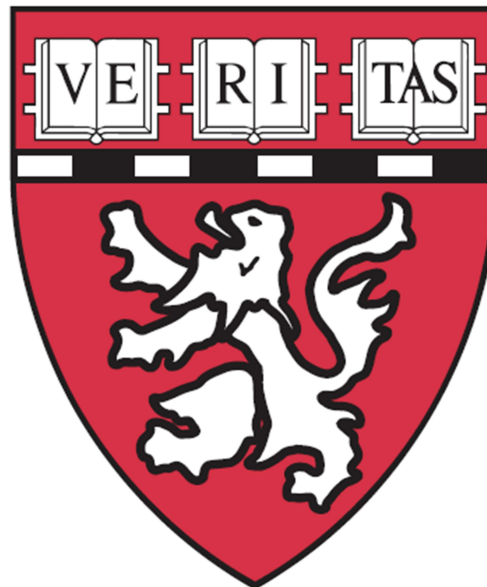


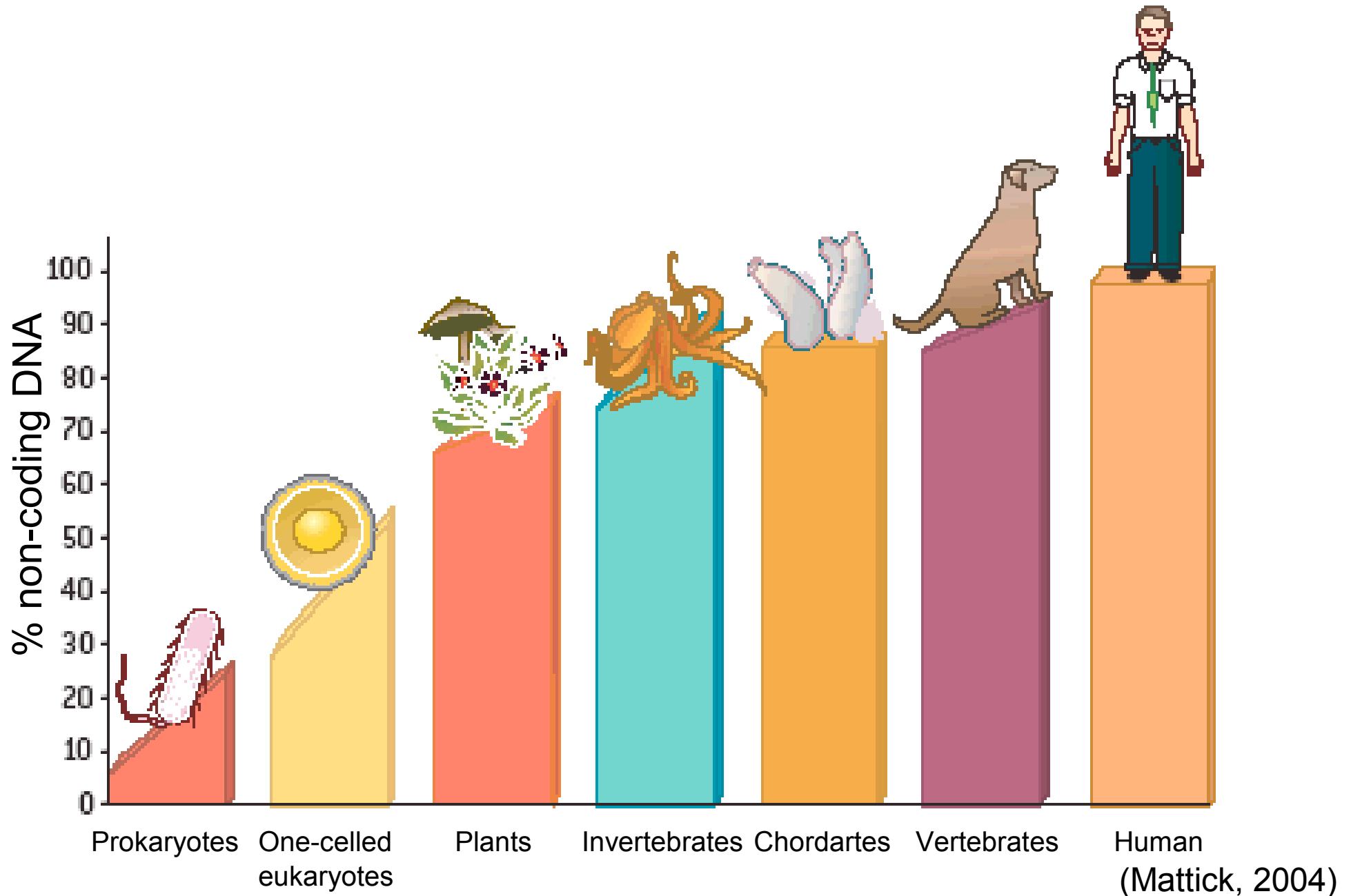
Probing the function of non-coding sequences in human and mouse using ChIP-seq and RNA-seq

Martin Hemberg

University of Pennsylvania
February 21, 2011



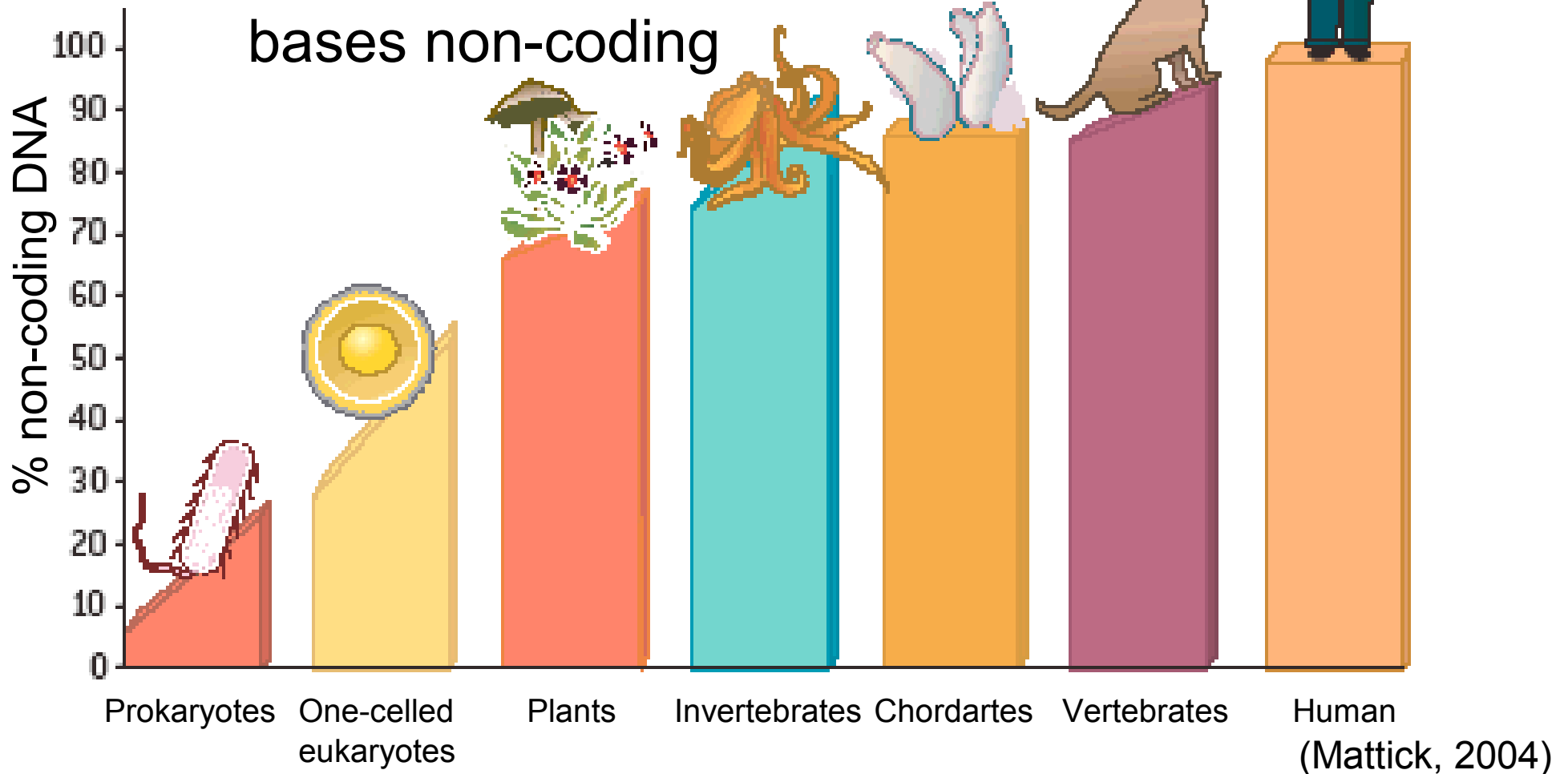
Most of the genome is **not** protein-coding



Most of the genome is **not** protein-coding

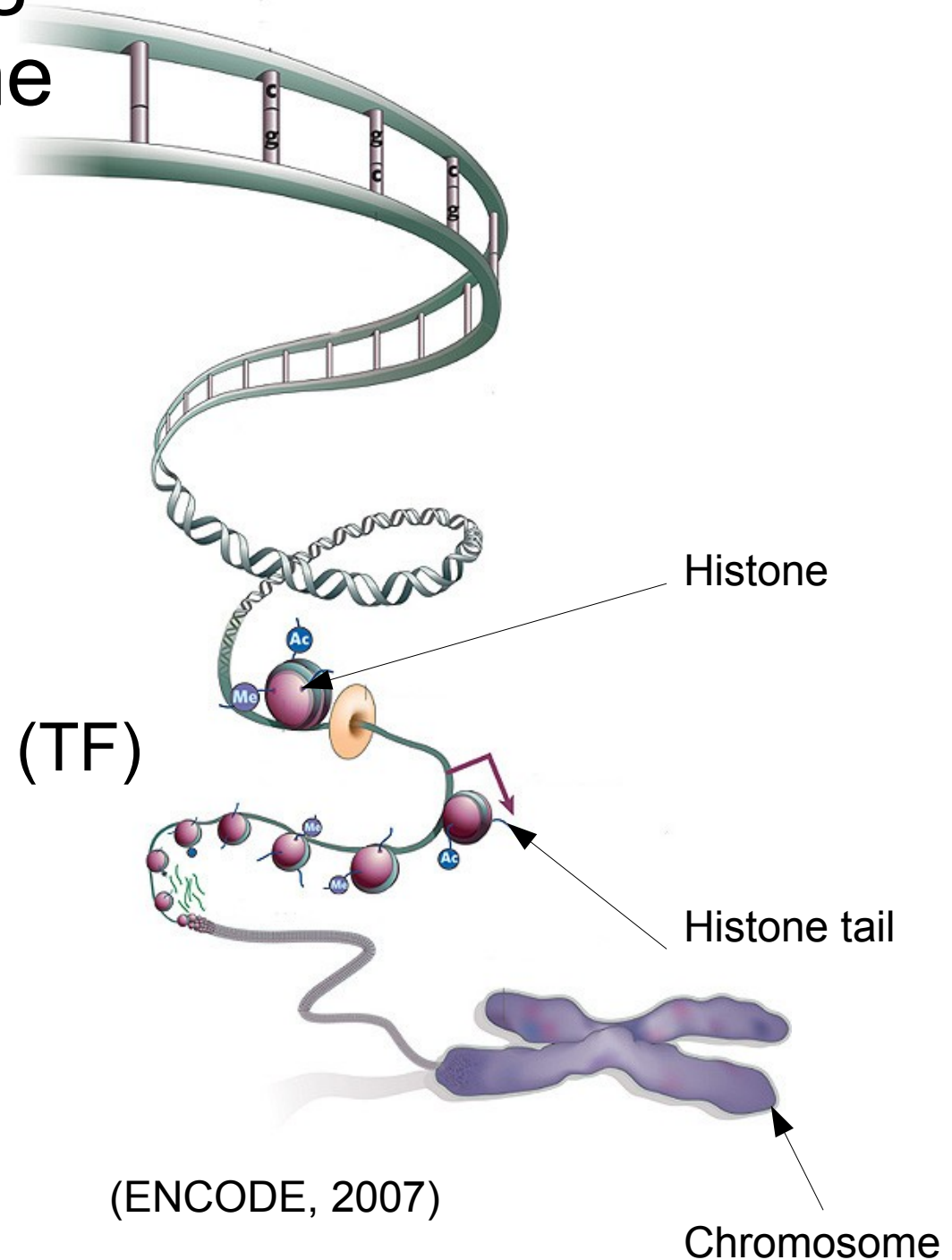
- 5% mammalian genome highly conserved

– 60% of highly conserved bases non-coding



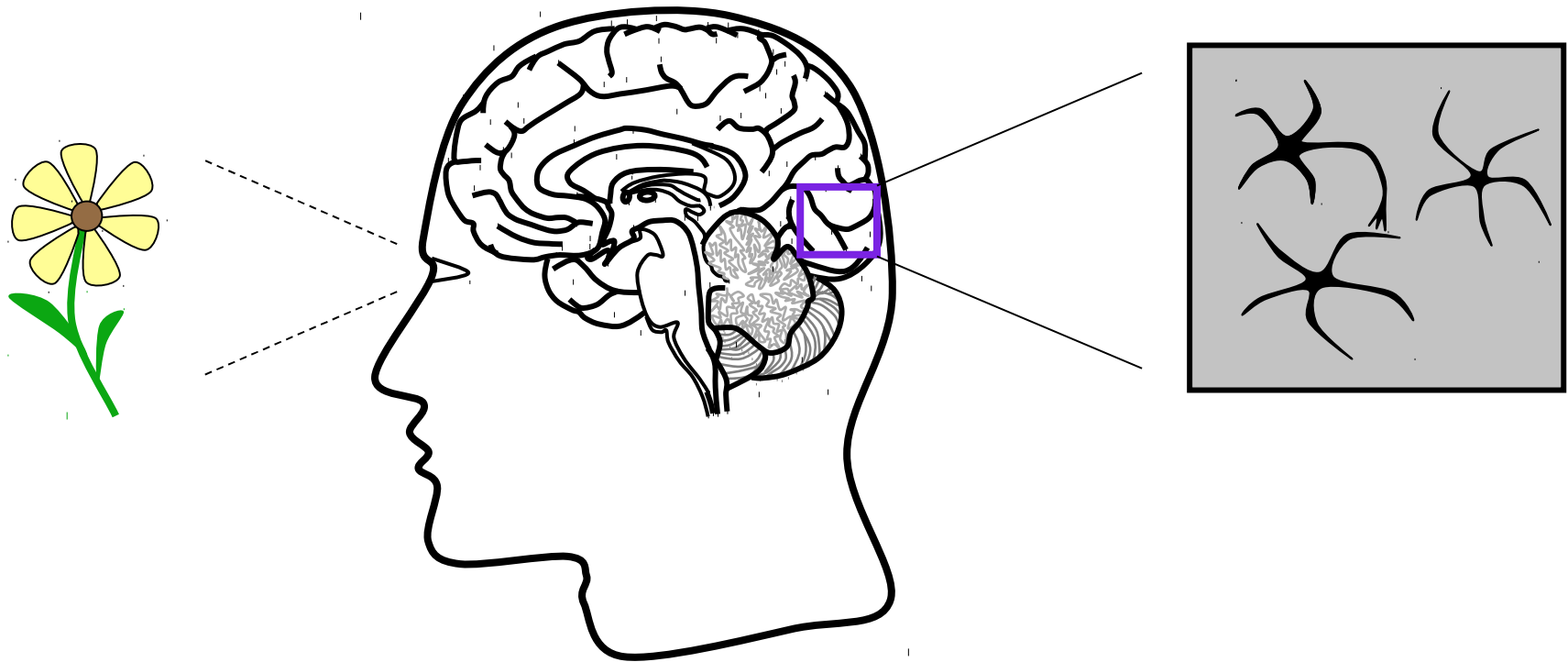
Additional layers of regulation determine the function of the genome

- Cell-type specific
 - DNA methylation
 - Post-translational modification of histone tails
 - Transcription factor (TF) binding

