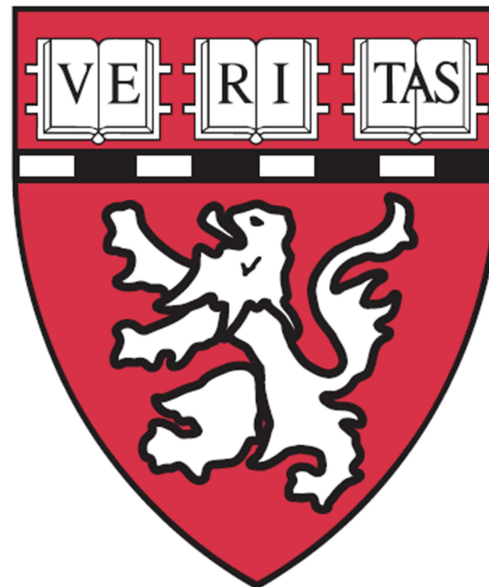


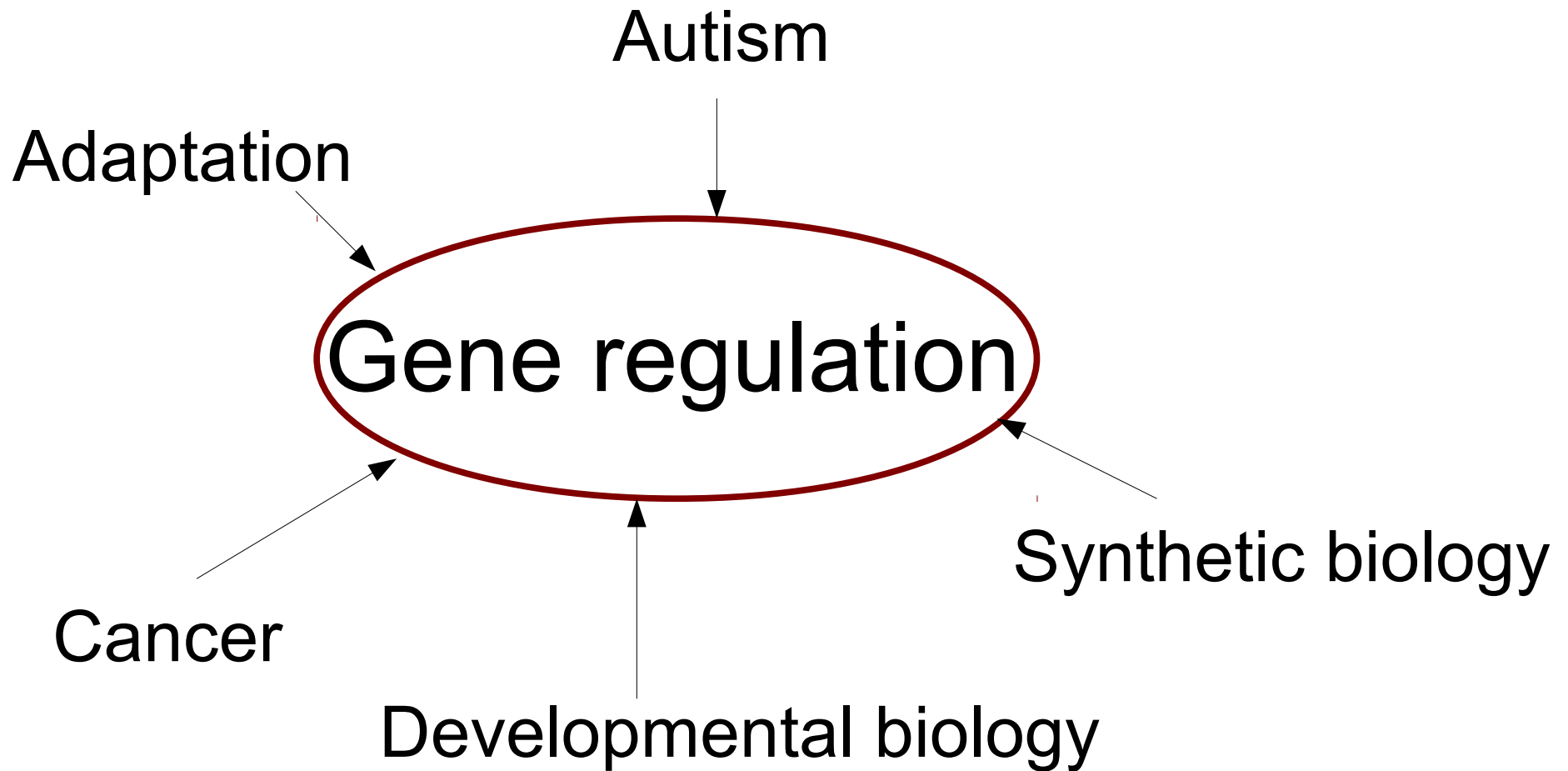
# Mechanisms and models of distal enhancers of inducible gene expression

Martin Hemberg

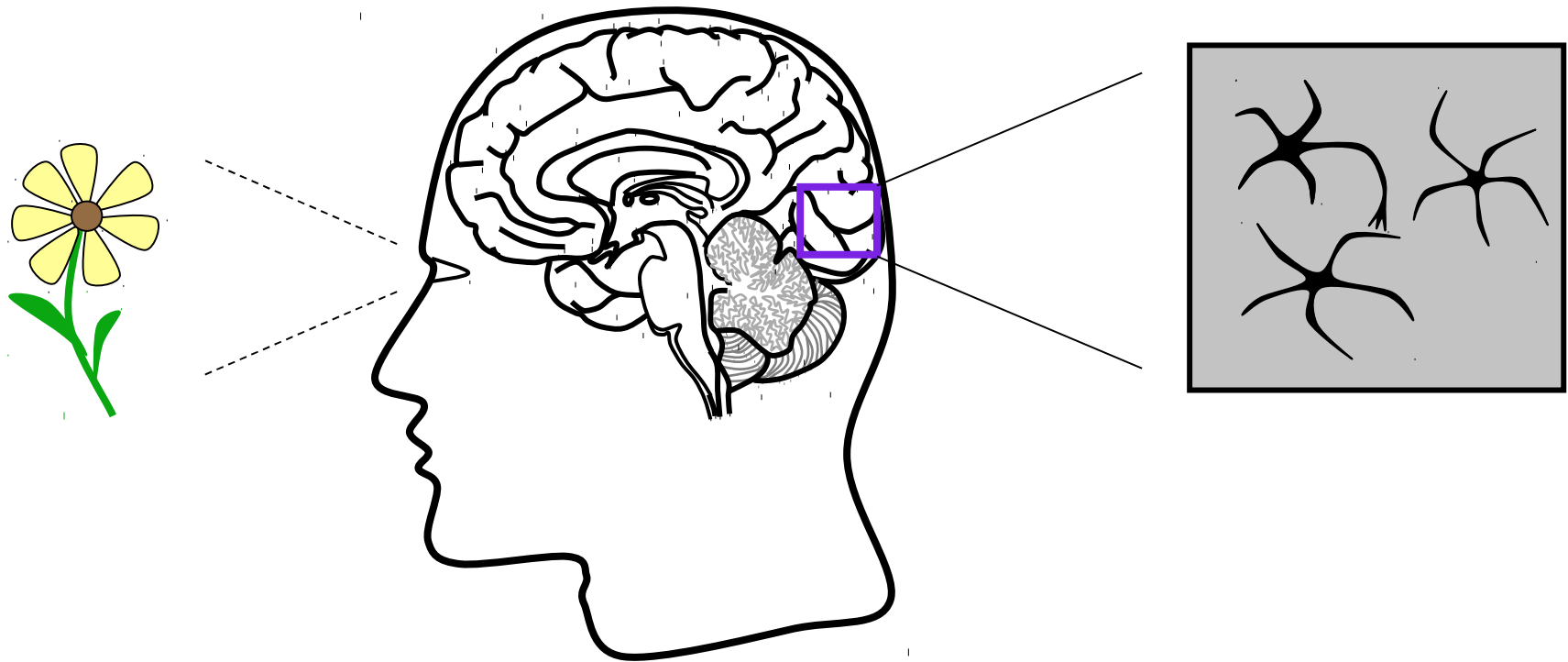
UC Berkeley  
February 28, 2012



# Why is gene regulation important?

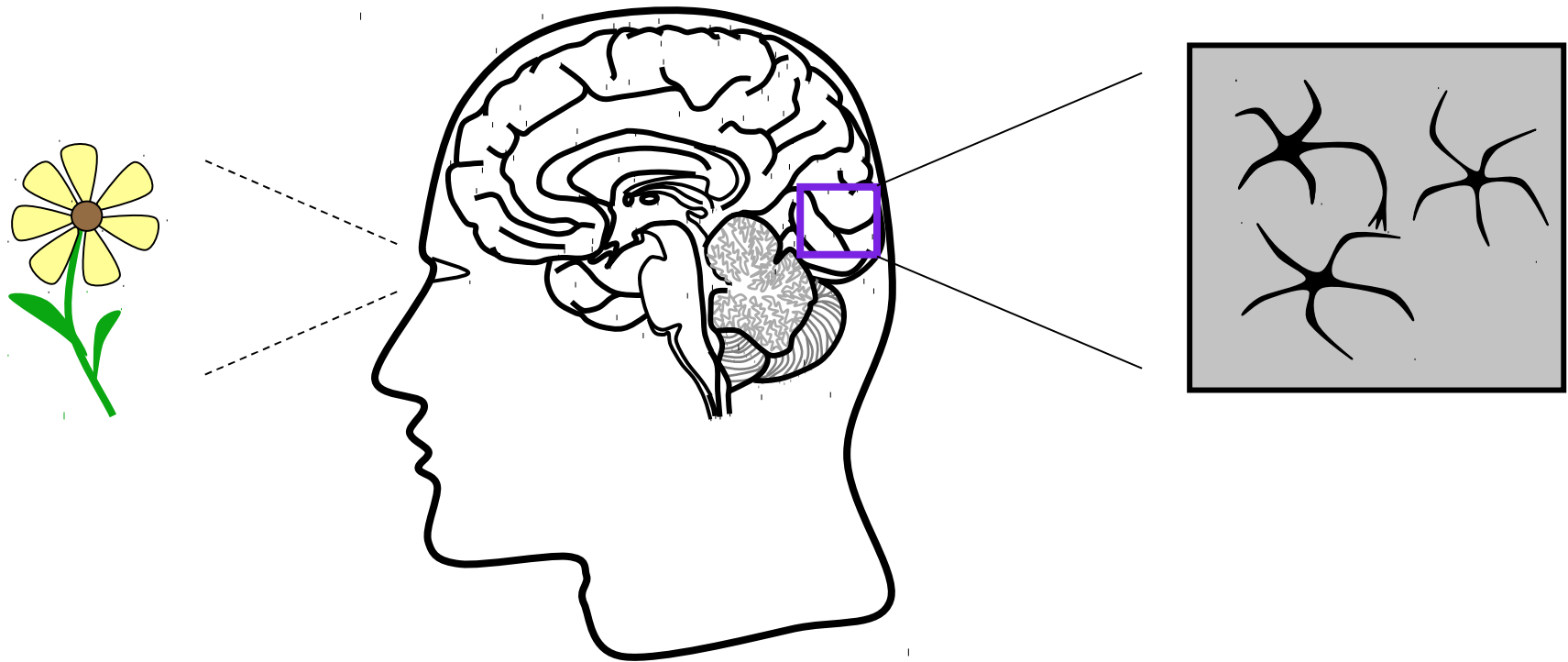


# Synapses change in response to external environmental stimuli



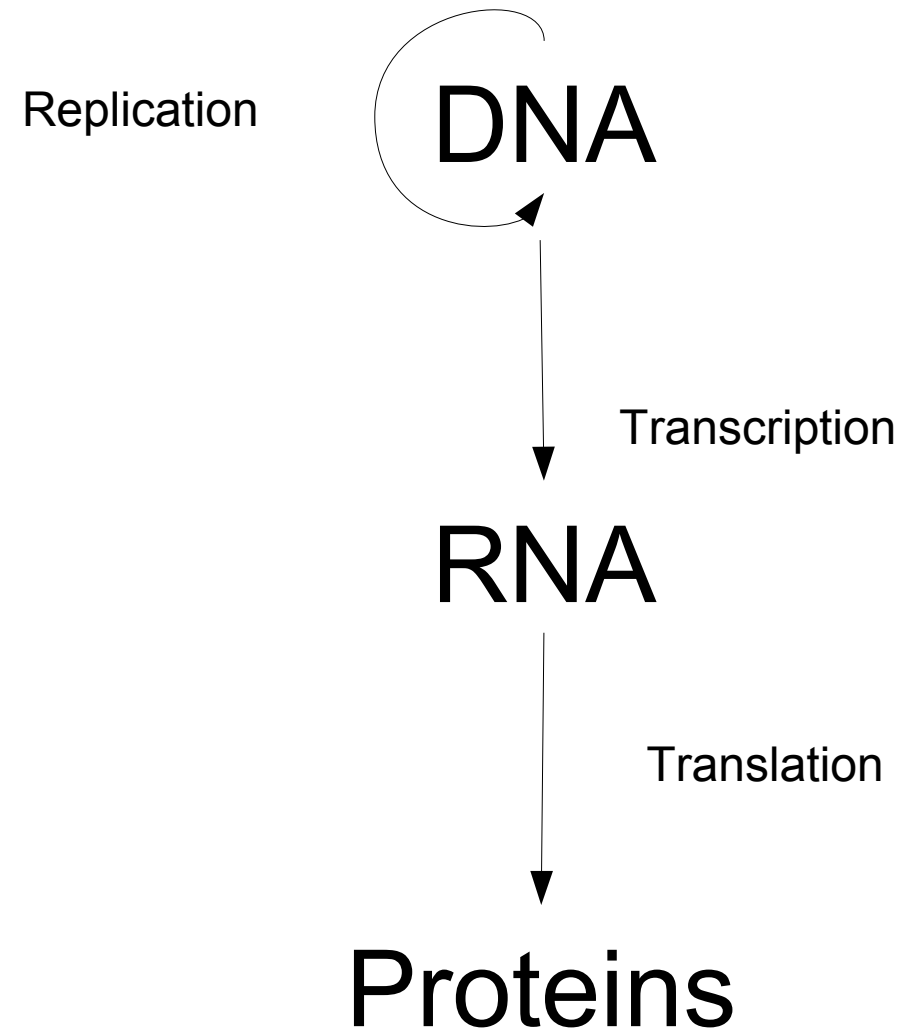
Hubel & Wiesel, 1970's

Changes in synapses are caused by changes in expression of ~1000 genes

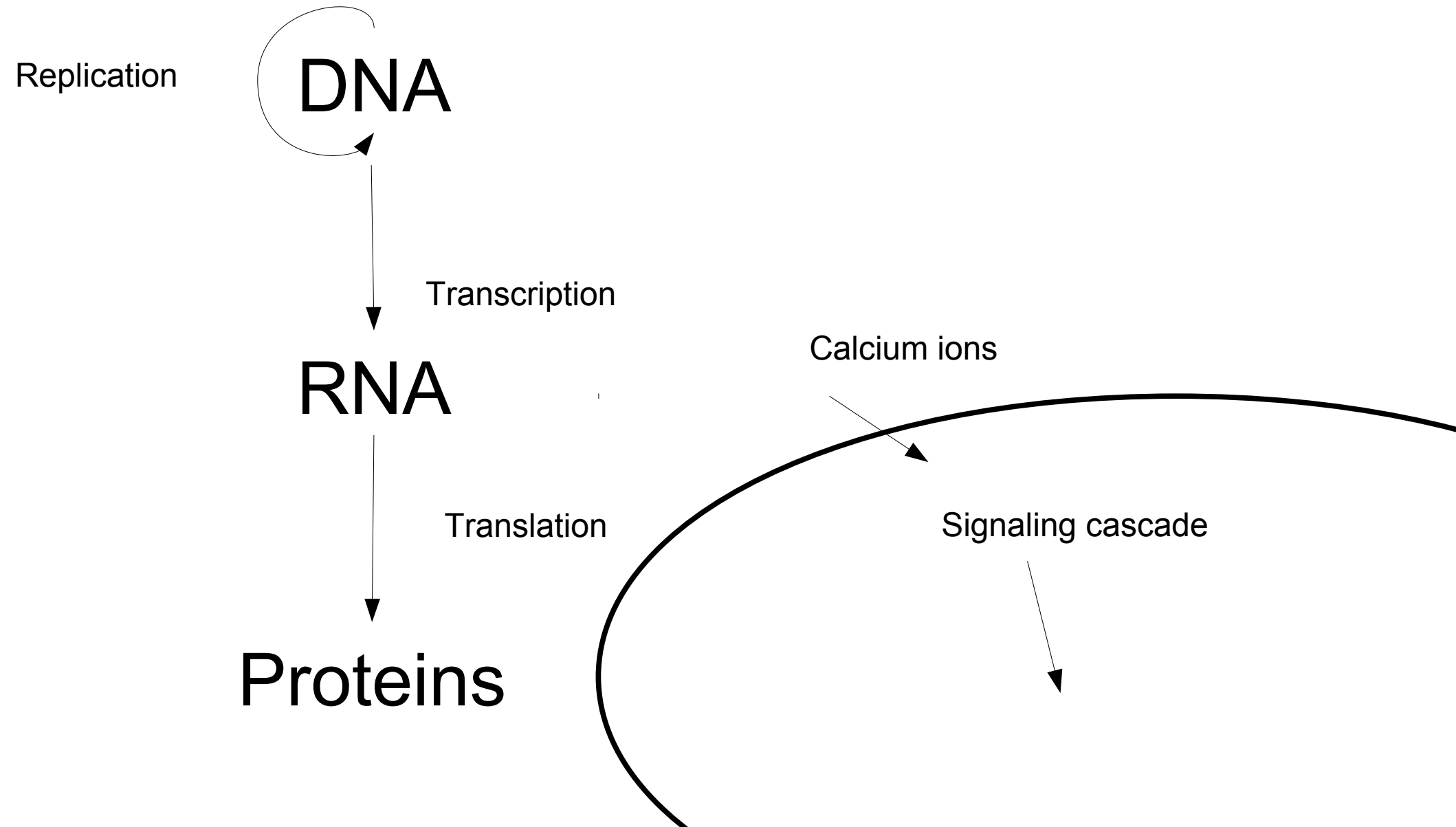


Hubel & Wiesel, 1970's

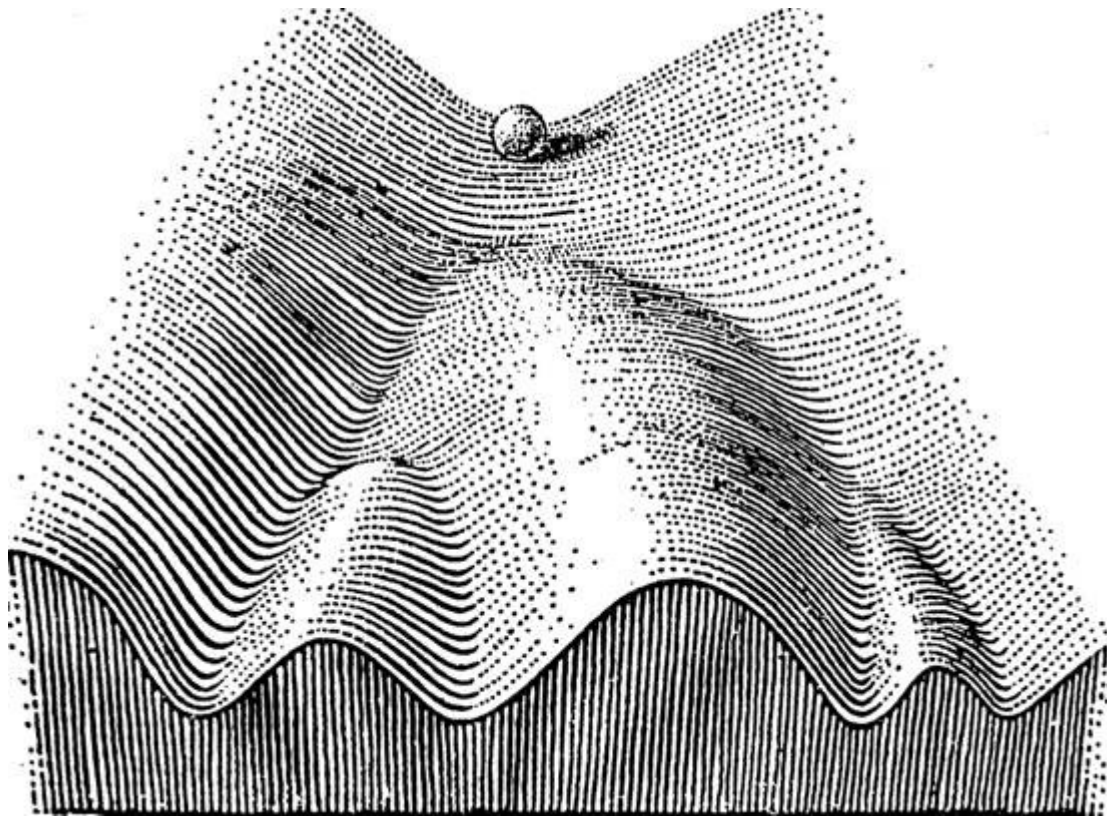
# What is gene expression?



# Activity-dependent gene expression triggered by influx of Ca



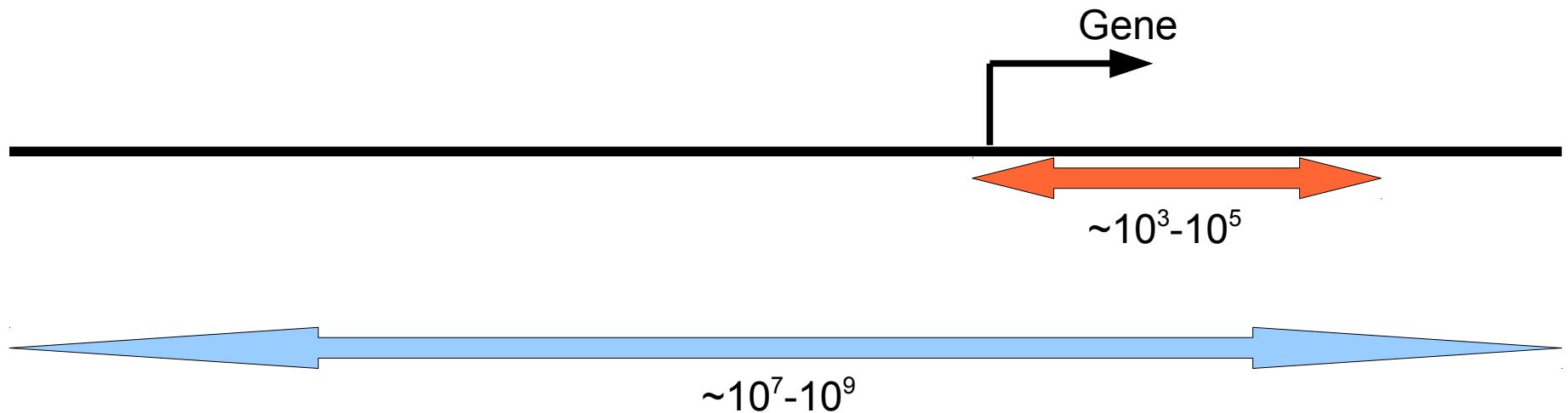
# What is gene regulation?



Waddington, 1953

# Mouse genome is large and has few genes

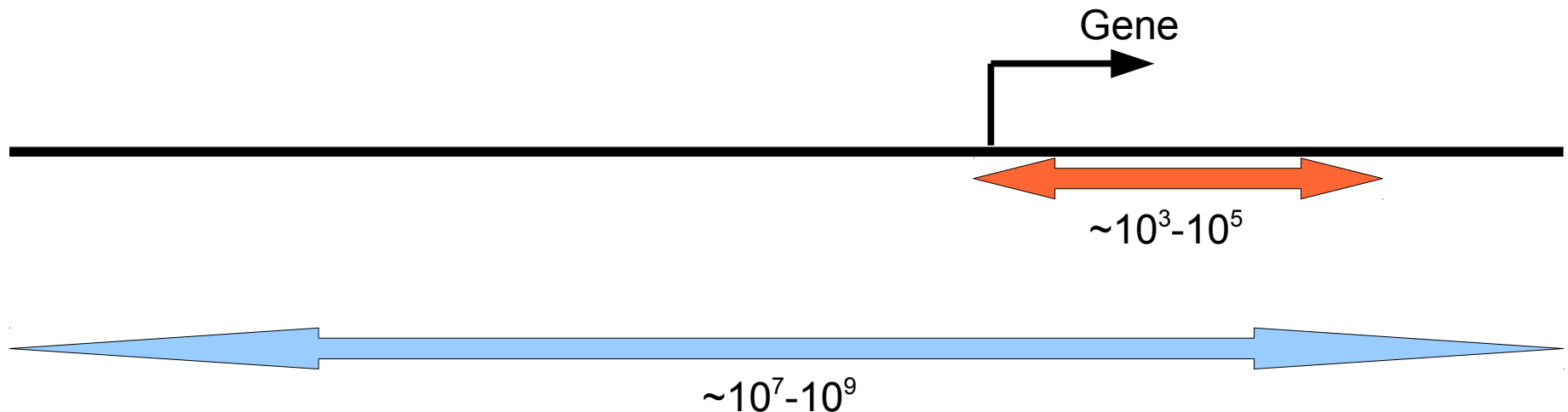
- ~25,000 genes
  - ~2% of DNA





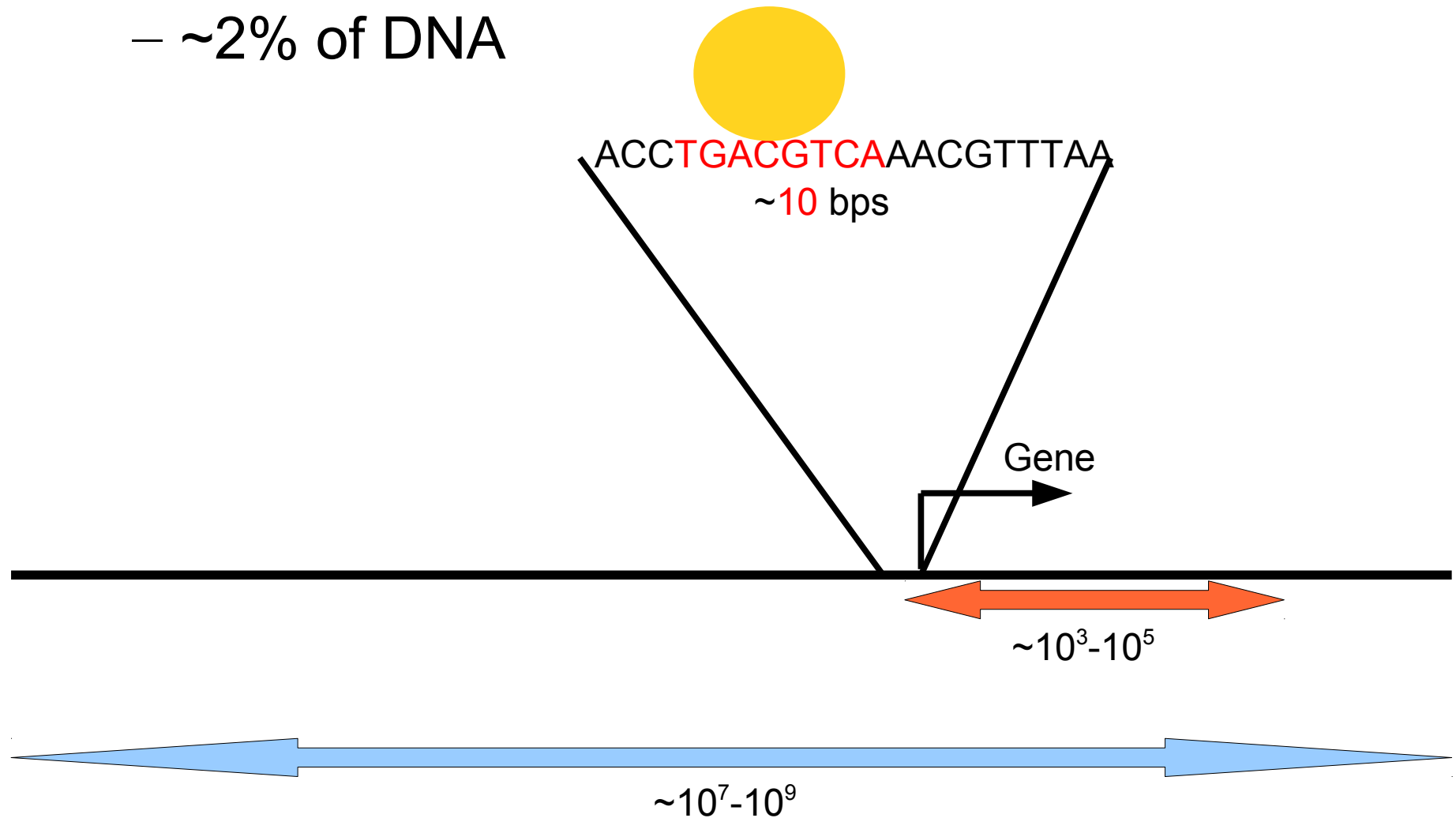
# Mouse genome is large and has few genes

- ~25,000 genes
  - ~2% of DNA
  - Bacteria  $\sim 10^6$  base pairs (bps)
    - $10^3$ - $10^4$  genes



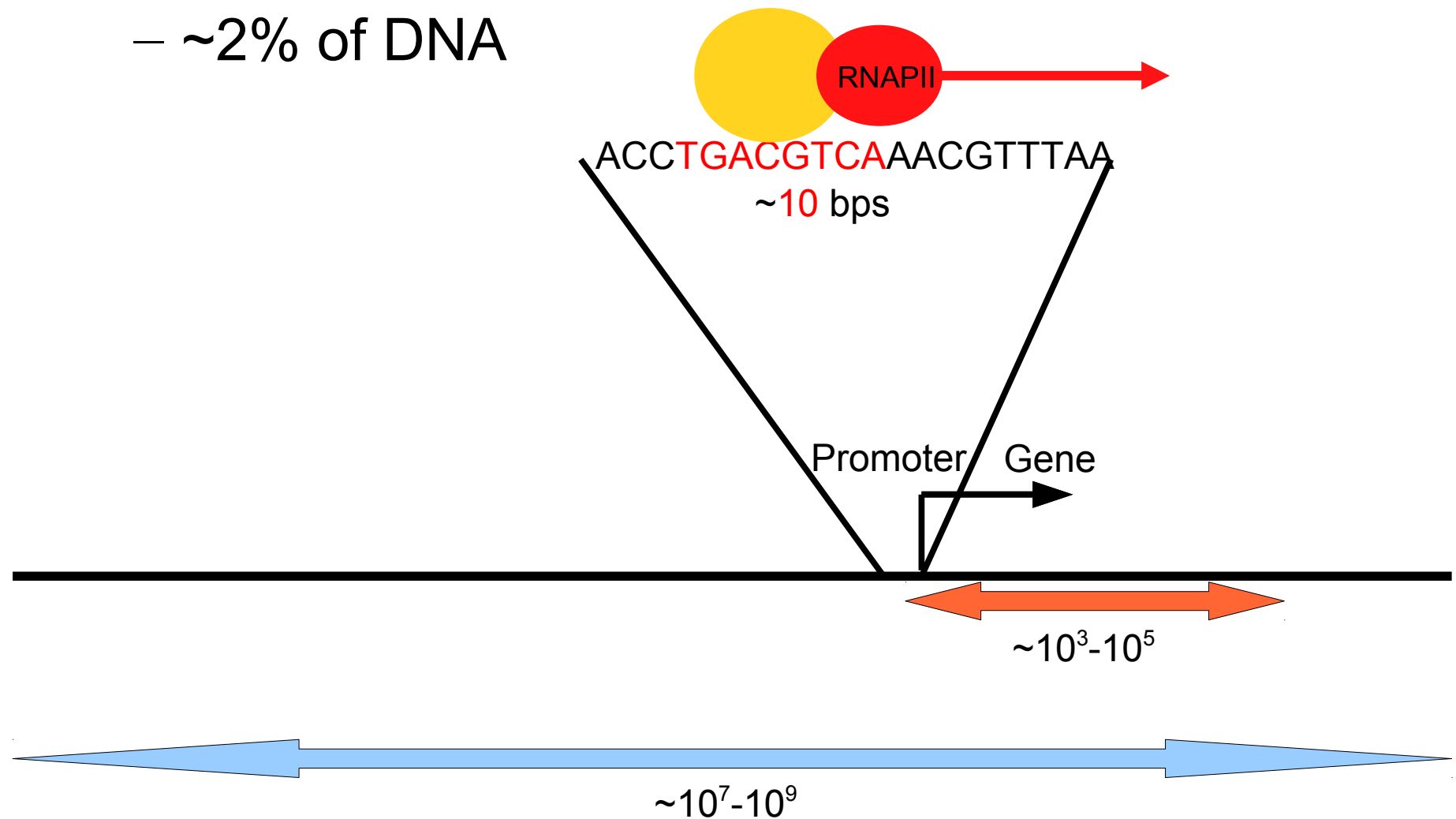
# Transcription Factors (TFs) bind to DNA motifs

- ~25,000 genes
  - ~2% of DNA

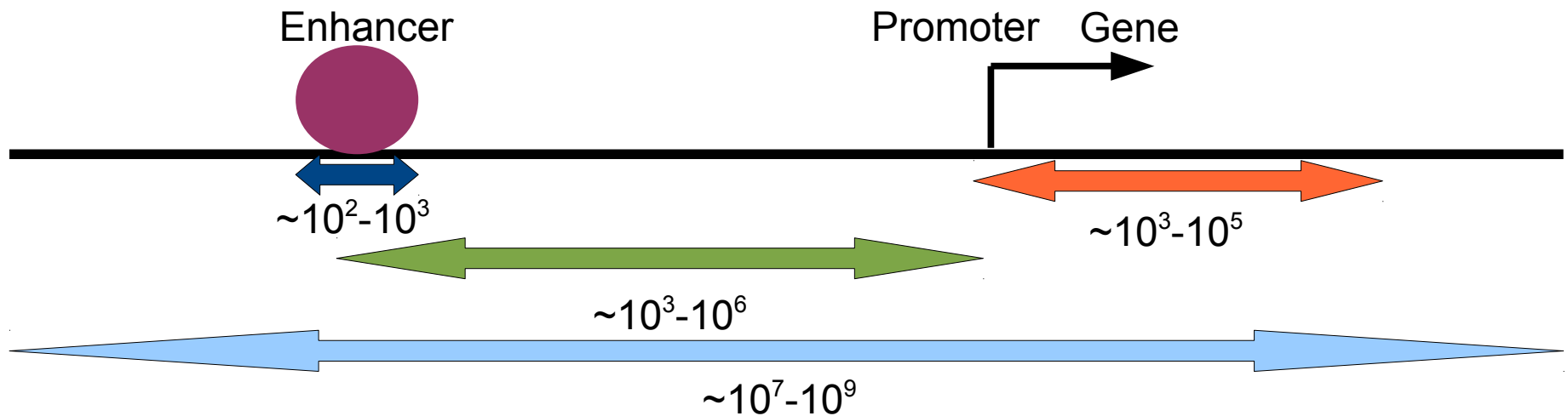


# Transcription factors bind at promoter to recruit RNA Polymerase II (RNAPII)

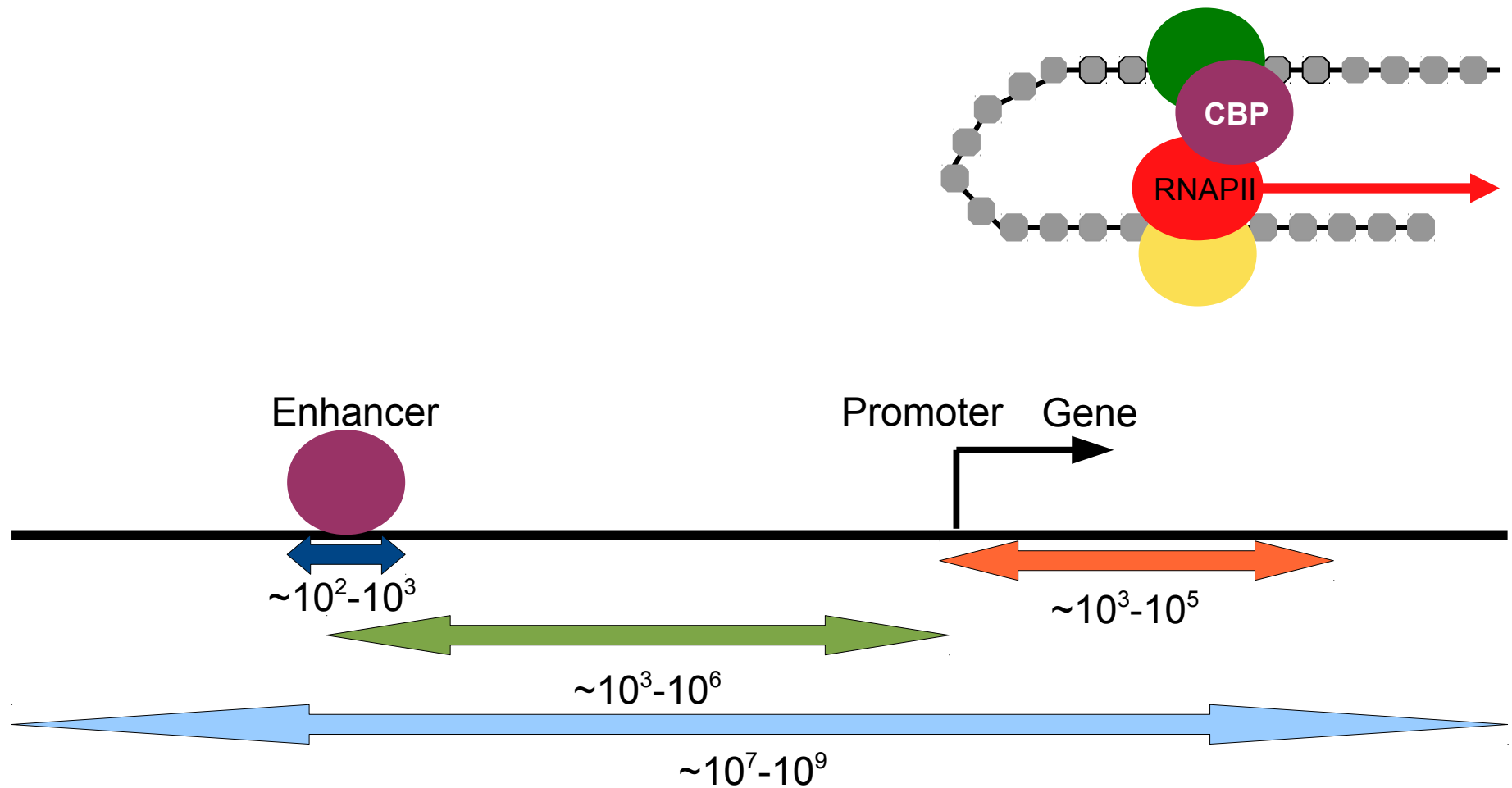
- ~25,000 genes
  - ~2% of DNA



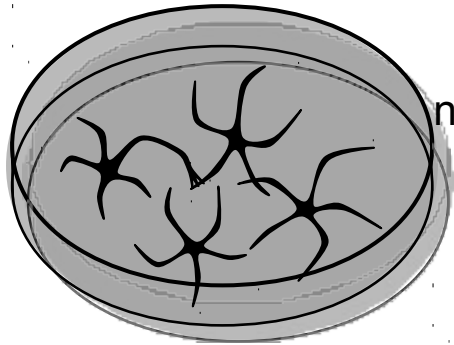
# Enhancers are distal regulatory sequences



