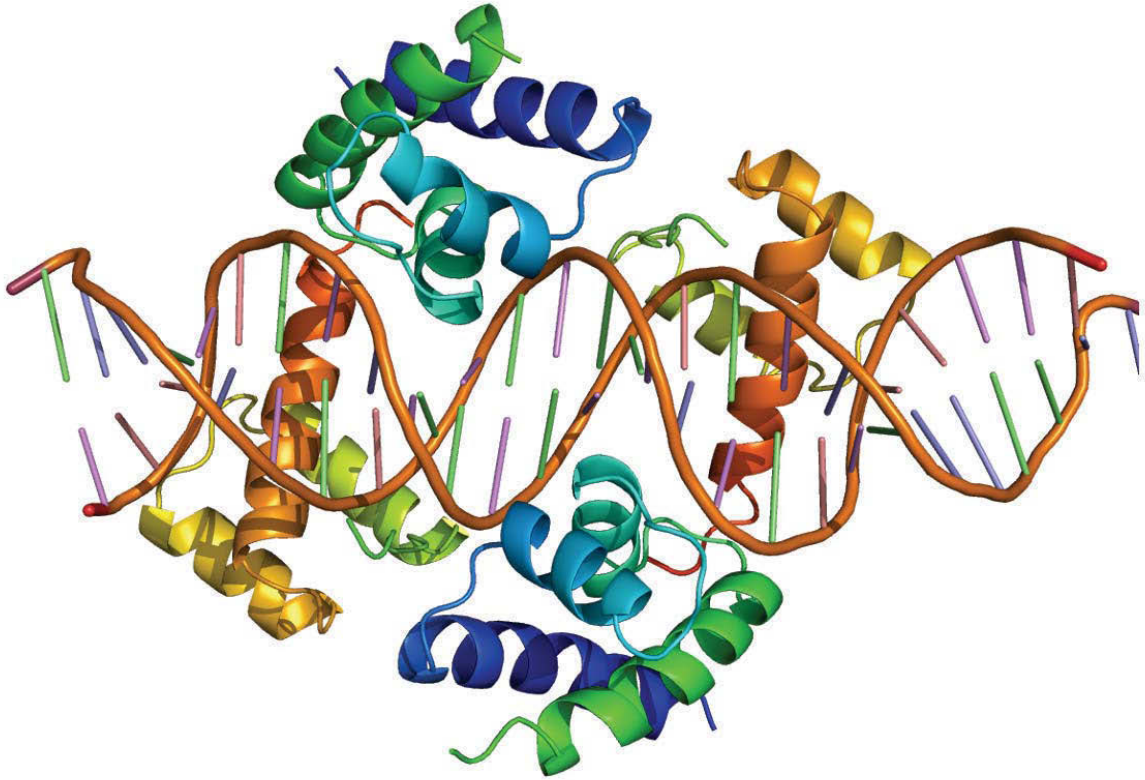
Lecture 7
ChIP-seq Analysis
Irreproducible Discovery Rate (IDR) Analysis

Foundations of Computational Systems Biology

David K. Gifford

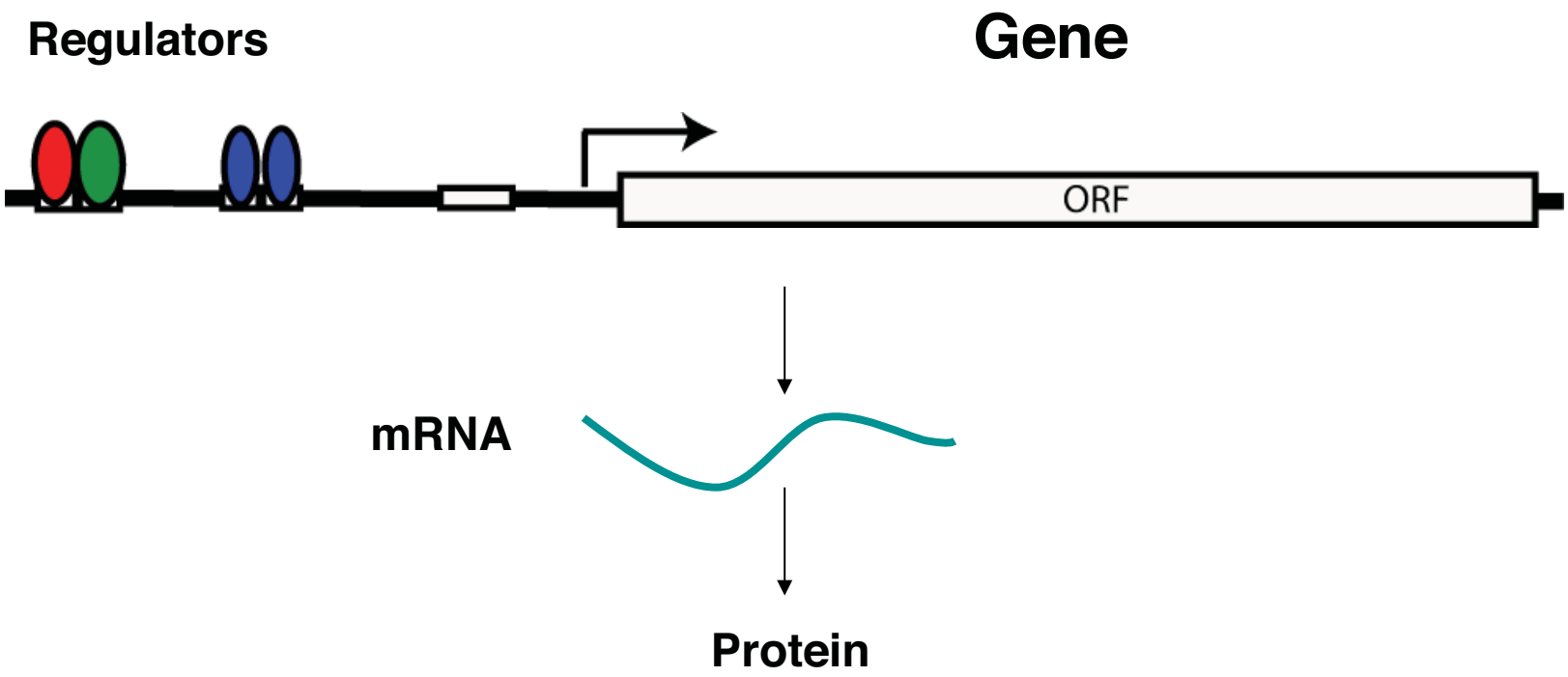
Transcription factors regulate gene expression



Transcription factors are proteins that bind to specific DNA sequences and act as molecular switches (Pit1 shown)

Humans have ~2000 gene regulators.

Gene Regulation: DNA -> RNA -> Protein

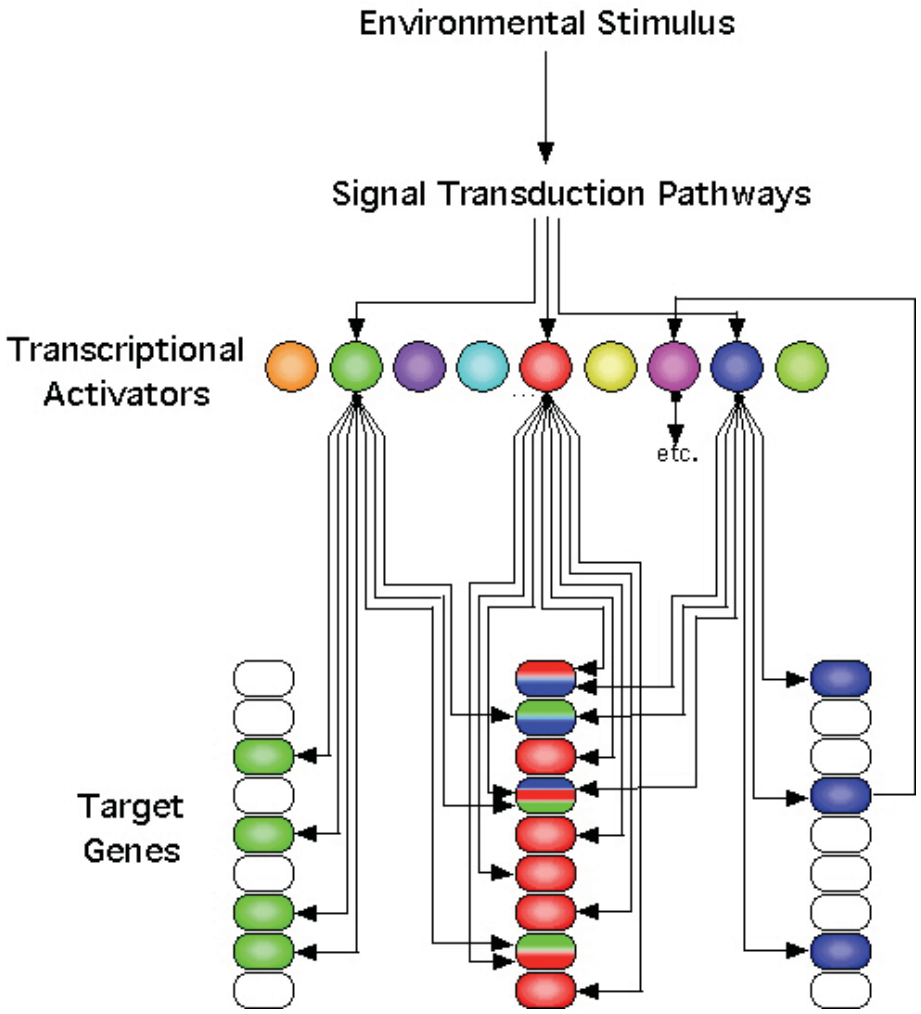


What are the gene regulators that control gene expression?
At what genes do these regulators operate?

Gene regulatory networks provide key insight into cellular function

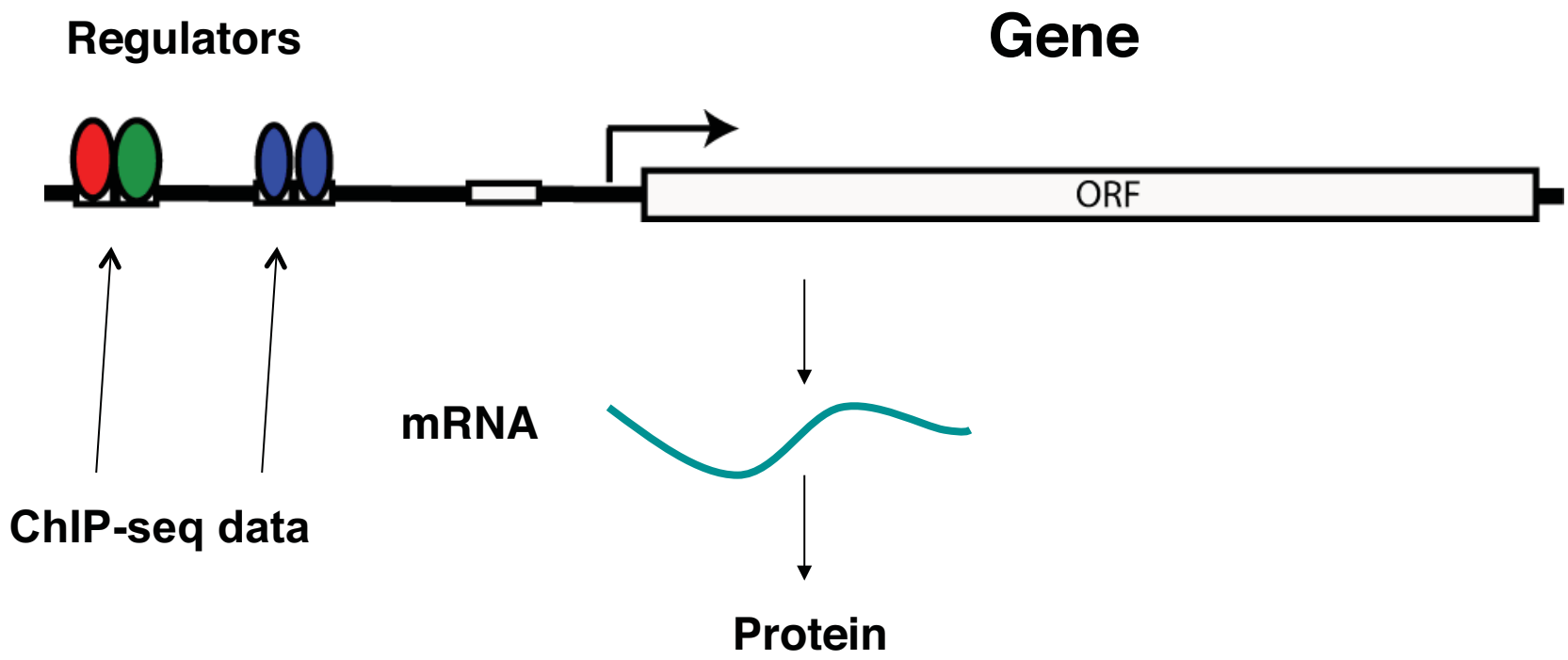
Transcriptional regulatory network information will:

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

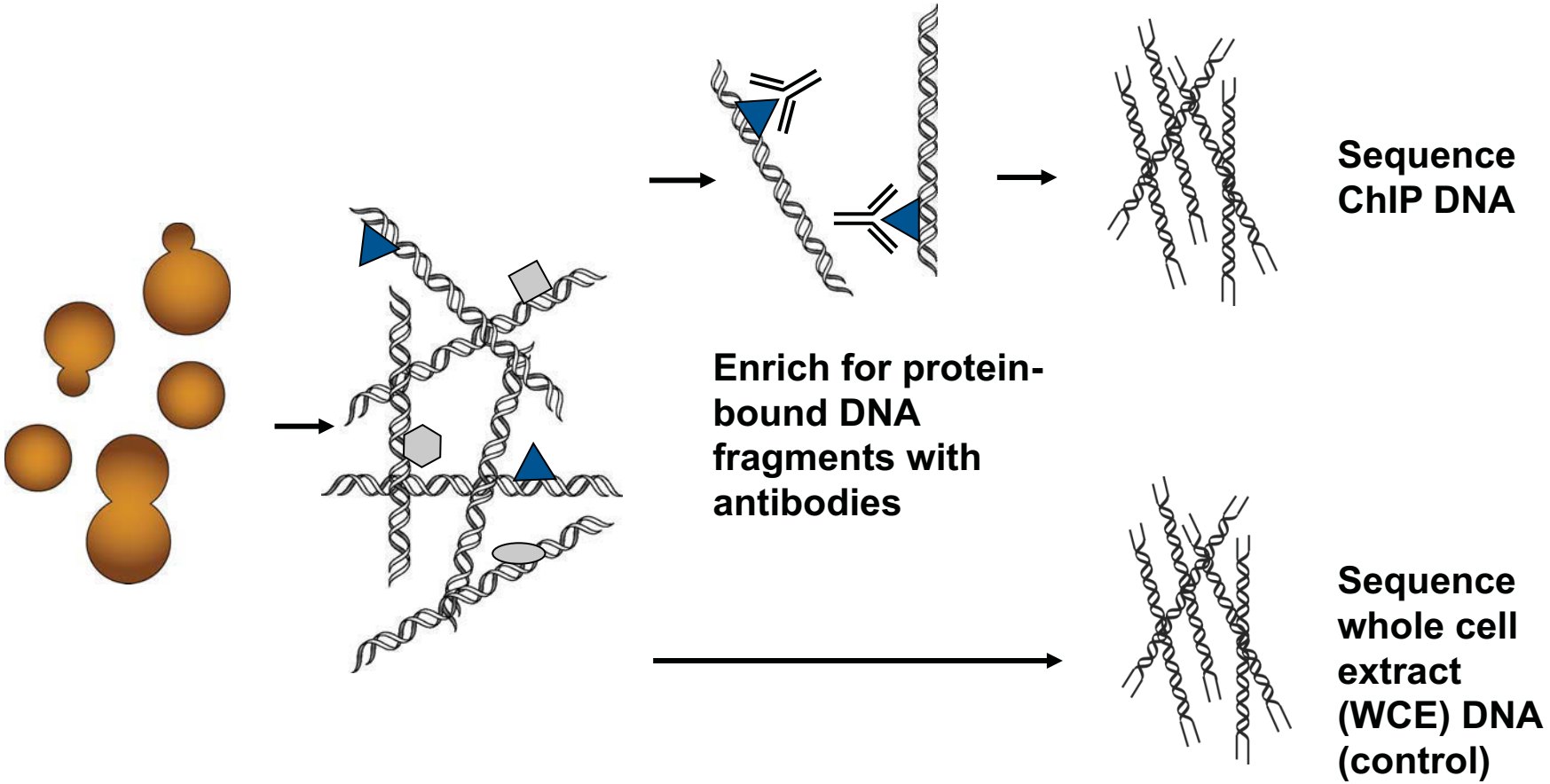


Courtesy of Richard Young. Used with permission.

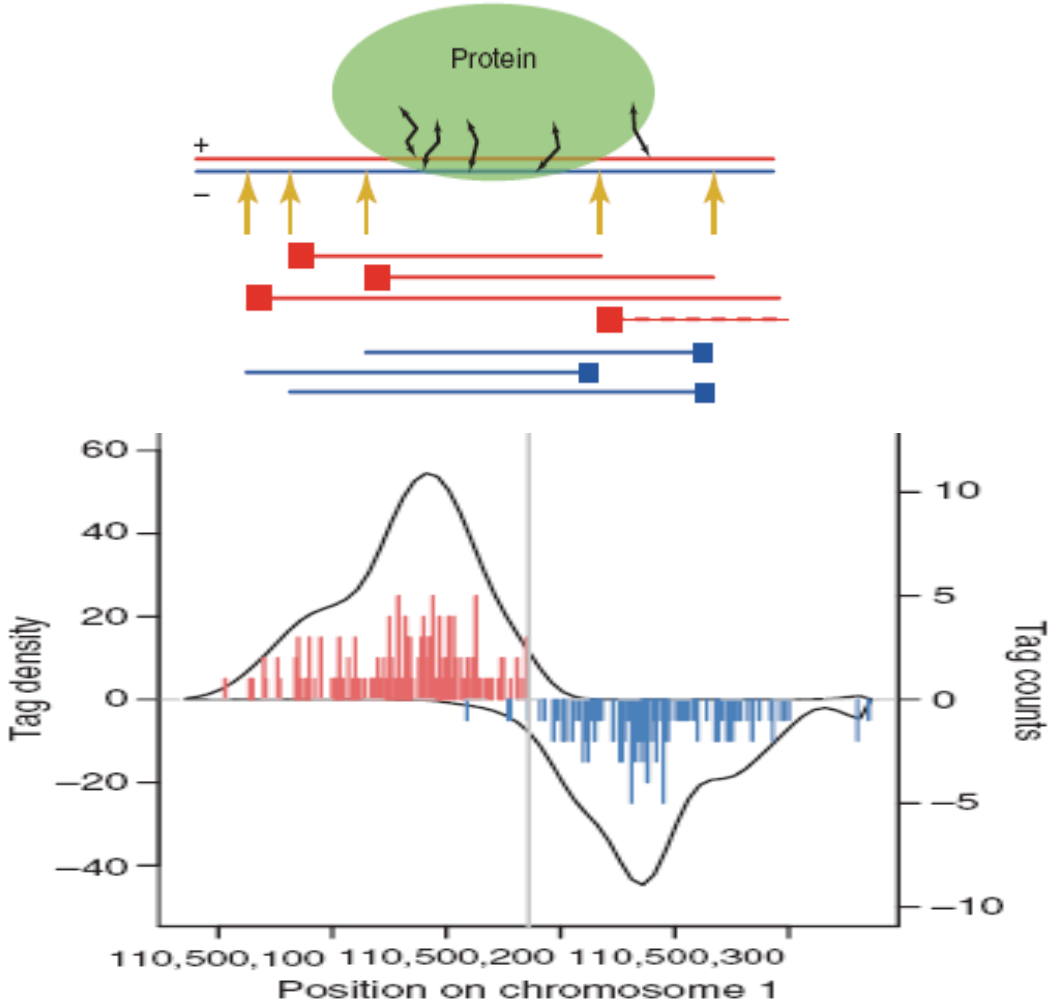
ChIP-seq data reveals where TFs bind to the genome



ChIP-seq protocol

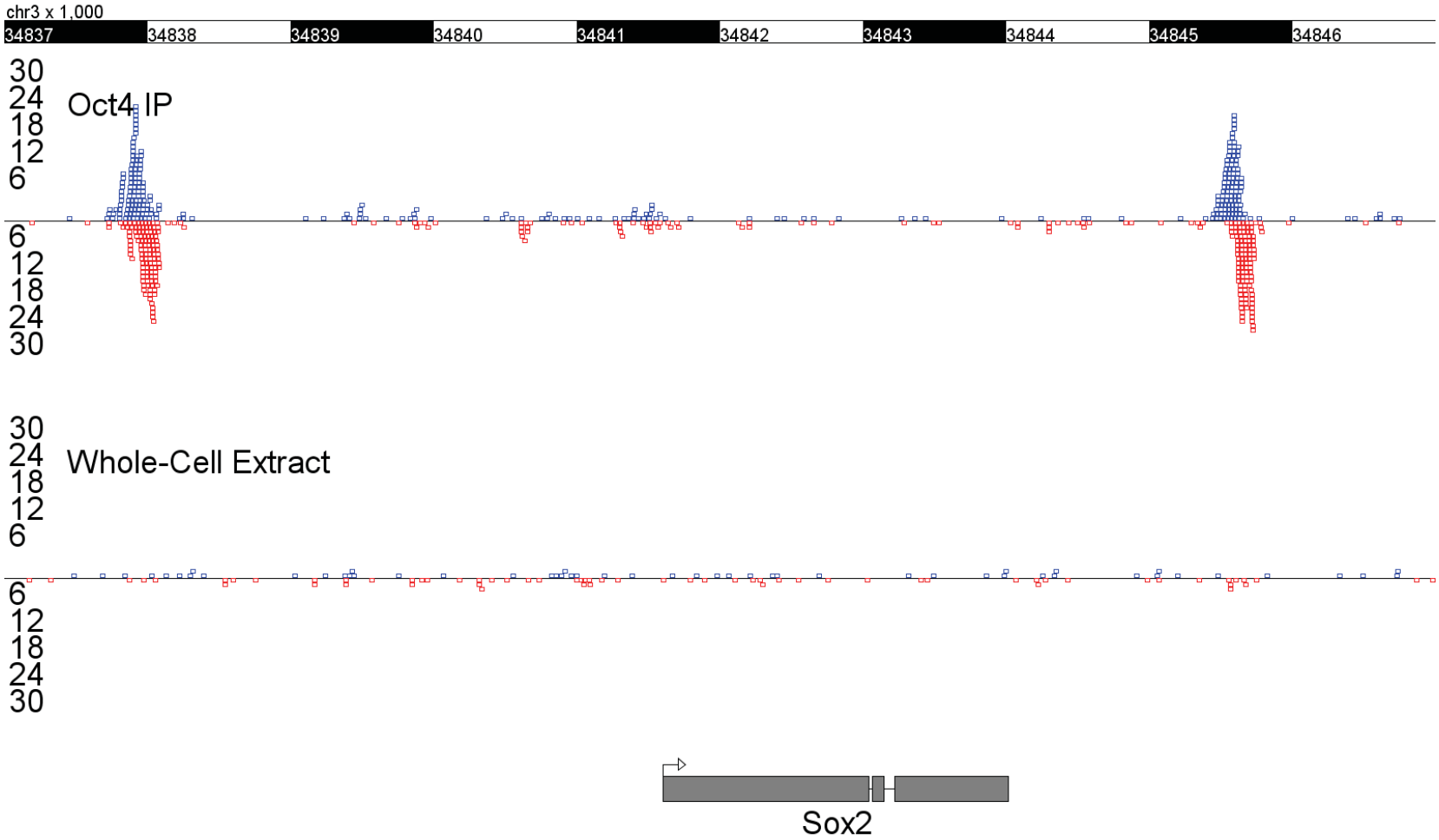


A binding event produces a distribution of reads around its site

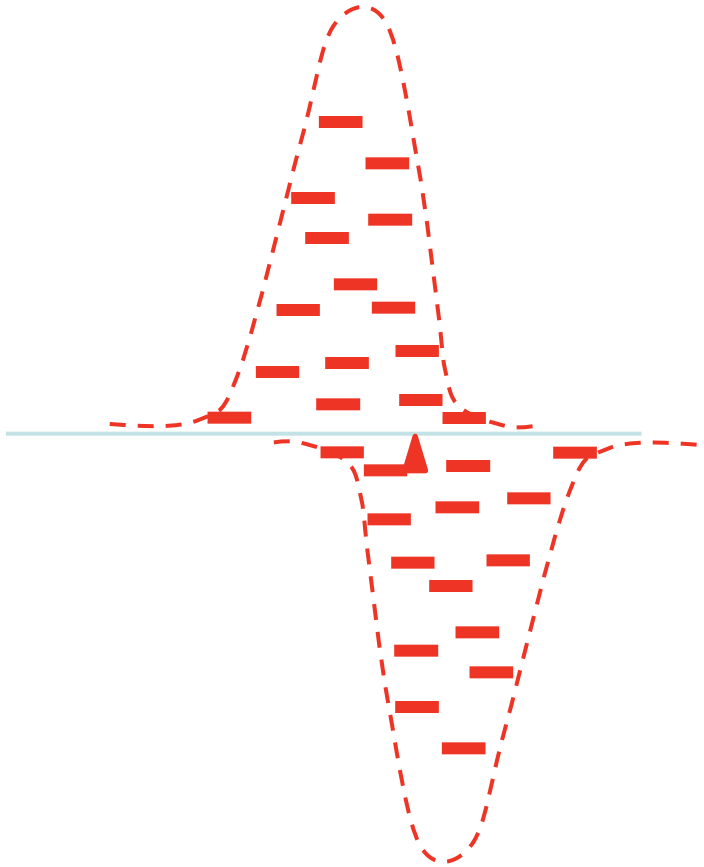


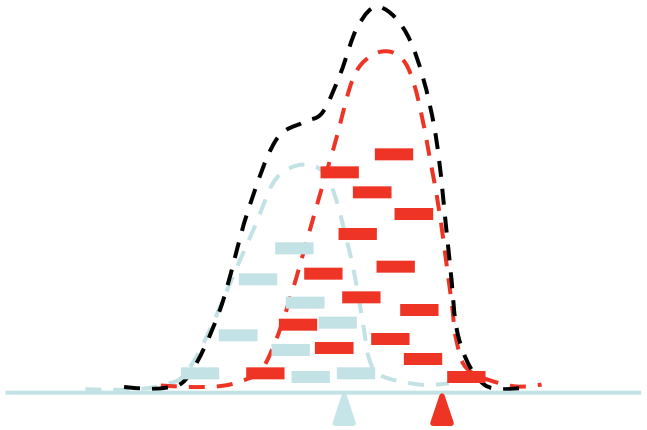
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Kharchenko, Peter V., Michael Y. Tolstorukov, et al. "Design and Analysis of ChIP-seq Experiments for DNA-binding Proteins." *Nature biotechnology* 26, no. 12 (2008): 1351-9.

