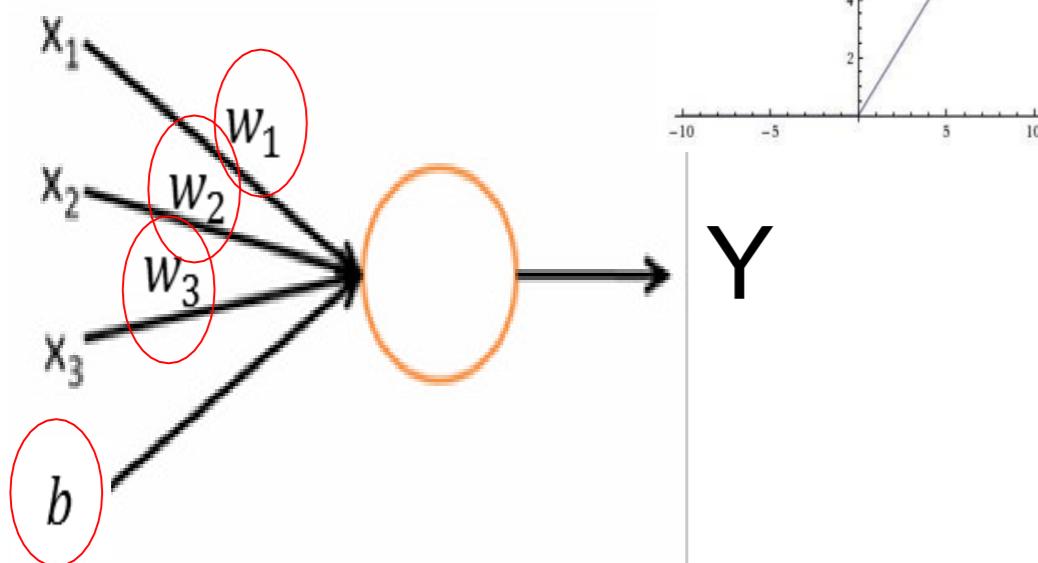
Training the neuron means learning the optimal w's and b

Artificial neuron can represent a motif

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

$$Z = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b$$

$Y = h(Z)$



parameters

Thresholded Motif

Scores
 $\max(0, W^*x)$

0	0	2.0	0	0	0
---	---	-----	---	---	---

Motif match Scores

$\sum(W^*x)$

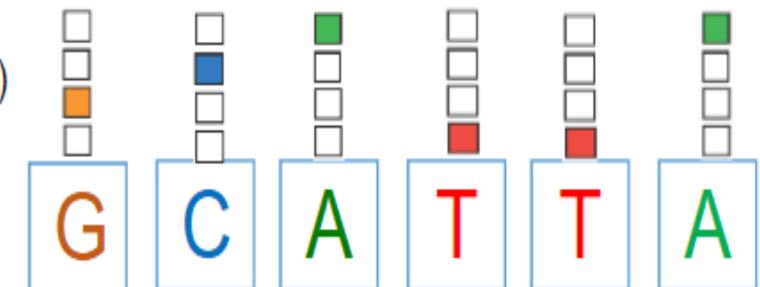
-2.2	-5.5	2.0	-4.3	-24	-17
------	------	-----	------	-----	-----

Scoring weights
 W

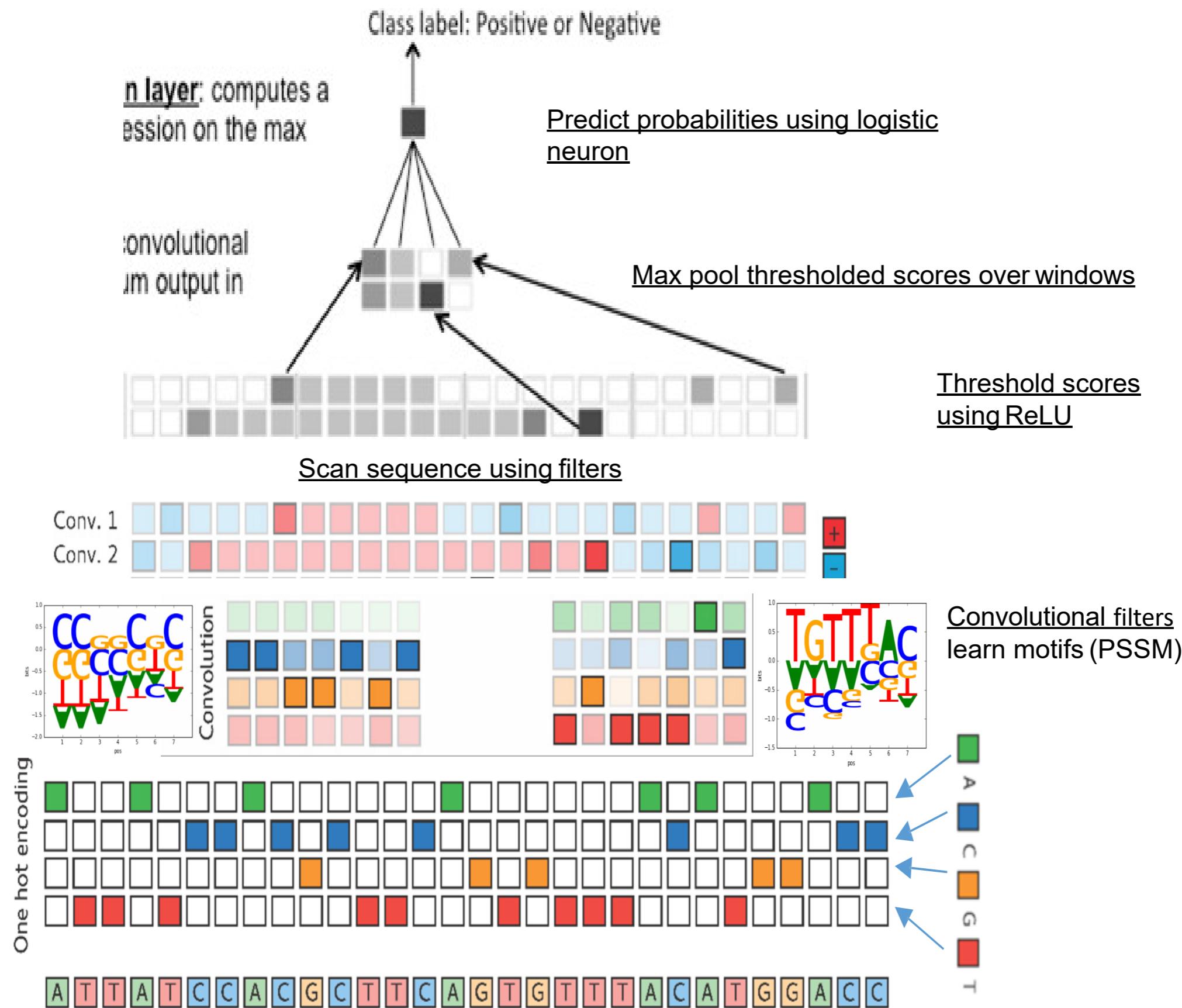
A	-5.7	-3.2	3.7	-3.2	3.7	0.6
C	0.5	-3.2	-3.2	-3.2	-3.2	-5.7
G	0.5	3.7	-3.2	-3.2	-3.2	-5.7
T	-5.7	-3.2	-3.2	3.7	-3.2	0.5

One-hot encoding (X)

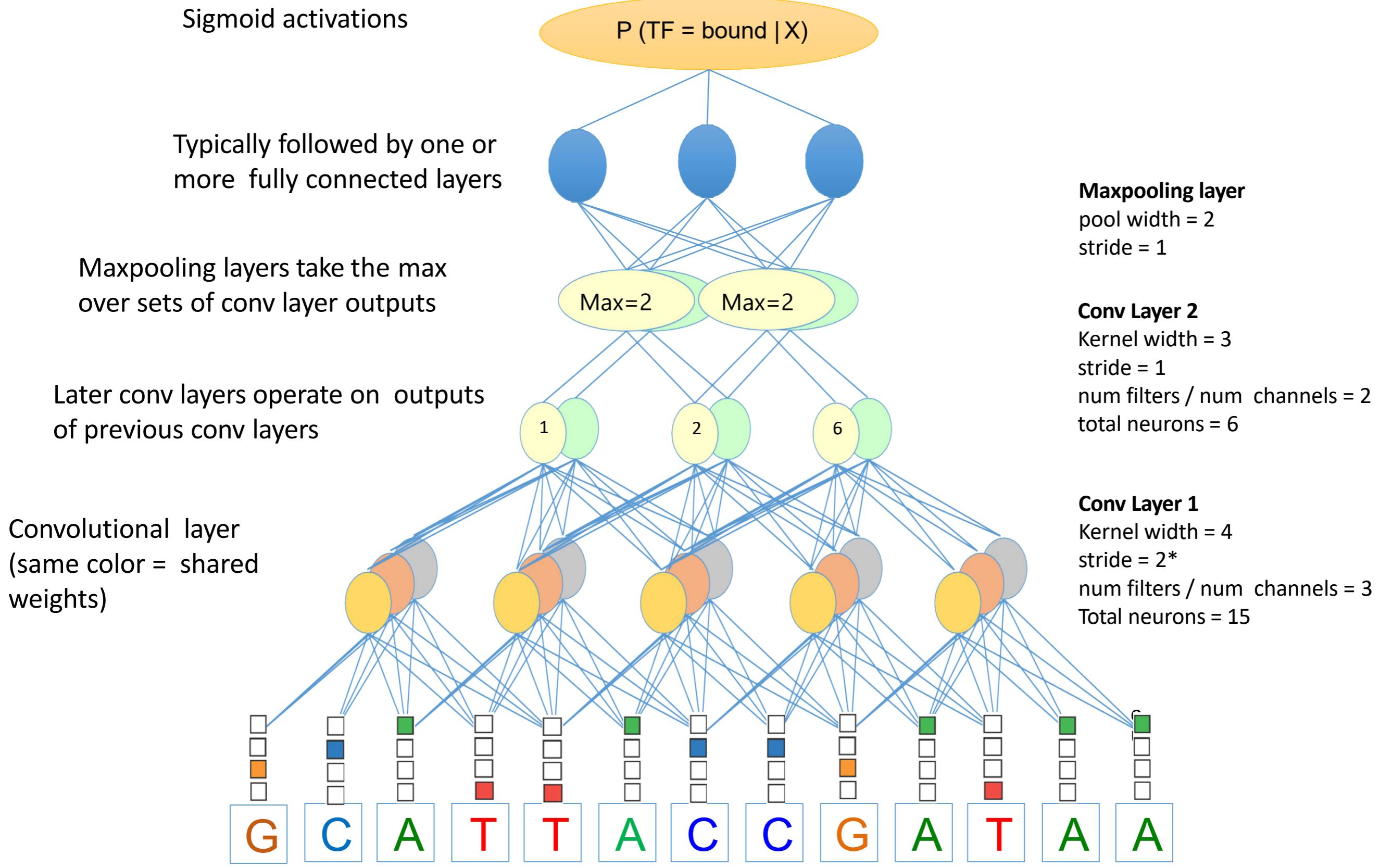
Input sequence



Biological motivation of Deep CNN



Deep convolutional neural network



*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

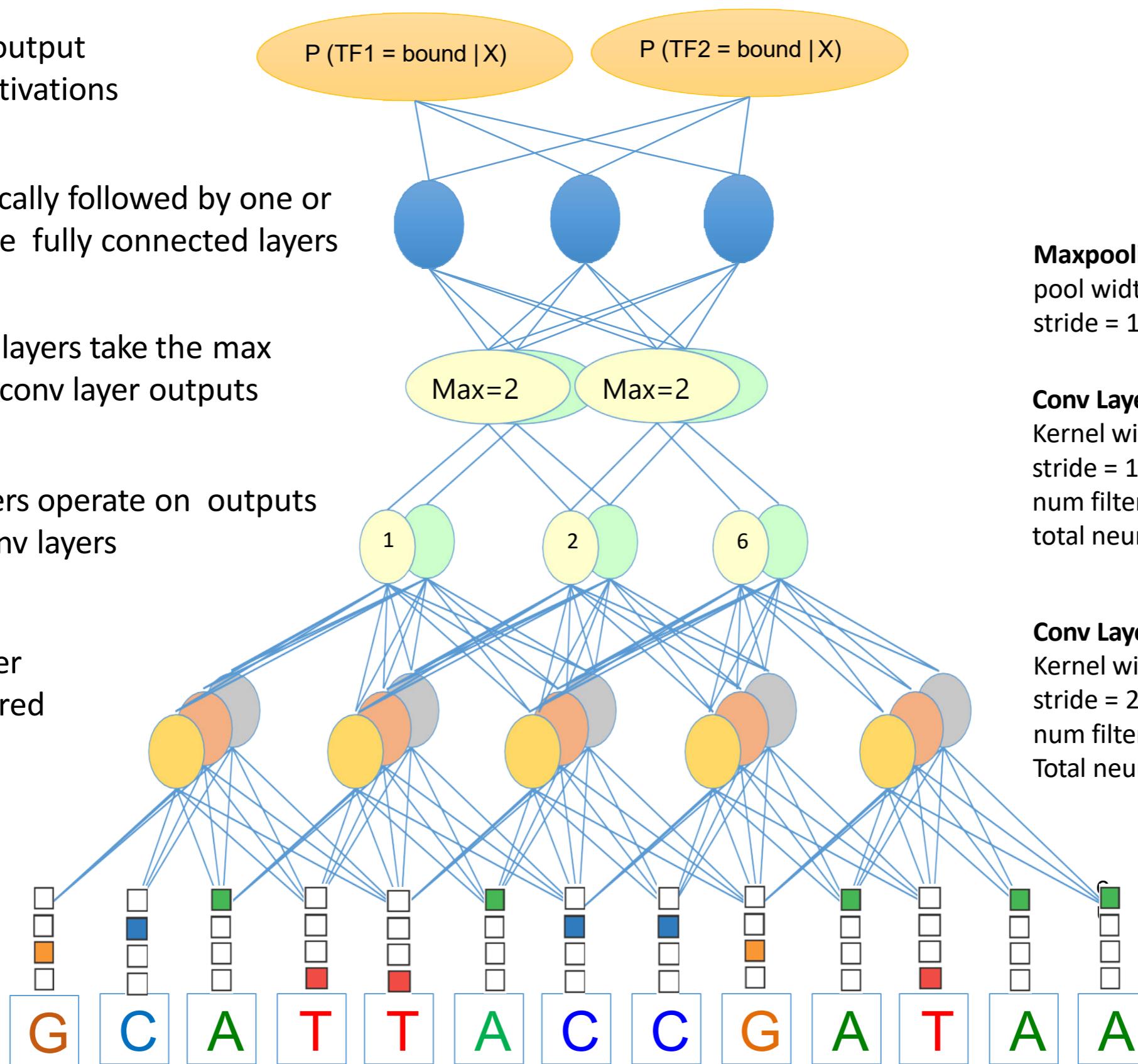
Later conv layers operate on outputs of previous conv layers