# Enhancers characterized by **CBP** binding



# Cultured mouse cortical neurons for genome-wide study of activity dependent gene expression

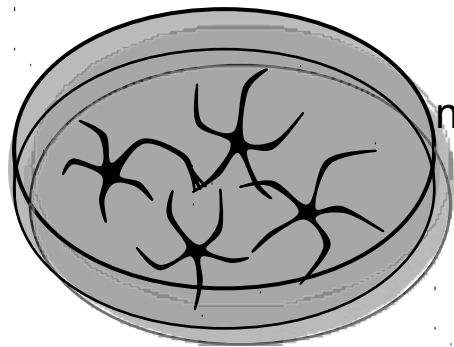


mouse cortical  
neurons

neuronal activation via potassium chloride (**KCl**) depolarization



# Potassium chloride (**KCl**) stimulation induces cells to change state

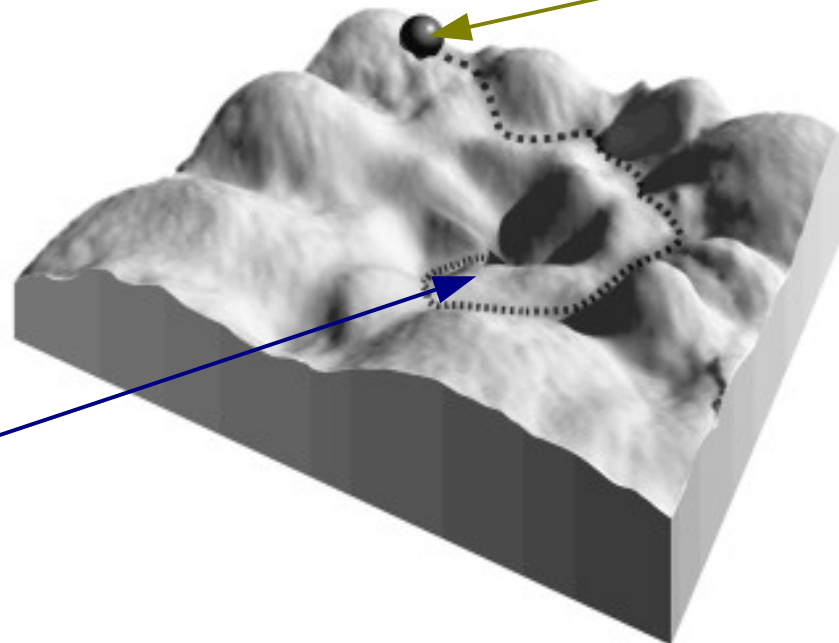


mouse cortical neurons

neuronal activation via potassium chloride (**KCl**) depolarization

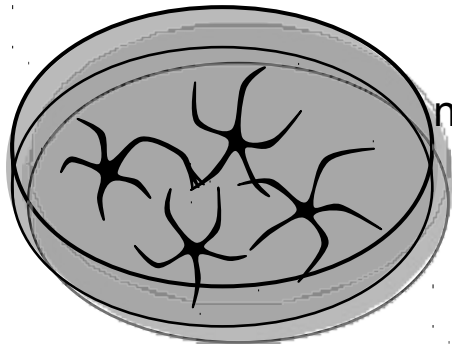


Unstimulated (un)



Stimulated (**KCl**)

# Genome-wide data obtained using high-throughput sequencing



mouse cortical neurons

neuronal activation via potassium chloride (**KCl**) depolarization

- KCl

ChIP-Seq  
RNA-Seq

+ KCl

ChIP-Seq  
RNA-Seq

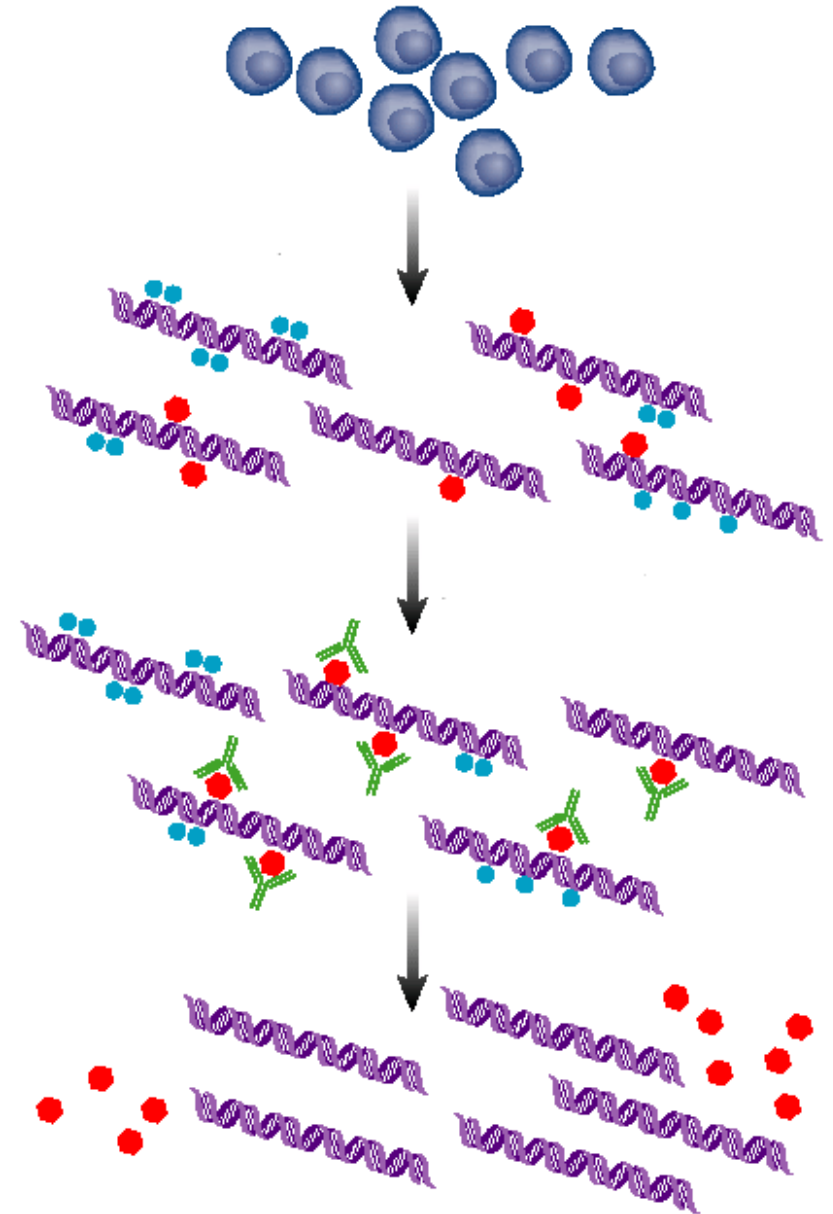


Jesse Gray  
Tae-Kyung Kim  
Greenberg Lab



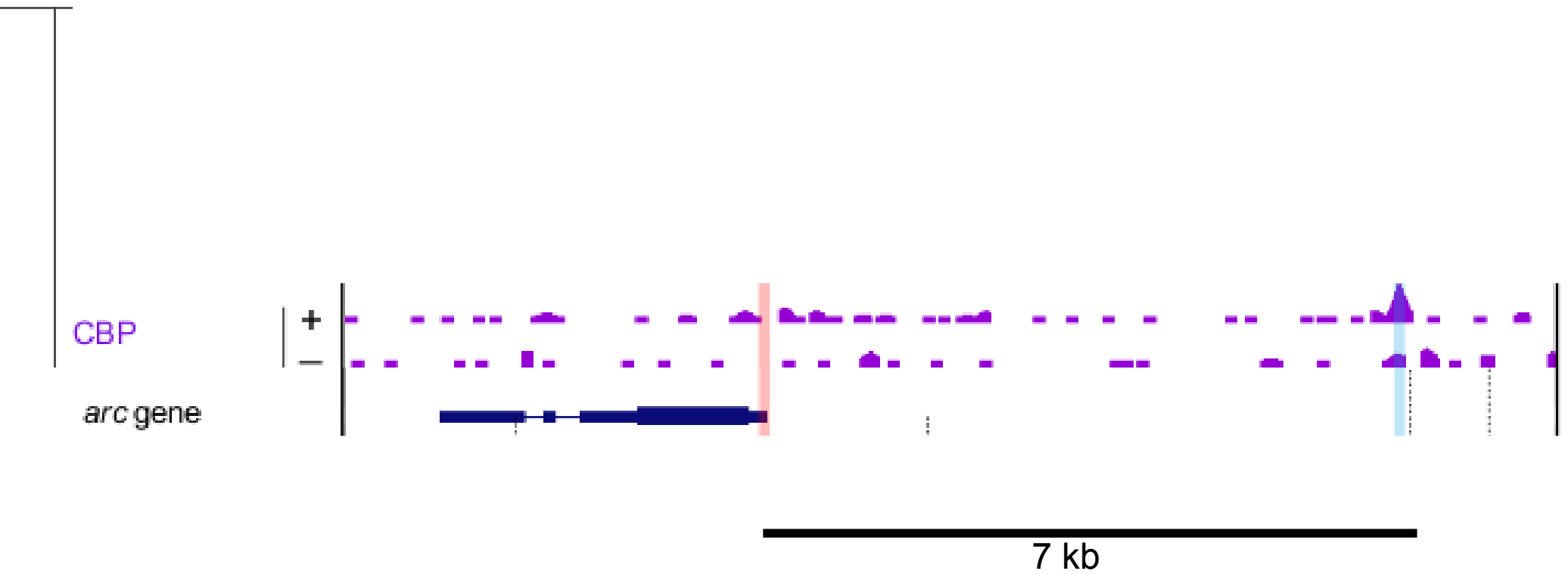
# Chromatin immunoprecipitation and sequencing (**ChIP-Seq**) finds protein binding sites *in vivo*

- Short **reads** mapped to reference genome
- #reads ~ binding
- ~10 M reads
- Unbiased



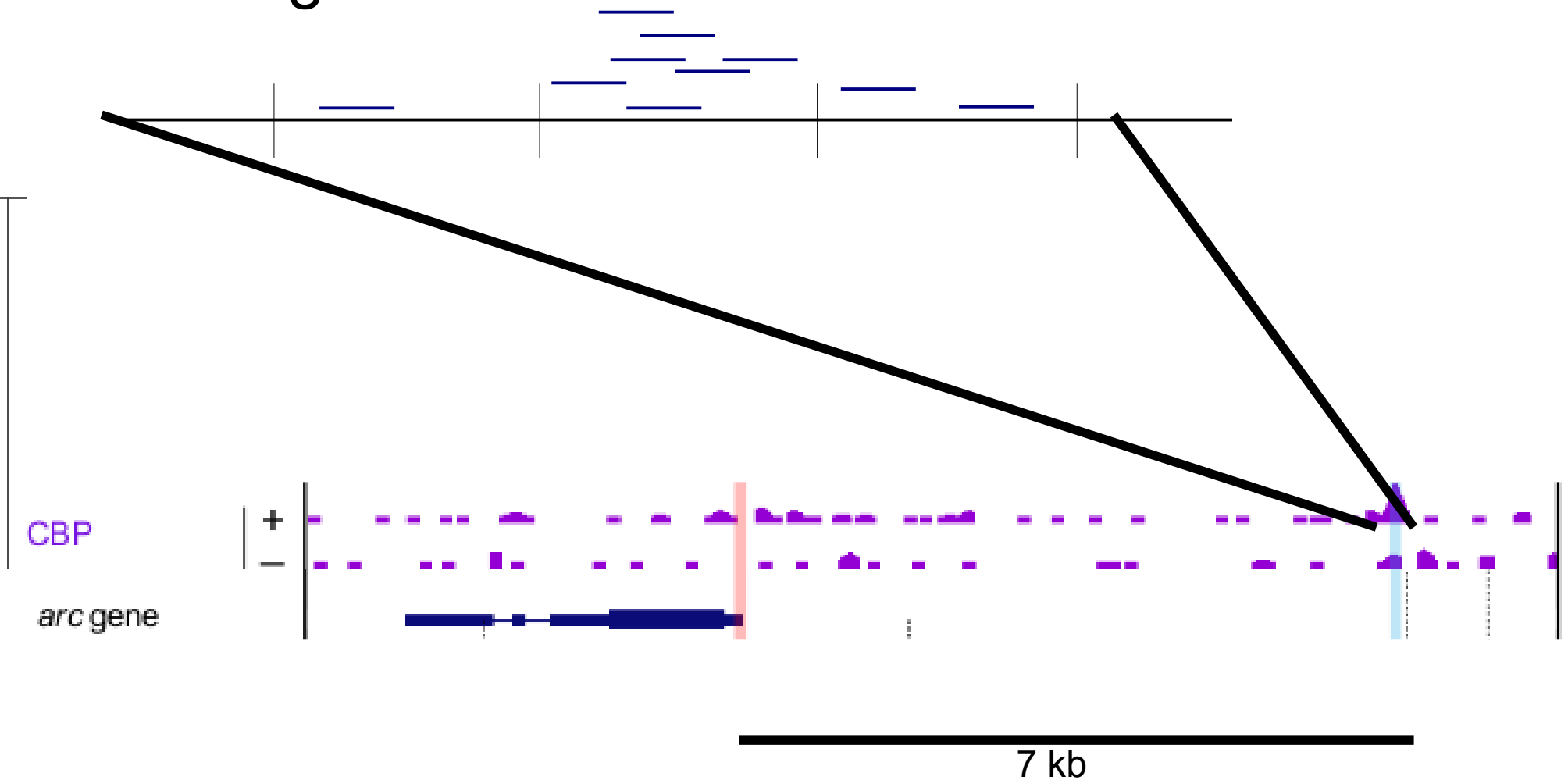
(Mardis, 2007)

# Inducible CBP binding at enhancers

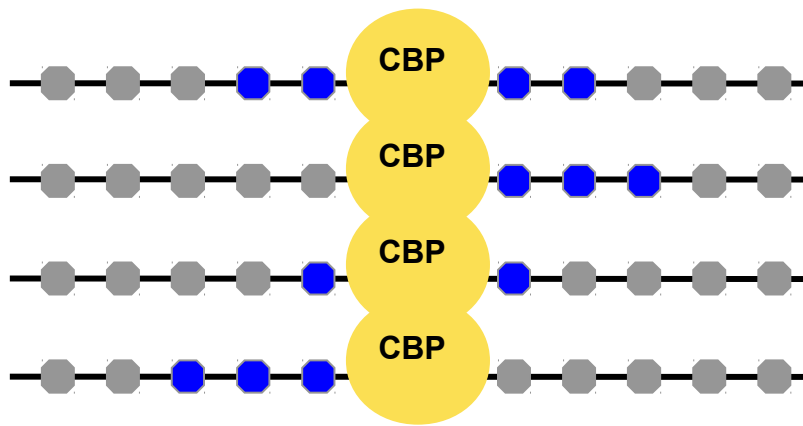


# Identifying ~28,000 CBP binding sites in two replicate experiments

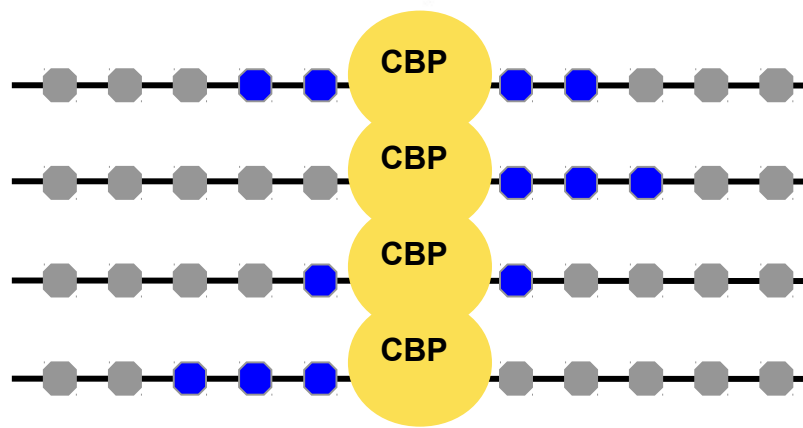
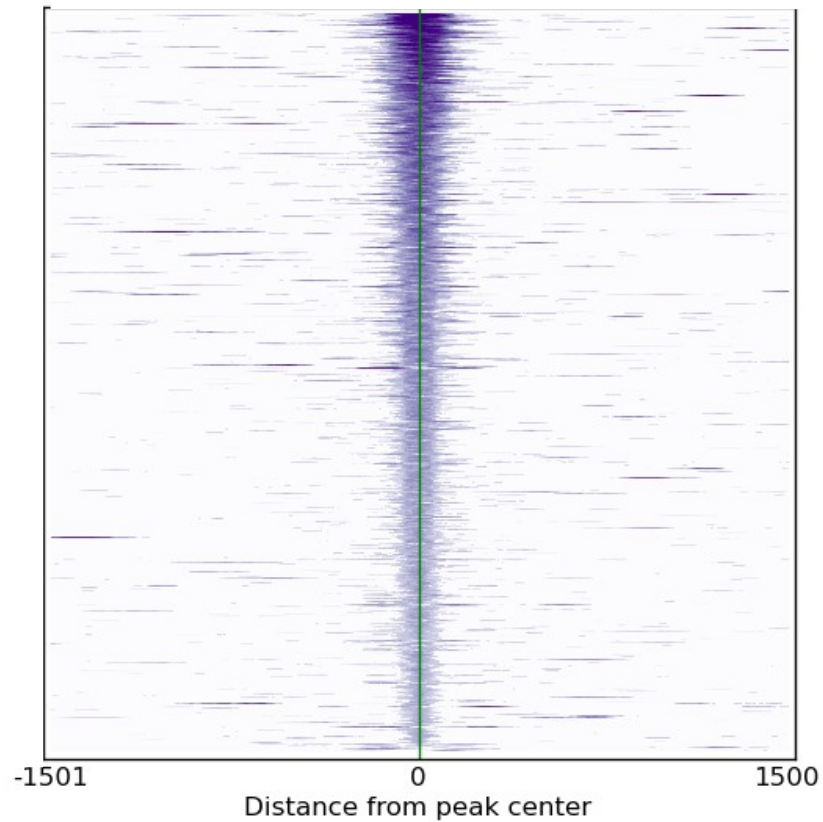
- Regions that have significantly more CBP than background



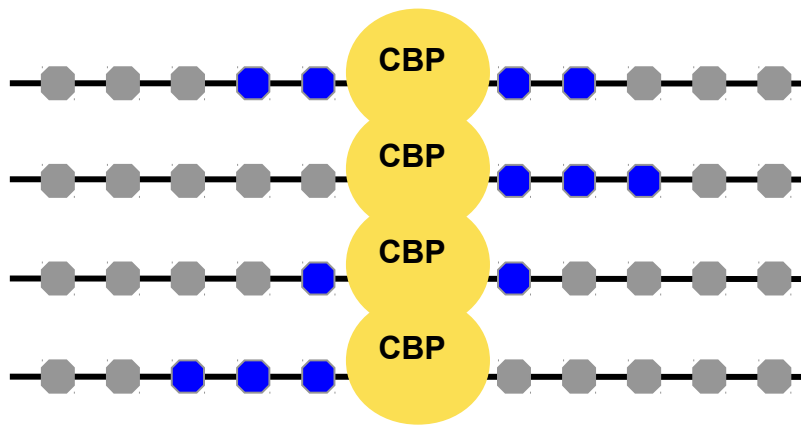
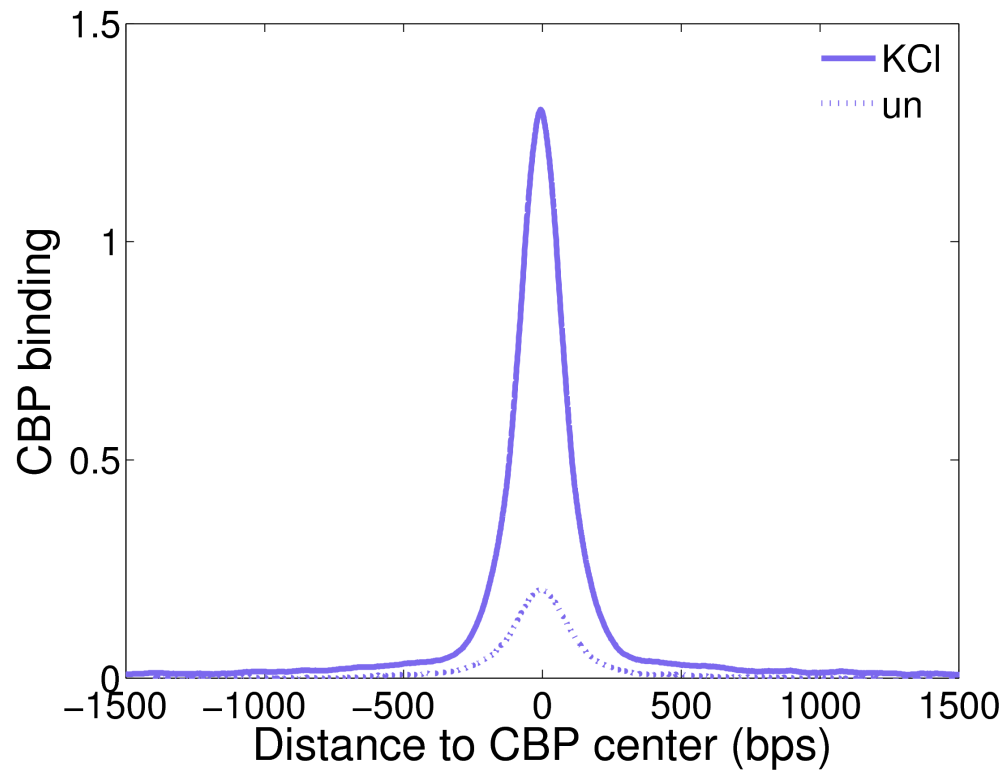
# Aligning CBP peaks to calculate binding profiles



# Aligning CBP peaks to calculate binding profiles

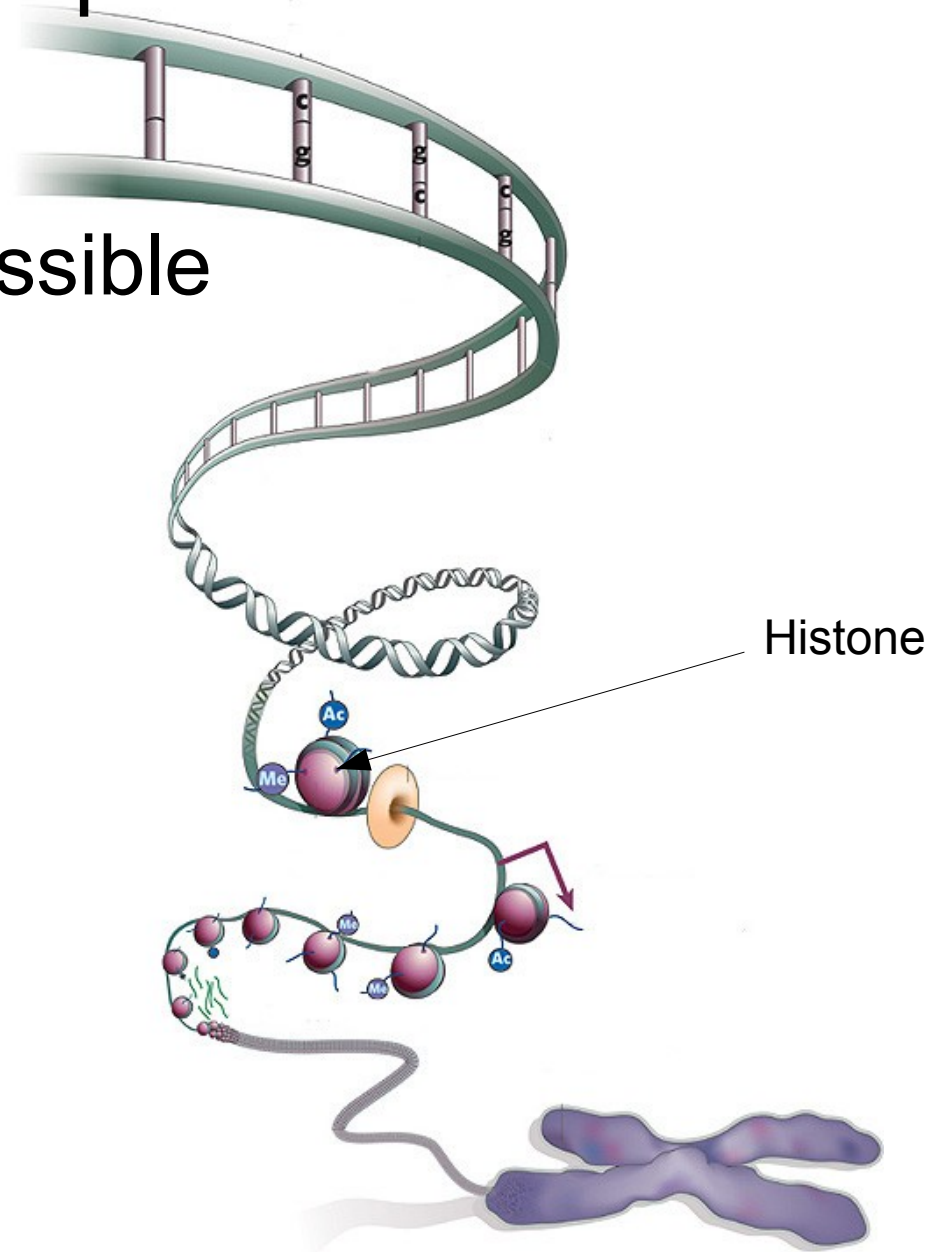


# Average profile of CBP binding



# Histones prevent transcription factors from binding to DNA

- ~100 k loci or 1% accessible
  - Open chromatin
  - Cell-type specific

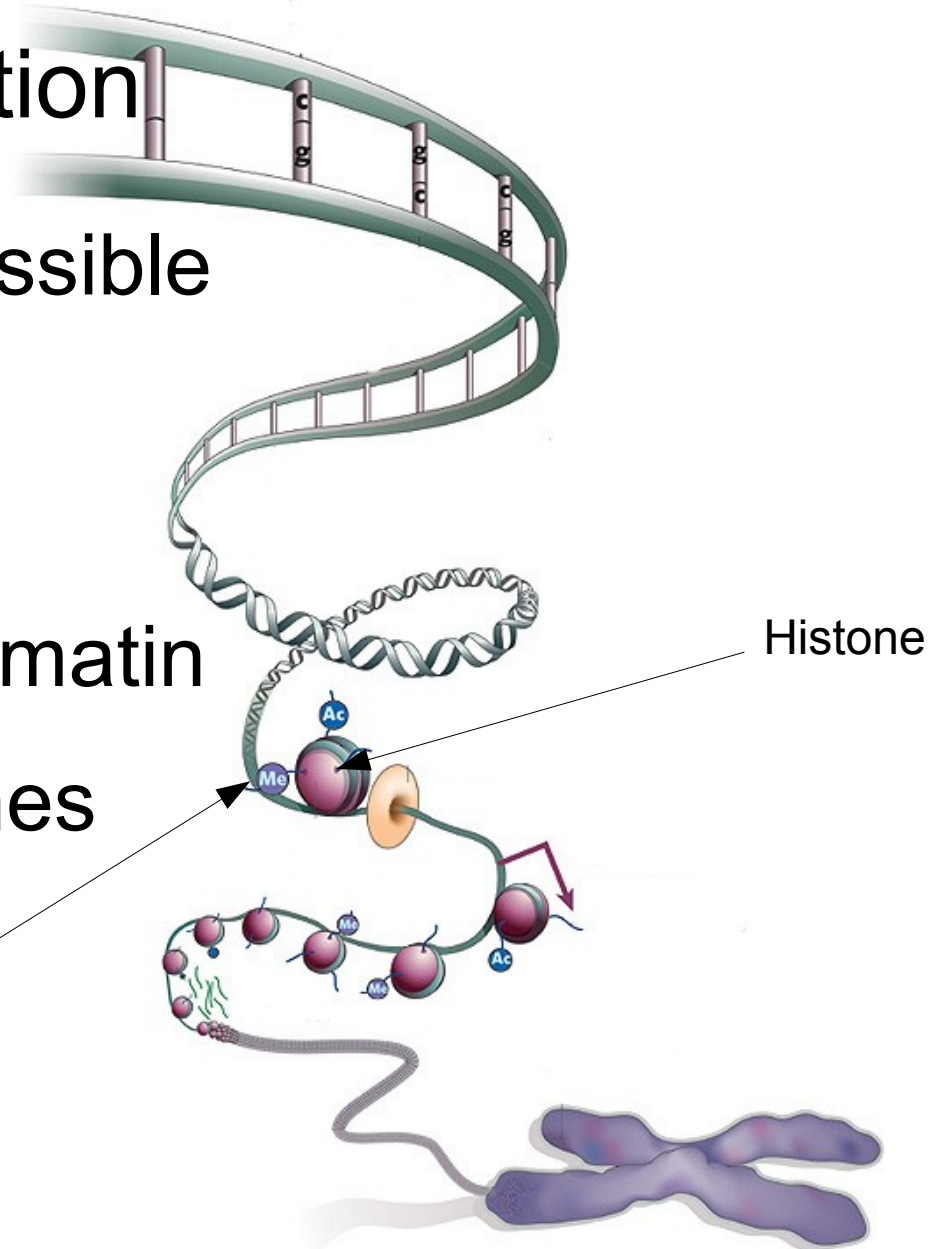


(ENCODE, 2007)

# Post-translational modifications of histone tails correlate with function

- ~100 k loci or 1% accessible
  - Open chromatin
  - Cell-type specific
- **H3K4me1** – open chromatin
- **H3K4me3** – active genes

Methyl group

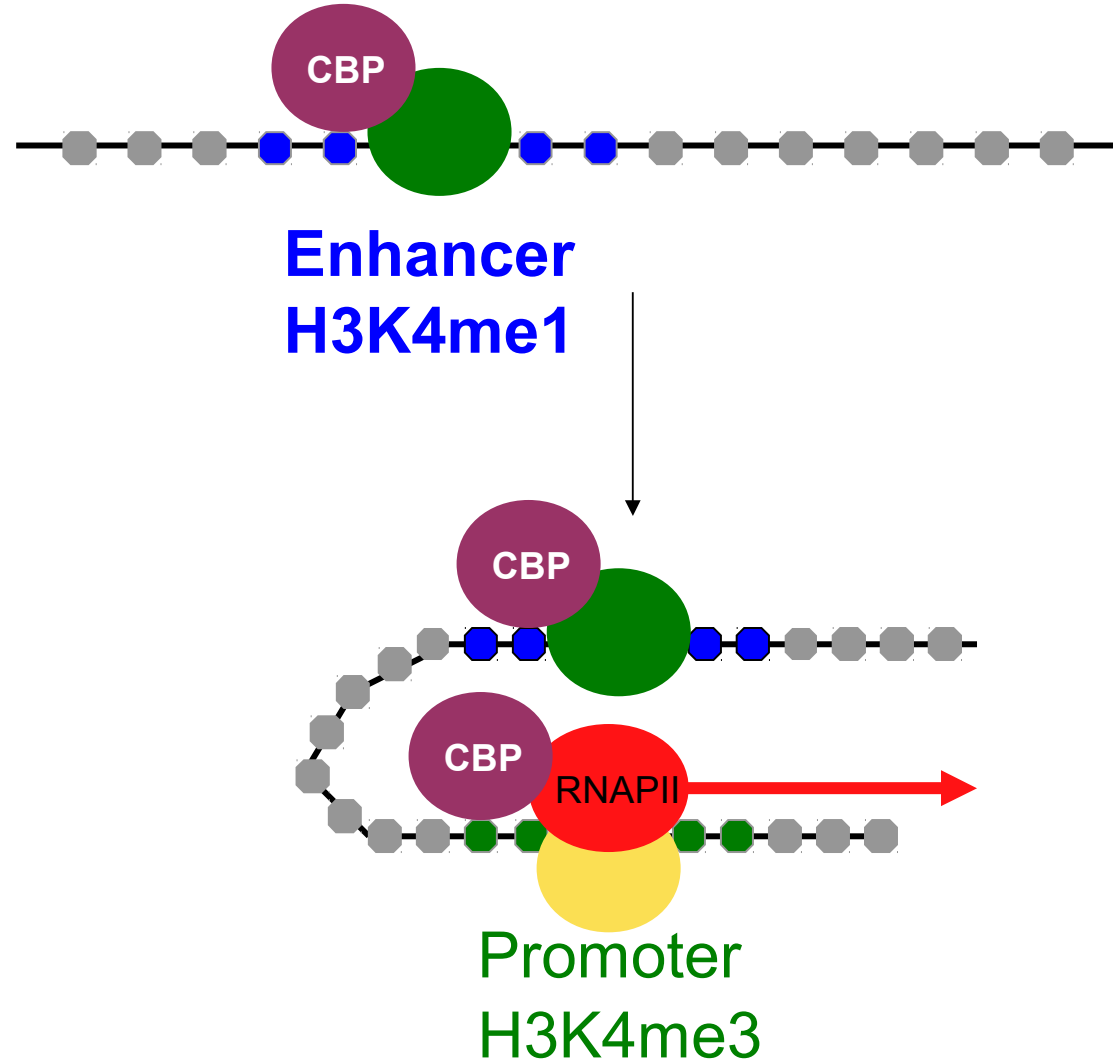


(ENCODE, 2007)

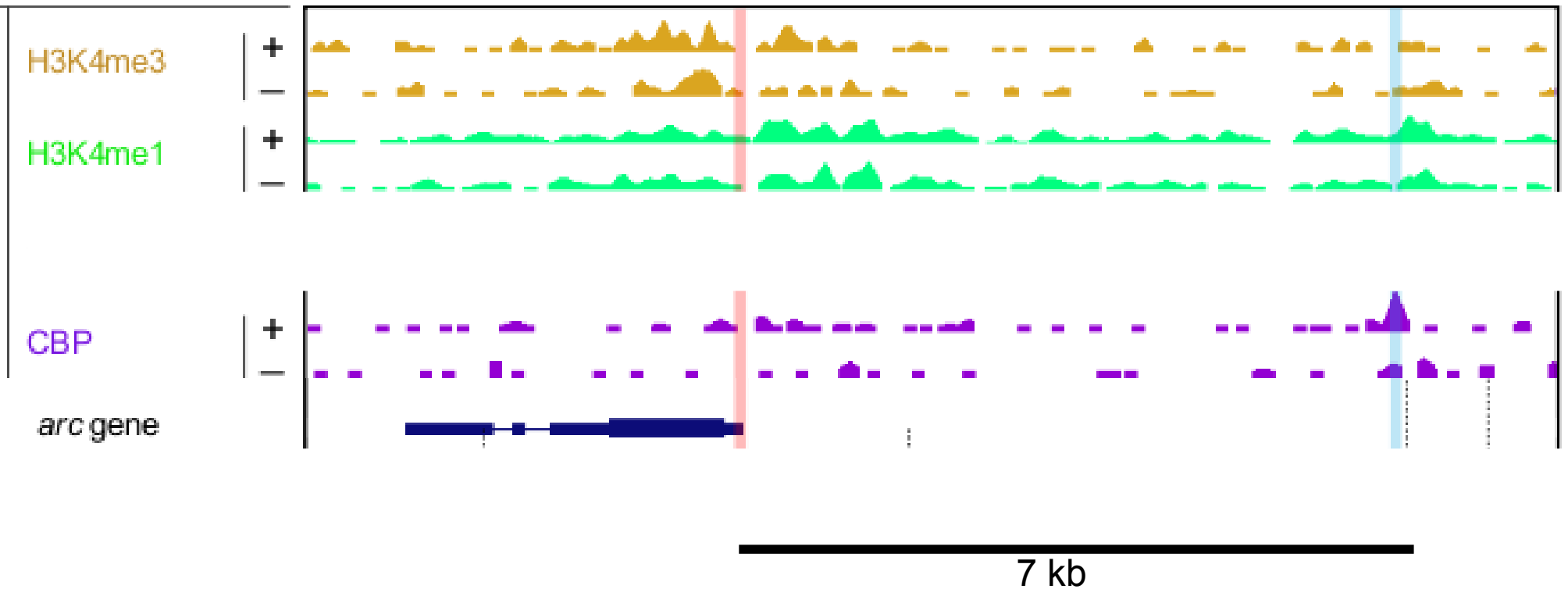


# A combination of CBP and histone modifications identifies putative enhancers

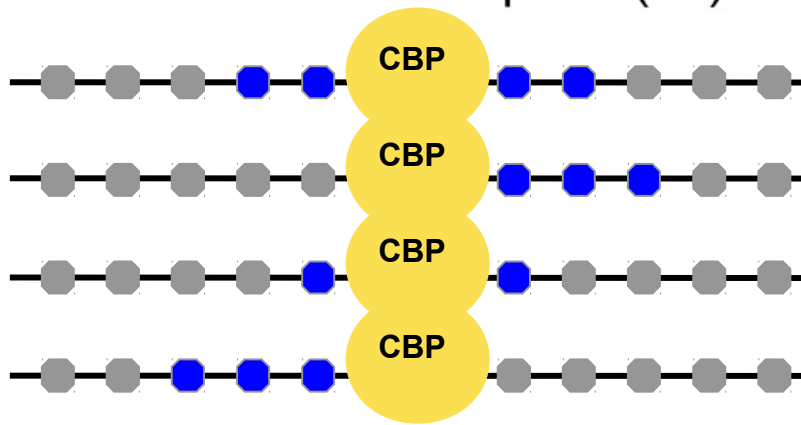
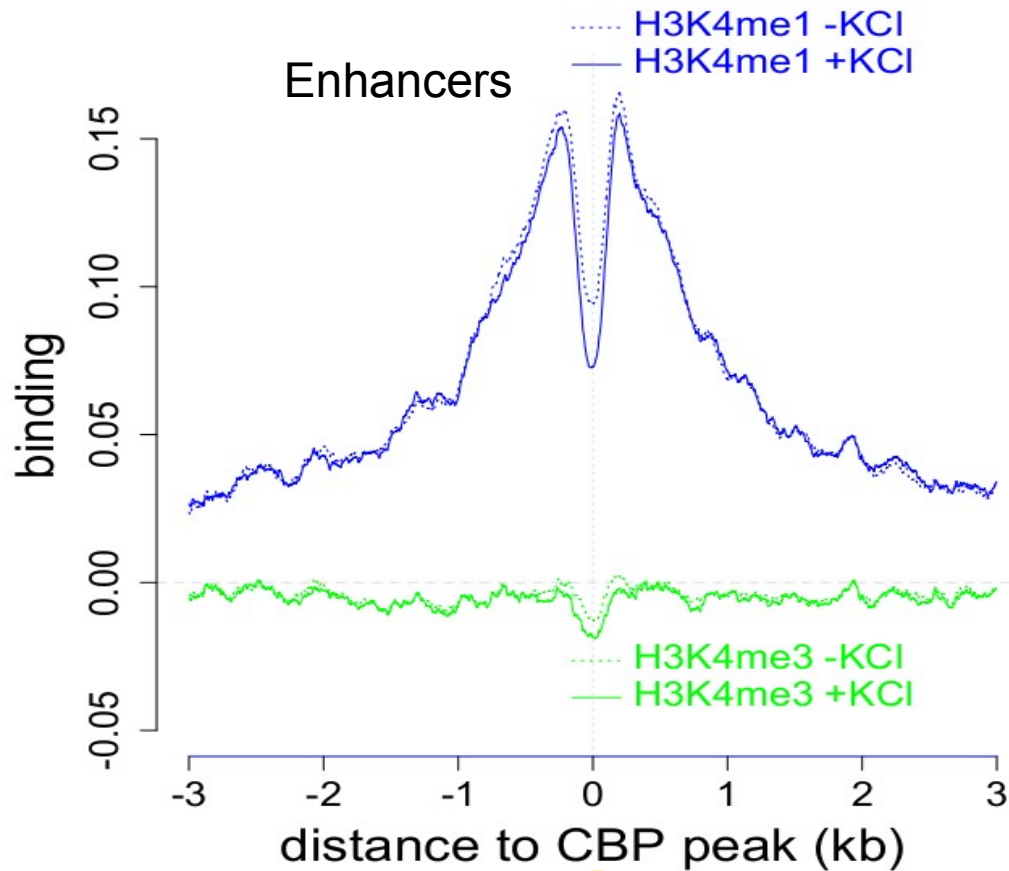
- **CBP** binding
- **H3K4me1** flanking
- **H3K4me3** absent
  - Many unannotated promoters in the genome



