Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford
Lecture 4

February 14, 2019

Deep Networks, Convolutional Networks, and Recurrent Networks

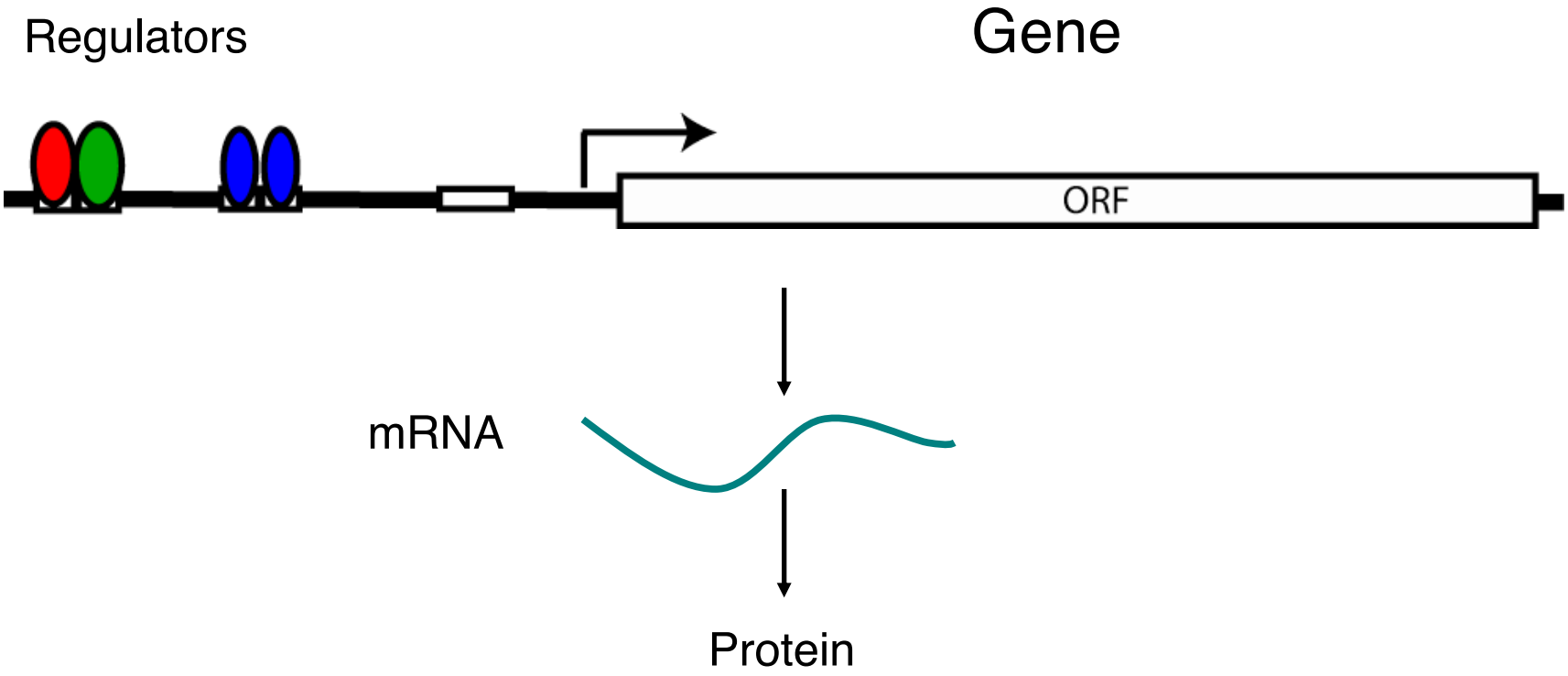
http://mit6874.github.io

What's on tap today!

- Transcription factor function
- Expectation-maximization (EM) methods for TF binding and motif discovery
- Deep Learning methods for TF binding and motif discovery

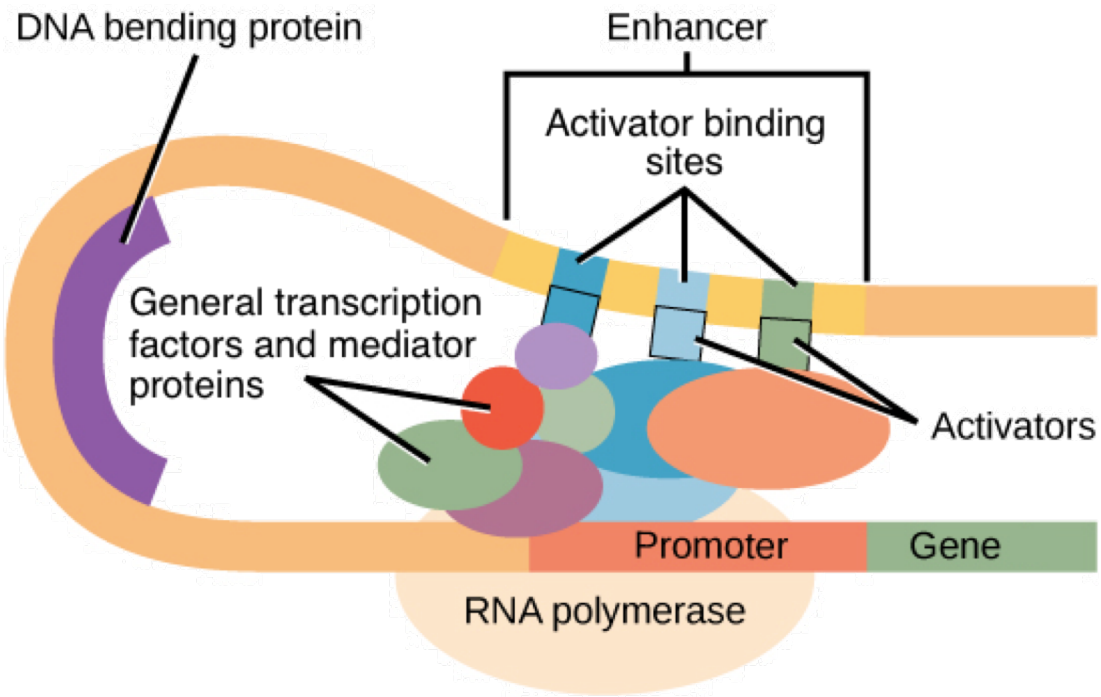
Transcription factors implement genomic regulation

Gene Regulation: DNA > RNA > Protein

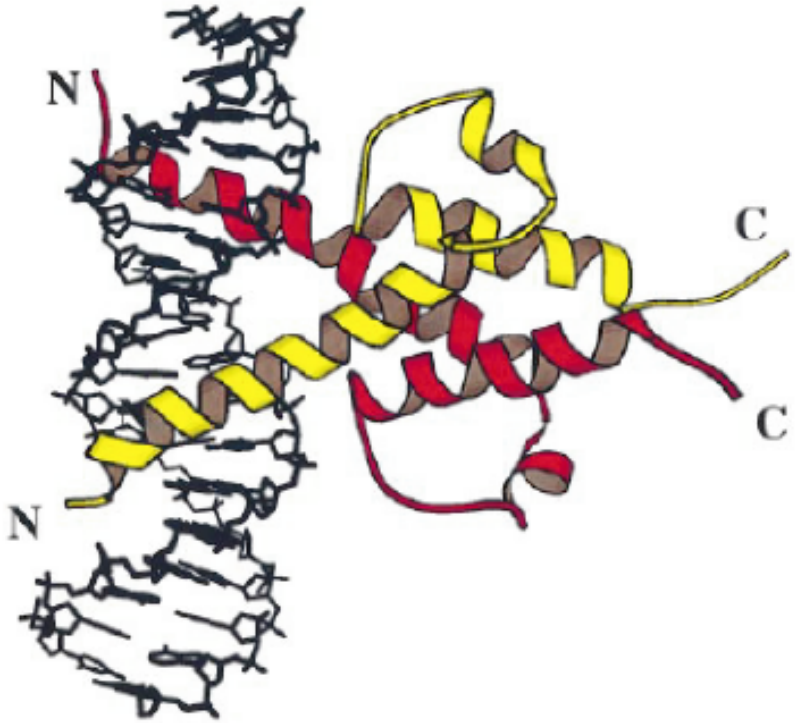


What are the gene regulators that control gene expression?
At what genes do these regulators operate?

DNA-protein binding is essential to cellular function

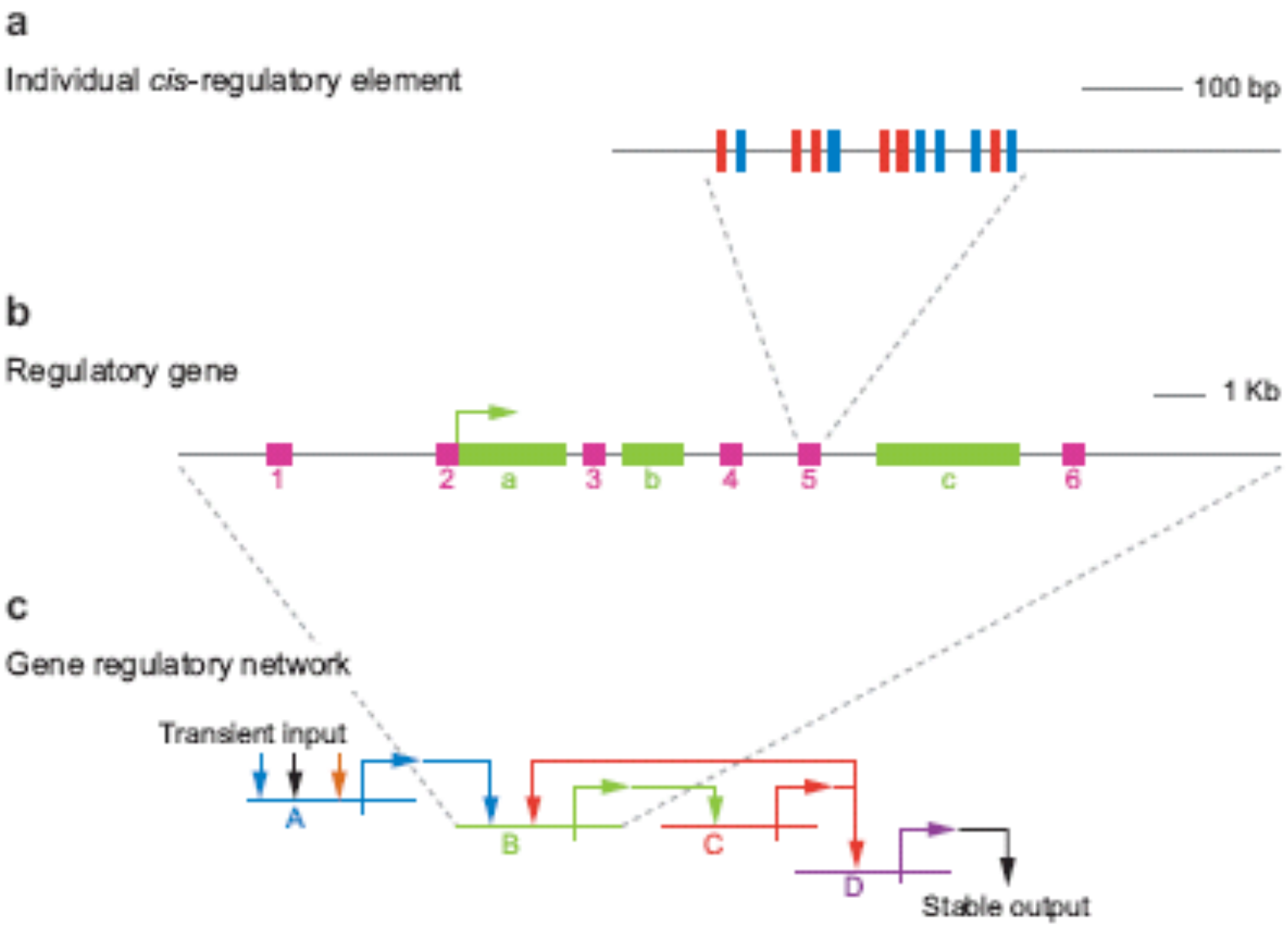


Transcription factors bind specific sequences

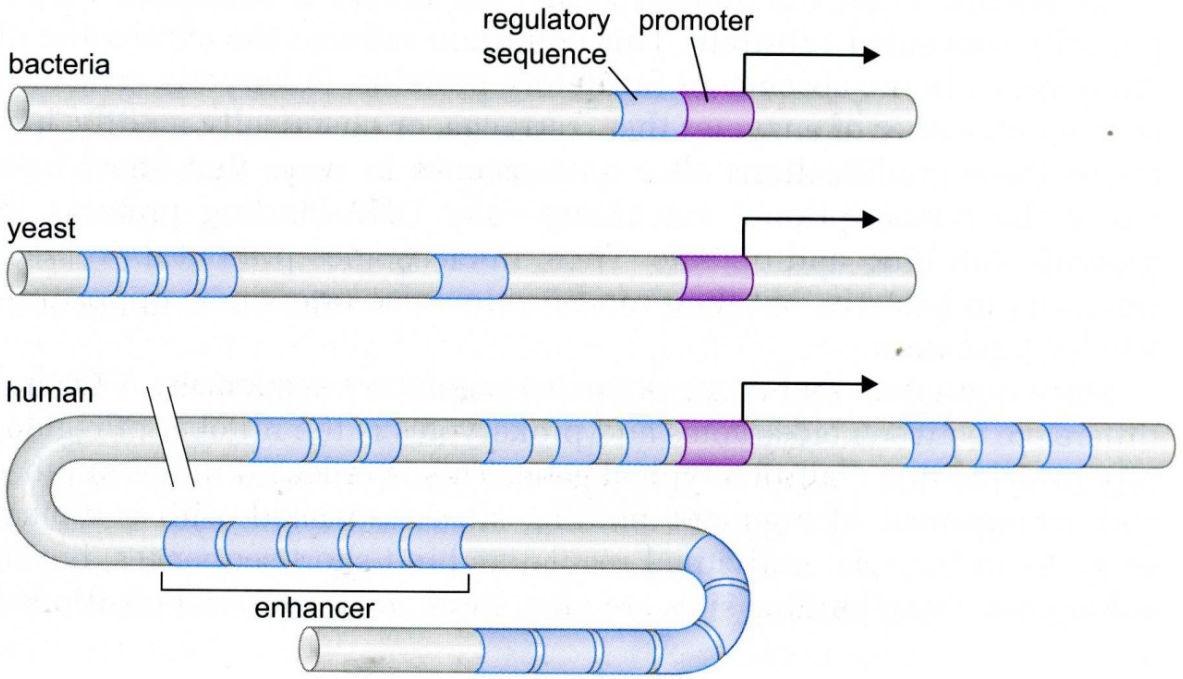


Protein molecules that bind to specific DNA sequences and act as molecular switches to turn genes on or off.

Humans have ~2000 transcription factors.



Combinatorial control lies at the heart of the complexity and diversity of eukaryotes

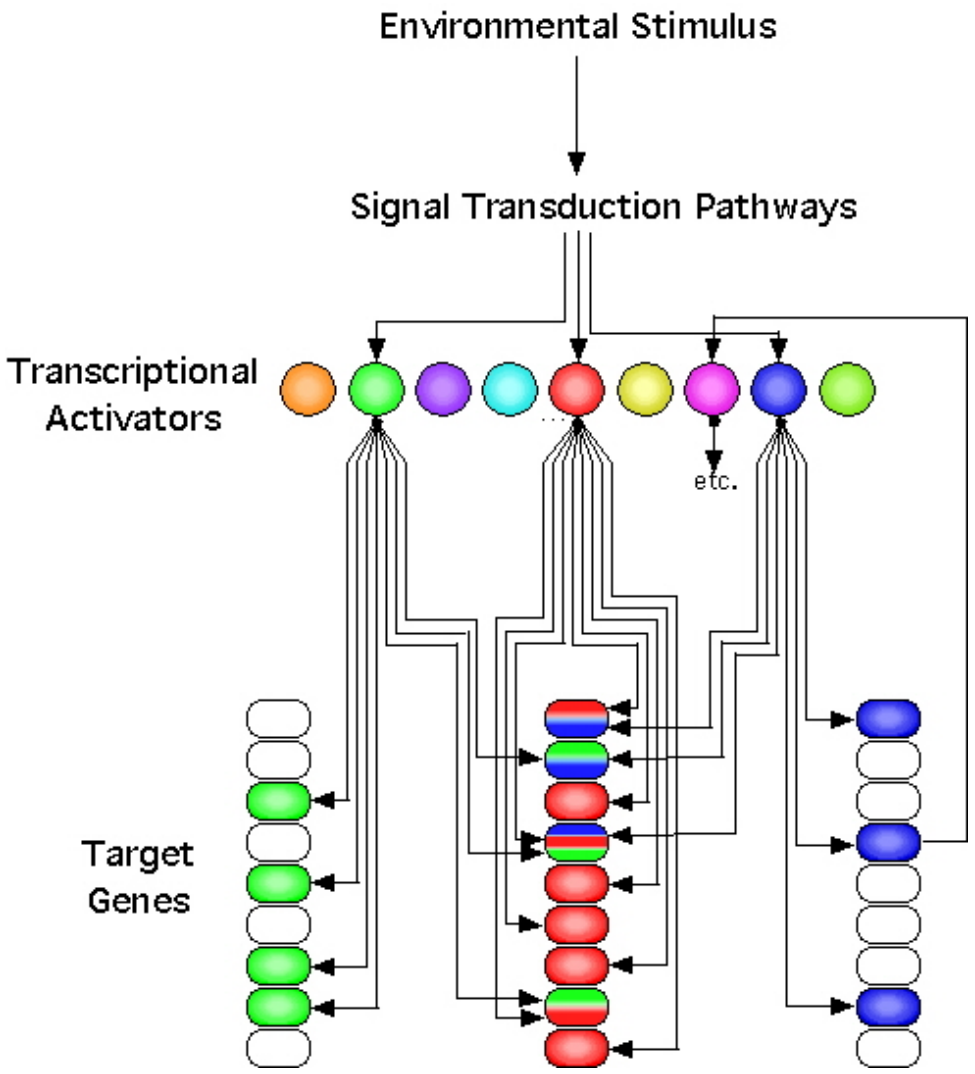


(Molecular biology of the gene, 6ed)