Data from two binding events
mES cell Oct4 ChIP Seq



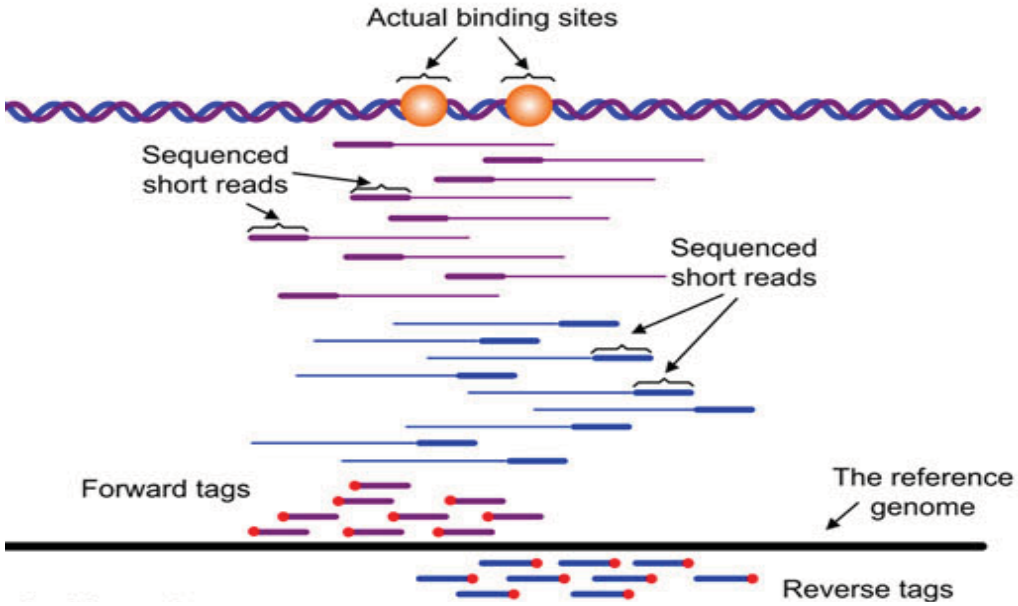
The spatial distribution of reads can be used to improve spatial resolution of prediction and de-convolve joint binding events





ChIP-Seq reads are independently generated from a set of spatially discrete binding events

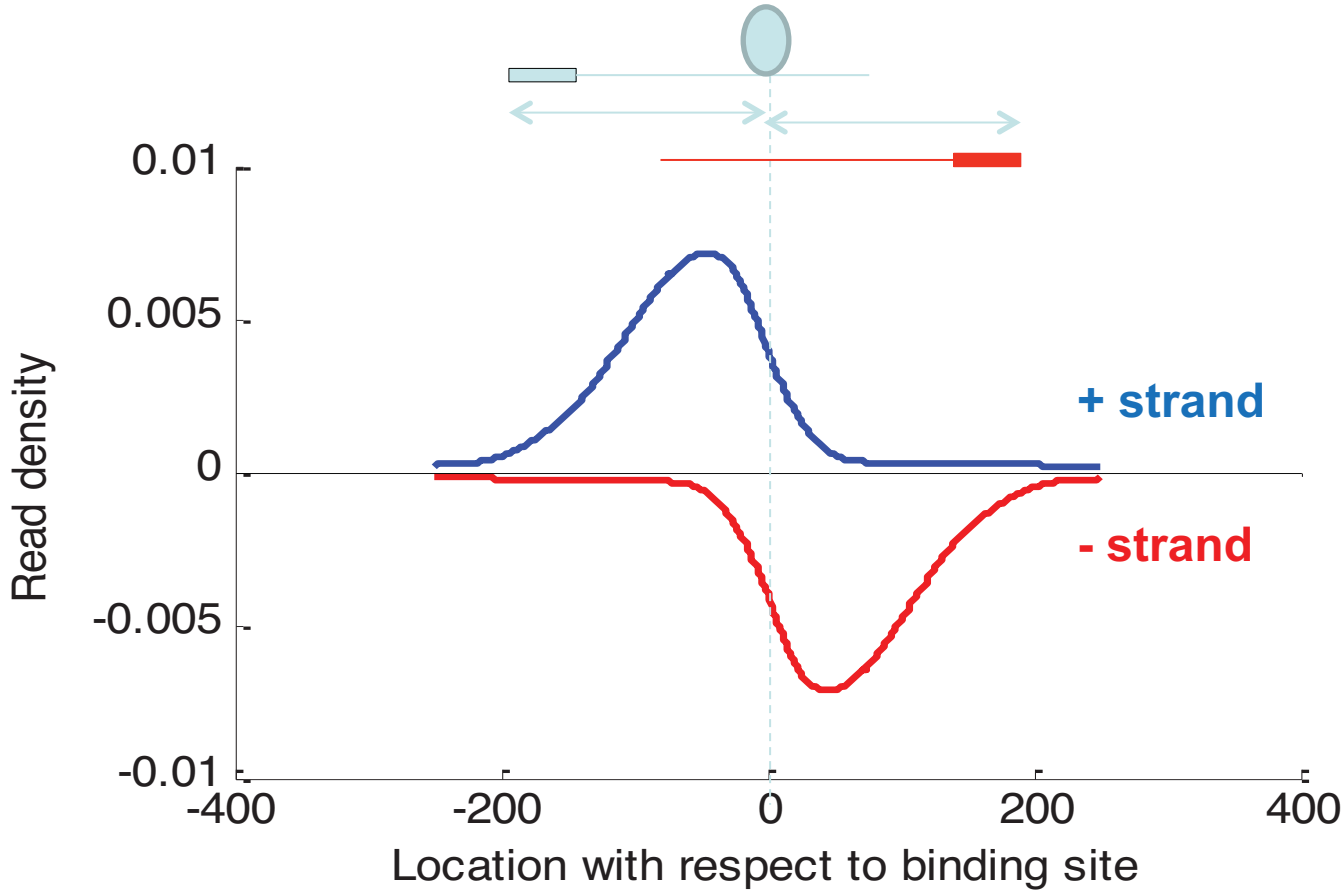
GPS addresses the challenges in ChIP-Seq analysis



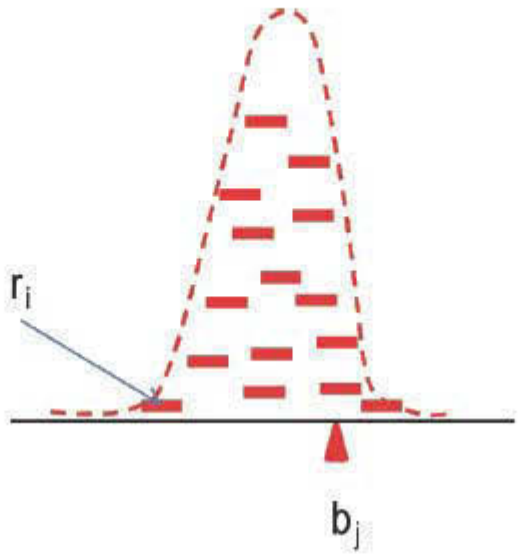
ChIP DNA are randomly fragmented → **Model the spatial distribution of the reads**

Mixture of Reads from different events → **Construct a mixture model**

GPS estimates the spatial distribution of the reads



GPS estimates the spatial distribution of the reads



$$p(r_i|b_j) = p(r_i|z_{ij} = 1) = emp((-1)^{s_i}(r_i - b_j))$$

r_i: a read at position r_i

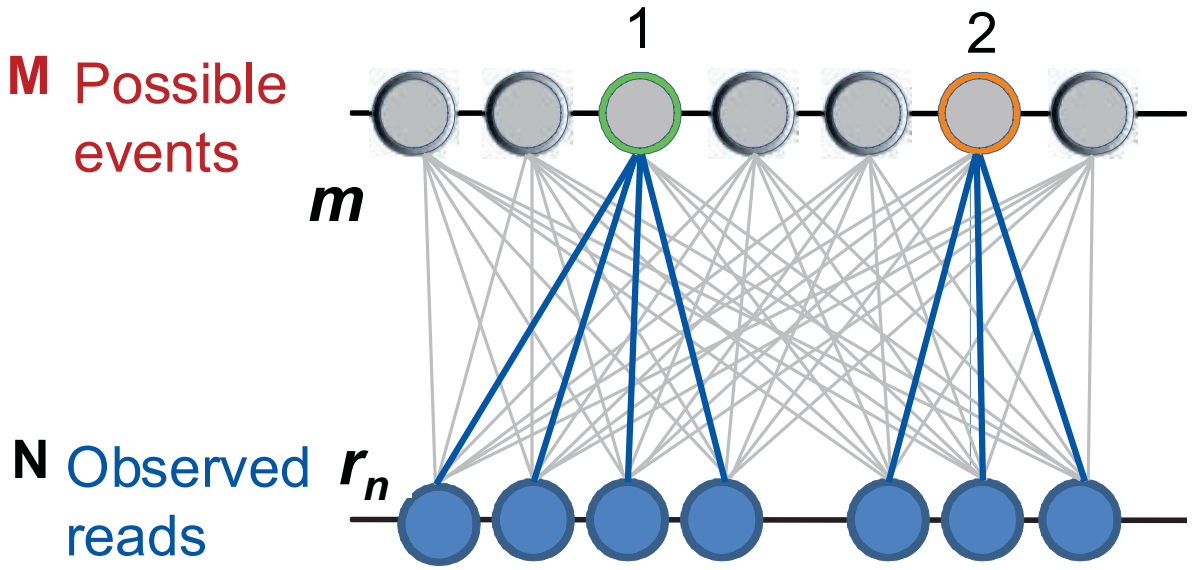
b_j: a binding event at position b_j

$emp(d)$: the empirical spatial distribution

S_i = 0 for forward strand
= 1 for reverse strand

GPS probabilistically models ChIP-Seq read spatial distribution using a mixture model

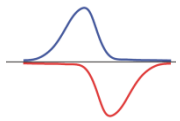
(single-base resolution)



Likelihood of observed reads

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

Prob. of event m
Mixing prob.