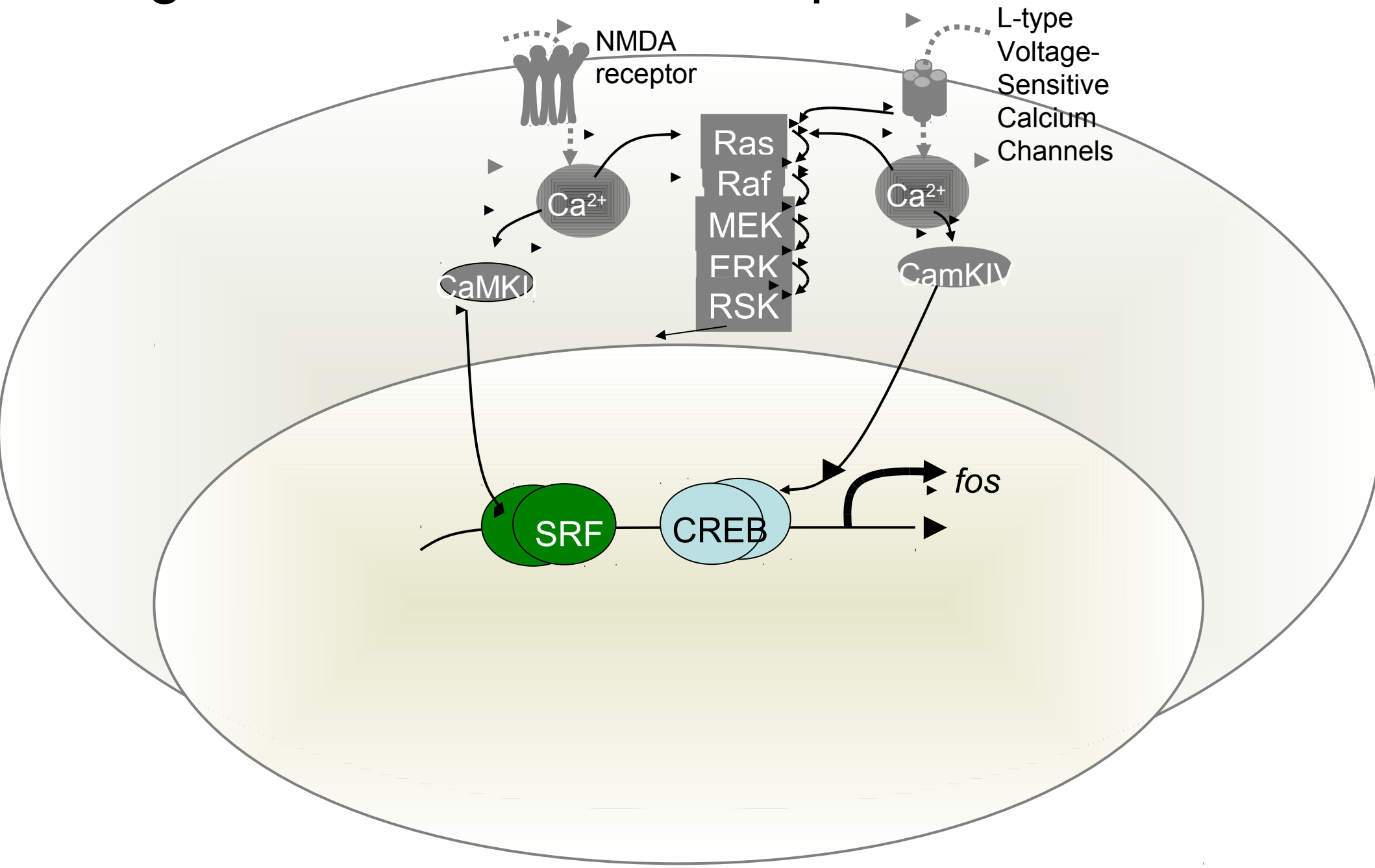


Activity dependent gene expression

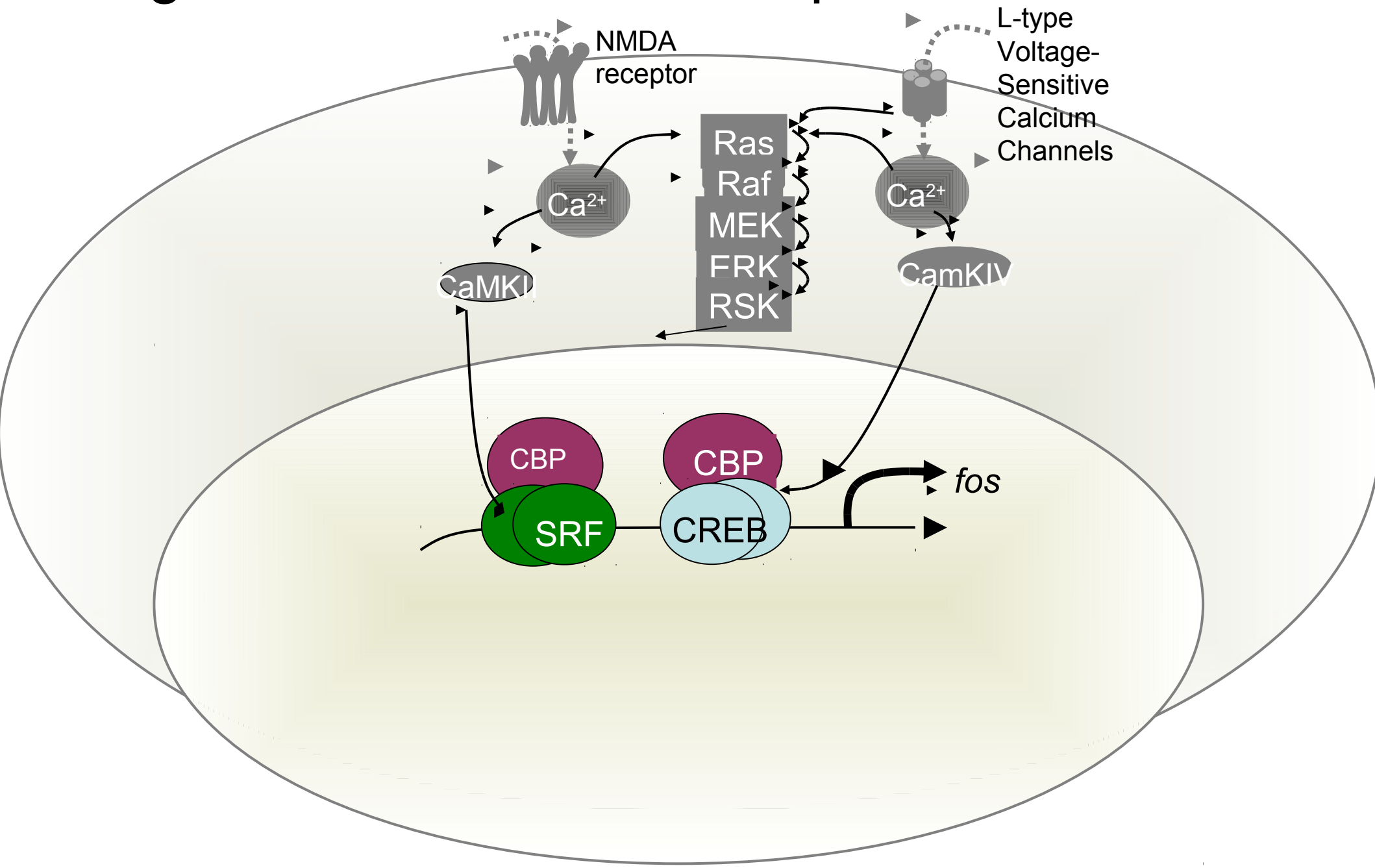
- Sensory experience shapes wiring in the brain
 - Synapses and patterns of neuronal activity changed



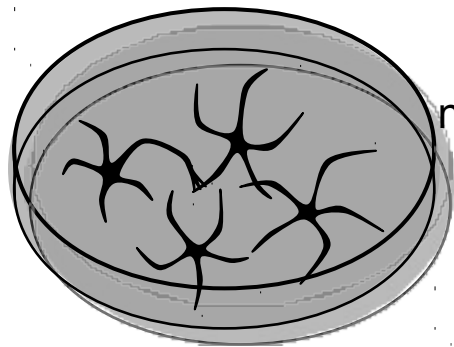
Regulation of c-fos transcription



Regulation of c-fos transcription



An experimental system for genome-wide study of activity dependent gene expression



mouse cortical neurons

neuronal activation via potassium chloride (KCl) depolarization

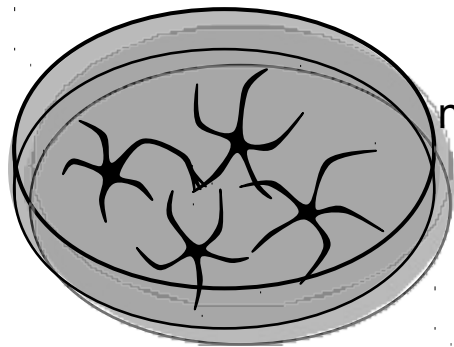
- KCl

+ KCl

ChIP-Seq
RNA-Seq

ChIP-Seq
RNA-Seq

An experimental system for genome-wide study of activity dependent gene expression



mouse cortical neurons

neuronal activation via potassium chloride (KCl) depolarization

- KCl

+ KCl

ChIP-Seq
RNA-Seq

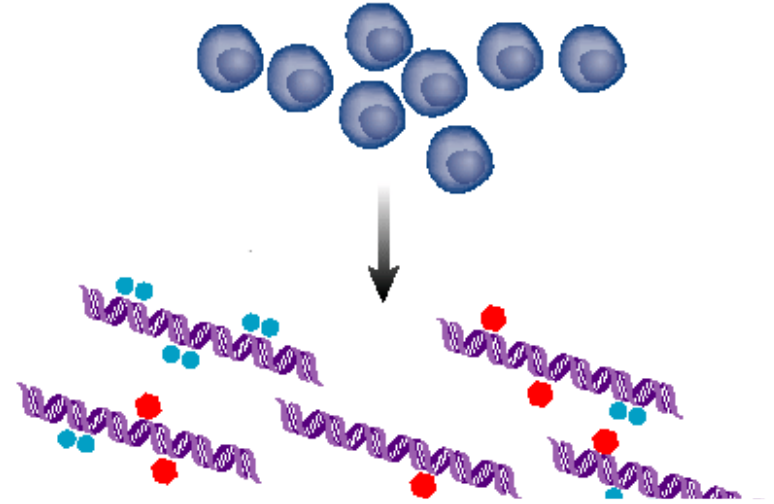
ChIP-Seq
RNA-Seq



Jesse Gray
Tae-Kyung Kim
Greenberg Lab

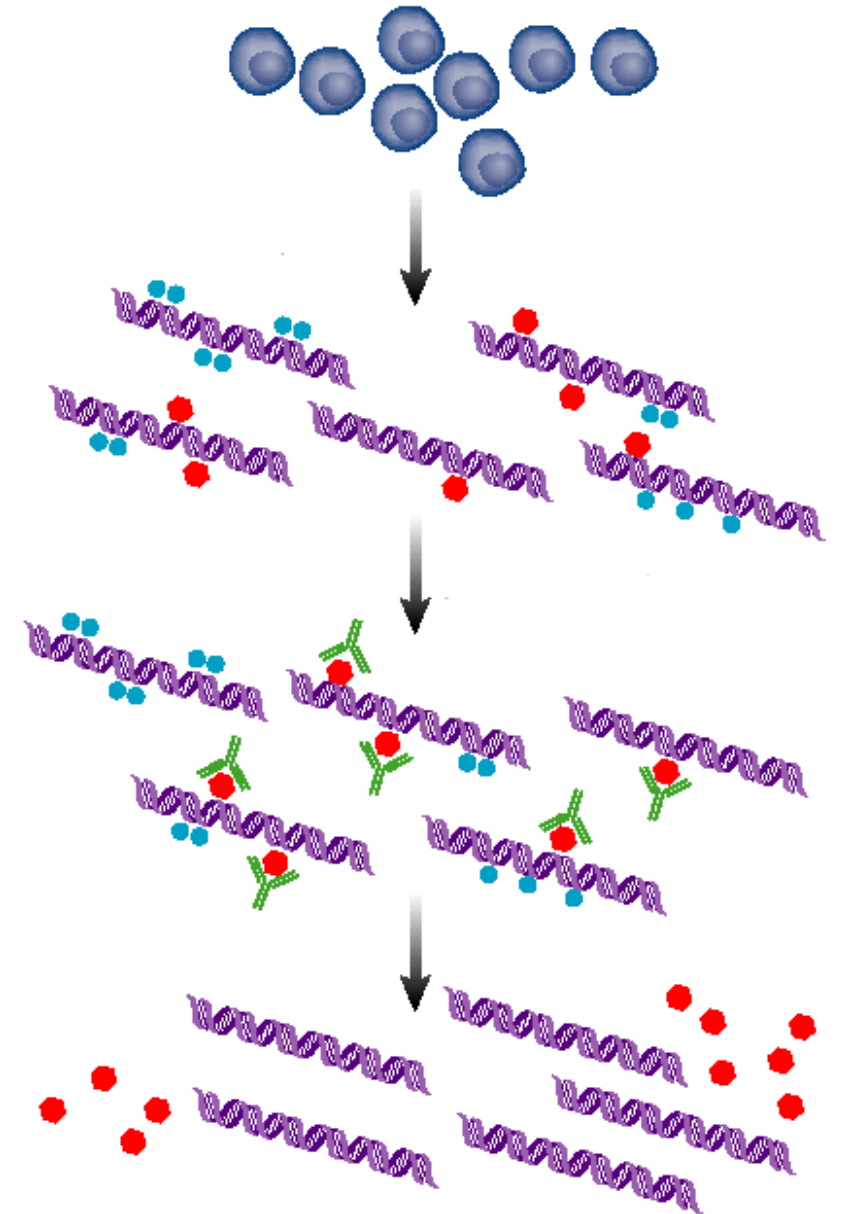
Chromatin immunoprecipitation and sequencing (ChIP-Seq) finds TF binding sites *in vivo*

- Cross-link TF
- Fragment DNA



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

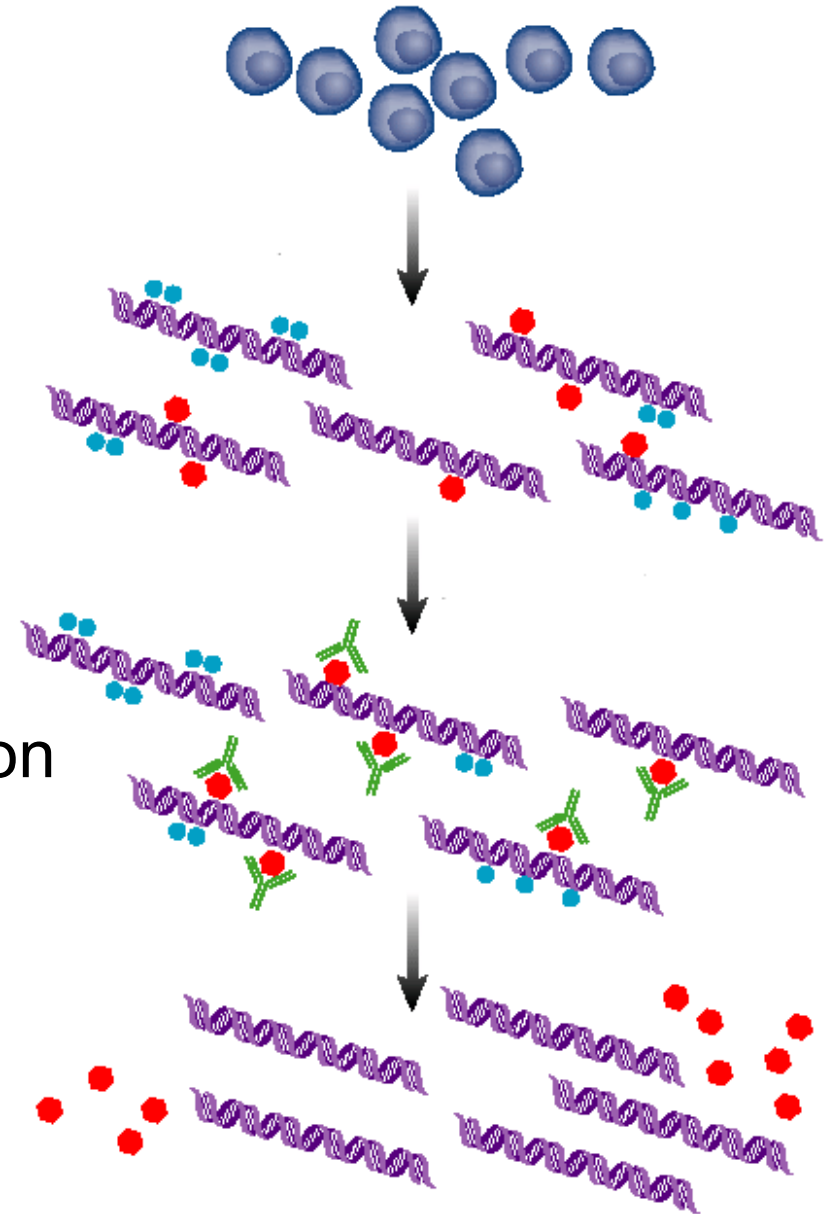
- Cross-link TF
- Fragment DNA
- Extract with antibody
- Reverse crosslink
- Sequence fragments



(Mardis, 2007)

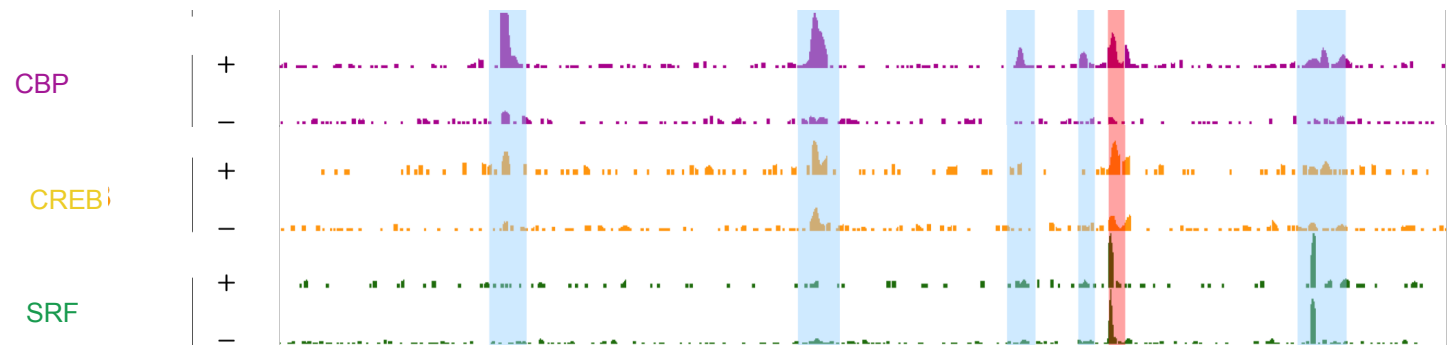
Chromatin immunoprecipitation and sequencing (ChIP-Seq) finds TF binding sites *in vivo*

- Cross-link TF
- Fragment DNA
- Extract with antibody
- Reverse crosslink
- Sequence fragments
 - Before and after KCl stimulation
 - CREB, SRFCBP, RNAPII
H3K4me3, H3K4me1
 - Input



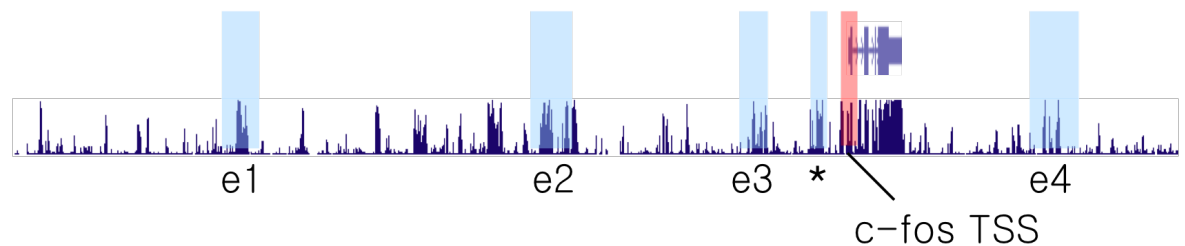
(Mardis, 2007)

CBP binding depends strongly on activity at the *fos* promoter and flanking loci



20 kb

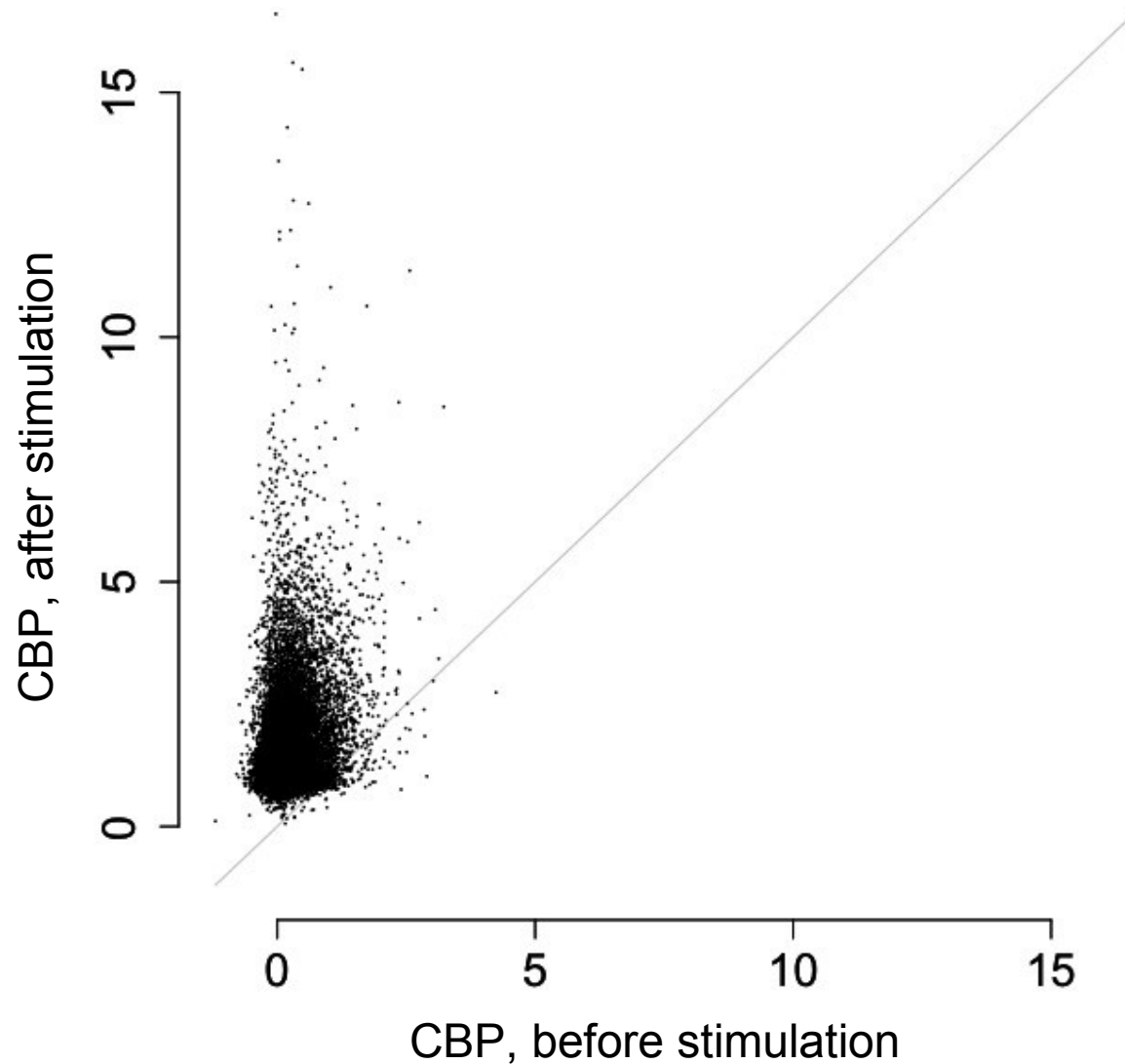
c-fos gene locus
conservation



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

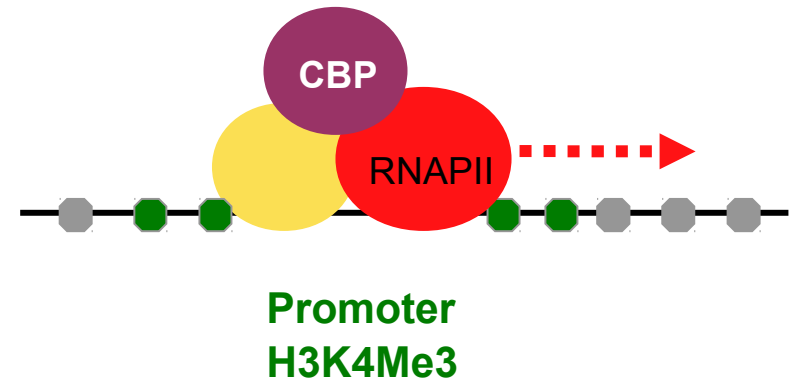
- Assume both ChIP and input $\sim \text{Poisson}(\lambda_c)$
 - ~ 1 read/100 bps
 - Sliding window of size 240 bps
- $Z = \# \text{ChIP reads} - \# \text{Input reads} \sim \text{Skellam}(\lambda_c, \lambda_i)$
- False Detection Ratio (FDR) determines cut-off
 - Controls for expected number of false positives