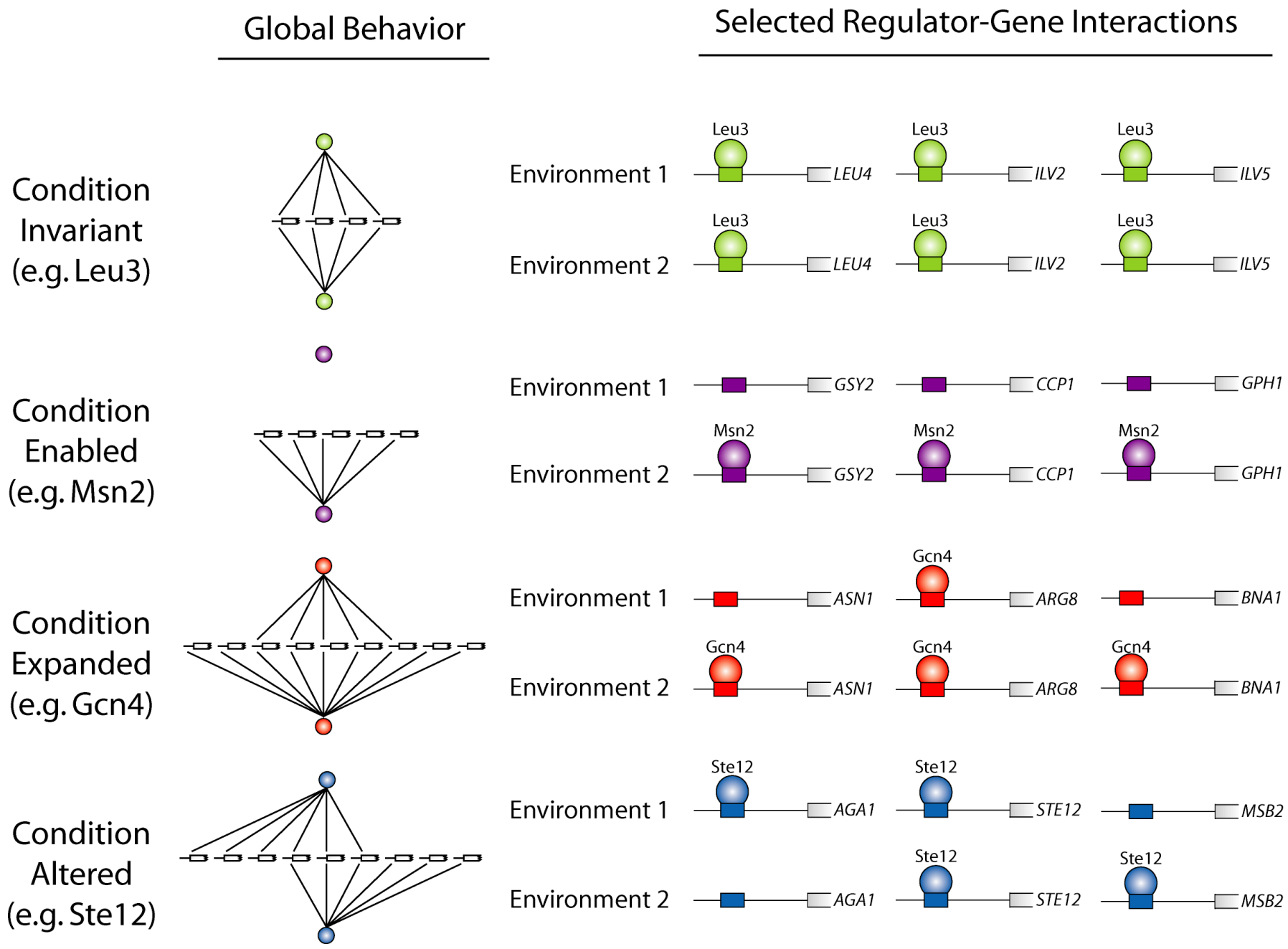
Why Map Transcriptional Regulatory Networks?

Transcriptional regulatory network information will:

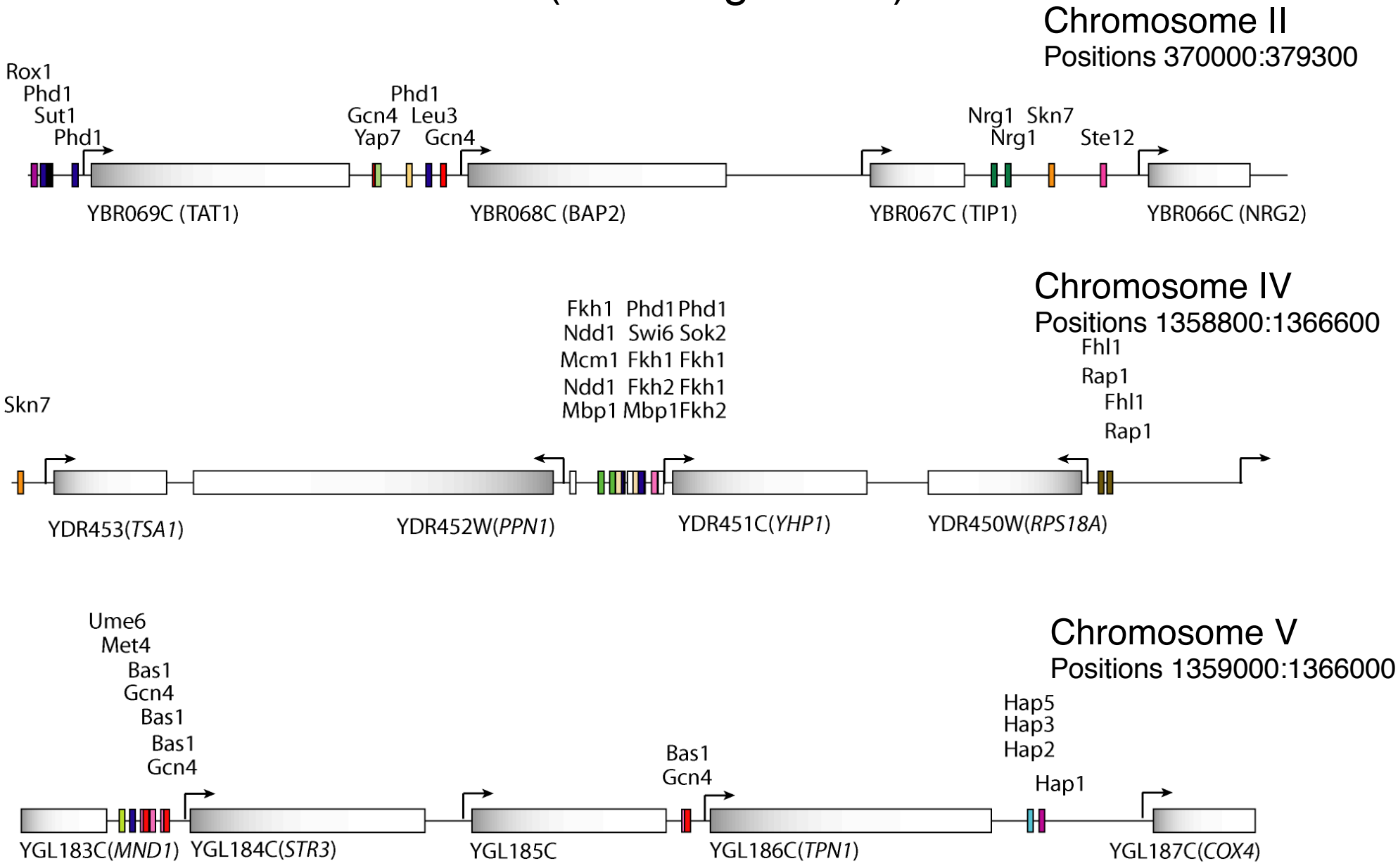
- reveal how cellular processes are connected and coordinated
- suggest new strategies to manipulate phenotypes and combat disease



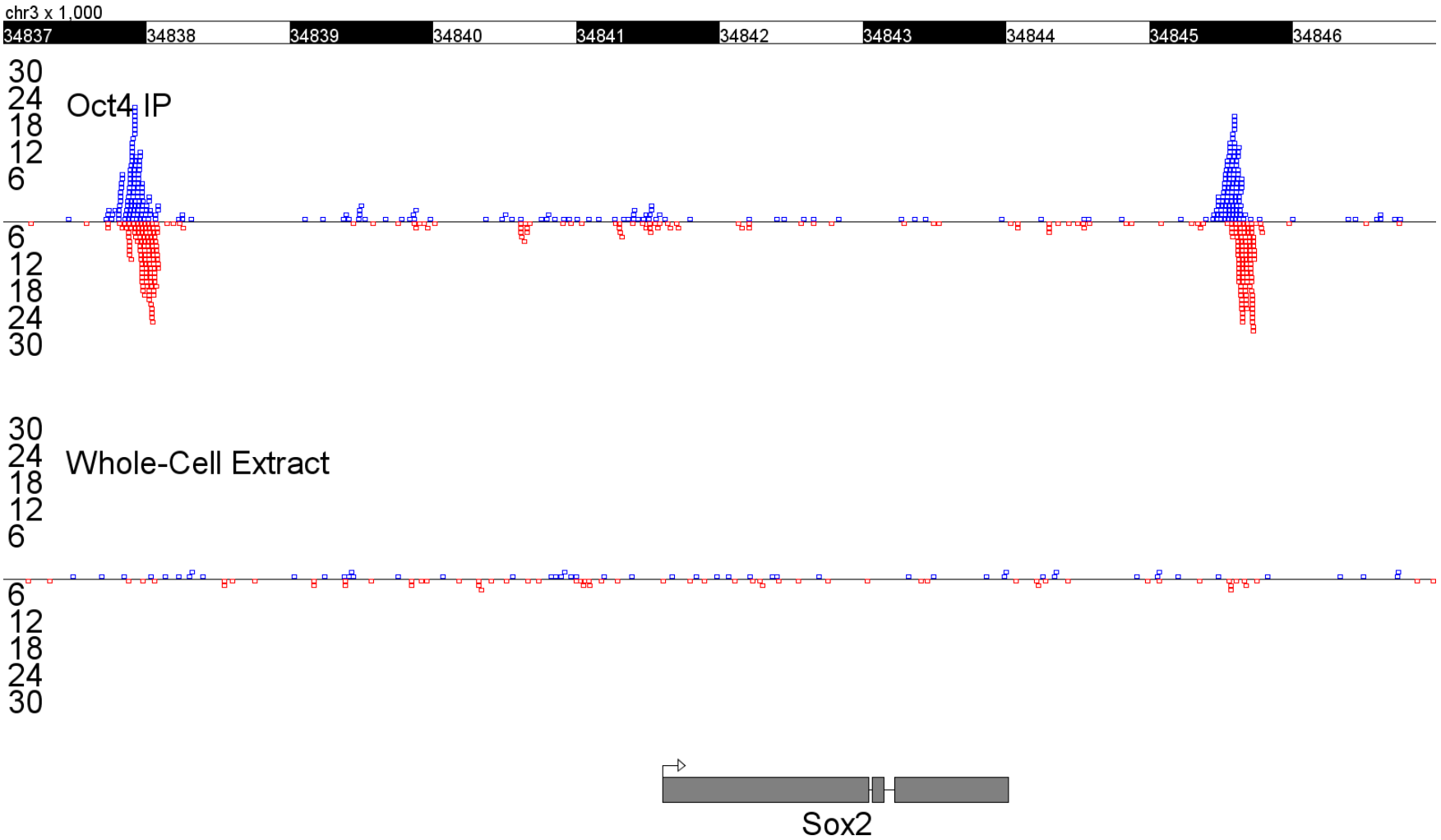
Environment-Specific Regulator Behaviors



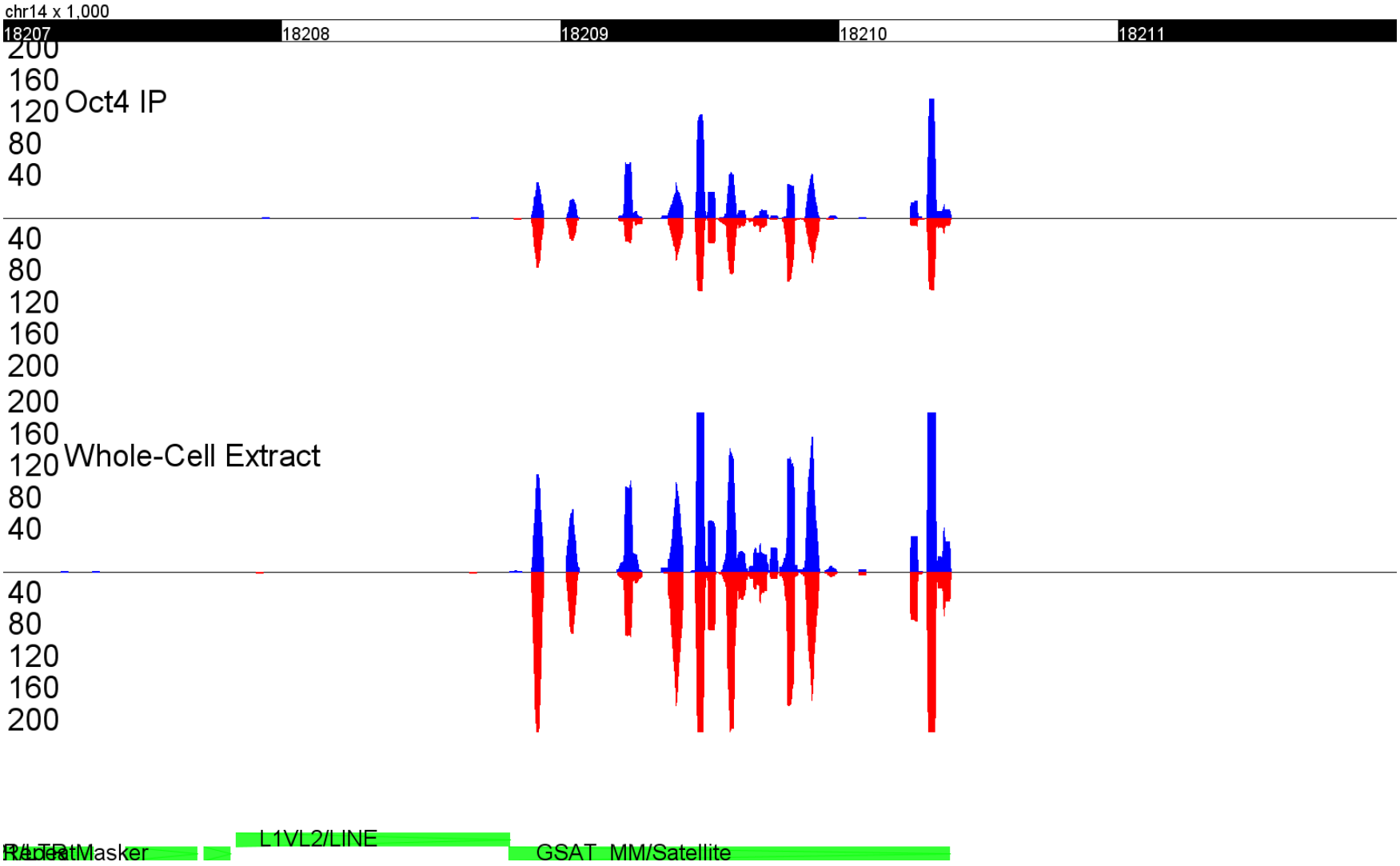
Sample of the Yeast Draft Transcriptional Regulatory Code (~150 regulators)



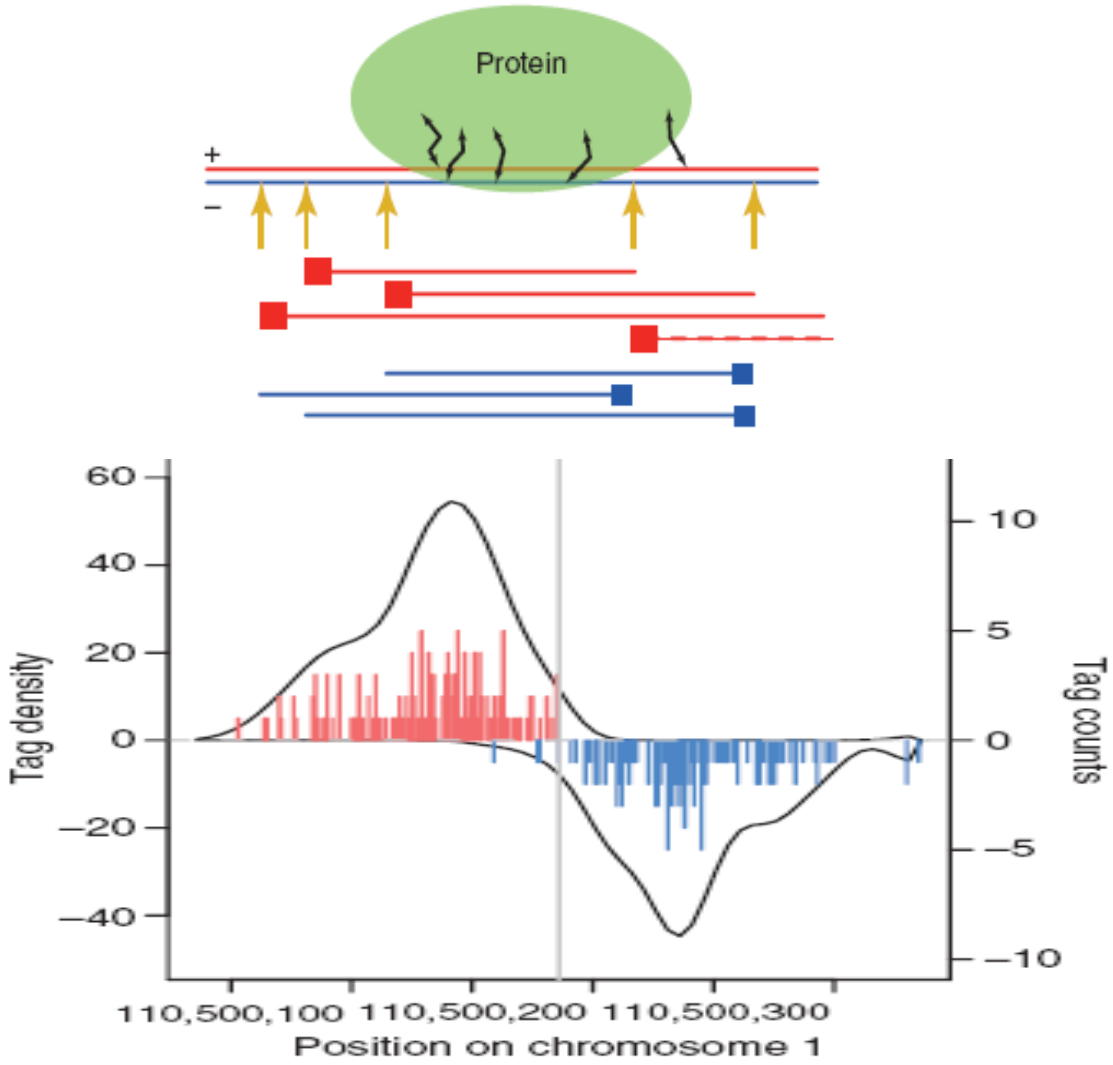
mES cell Oct4 ChIP Seq displays distinct binding events