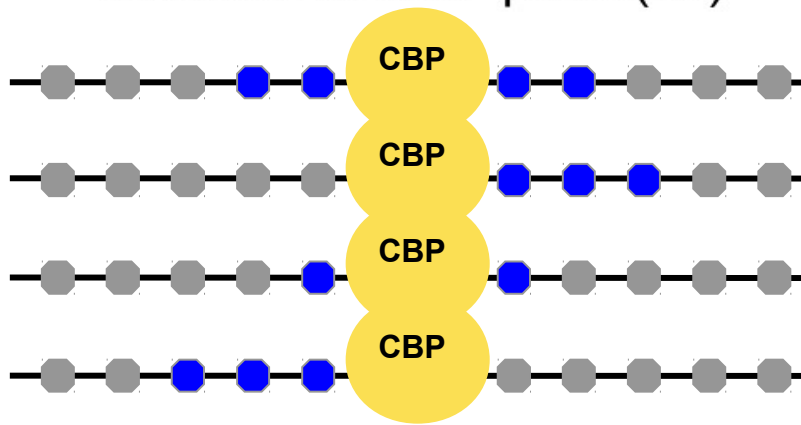
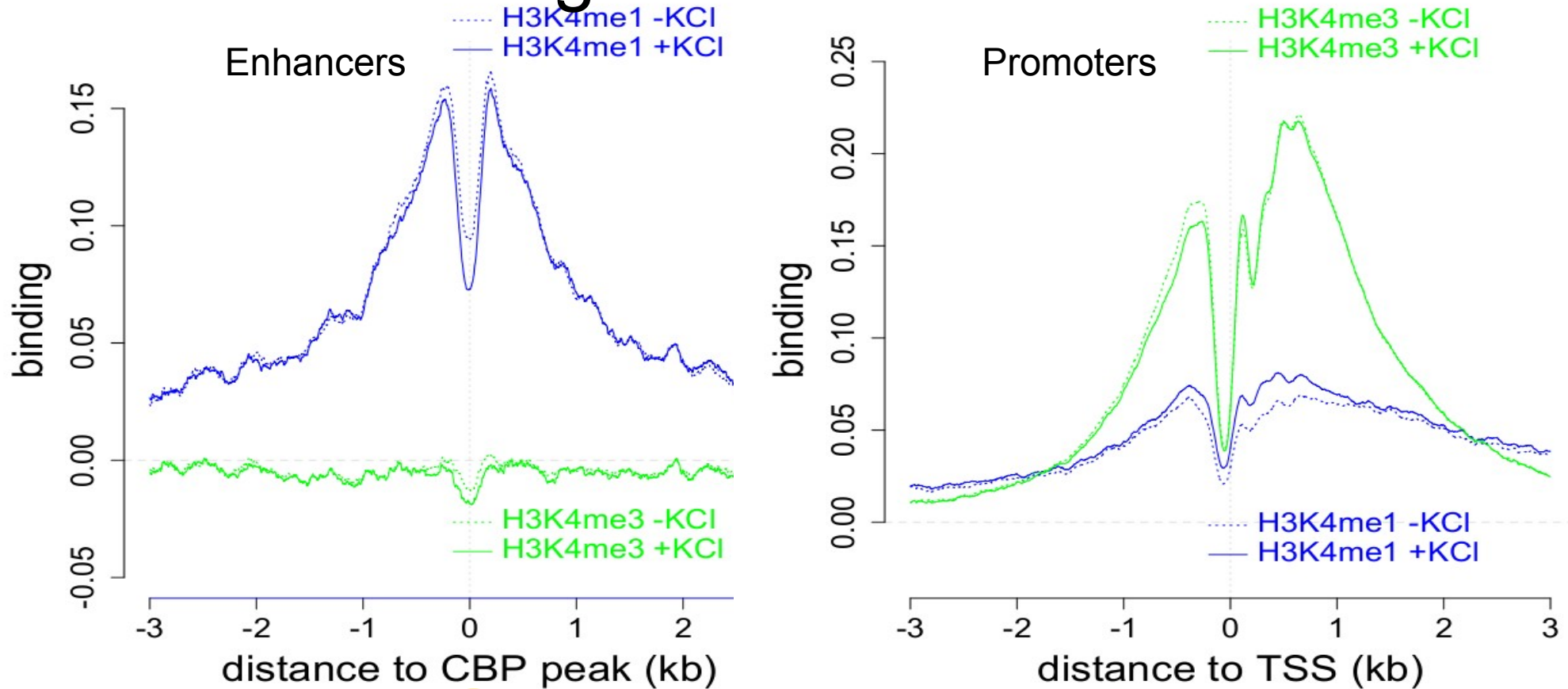
Distal CBP peaks have high levels of H3K4me1 and low levels of H3K4me3



# Enhancers have high levels of H3K4me1 and low levels of H3K4me3

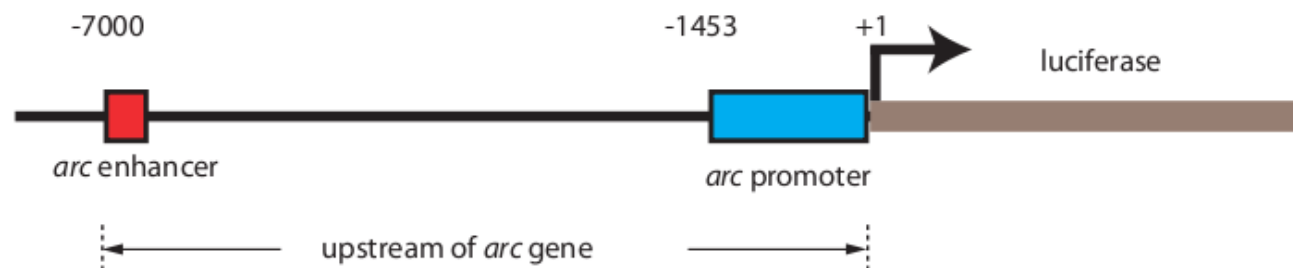


# Transcription start sites have high levels of H3K4me1 and high levels of H3K4me3

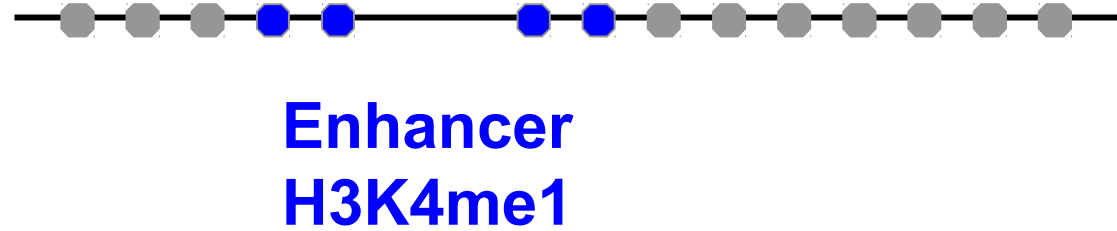


We identified ~12,000 activity-dependent enhancers throughout the genome

- **CBP** peak
- **High** levels of flanking **H3K4me1**
- **Low** levels of **H3K4me3**
  - 8/8 tested activity-dependent enhancers were validated using a luciferase assay

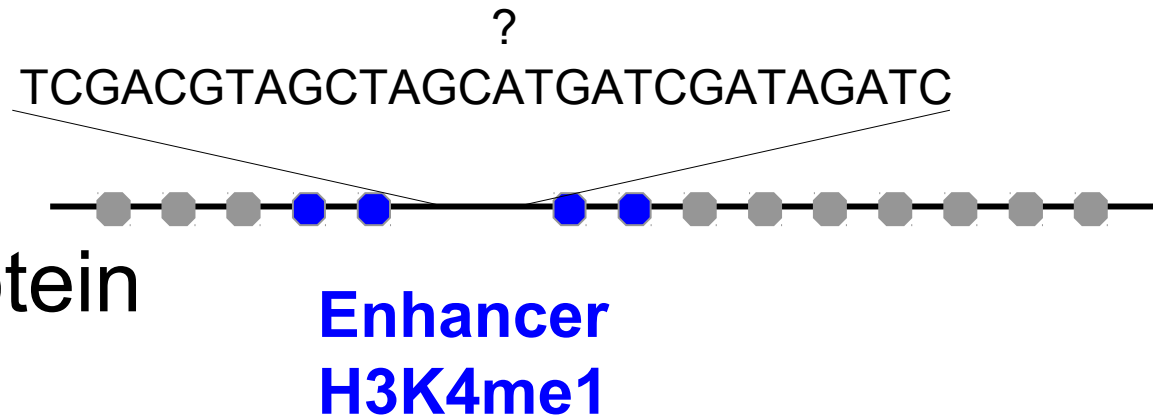


# A quantitative model of distal enhancers



# What TFs bind to enhancers?

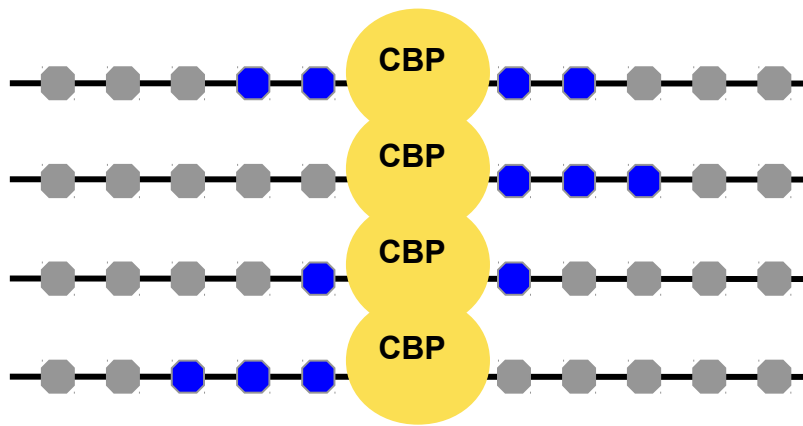
- CBP -  
CREB Binding Protein
  - >50 partners



# What motifs are enriched at enhancers?

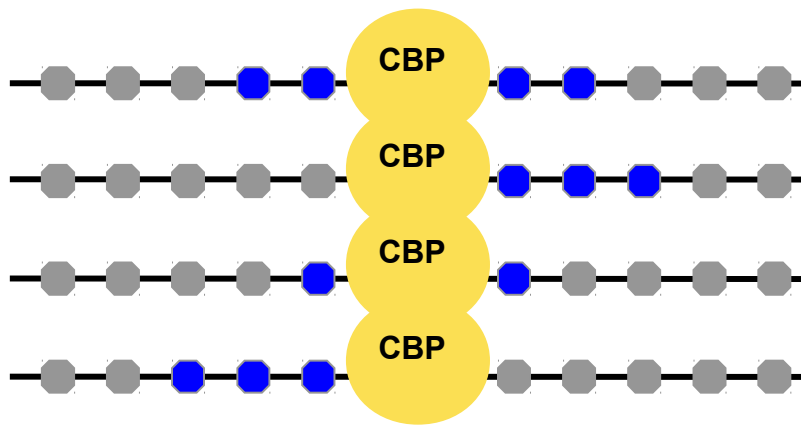
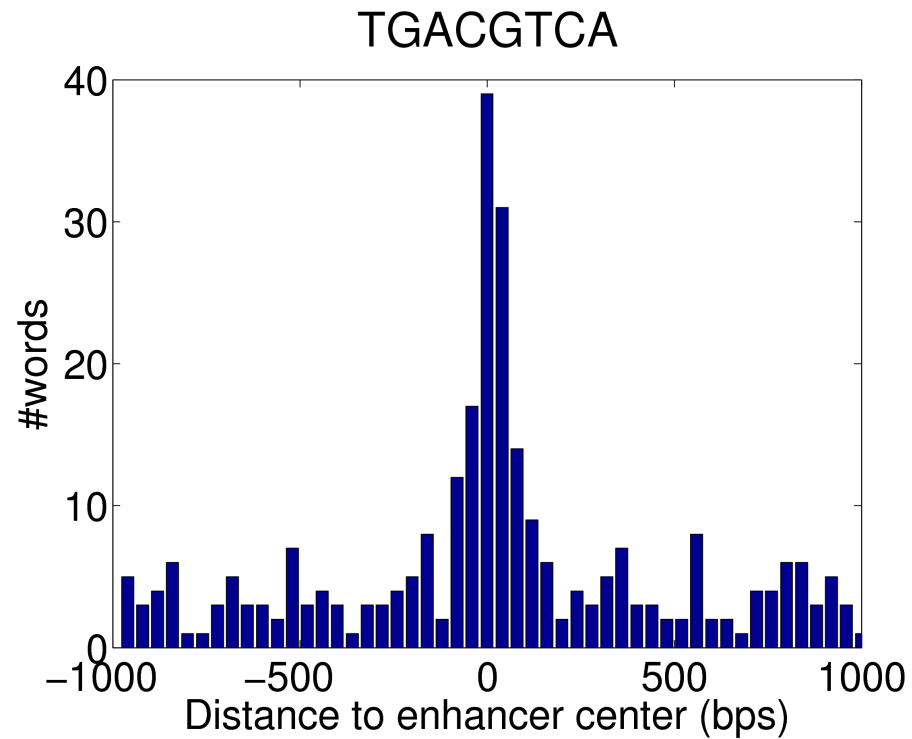
- Calculate enrichment relative to flanks

TCAGGCTGATGACGTCAAACCGTCGTTA  
ACCTTTTGACGTCAAATTTACGCTAGTAT  
TCGACGTAGCTAGCATGATCGATAGATC  
CGTGACGTCAAGTGCTCGTAAATCATAAG



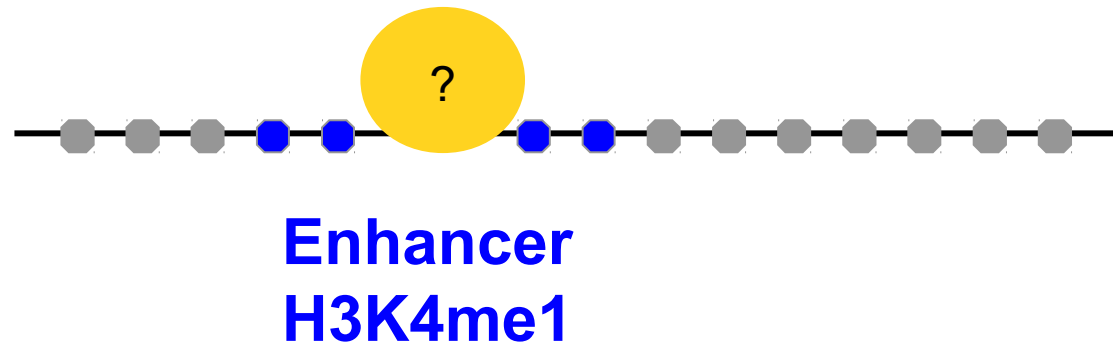


~100 enriched motifs found at enhancers

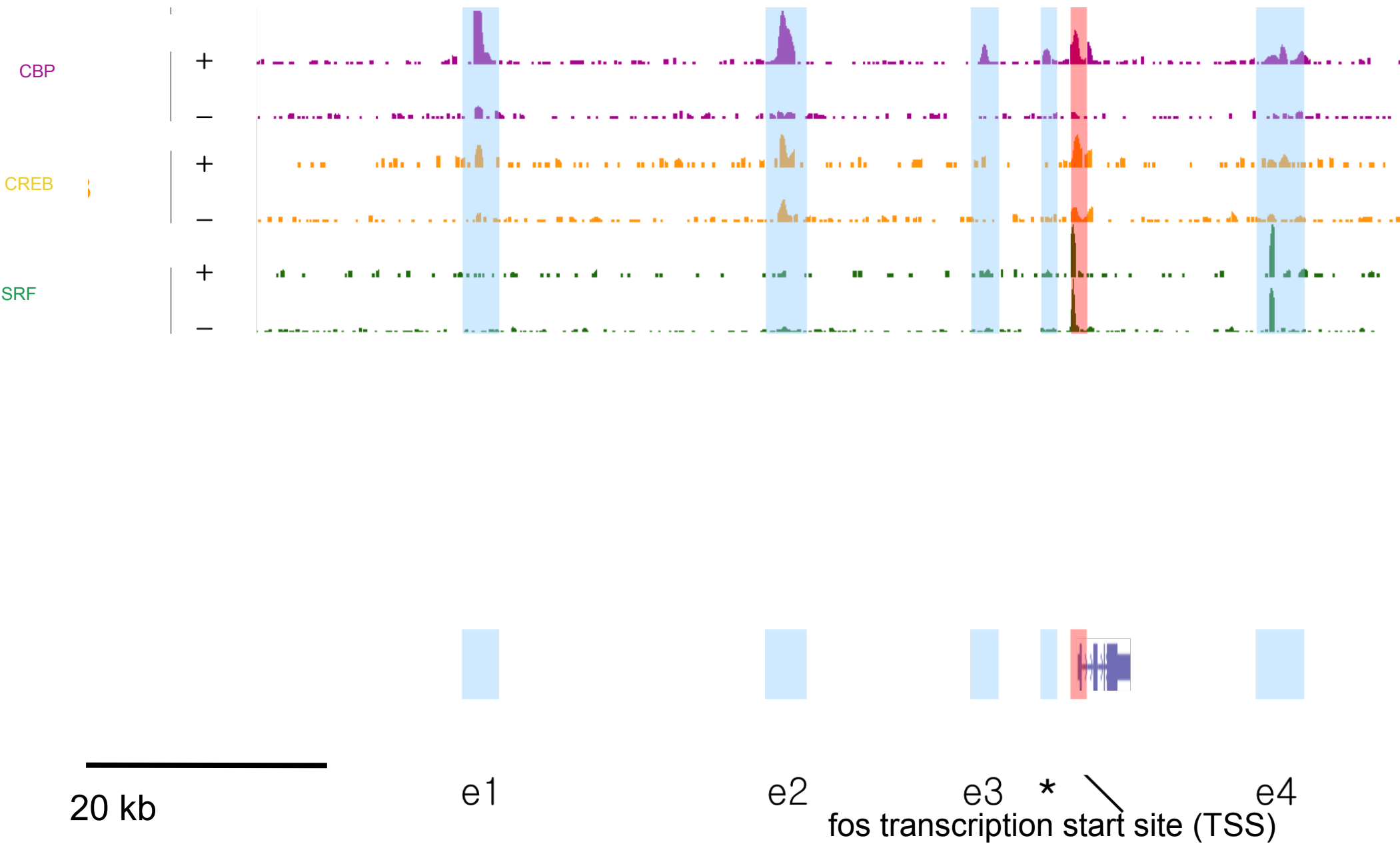


# What TFs are bound at enhancers?

- Enriched for ~100 sequence motifs

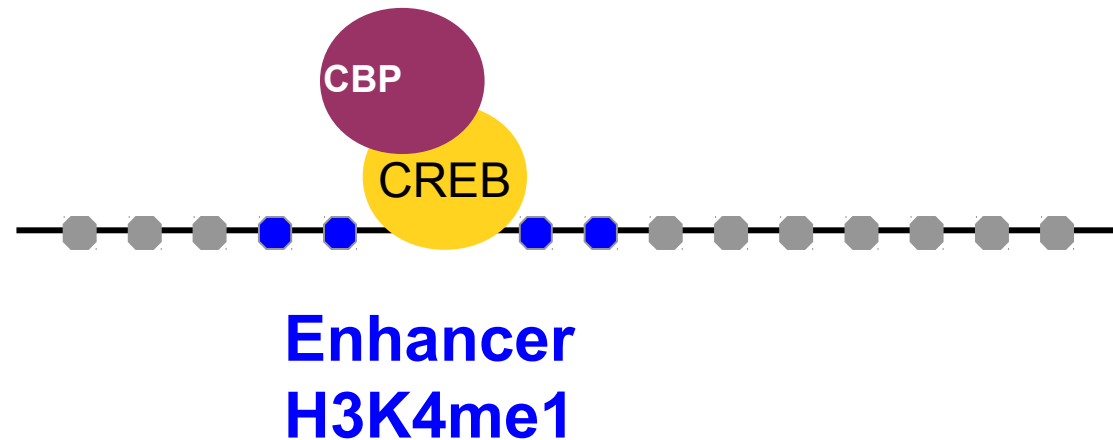


# SRF and CREB binding at Fos enhancers

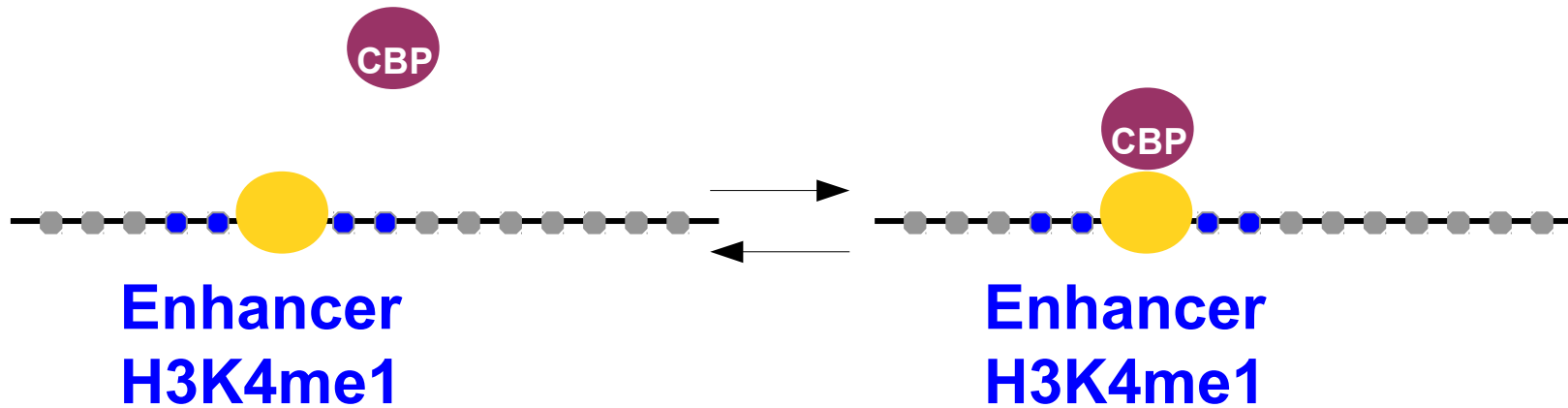


# Is CBP binding determined by other TFs?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence

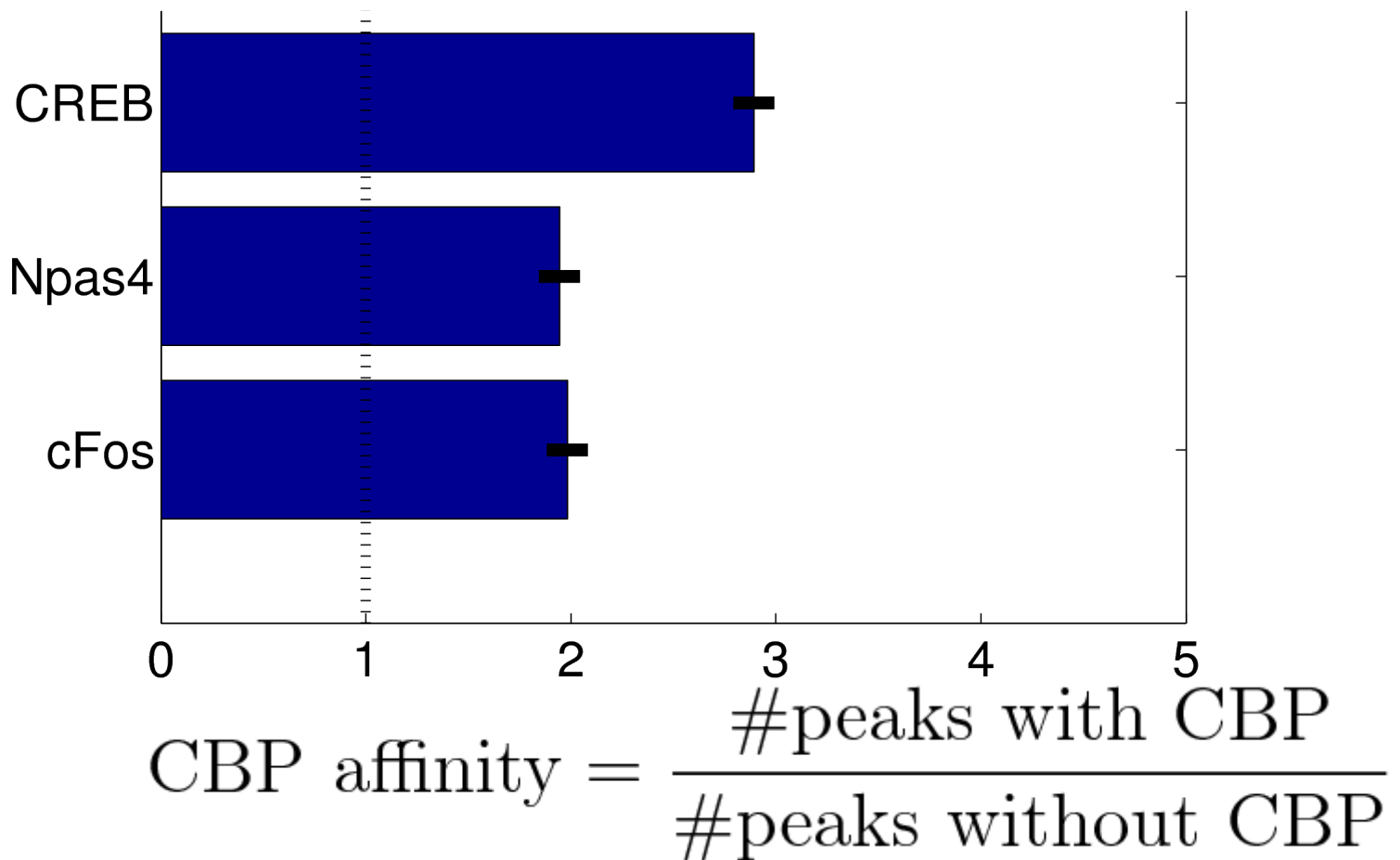


# TFs compete for CBP

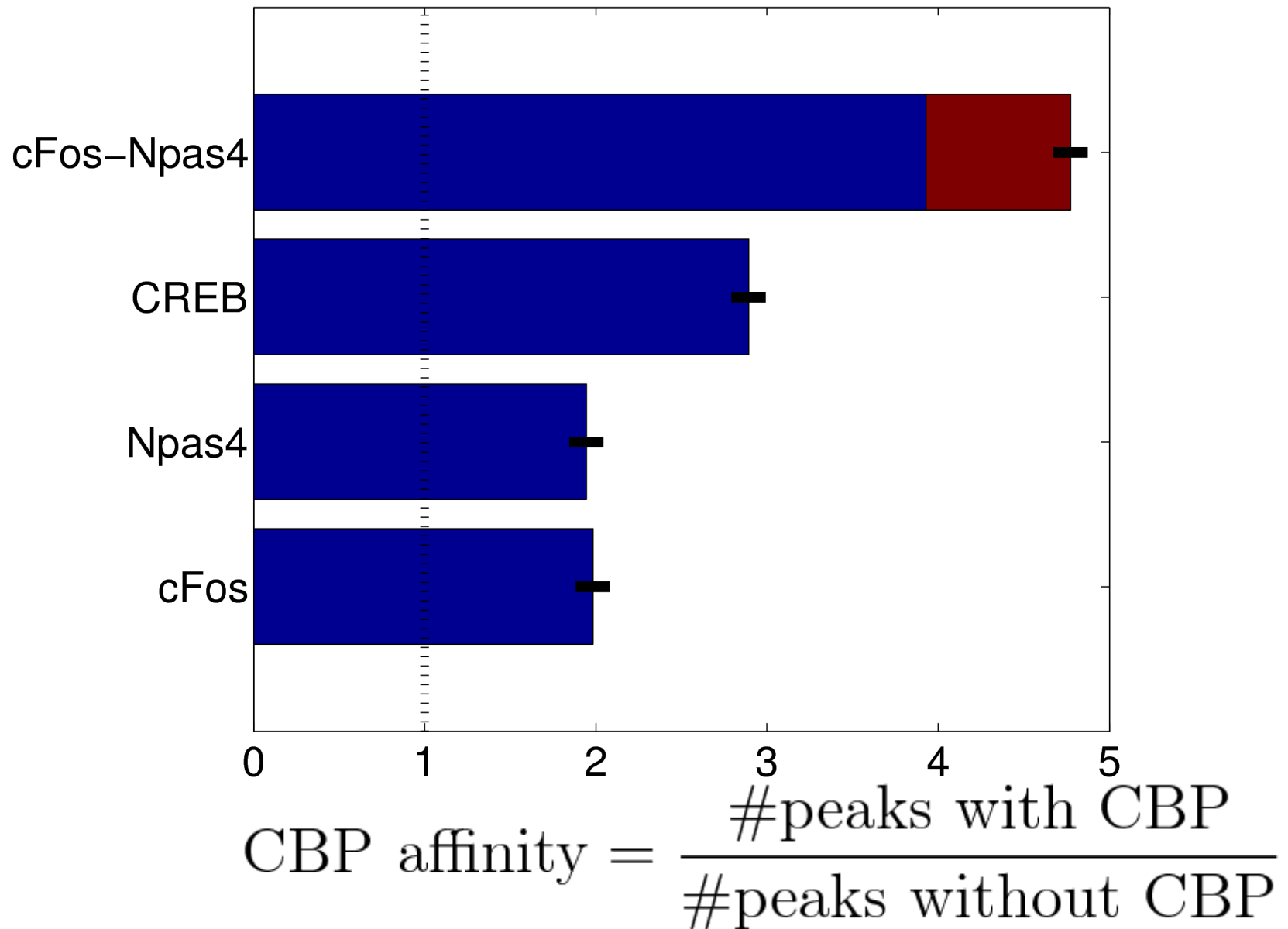


$$\text{CBP affinity} = \frac{\# \text{peaks with CBP}}{\# \text{peaks without CBP}}$$

# CBP binding determined by affinity of TF

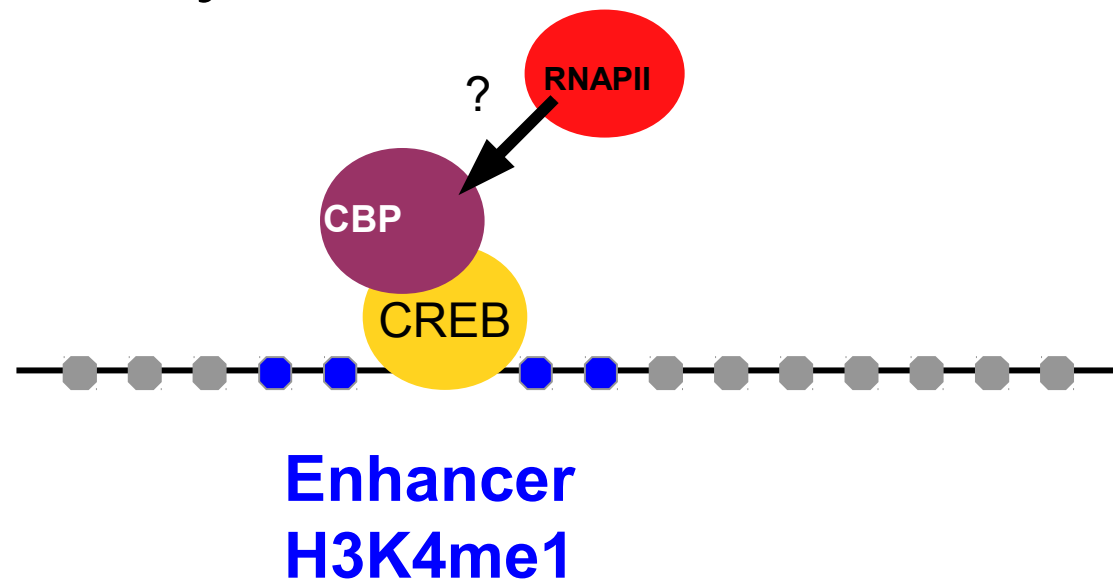


# Synergistic effects for combinations of TFs



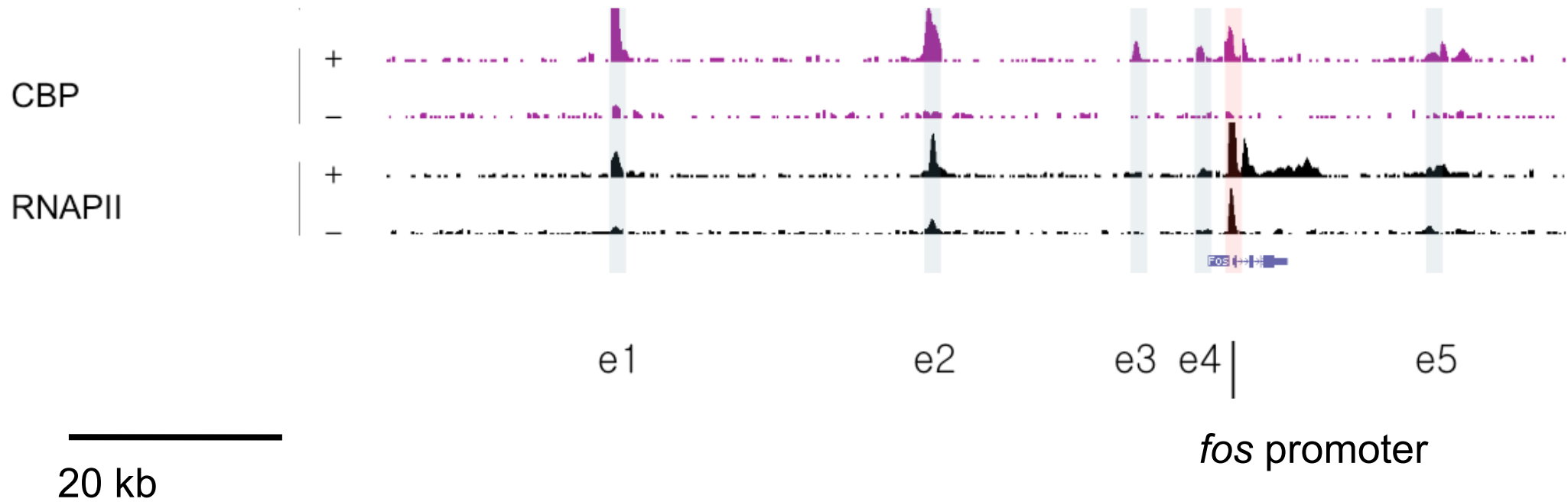
# What is the function of CBP at enhancers?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs

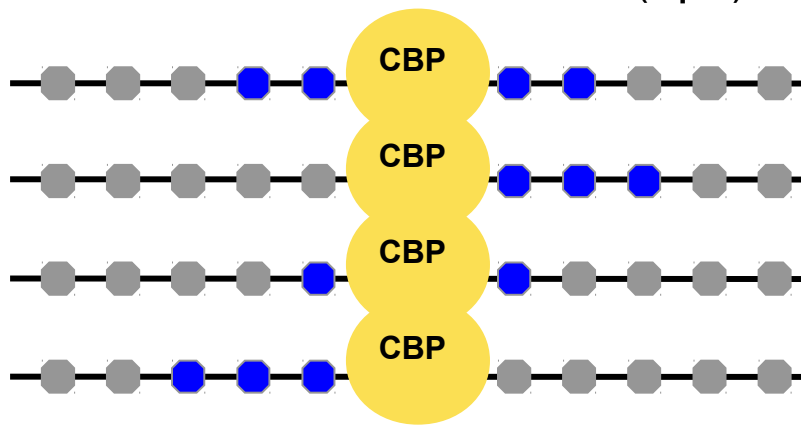
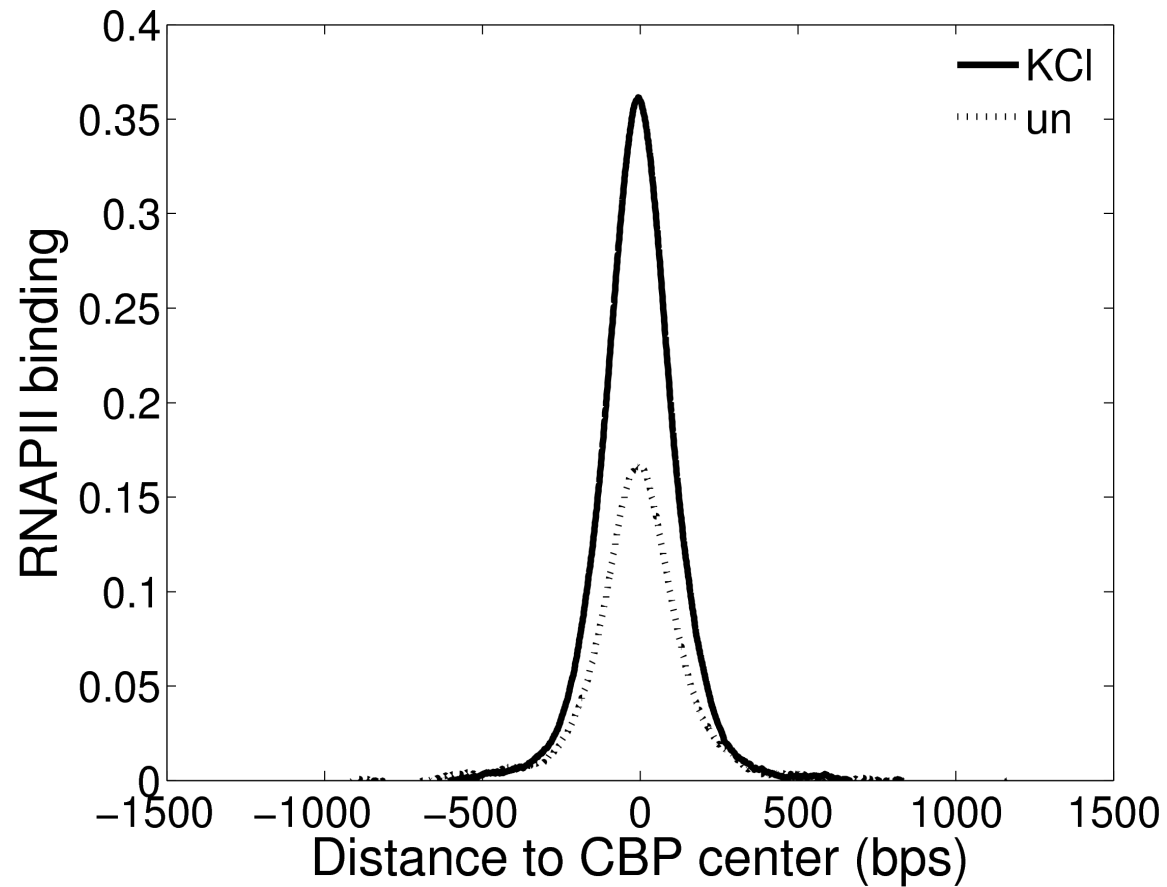