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

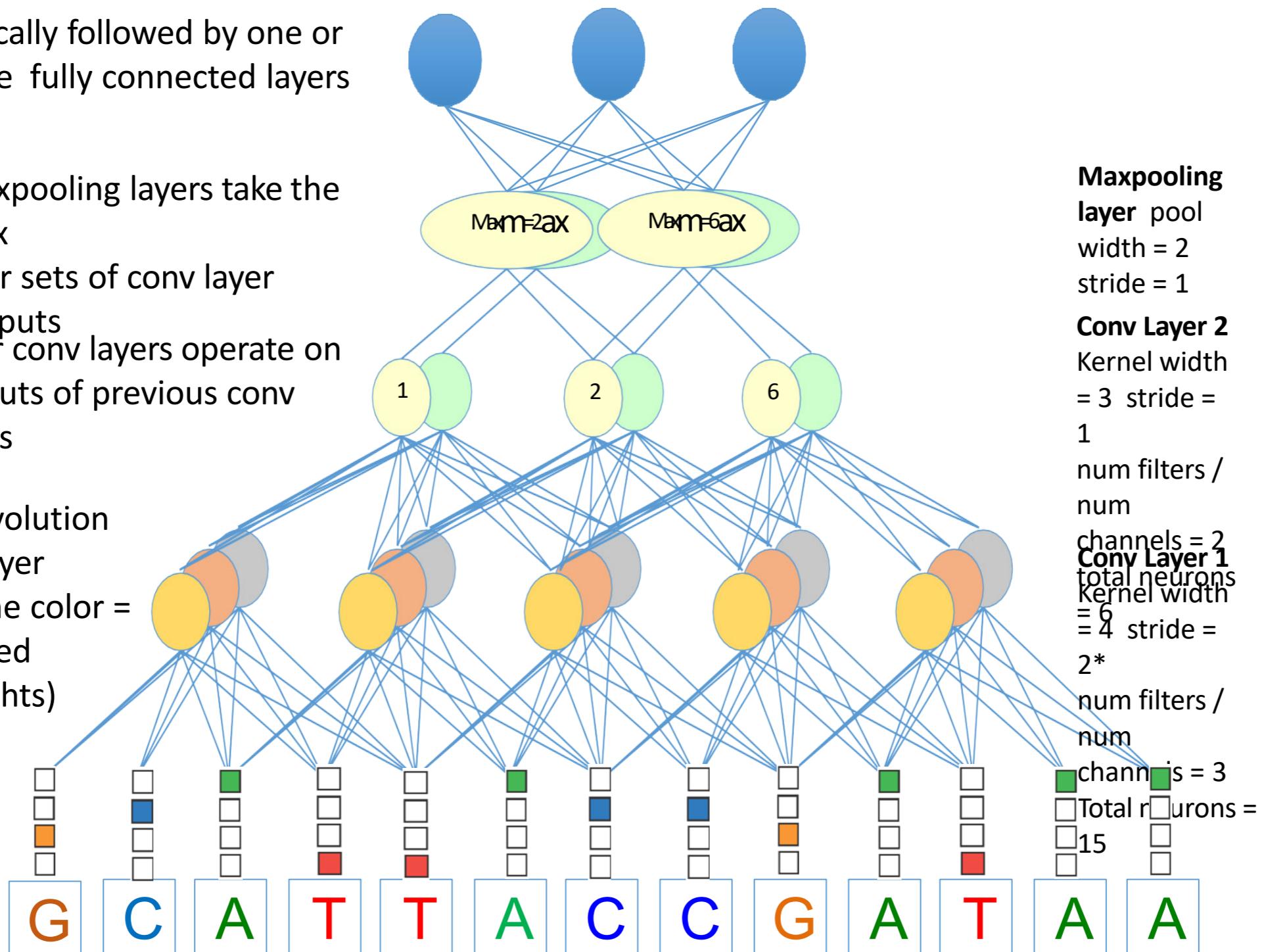


Multi-task CNN

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)



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
- Three-dimensional chromatin structure, Hi-C, ChIA-PET, TADs, Loop Extrusion

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 \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow Motif matches

4. Regulatory Genomics CNNs/RNNs in Practice: Diverse Architectures

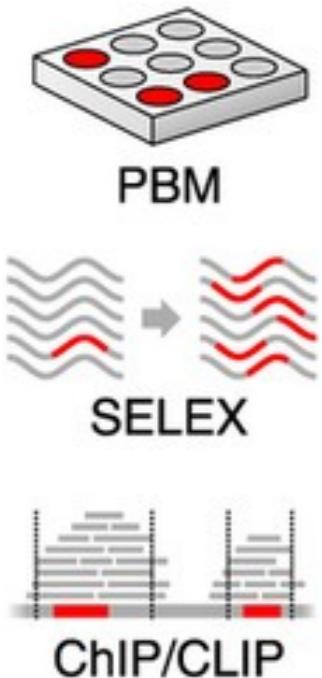
- DeepBind: Learn motifs, use in (shallow) fully-connected layer, mutation impact
- DeepSea: Train model directly on mutational impact prediction
- Basset: Multi-task DNase prediction in 164 cell types, reuse/learn motifs
- ChromPuter: Multi-task prediction of different TFs, reuse partner motifs
- DeepLIFT: Model interpretation based on neuron activation properties
- DanQ: Recurrent Neural Network for sequential data analysis

5. Guest Lecture: David Kelley on Basset and Deep Learning for Hi-C looping

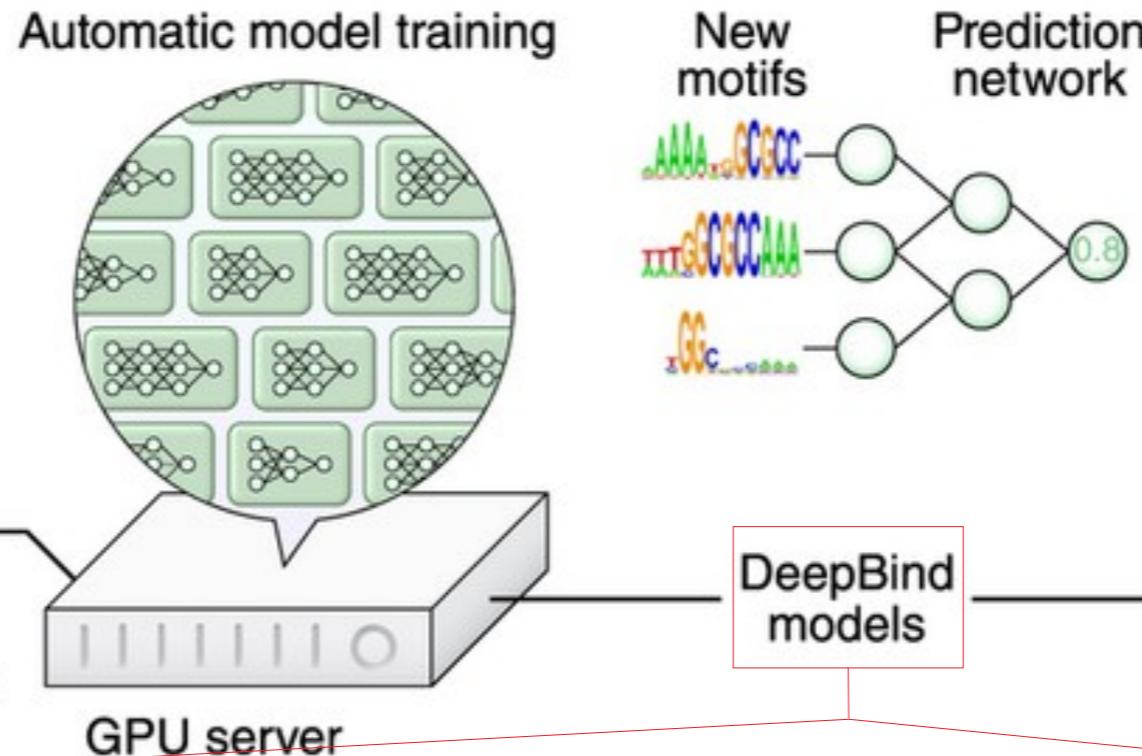
4. Regulatory Genomics CNNs in Practice: (a) DeepBind

DeepBind

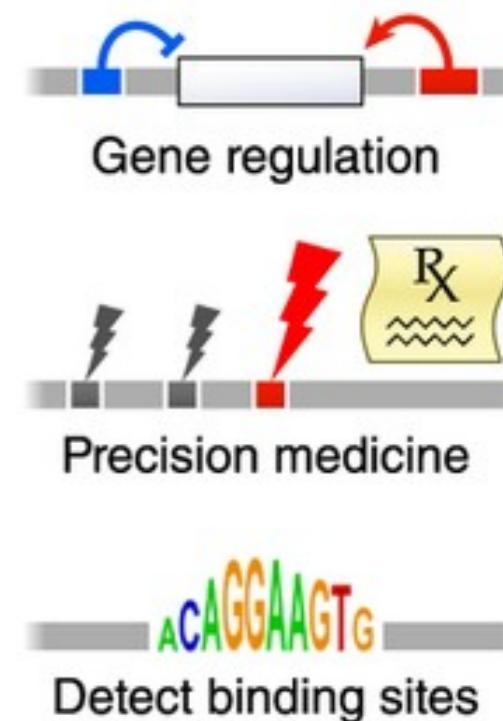
1. High-throughput experiments



2. Massively parallel deep learning



3. Community needs



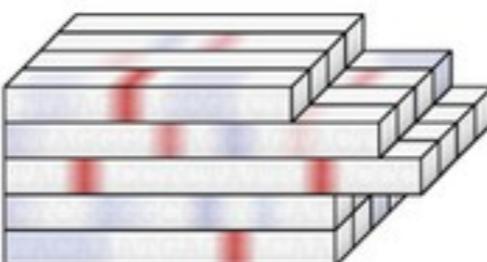
Current batch of inputs

CTAACGCACCGTCT
TTAGGGGCACCAGTACT
TAGCACCTCTATTGCACCC
CTCGGGGCCCTGCAT
TACAAATGAGCACAA

Convolve

Motif detectors

Motif scans

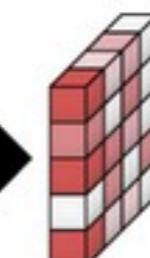


Rectify

Pool

Thresholds

Features



Neural network

Weights

Outputs

Targets

Current model parameters

Parameter updates

Update +

Backprop

Prediction errors

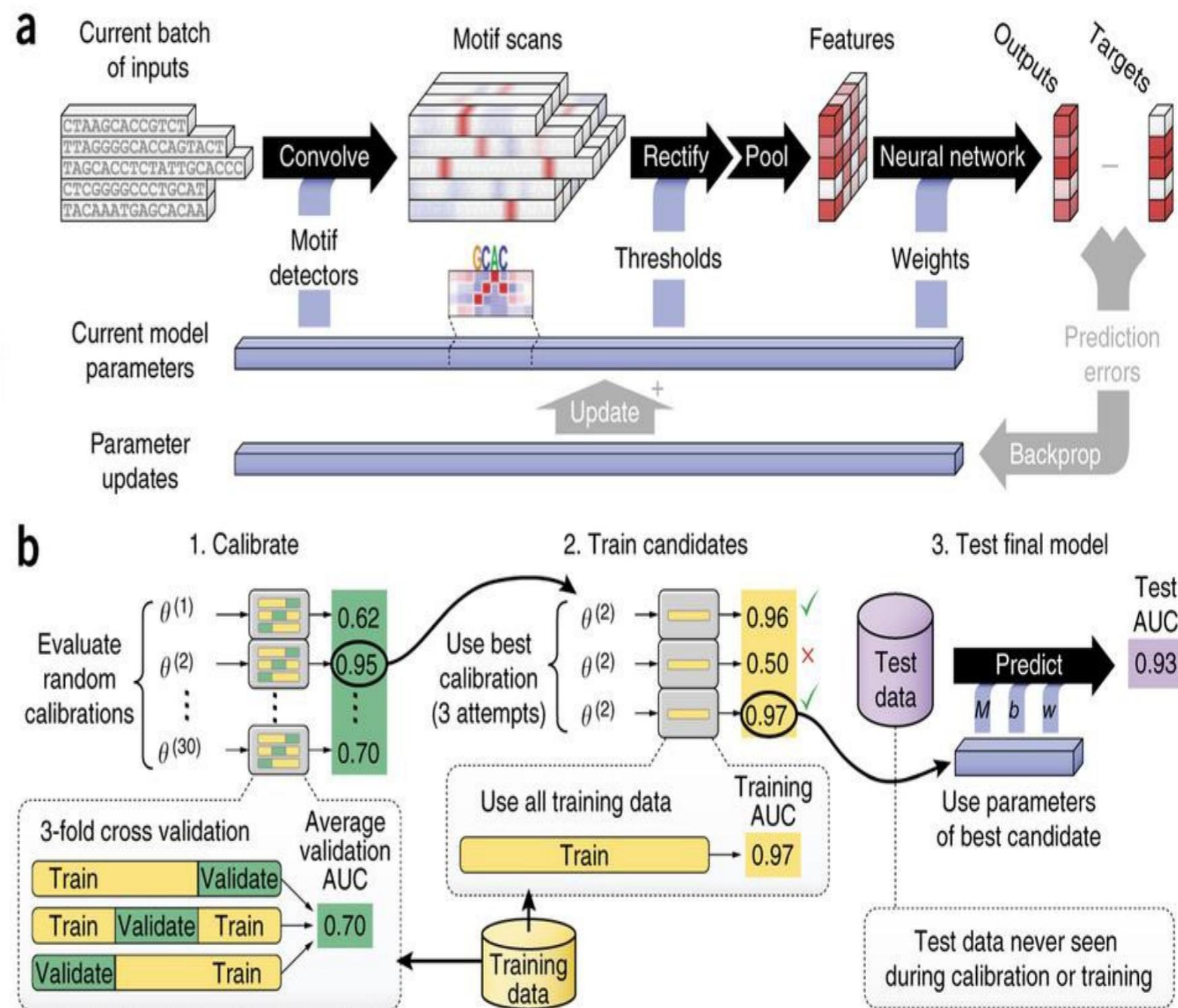
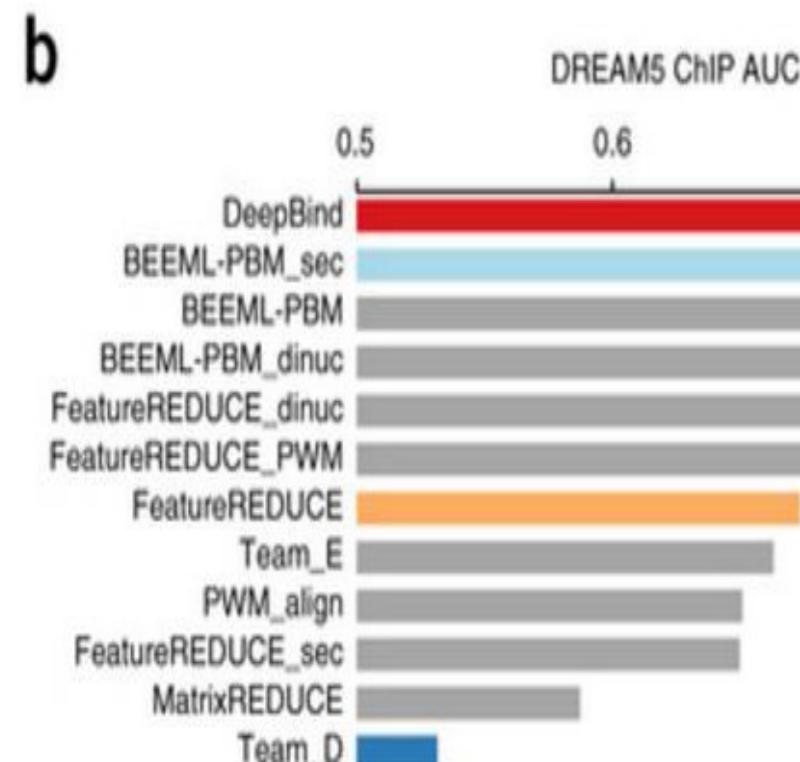
Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