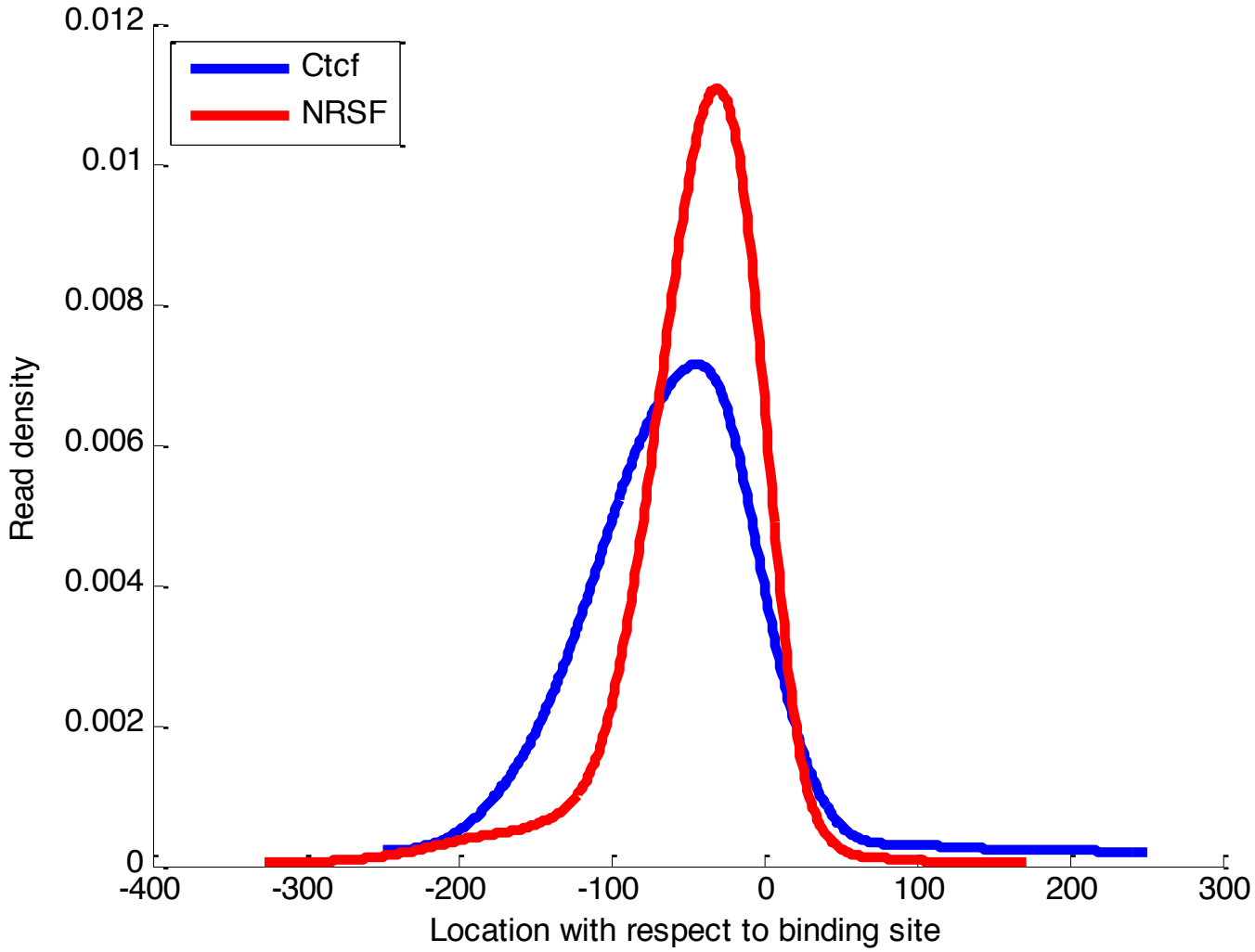
Repetitive “blacklisted” regions are not considered and are gaps in our knowledge of genomic function



Chromatin Immunoprecipitation (ChIP) sequencing (ChIP-seq) reveals genome-protein interactions

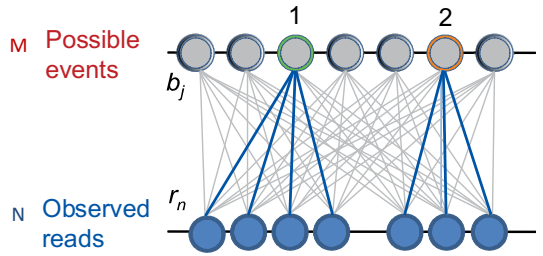


The read spatial distribution can be learned



Motif-based positional prior biases the binding event prediction

Mixture model



$$p(R \mid \pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n \mid j), \quad \sum_{m=1}^{M} \pi_m = 1$$

Position specific priors

- Events are sparse
- Events occurs more likely at motif positions

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_s + \alpha_m}$$

α_s : uniform sparse prior parameter governing the degree of sparseness, $\alpha_s > 0$;
α_m : position specific motif-based prior

Solve the model by maximizing the regularized likelihood of observed reads using the Expectation-Maximization (EM) algorithm:

E step

$$\gamma(z_n = m) = \frac{\pi_j p(r_n \mid m)}{\sum_{m'=1}^{M} \pi_{j'} p(r_n \mid m')}$$

M step

$$\hat{\pi}_m^{(i)} = \frac{\max(0, N_m - \alpha_S + \alpha_m)}{\sum_{m'=1}^{M} \max(0, N_{m'} - \alpha_S + \alpha_m)}$$

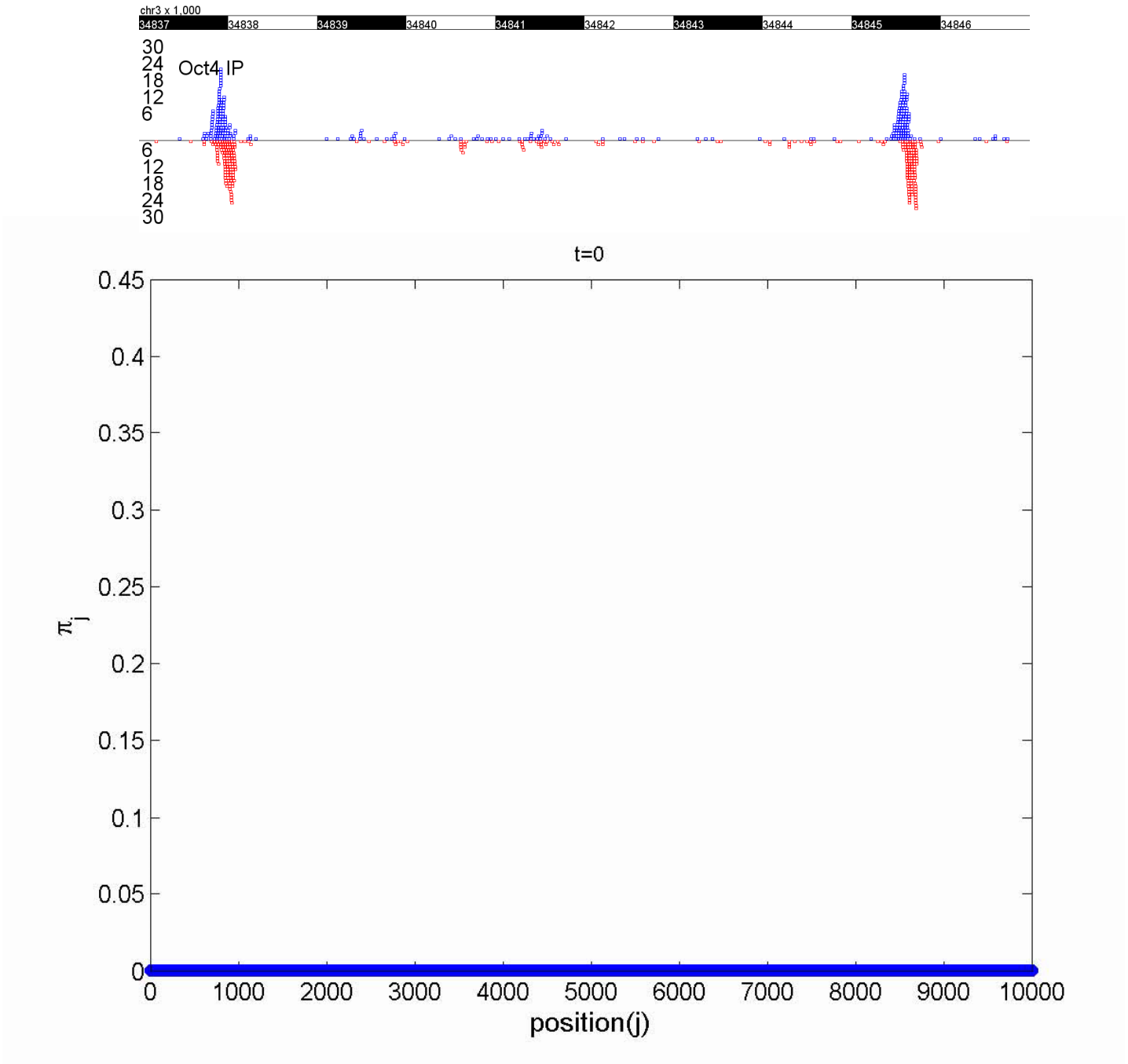
$$N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

γ is fractional responsibility of location m for read n

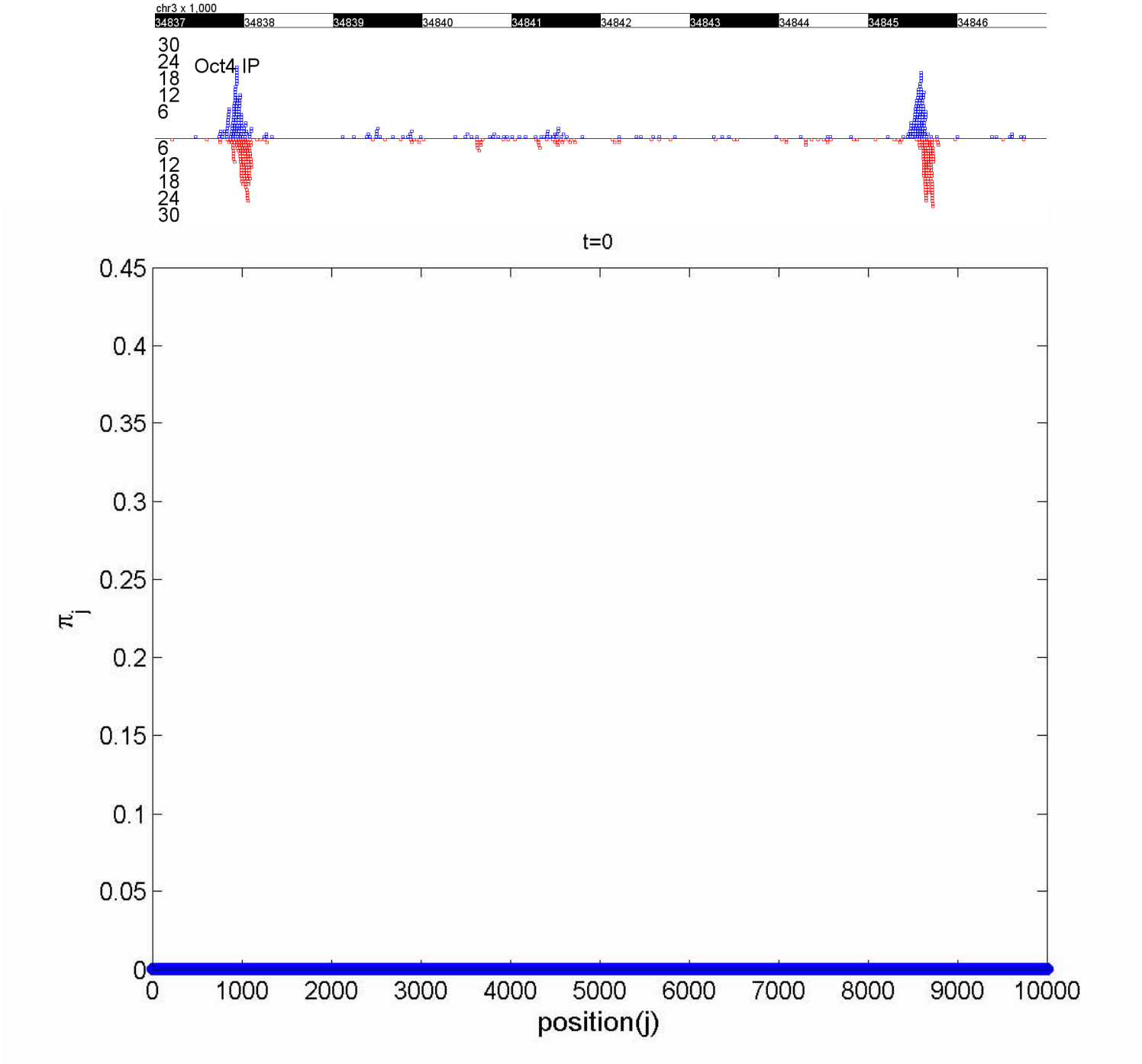
π is fraction of reads produced by location m

Component elimination

EM –
no prior


chr3 x 1,000
34837
34838
34839
34840
34841
34842
34843
34844
34845
34846
Oct4 IP
30
24
18
12
6
6
12
18
24
30
t=0
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
π_j
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
position(j)

EM – Sparse prior



Sequence logos describe what is bound

$S_{b,i} =$

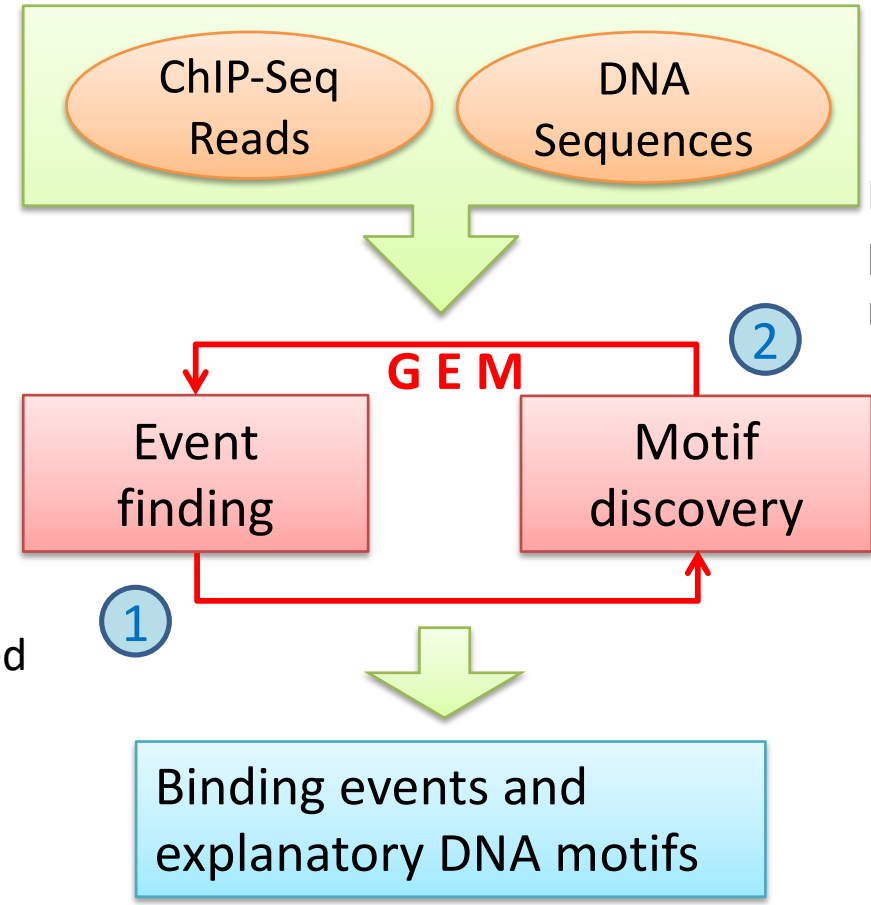
Logo height of base b at position i

$f_{b,i} =$ Fraction of base b at position i

$$I_i = 2 + \sum_{b \in \{A,C,G,T\}} f_{b,i} log_2 f_{b,i}$$

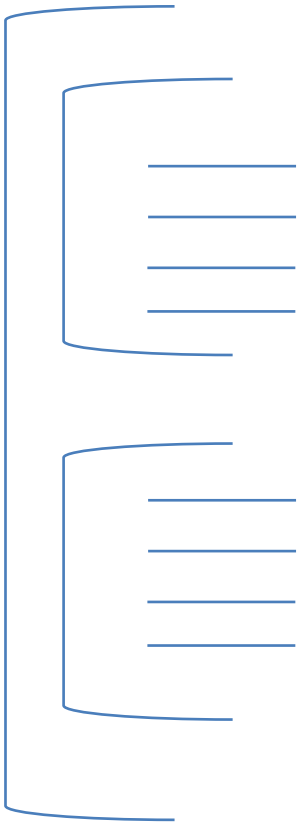
$$S_{b,i} = f_{b,i}\, I_i$$

Genome-wide Event finding and Motif discovery



Chance of a sequence occurring k times in bound set by chance (hypergeometric)