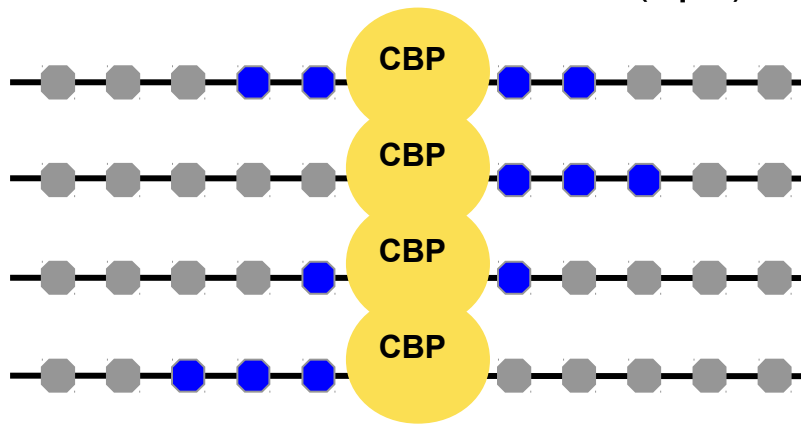
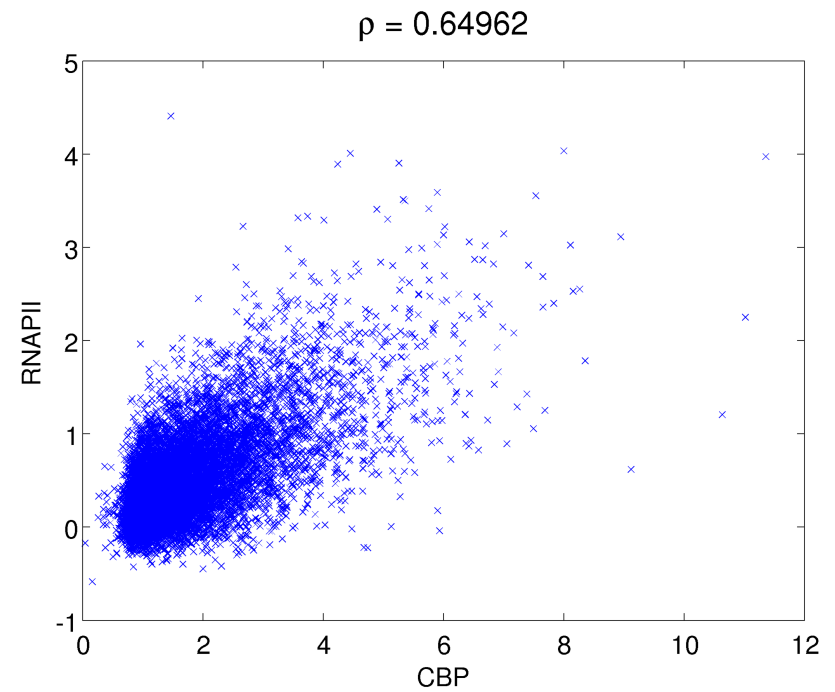
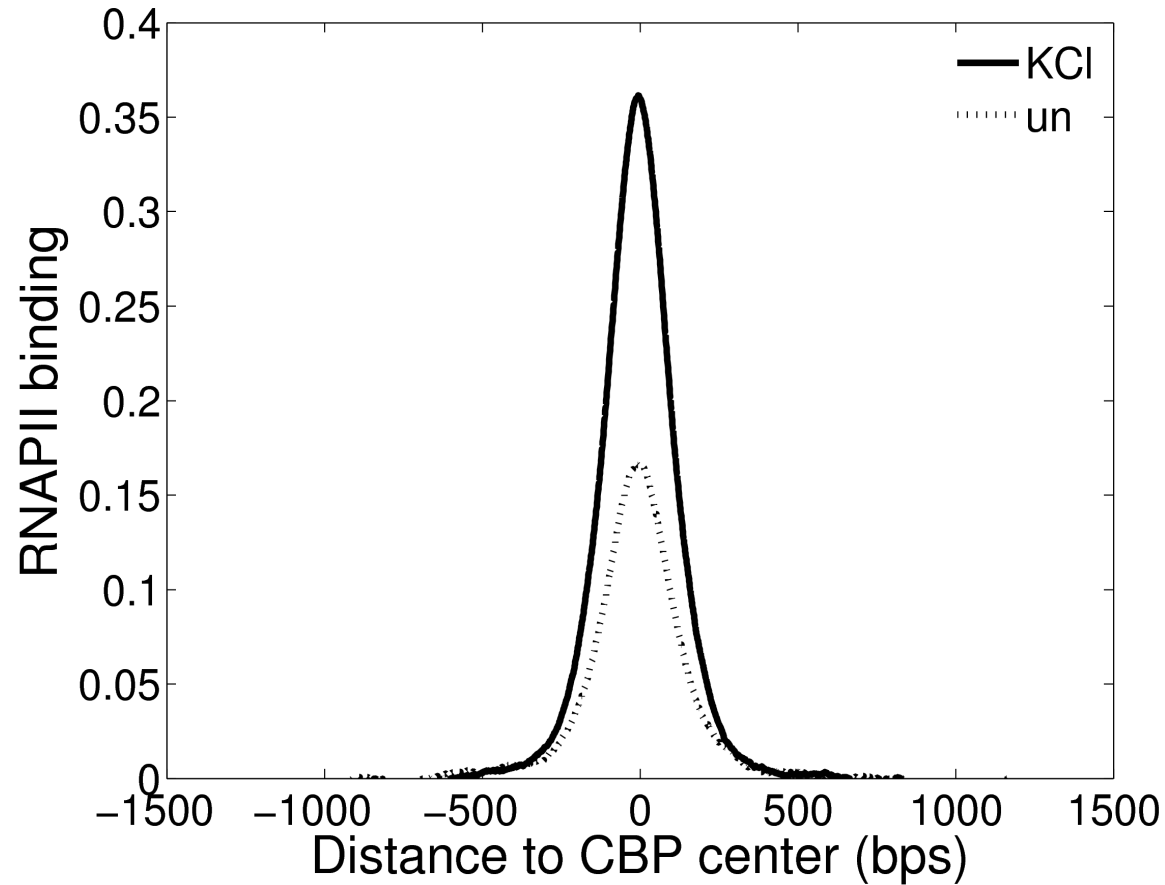
# RNAPII is recruited to CBP binding sites at the *fos* locus



# RNAPII is recruited at enhancers

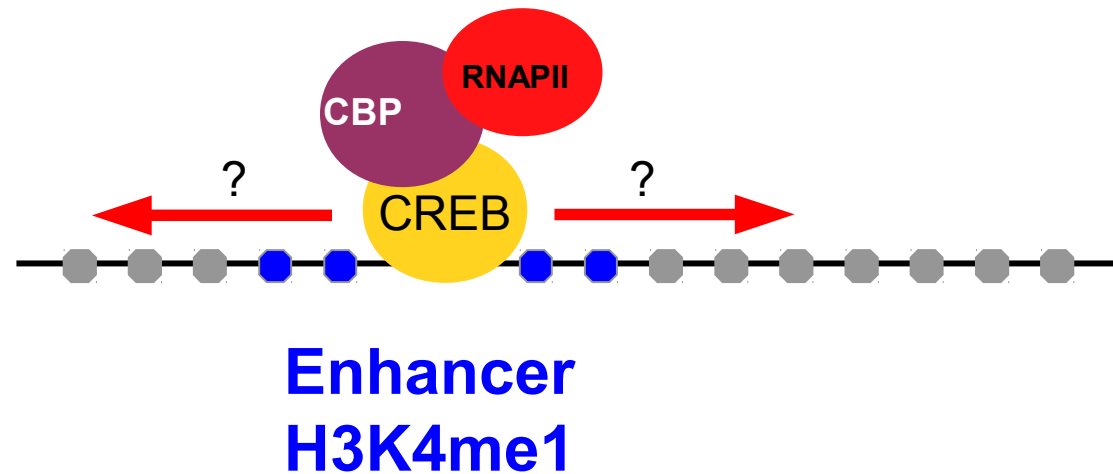


# RNAPII is correlated with CBP



# What is the function of RNAPII at enhancers?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII



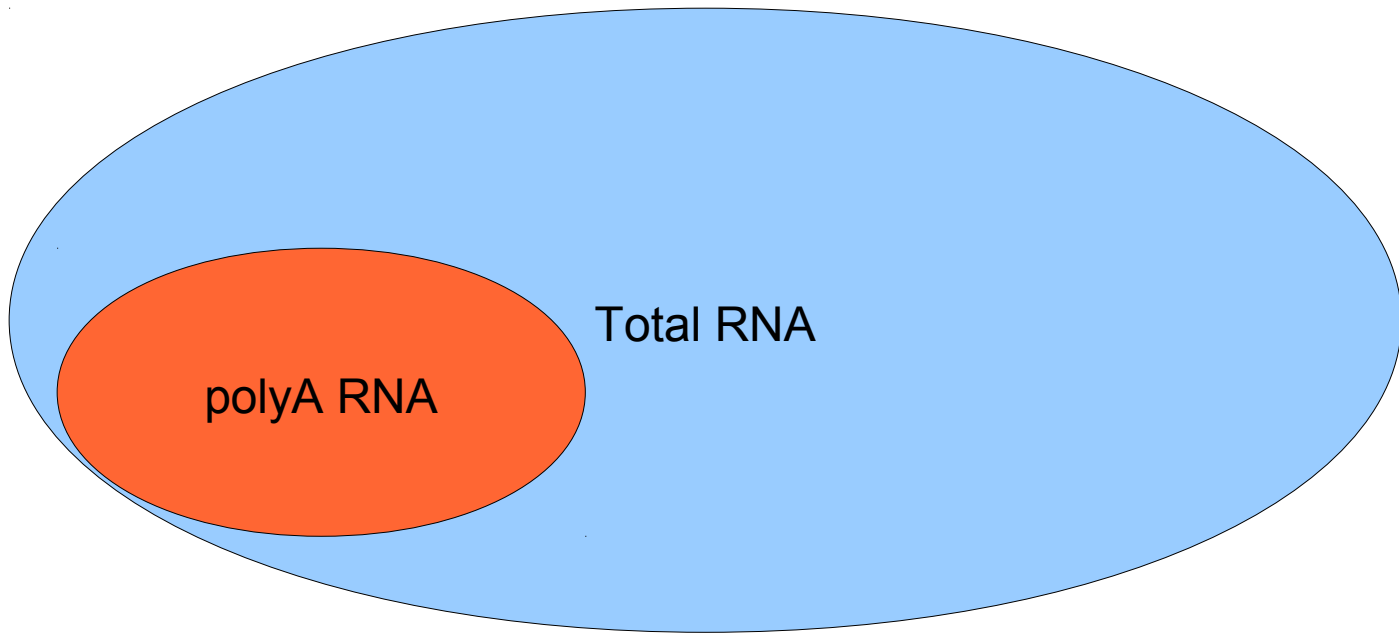
# RNA-Seq finds transcribed parts of the genome

- Short **reads** mapped to reference genome
- #reads ~ RNA



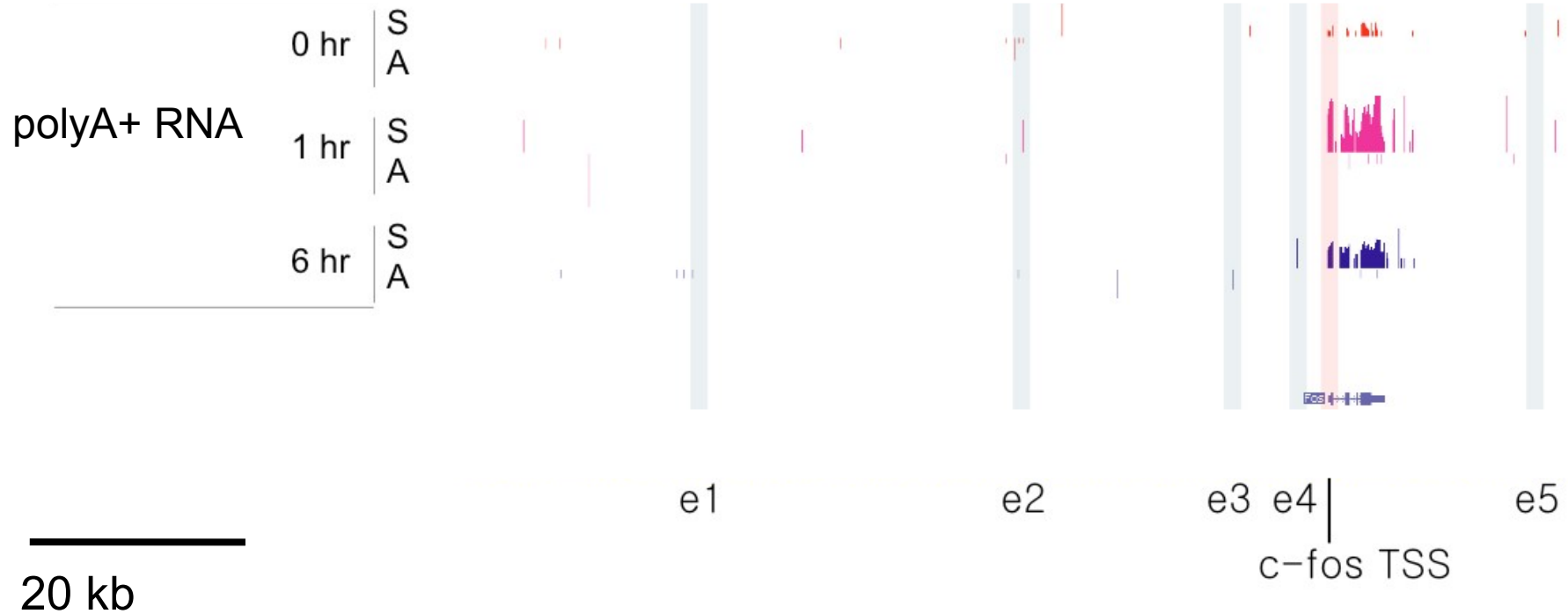
# polyA tail is added to messenger RNAs (mRNAs)

ACGUUUGUACCUAGCUAGCUUACGAG AAAAAAAAAAAAAAAAAAAAAA

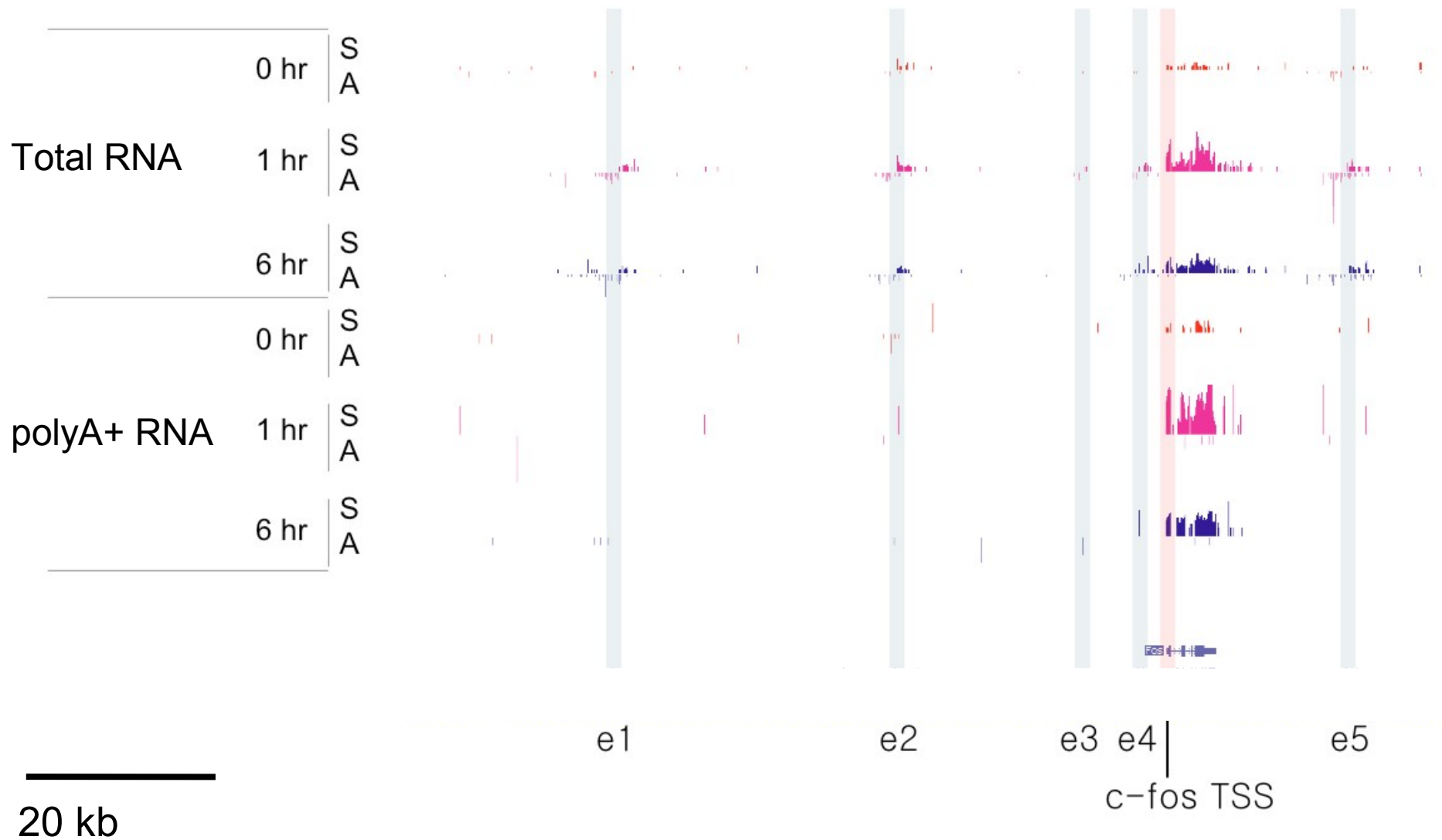


# Transcription of mRNA at the *fos* locus

ACGUUUGUACCUAGCUAGCUUACGAG AAAAAAAAAAAAAAAAAAAAAA

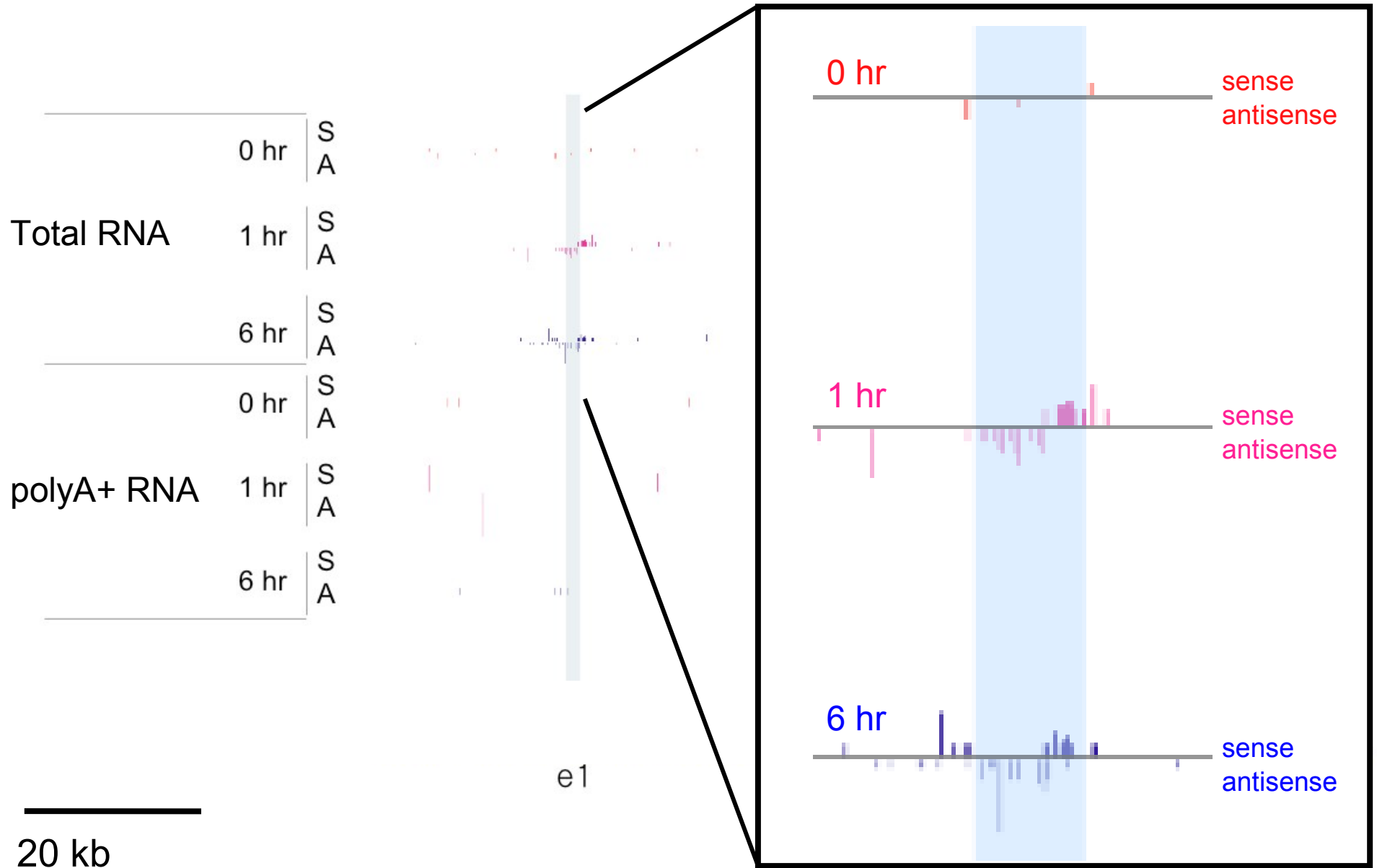


# Transcription of total RNA at the *fos* locus



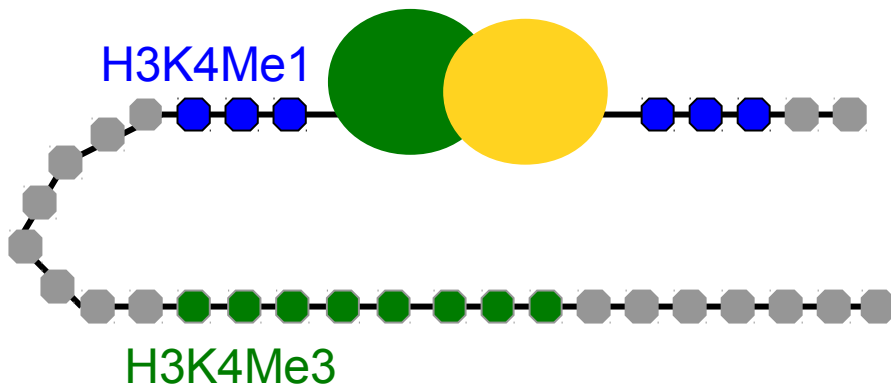


# Transcription at enhancers is activity-dependent

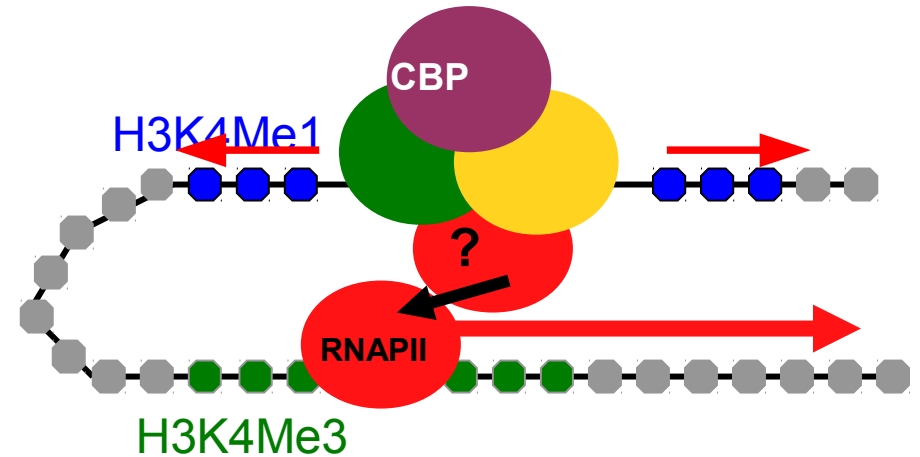


# Enhancer RNAs (eRNAs) novel species

Before neuronal activation

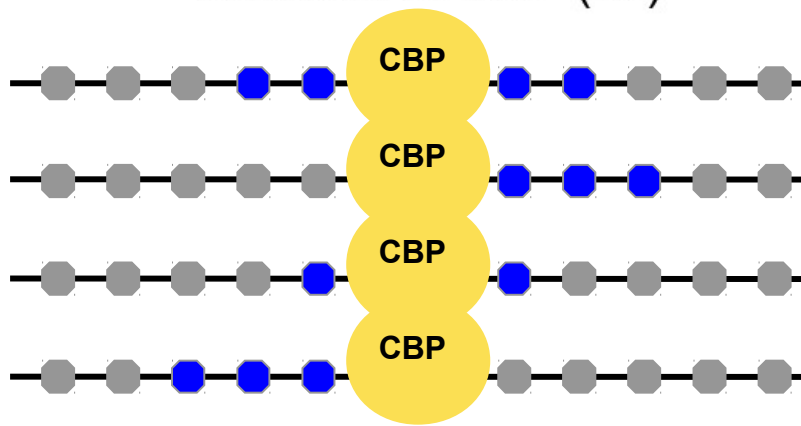
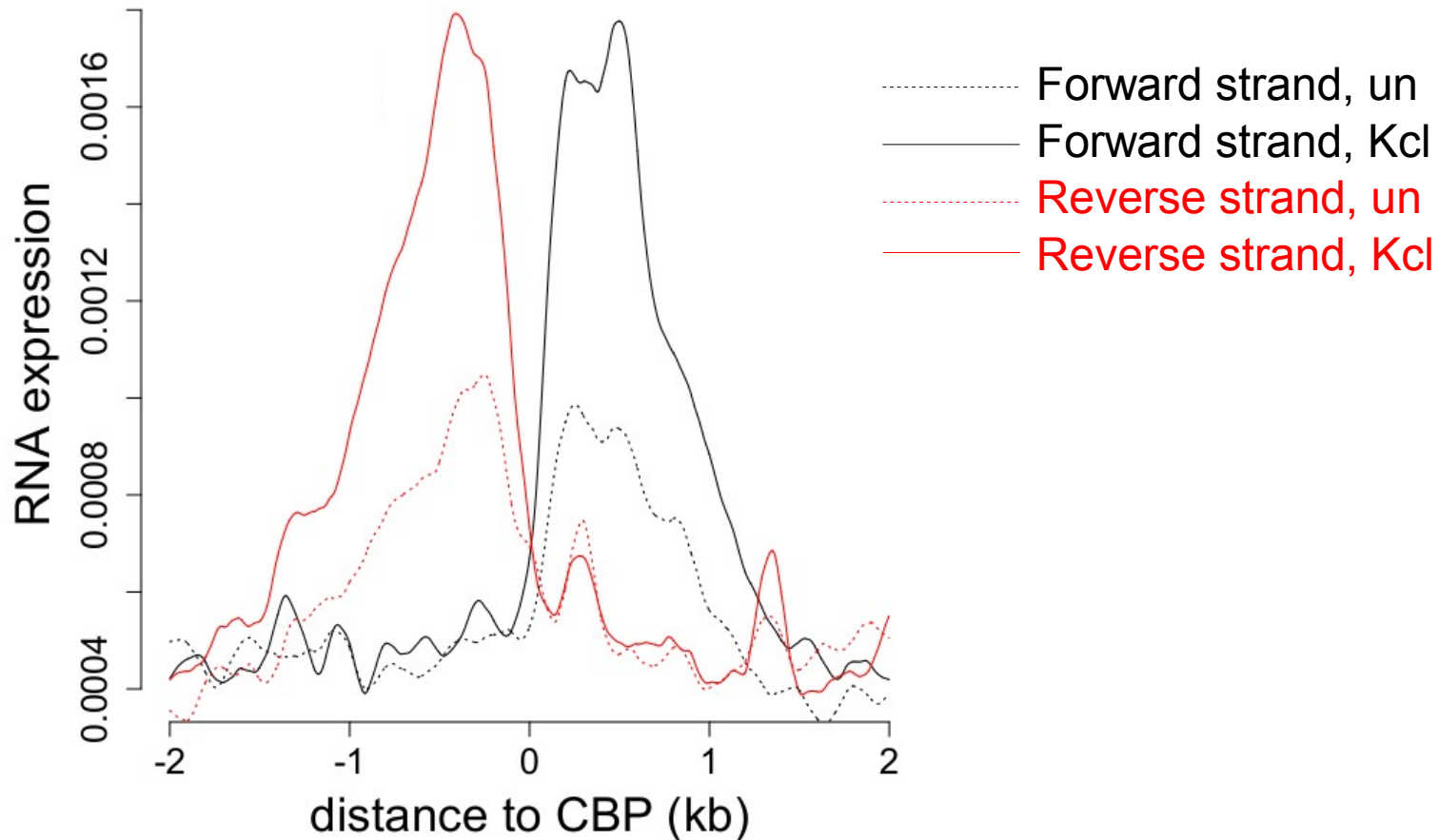


After neuronal activation

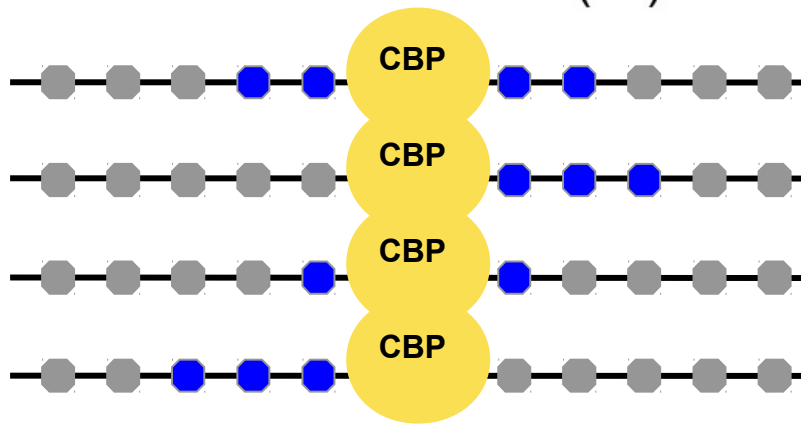
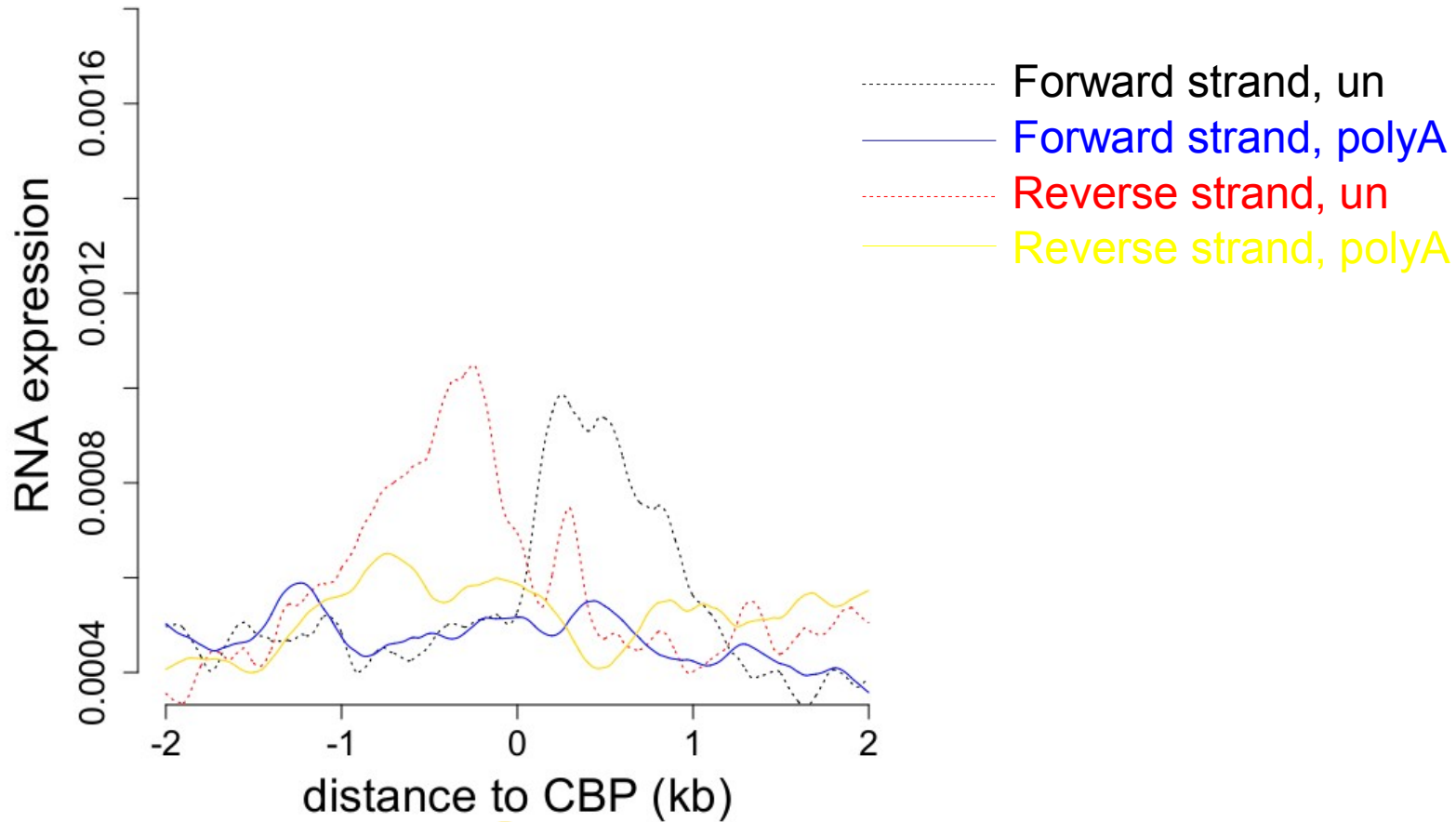


- mRNA, rRNA, tRNA, miRNA, snRNA, snoRNA, siRNA, piRNA, lncRNA, ... ?

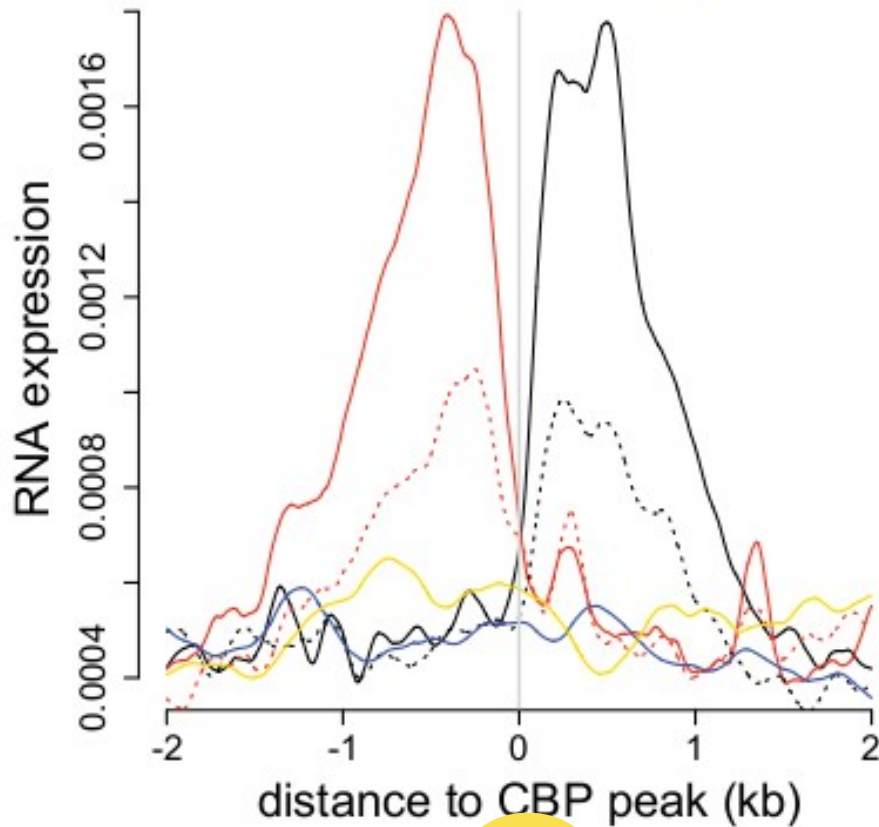
# eRNAs are induced by activity



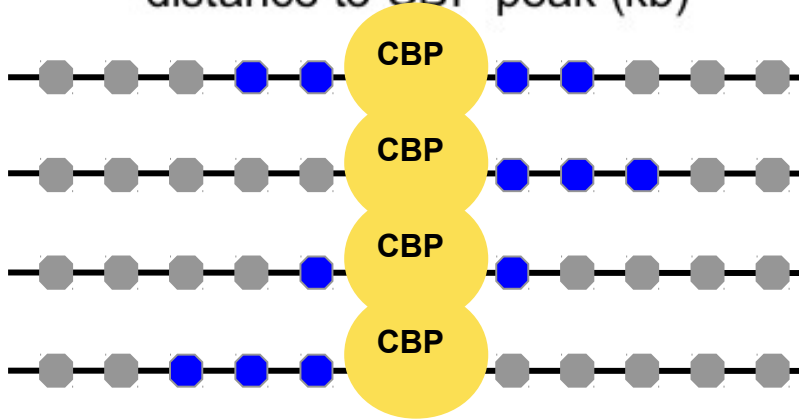
# eRNAs are not polyadenylated



# eRNAs are 100-fold lower than mRNAs



- 1 in 10k reads eRNA
- Inducible
- ~1 kb
- Bidirectional
- No polyA-tail
- Not protein-coding

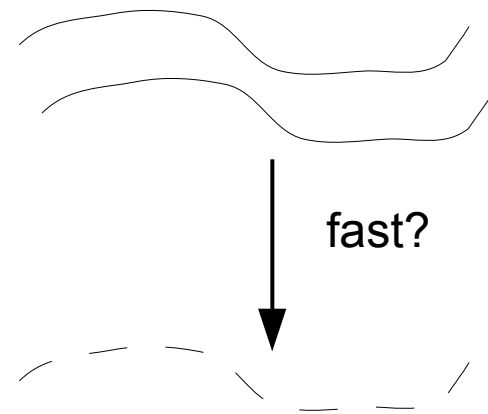
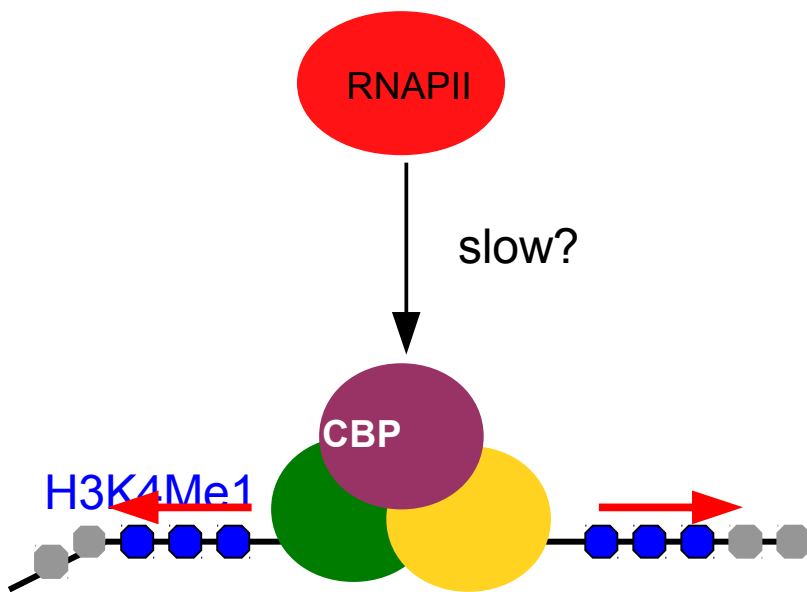


# Why do eRNAs have such low abundance?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII
- RNAPII synthesizes eRNAs

# Why do eRNAs have such low abundance?

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA



# A simple model of transcription

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \underbrace{\frac{P_M k}{L_M}}_{\text{production}} - \underbrace{\frac{M}{\tau_M}}_{\text{decay}}$$



# Parameters are straightforward to estimate

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} M - \frac{M}{\tau_M}$$

The diagram illustrates the biological meaning of the parameters in the mRNA production and decay equation. Arrows point from the following labels to the corresponding terms in the equation:

- RNAPII** points to the  $P_M$  parameter in the production term.
- Length of transcript** points to the  $L_M$  parameter in the production term.
- Elongation rate** points to the  $k$  parameter in the production term.
- half-life** points to the  $\tau_M$  parameter in the decay term.
- mRNA** points to the  $M$  variable in the decay term.