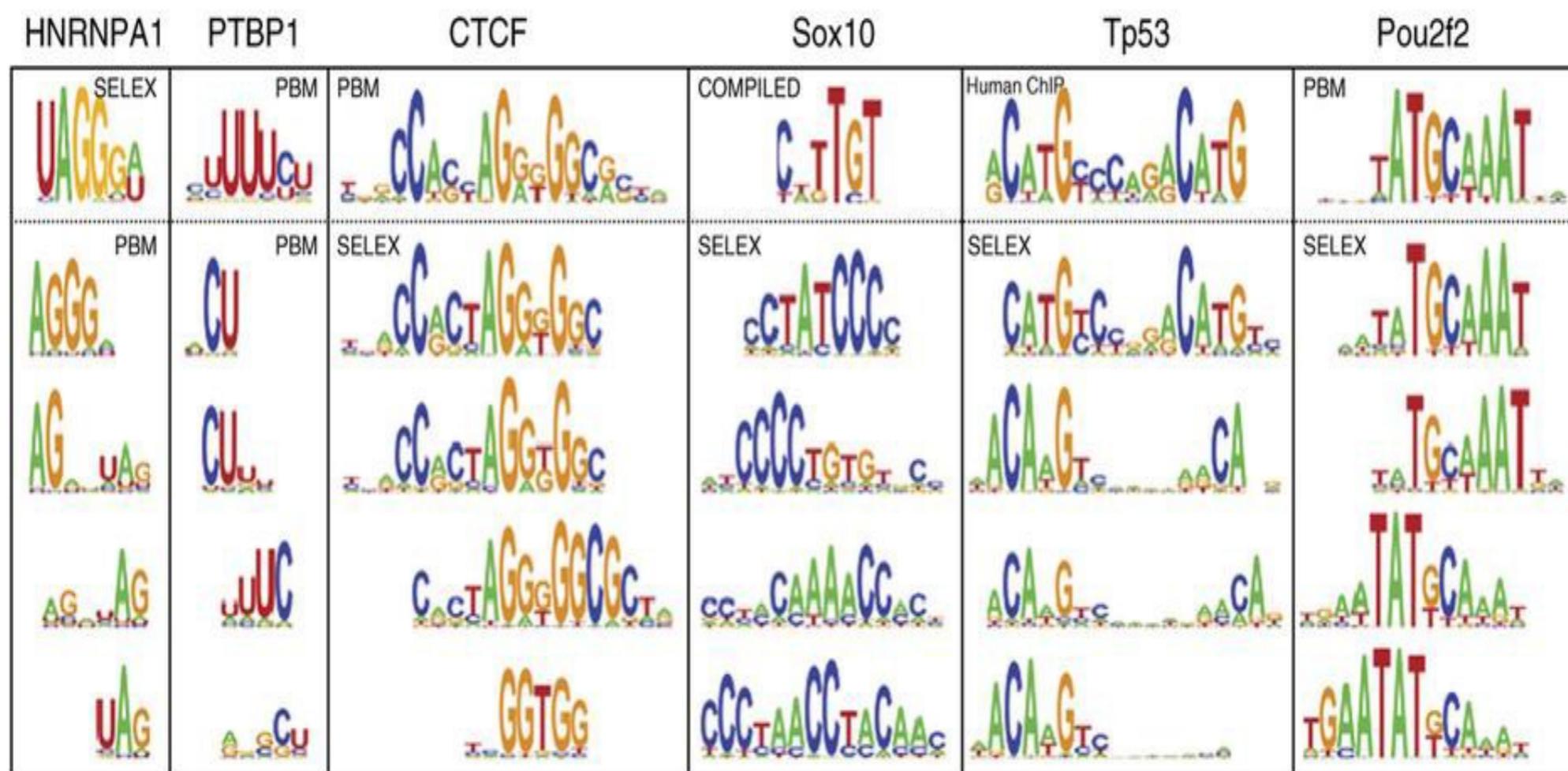
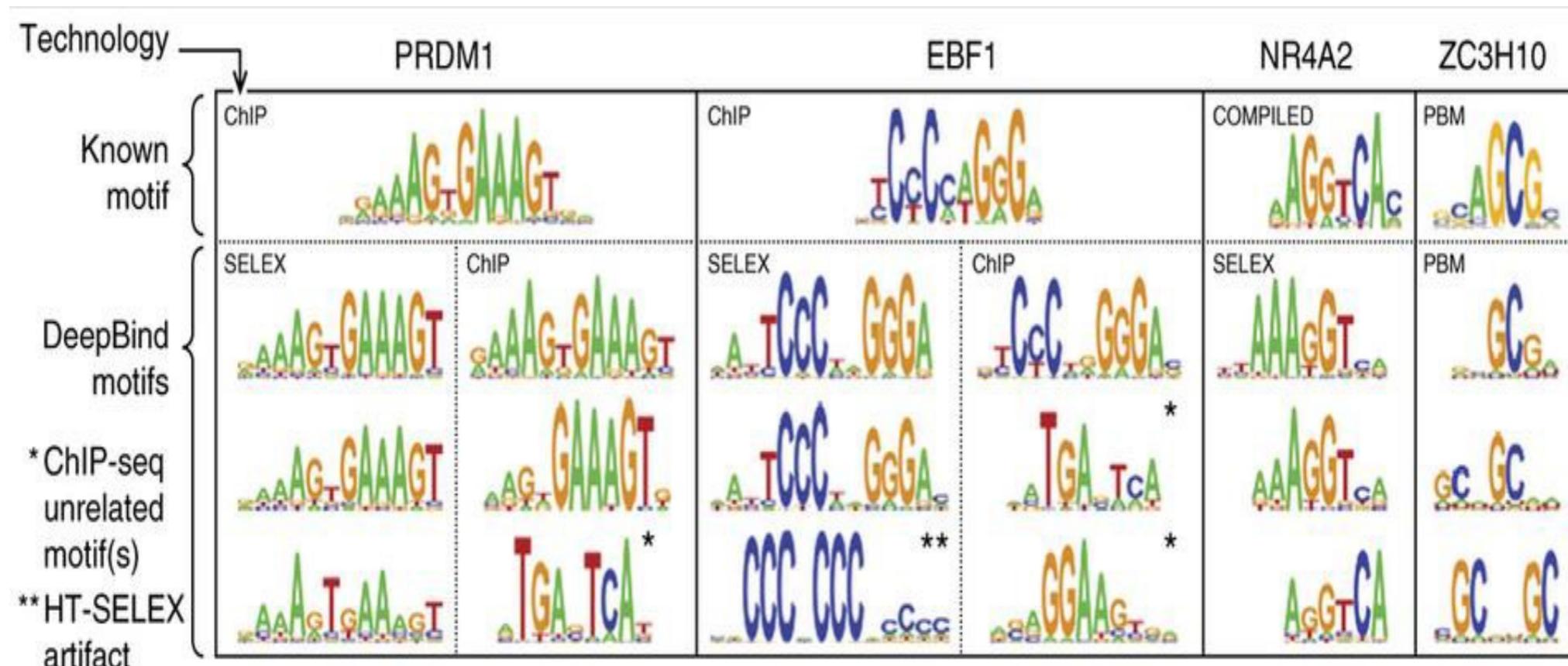
Babak Alipanahi, Andrew Delong, Matthew T Weirauch & Brendan J Frey

Affiliations | Contributions | Corresponding author

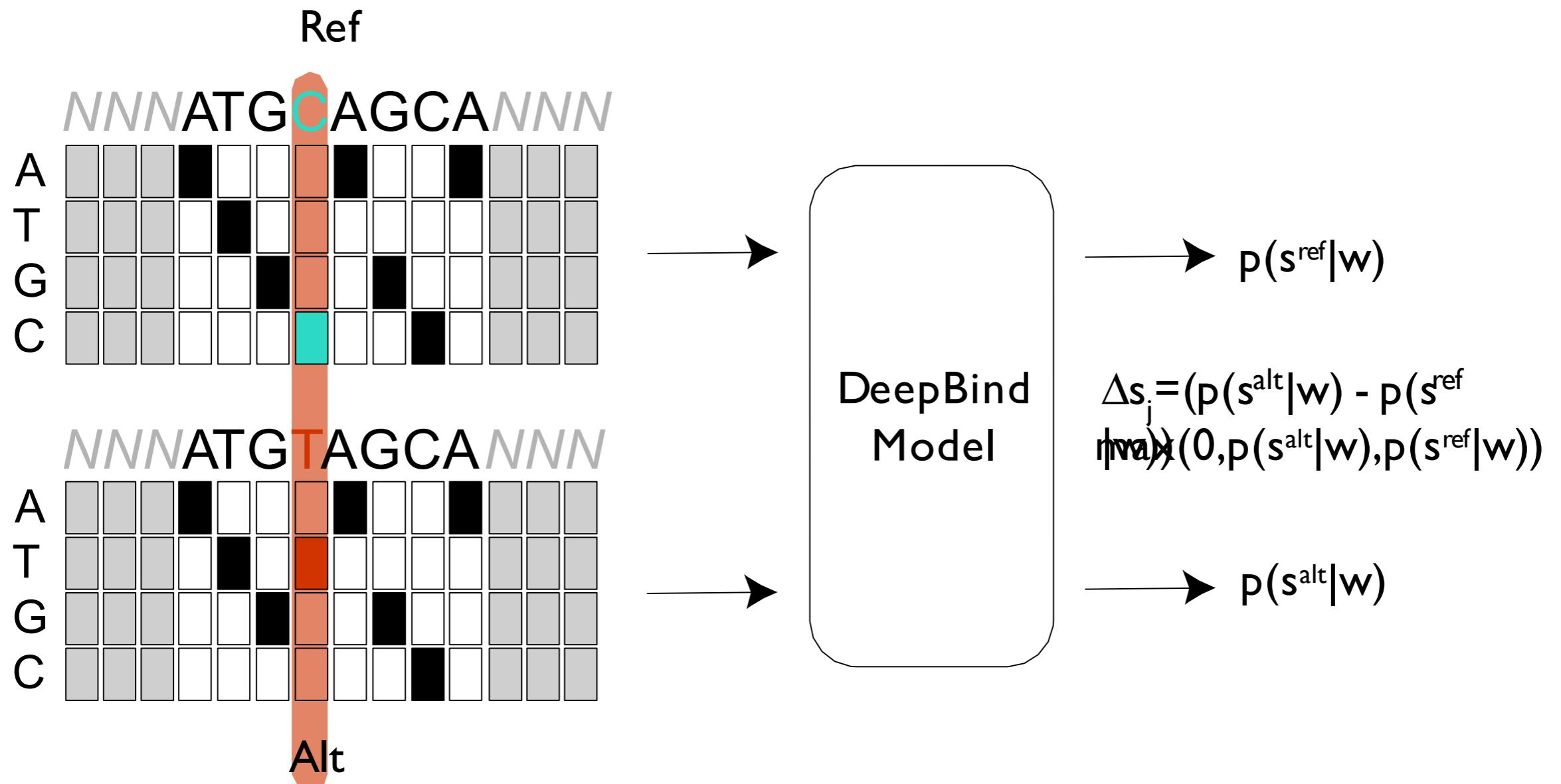
Nature Biotechnology 33, 831–838 (2015) | doi:10.1038/nbt.3300