Likelihood of observed reads

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

Read assignment is latent

$g(z_n = m) = 1$ **Read n came from event m**

$g(z_n = m) = 0$ **Read n did not come from event m**

$$\pi = \arg\max_{\pi} p(R \mid \pi)$$

Expectation-Maximization (EM) algorithm with component elimination

E step

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

γ $(z_n=m)$ **: the fraction of read *n* assigned to event *m***

M step

$$\hat{\pi}_m^{(i)} = \frac{N_m}{\sum_{m'=1}^{M} N_{m'}}$$

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

N_m* : the effective number of reads assigned to event *m

Expectation-Maximization (EM) algorithm with component elimination

Initialization

$$\pi_j = \frac{1}{M}$$

Strength of binding event at end

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

N_m : the effective number of reads assigned to event *m*

Expectation-Maximization (EM) algorithm with component elimination

E step

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

$\gamma(z_n=m)$: the fraction of read *n* assigned to event *m*

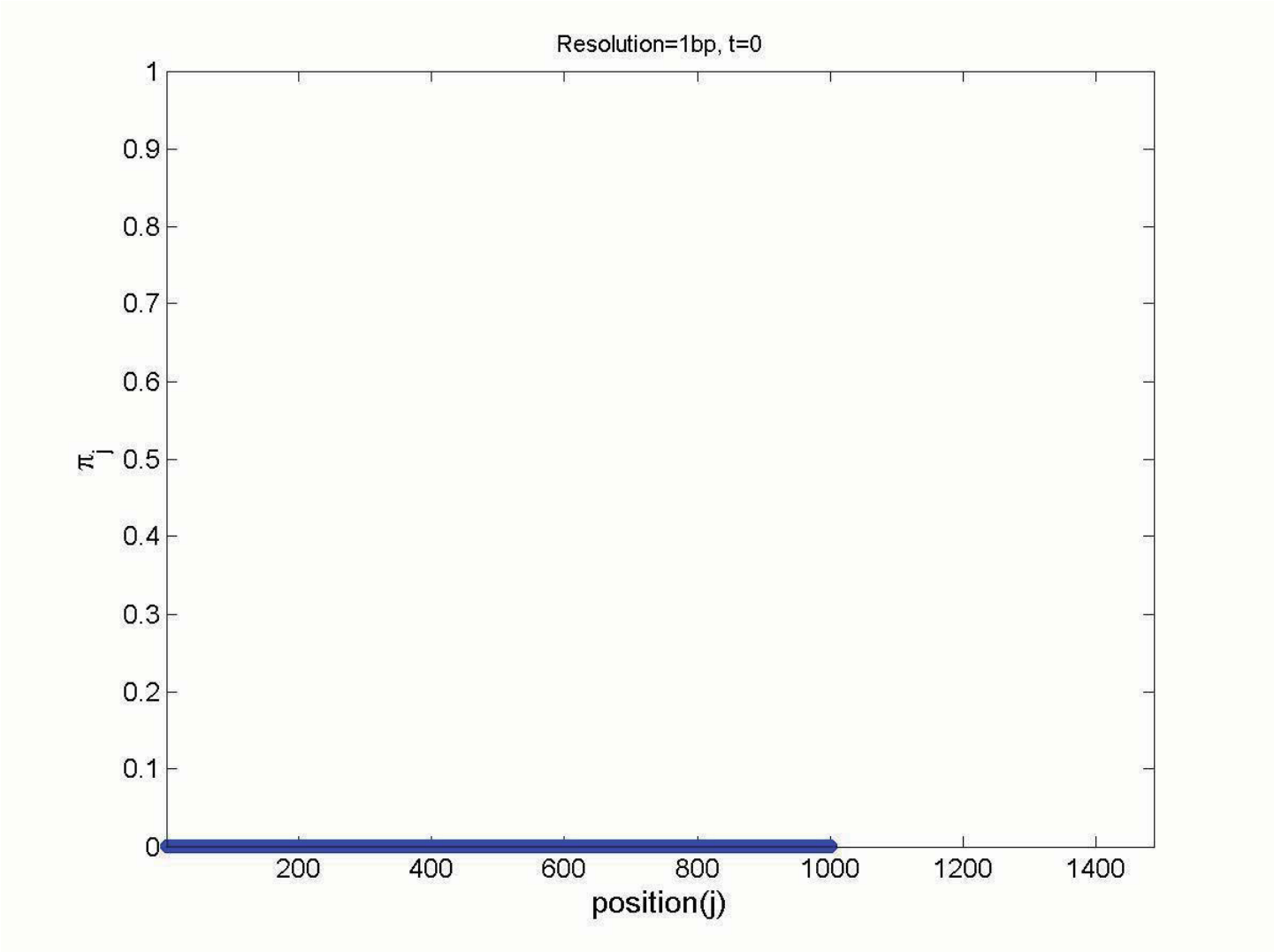
M step

$$\hat{\pi}_m^{(i)} = \frac{N_m}{\sum_{m'=1}^{M} N_{m'}}$$

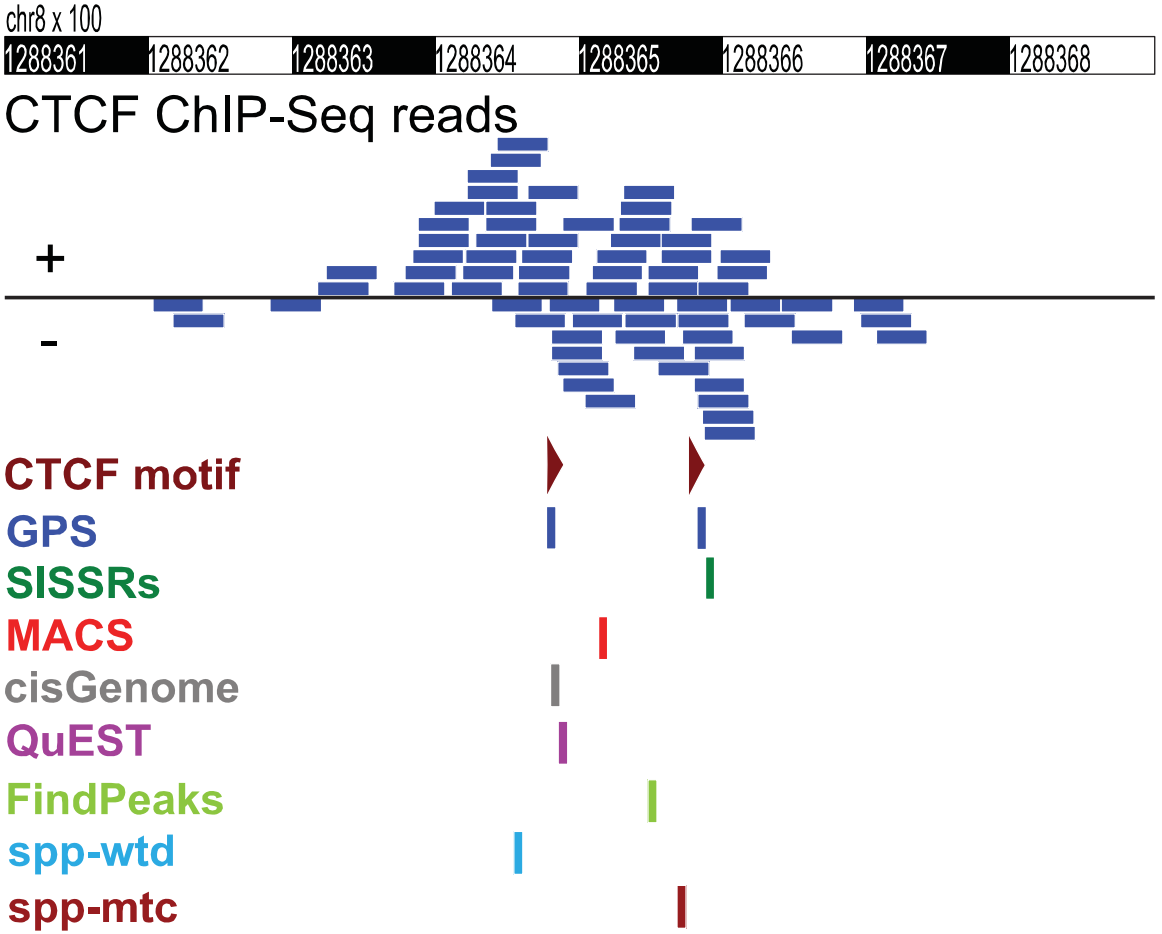
$$N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

N_m : the effective number of reads assigned to event *m*

Synthetic data, EM, no prior (events at 500 and 550 bp)



GPS deconvolves homotypic events and improves spatial accuracy



Example of a predicted joint CTCF event that contains coordinately located CTCF motifs

Likelihood of observed reads

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

A sparse prior on mixture components (binding events)

$$p(\pi) \propto \prod_{m=1}^{M} \frac{1}{(\pi_m)^{\alpha}}, \alpha > 0$$

(Figueiredo and Jain, 2002)

Expectation-Maximization (EM) algorithm with component elimination

E step

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

γ (z_n=m) : **the fraction of read *n* assigned to event *m***

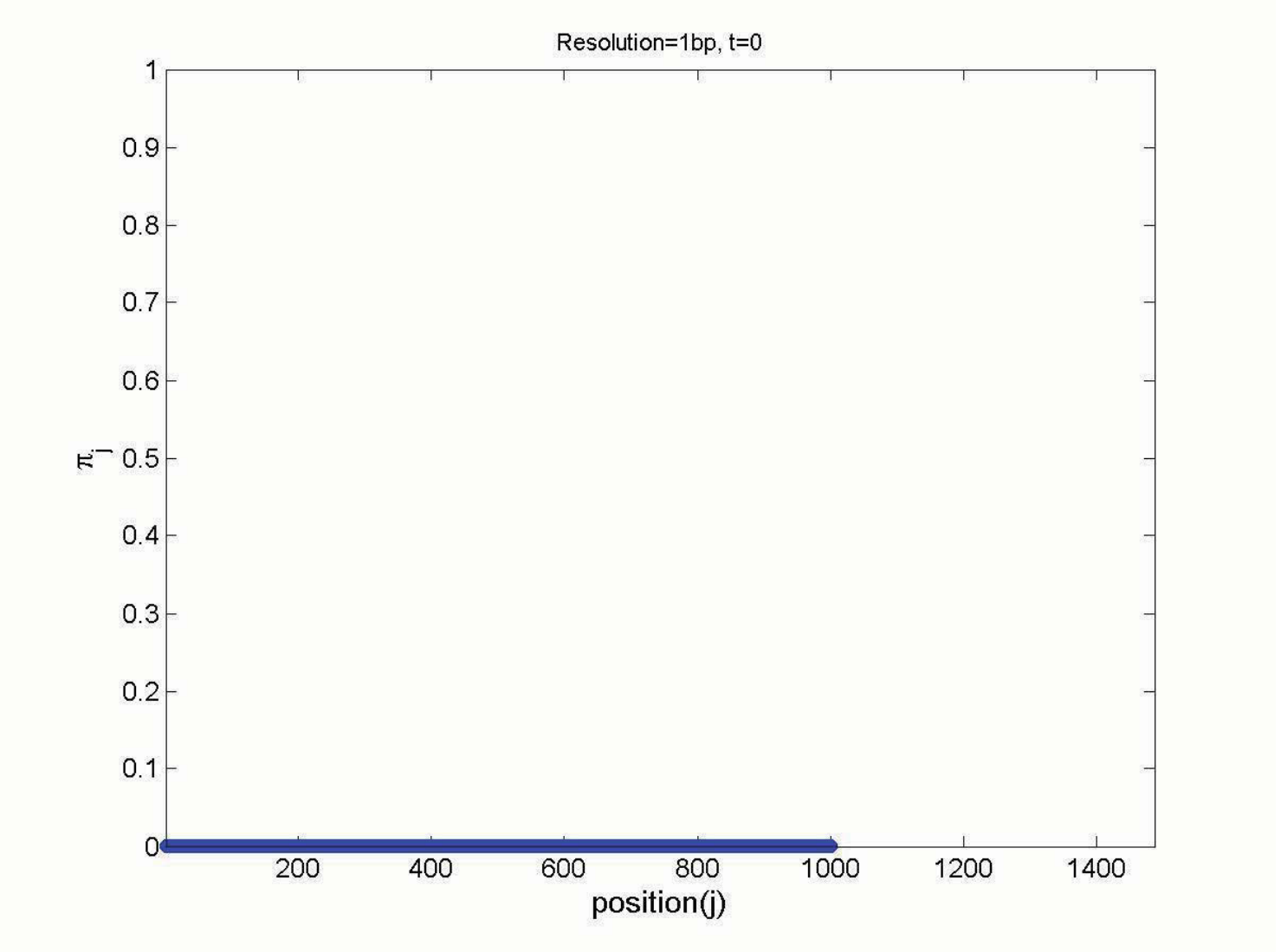
M step

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

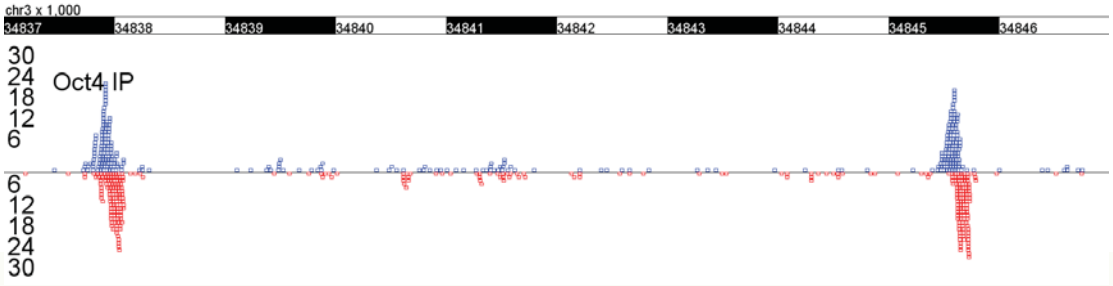
$$N_m = \sum_{n=1}^{N} \gamma(z_n = m)$$

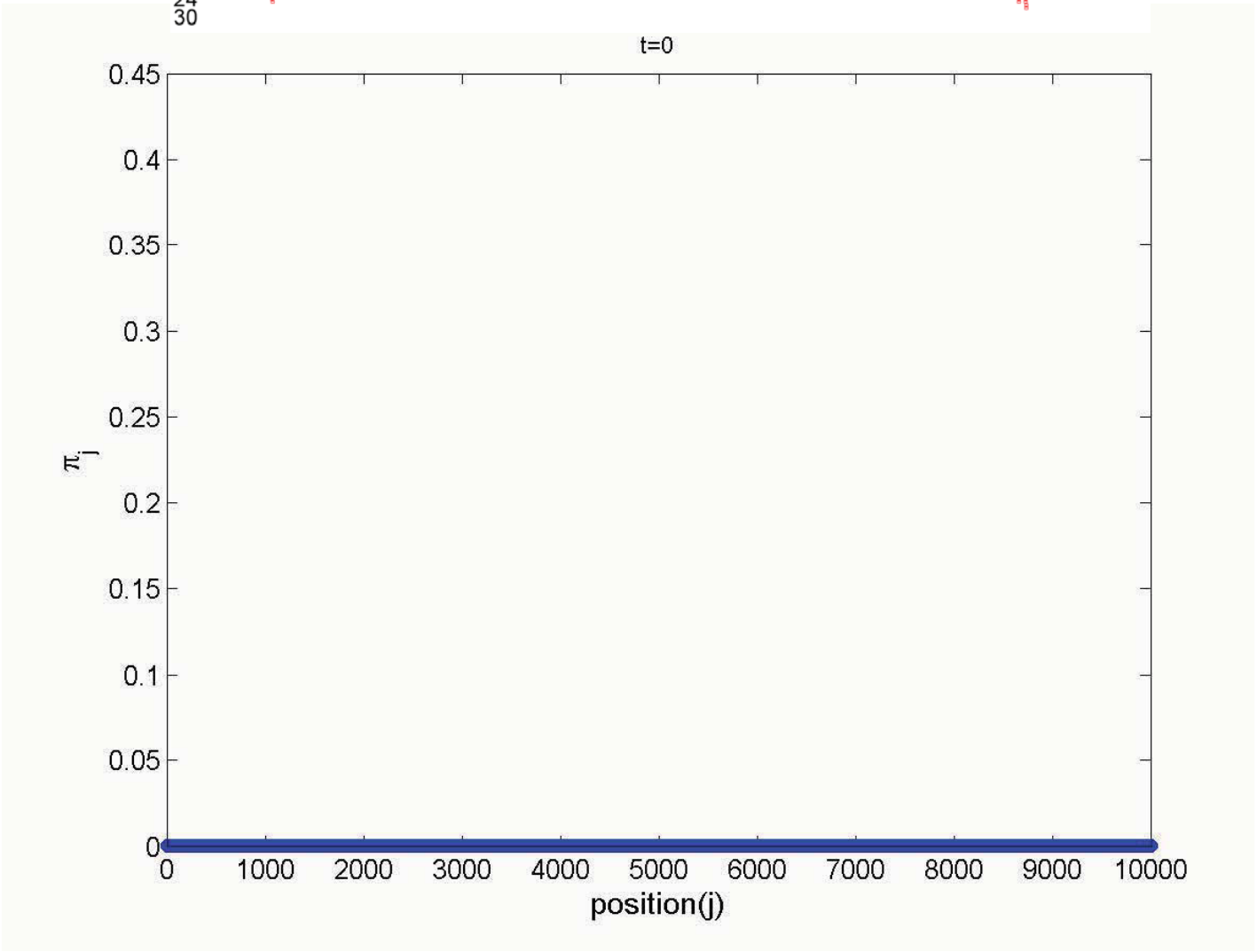
N_m* : the effective number of reads assigned to event *m

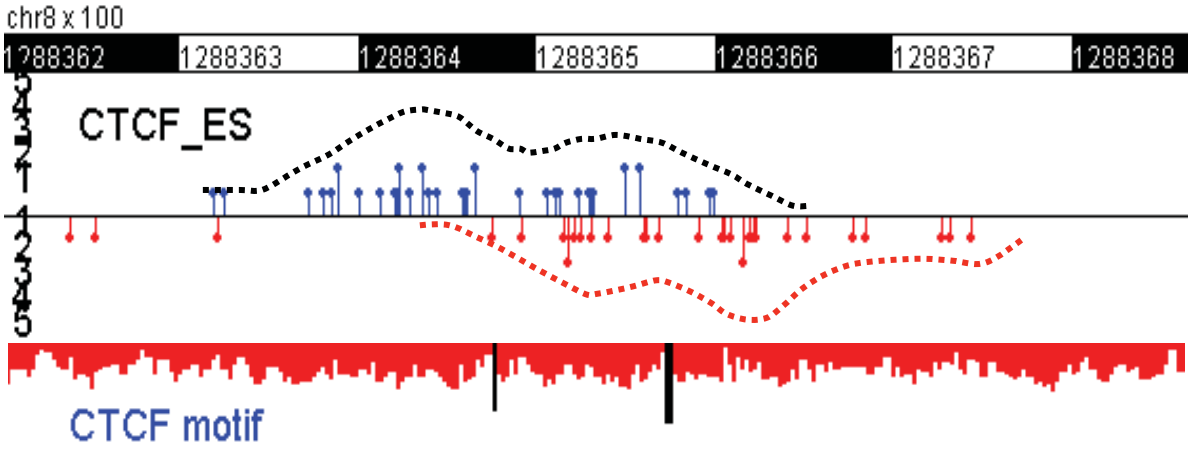
Synthetic data, EM, sparse prior (events at 500 and 550 bp)