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



Only ~3000 CBP peaks at promoters

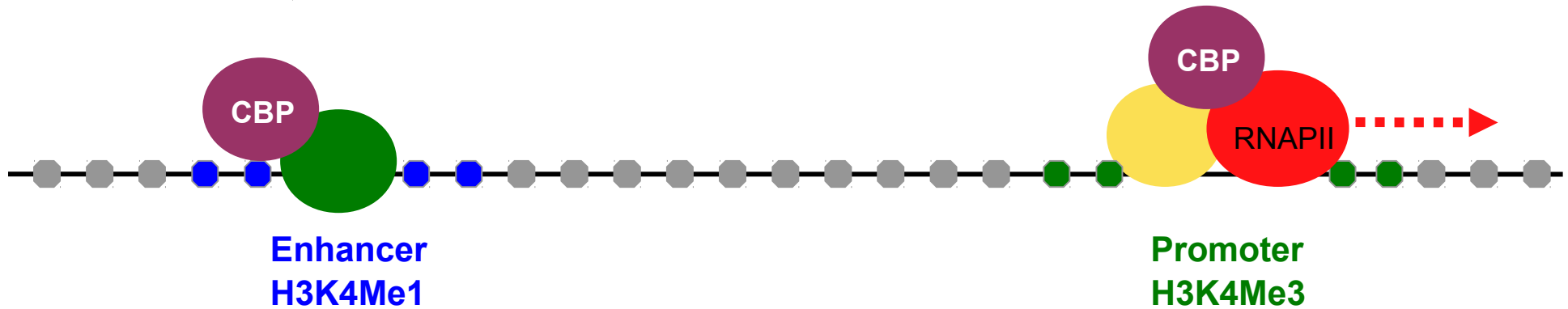
~3,000



CBP hypothesized to bind at enhancers

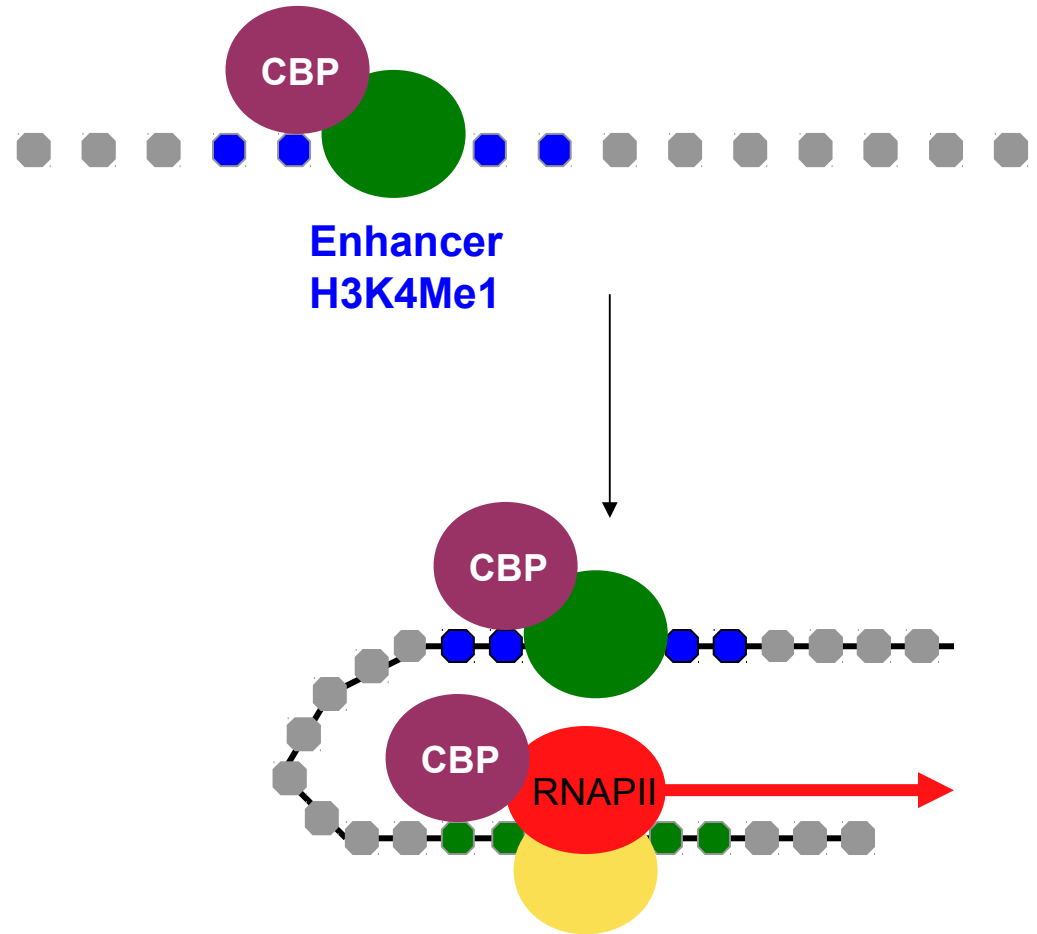
~25,000?

~3,000



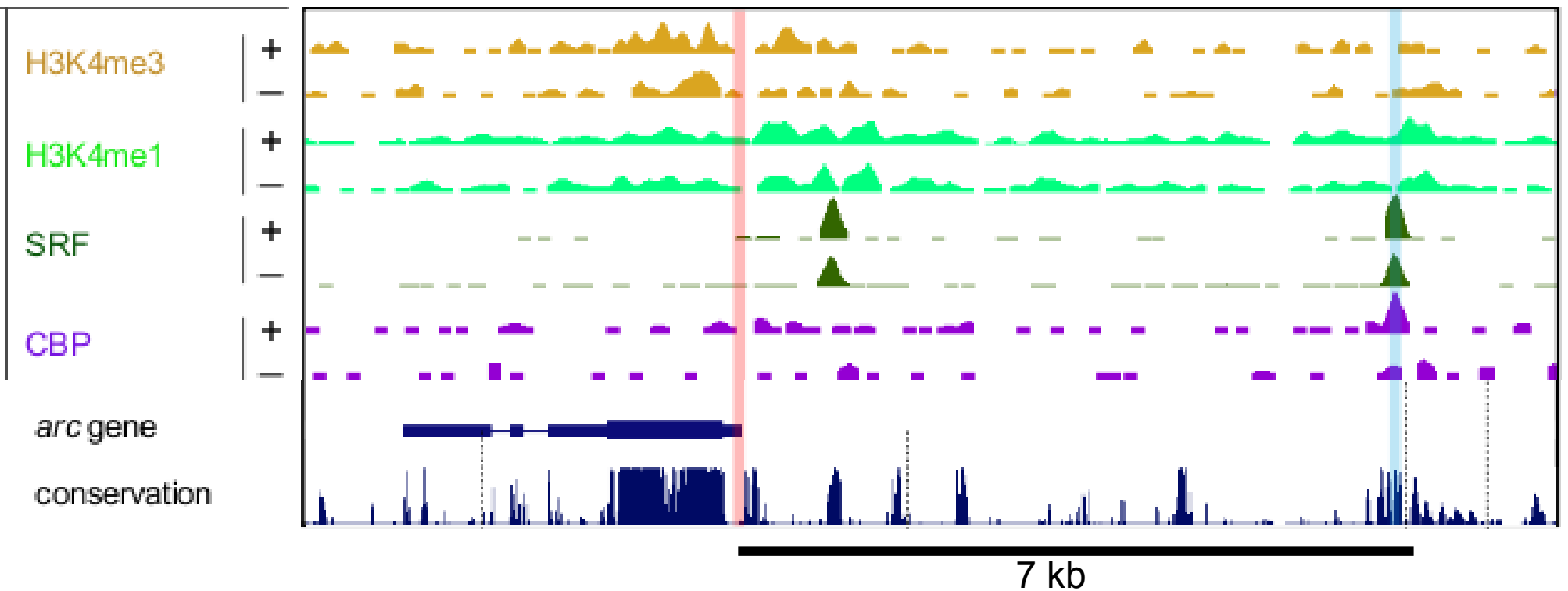
Enhancers are distal TF binding sites

- Various mechanisms for interaction with promoters suggested
- Marked by high levels of H3K4me1

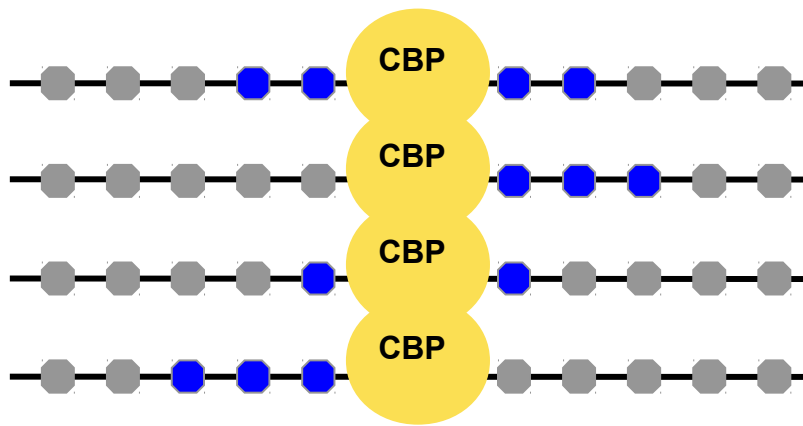


ENCODE, 2007
Heintzman et al, 2007
Roh et al, 2005
Visel et al, 2009

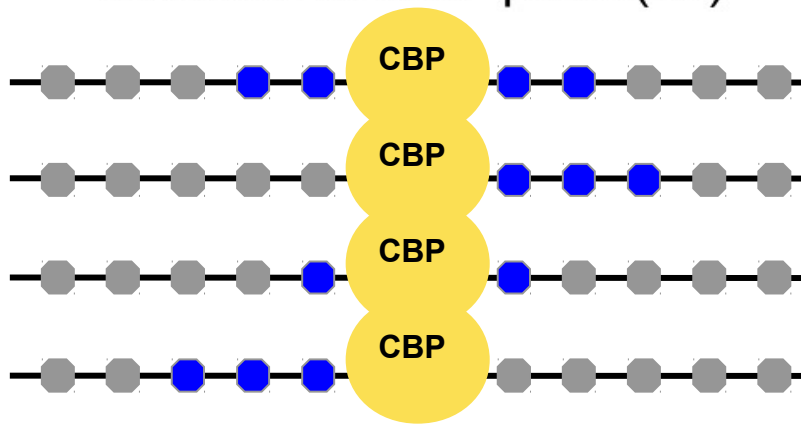
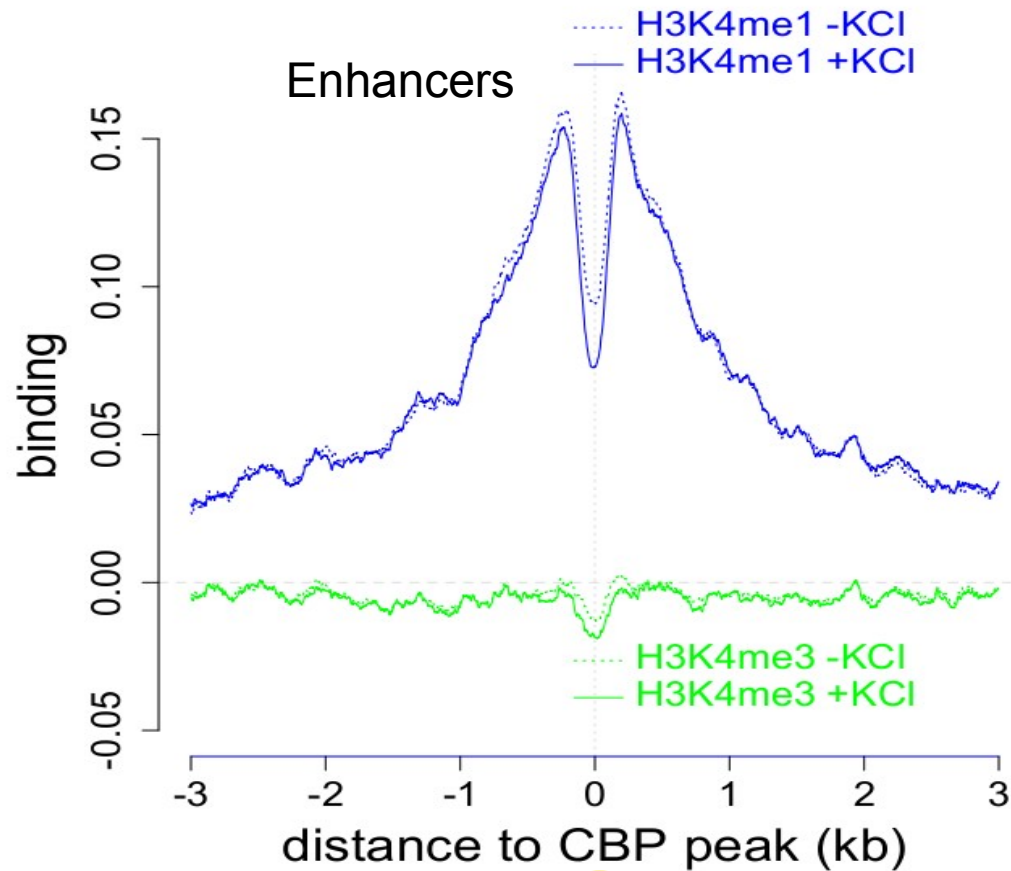
Distal CBP peaks have high levels of H3K4me1 but not H3K4me3



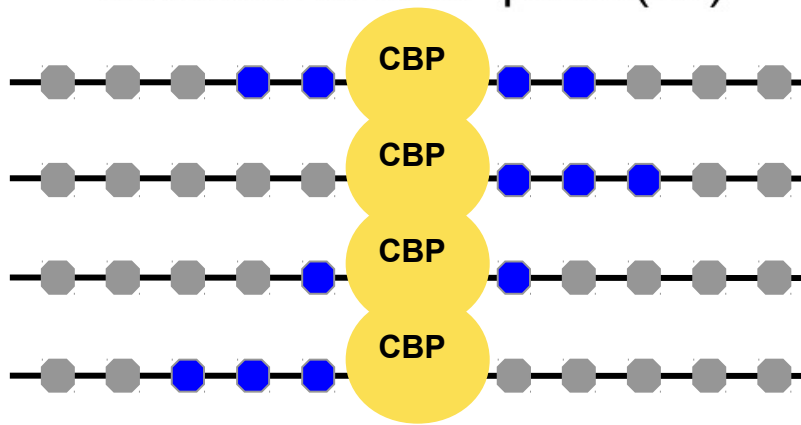
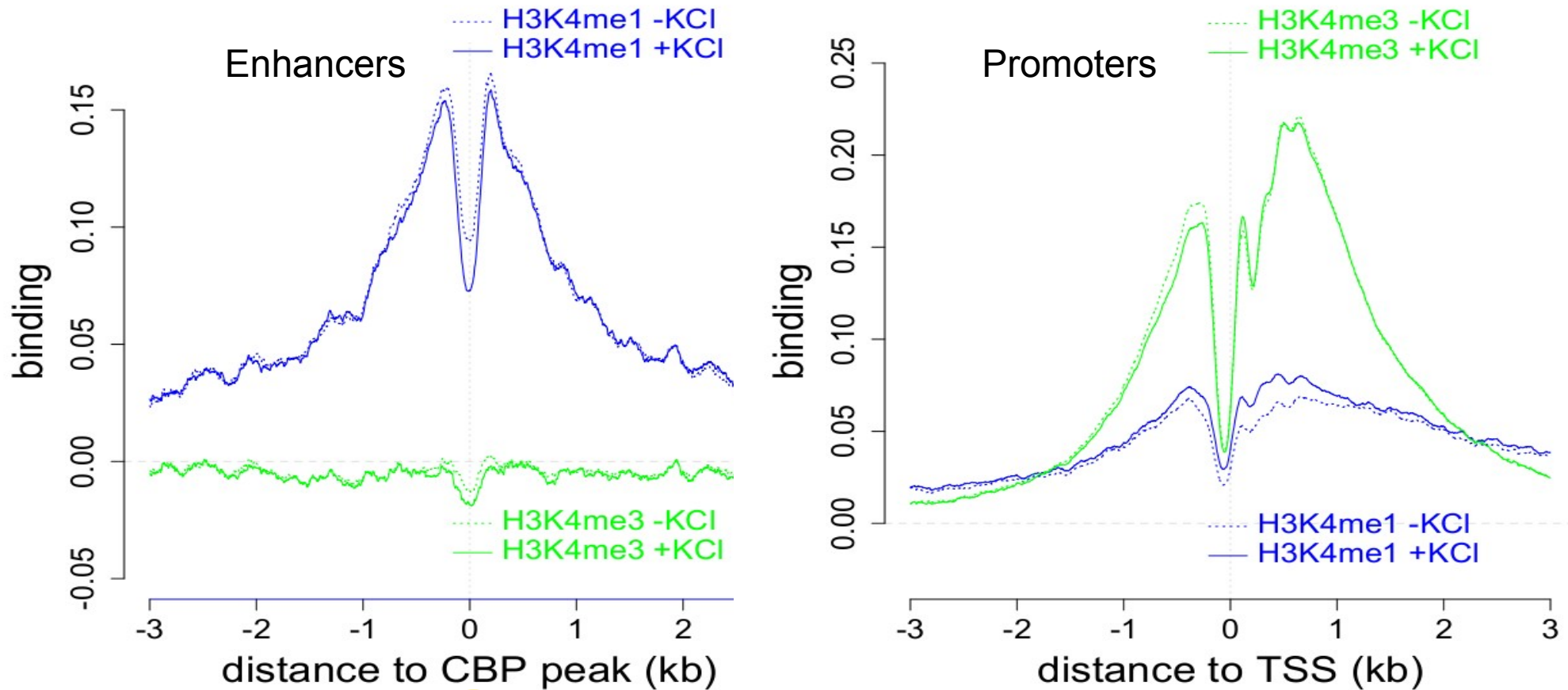
Aligning CBP peaks to calculate average binding profiles