## Similar expression for eRNA levels

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$
$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

# Half life of eRNAs relative to mRNAs

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$
$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

$$\frac{\tau_E}{\tau_M} = \frac{E^*}{M^*} \frac{L_E}{L_M} \frac{P_M}{P_E}$$

eRNAs half life is less than half an hour

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$
$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

$$\frac{\tau_E}{\tau_M} = \frac{E^*}{M^*} \frac{L_E}{L_M} \frac{P_M}{P_E}$$

$$\tau_E \sim 10^{-2} \times 1 \times 2 \times \tau_M \sim 2 \times 10^{-2} \times 600\text{min} = 12\text{min}$$

# Estimate consistent with experiments

- eRNA production much slower than mRNA
- eRNA decay much faster than mRNA

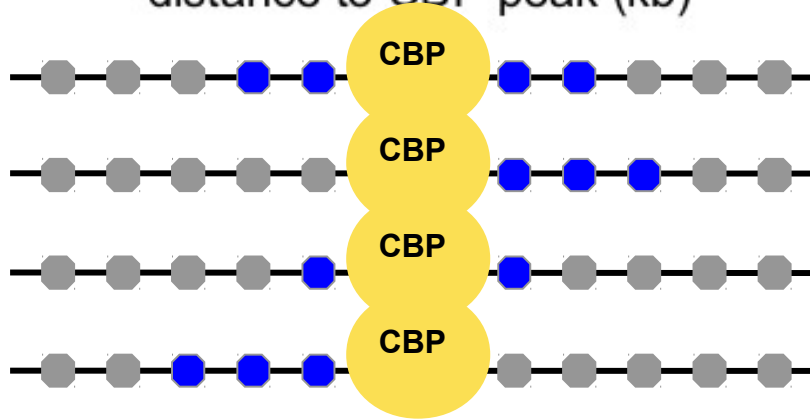
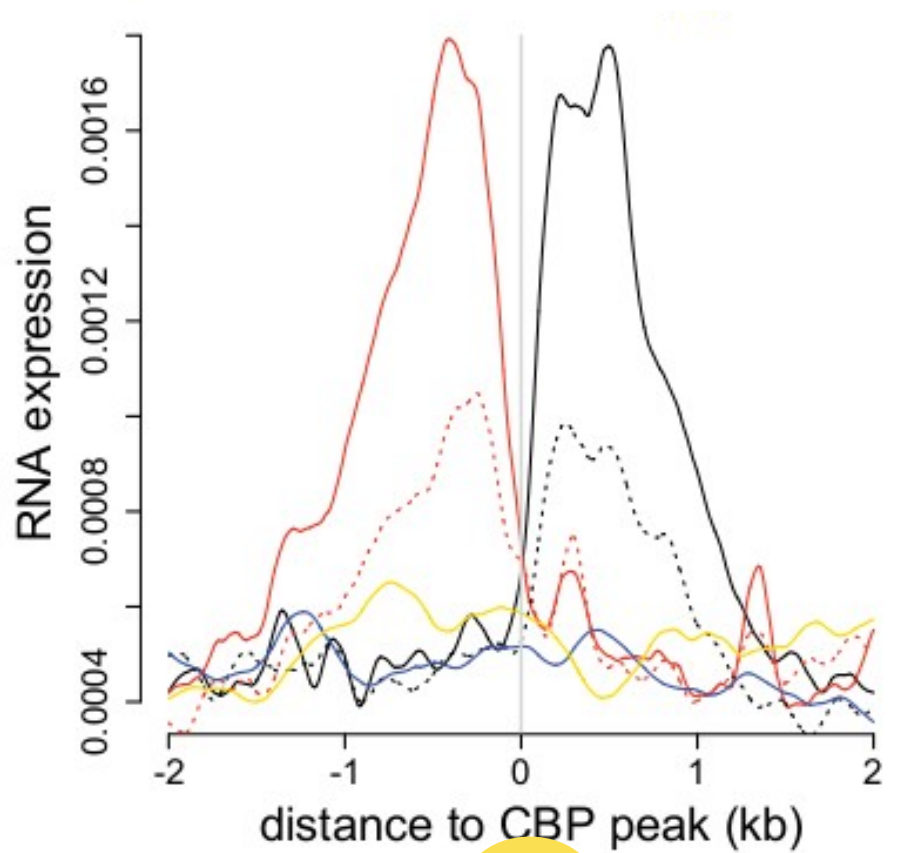
$$\frac{dM}{dt} = \frac{P_M k}{L_M} - \frac{M}{\tau_M}$$
$$\frac{dE}{dt} = \frac{P_E k}{L_E} - \frac{E}{\tau_E}$$

$$\frac{\tau_E}{\tau_M} = \frac{E^*}{M^*} \frac{L_E}{L_M} \frac{P_M}{P_E}$$

$$\tau_E \sim 10^{-2} \times 1 \times 2 \times \tau_M \sim 2 \times 10^{-2} \times 600\text{min} = 12\text{min}$$

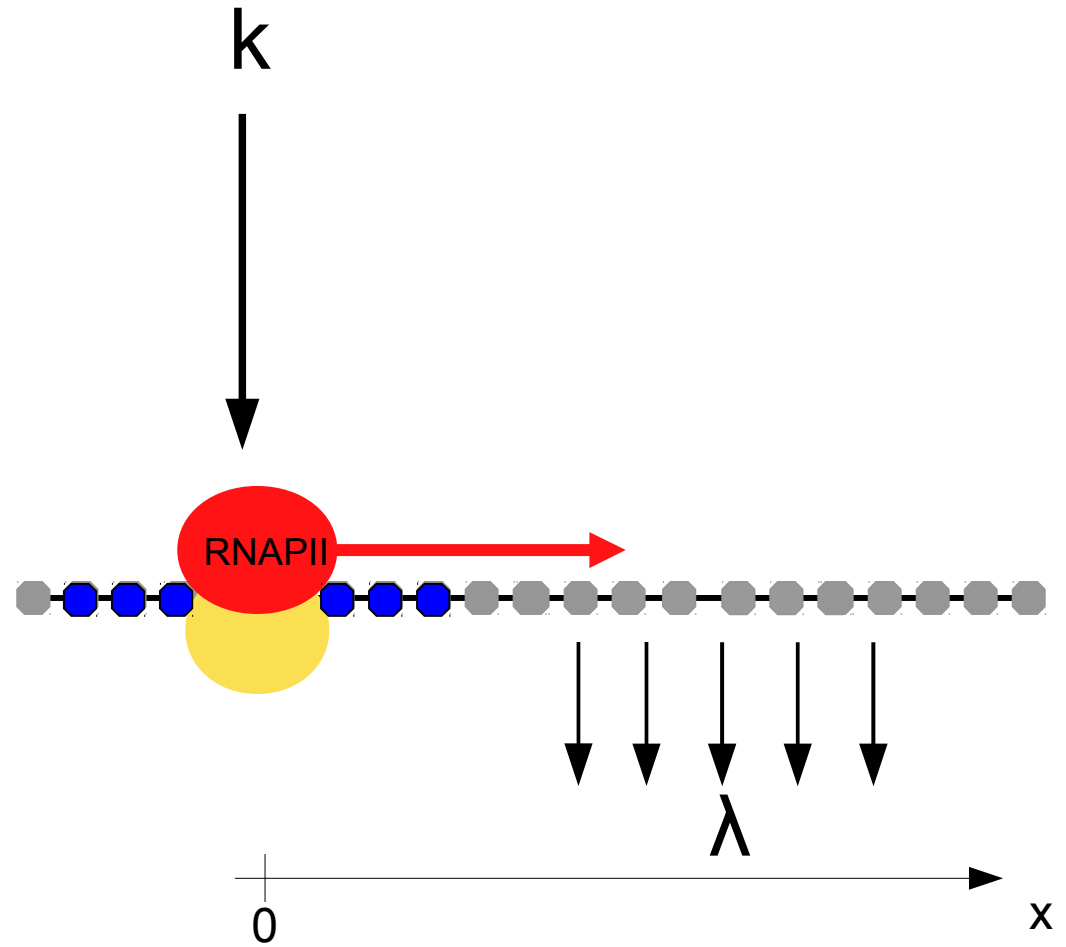
Finally we measured the stability of these transcripts using an actinomycinD chase. In comparison to both the mRNAs generated by the associated protein-coding genes and some known lncRNAs (like Xist and Neat), the upstream non-coding transcripts were very unstable, being reduced by 80% to 90% after a 30 min actinomycinD treatment (indicating a half-life lower than 7.5 min) (Figure 3D and Figure S3). High instability of a subset of lncRNAs both in yeast and mammals mainly depends on degradation by the nuclear exosome [39,40] and often results in the generation of more stable short RNA products [41], which in principle might be responsible for downstream functional effects.

# A mechanistic model of eRNA levels



RNAPII binds and falls off at a constant rate

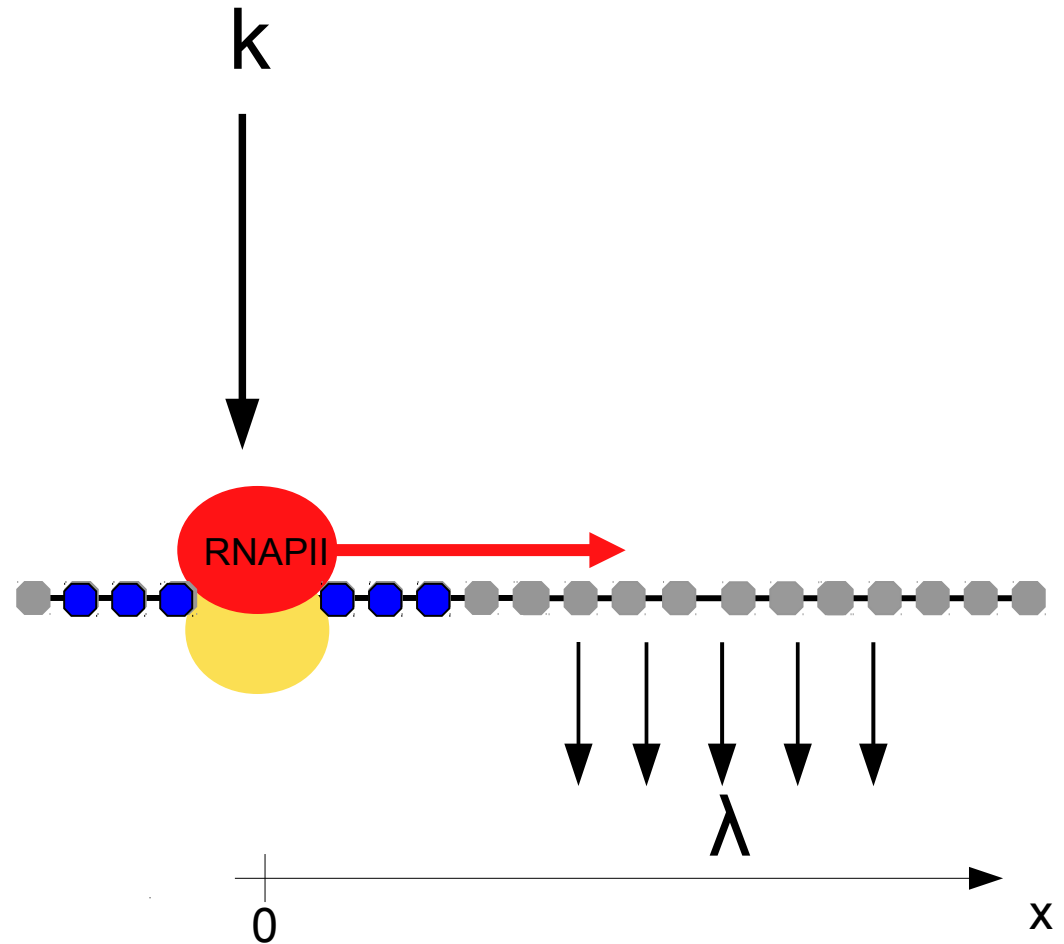
$$\frac{dP}{dx} = k - \lambda P$$



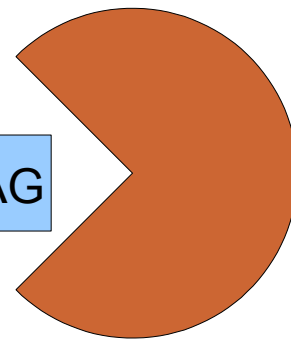
# eRNA production proportional to RNAPII

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



ACGUUUGUACCUAGCUAGCUUACGAG

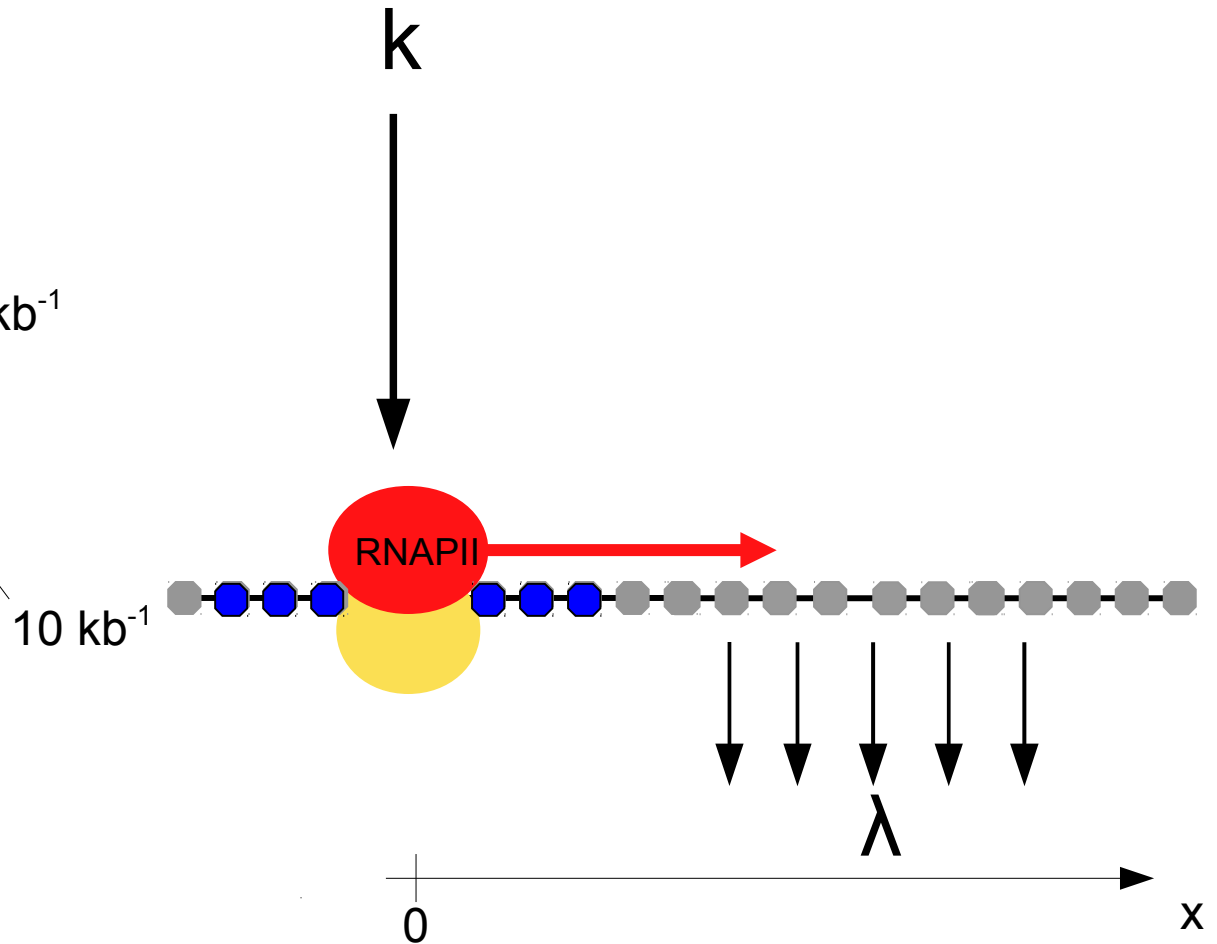




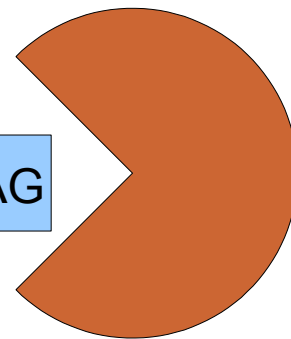
# Parameters can be estimated from literature

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



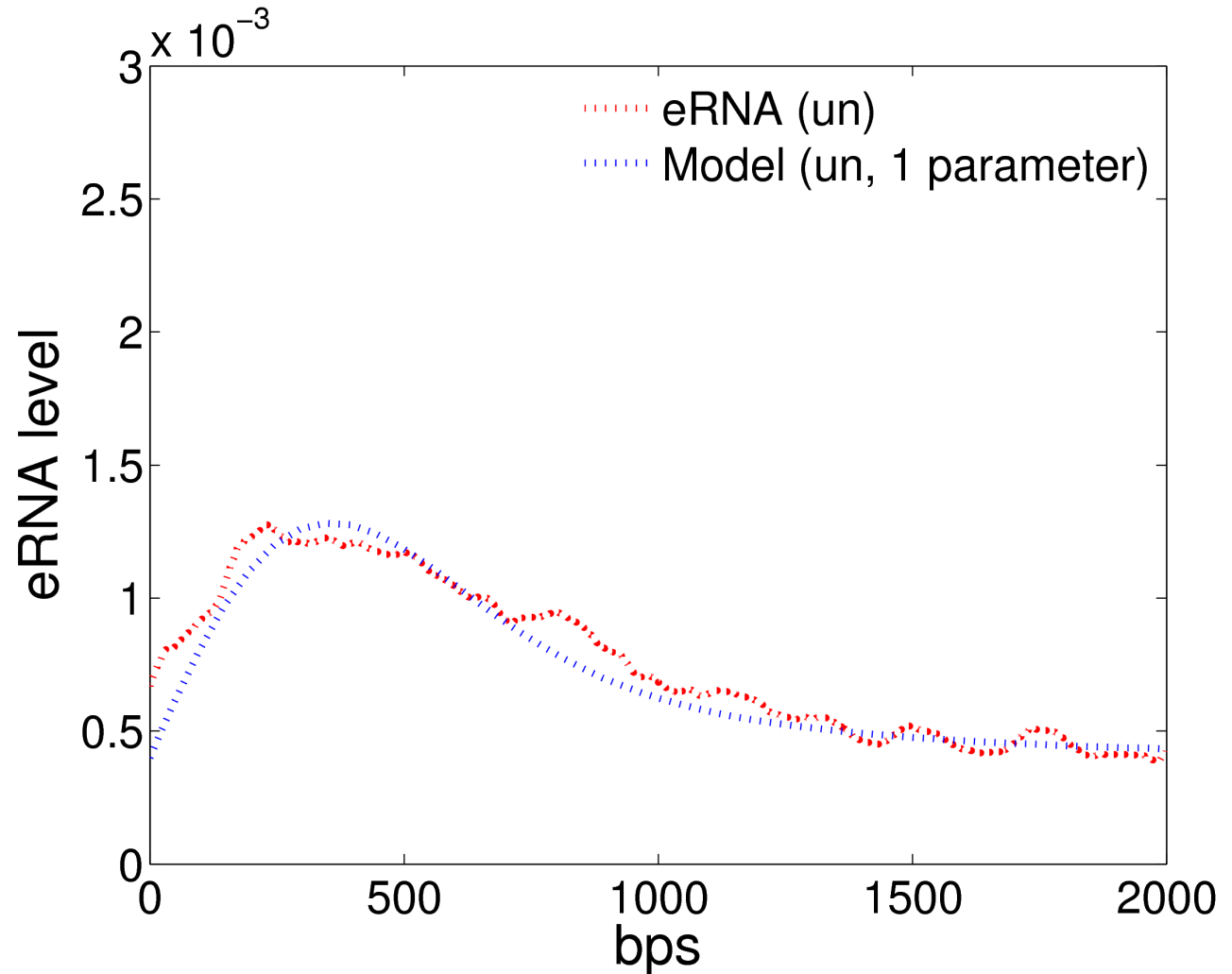
ACGUUUGUACCUAGCUAGCUUACGAG



# eRNA levels can be accurately predicted

$$\frac{dP}{dx} = k - \lambda P$$

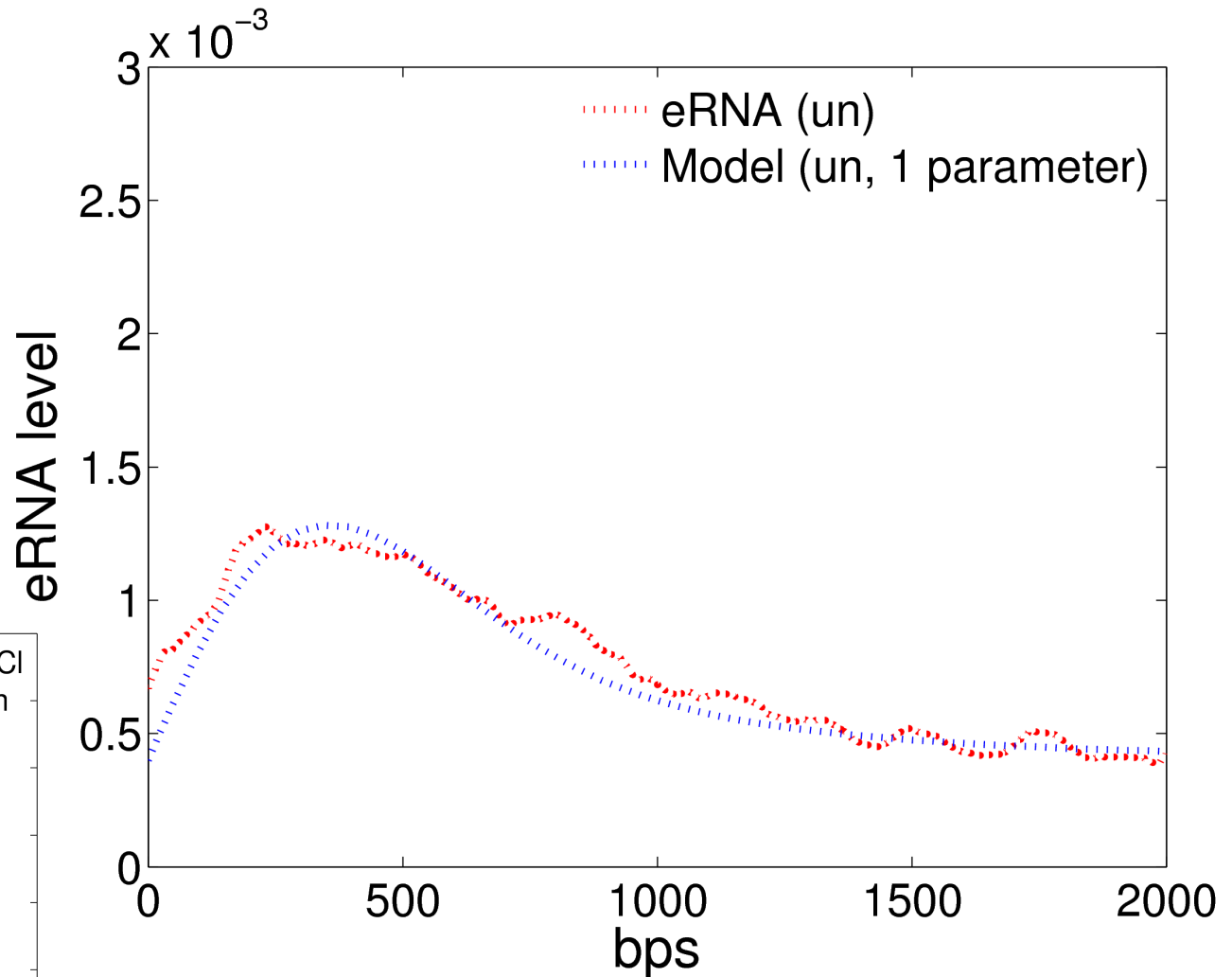
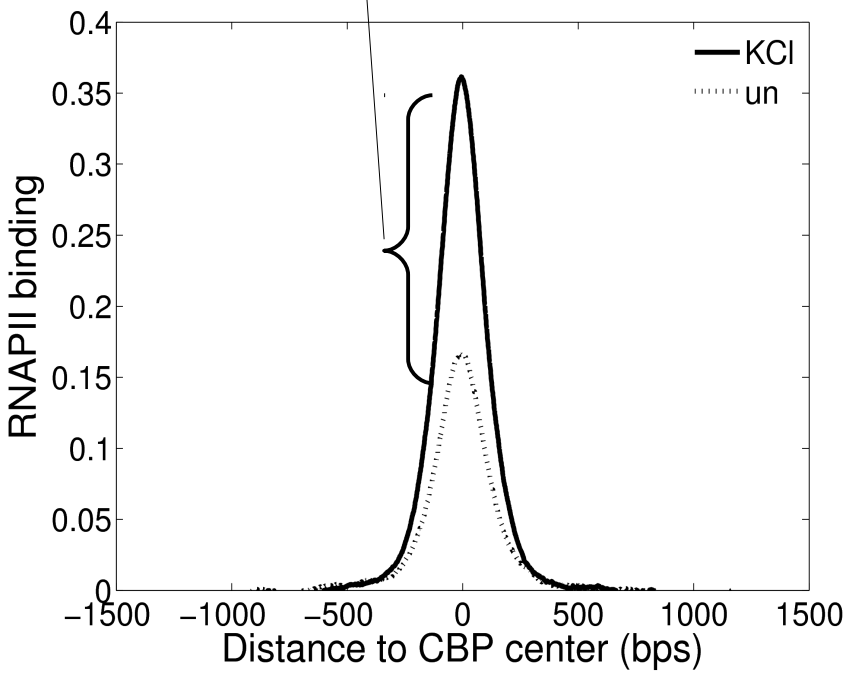
$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



# eRNA levels can be accurately predicted

$$\frac{dP}{dx} = k - \lambda P$$

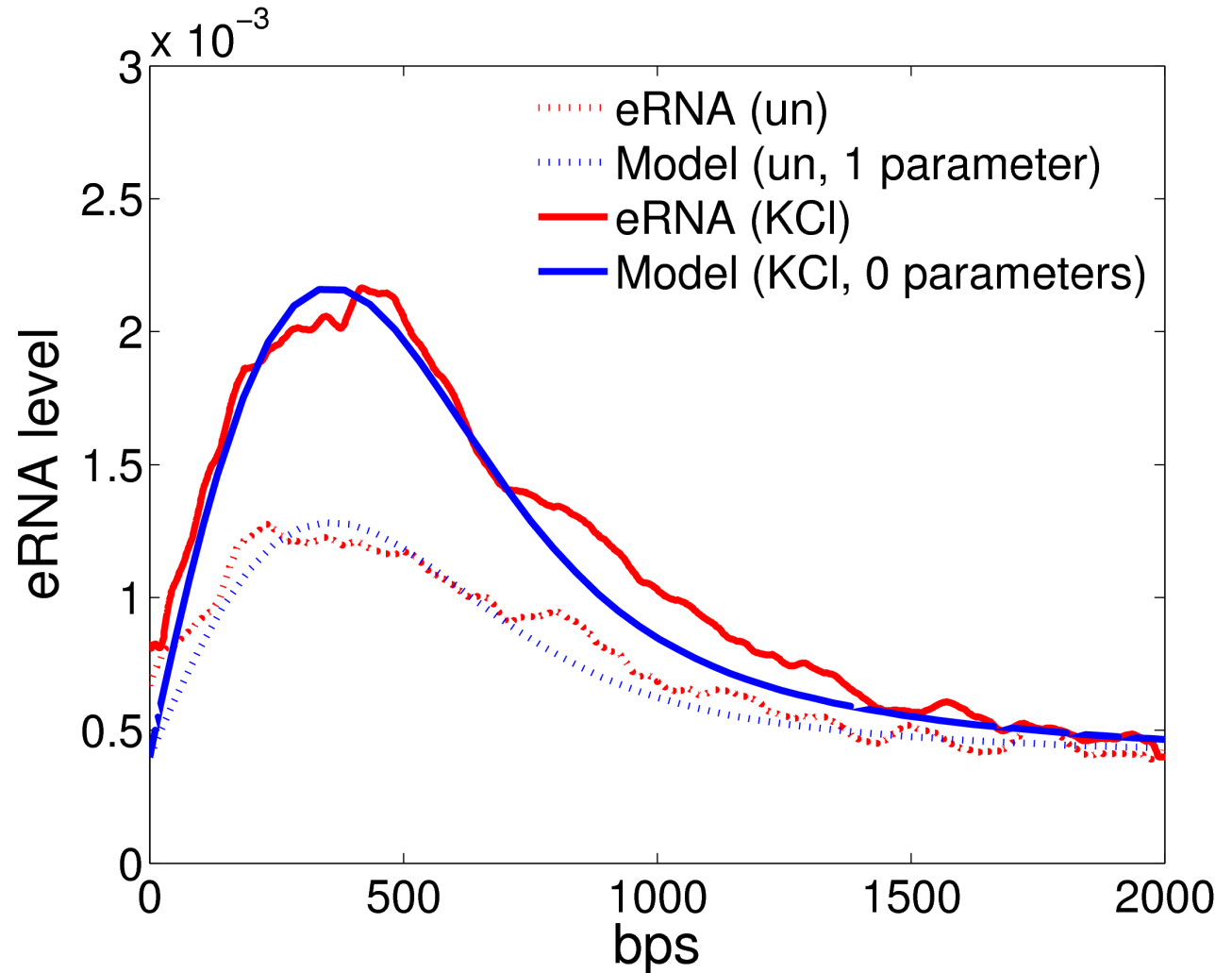
$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



# eRNA levels can be accurately predicted

$$\frac{dP}{dx} = k - \lambda P$$

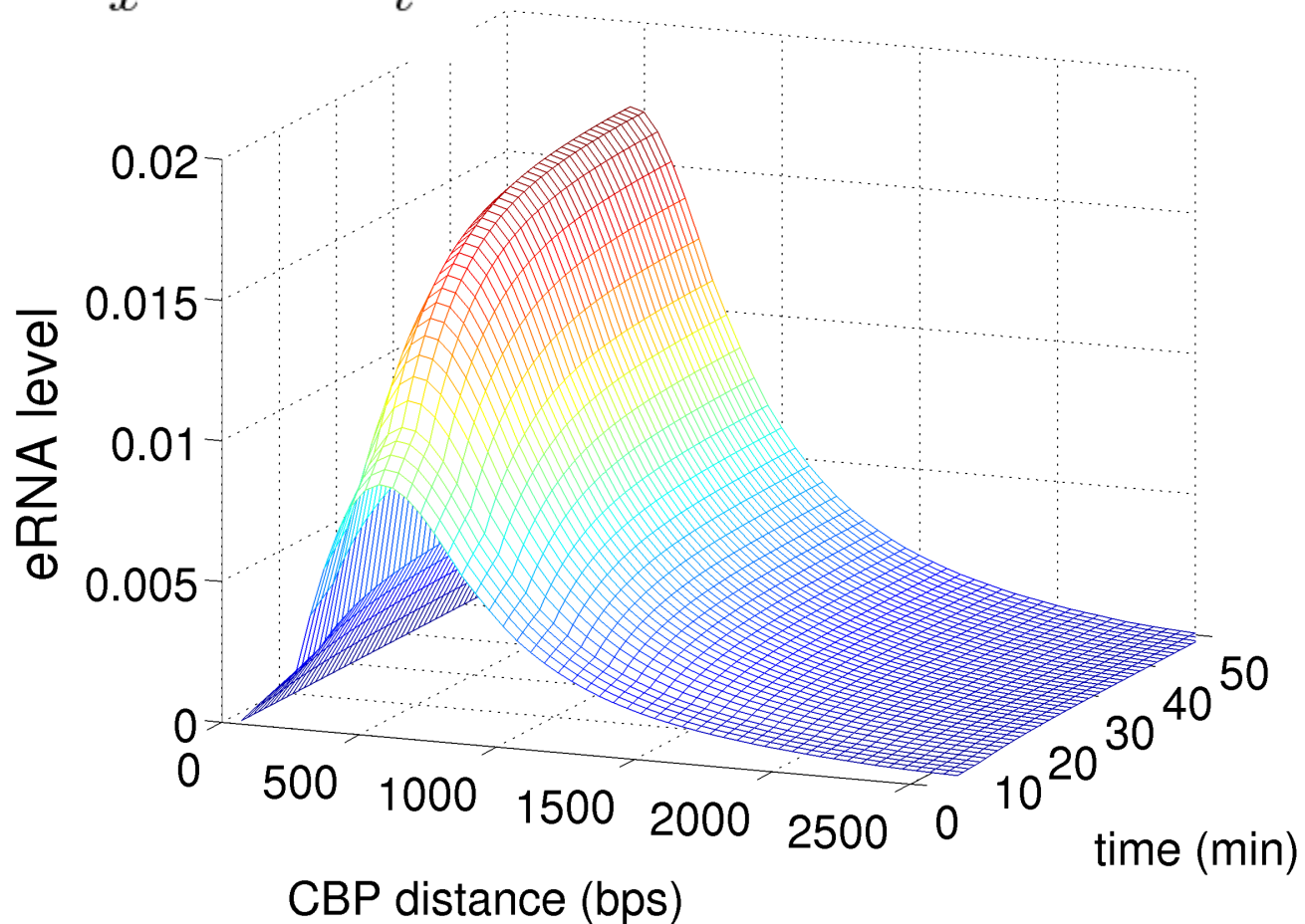
$$\frac{dE}{dx} = \gamma P(x) - \delta x E$$



# A PDE for eRNA levels

$$\frac{\partial P}{\partial x} + \frac{\partial P}{\partial t} = k(x, t) - \lambda_x P - \lambda_t P$$

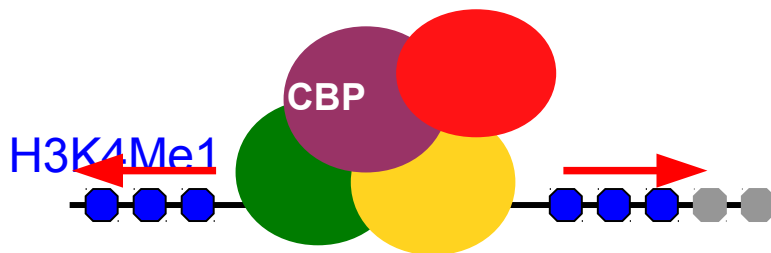
$$\frac{\partial E}{\partial x} + \frac{\partial E}{\partial t} = \gamma P(x, t) - \delta_x x E - \delta_t t$$



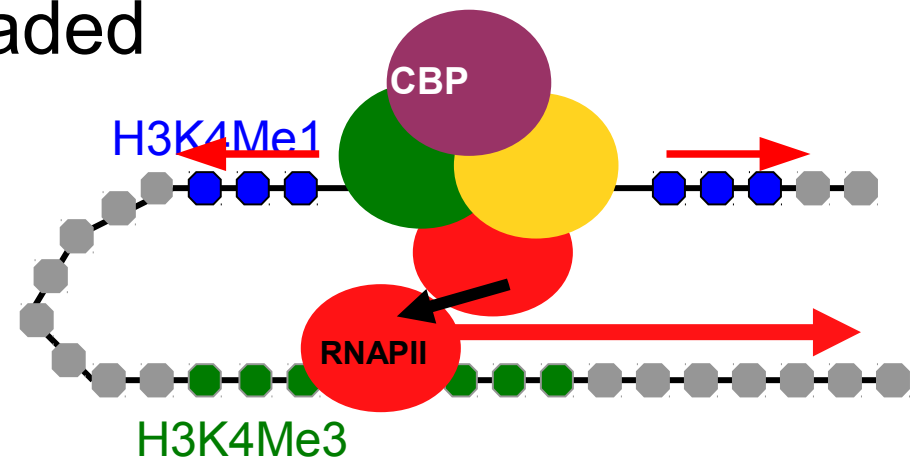
# Enhancers bind RNAPII and produce eRNA, but do they require the promoter?

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII
- RNAPII synthesizes eRNAs

– eRNAs are rapidly degraded

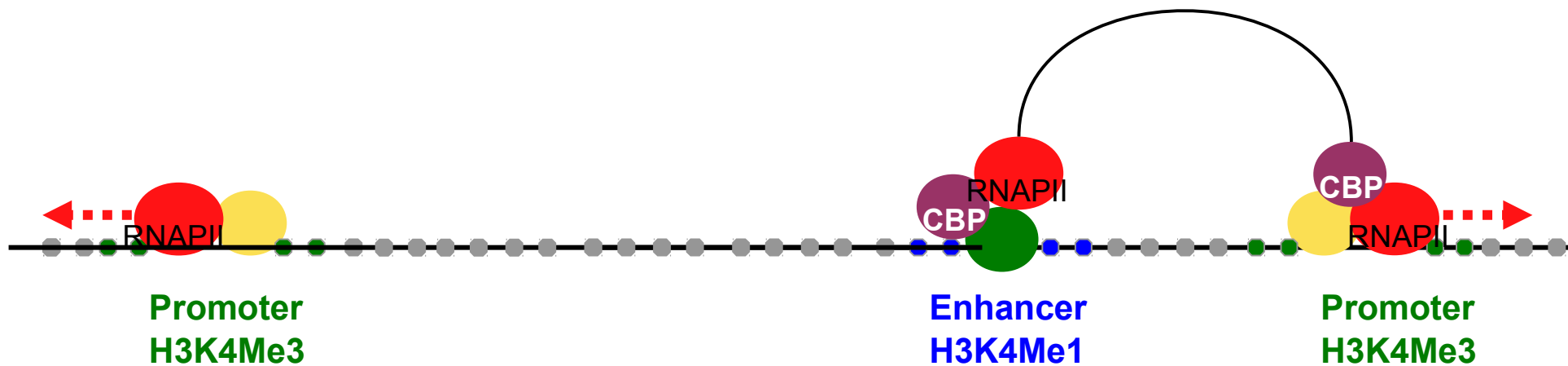


?