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



K Bound

N total

N-K Unbound

k occurrences of a sequence in K bound sequences

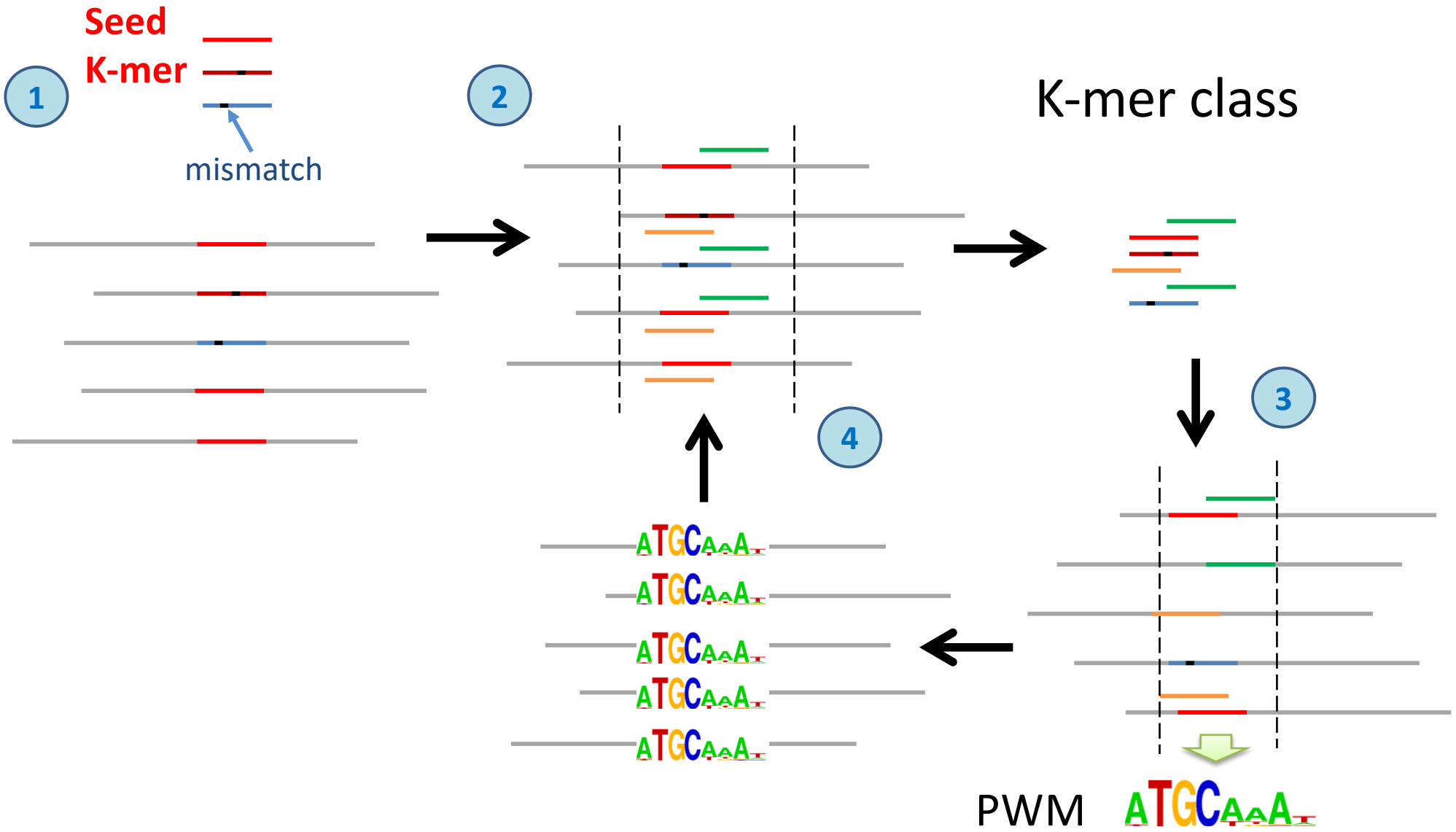
n-k occurrences of a sequence in N-K unbound sequences

Initial set of over-represented kmers

- N is the total number of positive and negative training sequences,
- N+ is the number of positive training sequences,
- n is the number of positive and negative training sequences containing the k-mer (positive and negative hit count), and
- l is the number of positive training sequences containing the k-mer (positive hit count). Component k-mers are required to have a HGP less than 10^{-5}.

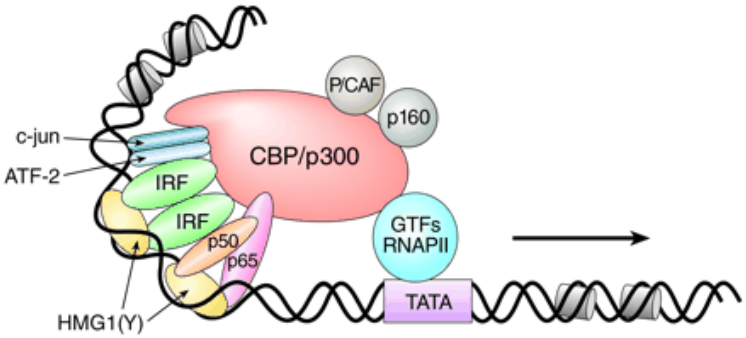
$$hgp = \sum_{l=n_+}^{\min(N_+, n)} \frac{\binom{N_+}{l}\binom{N-N_+}{n-l}}{\binom{N}{n}}$$

K-mer class motif discovery

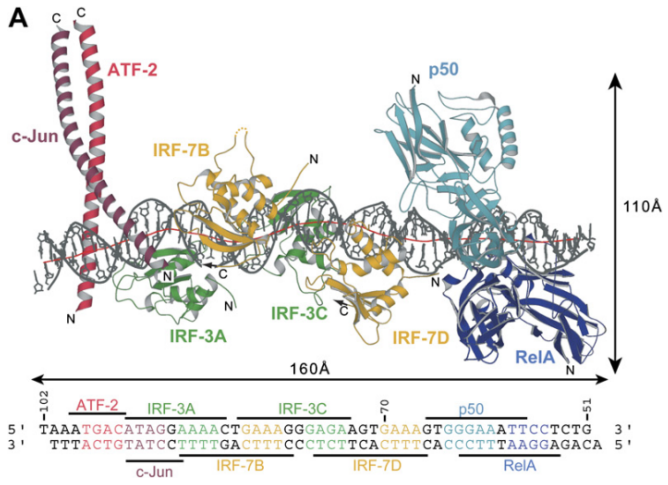


The spatial arrangement of transcription factor binding is critical

The IFN-β enhanceosome



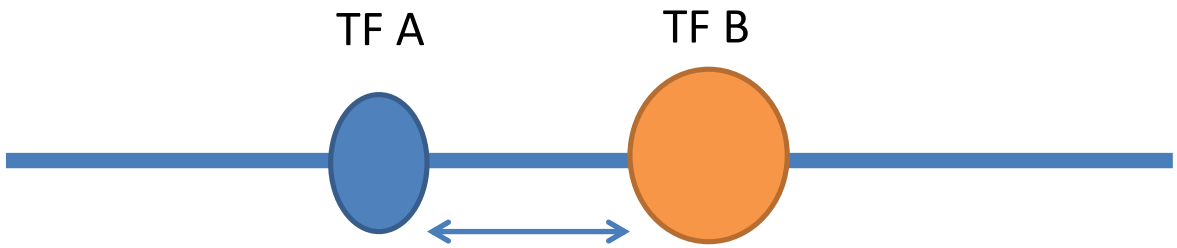
(Vo and Goodman, JBC, 2001)



(Panne, Cell, 2007)

- Single point mutations disable the enhancer
- No major protein-protein interaction

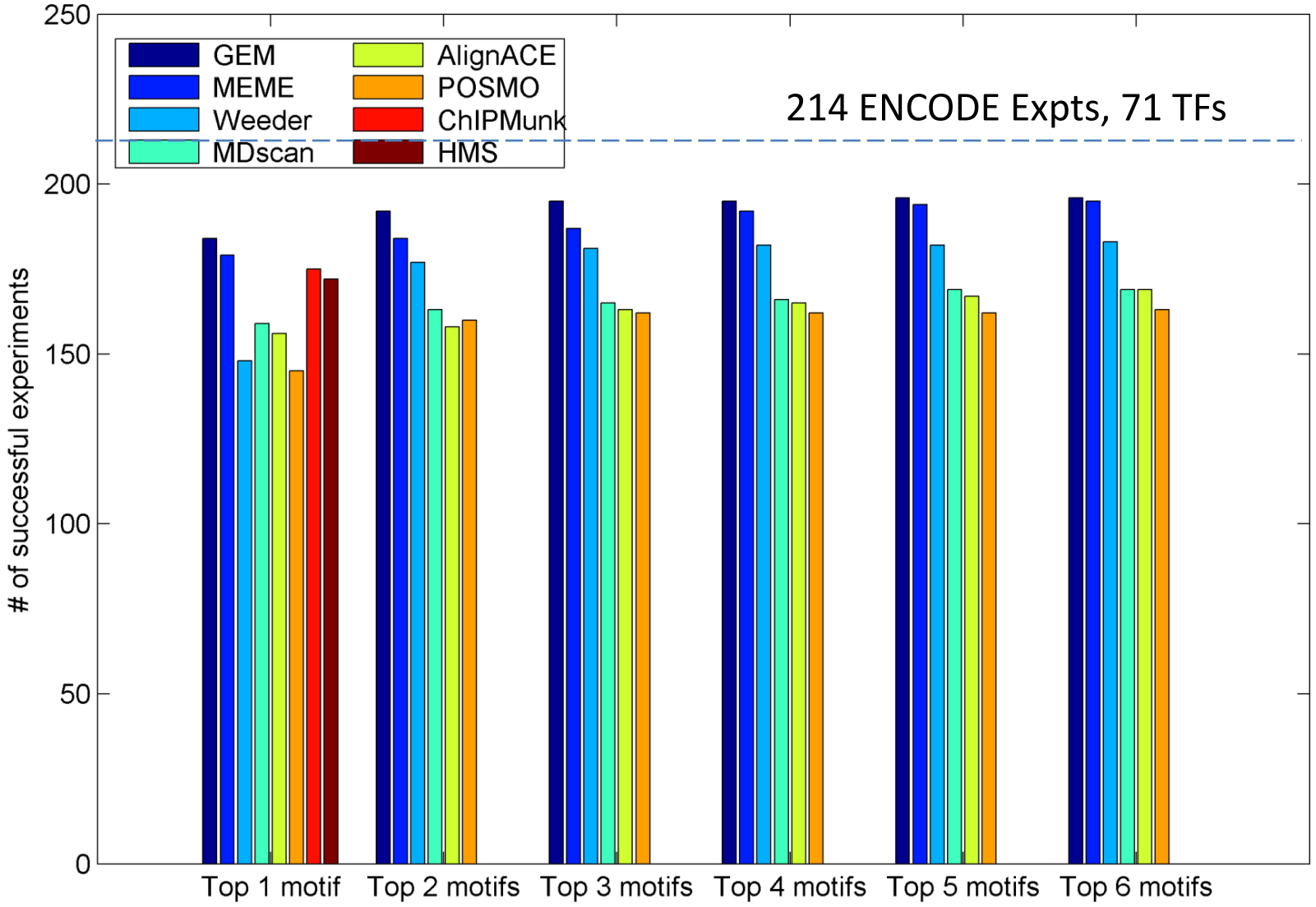
A precise genome wide characterization of *in vivo* spacing constraints between key transcription factors would reveal key aspects of the gene regulation.



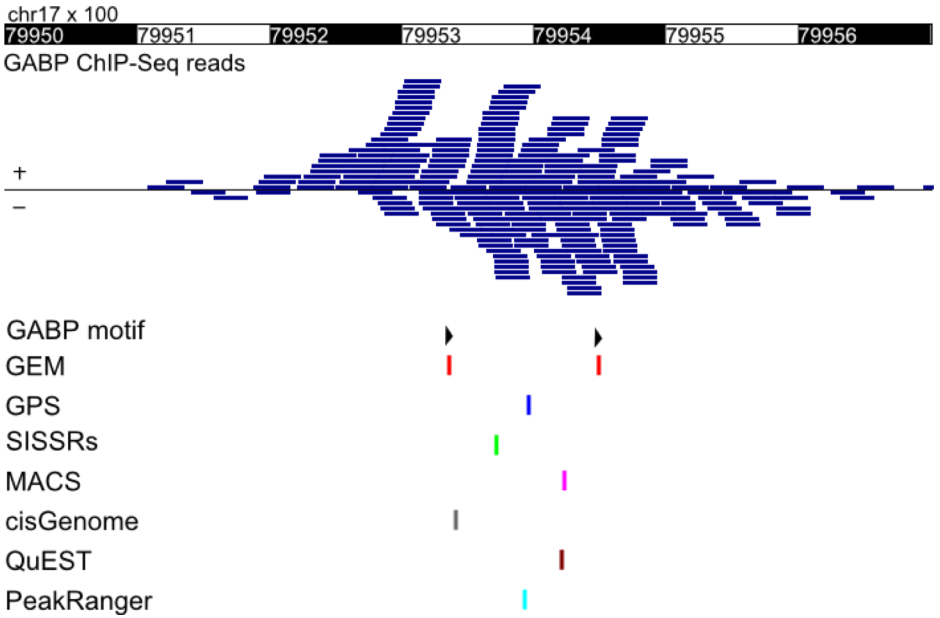
Evaluating GEM motif discovery

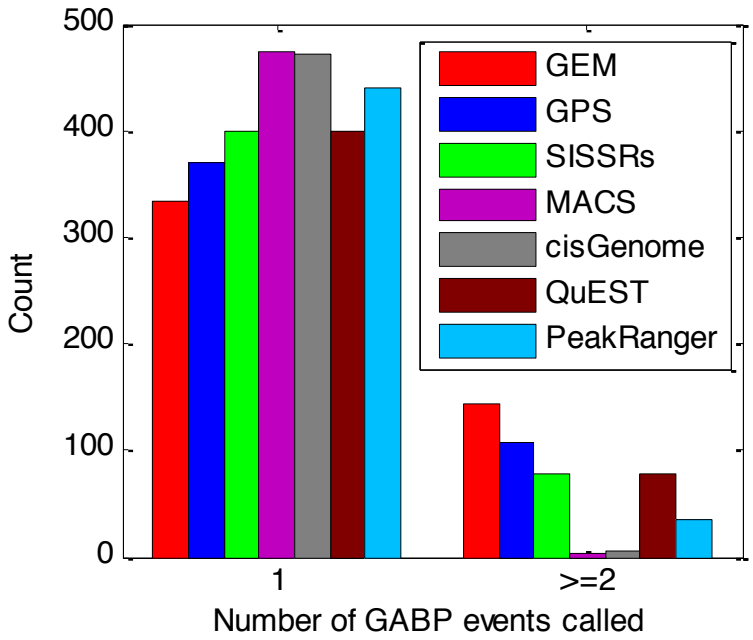
- Apply to a large set of ENCODE ChIP-Seq data
- Compare with 6 other methods
- Collect motif PWMs from public databases
- Compare discovered motifs with known motifs using STAMP

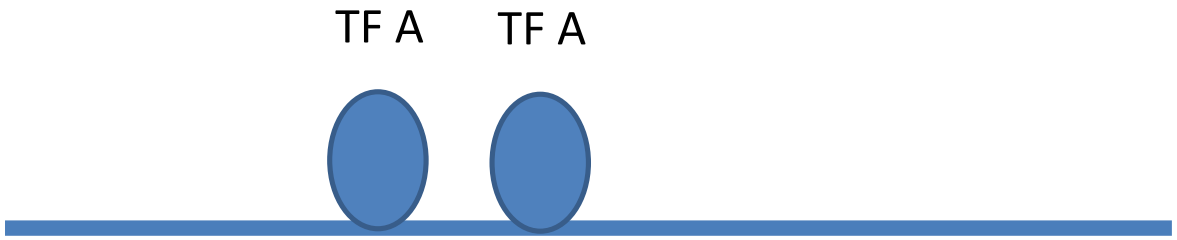
GEM motif discovery outperforms other methods when detecting motifs in ChIP-Seq data



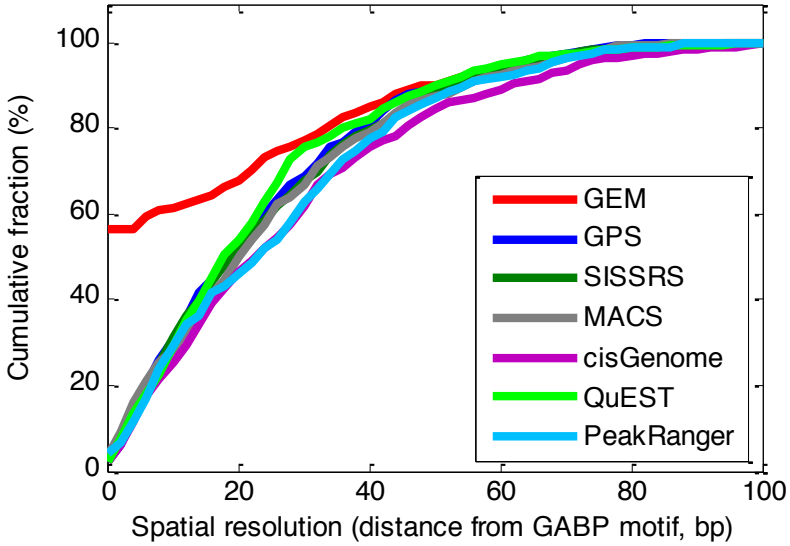
GEM improves the spatial accuracy in resolving proximal binding events.