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

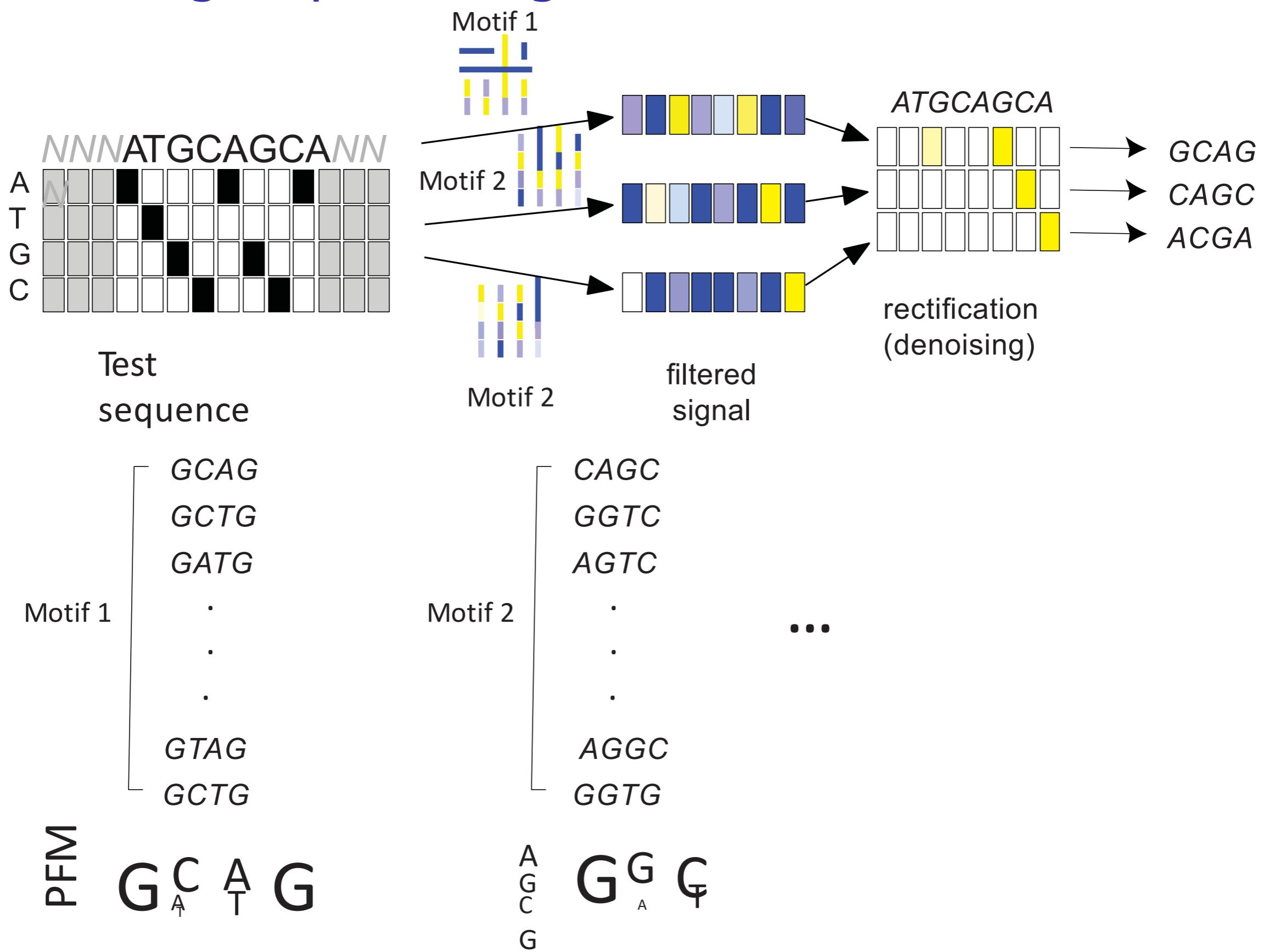




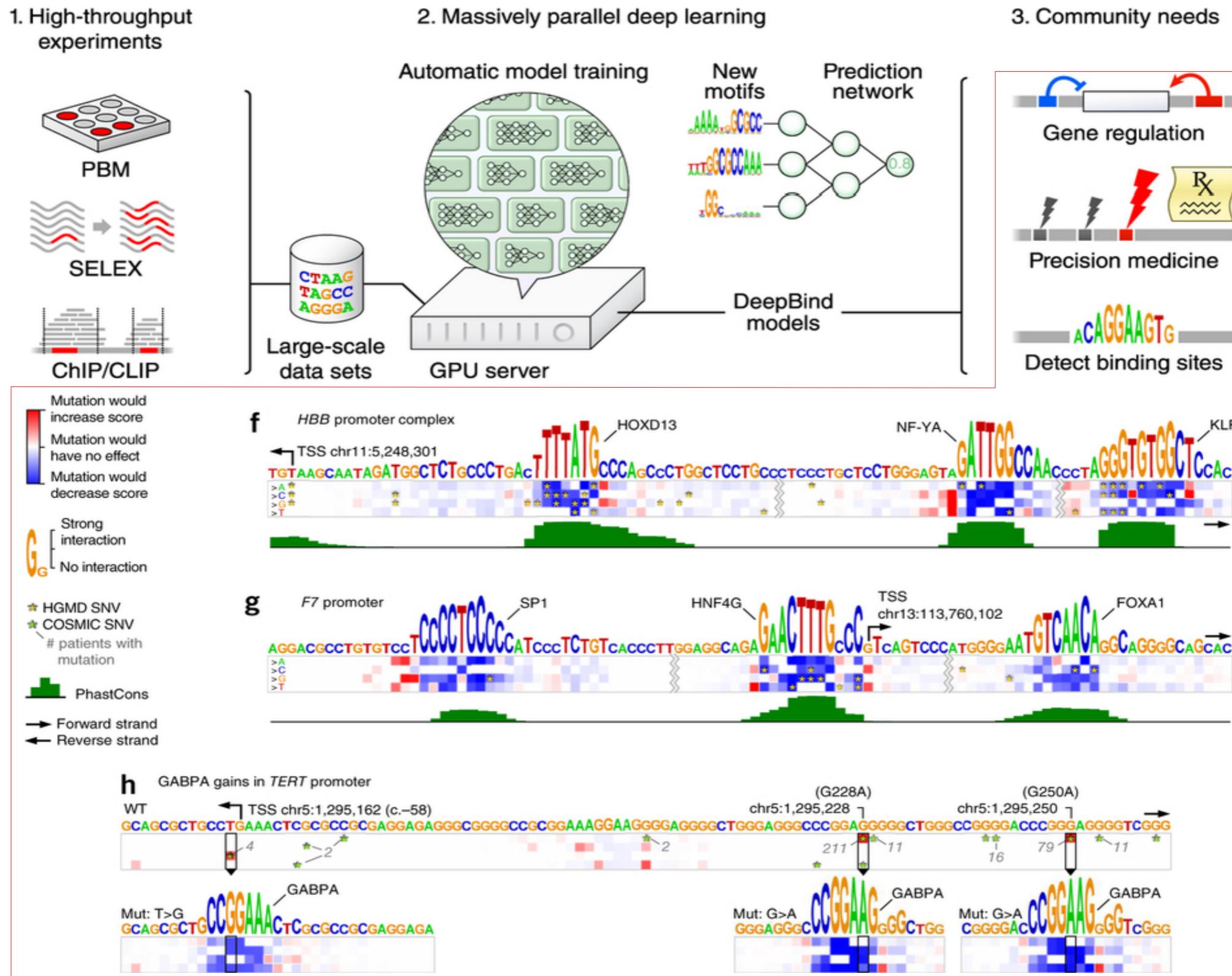
Constructing mutation map



Constructing sequence logo



Predicting disease mutations



DeepBind summary

The key deep learning techniques:

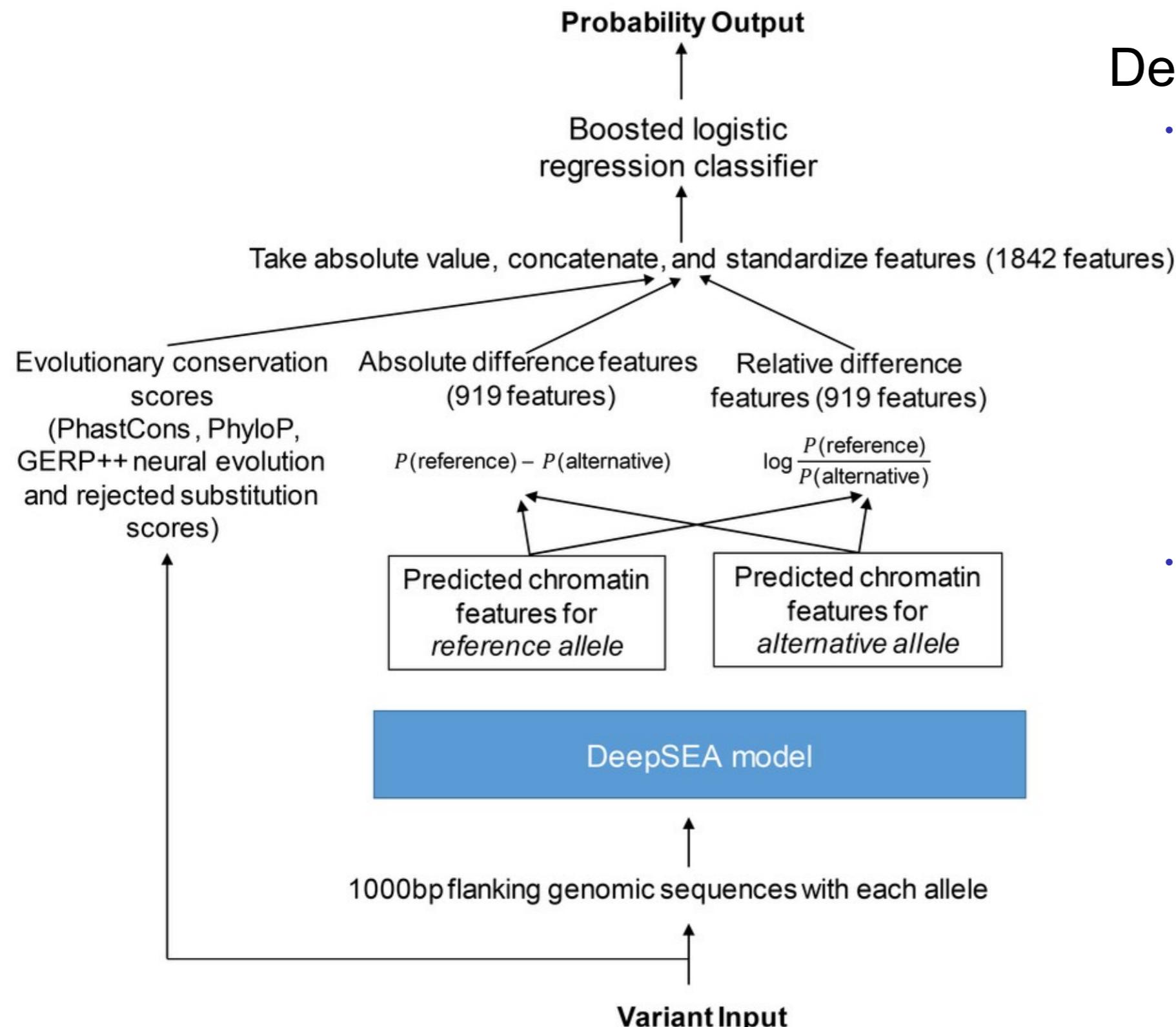
- Convolutional learning
- Representational learning
- Back-propagation and stochastic gradient
- Regularization and dropout
- Parallel GPU computing especially useful for hyperparameter search

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



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

Regulatory Genomics CNNs in Practice: (c) Basset

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

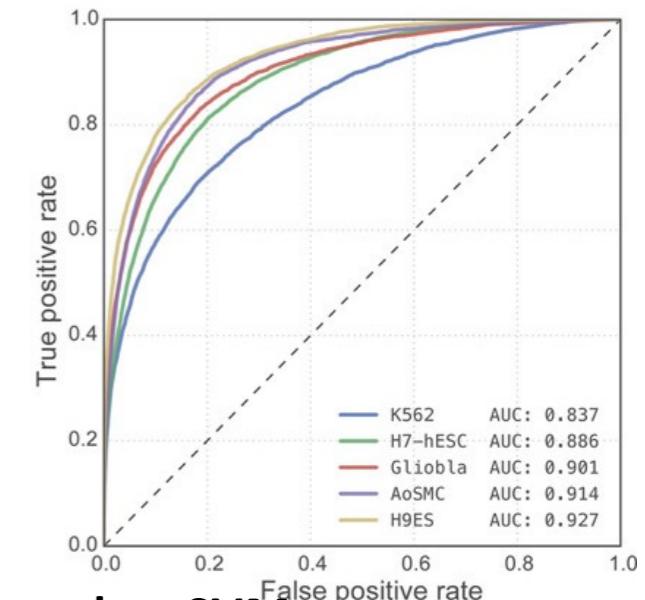
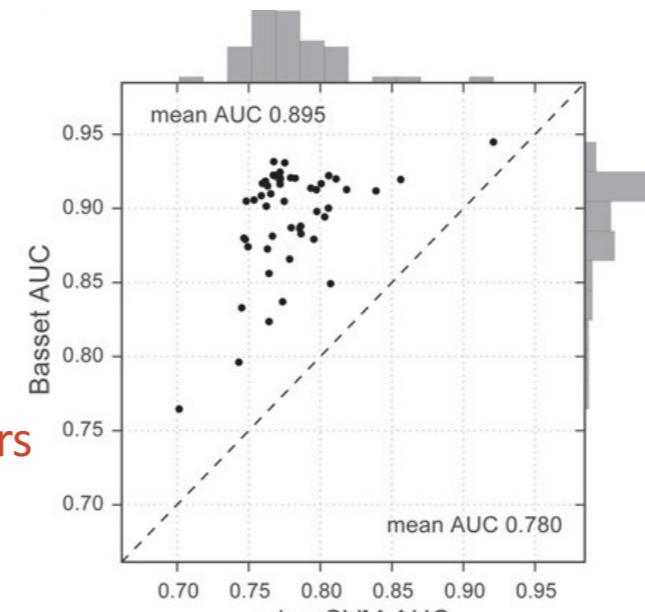
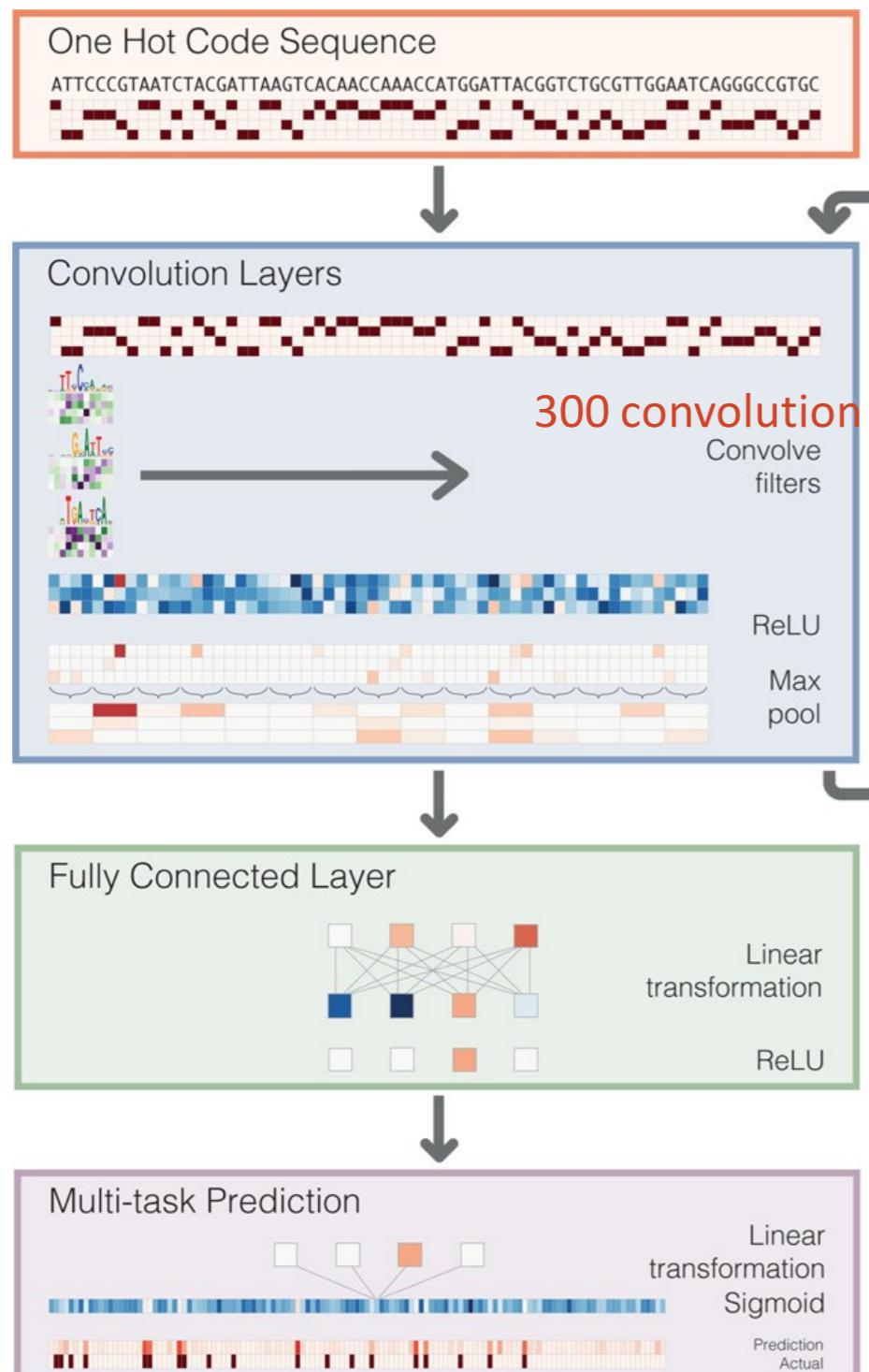
David R. Kelley