Most CBP peaks have high levels of H3K4me1 but not H3K4me3



Transcription start sites (TSSs) have high levels of H3K4me1 and H3K4me3

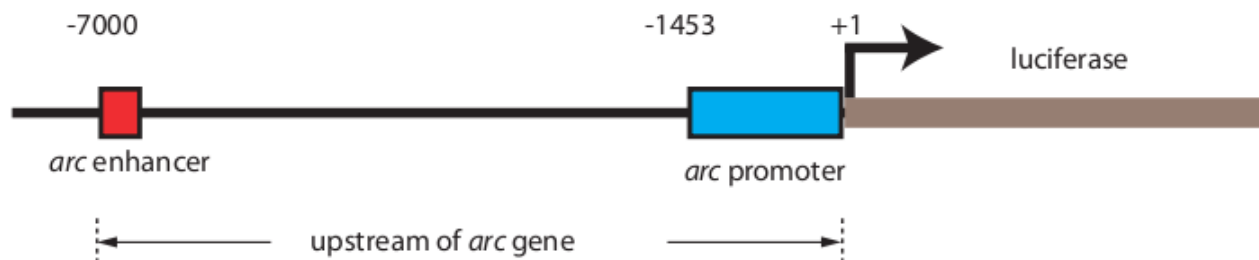


Identifying 5130 activity regulated enhancers

- CBP peak
- High levels of flanking H3K4me1
- Low levels of H3K4me3
- >1 kb from annotated promoter

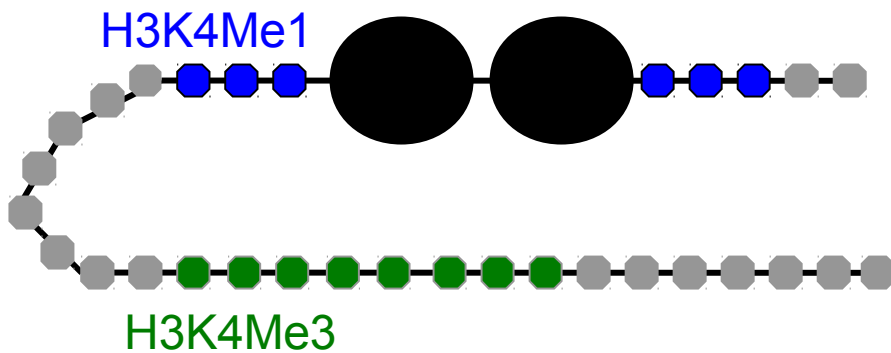
Identifying 5130 activity regulated enhancers

- CBP peak
- High levels of flanking H3K4me1
- Low levels of H3K4me3
- >1 kb from annotated promoter
 - 8/8 validated in luciferase assay
 - ~7000 intragenic enhancers

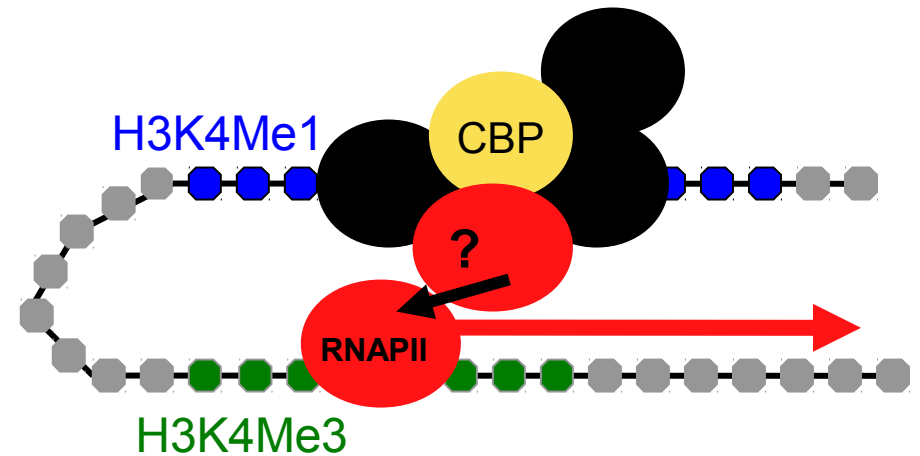


Properties of activity regulated enhancers

Before neuronal activation

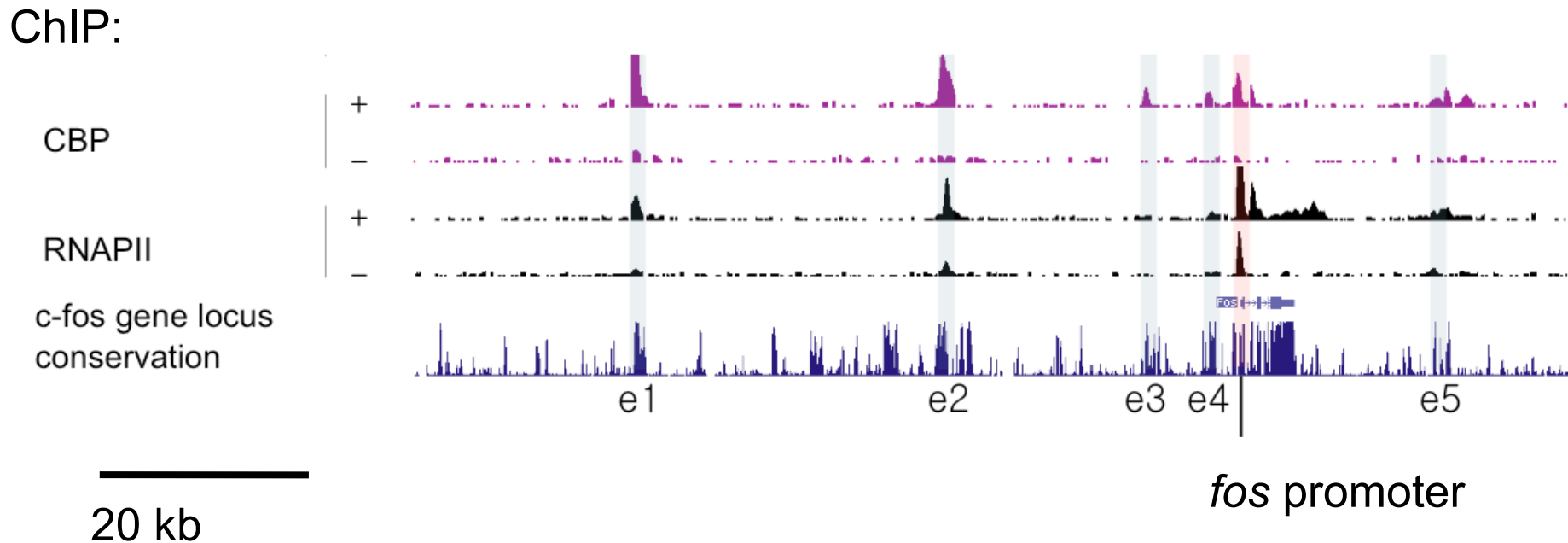


After neuronal activation

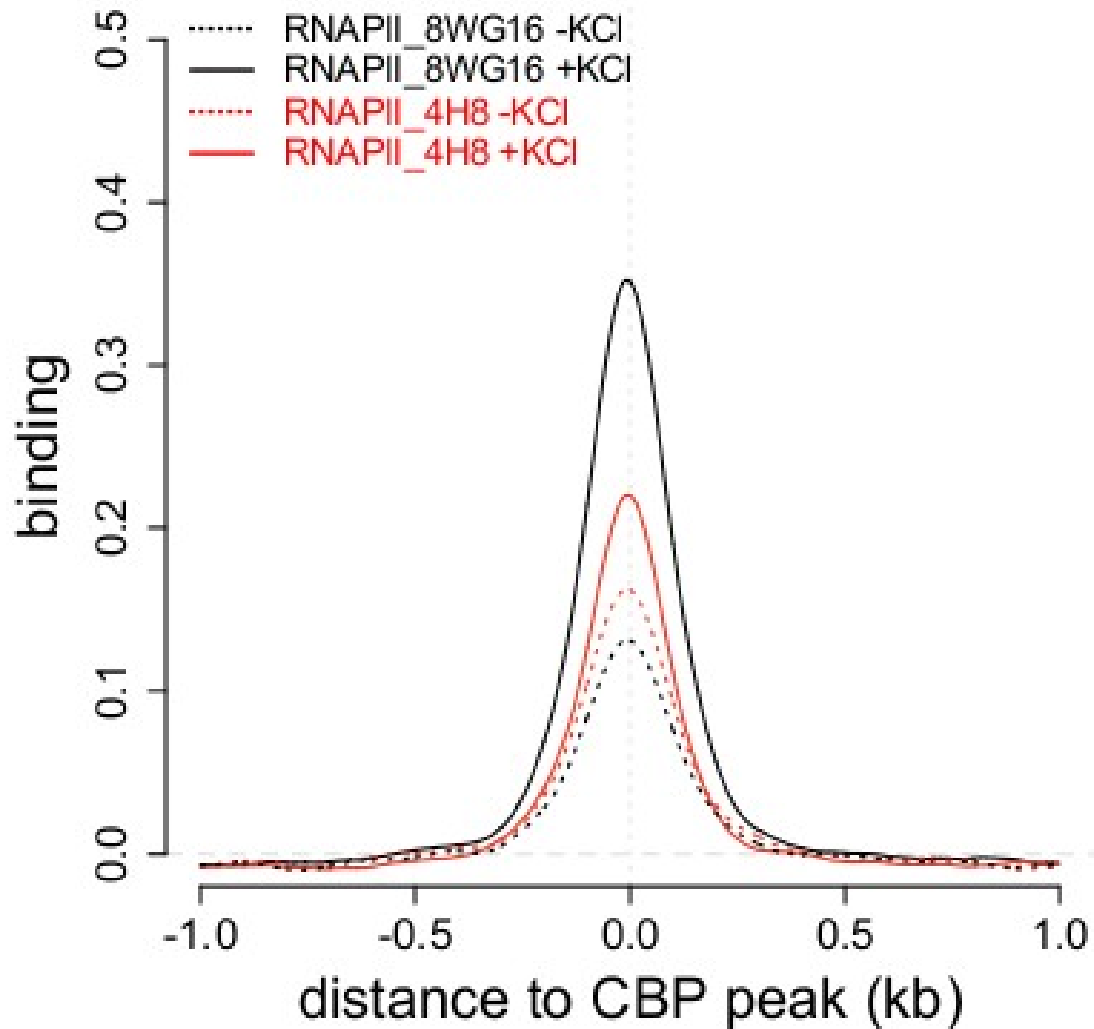


- Does RNAPII bind at enhancers?

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

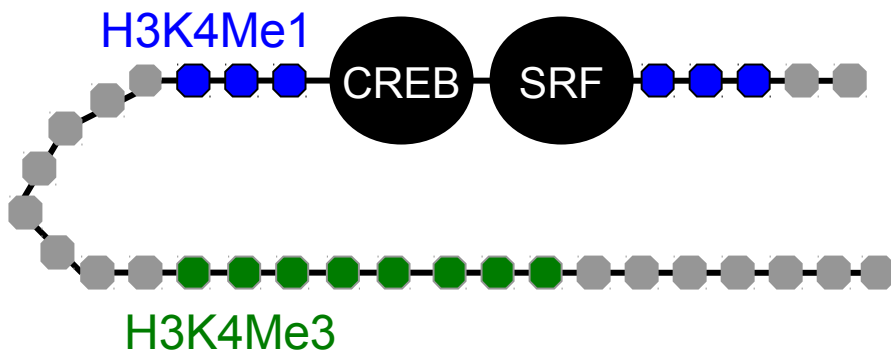


RNAPII is recruited at all enhancers

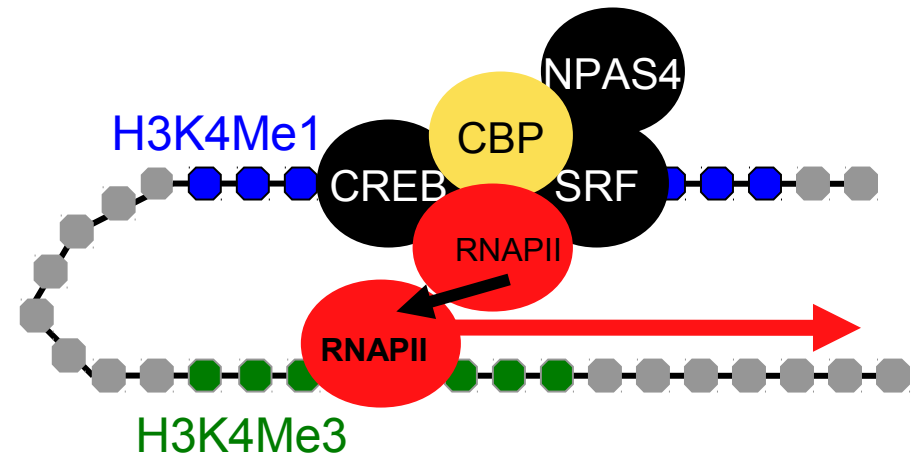


Properties of activity regulated enhancers

Before neuronal activation



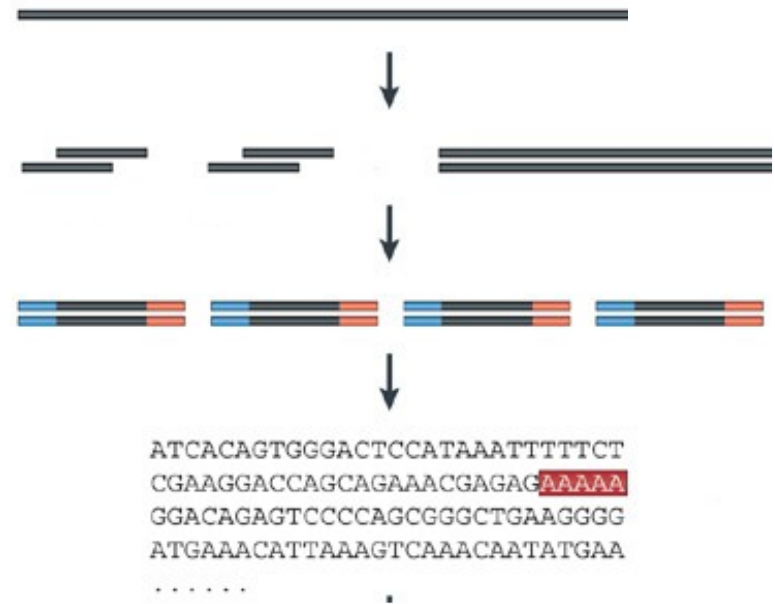
After neuronal activation



- Does RNAPII bind at enhancers?
- Are transcripts produced at enhancers?

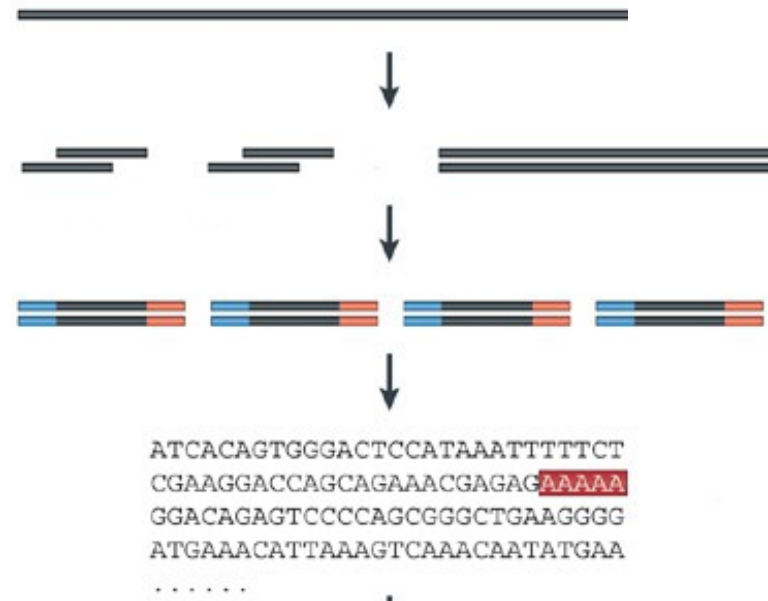
RNA-Seq reveals which parts of the genome are transcribed

- Fragment
- RNA → cDNA
- 35 bp reads mapped to genome

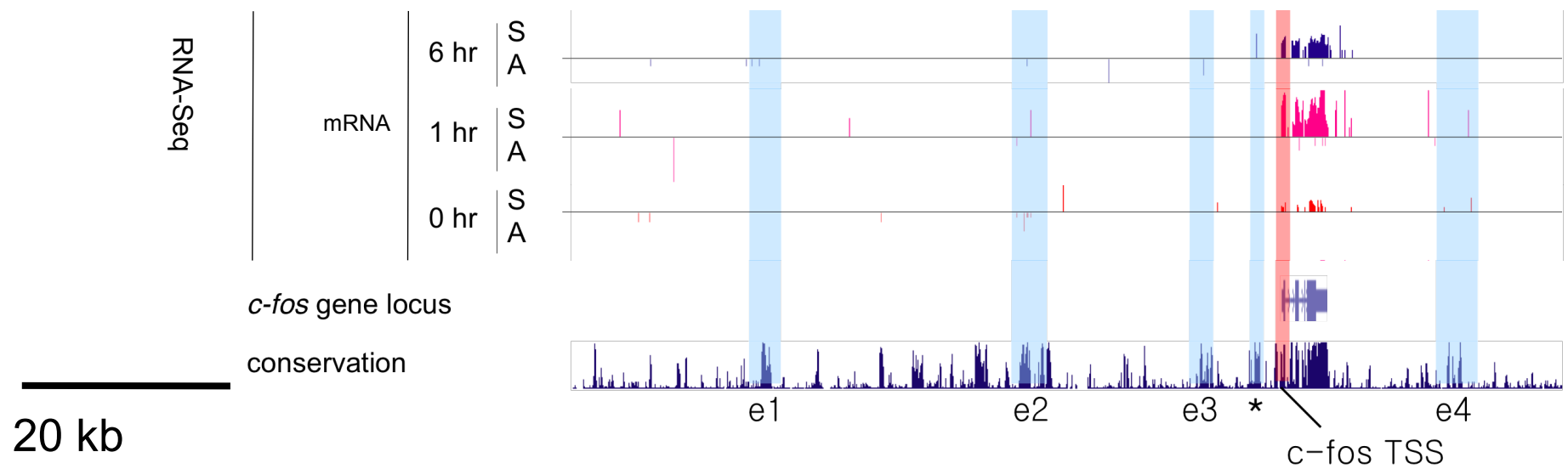


RNA-Seq reveals which parts of the genome are transcribed

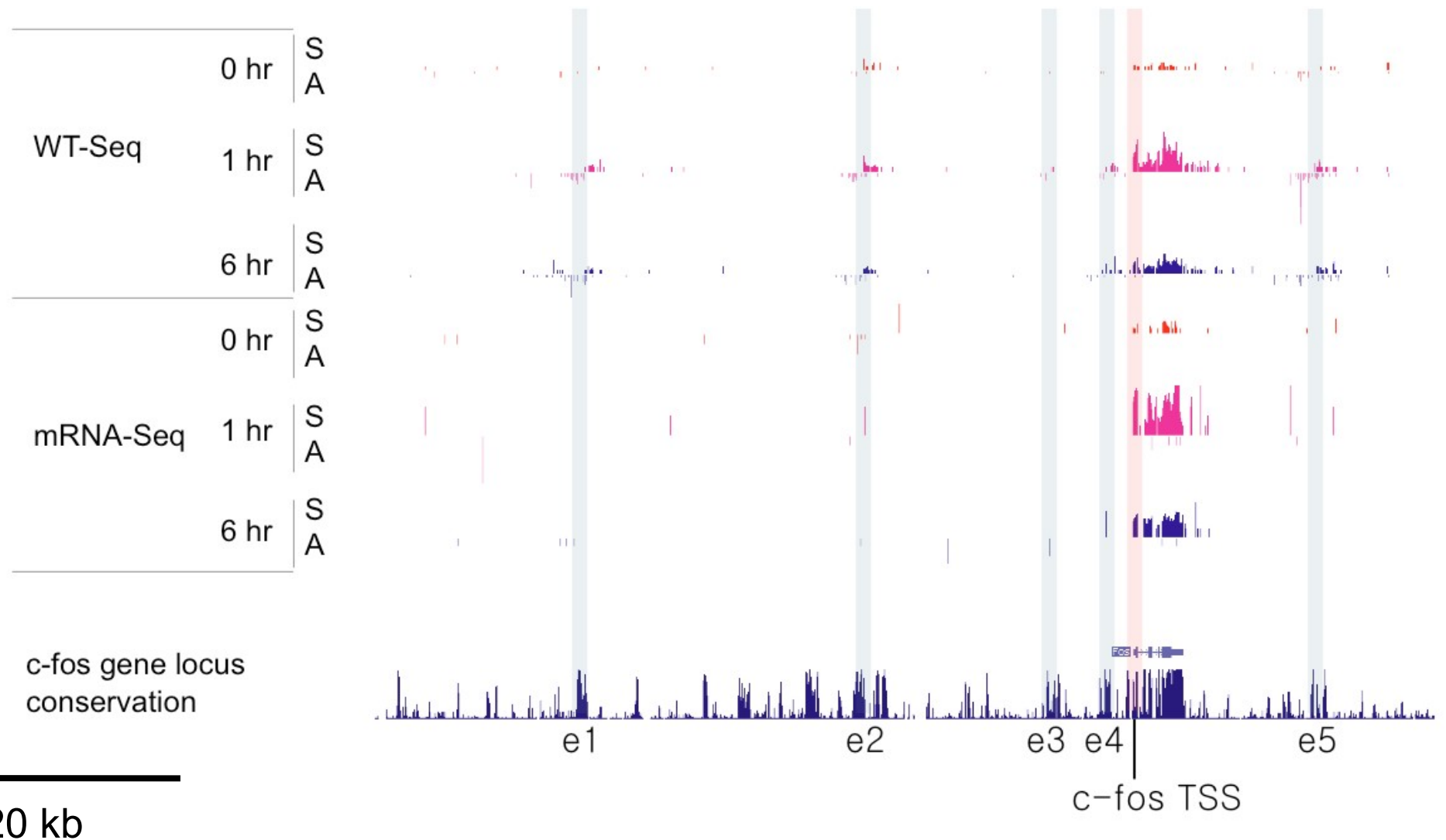
- Fragment
- RNA → cDNA
- 35 bp reads mapped to genome
 - Before and after KCl
 - Total RNA and polyA+