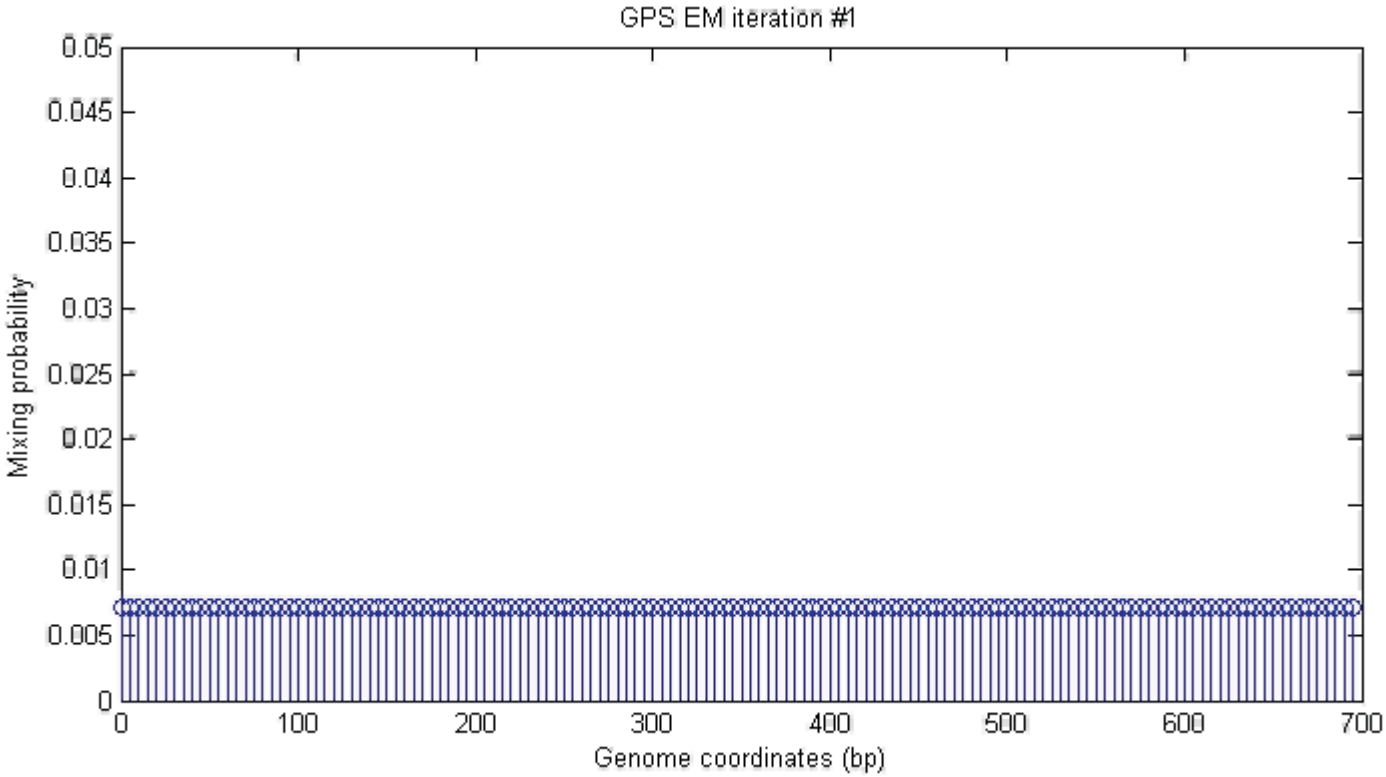


EM – Sparse prior

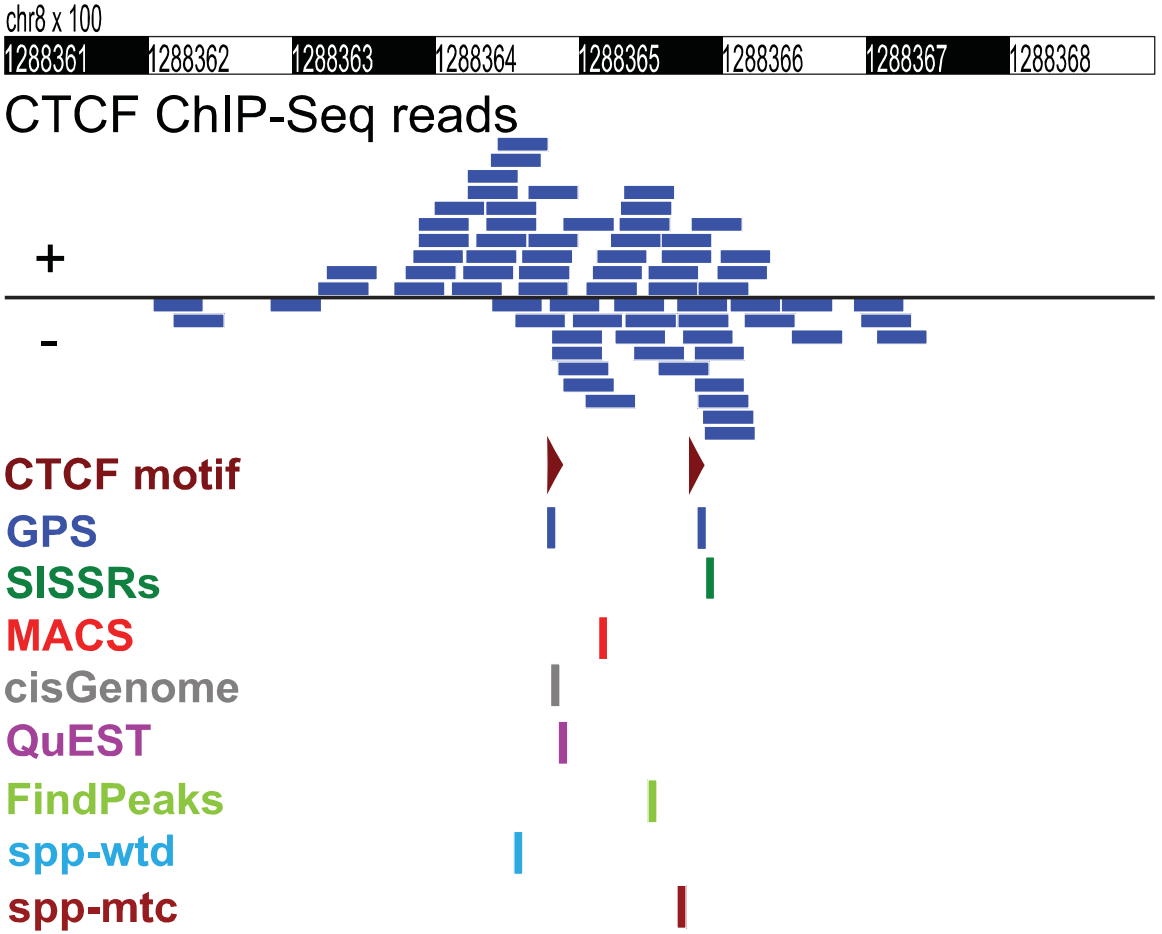





chr8 x 100
1288362
1288363
1288364
1288365
1288366
1288367
1288368
CTCF_ES
CTCF motif

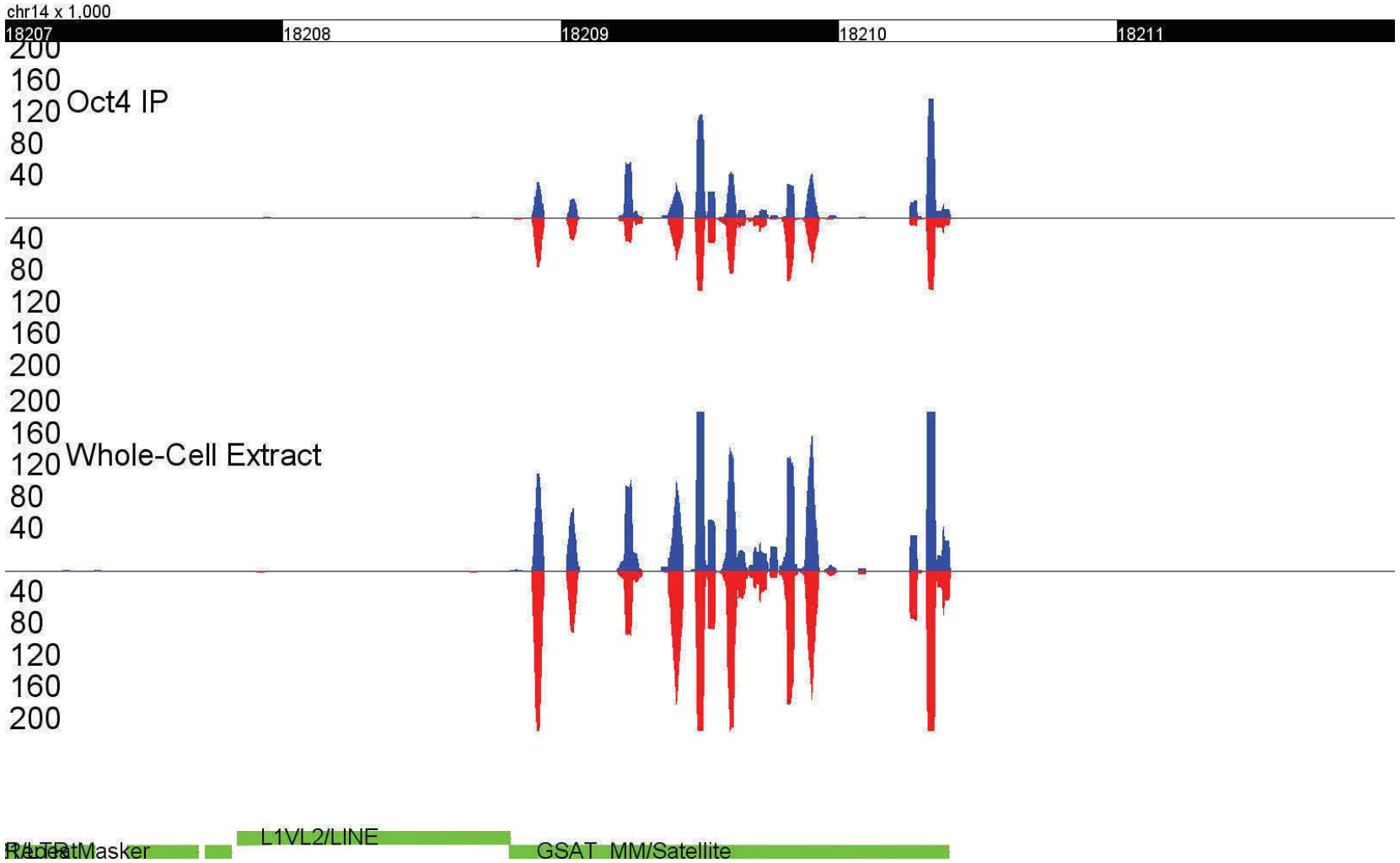

GPS EM iteration #1
Mixing probability
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
100
200
300
400
500
600
700
Genome coordinates (bp)

GPS deconvolves homotypic events and improves spatial accuracy



Example of a predicted joint CTCF event that contains coordinately located CTCF motifs

mES cell Oct4 ChIP Seq



We compute a p-value with a binomial test for significance

Null Model –

F(k,n,P) - Probability n-k reads observed in IP channel by chance with k reads observed in control. P = 0.5 equal chance reads occurred in control and IP channels for null model.

$$F(k,n,P) = \sum_{l=0}^{\lceil k \rceil} \binom{n}{l} P^l (1-P)^{(n-l)}$$

k: scaled control read count

n: total count of IP and scaled control reads

P: probability that reads occur from IP data, $P = 0.5$.

We determine significant events by Benjamini Hochberg at a desired false discovery rate (FDR)

Benjamini-Hochberg correction

$$Q - value = P - value \times \frac{Count}{Rank}$$

Count: total number of binding events tested.

Rank: Rank of event in list list of p-values, from most significant (rank = 1) to least (rank = Count)

Accept events (reject null) of rank = 1 .. k up to the point that the Q-value is greater than the desired FDR.

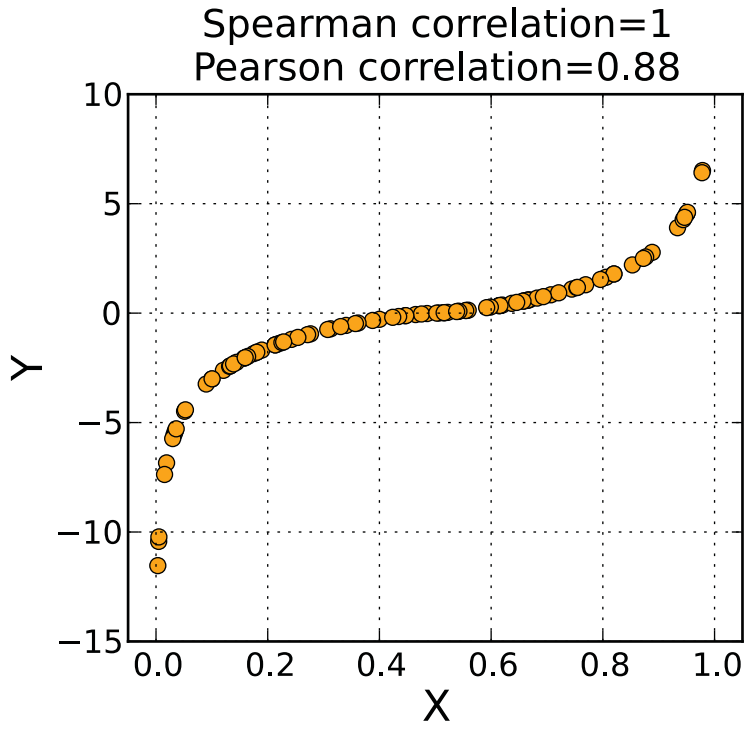
Irreproducible Discovery Rate (IDR) Analysis

- We have two replicates of an experiment
- How do we choose events are consistent in the two replicates?