Jasper Snoek

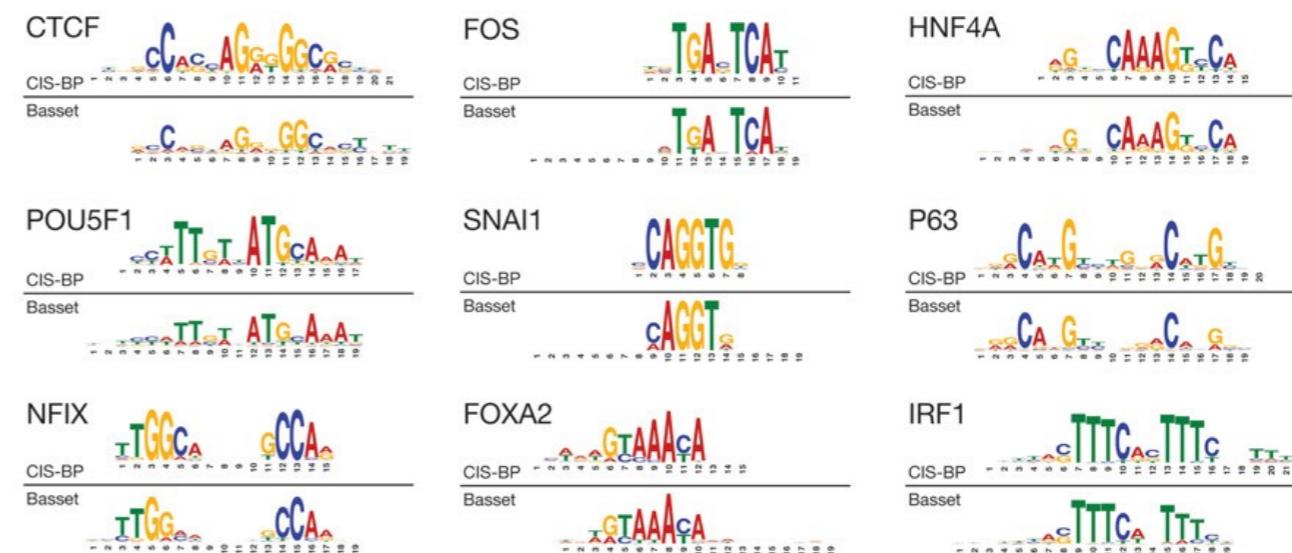
John L. Rinn

Genome Research, March 2016

Basset



CNN-based Basset outperforms gkm-SVM



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

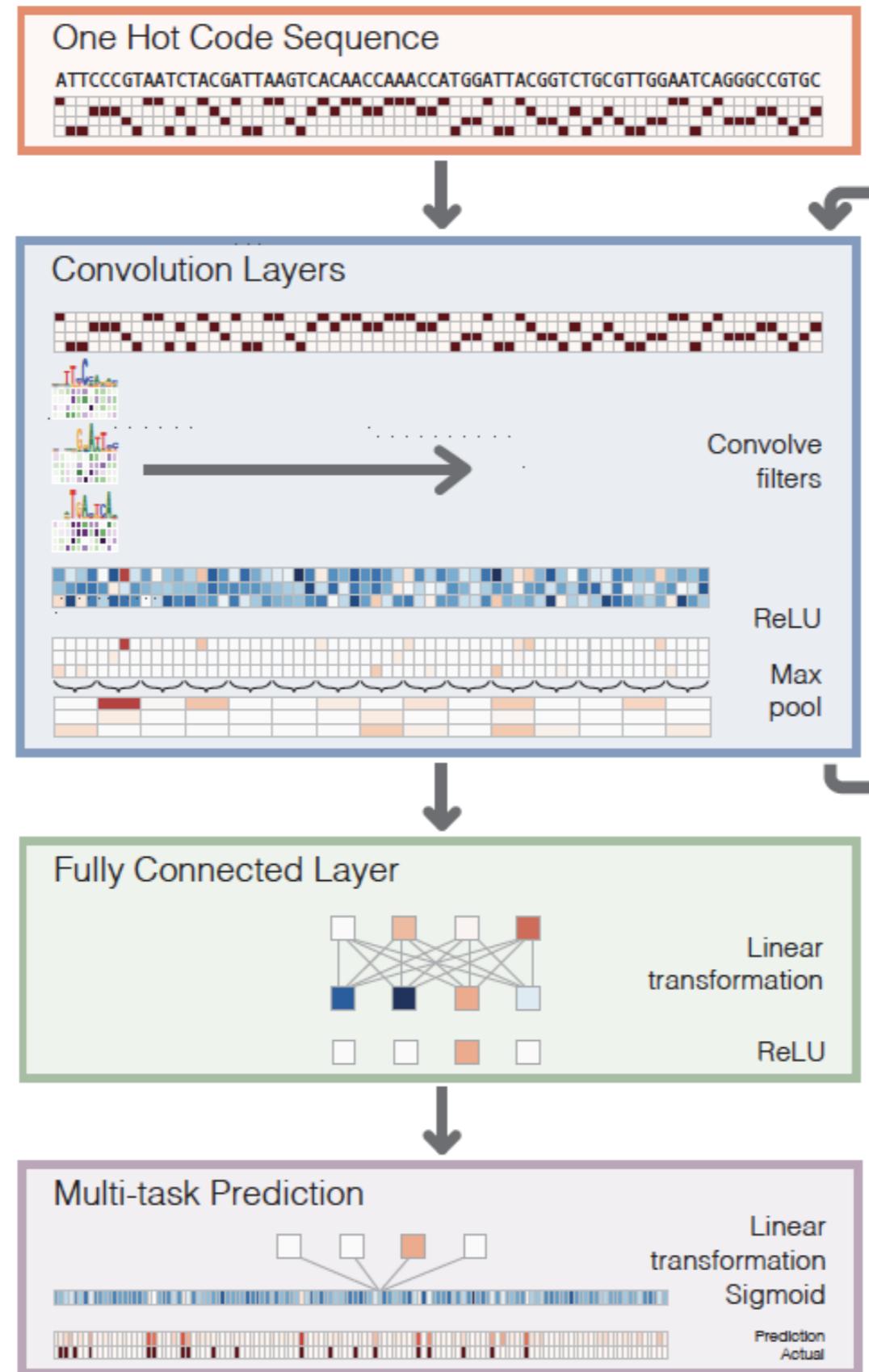
**Simultaneously
predicting DNase sites in
164 cell types**

[Kelley et al., 2016]

Bassett architecture for accessibility prediction

Input:
600 bp

1.9 million
training
examples



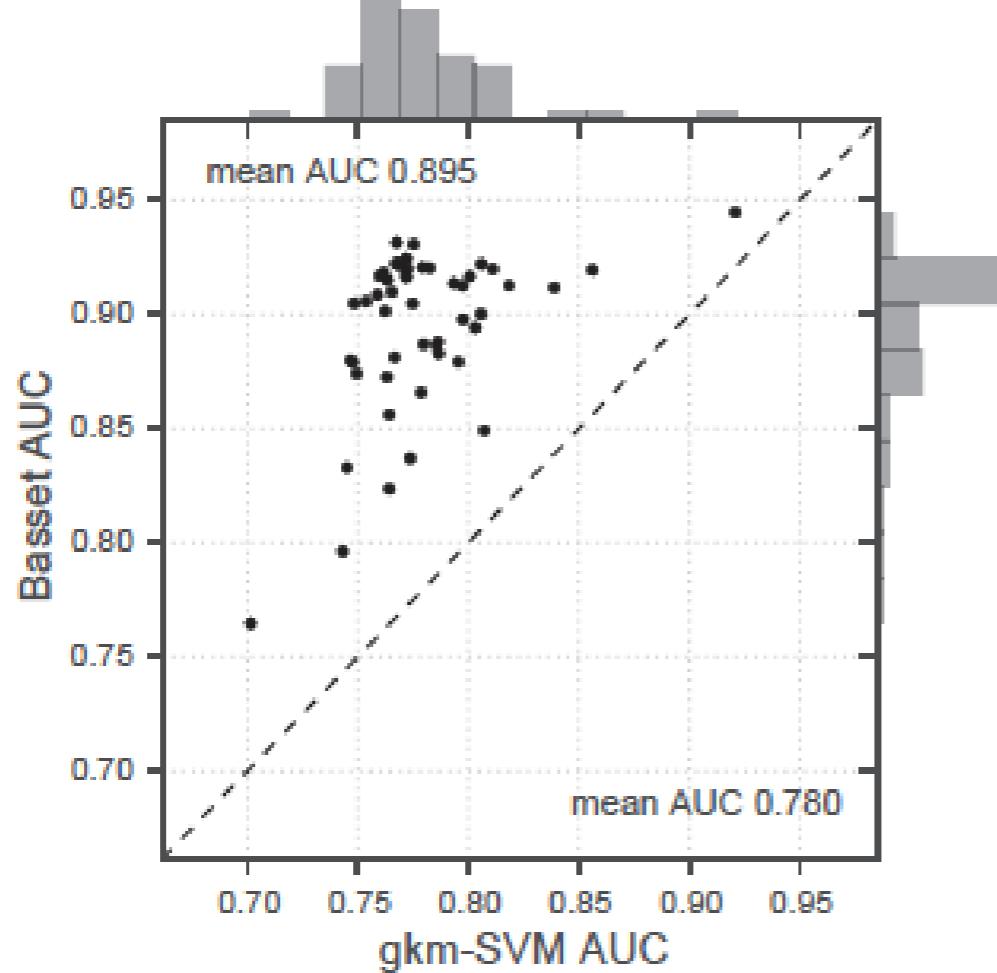
300 filters
3 conv layers
3 FC layers

3 fully connected layers

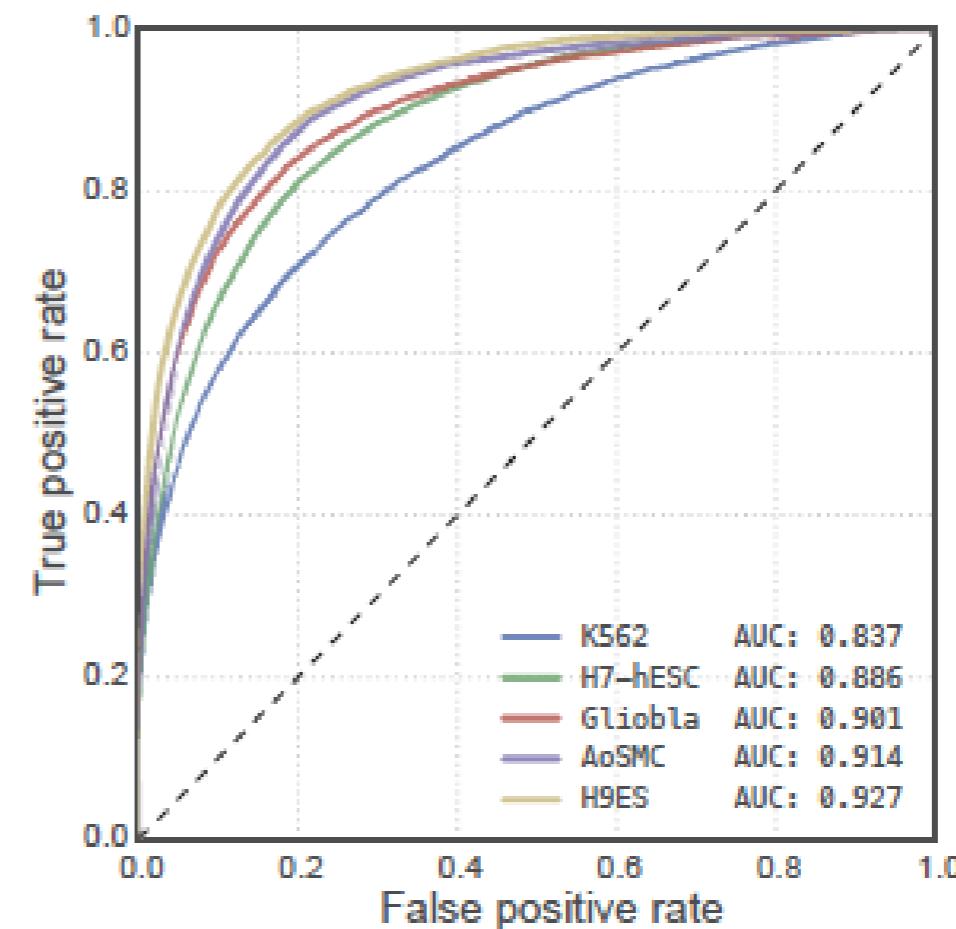
168 outputs
(1 per cell type)