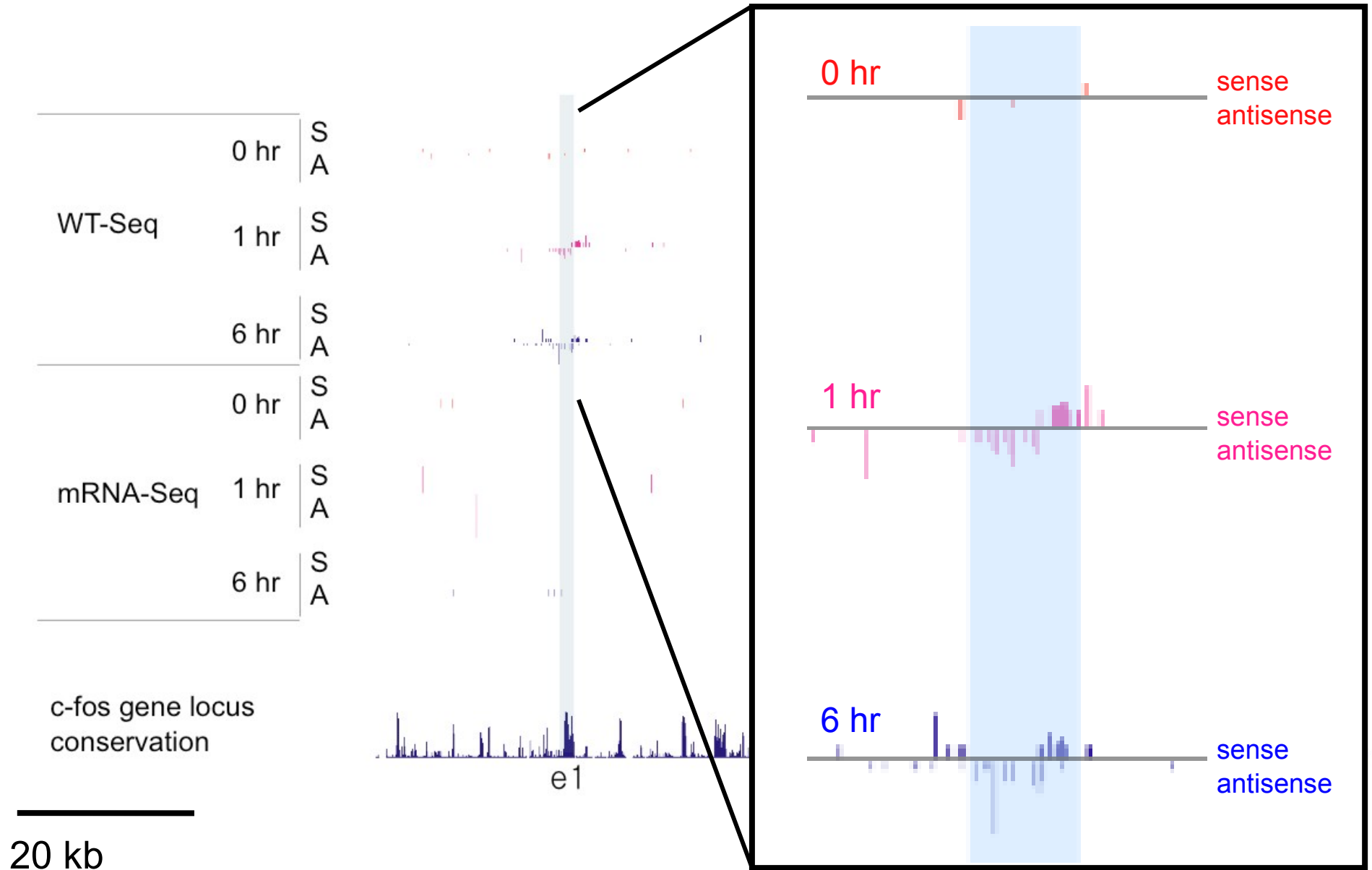
mRNA levels are strongly induced at the *fos* locus



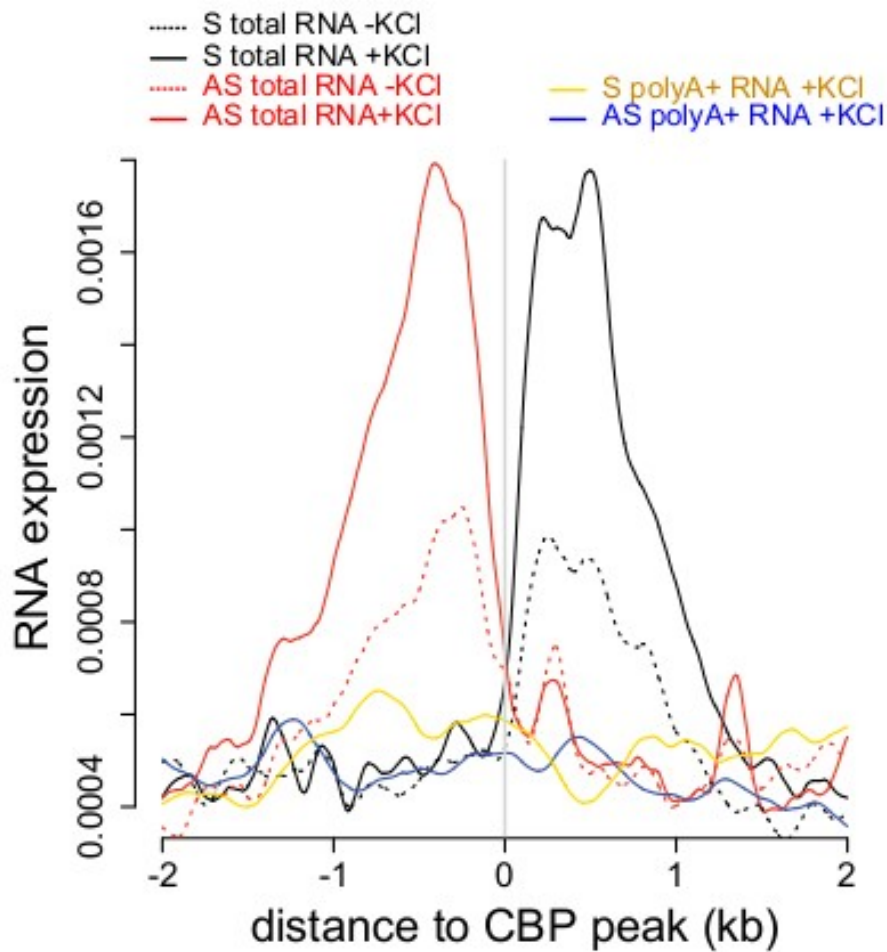
Transcription of enhancer RNA (eRNA) at the *fos* locus



Transcription of enhancer RNA (eRNA) at the *fos* locus



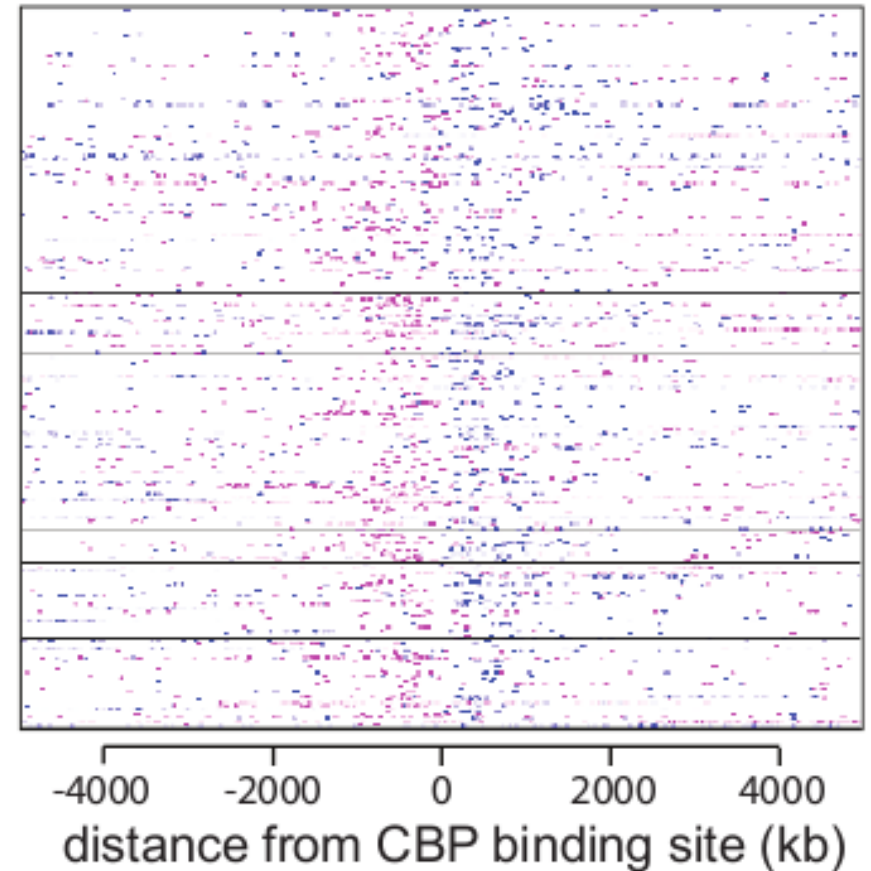
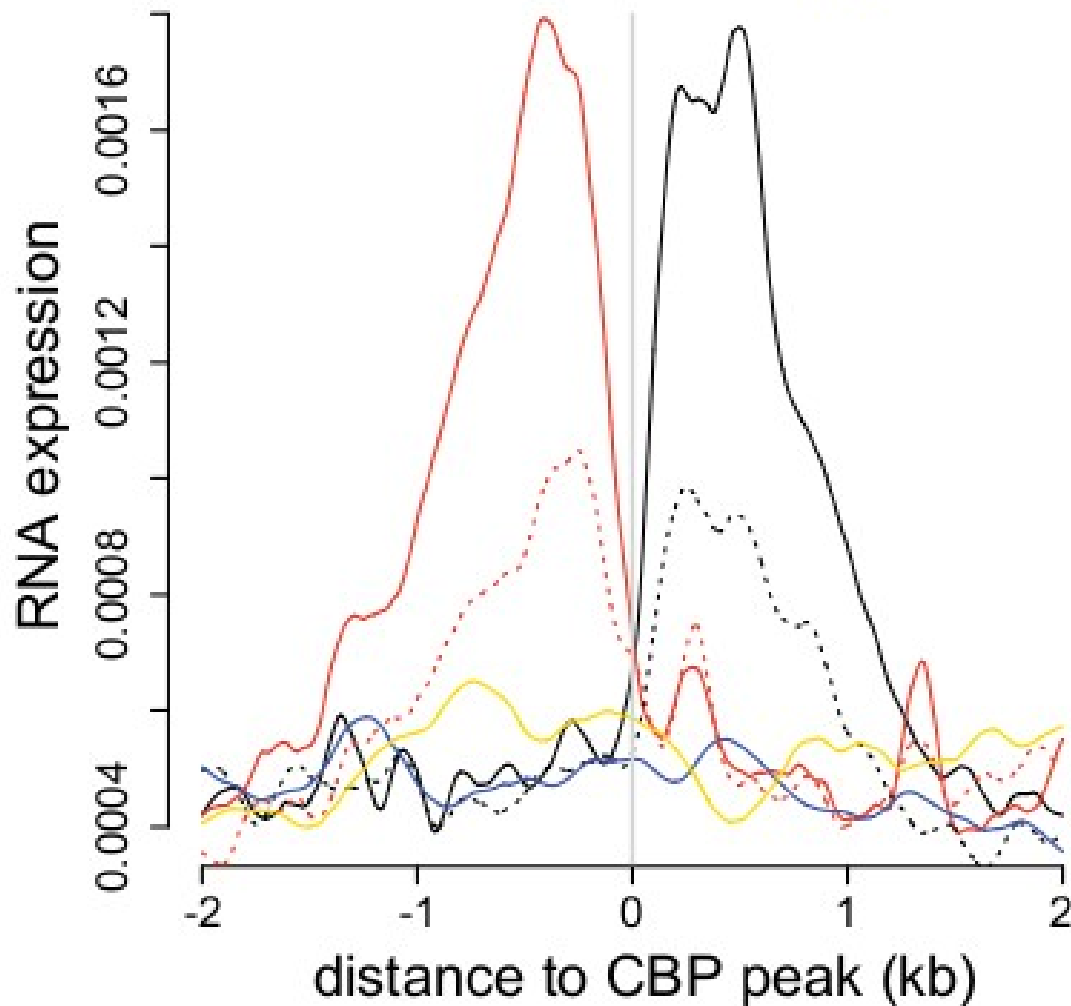
Genome-wide profile of transcription at enhancers



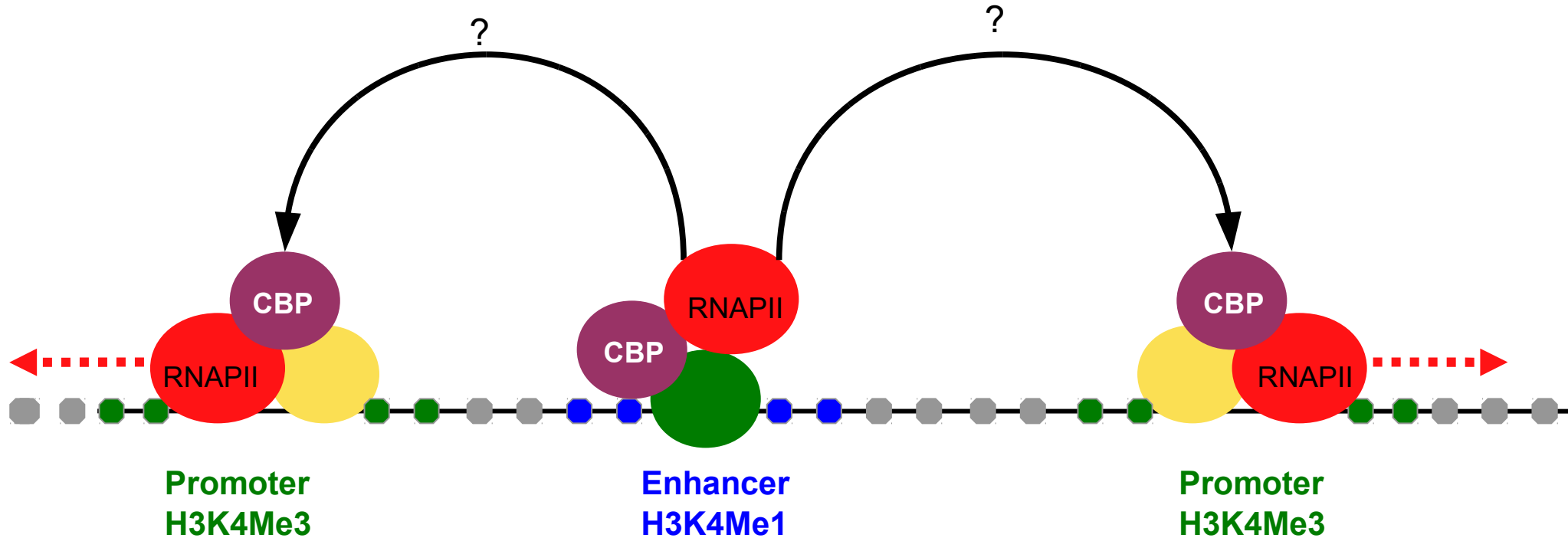
- Low expression
- ~1.5 kb
- Bidirectional
- No polyA
- Not protein-coding

Genome-wide profile of transcription at enhancers

- S total RNA -KCl
- S total RNA +KCl
- AS total RNA -KCl
- AS total RNA +KCl
- S polyA+ RNA +KCl
- AS polyA+ RNA +KCl

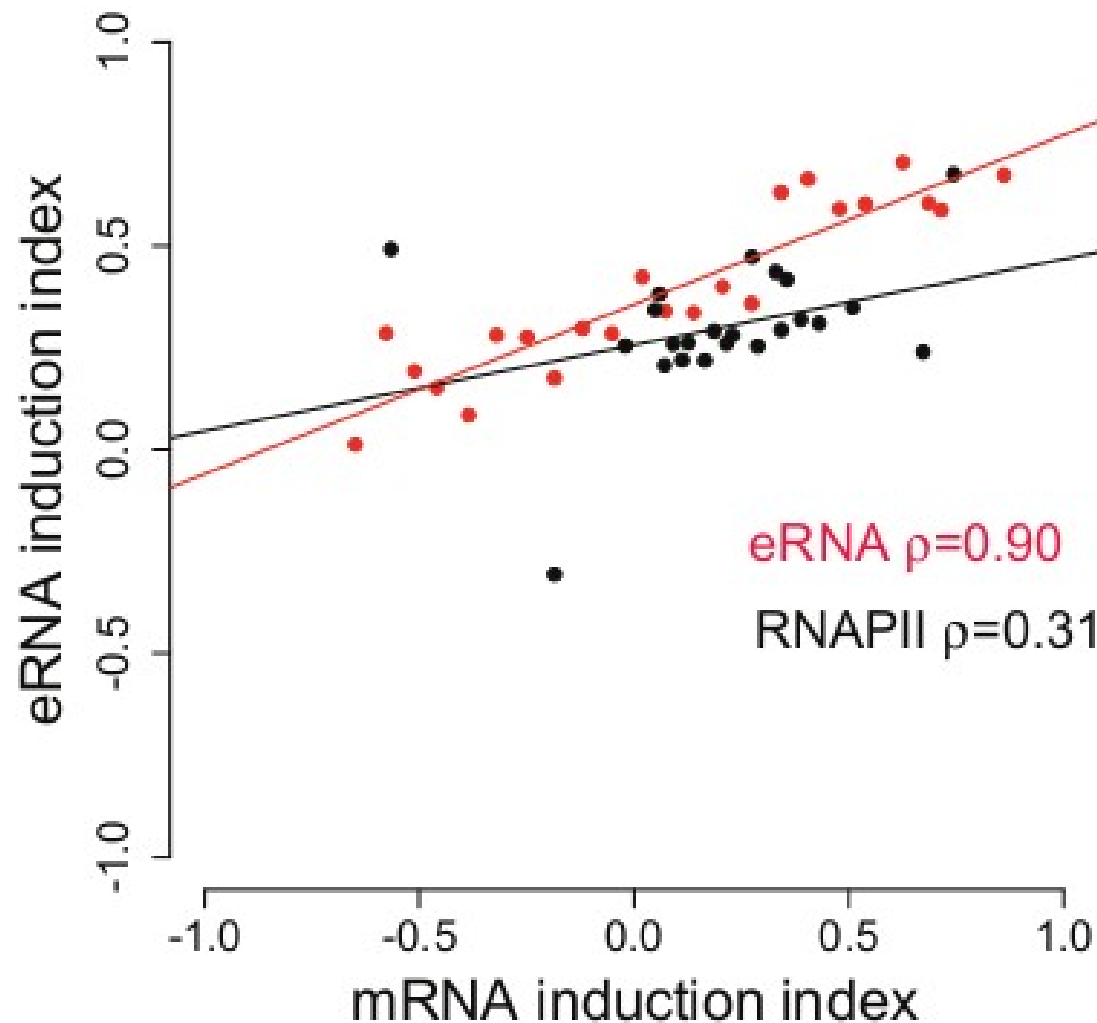


How do eRNA levels relate to mRNA levels?