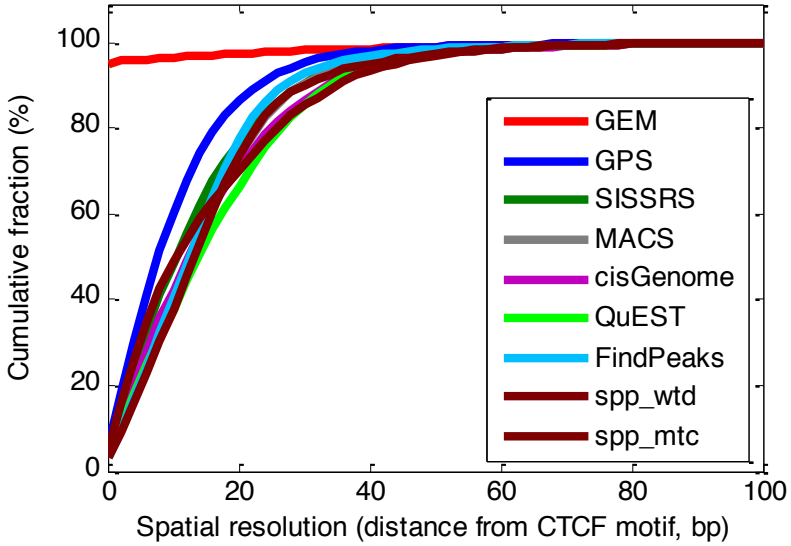


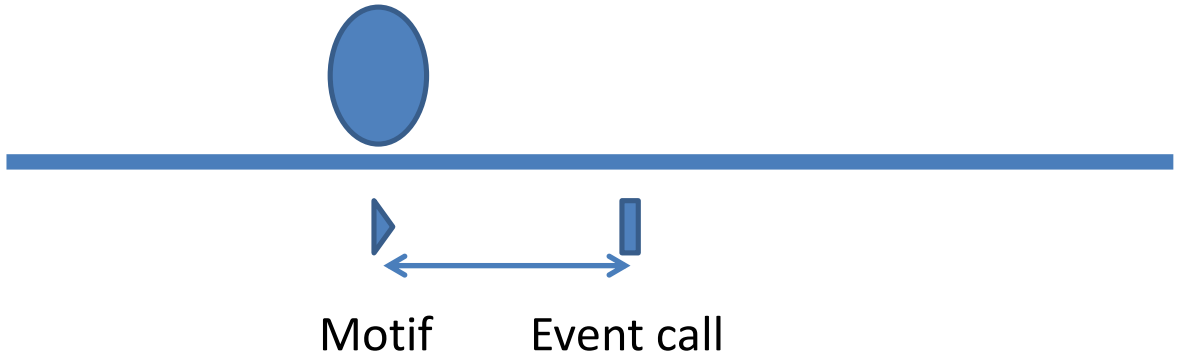




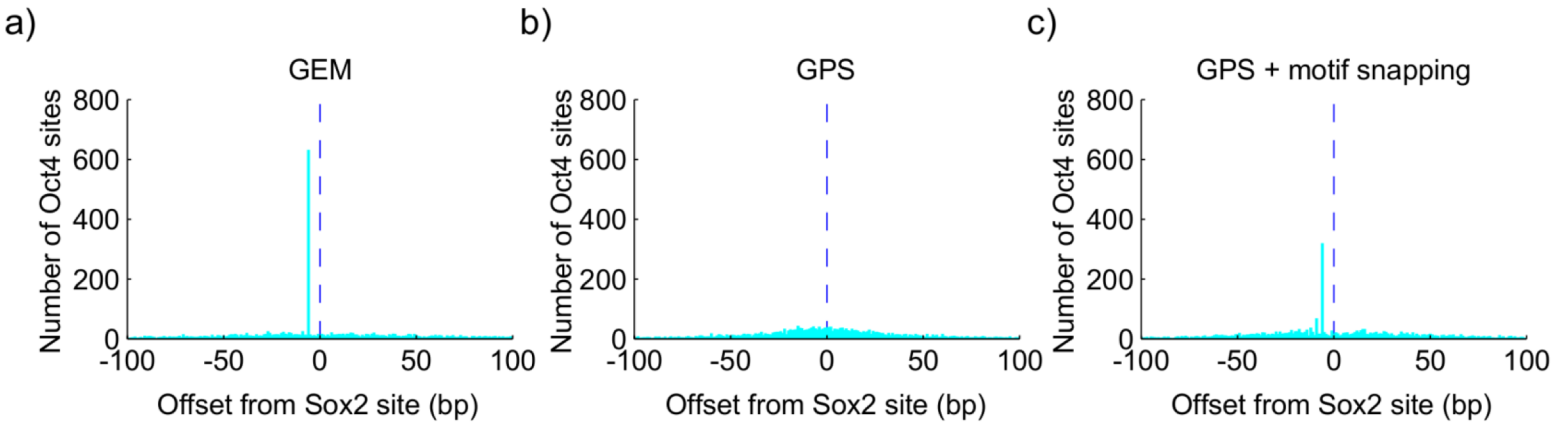
GEM improves spatial accuracy in binding event prediction

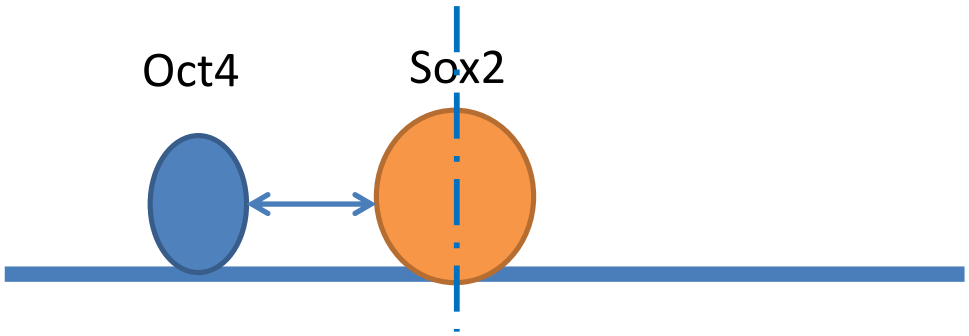


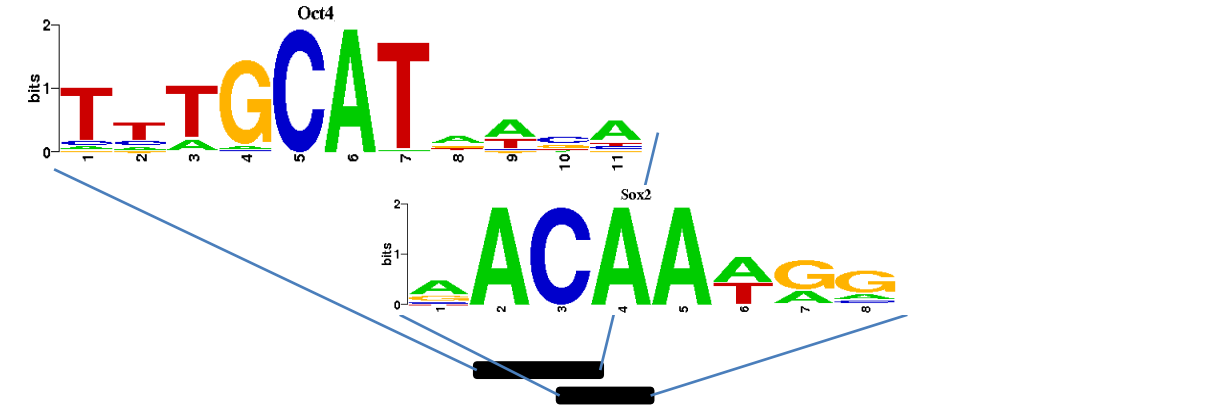


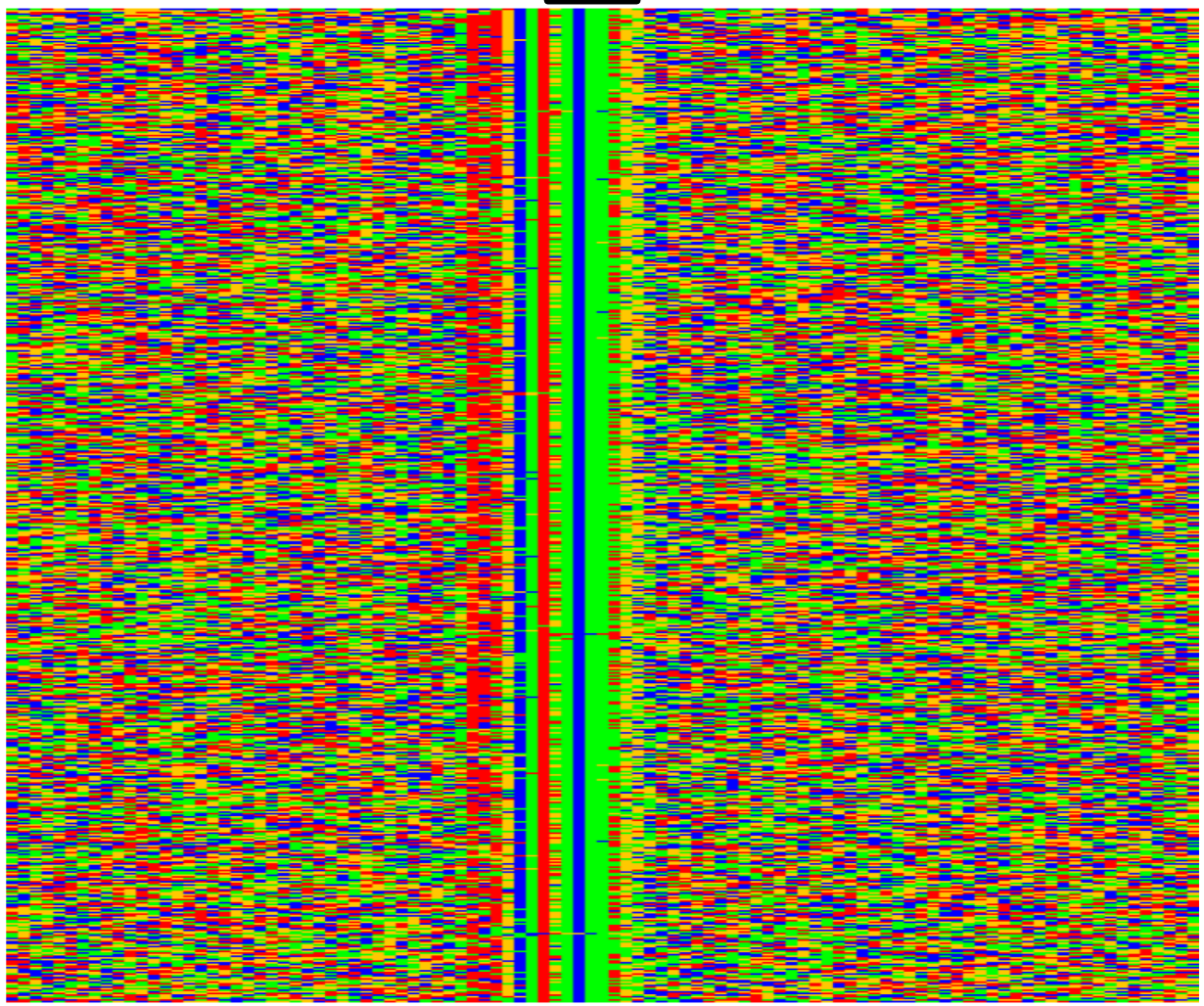


GEM reveals transcription factor spatial binding constraints



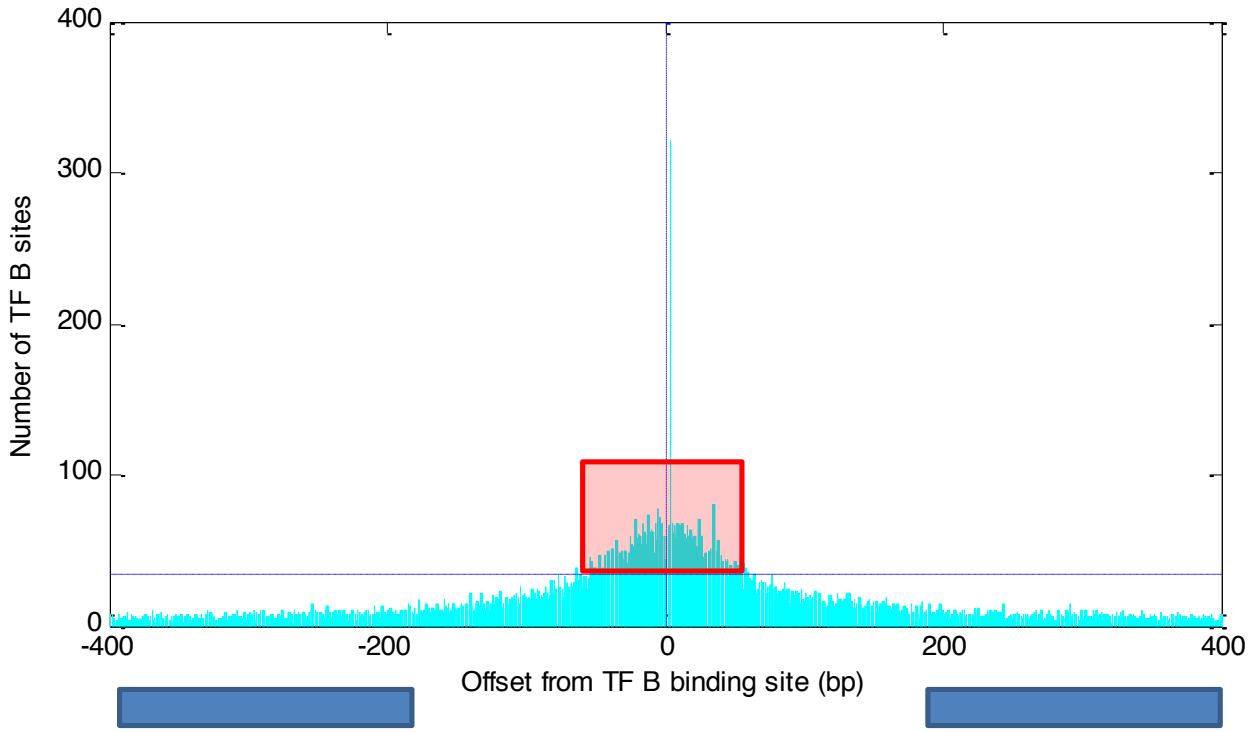



Oct4
bits
Sox2
bits



What are significant spacings?

- Compute average number of motifs in background region [200bp 400bp] and [-400bp -200bp]
- Use Poisson CDF to compute p-value of number of occurrences at each location in [-100bp 100bp]
- Bonferroni correct each p-value (p-value x 201)
- We choose that a corrected p-value is significant at 10^-8