Bassett AUC performance vs. gkm-SVM

B



C

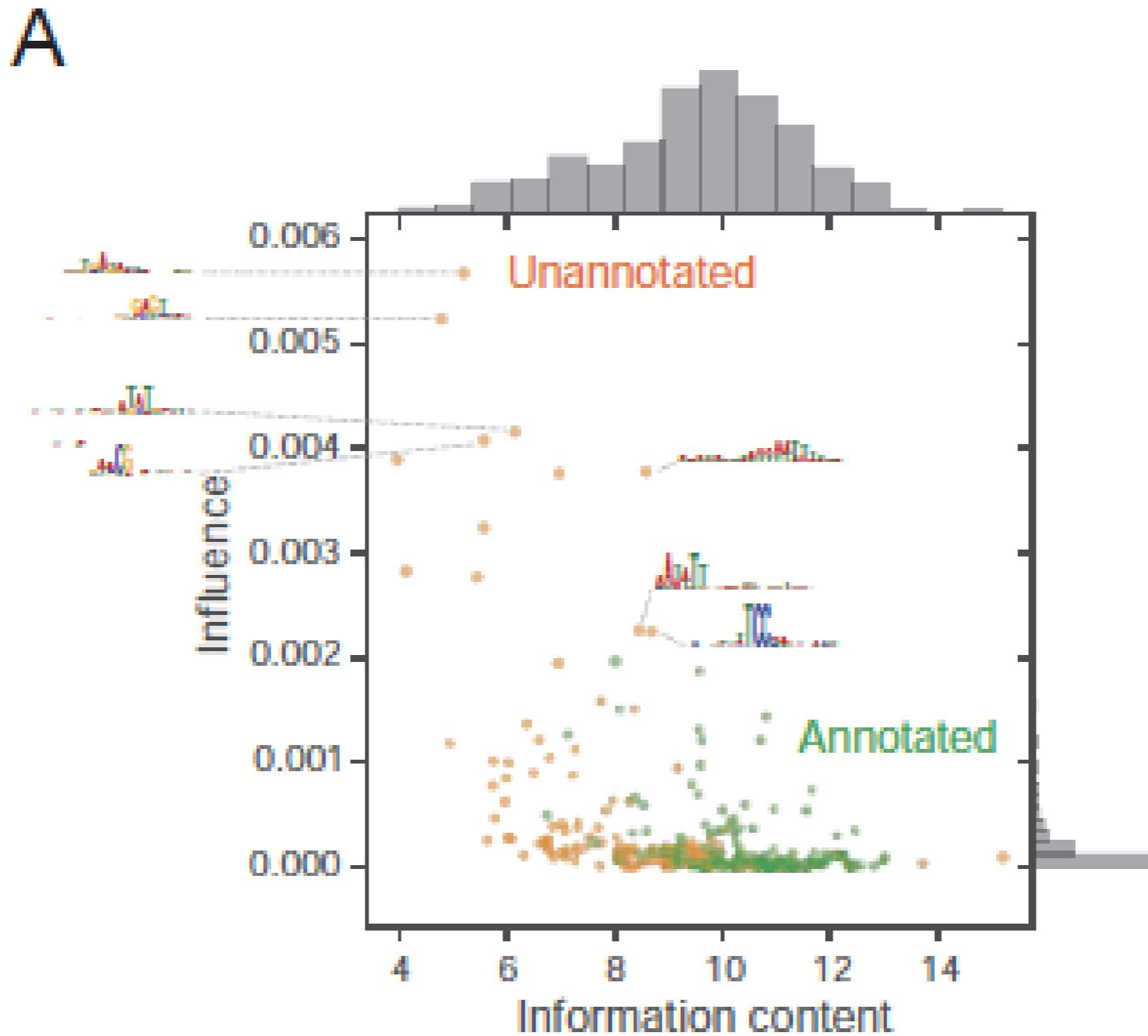


45% of filter derived motifs are found in the CIS-BP database

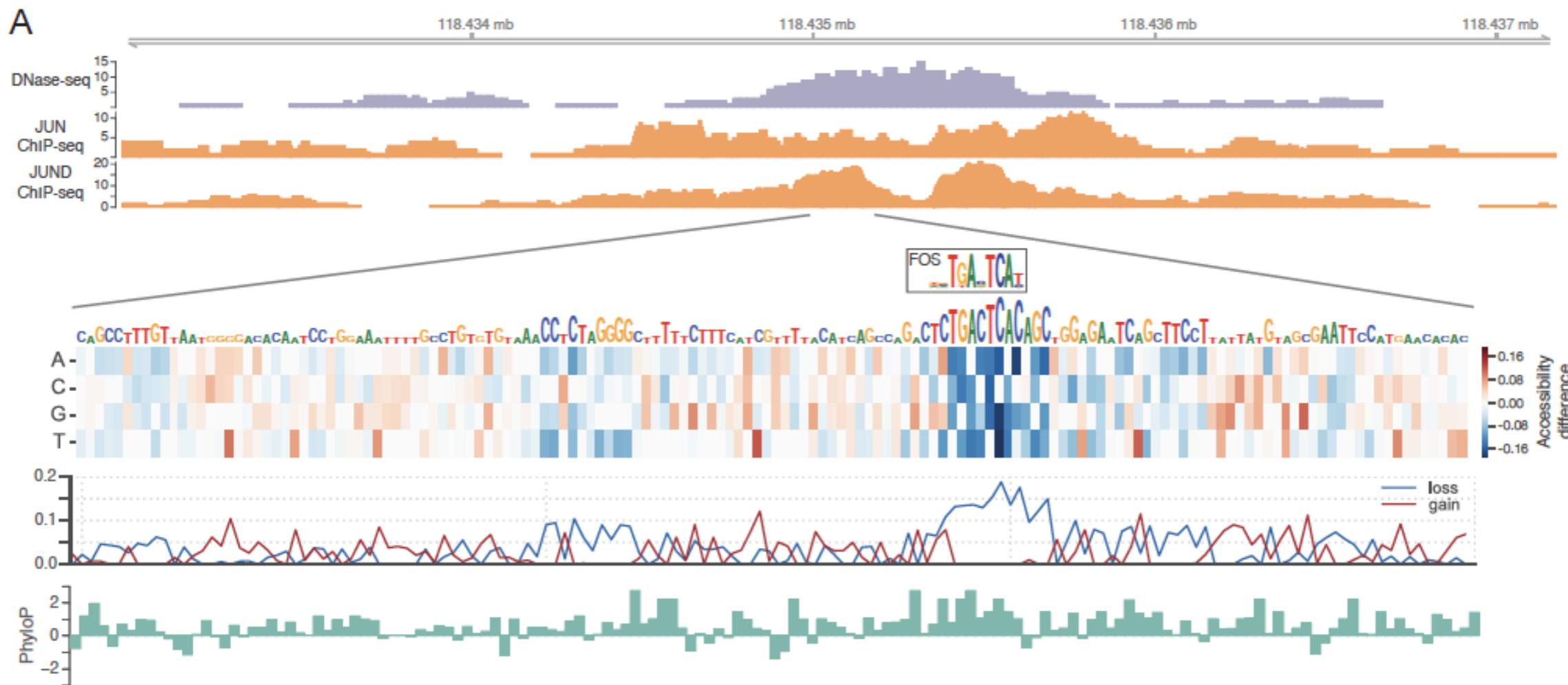


Motifs created by clustering matching input sequences and computing PWM

Motif derived from filters with more information tend to be annotated

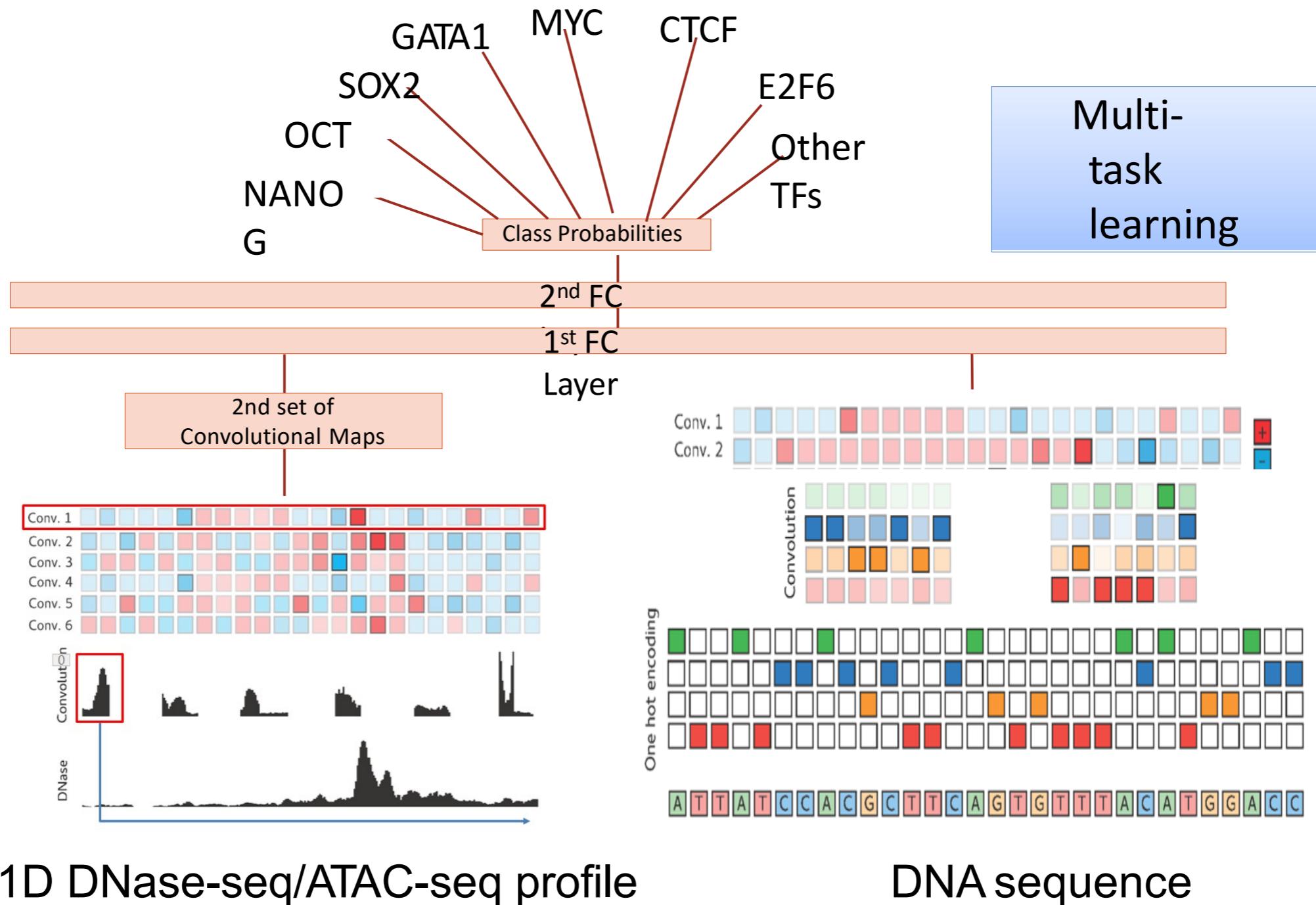


Computational saturation mutagenesis of an AP-I site reveals loss of accessibility



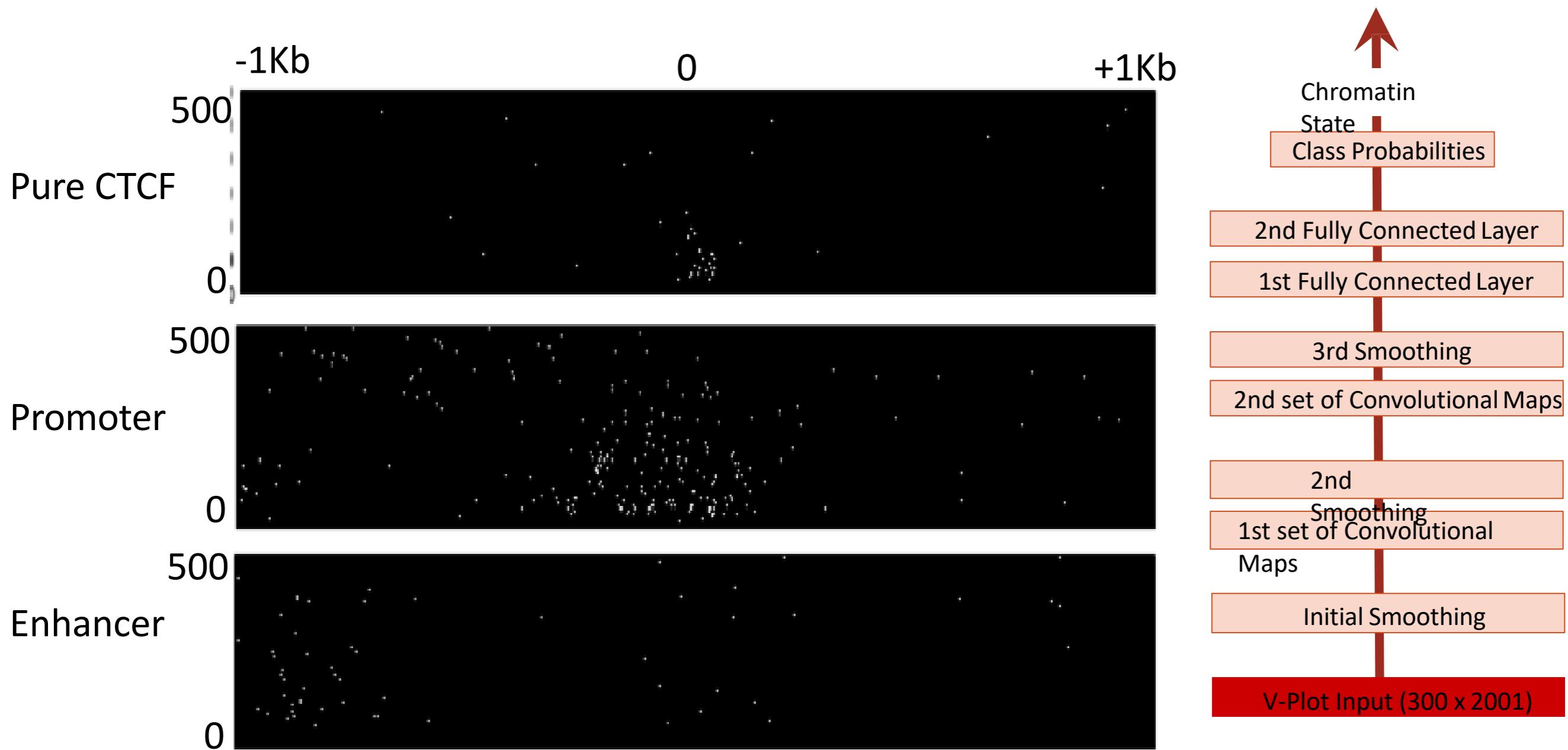
Regulatory Genomics CNNs in Practice: (d) Chromputer

ChromPuter

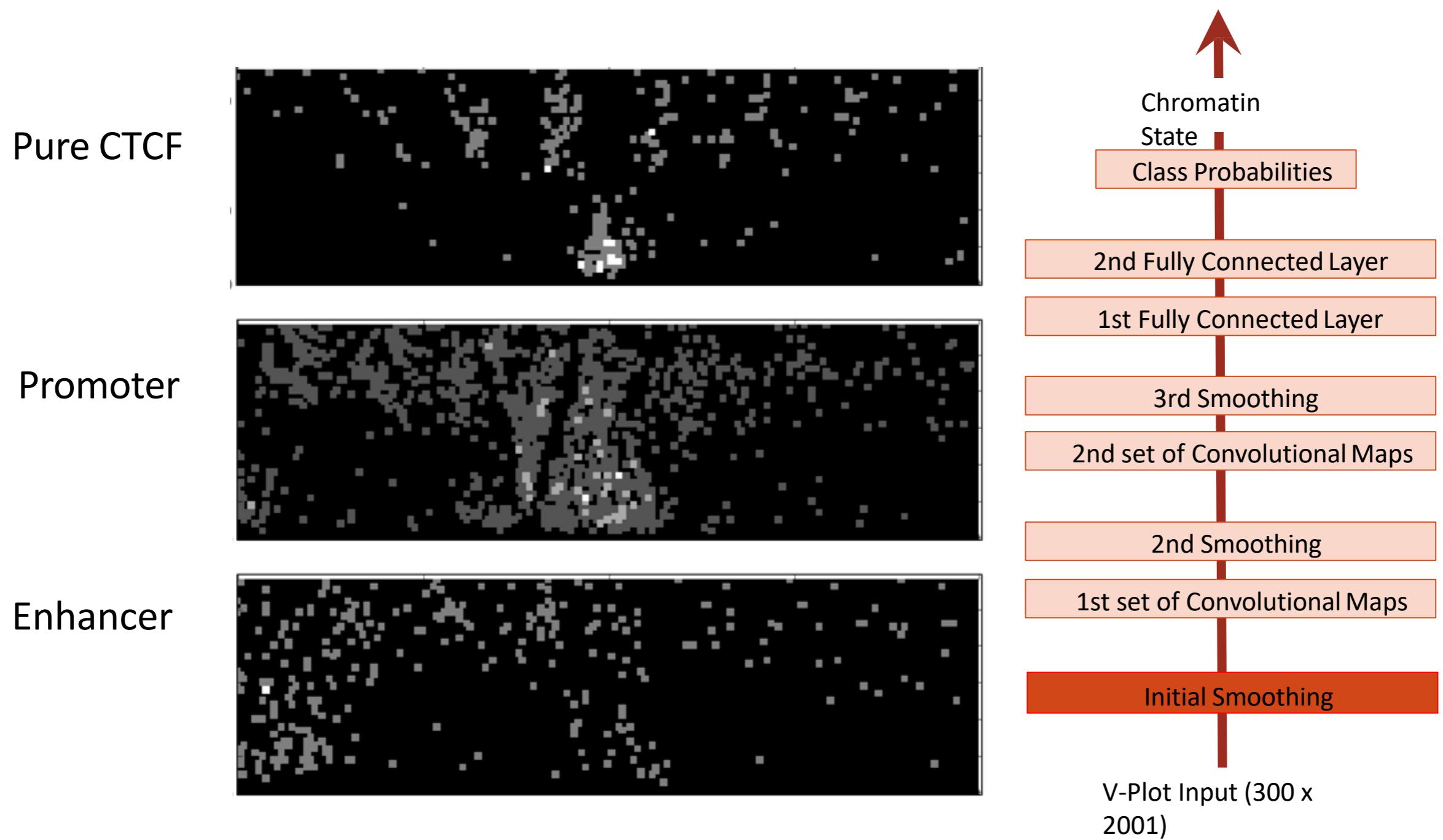


(Anshul Kundaje's group from Stanford)