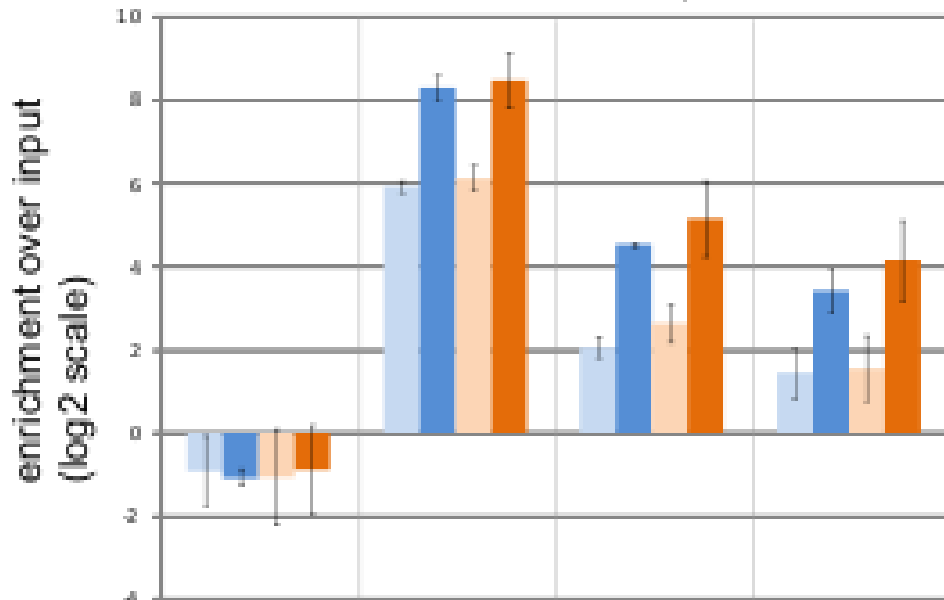
eRNA induction is correlated with induction of nearby mRNAs

$$\text{induction index} = (KCl^+ - KCl^-) / (KCl^+ + KCl^-)$$



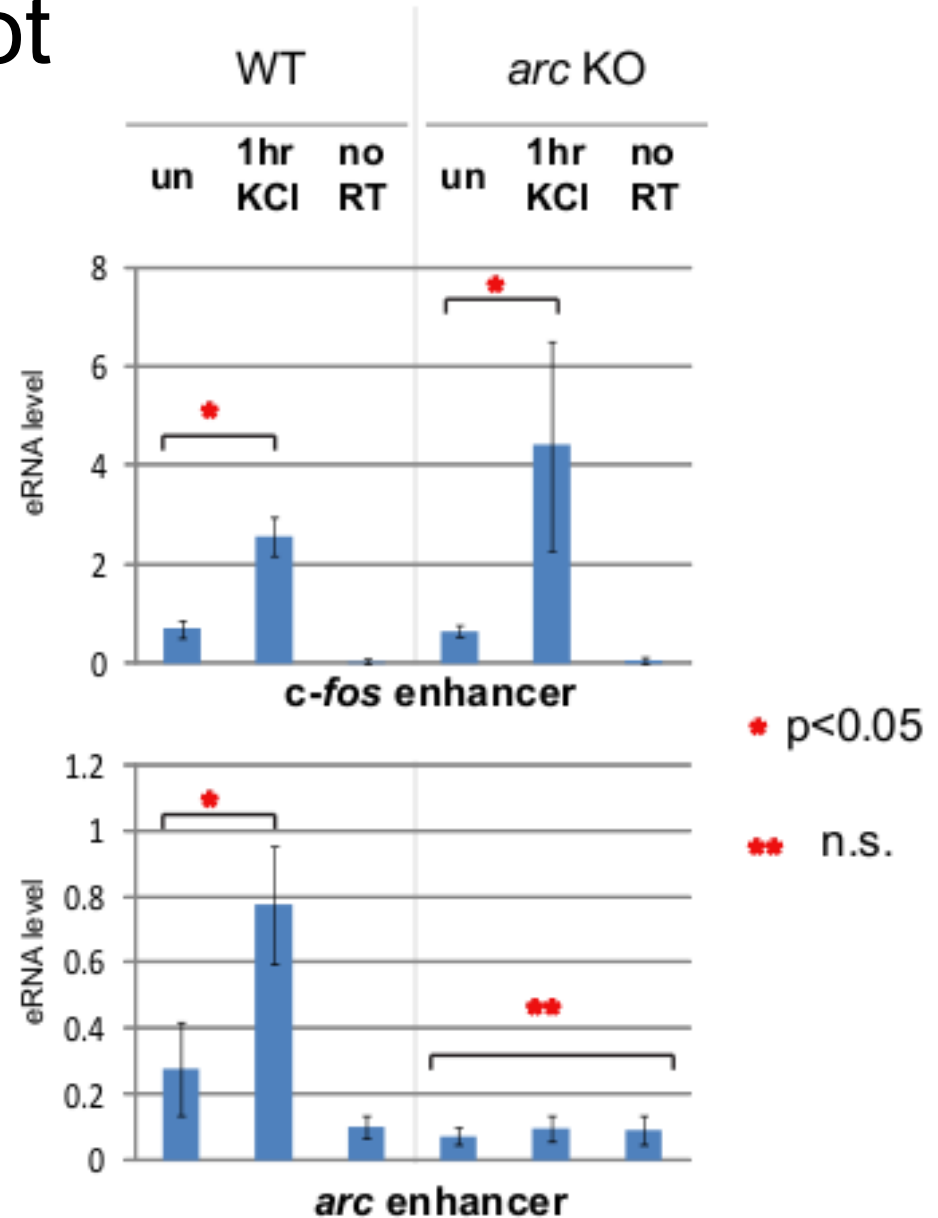
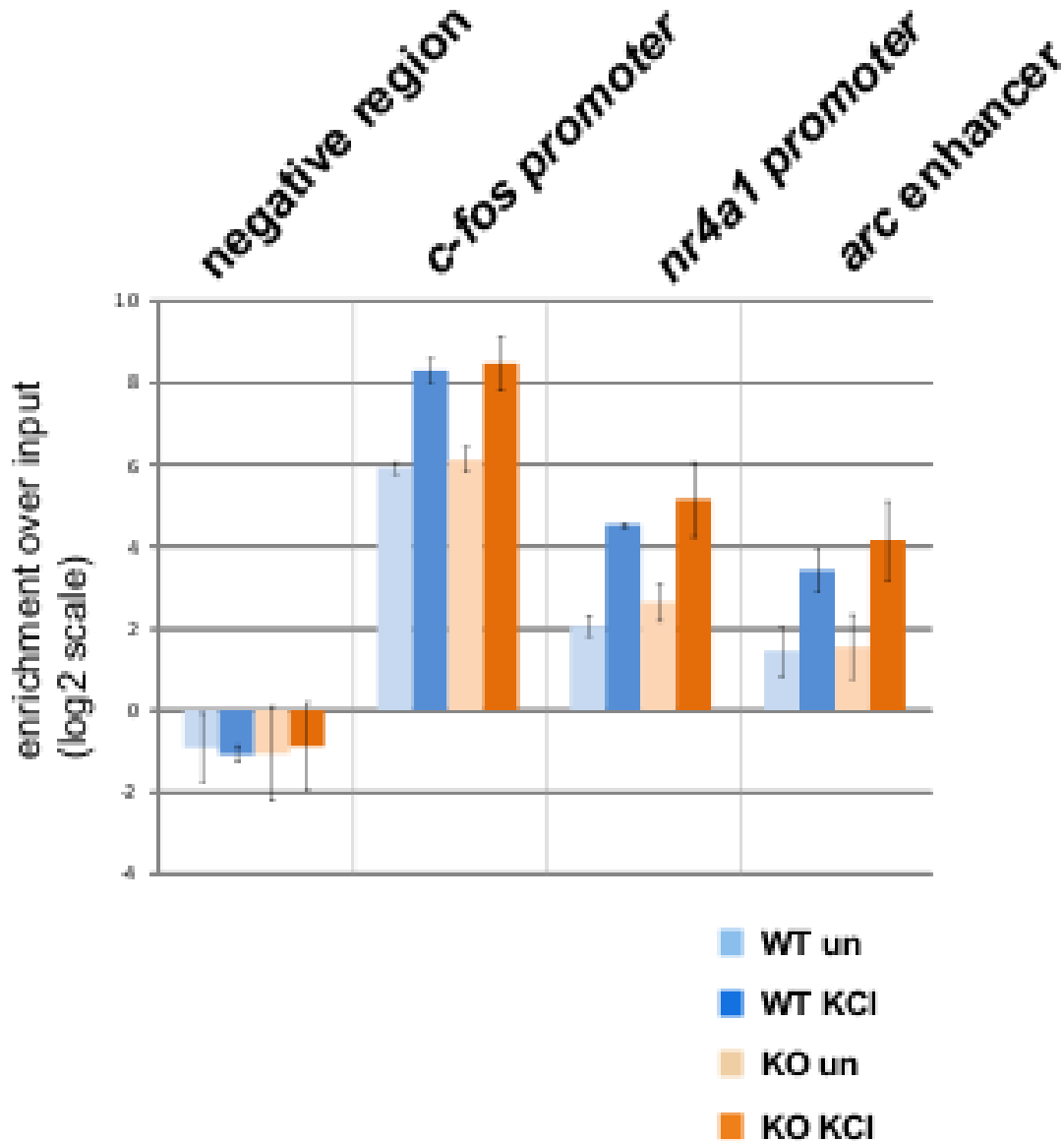
Knock-out experiment confirms that RNAPII recruitment is independent of the promoter

negative region
c-fos promoter
nr4a1 promoter
arc enhancer



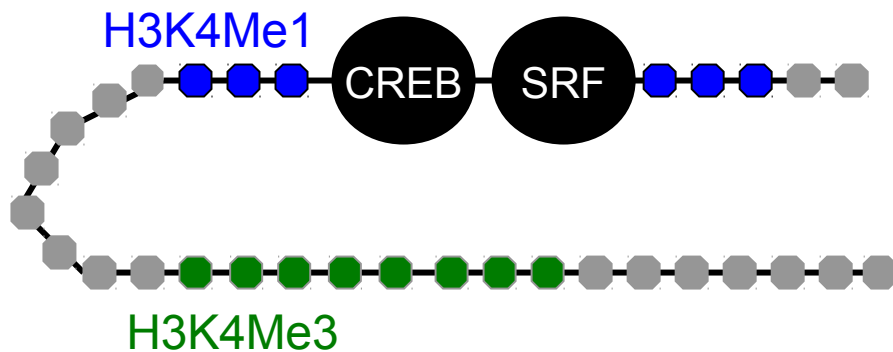
- WT un
- WT KCI
- KO un
- KO KCI

Knock-out experiment confirms that RNAPII recruitment is independent of the promoter but eRNA synthesis is not

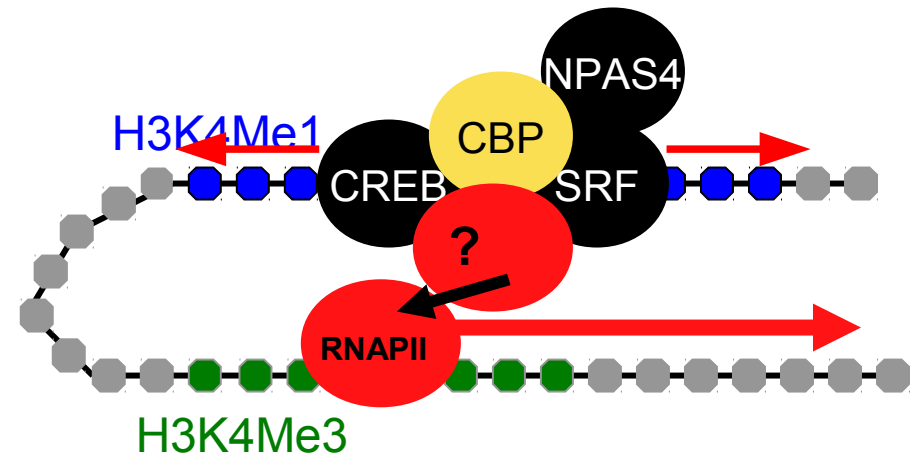


Summary

Before neuronal activation



After neuronal activation



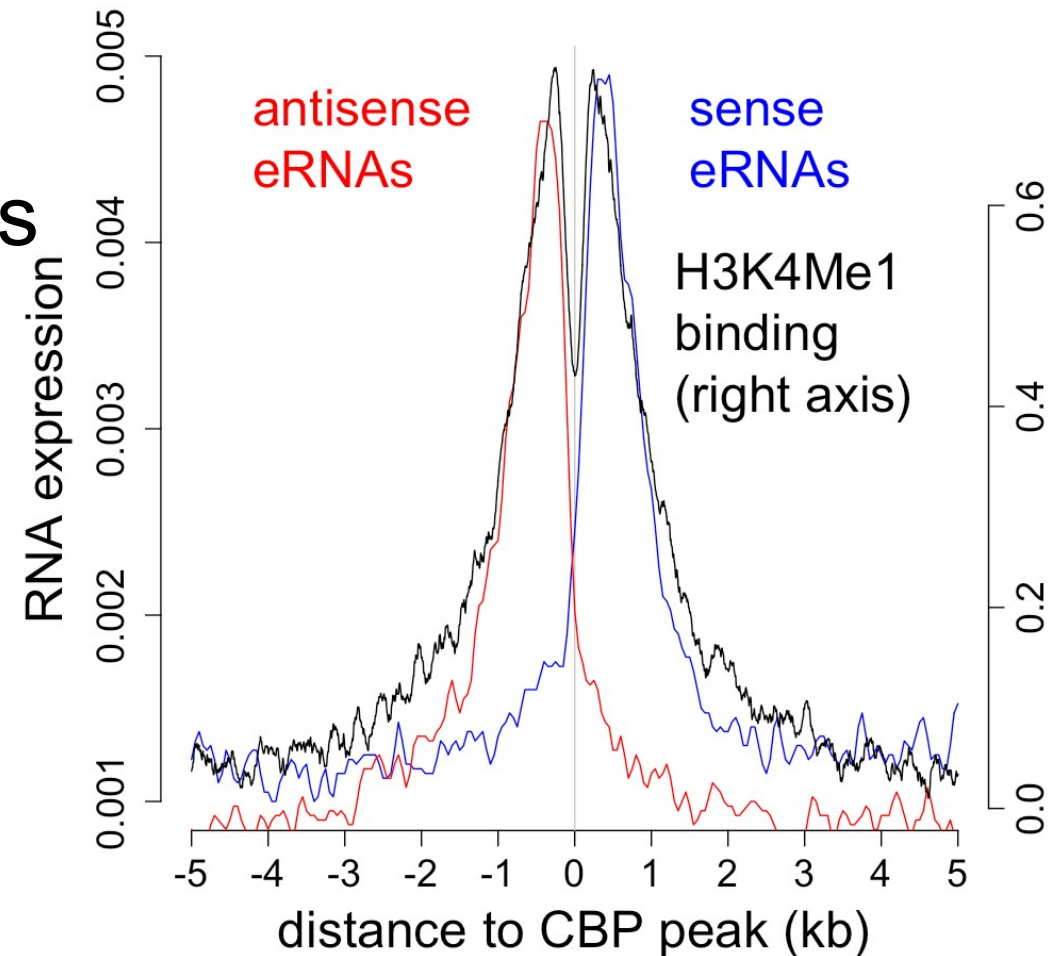
- Does RNAPII bind at enhancers? **YES**
- Are transcripts produced at enhancers? **YES**
- Is RNAPII recruitment independent? **YES**
- Is eRNA production independent? **NO**

We have not yet been able to determine the function of eRNAs

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

-George Bernard Shaw

- Noise
- Establish histone marks
- Transcript has function



eRNAs have been found in other cell types

doi:10.1038/nature09033

nature

ARTICLES

Widespread transcription at neuronal activity-regulated enhancers

Tae-Kyung Kim^{1*†}, Martin Hemberg^{2*}, Jesse M. Gray^{1*}, Allen M. Costa¹, Daniel M. Bear¹, Jing Wu³, David A. Harmin^{1,4}, Mike Laptewicz¹, Kellie Barbara-Haley⁵, Scott Kuersten⁶, Eirene Markenscoff-Papadimitriou^{1†}, Dietmar Kuhl⁷, Haruhiko Bito⁸, Paul F. Worley³, Gabriel Kreiman² & Michael E. Greenberg¹

Histone H3K27ac separates active from poised enhancers and predicts developmental state

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

OPEN ACCESS Freely available online

PLoS BIOLOGY

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

Francesca De Santa^{1,3}, Iros Barozzi^{1,3}, Flore Mietton^{1,3}, Serena Ghisletti¹, Sara Polletti¹, Betsabeh Khoramian Tusi¹, Heiko Muller¹, Jiannis Ragoussis², Chia-Lin Wei³, Gioacchino Natoli^{1*}

LETTER

doi:10.1038/nature09692

A unique chromatin signature uncovers early developmental enhancers in humans

Alvaro Rada-Iglesias¹, Ruchi Bajpai¹, Tomek Swigut¹, Samantha A. Brugmann¹, Ryan A. Flynn¹ & Joanna Wysocka^{1,2}

Summary I: Identified ~12,000 activity regulated enhancers that are transcribed

- Histone modification and CBP binding
- Transcription induction correlated with nearby promoter

What is the function of conserved non-coding sequences?

Evolution at Two Levels in Humans and Chimpanzees

Their macromolecules are so alike that regulatory mutations may account for their biological differences.

Mary-Claire King and A. C. Wilson

What is the function of conserved non-coding sequences?

Evolution at Two Levels in Humans and Chimpanzees

Their macromolecules are so alike that regulatory mutations may account for their biological differences.