Spearman's rank correlation provides a metric for replicate consistency but does not select events

- Consider two ranked lists of n detected events X and Y, one from each replicate, each ranked by scores from most significant to least significant.
- For matched event i ranks are x_i and y_i in X and Y

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$



Irreproducible Discovery Rate (IDR) Analysis

- $\Psi_n(t)$ is the fraction of the n events that are paired in the top n*t events in both X and Y It is roughly linear from t=0 to the point when events are no longer reproducible (not shared between replicates within the ranking)
- $\Psi'_n(t)$ is first derivative of $\Psi_n(t)$ with respect to t. It allows us to visualize when we transition from reproducible to irreproducible events as t increases

Irreproducible Discovery Rate (IDR) Analysis

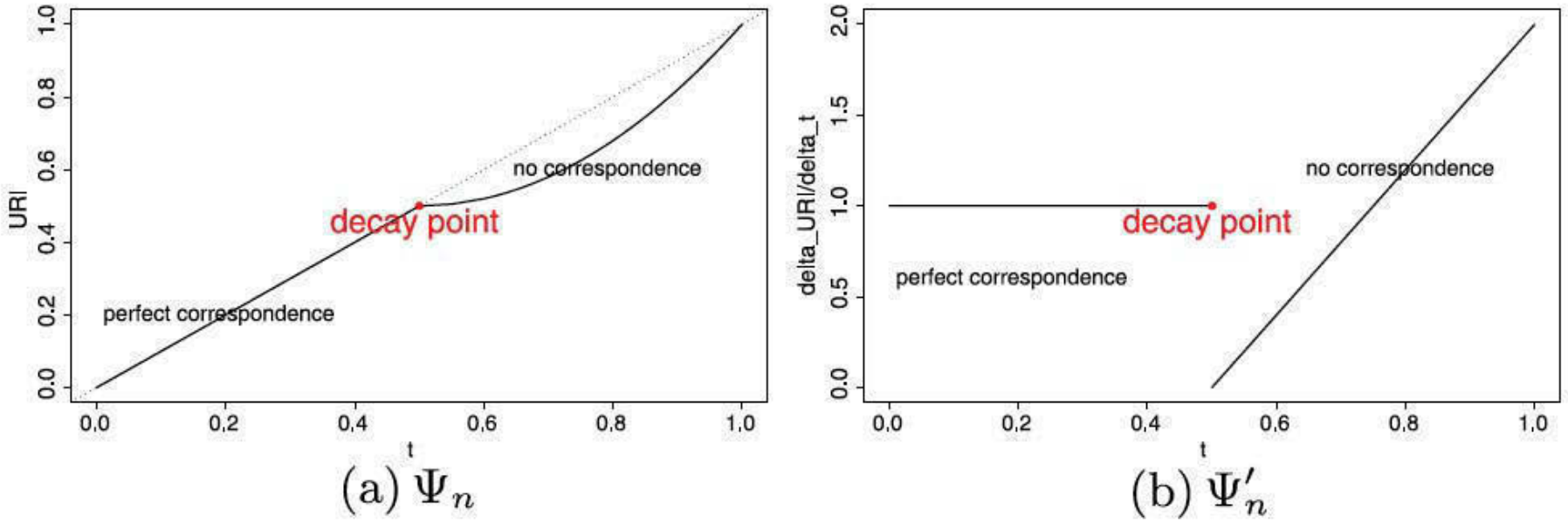


FIG. 1. *An illustration of the correspondence profile in an idealized case, where top 50% are genuine signals and bottom 50% are noise. In this case, all signals are ranked higher than noise; two rank lists have perfect correspondence for signals and no correspondence for noise.* (a) *Correspondence curve.* (b) *Change of correspondence curve.*

Source: Li, Qunhua, James B. Brown, et al. "Measuring Reproducibility of High-throughput Experiments." *The Annals of Applied Statistics* 5, no. 3 (2011): 1752-79.

MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS[1]

BY QUNHUA LI, JAMES B. BROWN, HAIYAN HUANG AND PETER J. BICKEL
University of California at Berkeley

Irreproducible Discovery Rate (IDR) Analysis

- Consider that the lists X and Y are a mixture of two kinds of events – reproducible and irreproducible.
- Model the ranking scores as a two component mixture and learn the parameters of the reproducible and irreproducible components
- For IDR α, select top l pairs using their scores such that the probability that the rate of pairs from the irreproducible part of the mixture is α

Irreproducible Discovery Rate Results

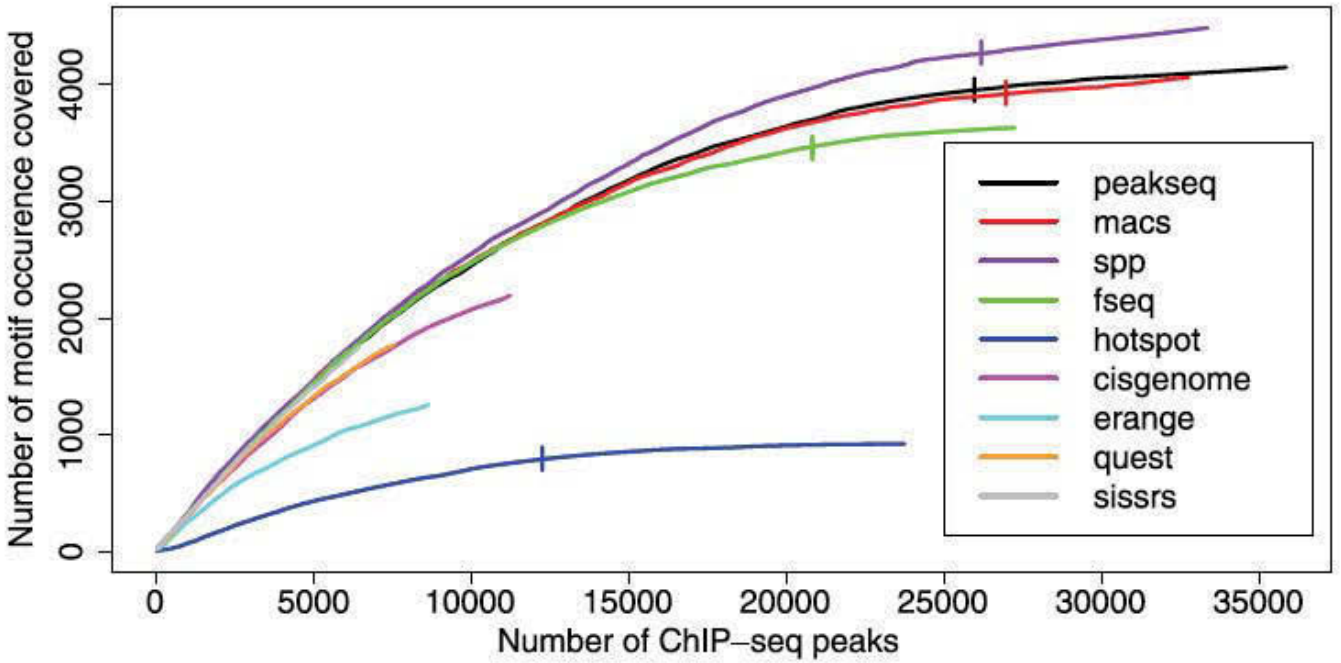


FIG. 7. *The coverage of high-confidence CTCF motif at different numbers of selected ChIP-seq peaks, plotted at various idr cutoffs for nine peak callers on a CTCF Chip-seq experiment from ENCODE. The bars on the curves of Peakseq, MACS, SPP, Fseq and Hotspot show the number of peaks selected at IDR = 0.05. No selection is made for the rest of the peak callers because model selection favors the one-component model for peaks identified by these callers.*

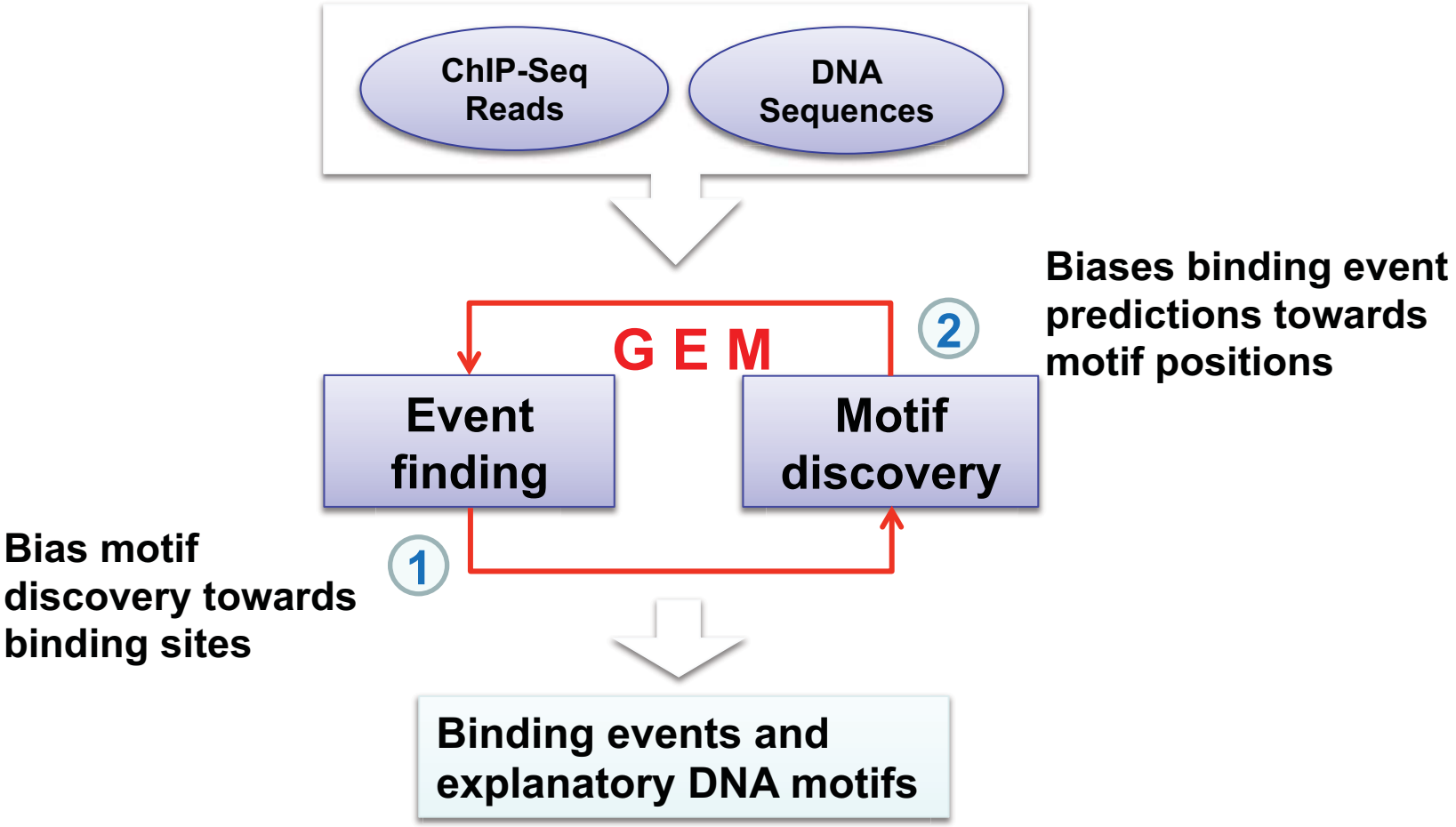
Source: Li, Qunhua, James B. Brown, et al. "Measuring Reproducibility of High-throughput Experiments." *The Annals of Applied Statistics* 5, no. 3 (2011): 1752-79.

MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS[1]

BY QUNHUA LI, JAMES B. BROWN, HAIYAN HUANG AND PETER J. BICKEL

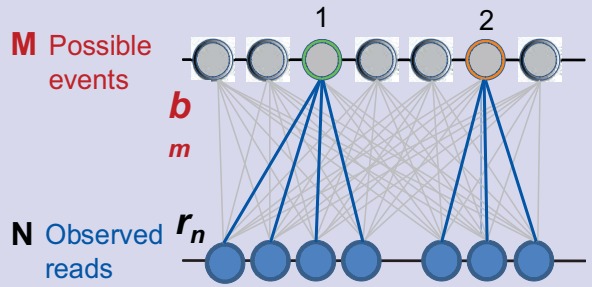
University of California at Berkeley

Genome-wide **E**vent finding and **M**otif discovery



Motif-based positional prior biases the binding event prediction

Mixture model



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

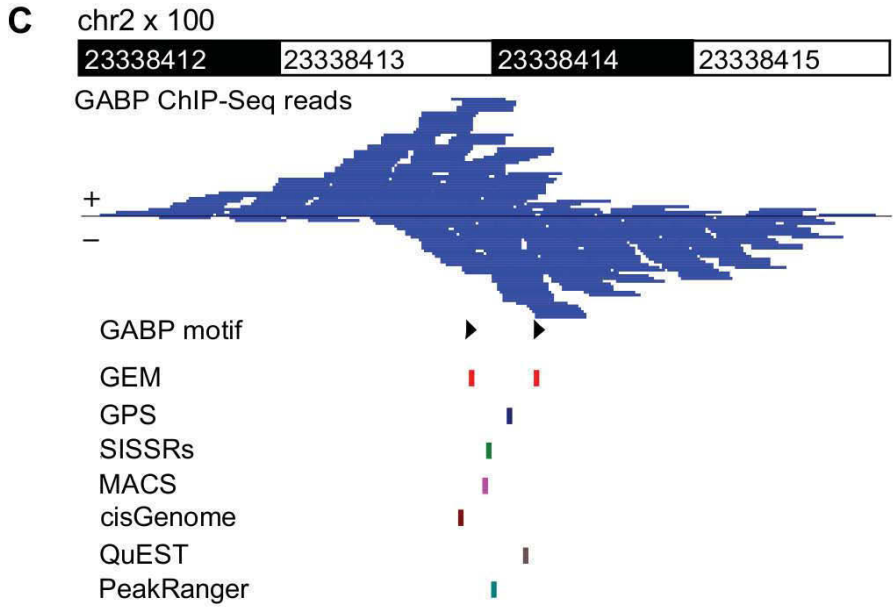
Position-specific priors

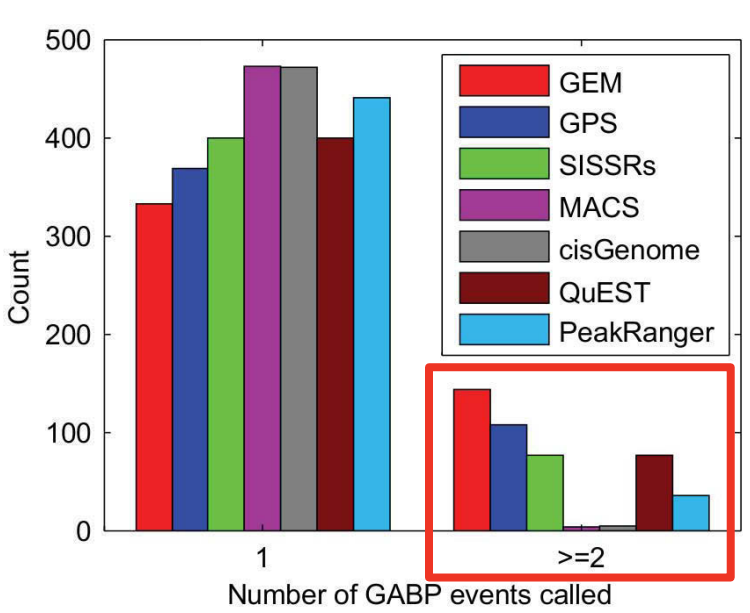
- **Events are sparse**
- **Events occurs more likely at motif positions**