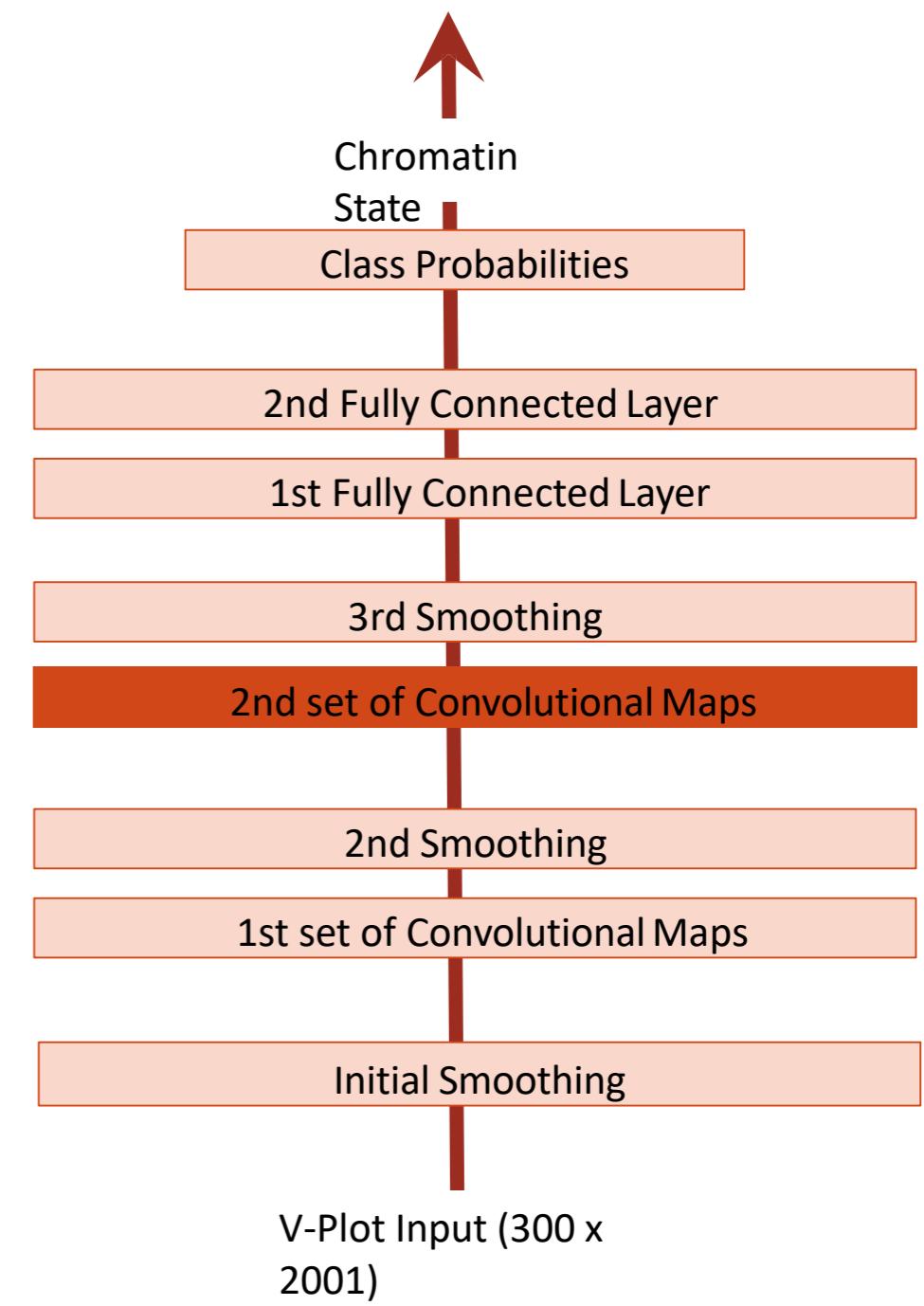
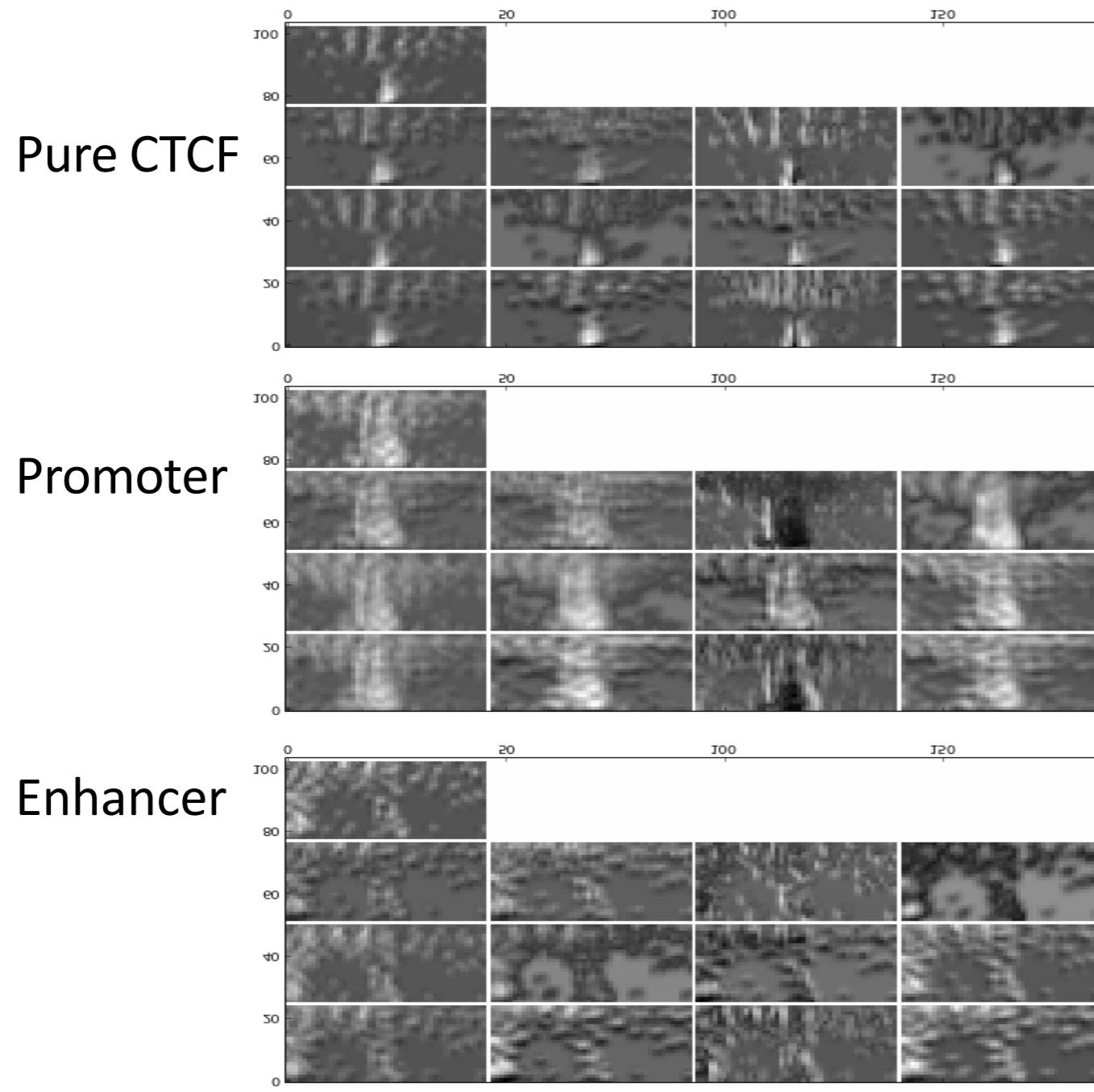
How does a deep conv. neural network transform the raw V-plot input at each layer



After initial pooling (smoothing)

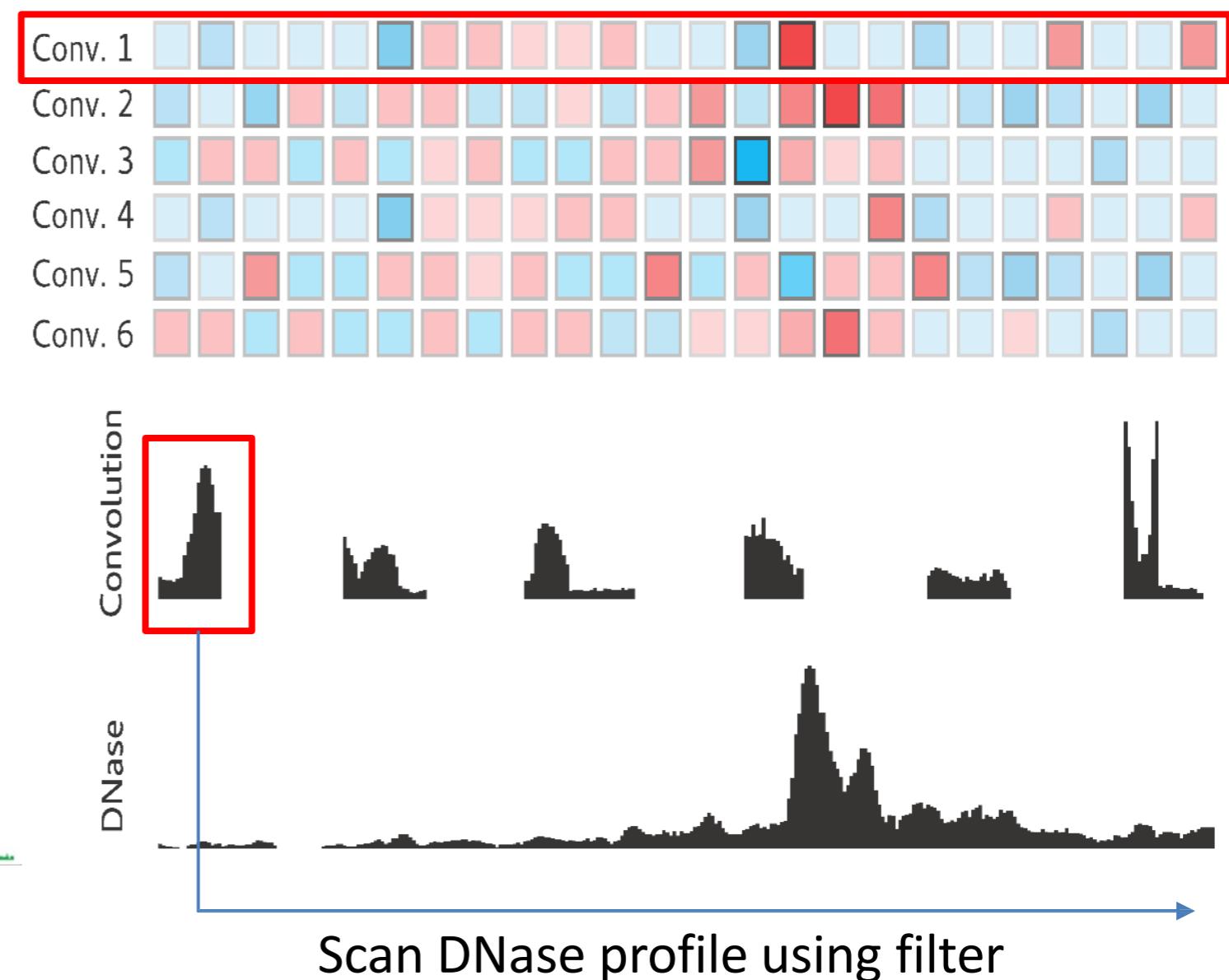
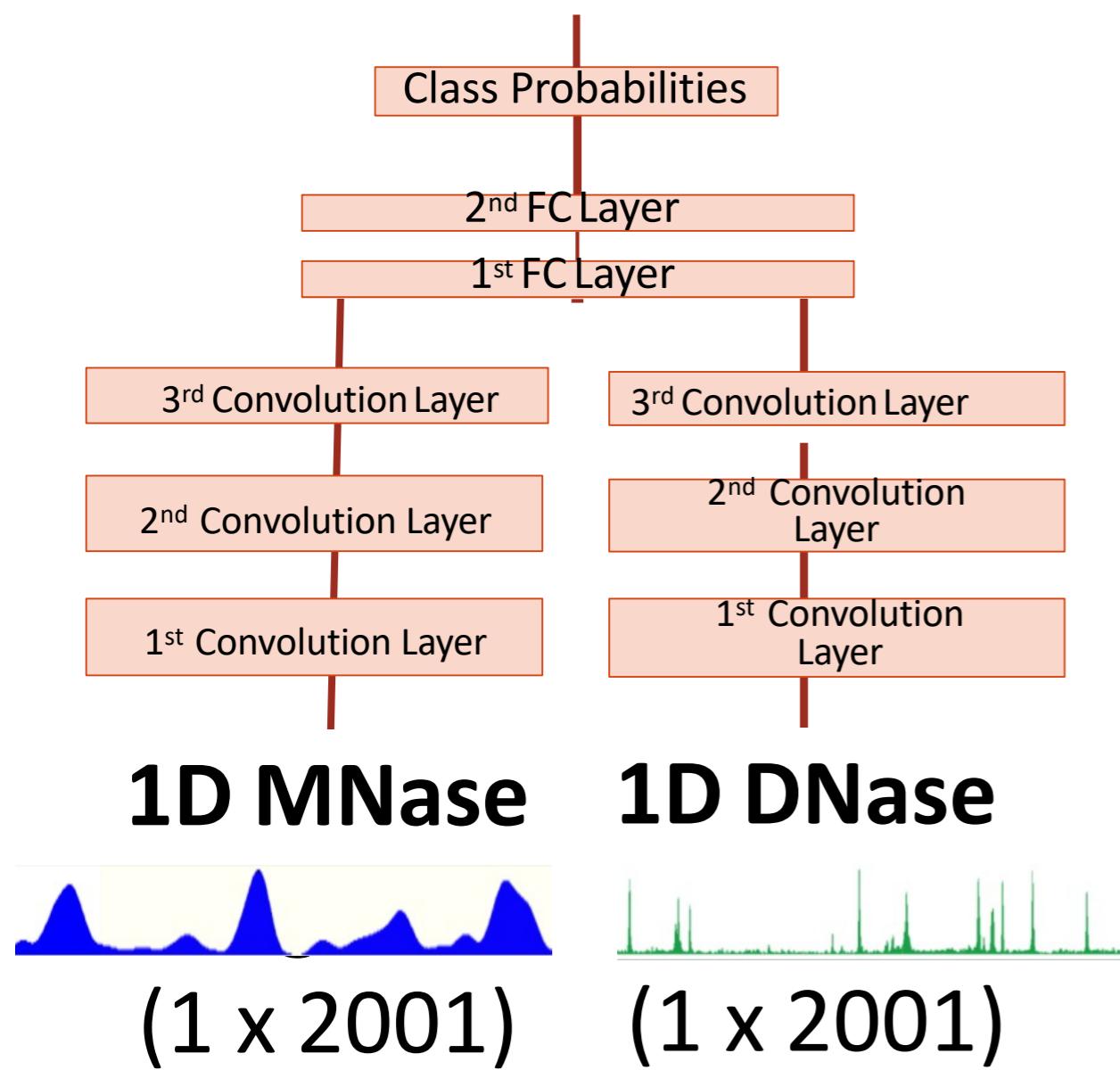


Second set of convolutional maps

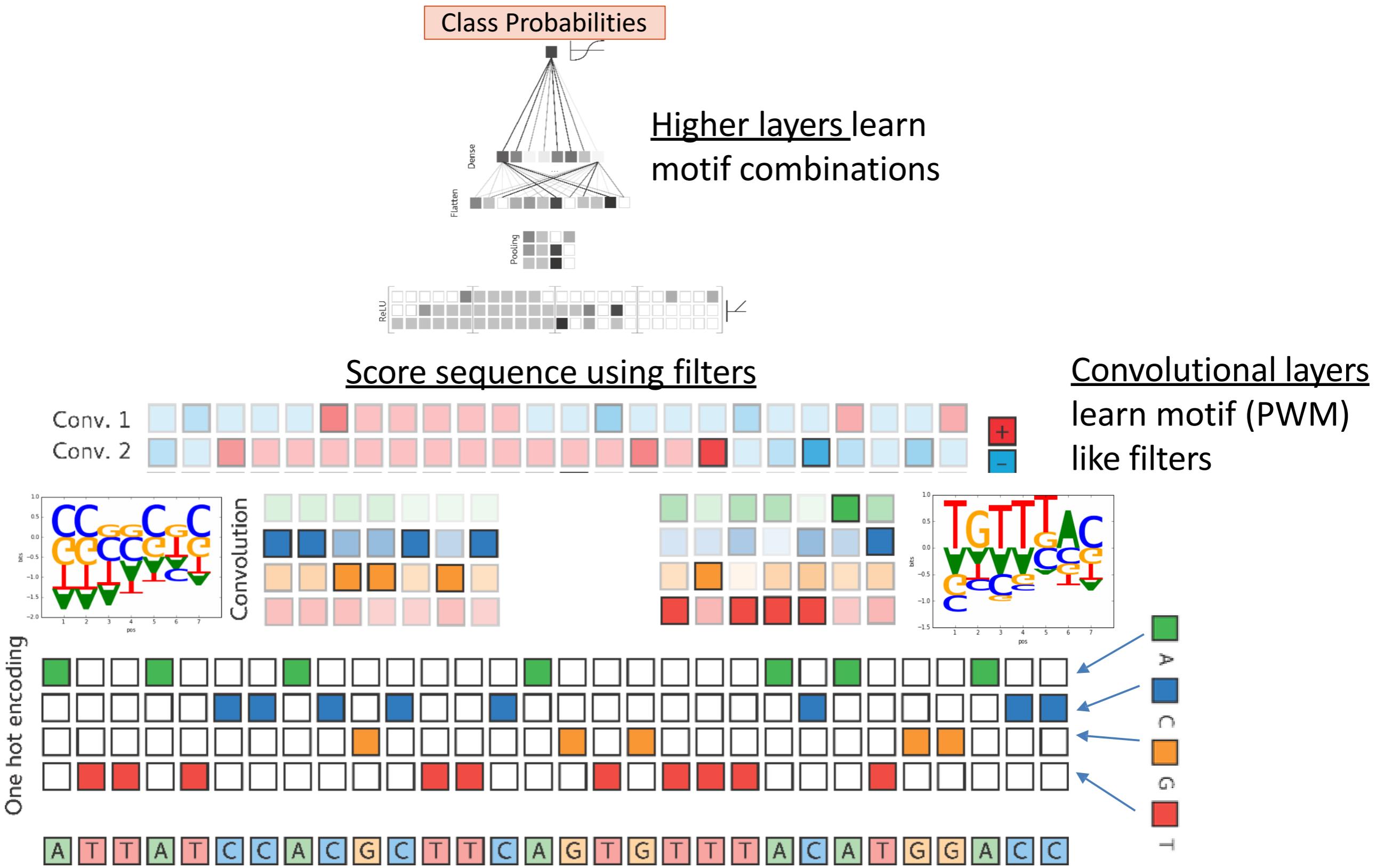


Learning from multiple 1D functional data (e.g. DNase, MNase)