Mary-Claire King and A. C. Wilson

Large-Scale Transcriptional Activity in Chromosomes 21 and 22

Philipp Kapranov,¹ Simon E. Cawley,¹ Jorg Drenkow,¹ Stefan Bekiranov,¹ Robert L. Strausberg,² Stephen P. A. Fodor,¹ Thomas R. Gingeras^{1*}

[OPEN ACCESS](#) Freely available online

PLoS BIOLOGY

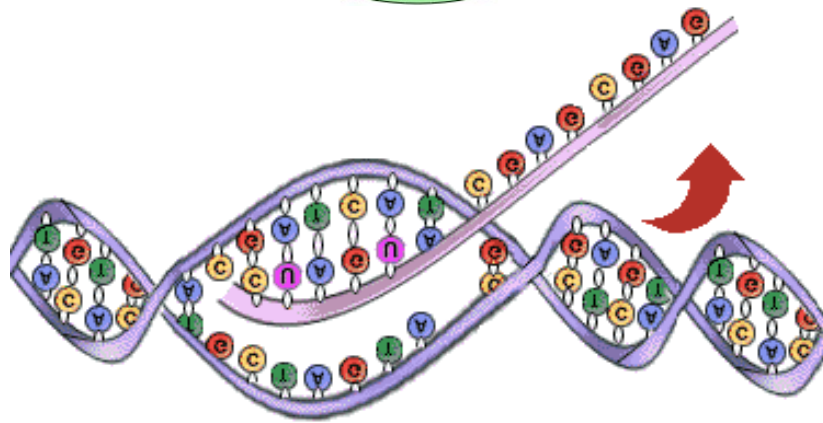
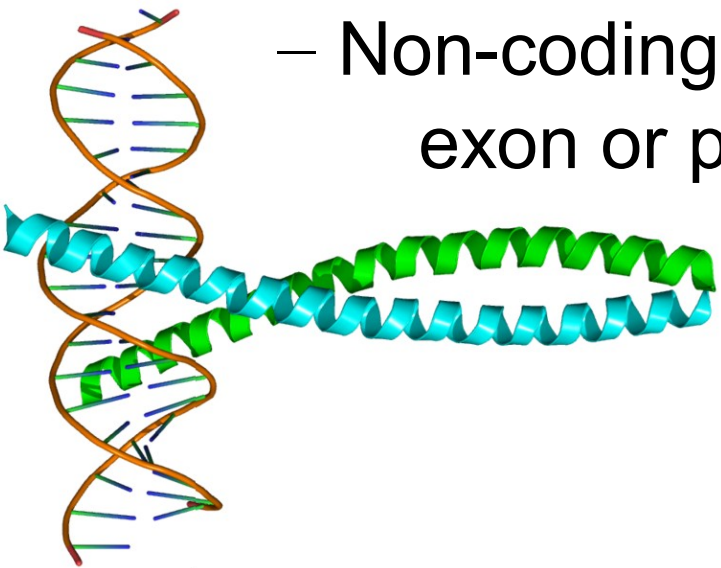
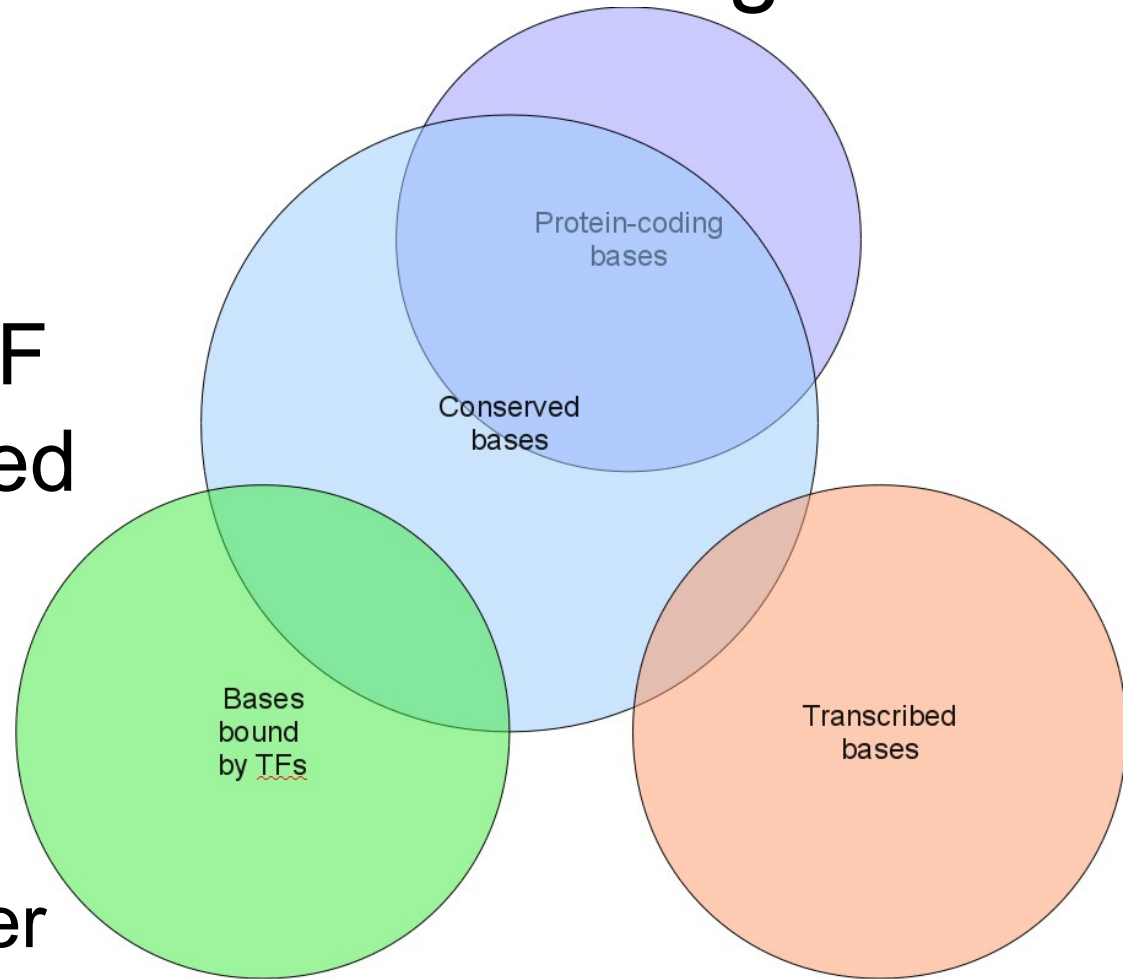
Most “Dark Matter” Transcripts Are Associated With Known Genes

Harm van Bakel¹, Corey Nislow^{1,2}, Benjamin J. Blencowe^{1,2}, Timothy R. Hughes^{1,2*}

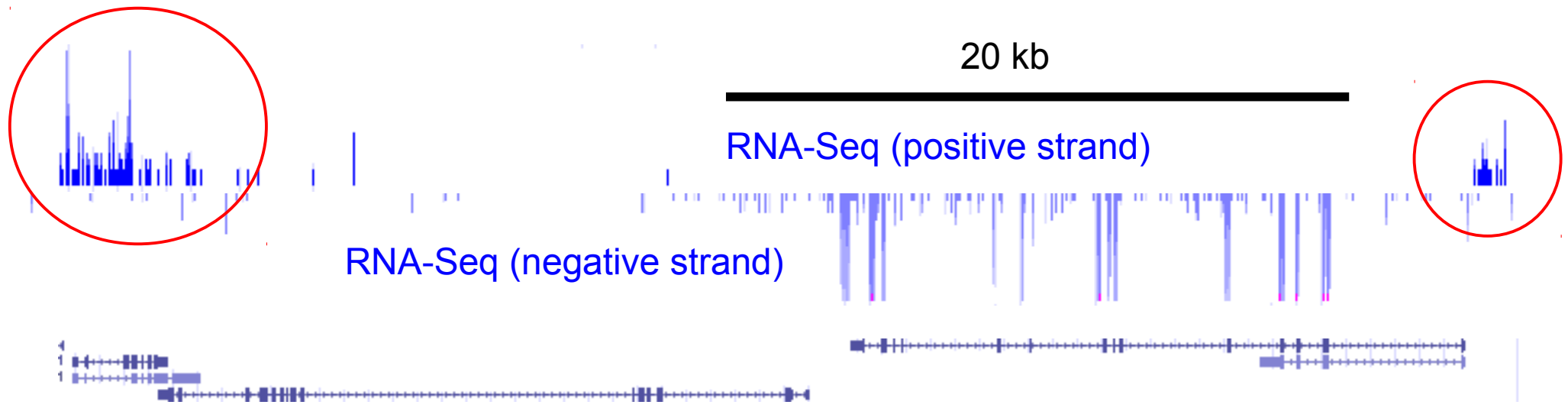
¹ Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, ² Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

What drives the conservation of extragenic regions?

- Compare extragenic transcription and TF binding to conserved bases
 - TF binding sites
 - Non-coding RNA exon or promoter

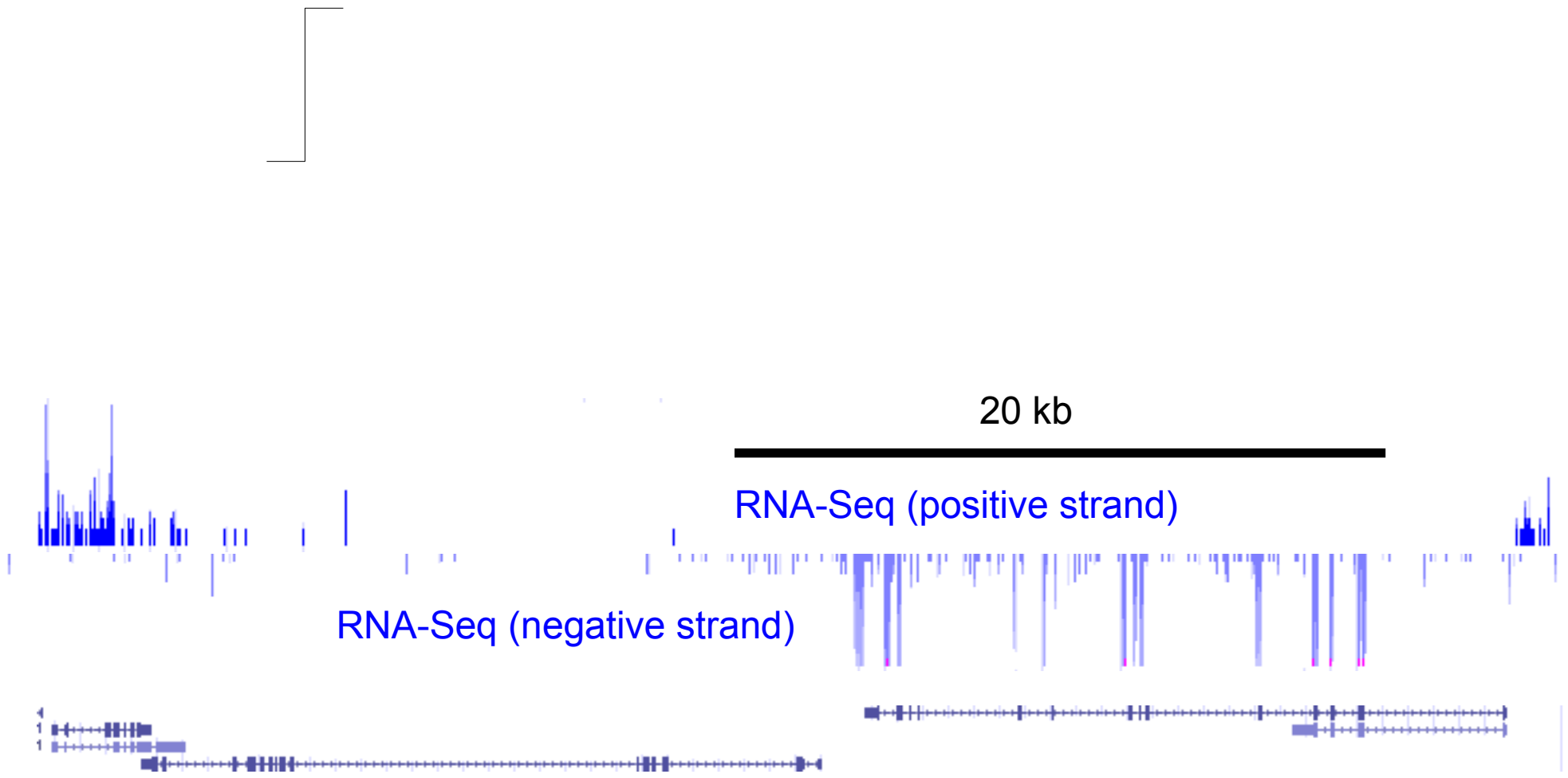


De novo identification of transcribed regions



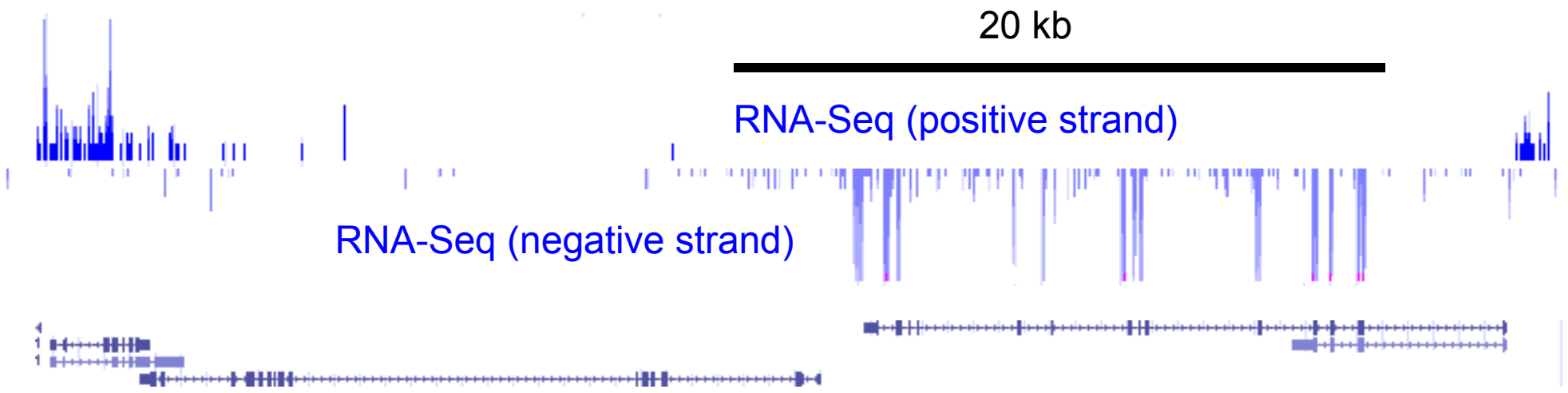
Using Haar-wavelets to identify transcribed regions (HaTriC)

- Find where read-density changes abruptly



Using Haar-wavelets to identify transcribed regions (HaTriC)

- Find where read-density changes abruptly
 - Consider multiple length scales

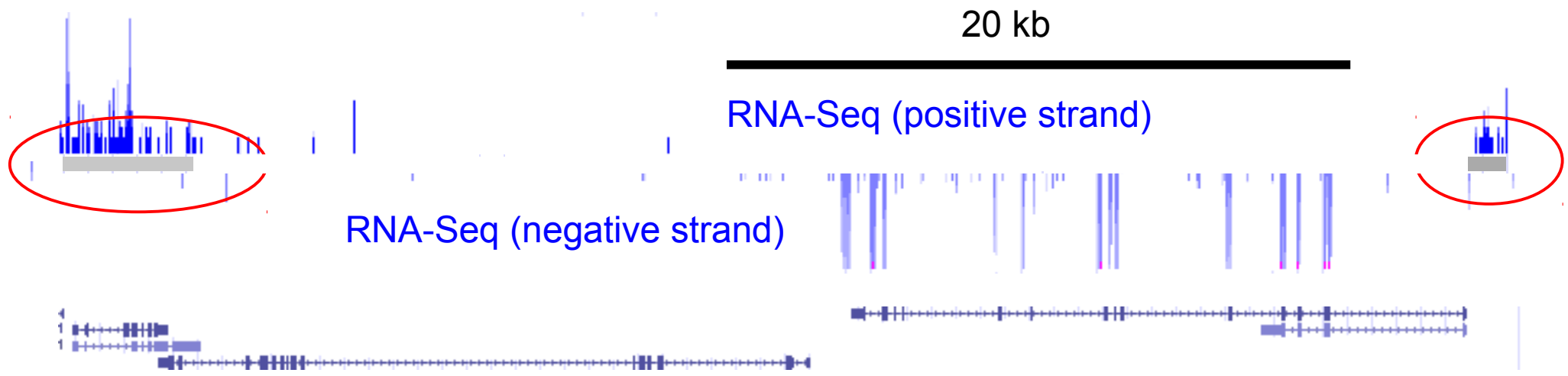


Using Haar-wavelets to identify transcribed regions (HaTriC)