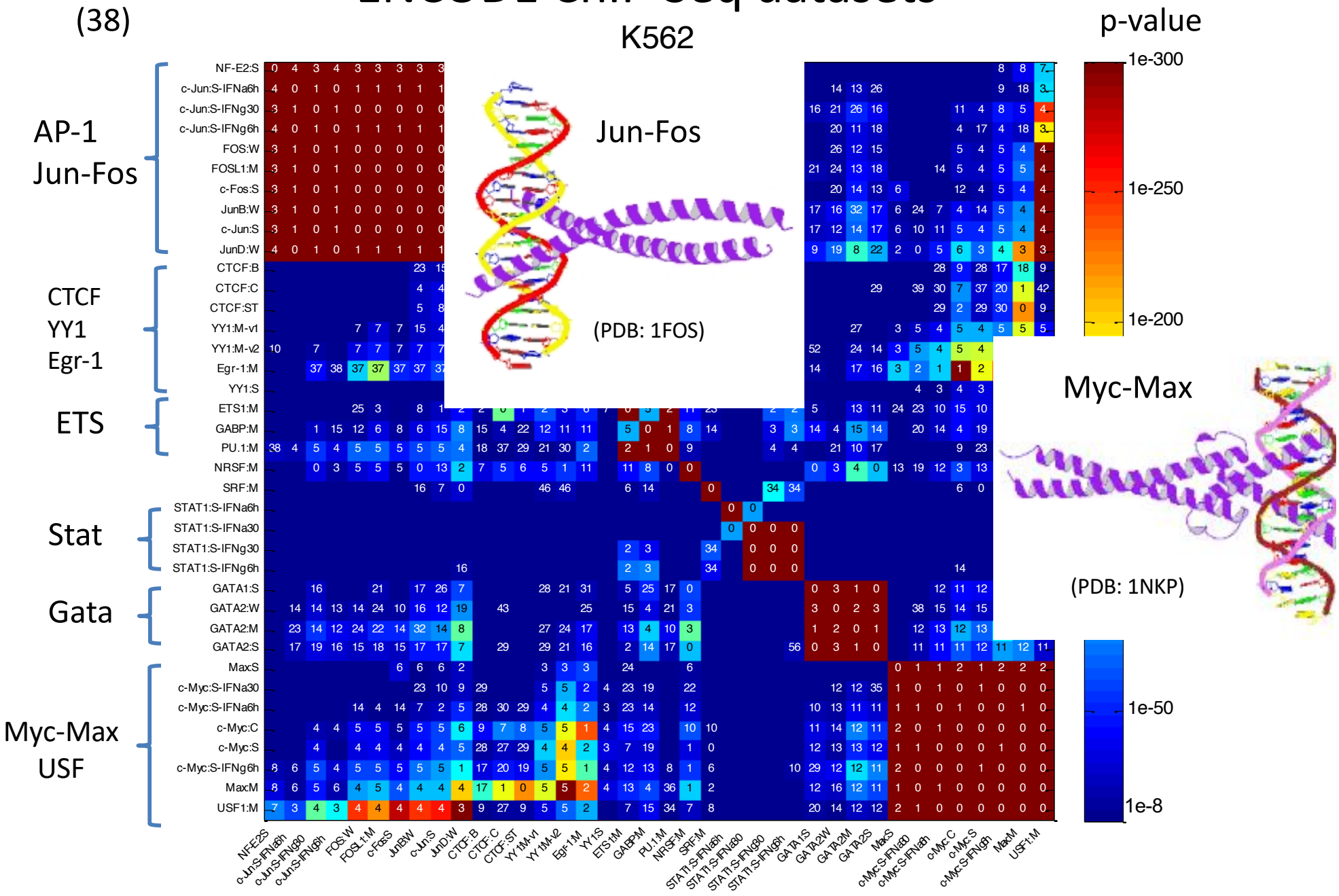


Spatial binding constraints detected from ENCODE ChIP-Seq datasets

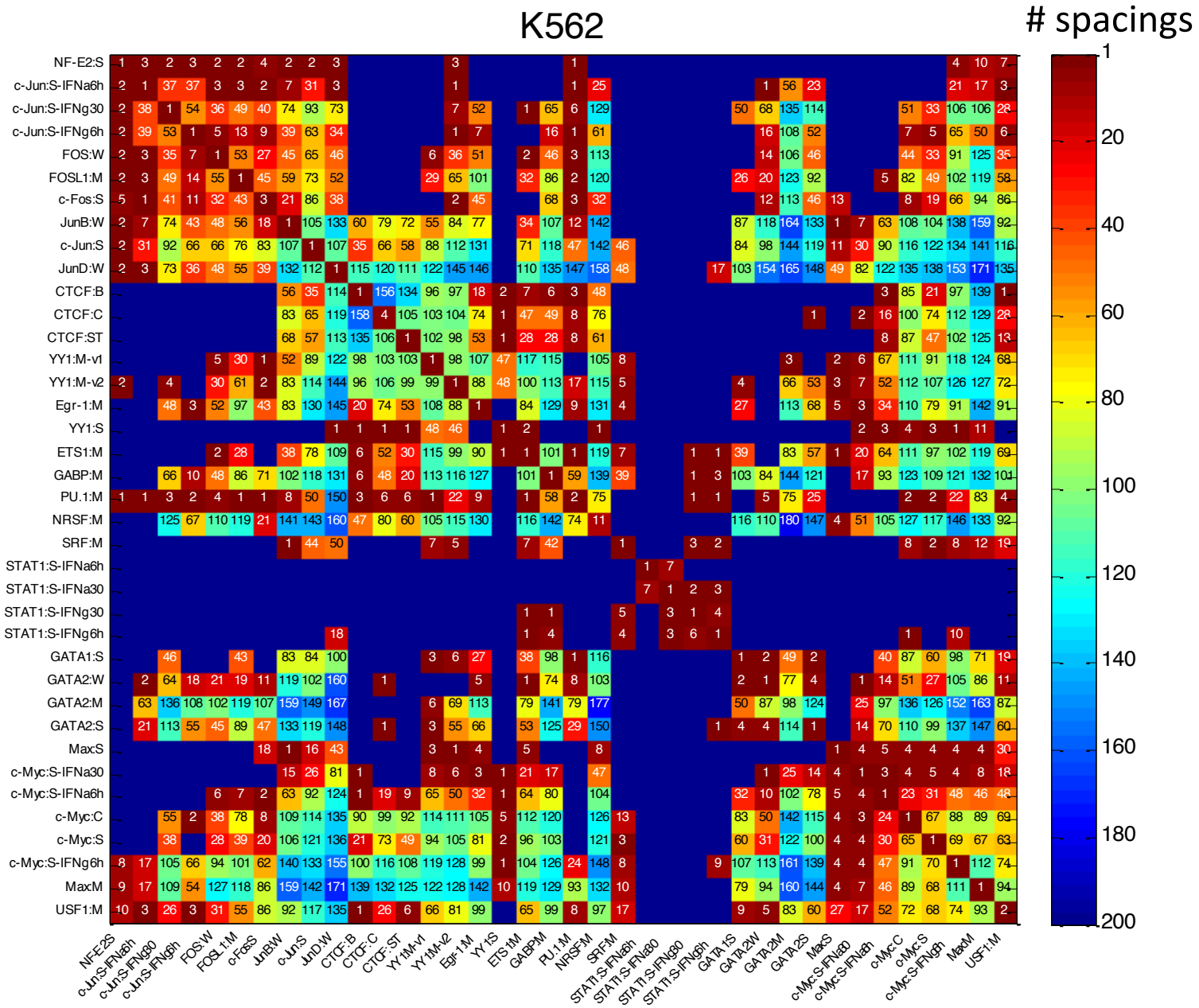
Cell type	Description	Expts	Constraint pairs
K562	leukemia	38	154
GM12878	lymphoblastoid	29	134
HepG2	liver carcinoma	21	86
HeLa-S3	cervical carcinoma	13	34
H1-hESC	embryonic stem cells	11	19

(355 distinct TF pairs)

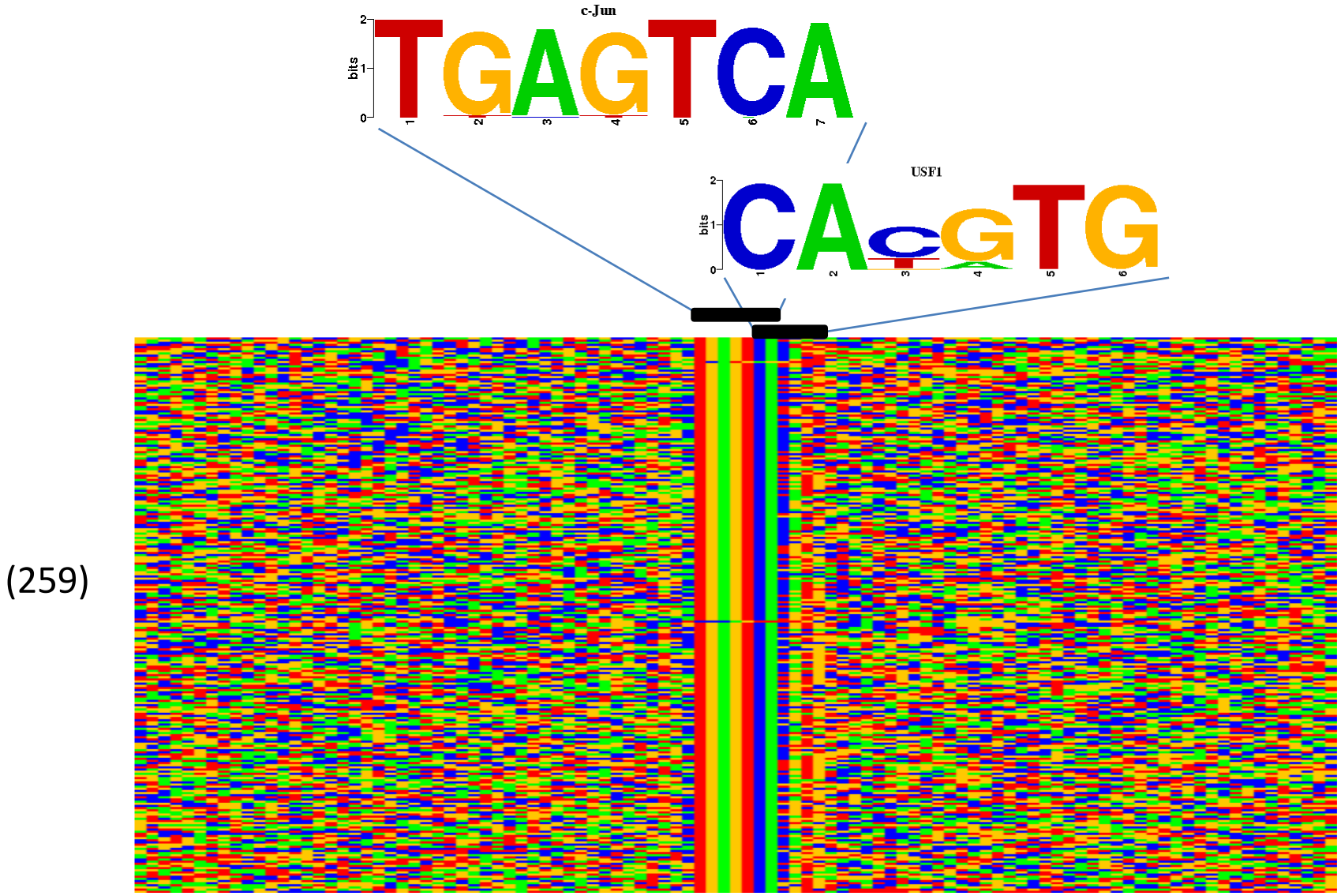
Spatial binding constraints detected from ENCODE ChIP-Seq datasets