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

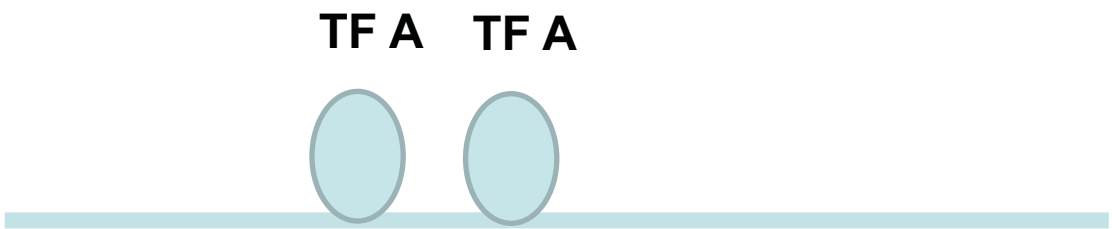
α_s : uniform sparse prior parameter governing the degree of sparseness, $\alpha_s > 0$;
α_m : position specific motif-based prior

GEM improves in resolving joint binding events

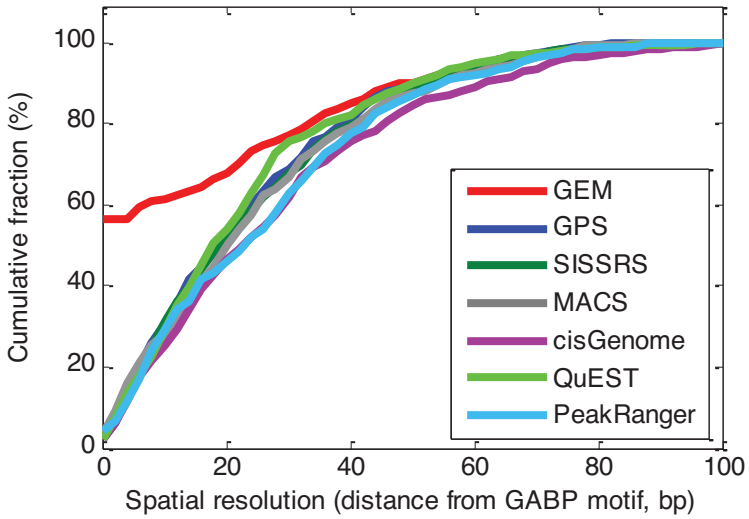




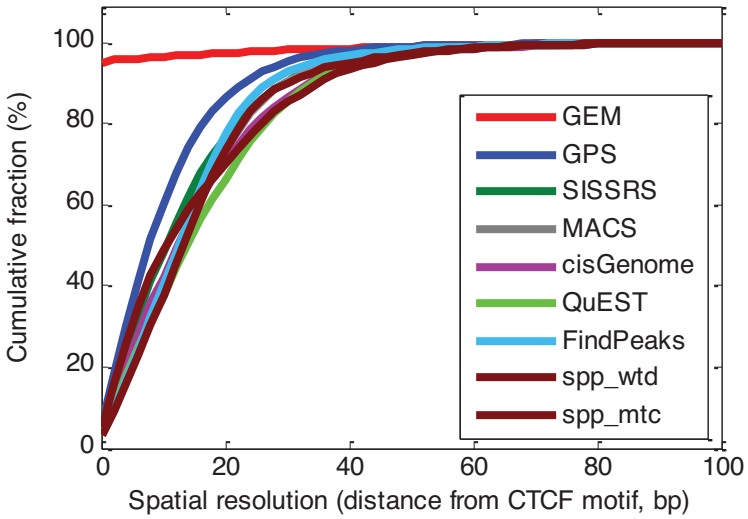
(Human GABP Data : Valouev et al., 2008)



GEM improves spatial accuracy in binding event prediction



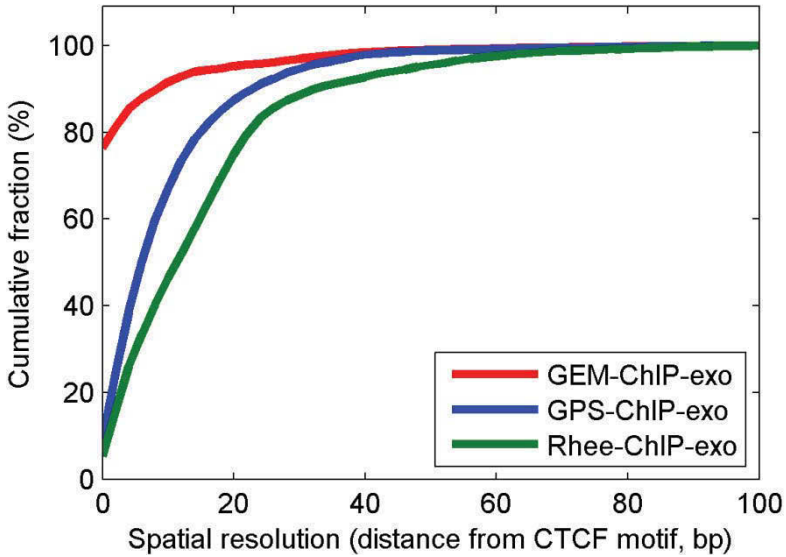
(Human GABP Data Valouev et al., 2008)

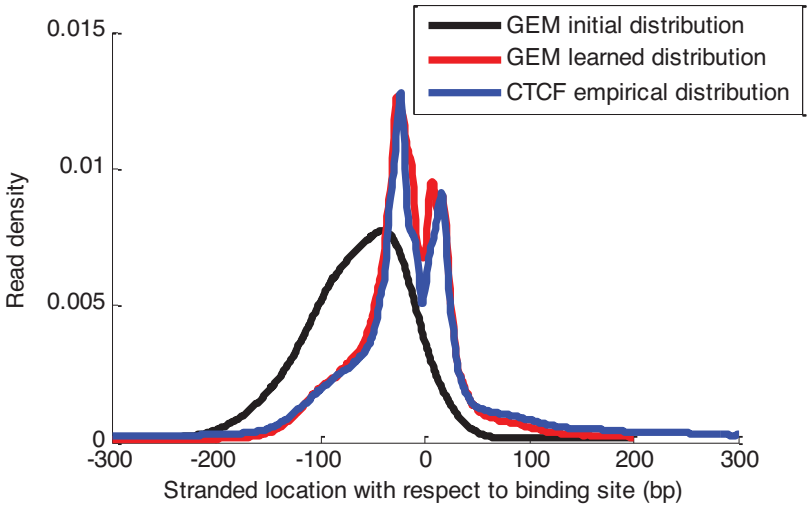


(Mouse CTCF data Chen , et. al. 2008)

Motif **Event call**

GEM improves the spatial resolution of ChIP-exo data event prediction



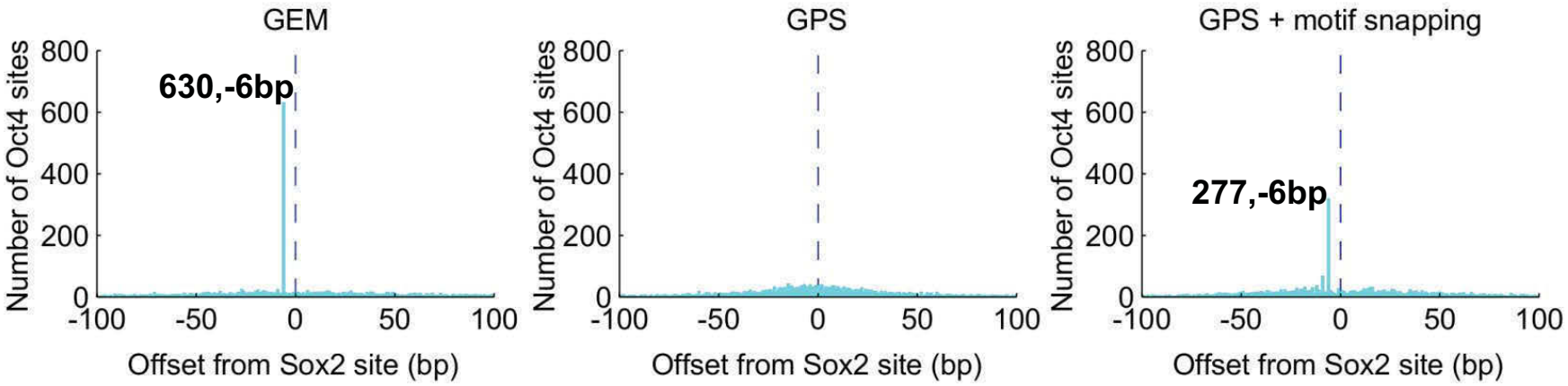


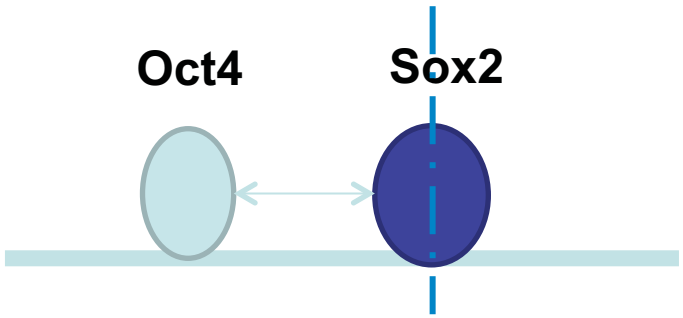
Motif **Event call**

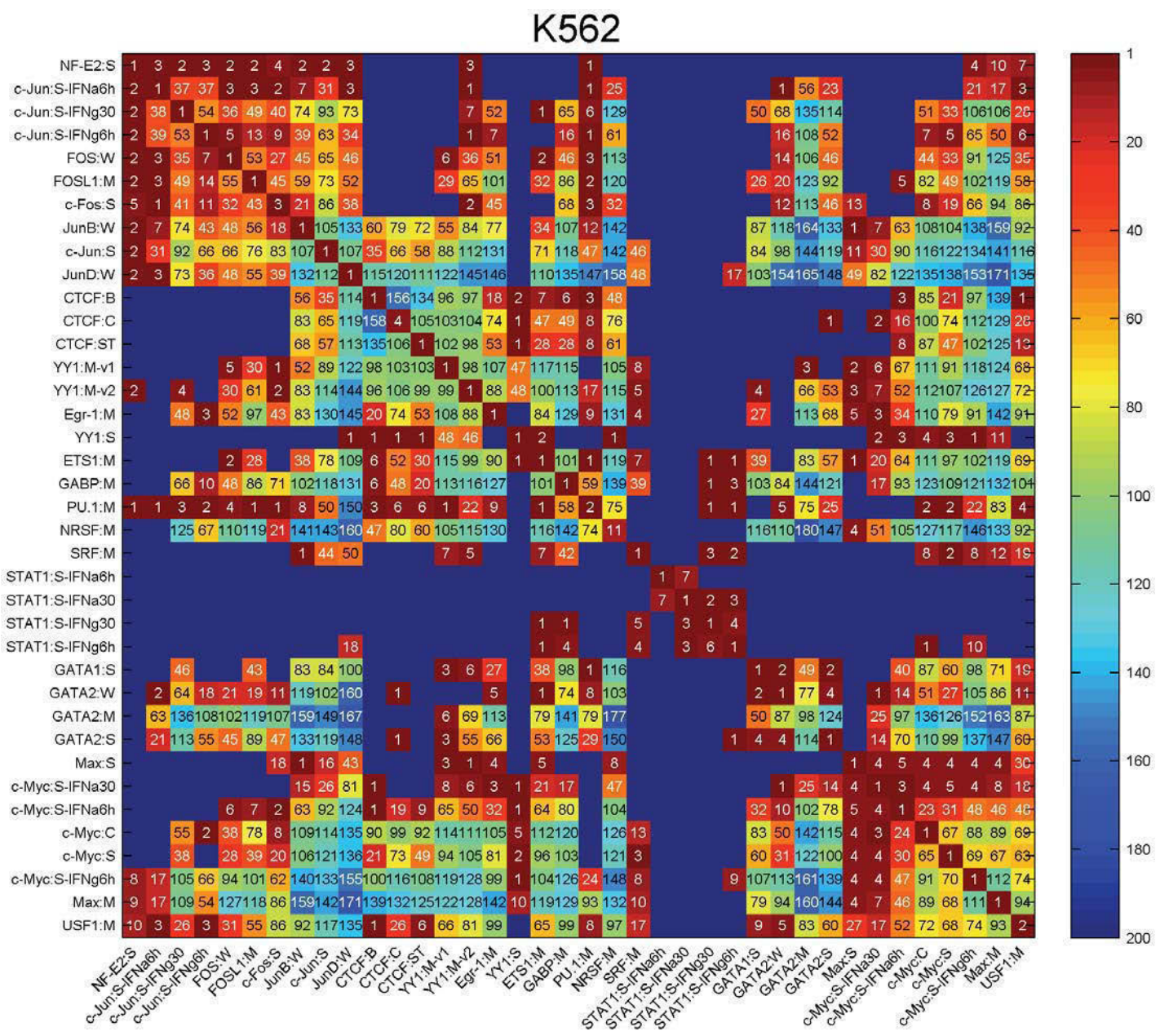
(Rhee and Pugh, 2011)

GEM reveals transcription factor spatial binding constraints

Total ~7500 Oct4 sites, ~2500 sites are within 100bp of Sox2 sites





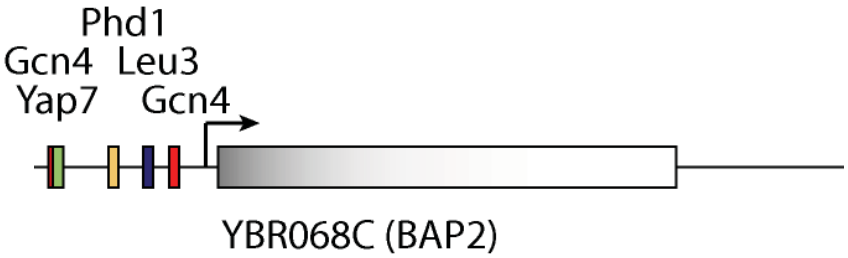


Courtesy of PLoS Computational Biology. License: CC-BY.
Source: Guo, Yuchun, Shaun Mahony, et al. "High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints." *PLoS Computational Biology* 8, no. 8 (2012): e1002638.