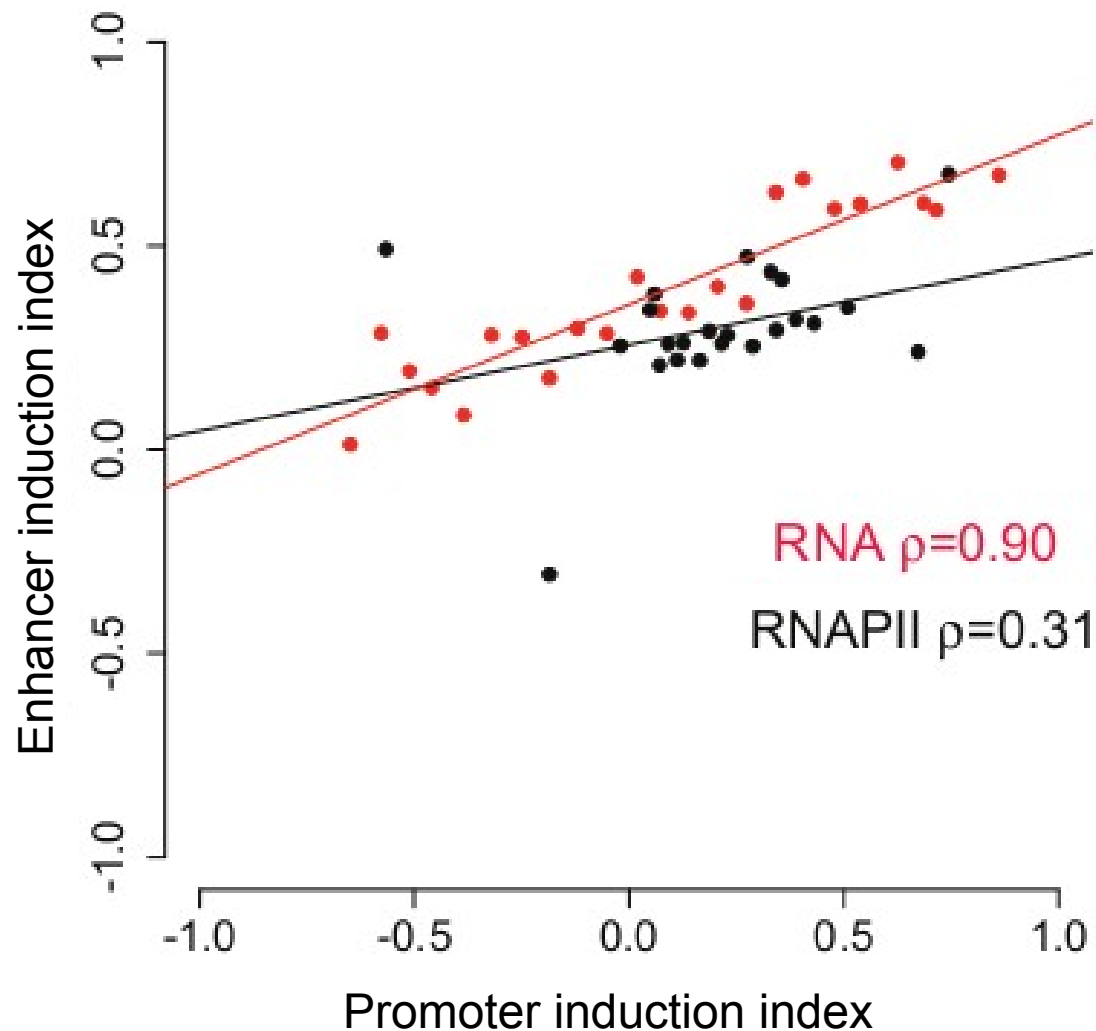
# Pair each enhancer with nearest promoter and compare RNAPII and RNA

$$\text{induction index} = \frac{KCl - un}{KCl + un}$$



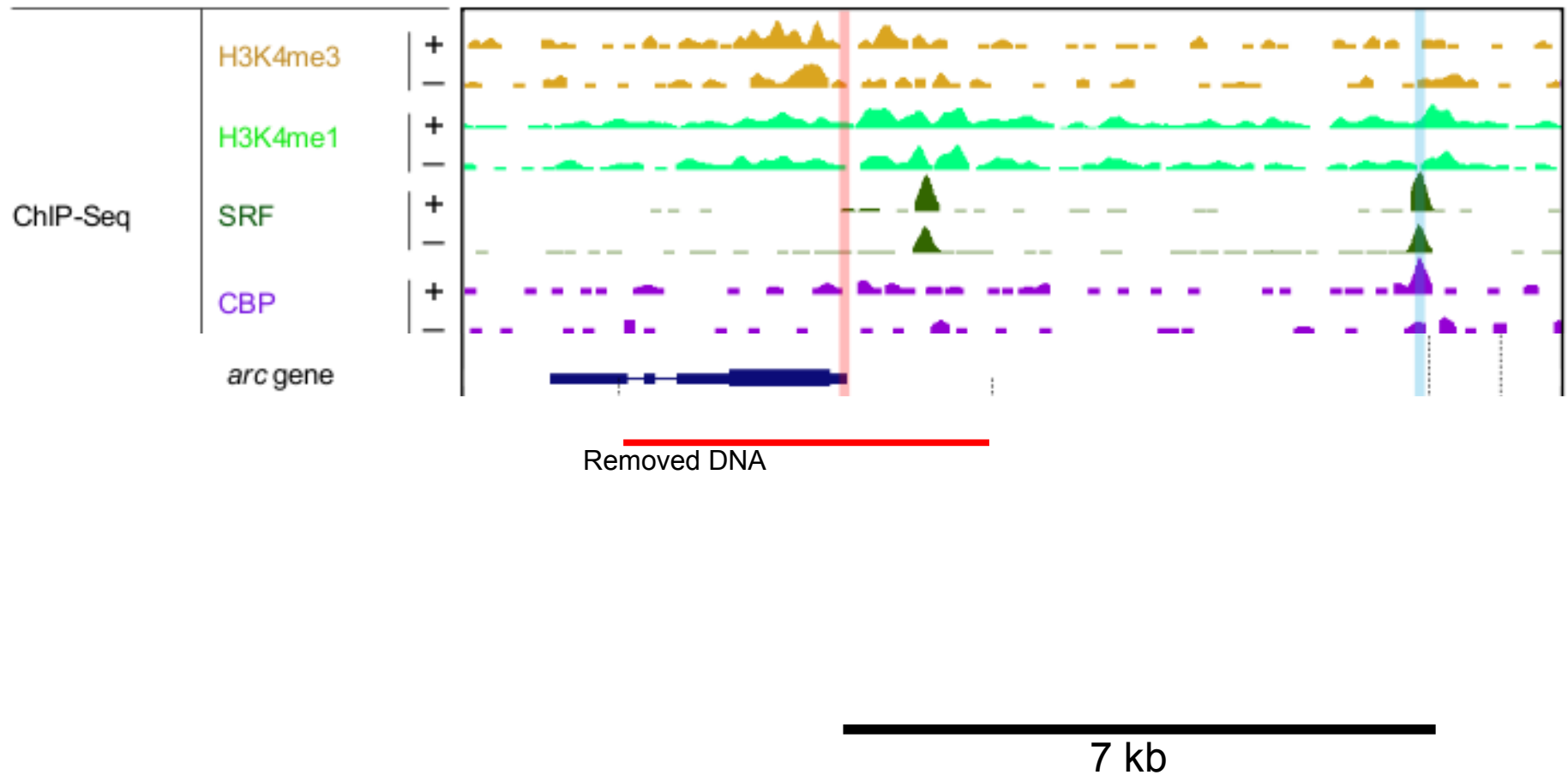
# eRNA induction is correlated with induction of nearby mRNAs but not RNAPII

$$\text{induction index} = \frac{KCl - un}{KCl + un}$$





Deletion of the Arc-promoter confirms that RNAPII recruitment is independent but eRNA transcription is not.

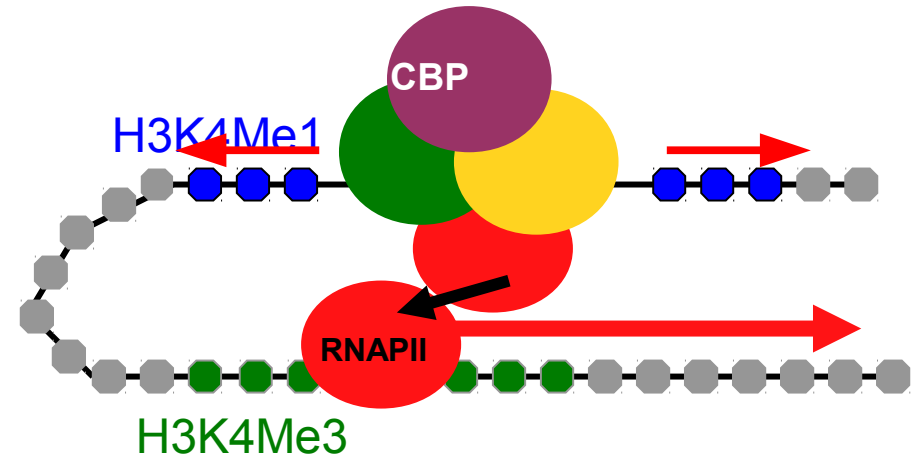


# Properties of activity-dependent enhancers

- Enriched for ~100 sequence motifs
- ChIP-seq reads predicted by sequence
- CBP binding determined by other TFs
- CBP recruits RNAPII
- RNAPII synthesizes eRNAs
  - eRNAs are rapidly degraded
  - eRNA induction correlated with mRNA induction
  - RNAPII recruitment independent of promoter

# What is the function of RNAPII at enhancers?

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter



Science is always wrong. It never solves a problem without creating ten more.

-George Bernard Shaw

# A simple model of RNAPII recruitment

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = \underbrace{k_p + Nk_e c}_{\text{Binding rate}} - \underbrace{\frac{P_M}{\tau}}_{\text{decay}}$$

$P$  – polymerase levels

$k_p$  – binding rate at promoter

$k_e$  – binding rate at enhancer

$N$  – number of enhancers

$c$  – contact probability

$\tau$  – RNAPII half life

# Steady state level of RNAPII is increased

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

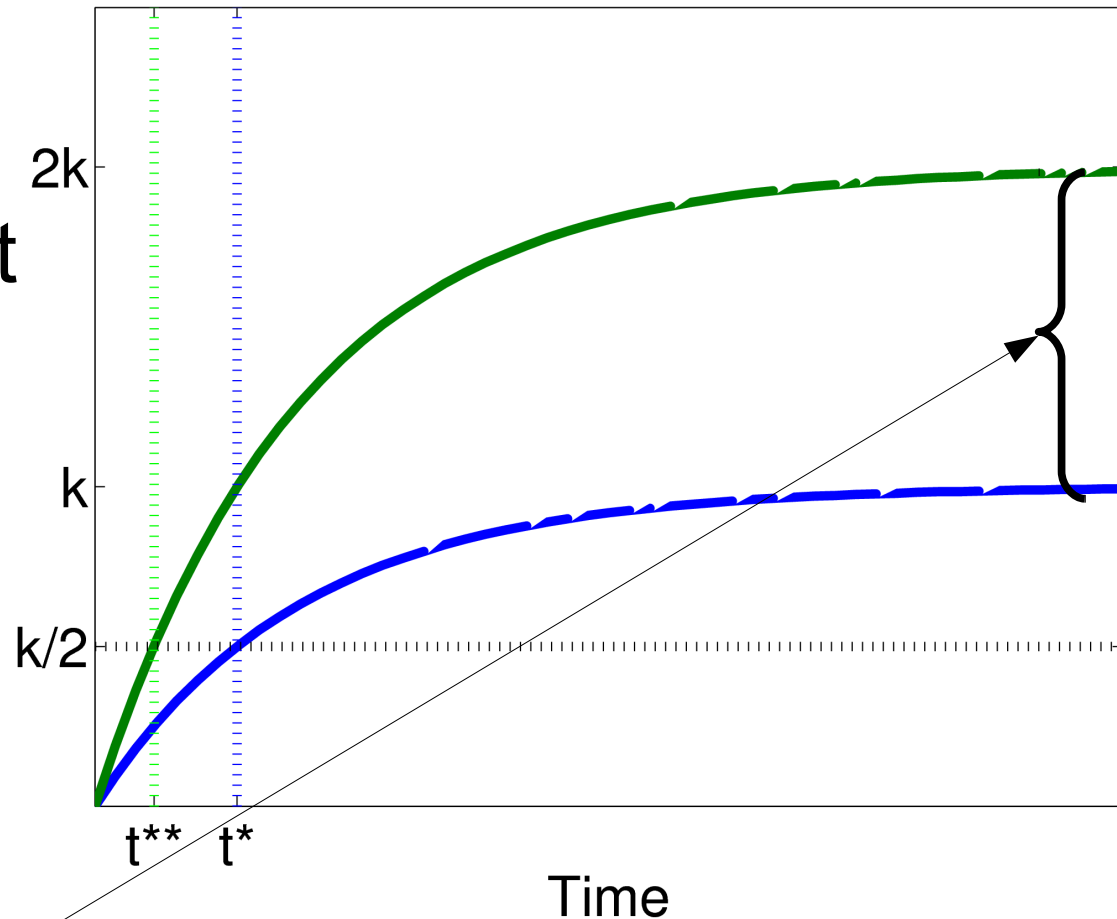
$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

$$P_M(t) = \underbrace{(k_p + Nk_e c)}_{\text{Steady state level}} (1 - e^{-t/\tau})$$

# Steady state level of RNAPII is increased

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

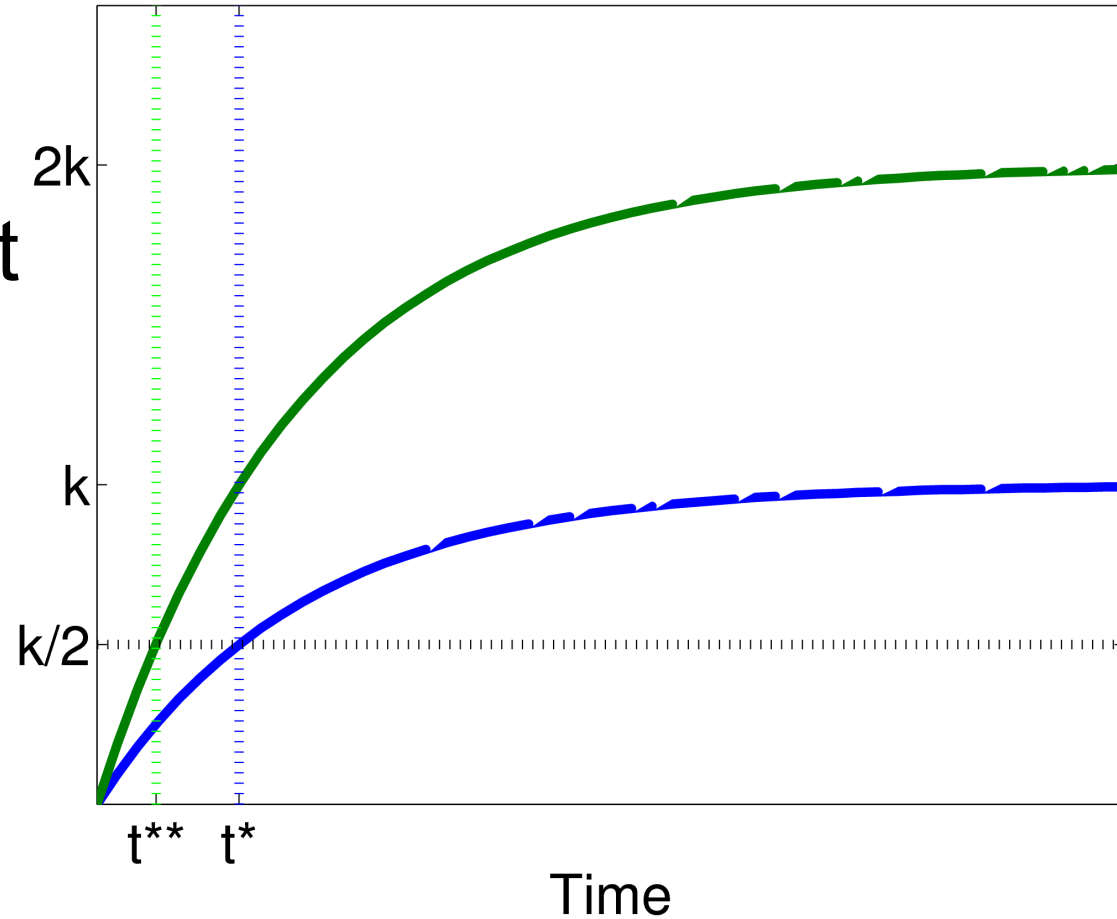


$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$

# Rise time is reduced

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

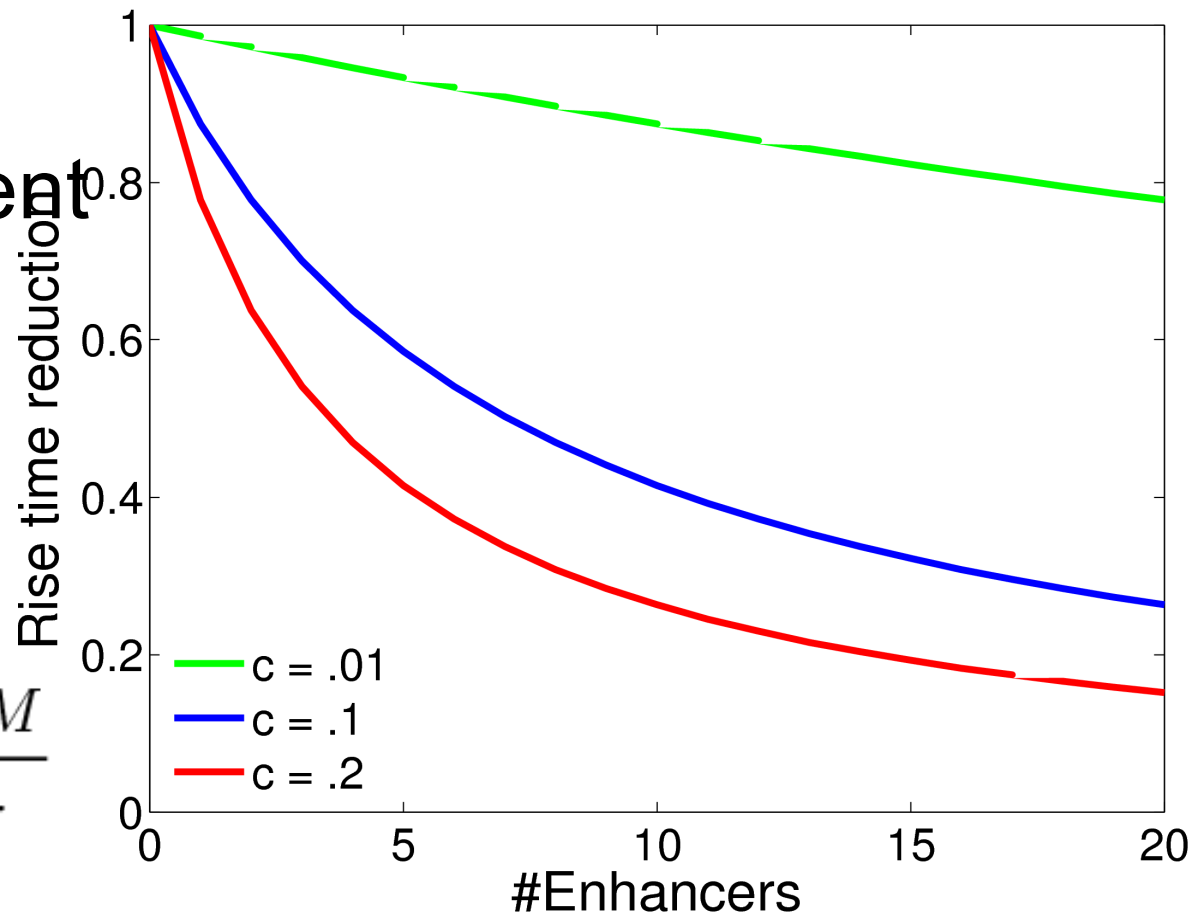


$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$

# Significant speed-up with ~5 enhancers

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$



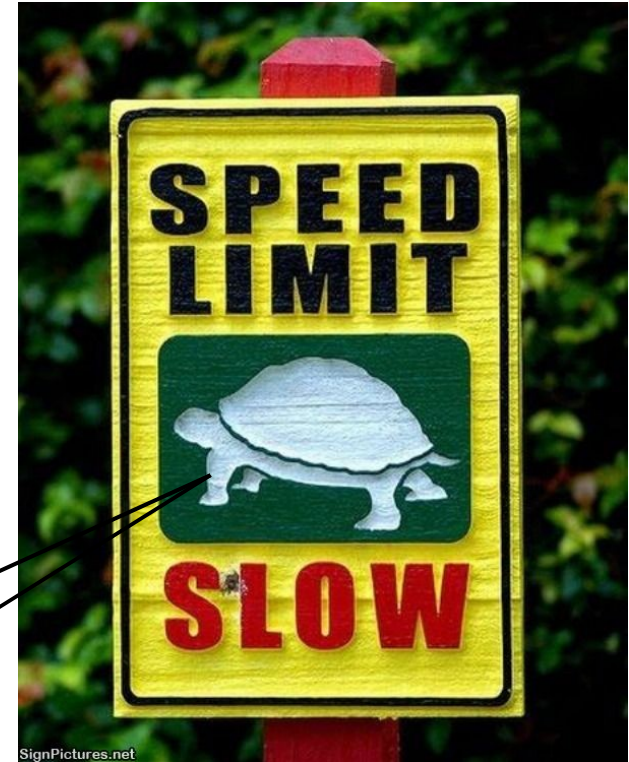
$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$



# Recruitment of RNAPII is diffusion limited

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$



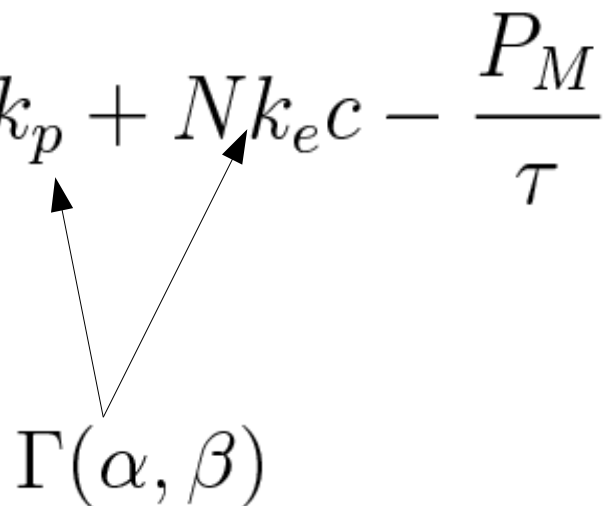
$$P_M(t) = (k_p + Nk_e c)(1 - e^{-t/\tau})$$

# Enhancers may reduce the noise in RNAPII

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$

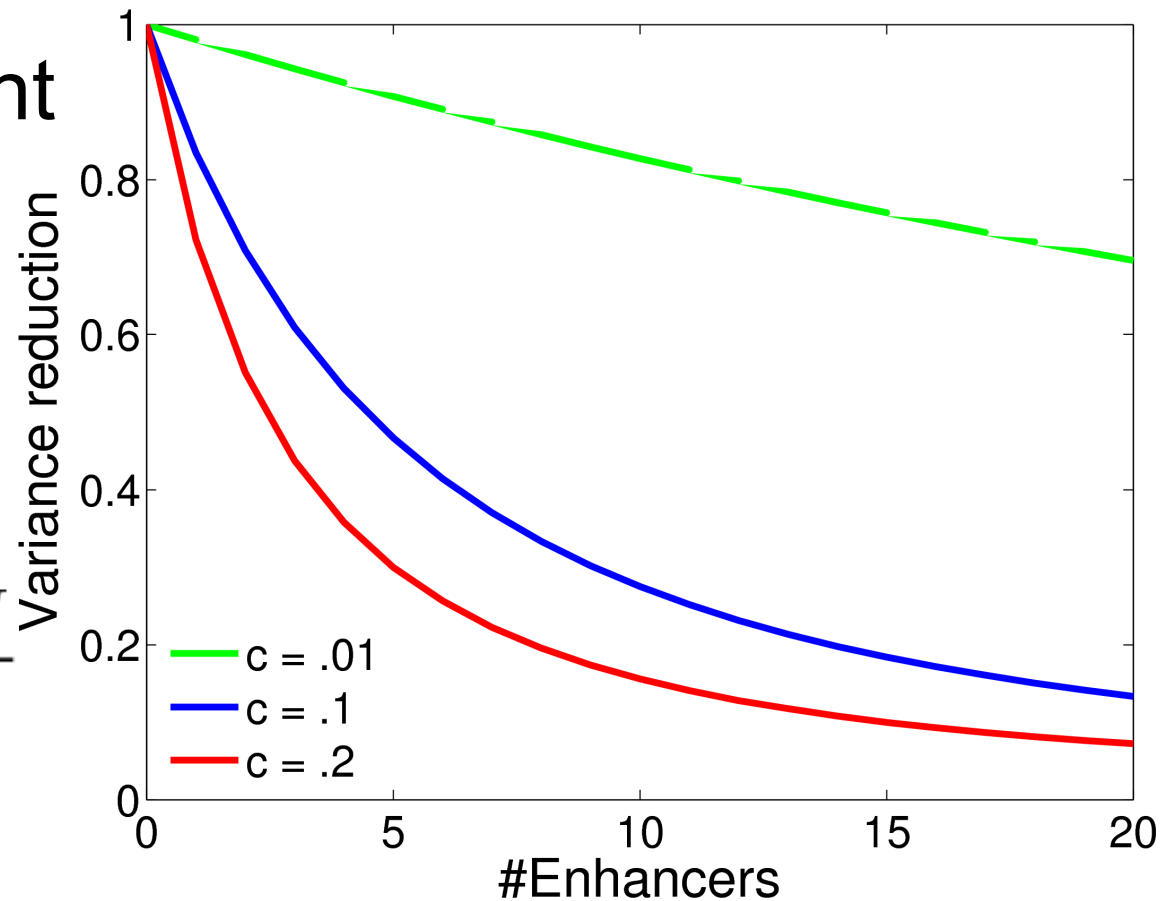
$\Gamma(\alpha, \beta)$

A diagram showing the function  $\Gamma(\alpha, \beta)$  at the bottom. Two arrows originate from this function: one points to the parameter  $k_p$  in the equation above, and the other points to the parameter  $Nk_e c$ .

# RNAPII noise reduction proportional to number of enhancers

- Transcribe eRNAs
- Speed up recruitment of RNAPII at promoter

$$\frac{dP_M}{dt} = k_p + Nk_e c - \frac{P_M}{\tau}$$



$$\frac{\text{Variance strong promoter}}{\text{Variance weak promoter with enhancers}} = \frac{\text{Var}[(1 + Nc)k]}{\text{Var}[k] + N\text{Var}[ck]} = \frac{(1 + Nc)^2 \text{Var}[k]}{(1 + Nc^2) \text{Var}[k]} \sim N$$

# What is the function of eRNAs?

- What is the function of RNAPII at enhancers?
  - Increase rate of RNAPII recruitment
    - Possibly faster than diffusion limit
  - Faster rise-time
  - Reduced noise
- What is the function of eRNAs?
  - Noise
  - Transcription establishes histone modifications
  - Transcript has function

# Establishing H3K4me1 levels at enhancers

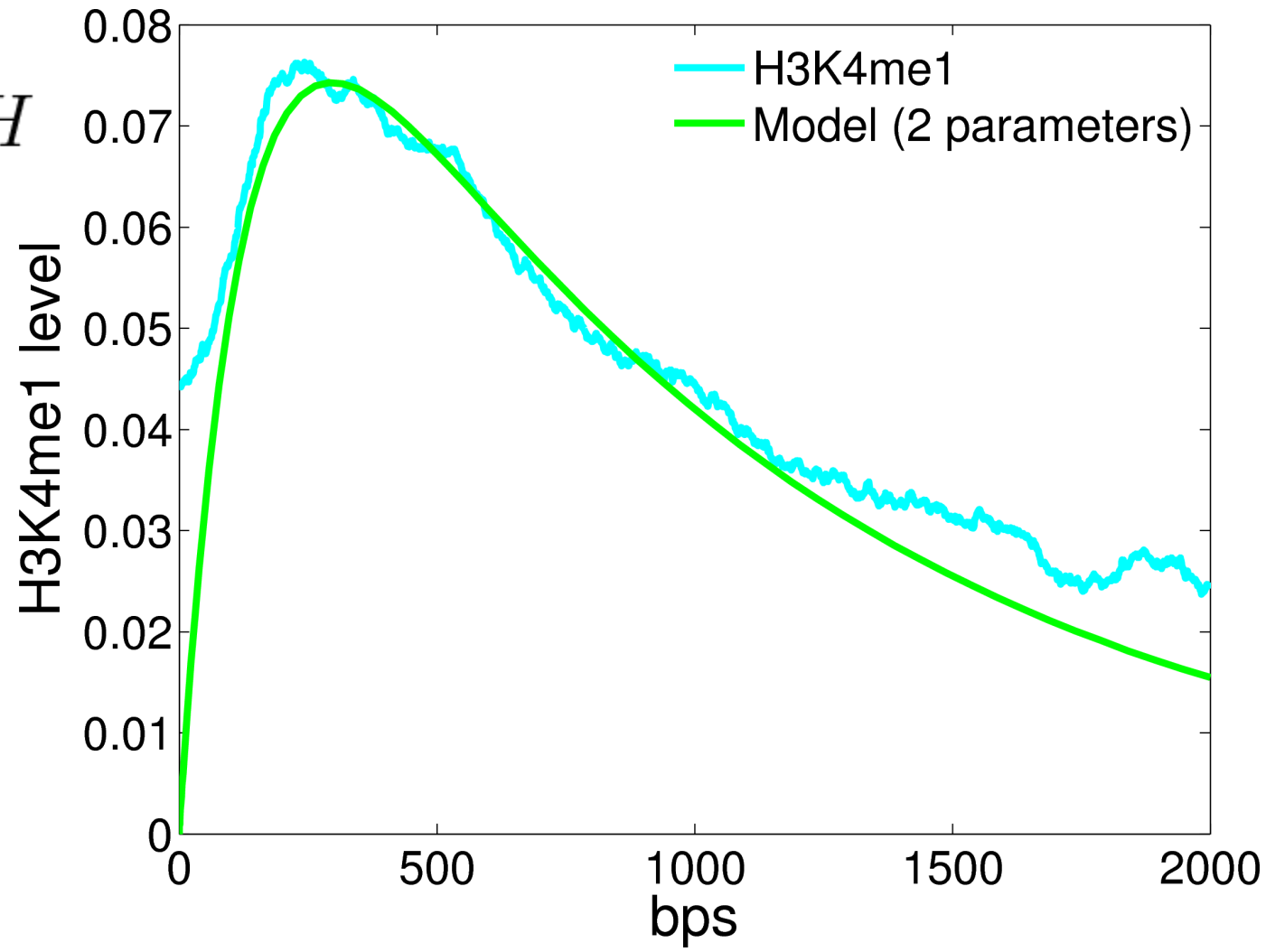
$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dH}{dx} = \kappa P(x) - \mu H$$

# Establishing H3K4me1 levels at enhancers

$$\frac{dP}{dx} = k - \lambda P$$

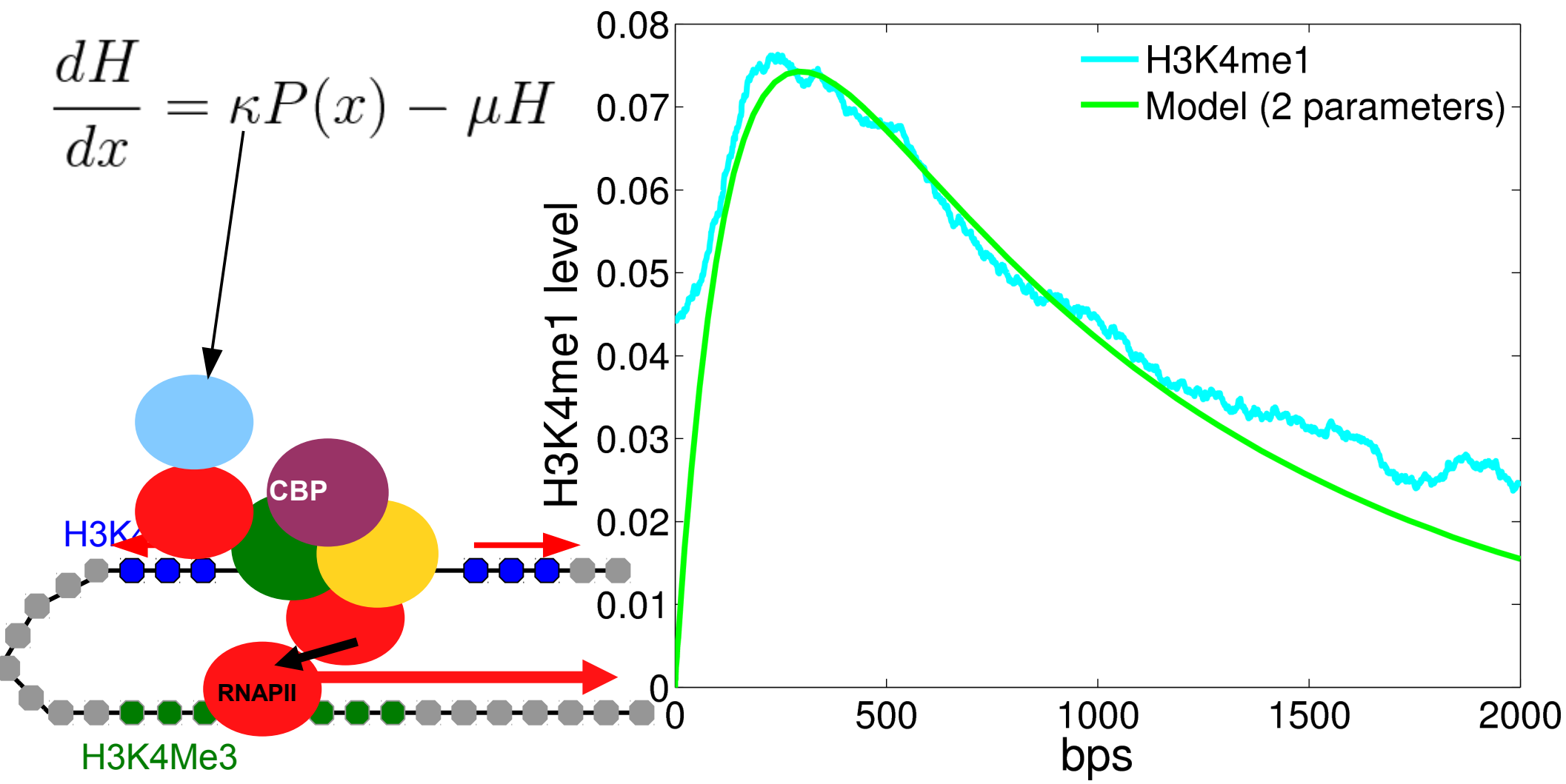
$$\frac{dH}{dx} = \kappa P(x) - \mu H$$



# Establishing H3K4me1 levels at enhancers

$$\frac{dP}{dx} = k - \lambda P$$

$$\frac{dH}{dx} = \kappa P(x) - \mu H$$



# eRNAs have been found in other cell types

doi:10.1038/nature09033

nature

ARTICLES

## Widespread transcription at neuronal activity-regulated enhancers

Tae-Kyung Kim<sup>1\*†</sup>, Martin Hemberg<sup>2\*</sup>, Jesse M. Gray<sup>1\*</sup>, Allen M. Costa<sup>1</sup>, Daniel M. Bear<sup>1</sup>, Jing Wu<sup>3</sup>, David A. Harmin<sup>1,4</sup>, Mike Laptewicz<sup>1</sup>, Kellie Barbara-Haley<sup>5</sup>, Scott Kuersten<sup>6</sup>, Eirene Markenscoff-Papadimitriou<sup>1†</sup>, Dietmar Kuhl<sup>7</sup>, Haruhiko Bito<sup>8</sup>, Paul F. Worley<sup>3</sup>, Gabriel Kreiman<sup>2</sup> & Michael E. Greenberg<sup>1</sup>

## Histone H3K27ac separates active from poised enhancers and predicts developmental state

Menno P. Creyghton<sup>a,1</sup>, Albert W. Cheng<sup>a,b,1</sup>, G. Grant Welstead<sup>a</sup>, Tristan Kooistra<sup>c,d</sup>, Bryce W. Carey<sup>a,e</sup>, Eveline J. Steine<sup>a,e</sup>, Jacob Hanna<sup>a</sup>, Michael A. Lodato<sup>a,c</sup>, Garrett M. Frampton<sup>a,e</sup>, Phillip A. Sharp<sup>d,e</sup>, Laurie A. Boyer<sup>e</sup>, Richard A. Young<sup>a,e</sup>, and Rudolf Jaenisch<sup>a,e,2</sup>

OPEN ACCESS Freely available online

PLoS BIOLOGY

## A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers

Francesca De Santa<sup>1,3</sup>, Iros Barozzi<sup>1,3</sup>, Flore Mietton<sup>1,3</sup>, Serena Ghisletti<sup>1</sup>, Sara Polletti<sup>1</sup>, Betsabeh Khoramian Tusi<sup>1</sup>, Heiko Muller<sup>1</sup>, Jiannis Ragoussis<sup>2</sup>, Chia-Lin Wei<sup>3</sup>, Gioacchino Natoli<sup>1\*</sup>

## LETTER

doi:10.1038/nature09692

## A unique chromatin signature uncovers early developmental enhancers in humans

Alvaro Rada-Iglesias<sup>1</sup>, Ruchi Bajpai<sup>1</sup>, Tomek Swigut<sup>1</sup>, Samantha A. Brugmann<sup>1</sup>, Ryan A. Flynn<sup>1</sup> & Joanna Wysocka<sup>1,2</sup>

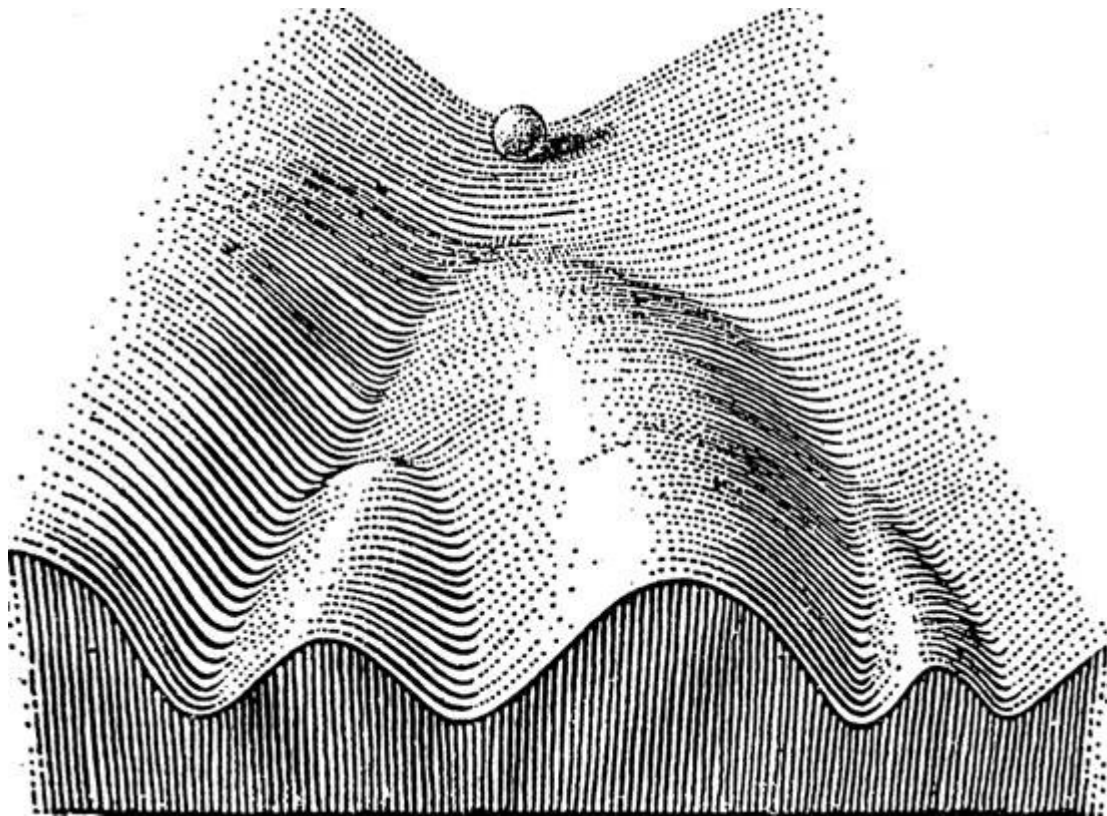


# Summary

- Identified ~12k activity-dependent enhancers
- Discovered and quantified novel mechanisms
  - Identified enriched motifs and bound TFs
  - Combinatorial code for CBP affinity
  - Recruitment of RNAPII at enhancers
  - Transcription at enhancers
    - Properties of eRNA
    - Model of RNAPII and eRNA levels
    - Interaction with promoter necessary
  - Function of RNAPII and eRNA

# Stochastic models of gene expression

- Transitions between stable states
- Noise



Waddington, 1953

## Master Equation (**ME**) description

$$\frac{dP_j}{dt} = \sum_i W_{ij} P_i(t) - W_{ji} P_j(t)$$

$P_j$  - **Probability** of having  $j$  molecules

$W_{ij}$  - **Transition rate** from  $i$  to  $j$

MCMC required for solving ME

$$\frac{dP_j}{dt} = \sum_i W_{ij} P_i(t) - W_{ji} P_j(t)$$

$P_j$  - **Probability** of having j molecules

$W_{ij}$  - **Transition rate** from i to j

- Use Markov Chain Monte Carlo (MCMC)
  - Gillespie's Stochastic Simulation Algorithm (**SSA**)

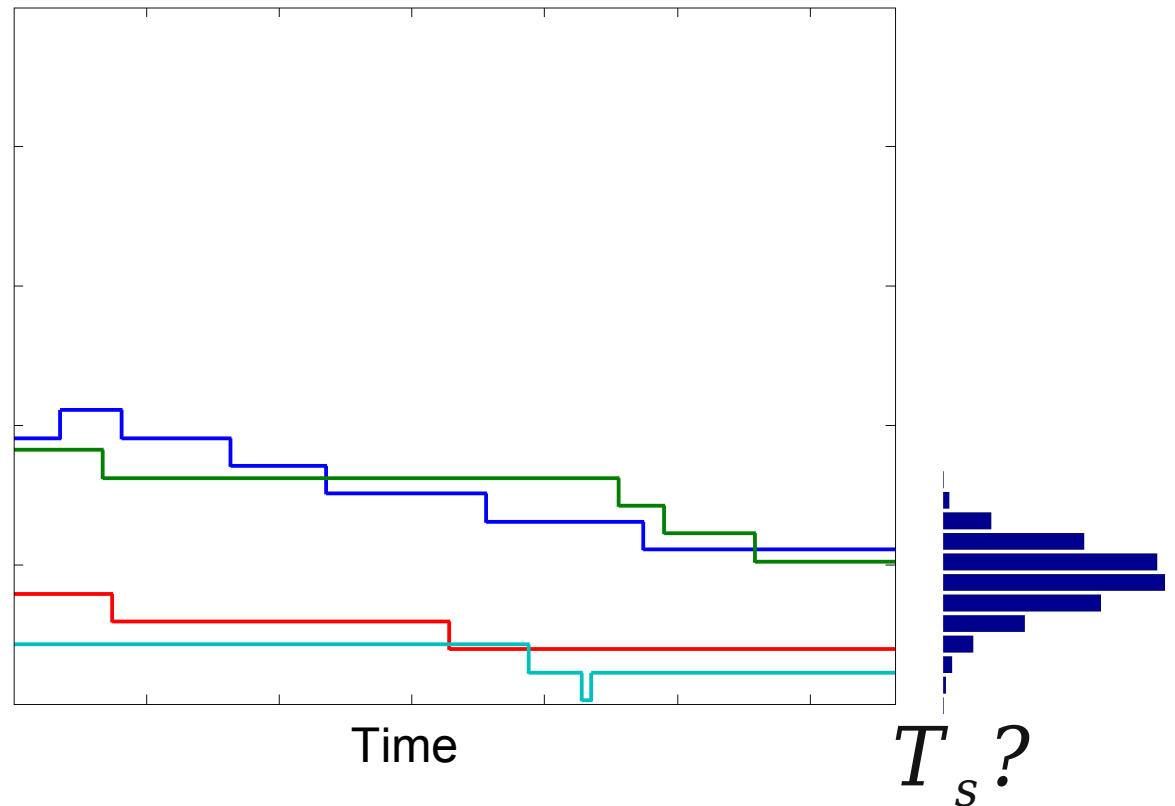
# How long do we need to run MCMC?