Spatial binding constraints detected from ENCODE ChIP-Seq datasets

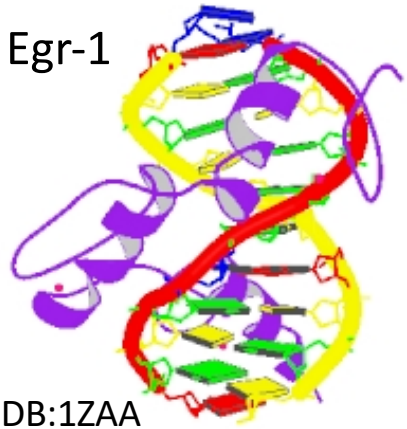


Cooperative binding: c-Jun/USF-1

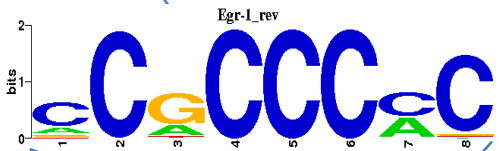


Competitive binding: CTCF/Egr-1

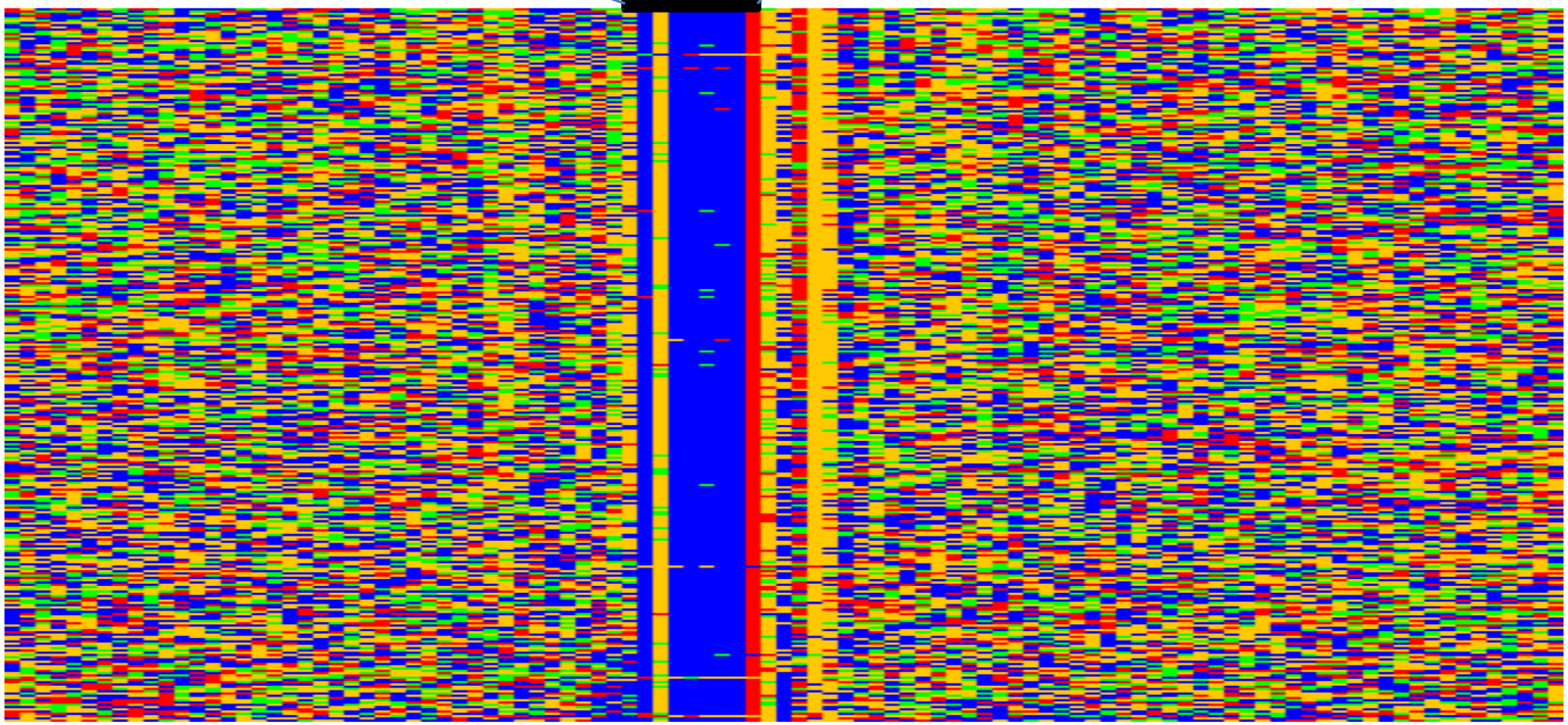
Egr-1



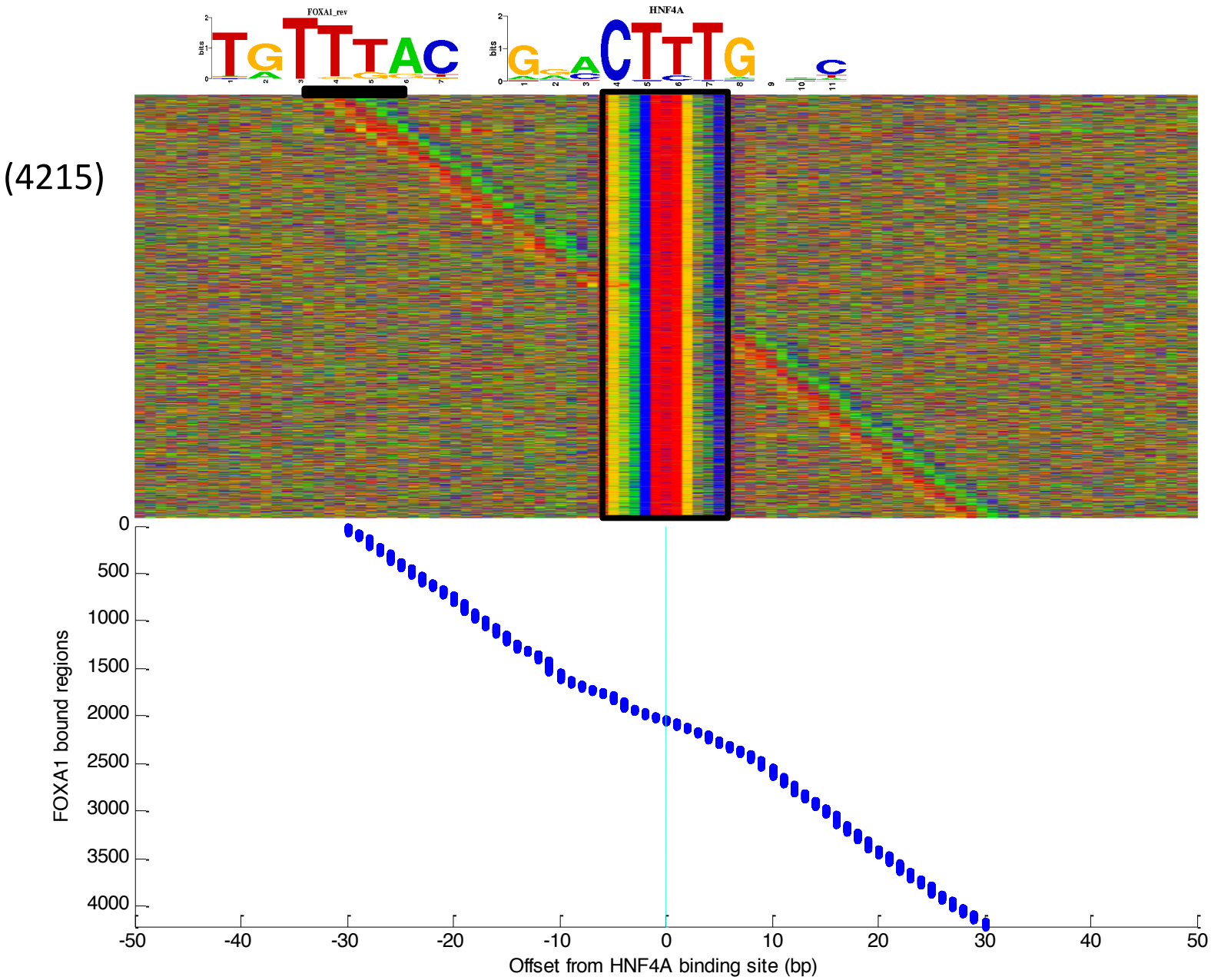
PDB:1ZAA



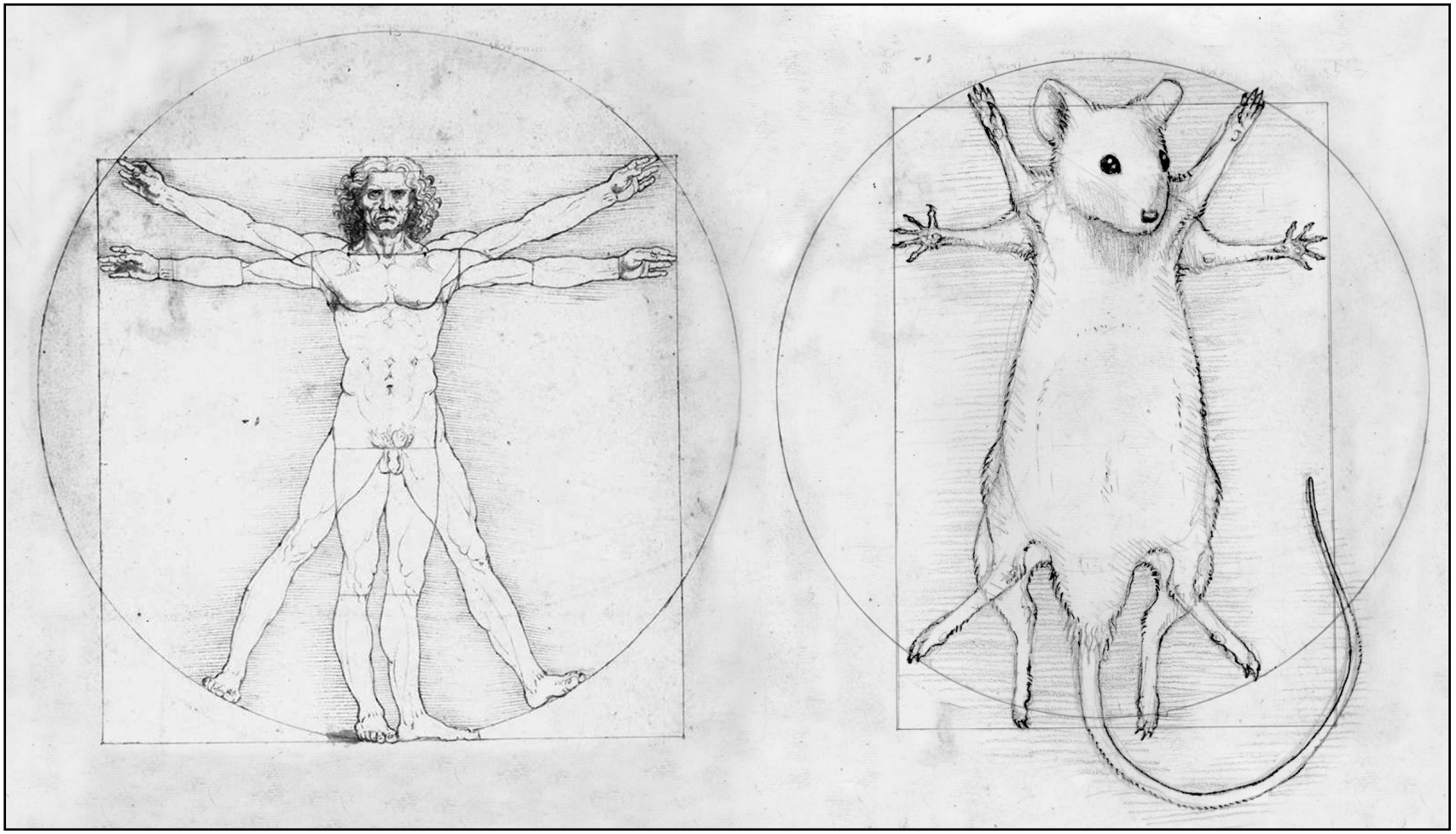
(315)



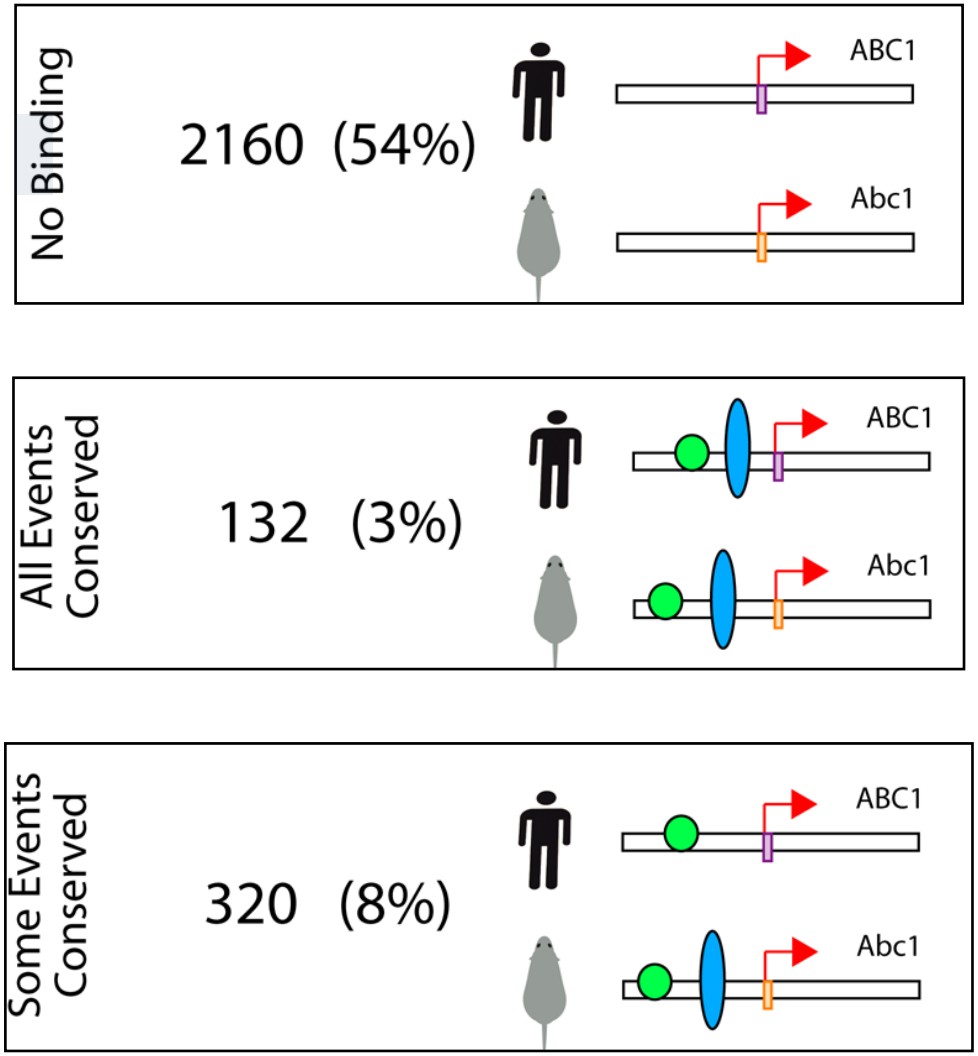
Collaborative binding: HNF4-α/FOXA1



Is conservation a good predictor of conserved binding events across species?



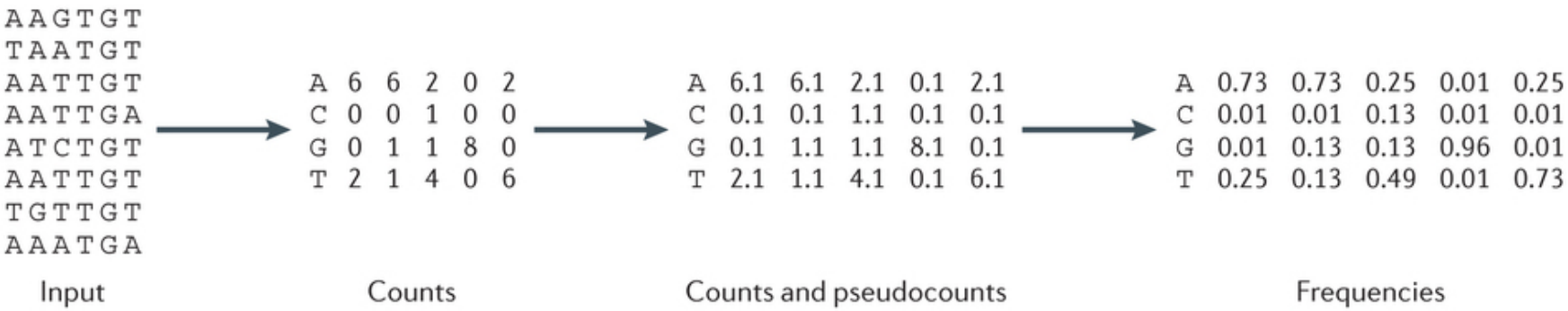
Promoter proximal binding is not well conserved in liver (FOXA2, HNF1A, HNF4A, HNF6)



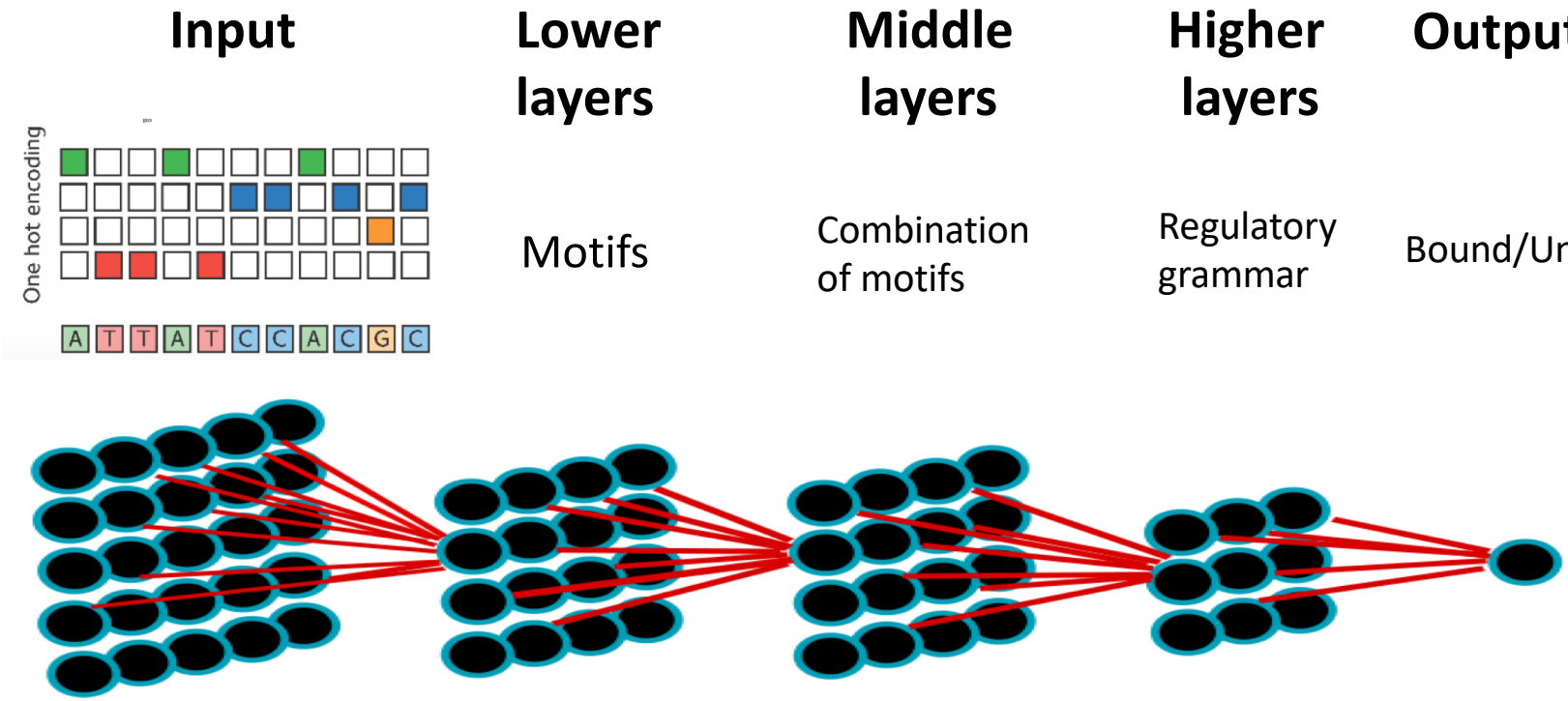
D. Odom, R. Dowell E. Fraenkel, D. Gifford Labs
Nature Genetics, 2007

Deep learning approaches to TF-DNA binding

Traditional DNA-protein binding models



One possible learned network structure



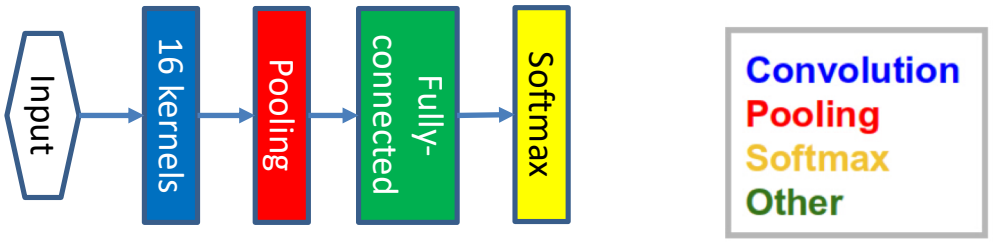
CNNs can outperform conventional approaches in modeling DNA-protein binding

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

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

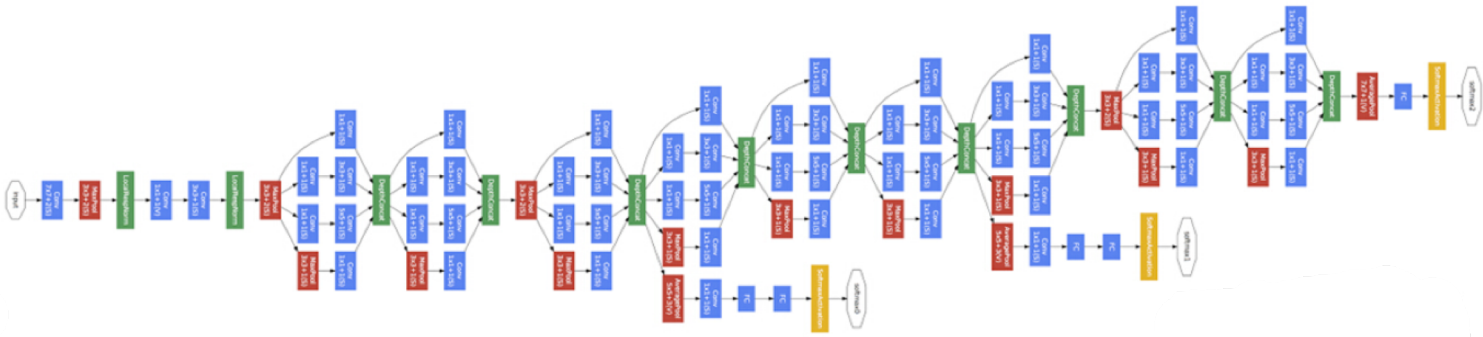
DeepBind (2015):

One convolutional layer with 16 kernels, maximum pooling window



DeepBind is “shallow learning” compared with other CNNs

GoogLeNet[1]
(Computer Vision)



DeepBind

Convolution
Pooling
Softmax
Other

[1] Szegedy et al. Going Deeper with Convolutions.

Open questions about deep learning for genomics

- What architectures work best to model DNA-protein binding?
- How “deep” should a network be?
- What components of the network contribute most to overall performance?
- Is the optimum network design specific to the task / experiment / TF?

Our approach

- We developed a framework to systematically benchmark CNN architectures on genomics tasks
- We analyzed the contribution of different network components
- We explored if the optimum architecture is task-specific
- We evaluated training data requirements

Systematic benchmarking is important

- Task should be meaningful
  - *Real vs. artificial sequences (DeepBind): motif discovery*
    - Simple
    - Learn motif from similar nucleotide background
    - Not generalizable to classify real bound sequences

Systematic benchmarking is important

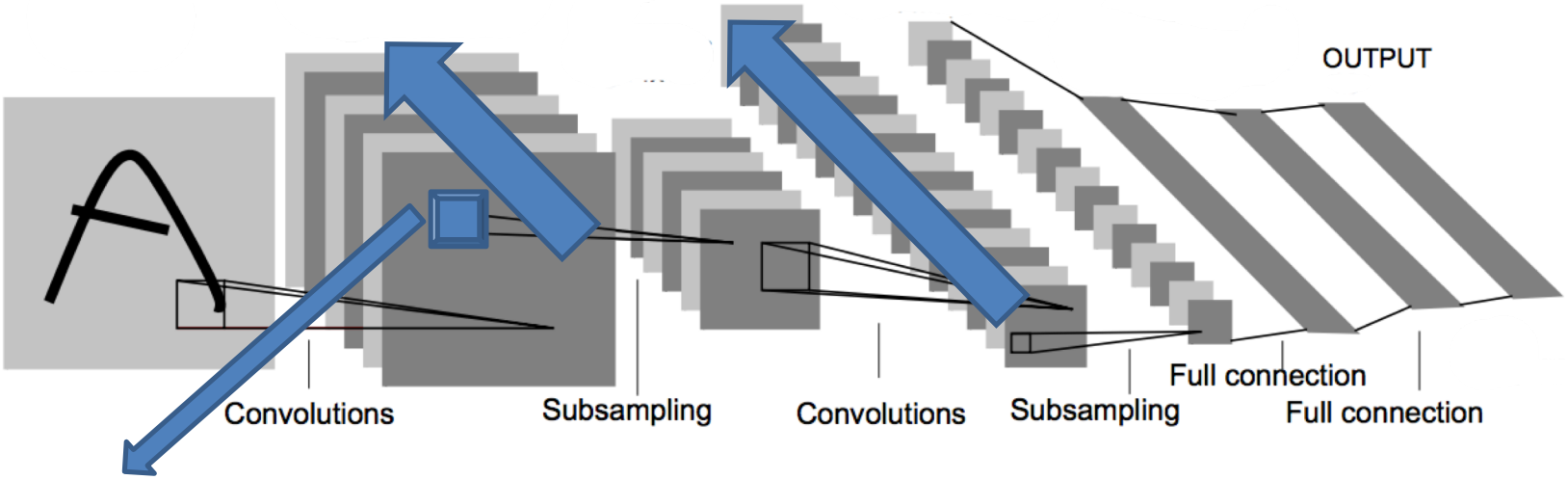
- Task should be meaningful
  - *Real vs. artificial sequences (DeepBind): motif discovery*
  - *Bound motif vs. unbound motif*: *motif occupancy*
    - Hard
    - Forces the model to learn better and higher-level sequence determinants