GEM Summary

- GEM incorporates motif information as a position-specific prior to bias binding event prediction
- GEM achieves exceptional spatial resolution, and further improves joint event deconvolution
- GEM systematic analysis reveals *in vivo* transcription factor spatial binding constraints in human and mouse cells, provides testable models for transcription factor interactions

Concept of a Transcriptional Regulatory Code



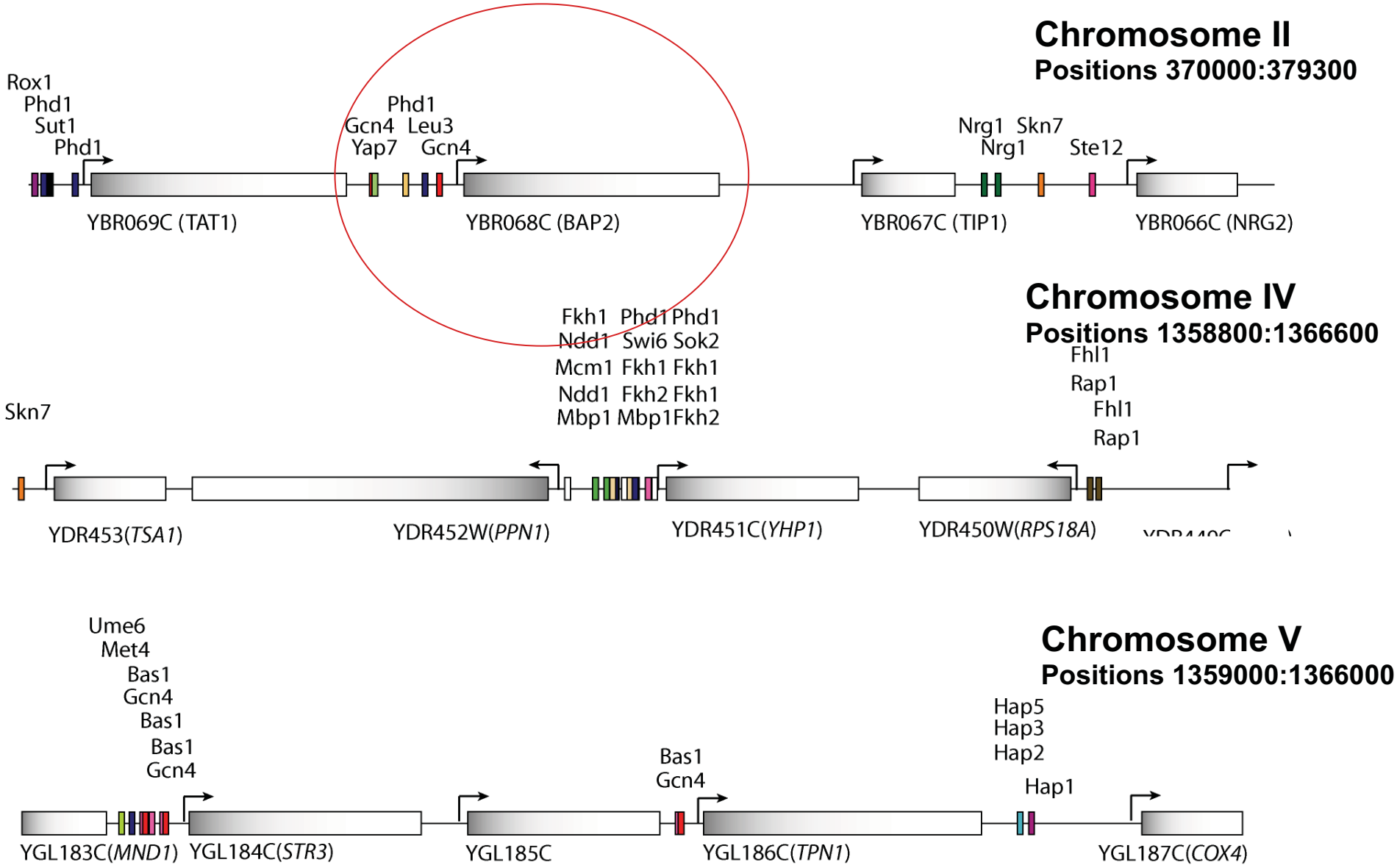
Harbison et al., Nature 431: 99 (2004)

What regulators contribute to control of each gene?

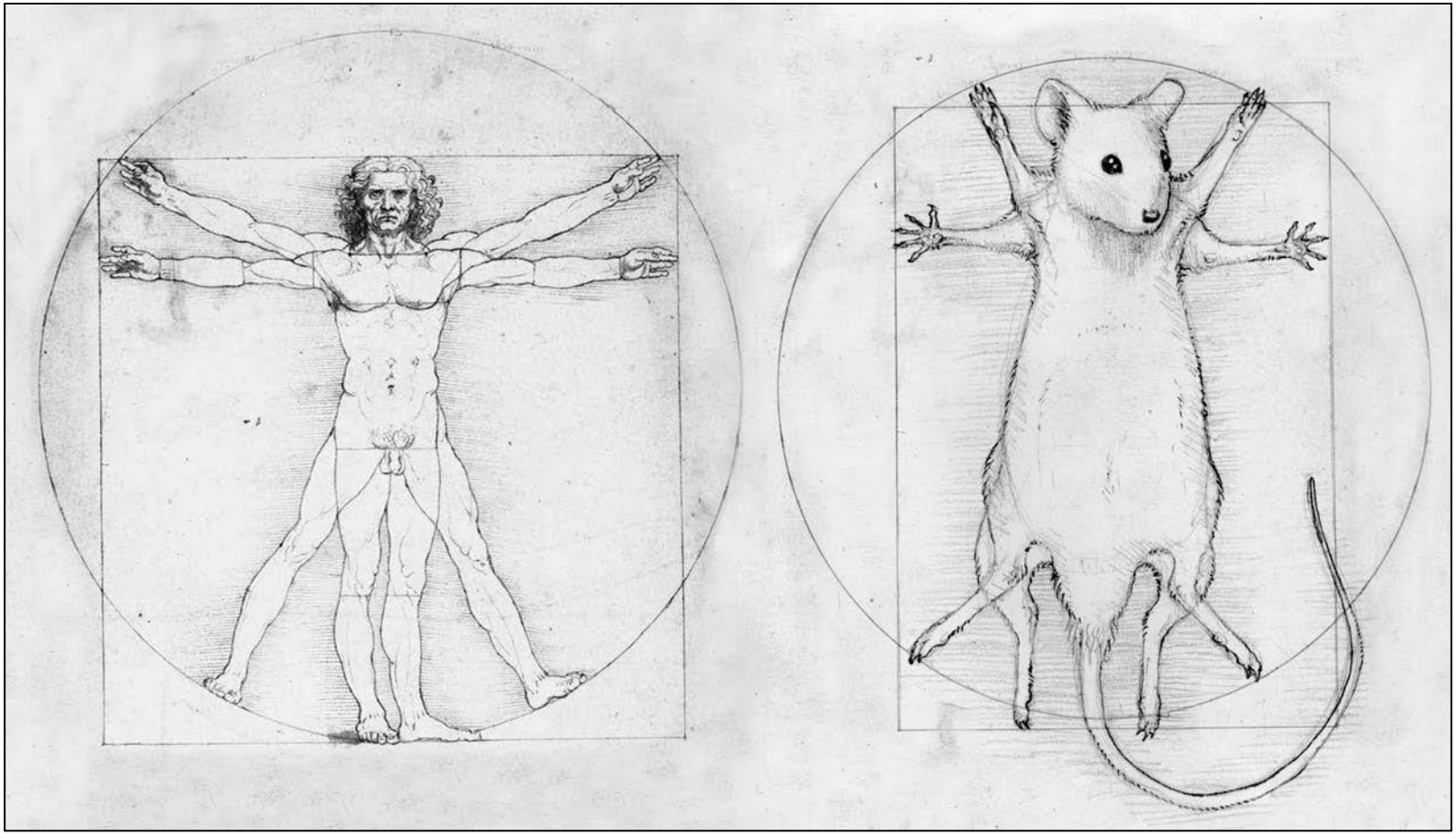
What sequences do they bind (cis-elements)?

When do the regulators bind these sequences?

Samples of the Draft Transcriptional Regulatory Coc

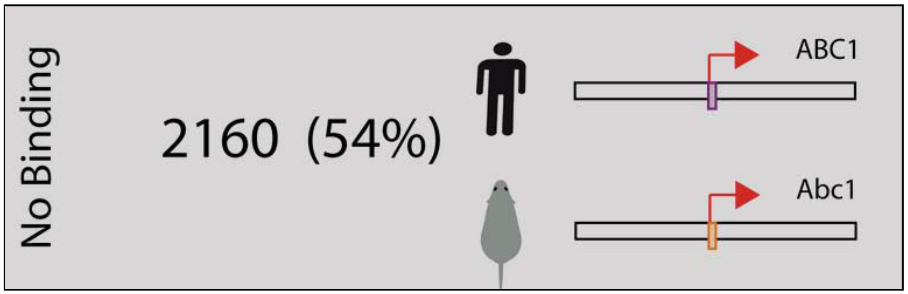


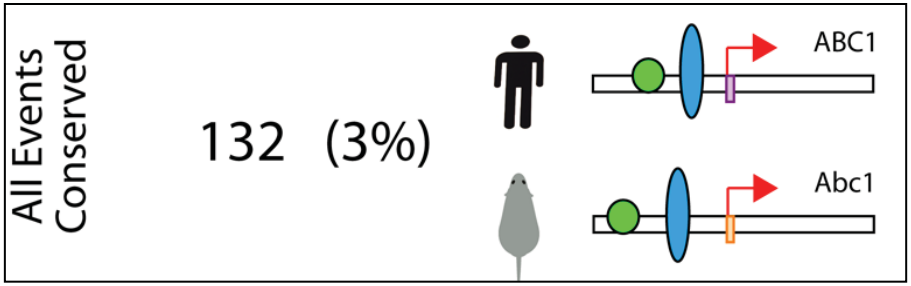
Is conservation a good predictor of conserved binding events across species?

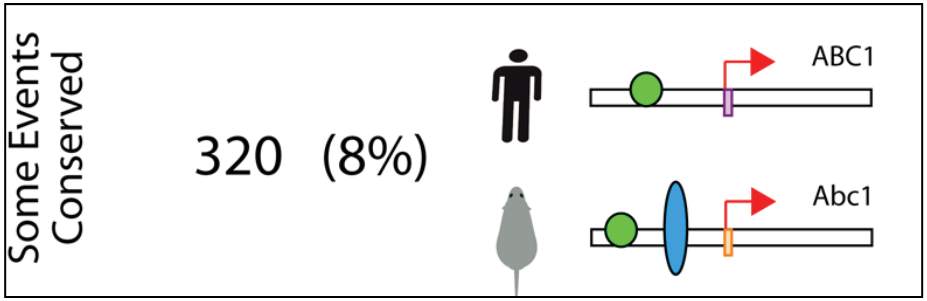


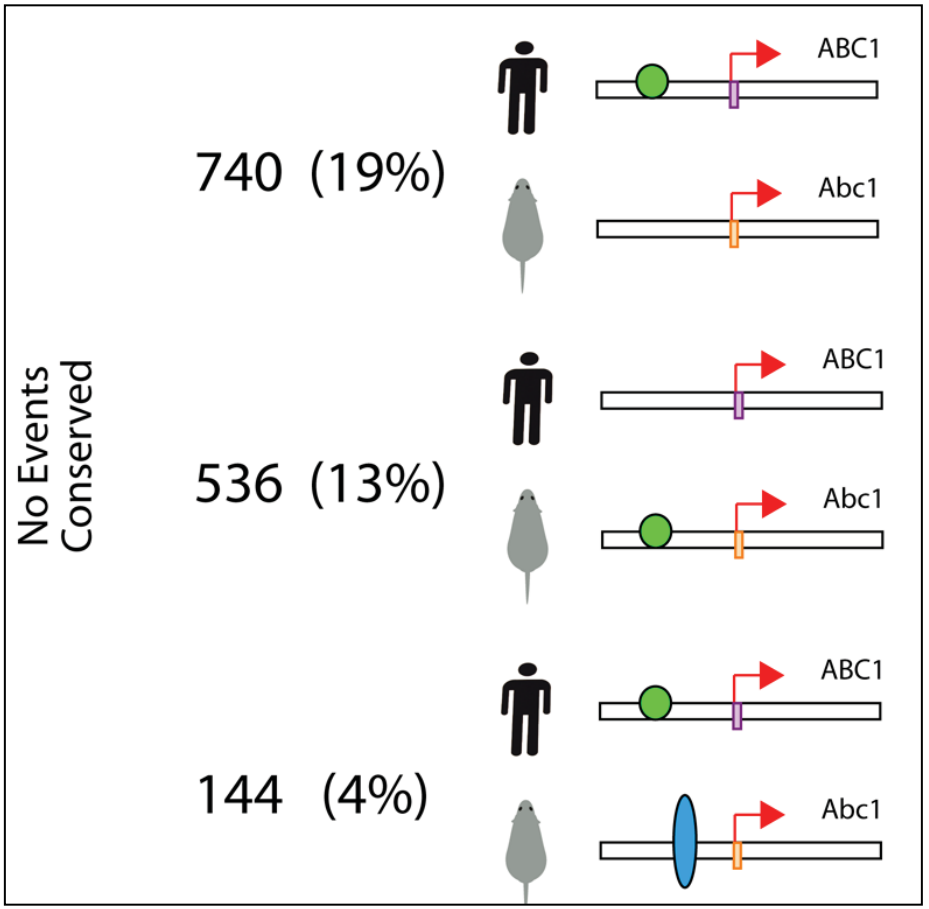
Mouse image courtesy of David Deen. Used with permission.

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

Source: Odom, Duncan T., Robin D. Dowell, et al. "Tissue-specific Transcriptional Regulation has Diverged Significantly between Human and Mouse." *Nature Genetics* 39, no. 6 (2007): 730-32.

FIN

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014