Chromatin State

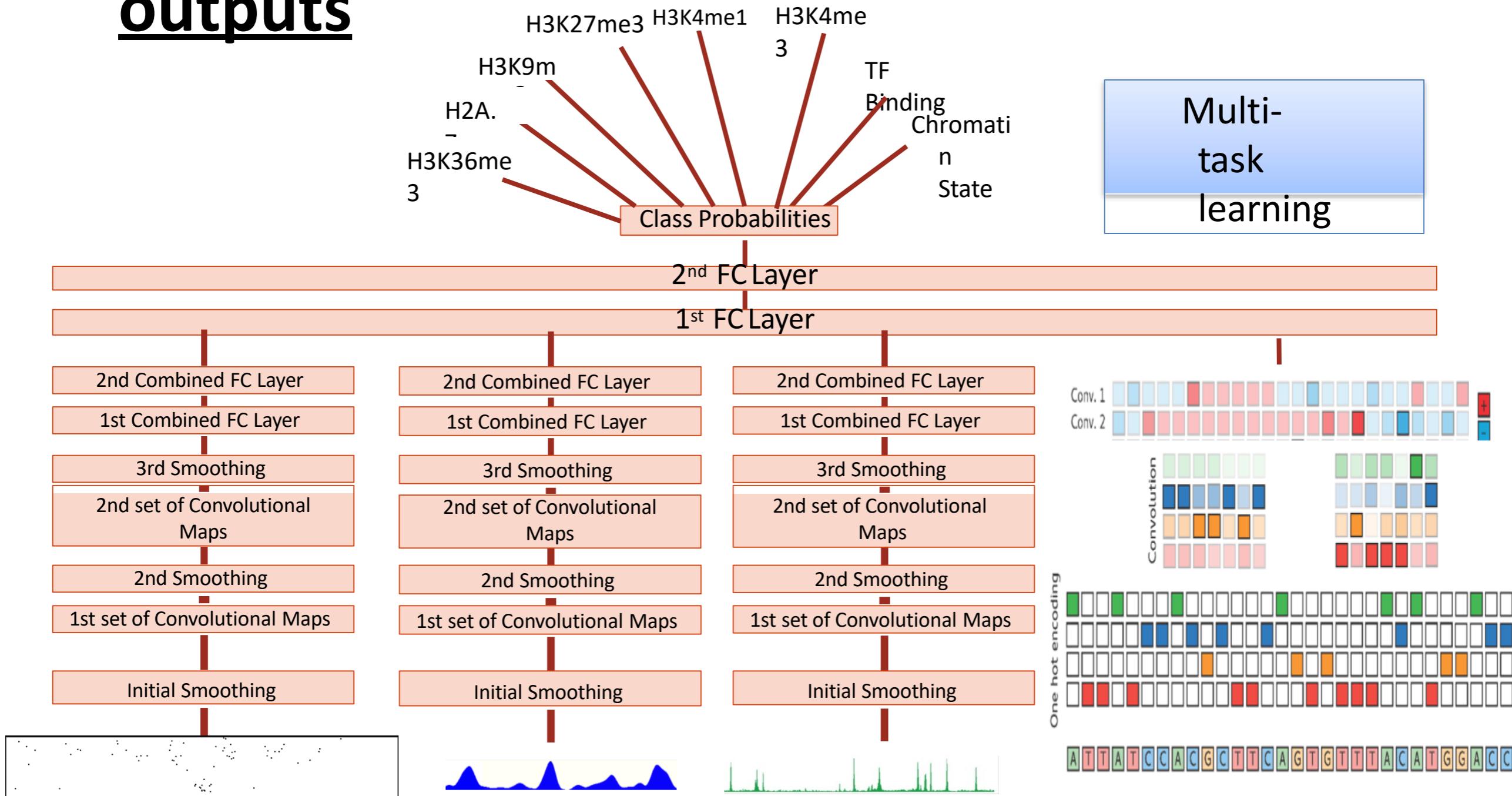


Learning from raw DNA sequence



THE CHROMPUTER

Integrating multiple inputs (1D, 2D signals, sequence) to simultaneously **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%
<u>Random Forest</u> : 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

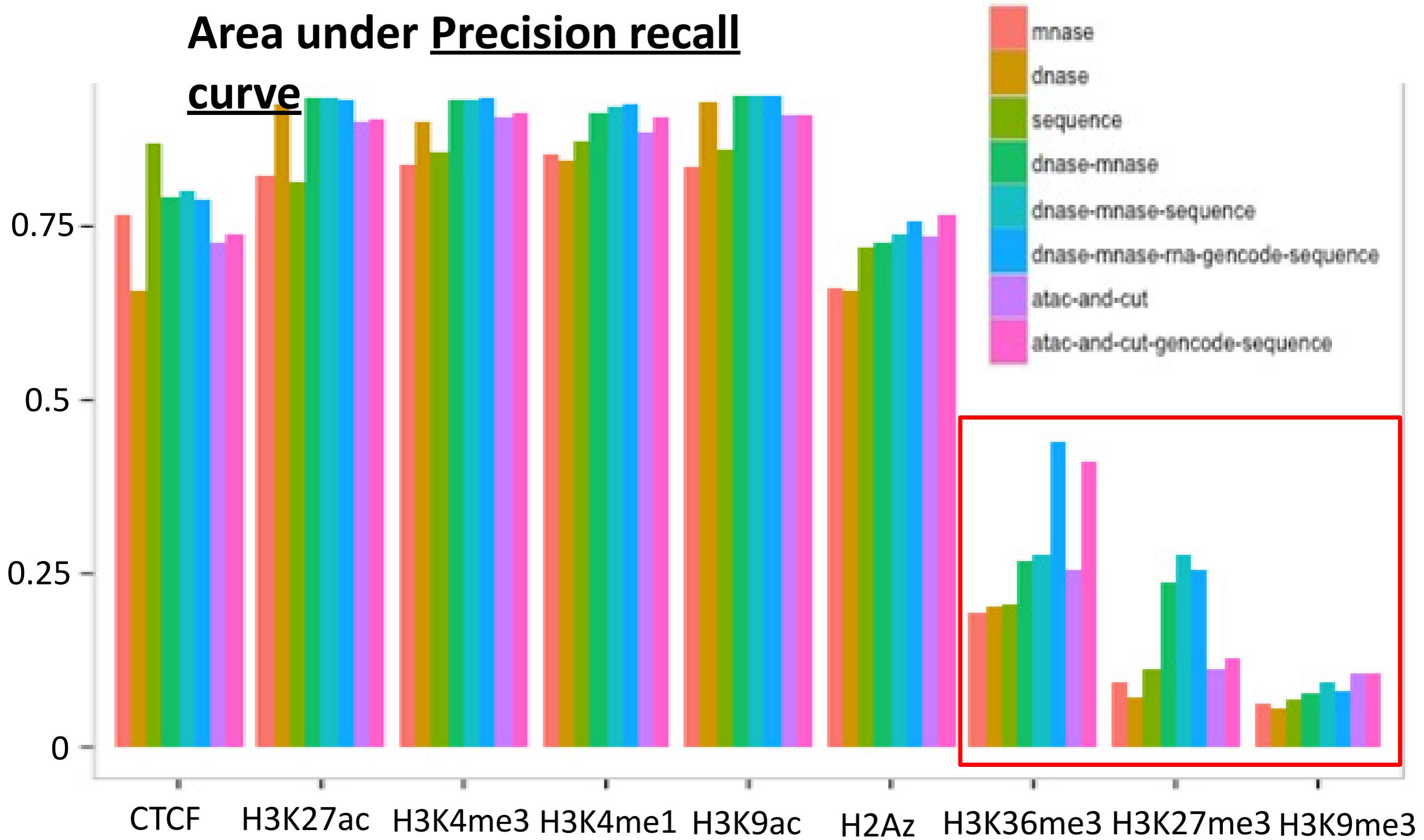
- Learn model on **DNase and MNase only**
- **Learn on GM12878, predict on K562 (and vice versa)**
- **Requires local normalization** to make signal comparable

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

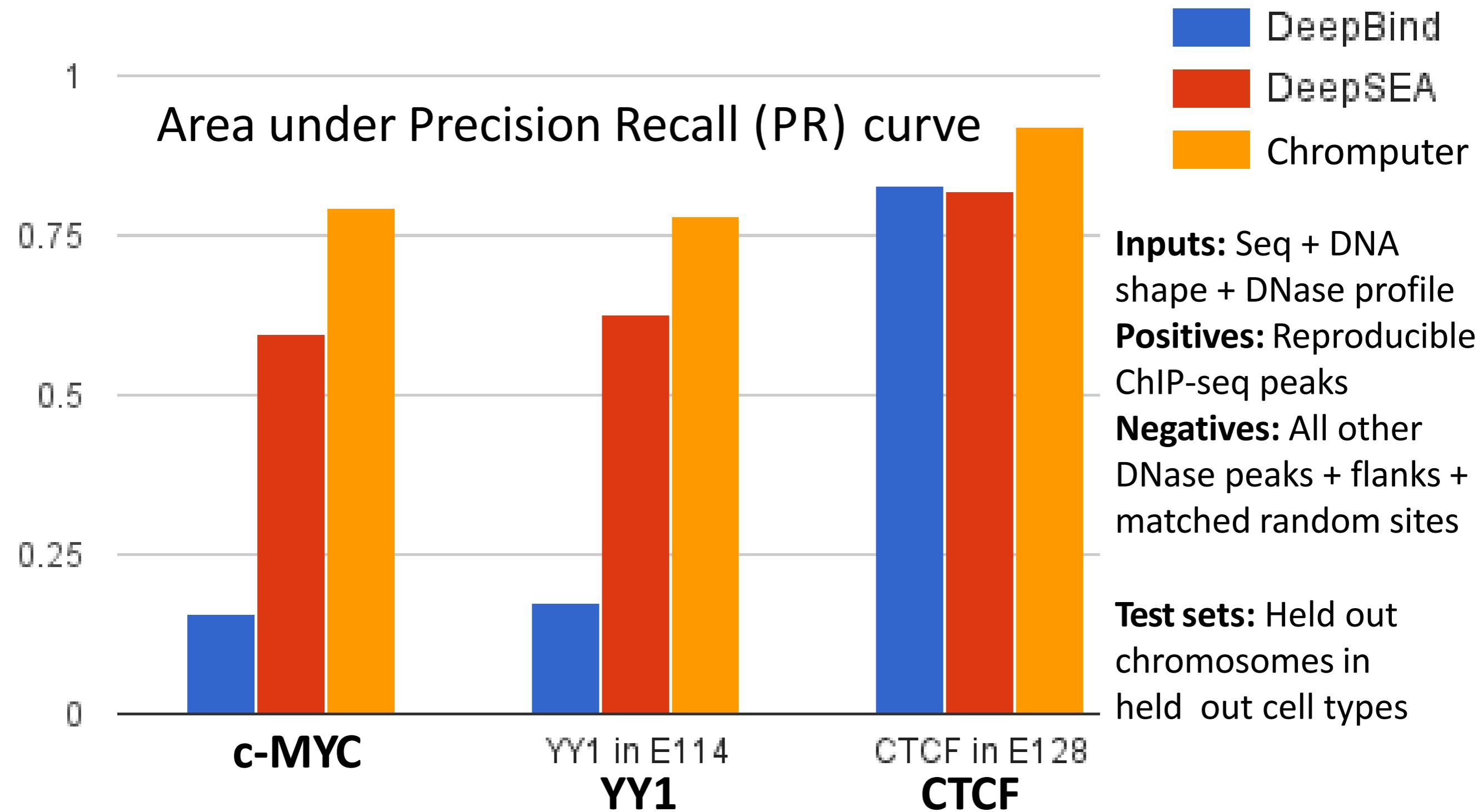
Predicting individual histone marks from ATAC/DNase/MNase/Sequence

Area under Precision recall

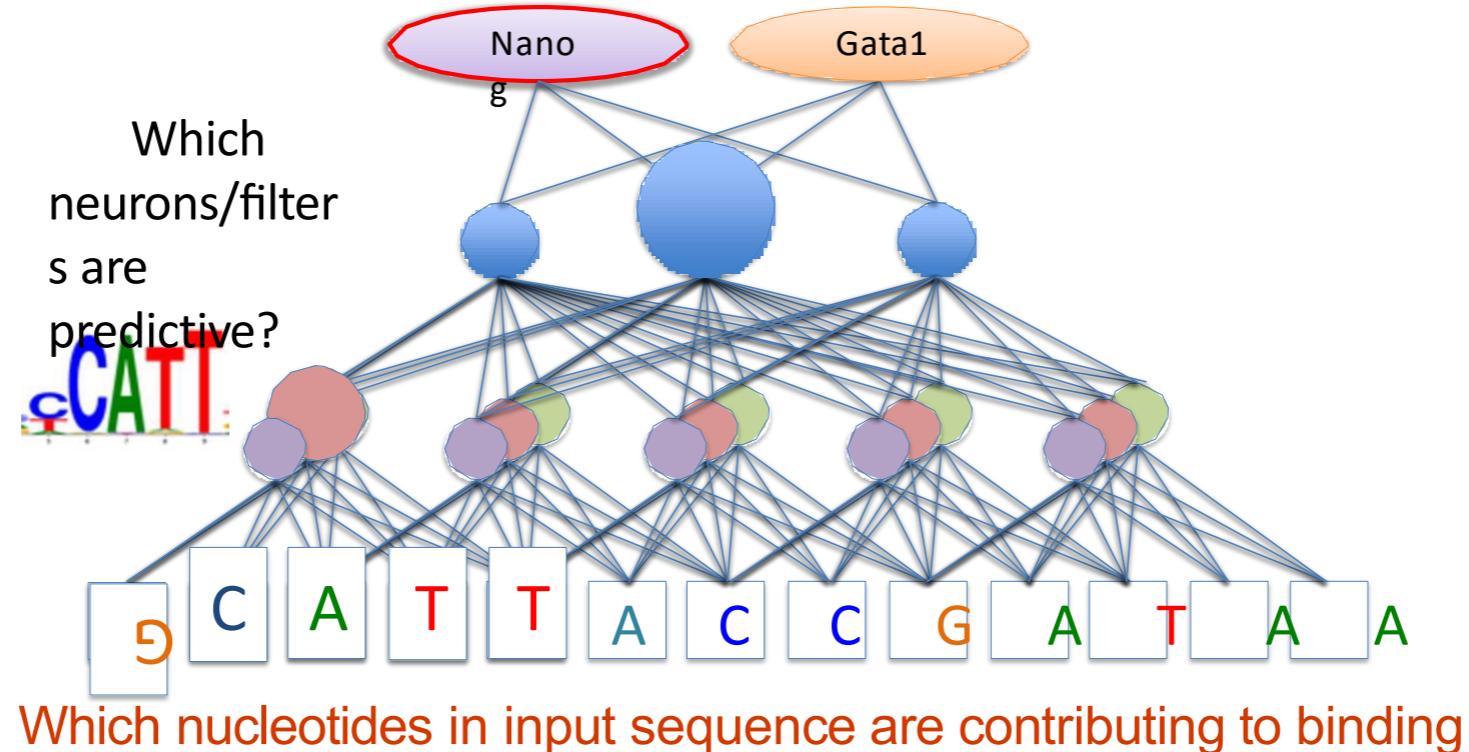
curve



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



DeepLift reveals feature importance at the input layer



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

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 \Leftrightarrow DNA letters. Patches/filters \Leftrightarrow Motifs. Higher \Leftrightarrow combinations
- Learning convolutional filters \Leftrightarrow Motif discovery. Applying them \Leftrightarrow 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