Systematic benchmarking is important

- Task should be meaningful
- Balance the number of positive and negative samples
- Control any artificial bias, location of the motif in the sample
- Conclusion should be the consensus across diverse TF ChIP-seq experiments (we used 690 from ENCODE)

CNNs have three important architectural dimensions to vary

More convolution kernels better capture feature diversity



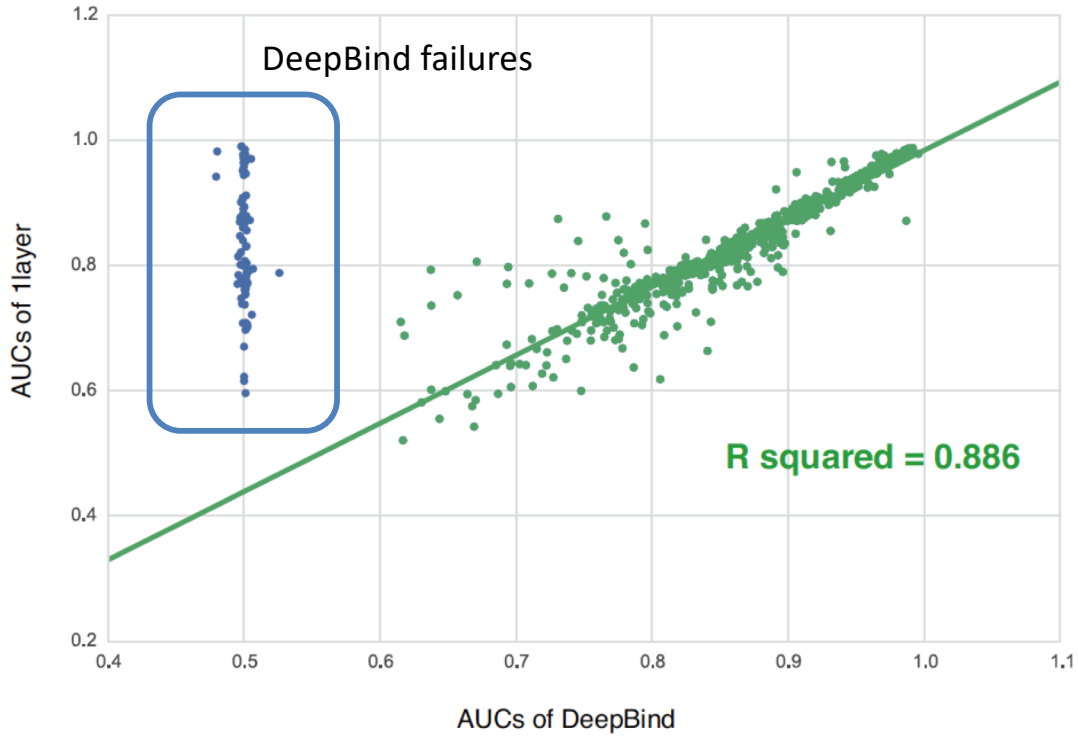
Smaller pooling window retains more location information

More layers can capture higher level features

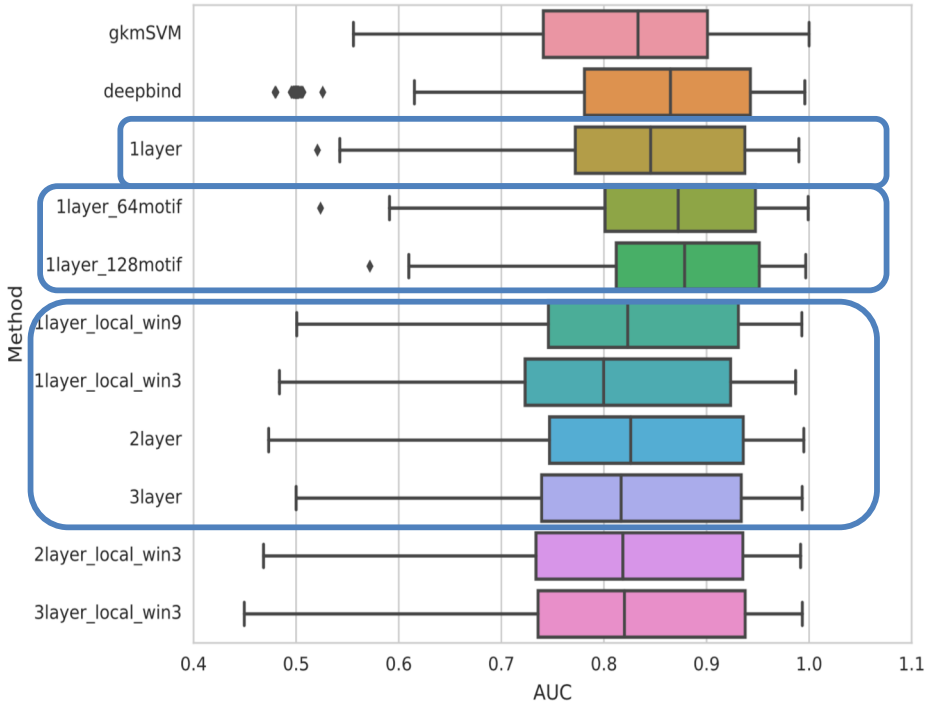
CNN architectures compared

Our Name	More Conv. Kernels	Deeper	Smaller pooling size
1layer (DeepBind)	-	-	-
1layer_64motif	✓	-	-
1layer_128motif	✓✓	-	-
1layer_local_win9	-	-	✓
1layer_local_win3	-	-	✓✓
2layer	-	✓	-
3layer	-	✓✓	-
2layer_local_win3	-	✓	✓✓
3layer_local_win3	-	✓✓	✓✓

Baseline model reproduces DeepBind

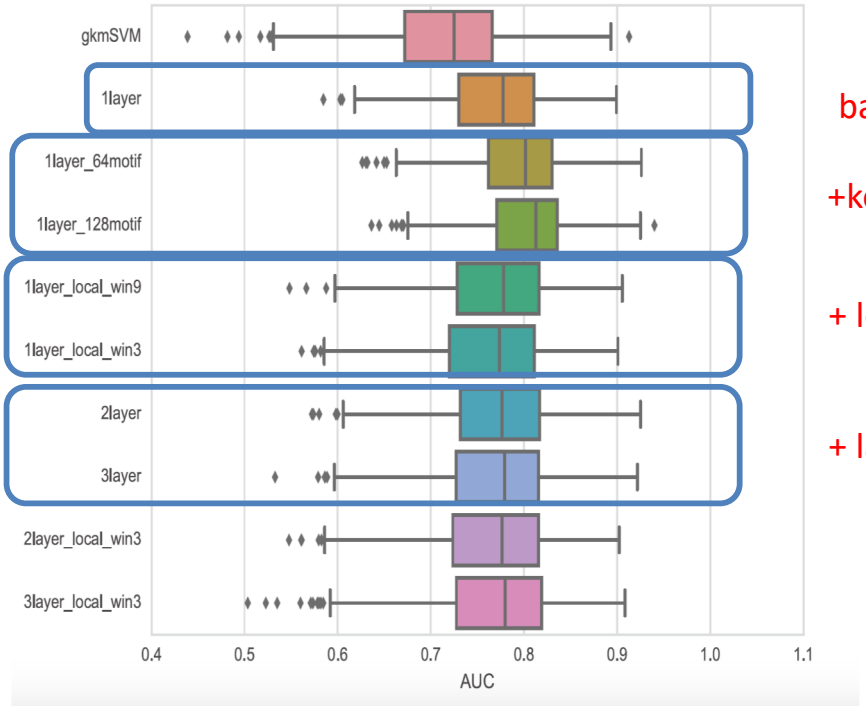


Simple models are best for a **motif discovery task**



- More convolutional kernels helps model motif diversity
- Smaller pooling size, more layers monotonically decrease performance
  - possibly because most determinants are low-level (motifs) and position-independent

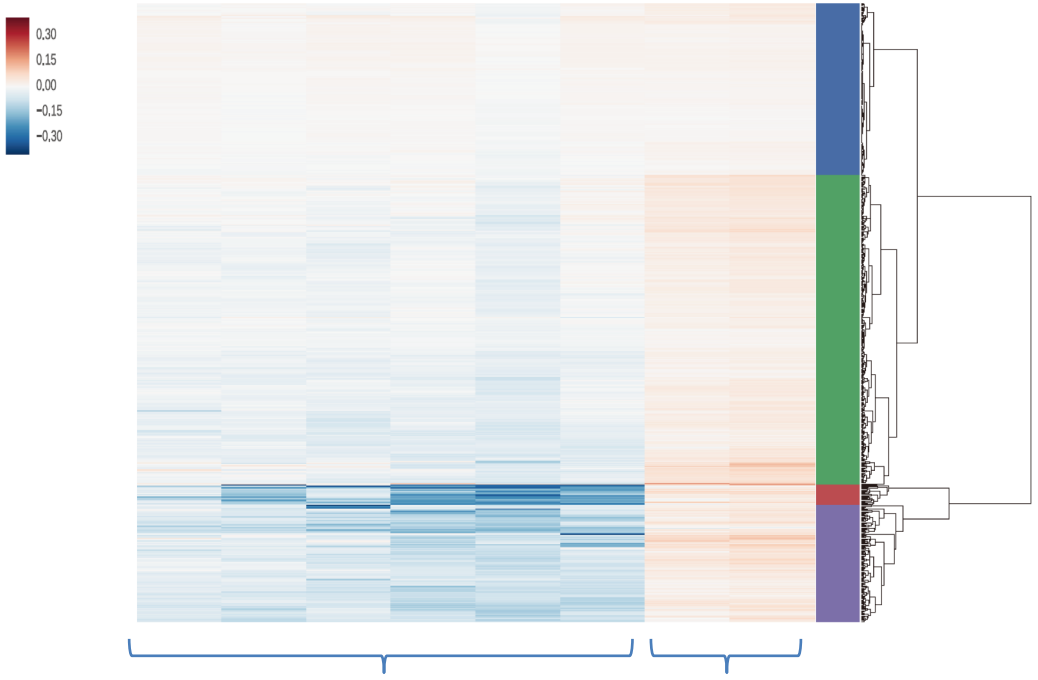
Depth improves performance in a **motif occupancy task**



- AUC decreases for all architectures
- More convolutional kernels help model the motif diversity
- Smaller pooling size slightly decreases the performance
- Deeper networks have slightly better performance
  - There are more high-level determinants that can be better modeled by deeper layers, consistent with the task design

Observed performance is experiment-specific

Performance relative to 1layer

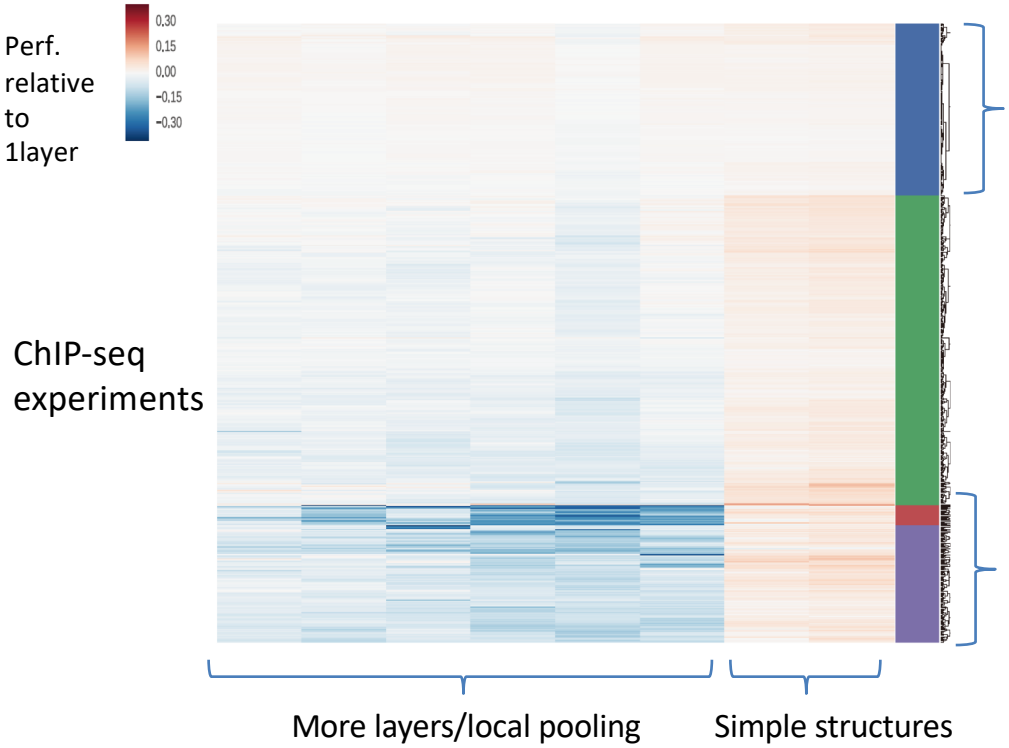


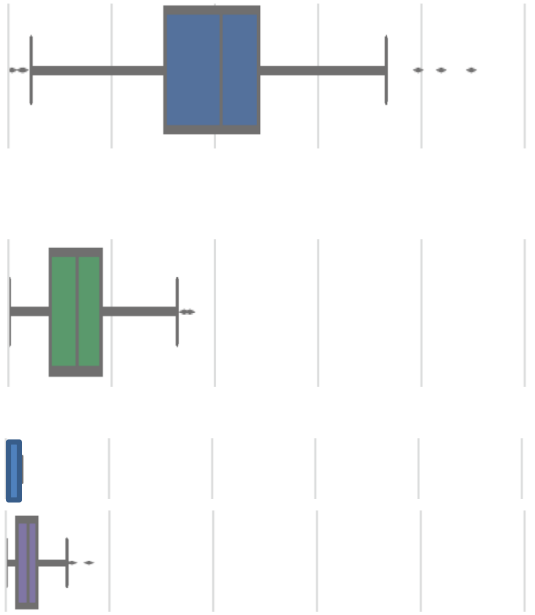
Small variance in performance

ChIP-seq experiments

Huge variance in performance

More complex networks require more training data

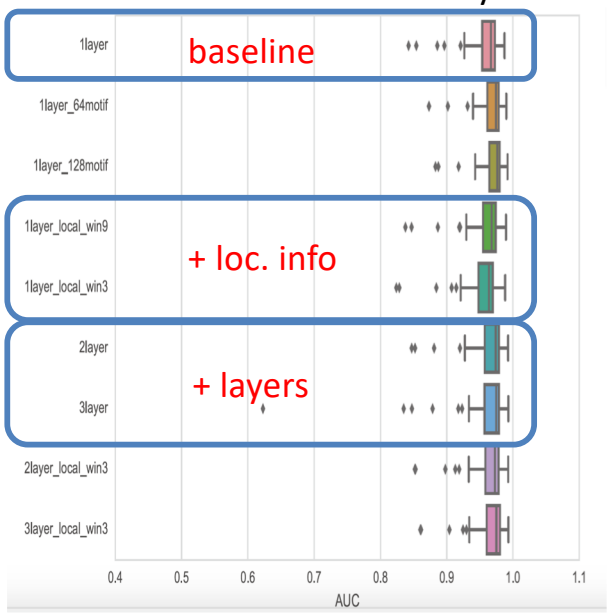




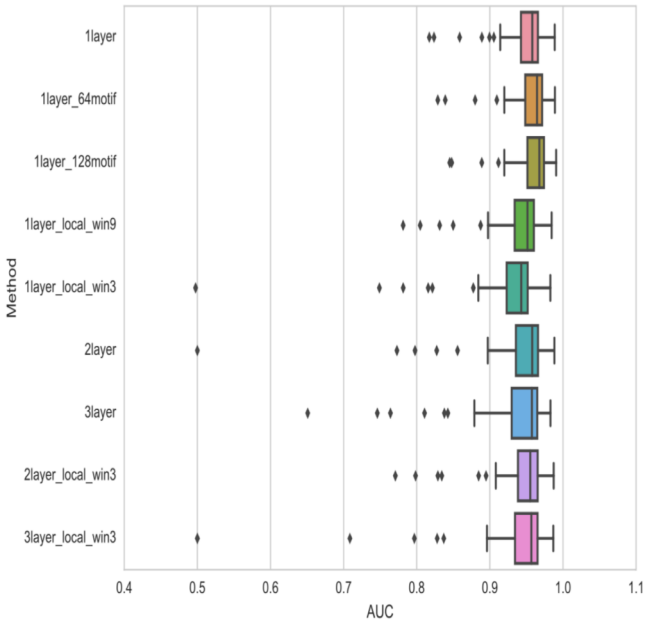
Training Example Count

Variance increases with fewer training examples

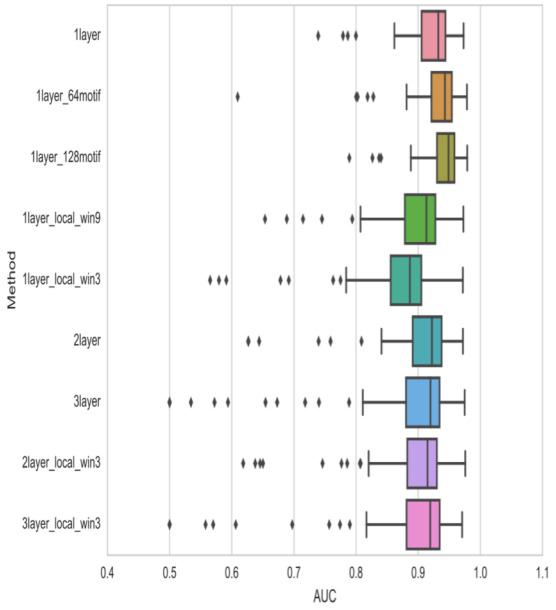
Performance on motif discovery task



80,000 training examples



20,000 training examples



5,000 training examples

CNNs can outperform conventional methods

- CNNs outperform conventional methods with the right structure
- The optimum structure is different from that in computer vision
- Different biological tasks and data yield different conclusions
- Understanding the problem at hand and comparing different structures is important to design a good CNN model for biology applications (http://cnn.csail.mit.edu)

FIN - Thank You