$$\frac{dP_j}{dt} = \sum_i W_{ij} P_i(t) - W_{ji} P_j(t)$$

$P_j$  - **Probability** of having  $j$  molecules

$W_{ij}$  - **Transition rate** from  $i$  to  $j$

#molecules



# Perfect sampling guarantees stationarity

$$\frac{dP_j}{dt} = \sum_i W_{ij} P_i(t) - W_{ji} P_j(t)$$

$P_j$  - **Probability** of having  $j$  molecules

$W_{ij}$  - **Transition rate** from  $i$  to  $j$

- Dominated Coupling From The Past SSA  
**proven** to reach stationary distribution

**BMC Systems Biology**



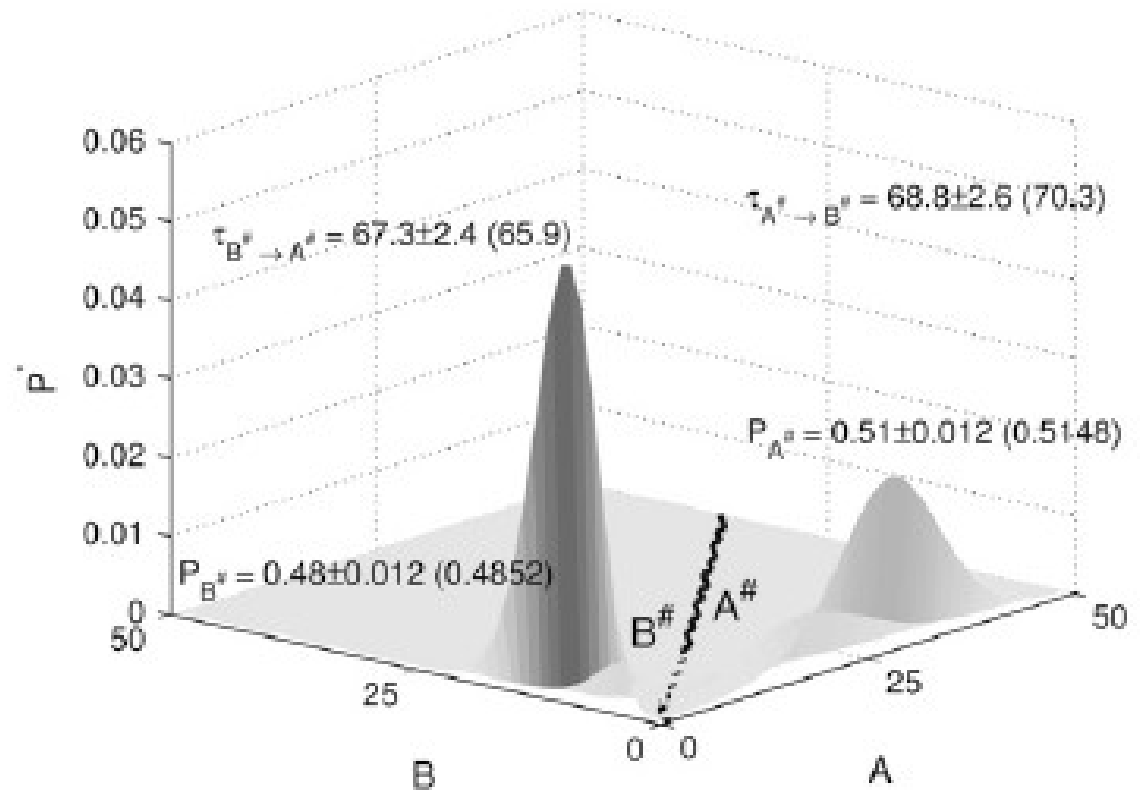
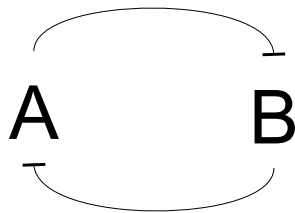
Methodology article

**Open Access**

**A Dominated Coupling From The Past algorithm for the stochastic simulation of networks of biochemical reactions**

Martin Hemberg<sup>1</sup> and Mauricio Barahona<sup>\*1,2</sup>

# Perfect sampling of transitions between steady states



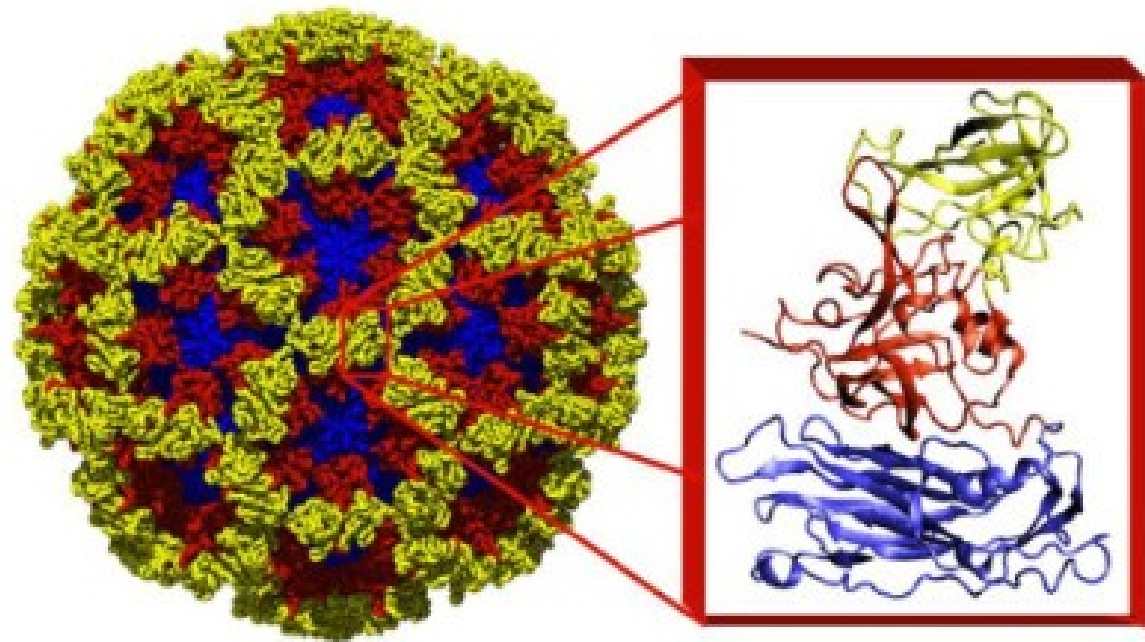
## Perfect Sampling of the Master Equation for Gene Regulatory Networks

Martin Hemberg and Mauricio Barahona

Department of Bioengineering and Institute for Mathematical Sciences, Imperial College London, London, United Kingdom

# Assembly of viral capsids

- Protect viral genome
  - Self-assembly
  - Identical subunits
  - Icosahedral symmetry



Biophysical Journal Volume 90 May 2006 3029–304:

## Stochastic Kinetics of Viral Capsid Assembly Based on Detailed Protein Structures

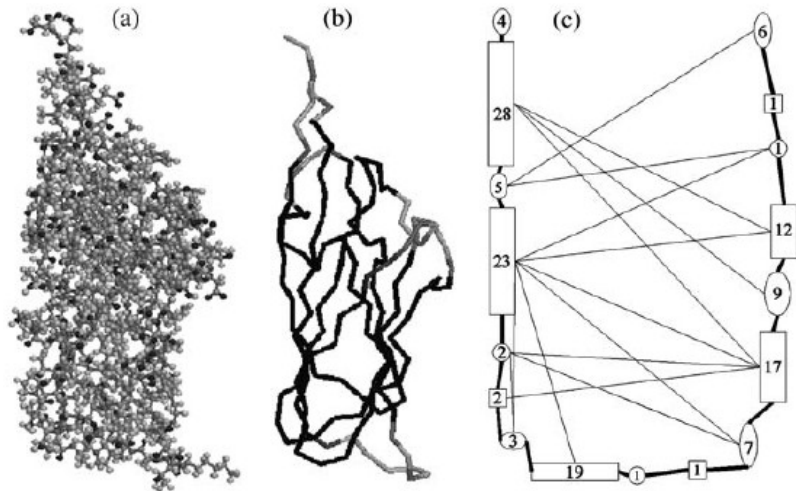
Martin Hemberg,\* Sophia N. Yaliraki,<sup>†</sup> and Mauricio Barahona\*

\*Department of Bioengineering and <sup>†</sup>Department of Chemistry, Imperial College London, London, United Kingdom



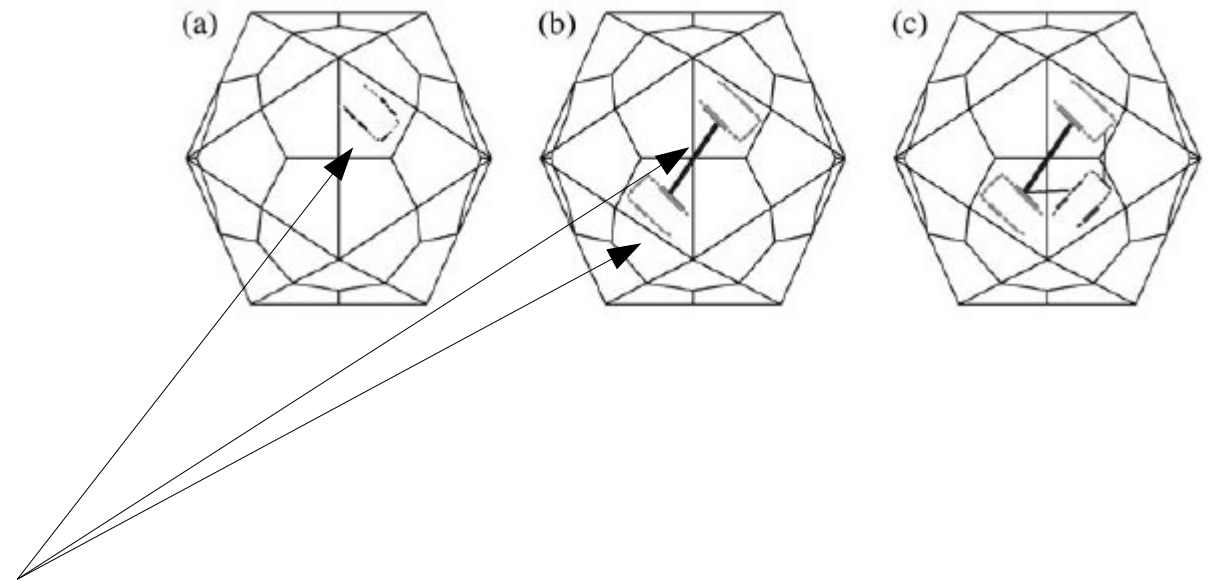
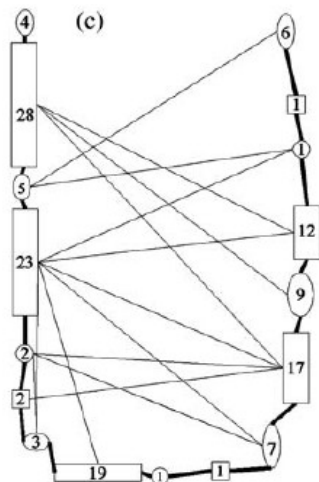
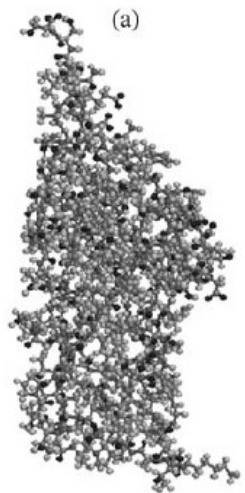
# Coarse-grained protein model

- Atomic-structure
- FIRST calculates rigidity of amino acids
- Identify ~20 rigid blocks



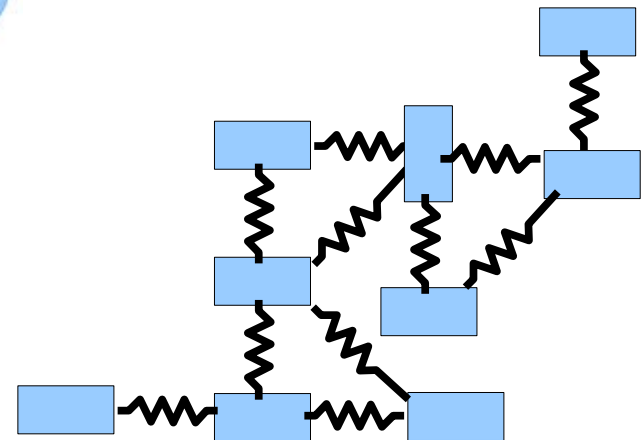
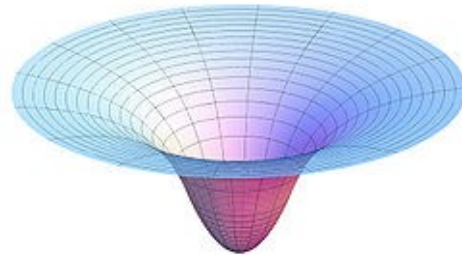
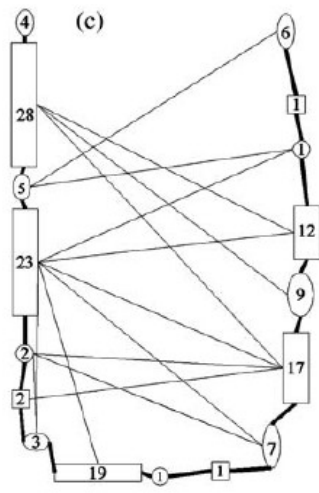
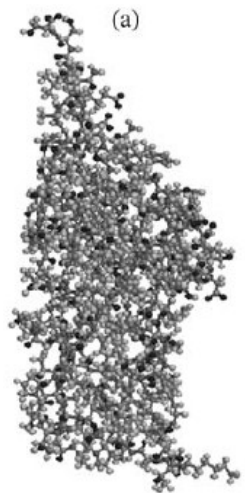
# Use reduced representation for aggregates

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
  - Association restricted by diffusion



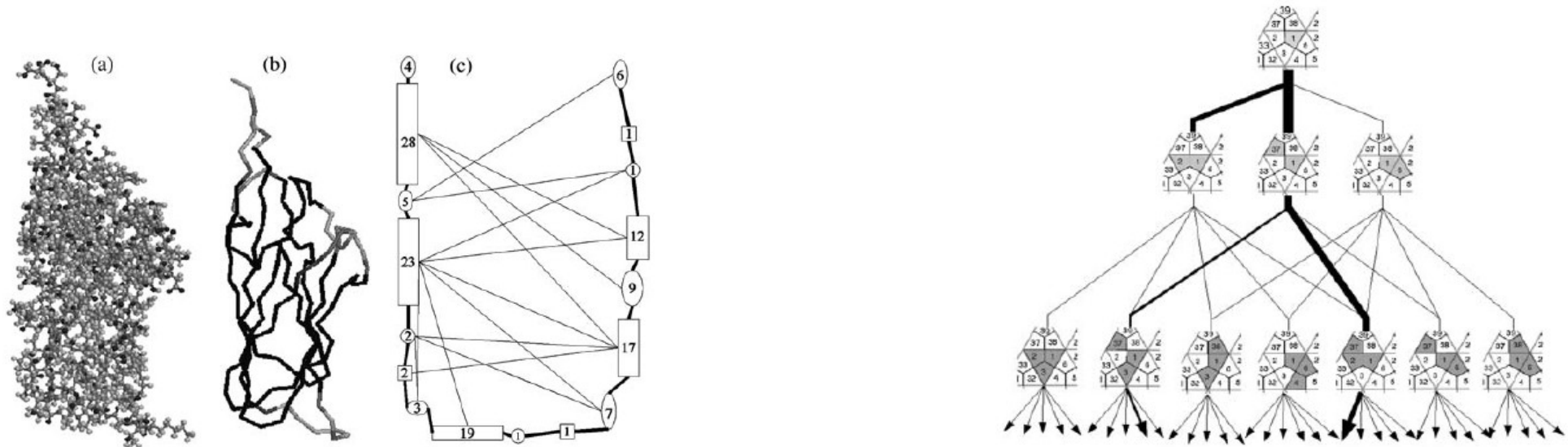
# Aggregates modeled as mass-spring graph

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
  - Association restricted by diffusion
  - Dissociation escape from multi-dimensional well



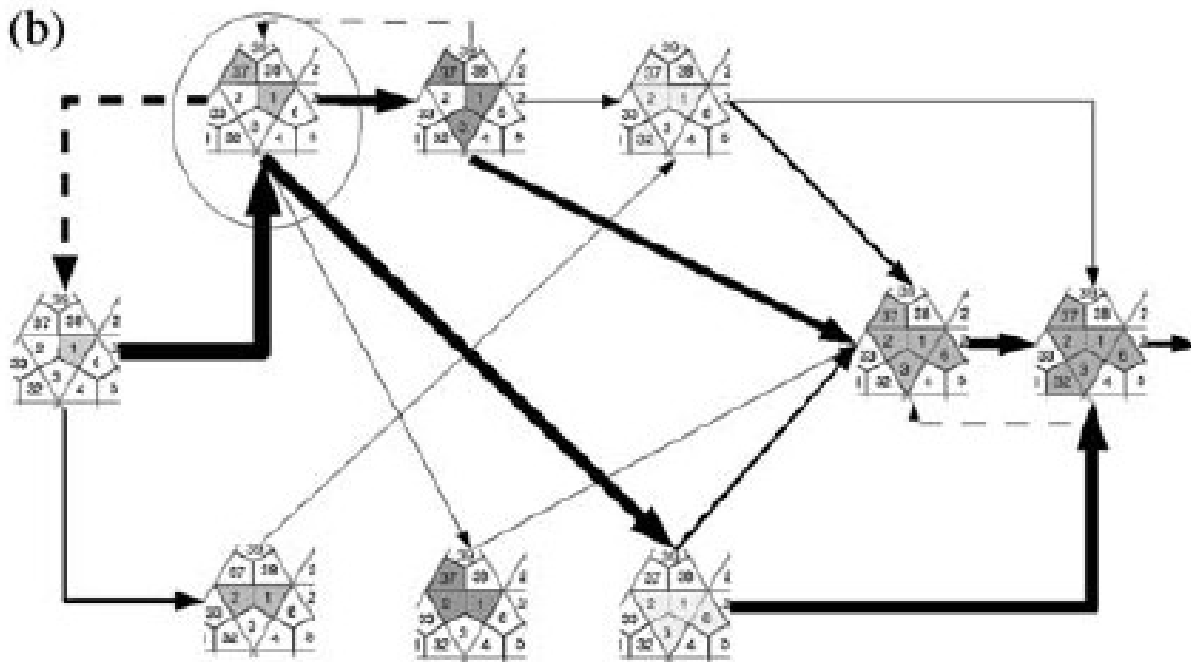
# All reactions cannot be enumerated

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
  - Association restricted by diffusion
  - Dissociation escape from multi-dimensional well



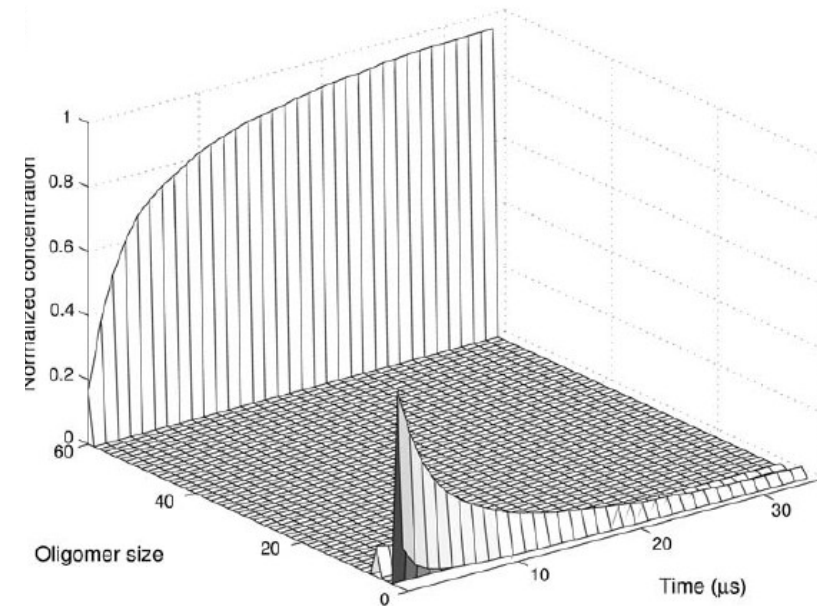
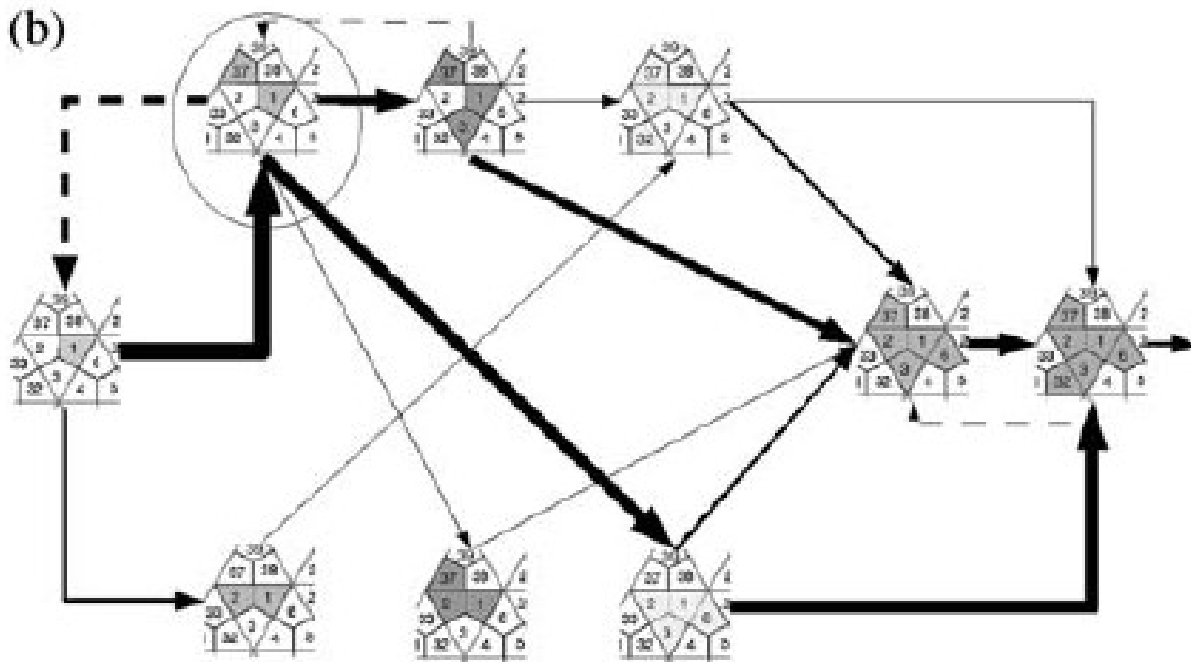
# Probabilistic sampling of assembly paths

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
  - Association restricted by diffusion
  - Dissociation escape from multi-dimensional well



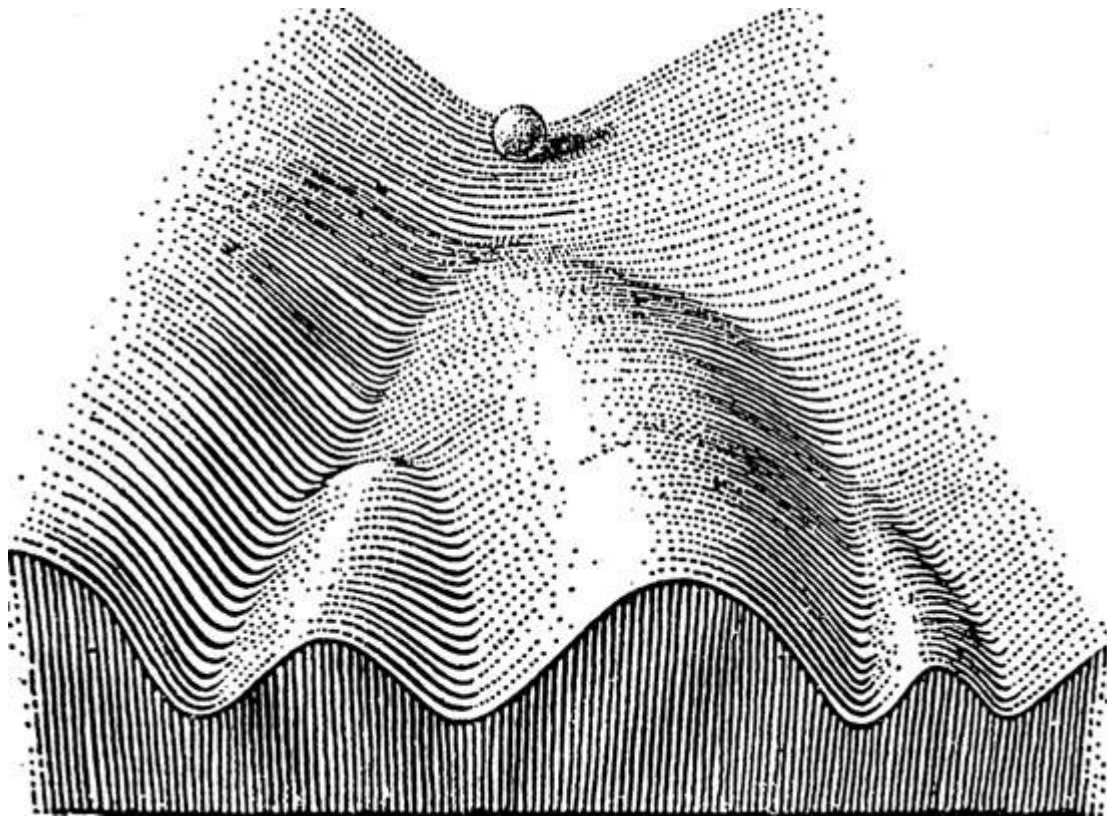
# Identify stable intermediaries

- Atomic-structure, identify rigid blocks
- Oligomer association and dissociation rates
  - Association restricted by diffusion
  - Dissociation escape from multi-dimensional well



# Future Work: Organizing principles of the genome

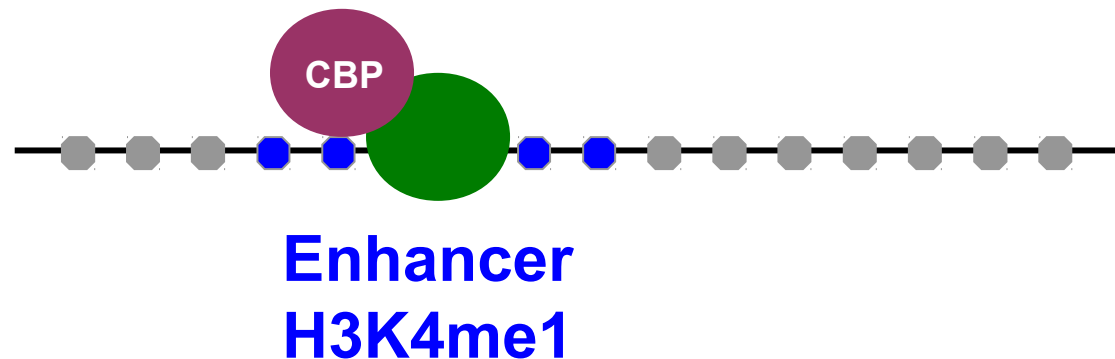
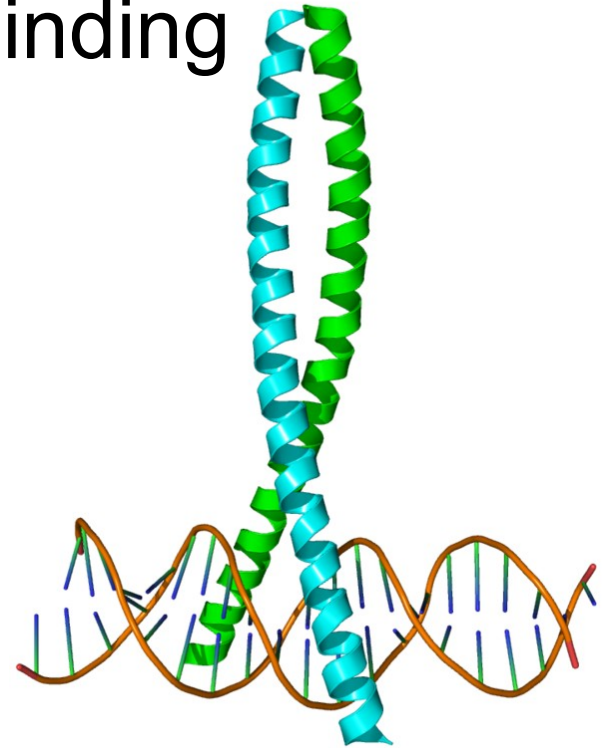
- Use genome-wide data to develop systems biology and biophysical models of gene regulation and gene expression



Waddington, 1953

# Develop biophysical models of TF binding

- Use ChIP-Seq for biophysical models of TF binding
- Relate looping to biophysical properties of DNA





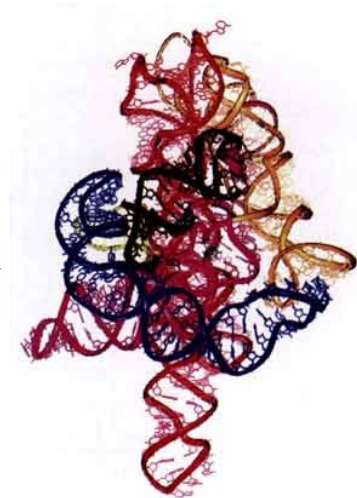
# Model stochastic gene expression for entire transcriptome

- Analytical models of gene expression noise
  - Parametric robustness
- Apply to genome-wide single-cell RNA-Seq

# Determine structure of RNAs

- Other species of novel non-coding RNAs
  - Identify structural motifs
- High-throughput sequencing of structure
  - PARS
  - SHAPE-Seq

.....ACGUCCAAAUUCCCUAGGCUCAAGGCAUUCGAUCGGGAUUAUA..... →



# Acknowledgements

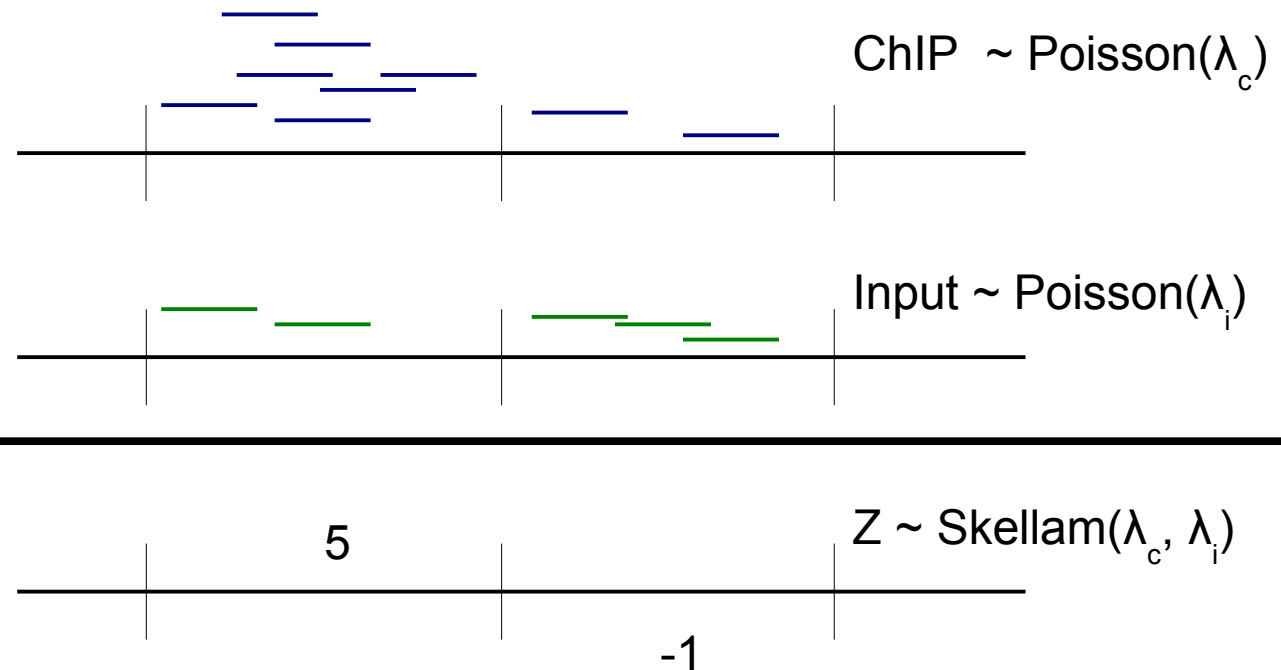
- Gabriel Kreiman
- Jesse Gray
- Tae-Kyung Kim
- Michael Greenberg
- Mauricio Barahona

**Thank You**

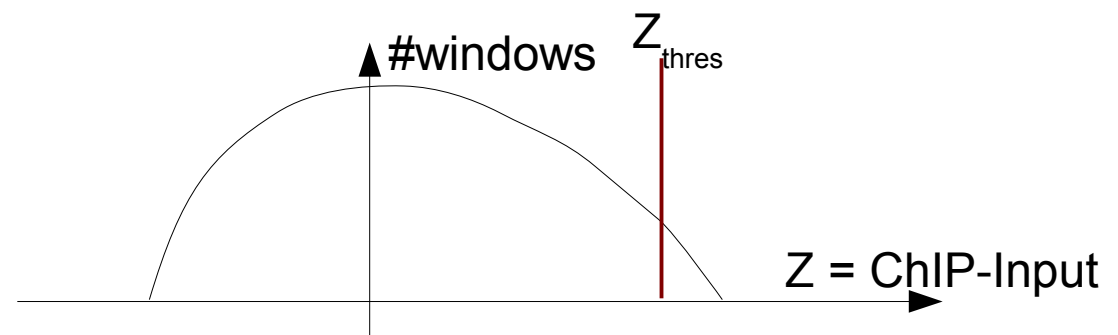


?

# Identifying regions with larger than expected number of ChIP-Seq reads



- False Detection Rate (FDR) determine threshold



# Use False Detection Ratio (FDR) to correct for multiple hypotheses

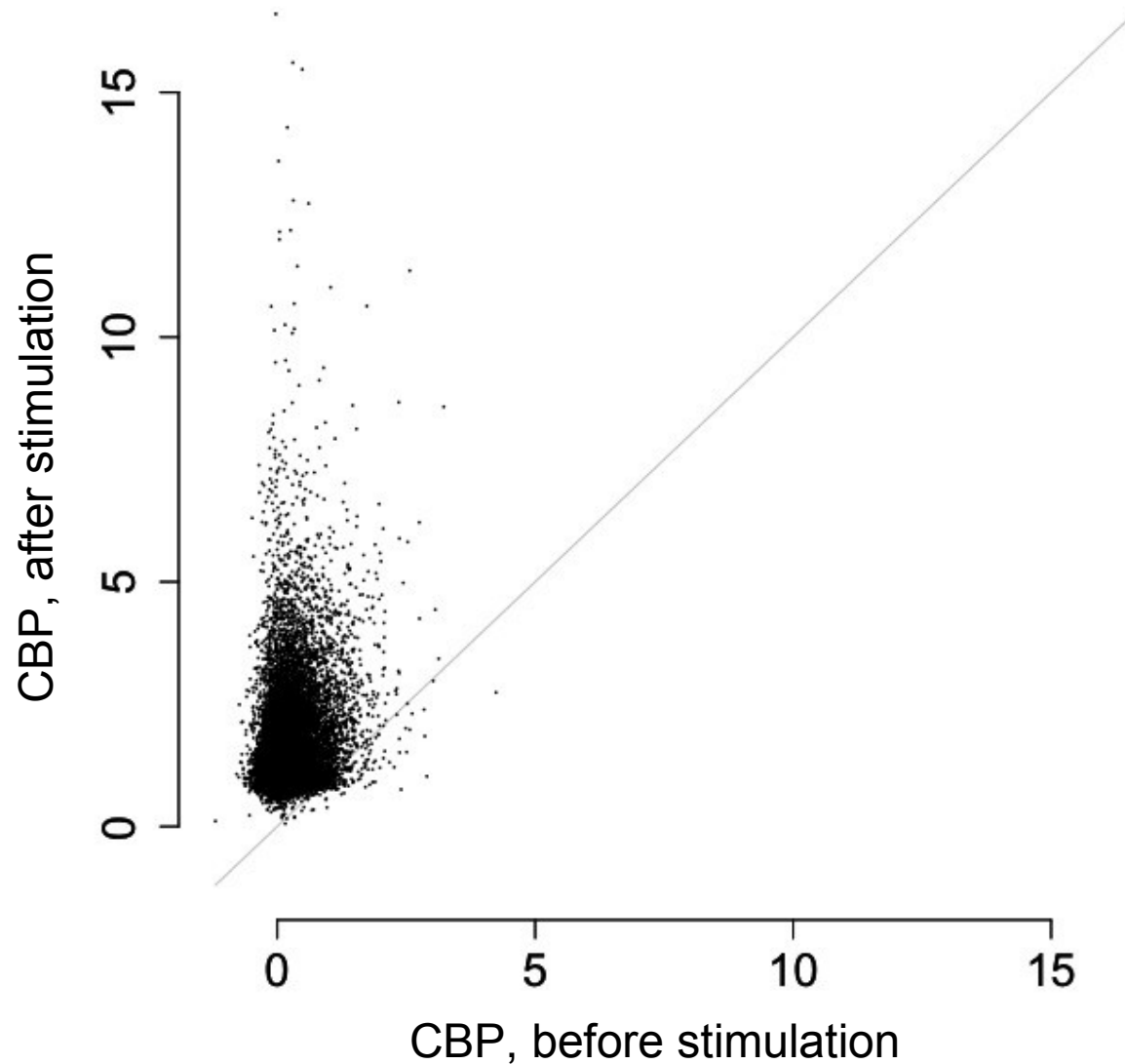
- $Z_i = \text{\#ChIP reads} - \text{\#input reads in window } i$
- $\sim 1 \text{ read}/100 \text{ bp}$ 
  - Assume  $\text{\#reads in window } P(k) = \lambda^k \exp(-\lambda)/k!$ 
    - Difference between two Poisson random variables
    - $Z_i \sim \text{Skellam}(z, \lambda_1, \lambda_2)$

$$p(x) = e^{-(\lambda_1 + \lambda_2)} (\lambda_1 / \lambda_2)^{x/2} I_x(2 \sqrt{\lambda_1 \lambda_2})$$

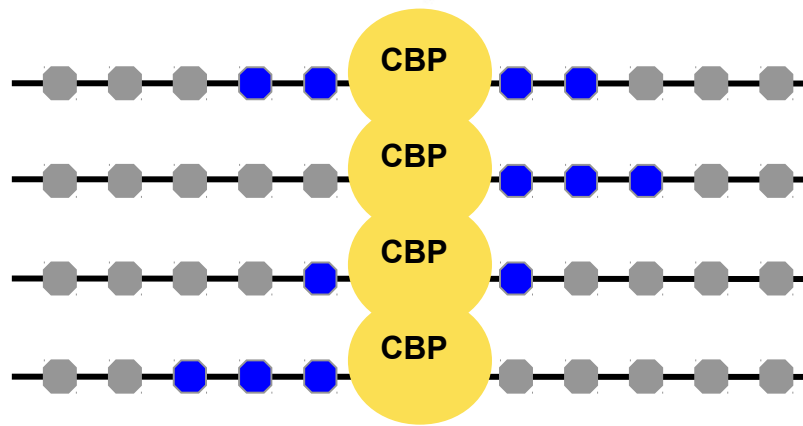
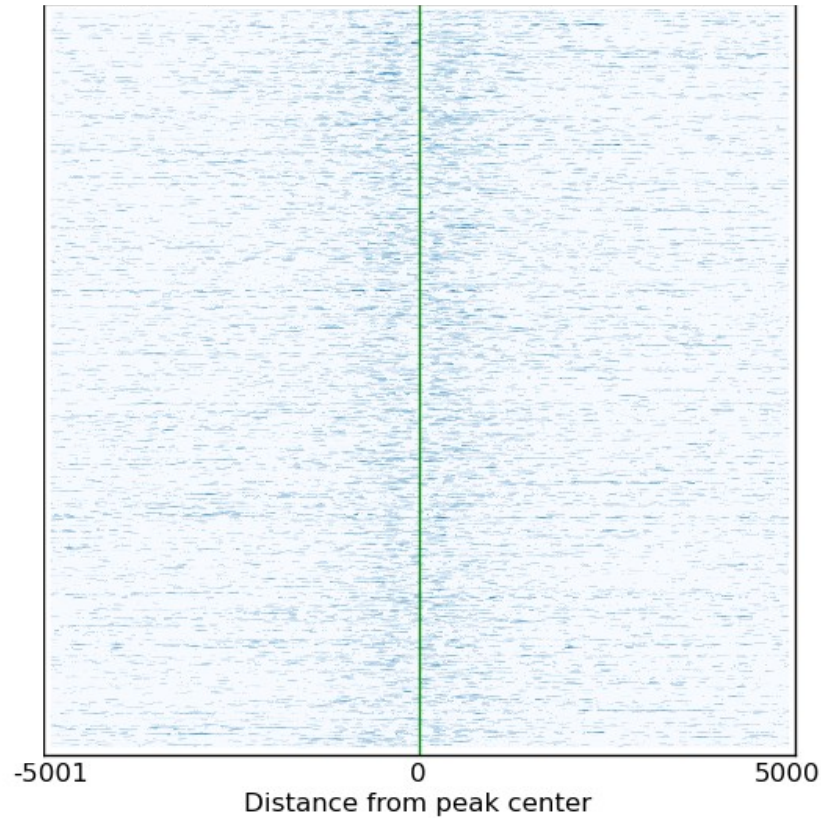
- Millions of windows need to be tested
  - FDR - expected fraction of false positives



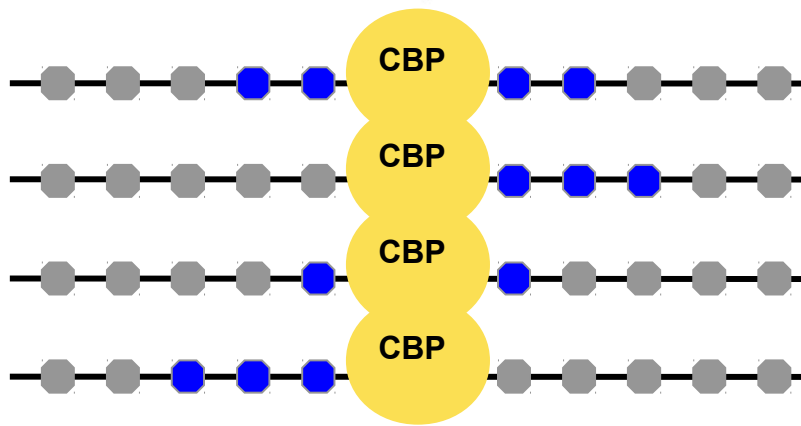
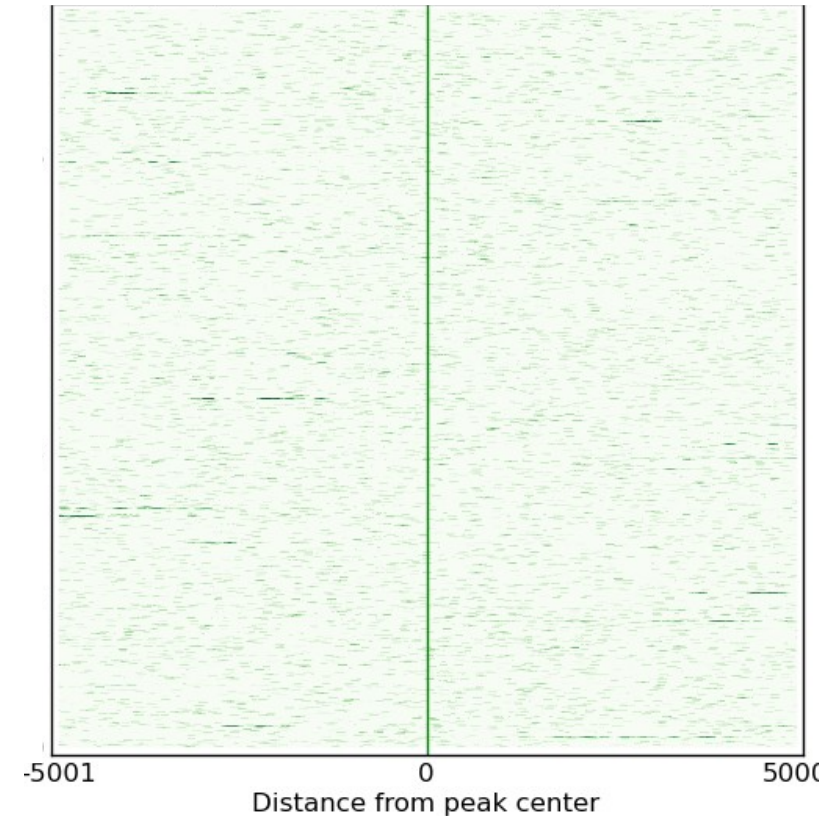
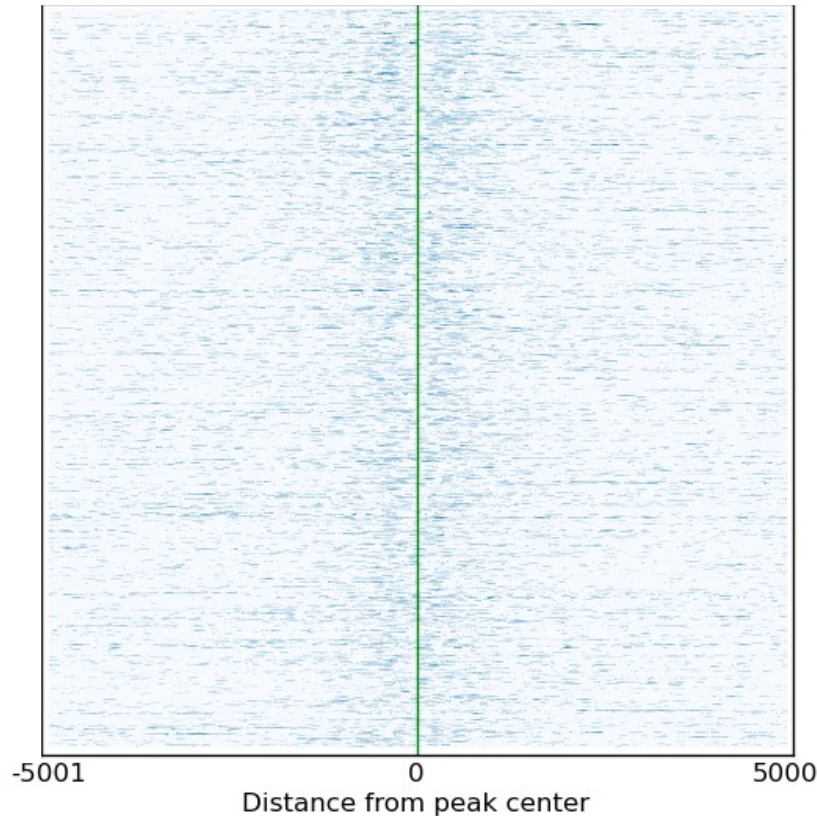
CBP binds in an activity regulated manner to  
~28,000 sites throughout the genome



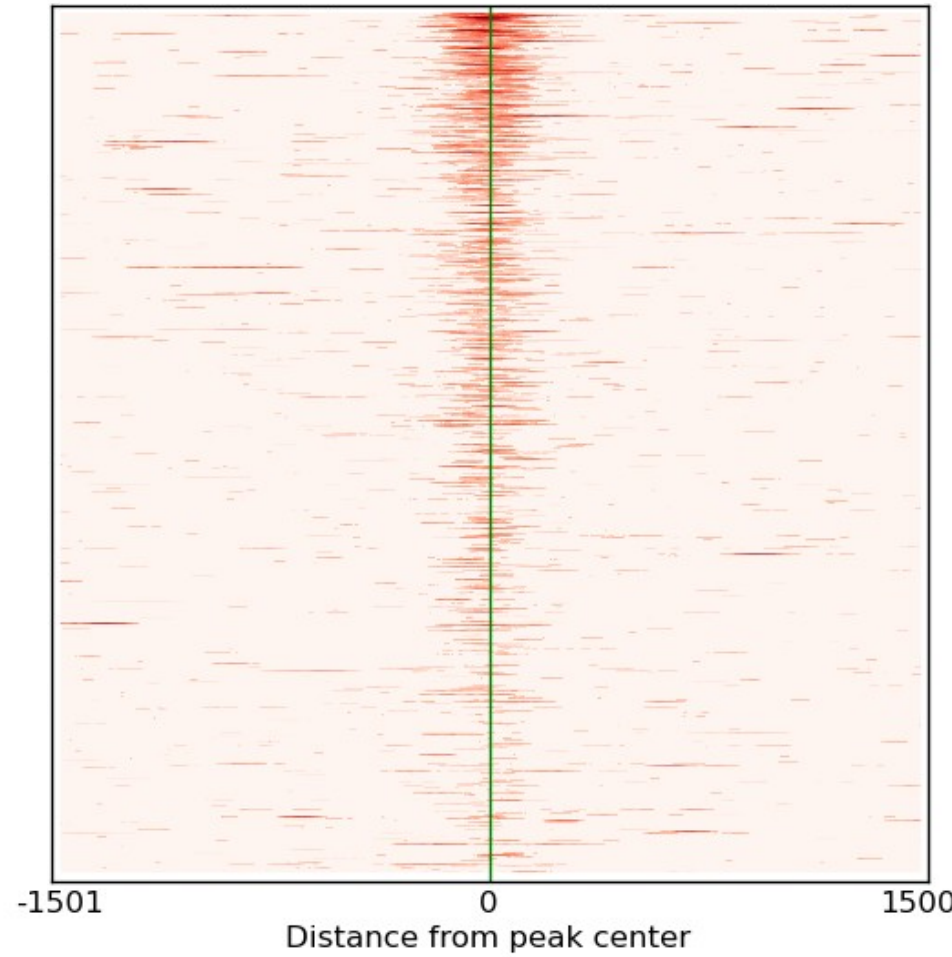
# Aligning CBP peaks to calculate H3K4me1 binding profiles



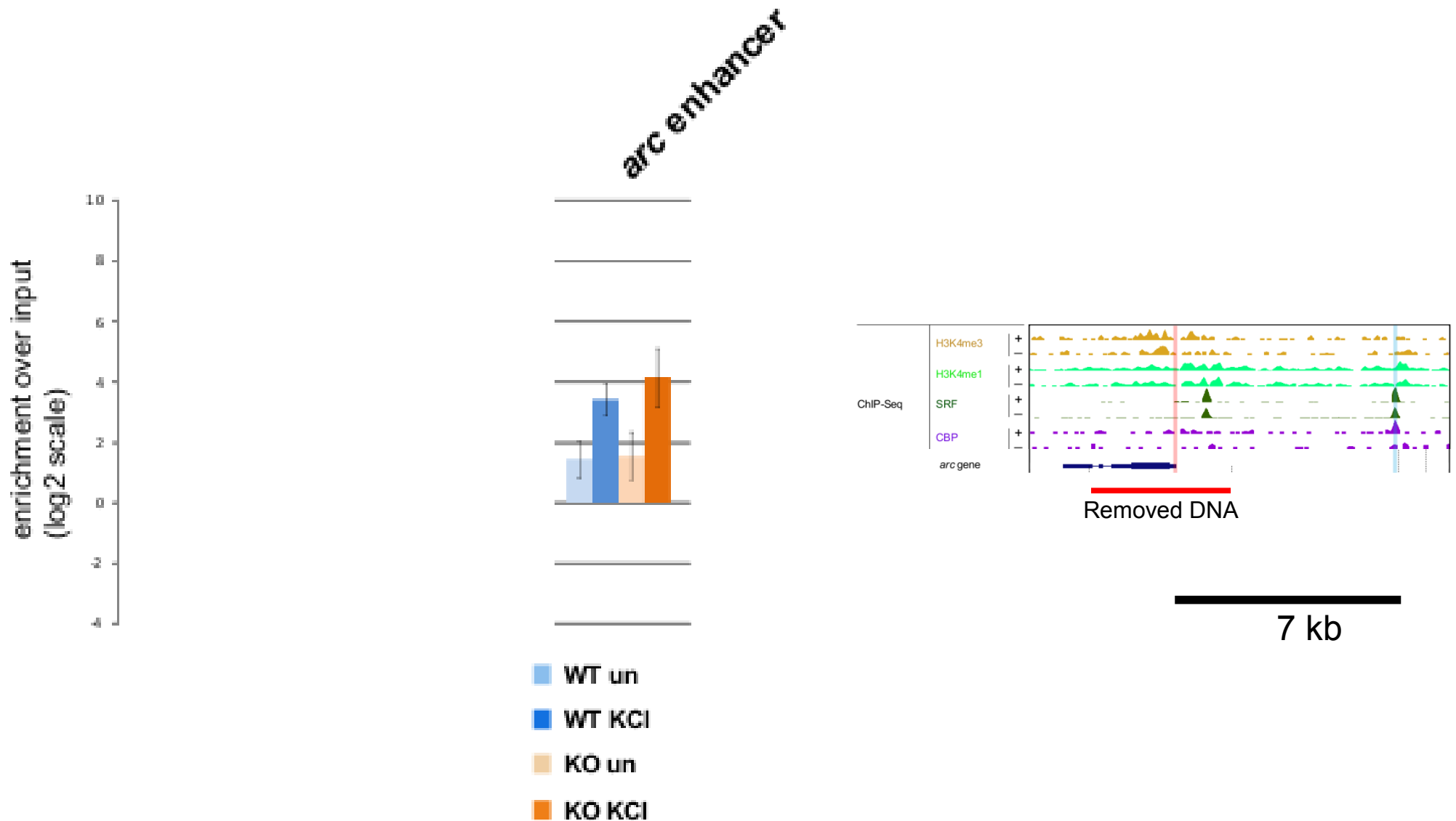
# Aligning CBP peaks to calculate H3K4me1 and H3K4me3 binding profiles



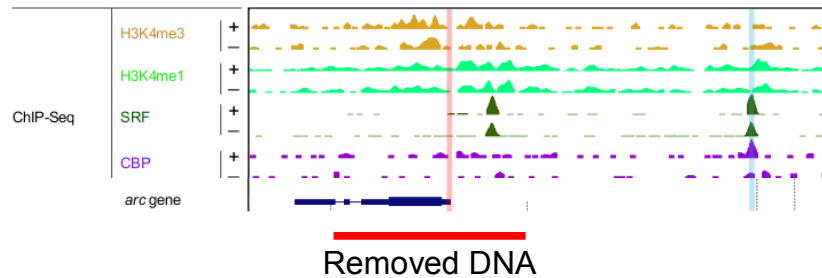
# RNAPII binds at activity-dependent enhancers



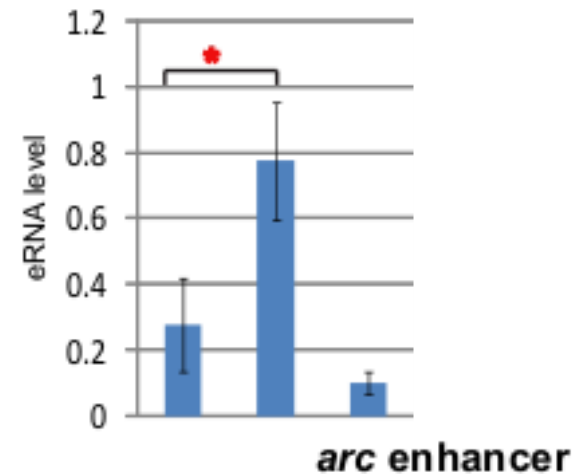
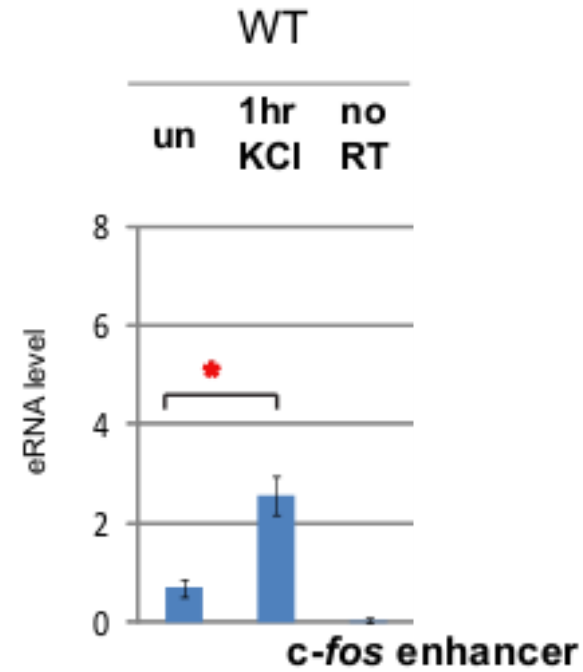
# RNAPII levels are unchanged at the enhancer in the mutant before and after KCl



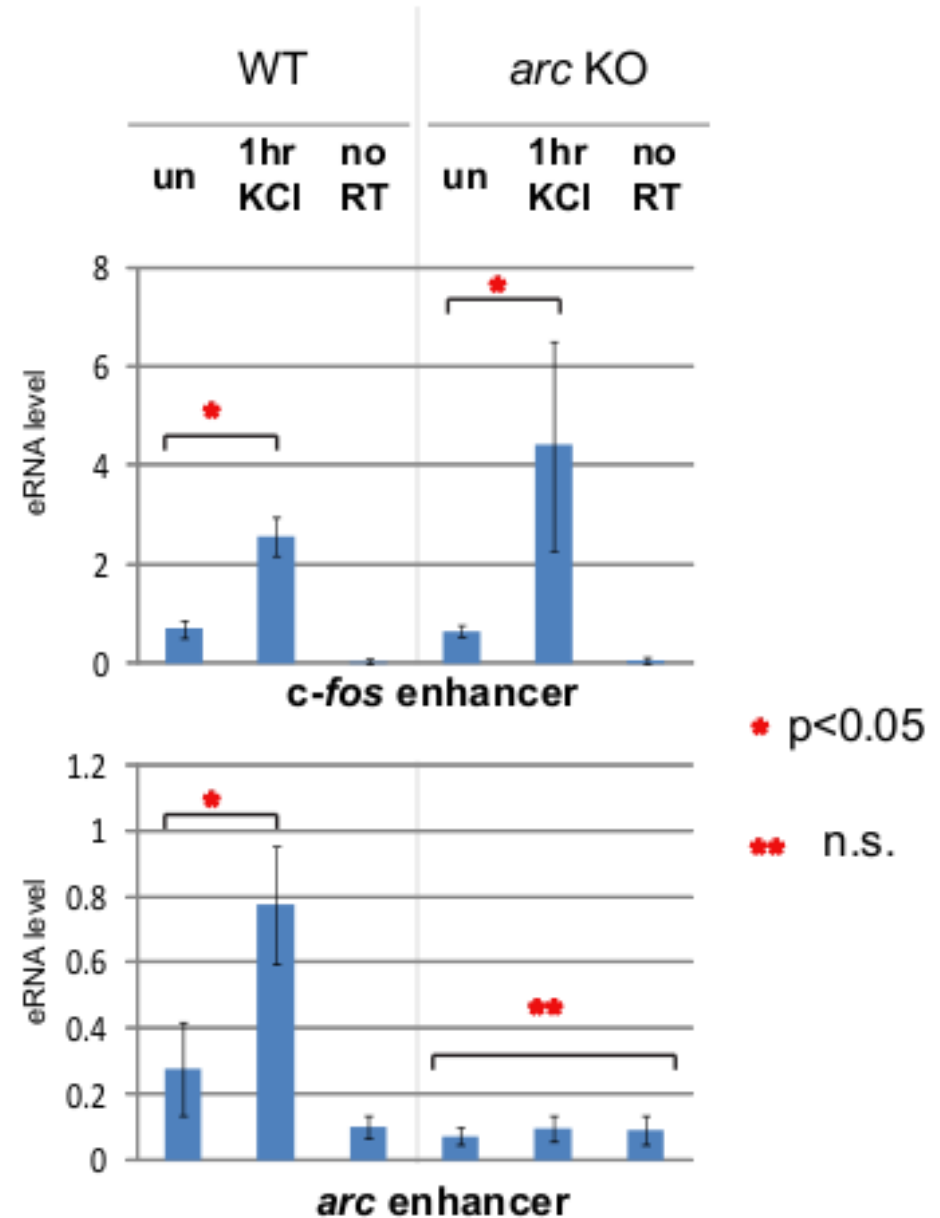
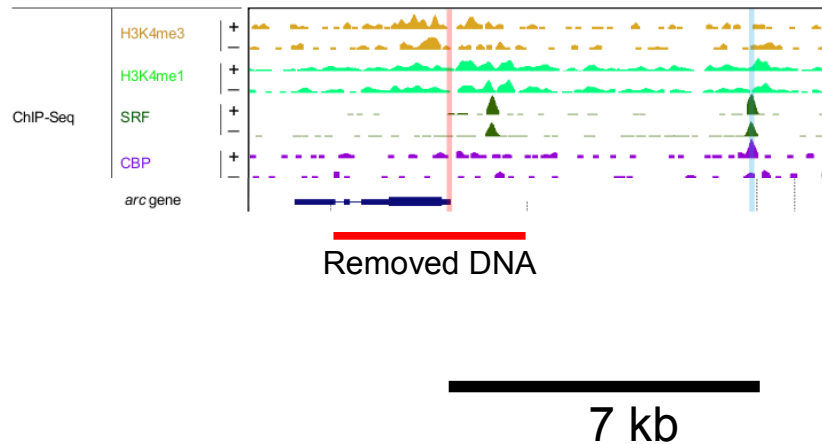
# Transcription at the Fos and Arc enhancers



7 kb

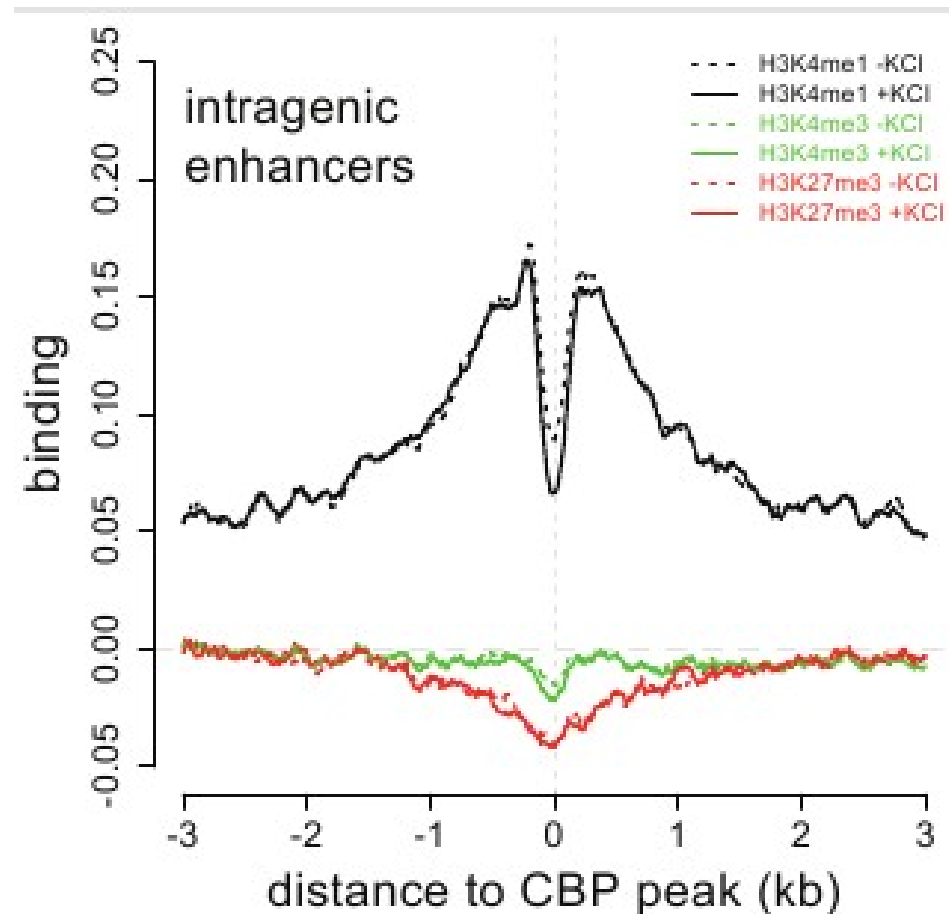


# No transcription at Arc enhancer in mutant



# Intragenic enhancers

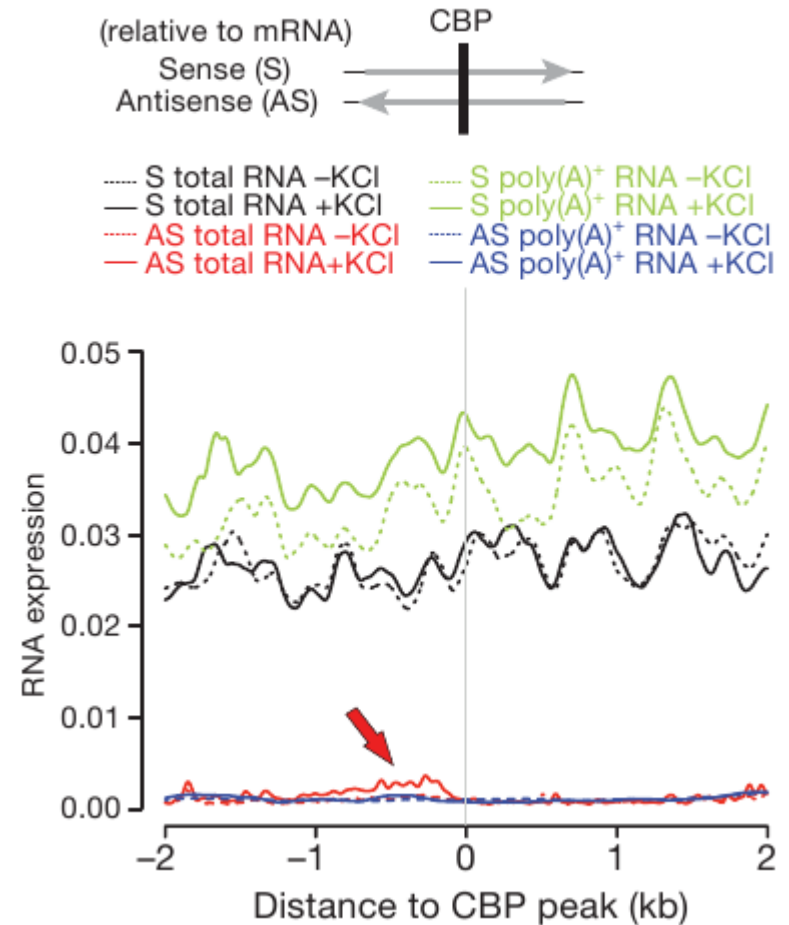
- ~7,000 enhancers overlapping introns
  - H3K4me1, but no H3K4me3





# Intragenic enhancers are also transcribed

- ~7,000 enhancers overlapping introns
  - No signal detectable on sense strand
  - Significant anti-sense transcription



~100 enriched motifs are found

<b>Word</b>	<b>Enrichment</b>	<b>Known TF</b>
TGASTCA	4.74	Fos/Jun
TGACGTCA	6.41	Creb
CTAWWWATA	3.34	Srf
TCGTG	1.56	Npas4
CTGCCAAA	3.34	?

# How abundant are eRNAs compared to mRNAs?

- Identify **all** transcripts in the genome
  - Wavelet-based algorithm for *de novo* detection of transcribed regions accounts for 99.8% of reads
    - Annotated RNAs ~ 98.3%
    - eRNAs ~ 0.02%
      - 1 in 10,000 reads is an eRNA read
      - mRNAs ~100 times more abundant

# Estimating the production rate of eRNAs

$$\frac{dE}{dt} = kN - \frac{E}{\tau_E}$$

$$k = \frac{E^*}{N\tau_E} \sim \frac{10^3}{10^4 \times 10^{-1}\text{h}} = 1\text{h}^{-1}$$

# Parameters for the eRNA fit

$$\lambda = \frac{k_{drop} \text{ s}^{-1}}{k_{elong} \text{ bp}^{-1} \text{ s}^{-1}} \sim \frac{2 \times 10^{-2}}{20} \text{ bp}^{-1} = 10^{-3} \text{ bp}^{-1}$$

$$\tau_{decay} = \tau_{find} + \tau_{bp} L$$

$$H(x, t) = \frac{k\kappa}{\mu_x(\mu_x - \lambda)} (e^{-\lambda x} - e^{-\mu_x x}) \times e^{-\mu t}$$

$$E(x) = \sqrt{\frac{\pi}{2\lambda}} \frac{\gamma k}{\lambda} e^{-\delta^2/2\lambda - \lambda x^2/2} i \left[ \text{erf}\left(\frac{\delta i - \lambda i x}{\sqrt{2\pi}}\right) - \text{erf}\left(\frac{\delta i}{\sqrt{2\lambda}}\right) \right]$$

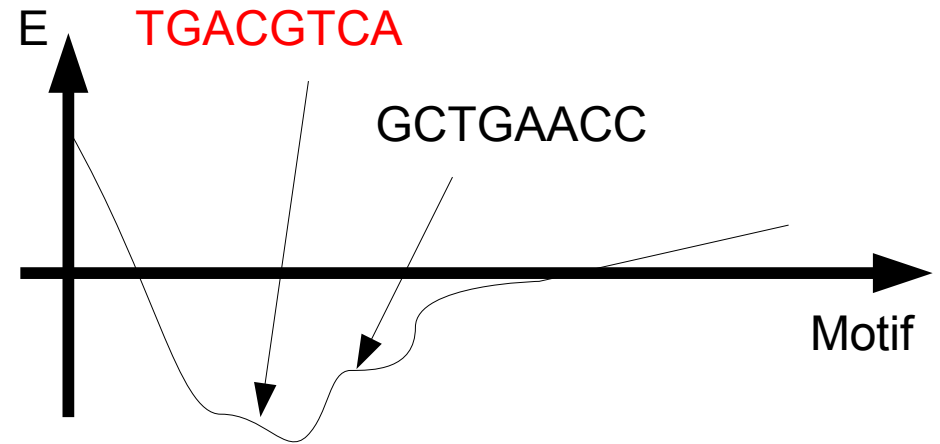
# Can the read count be predicted from sequence motifs?

$$R = f(E(\theta)) + \xi$$

#reads

Binding energy

noise



# Assume transfer function maximizes mutual information

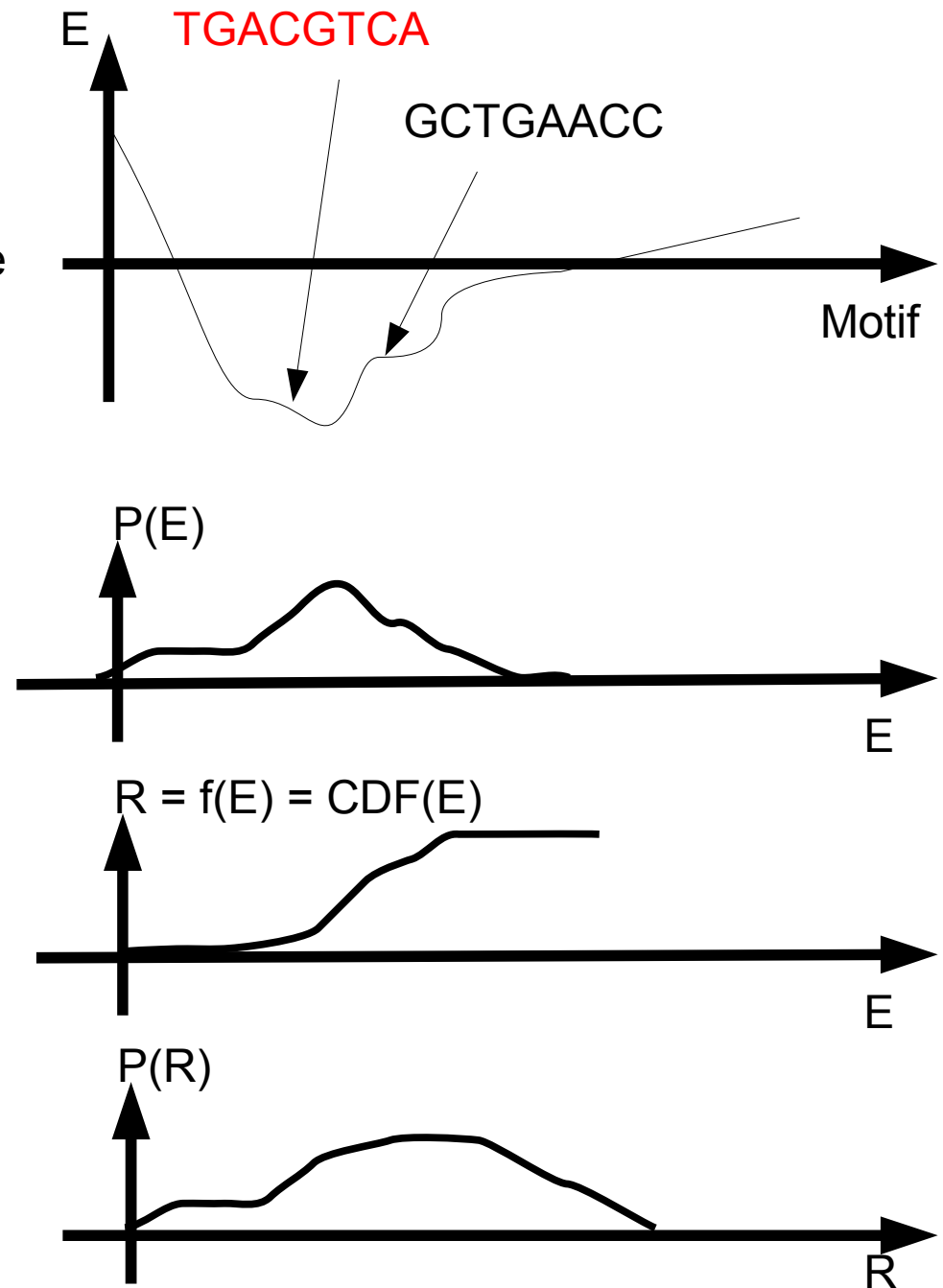
$$R = f(E(\theta)) + \xi$$

#reads  $\uparrow$   $R$

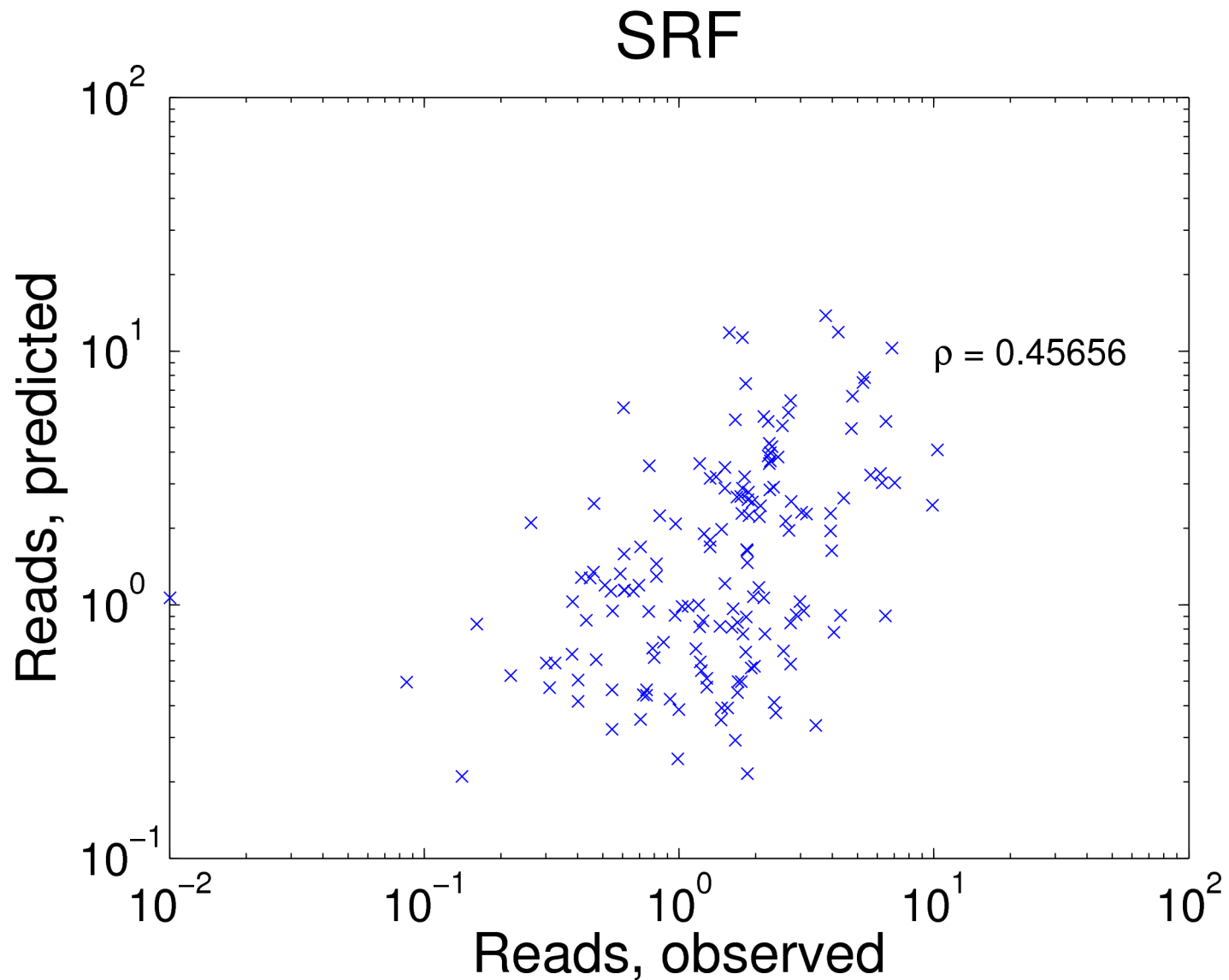
Binding energy (from JASPAR)  $\uparrow$   $E(\theta)$

noise  $\uparrow$   $\xi$

- $f$  monotone
- $\max I(R; E)$
- Noise small and gaussian



# Number of reads can be predicted by binding energy





# A PDE for histone levels

$$\frac{\partial H}{\partial x} + \frac{\partial H}{\partial t} = \kappa P(x, t) - \mu_x H - \mu_t H$$

$$H(x, t) = \frac{\kappa \kappa}{\mu_x (\mu_x - \lambda)} (e^{-\lambda x} - e^{-\mu_x x}) \times e^{-\mu_t t}$$