- Find where read-density changes abruptly
 - Consider multiple length scales

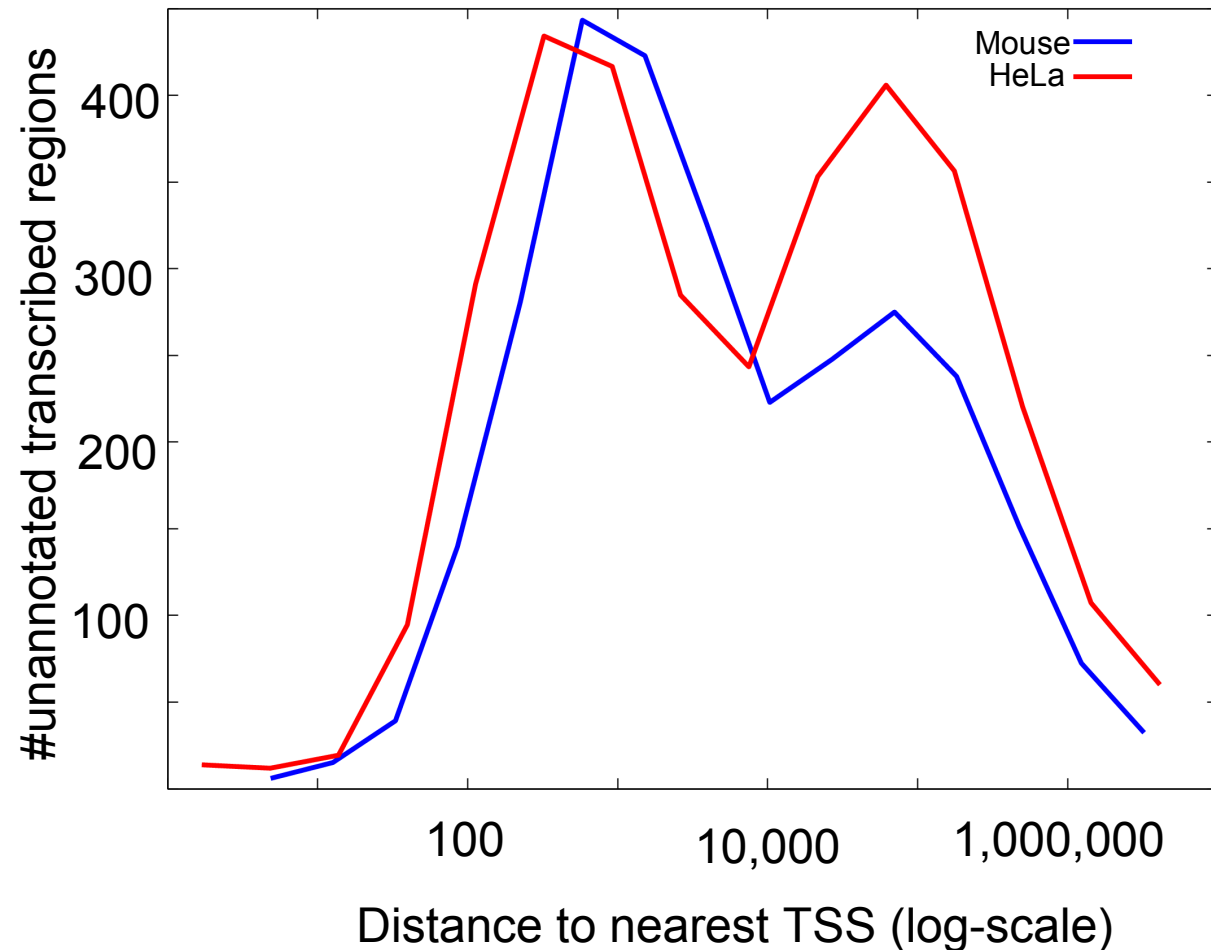


- Interleaving regions of high/low density
- Mask expressed regions and repeat



Most unannotated transcribed regions are found near annotated genes

- 80% of active genes correctly identified



Most reads are found in annotated genes

- Genic 88.66%
- Annotated ncRNA 3.87%

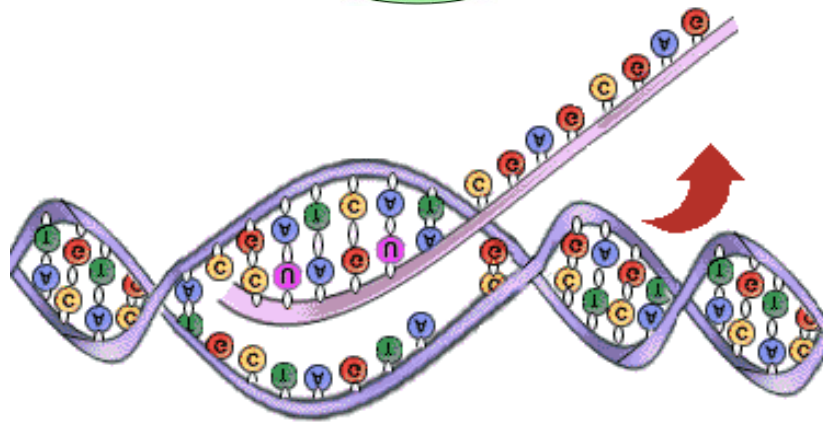
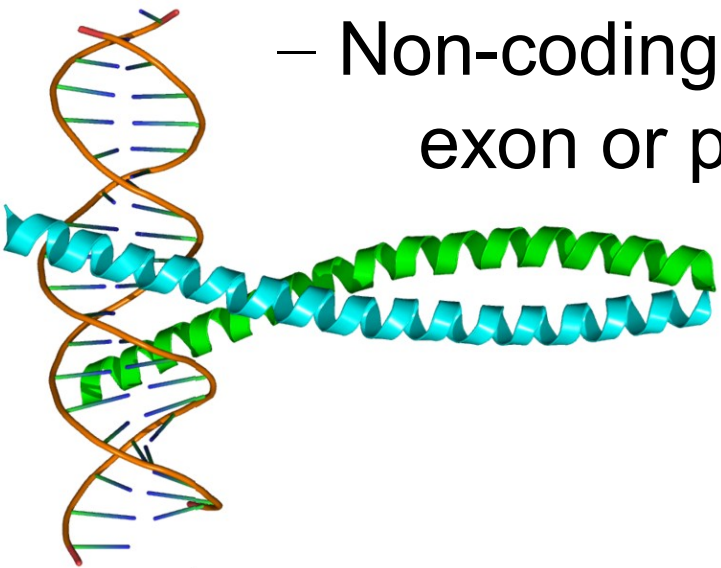
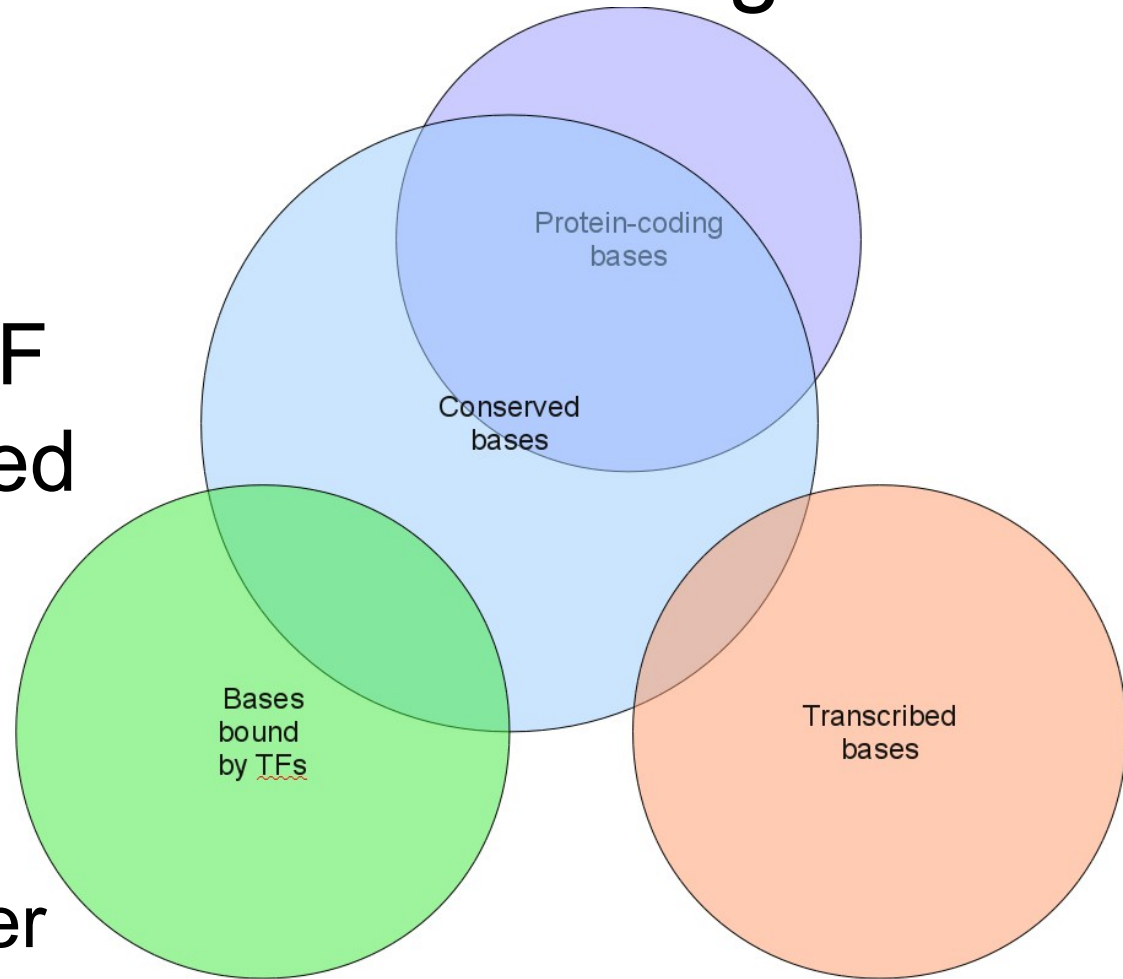
Transcribed regions account for 99.54% of all reads

- Genic 88.66%
- Annotated ncRNA 3.87%
- Upstream anti-sense 0.51%
- Anti-sense 0.23%
- Novel 5.28%
- eRNAs 0.04%
- Other regulatory factor binding sites 0.96%

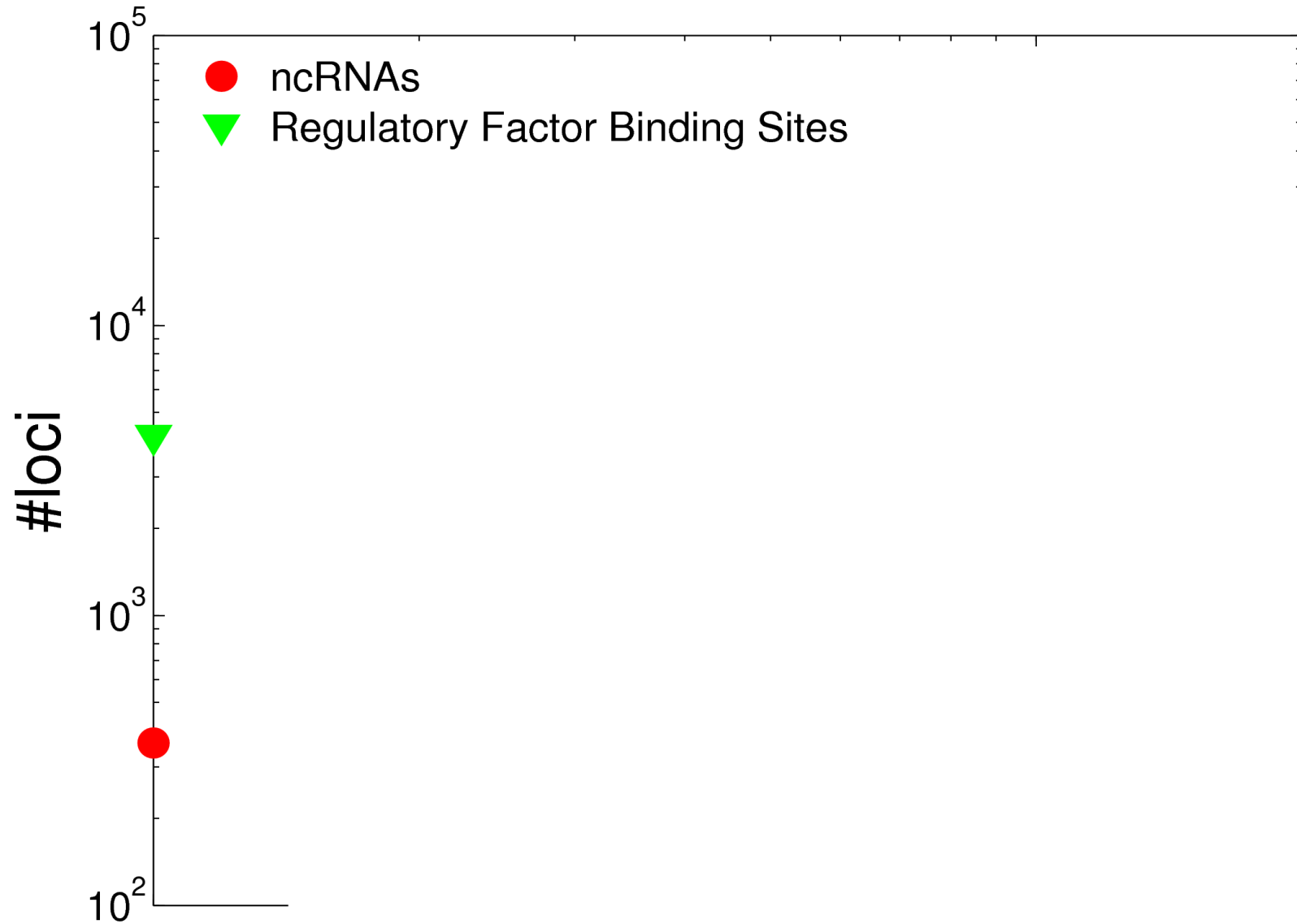
There are many extragenic regions transcribed at very low levels

What drives the conservation of extragenic regions?

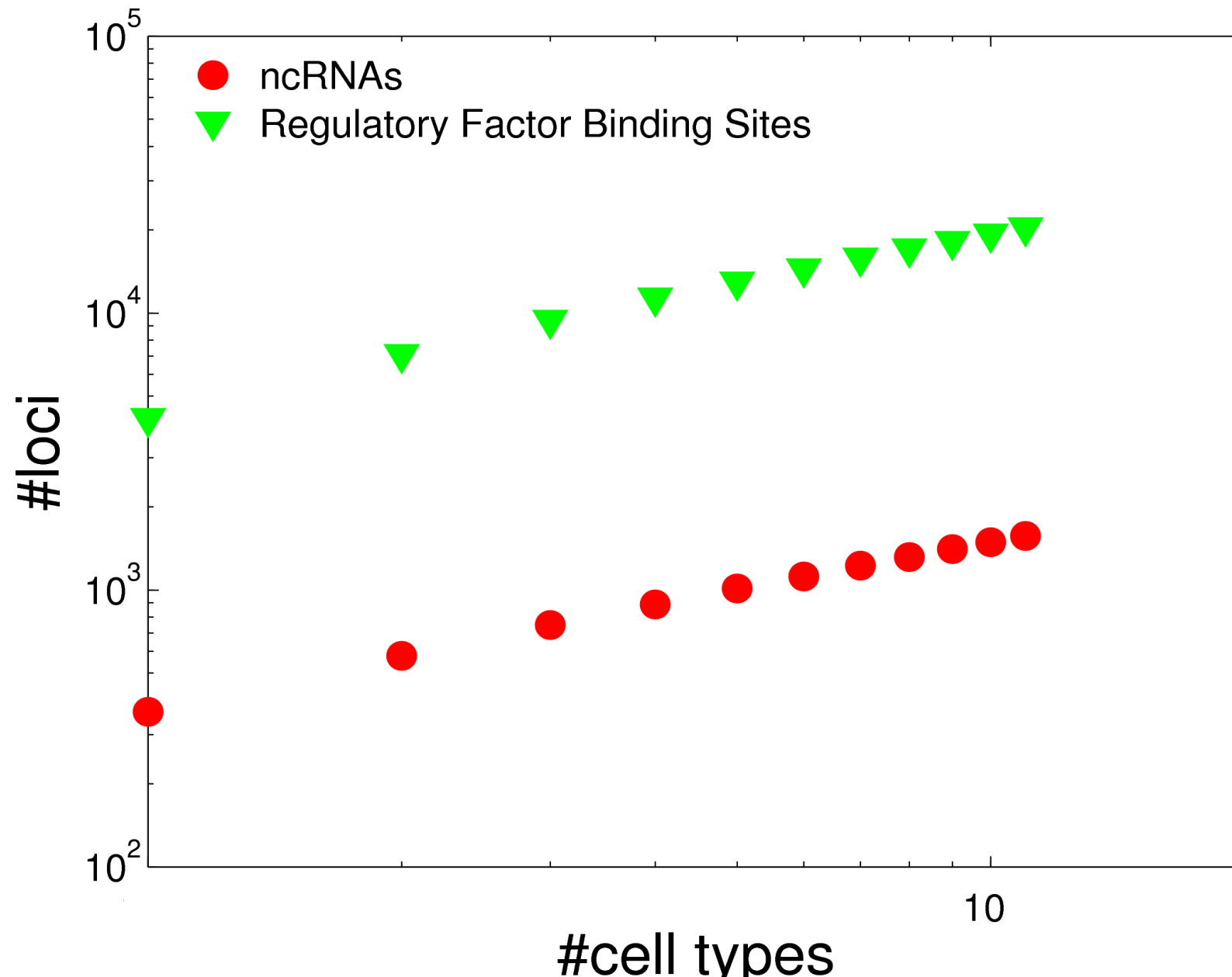
- Compare extragenic transcription and TF binding to conserved bases
 - TF binding sites
 - Non-coding RNA exon or promoter



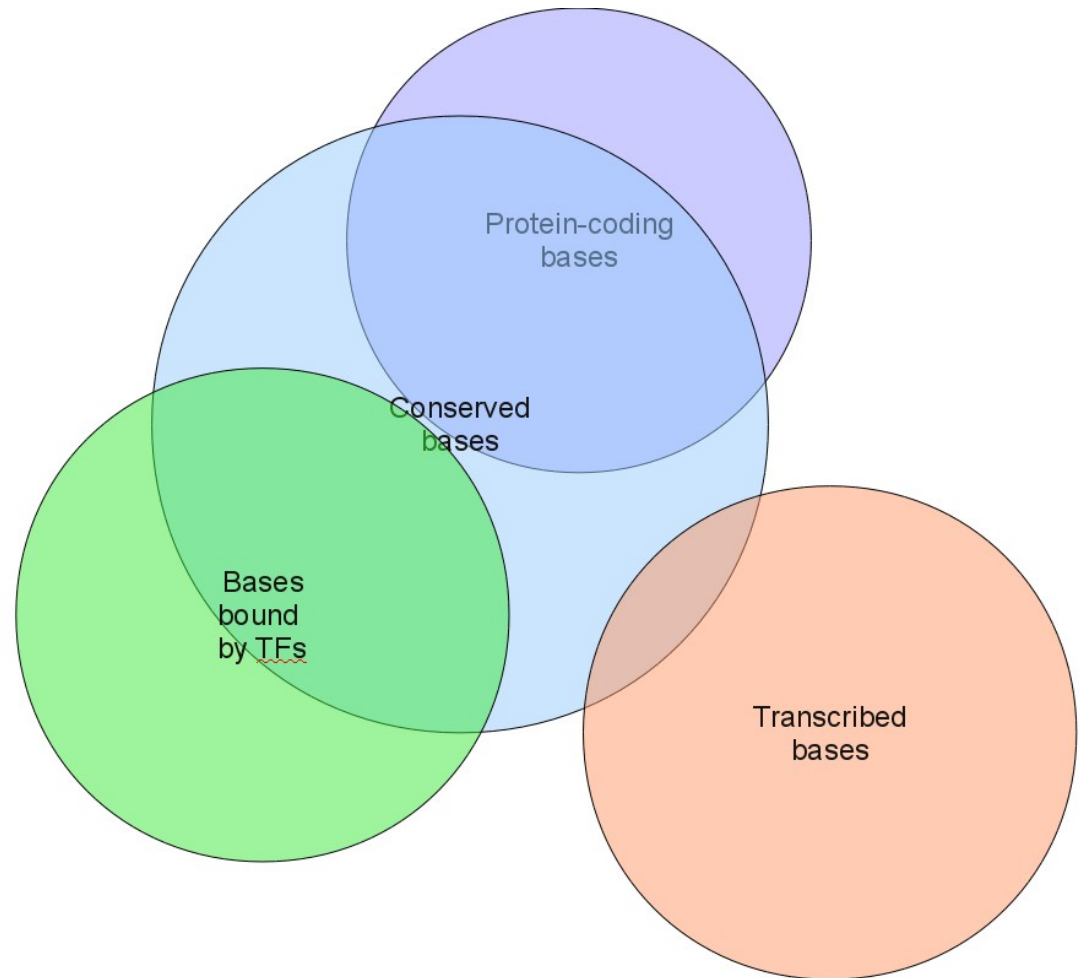
About 80% of conserved bases are transcription factor binding sites



About 80% of conserved bases are transcription factor binding sites

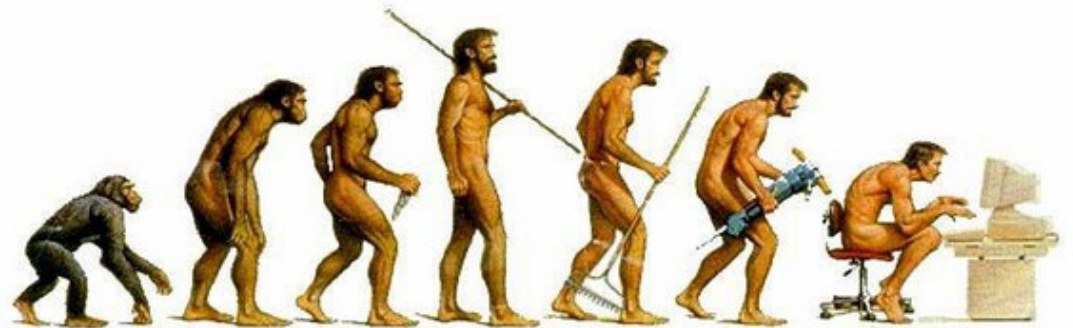
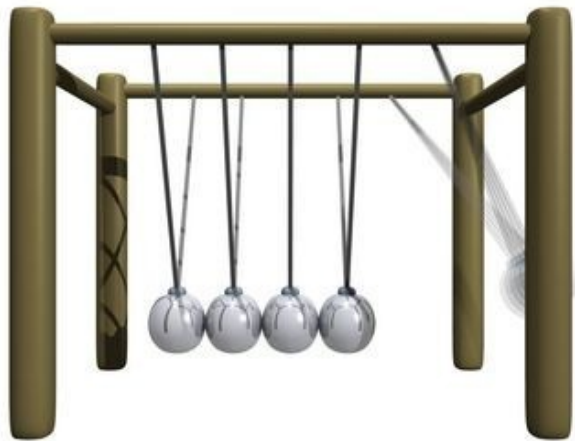


Summary II: *De novo* identification of transcribed regions suggests that most conservation is due to TF binding



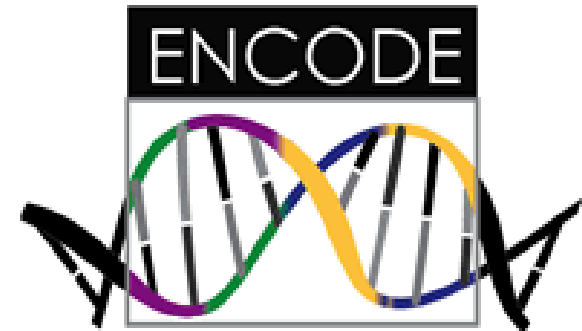
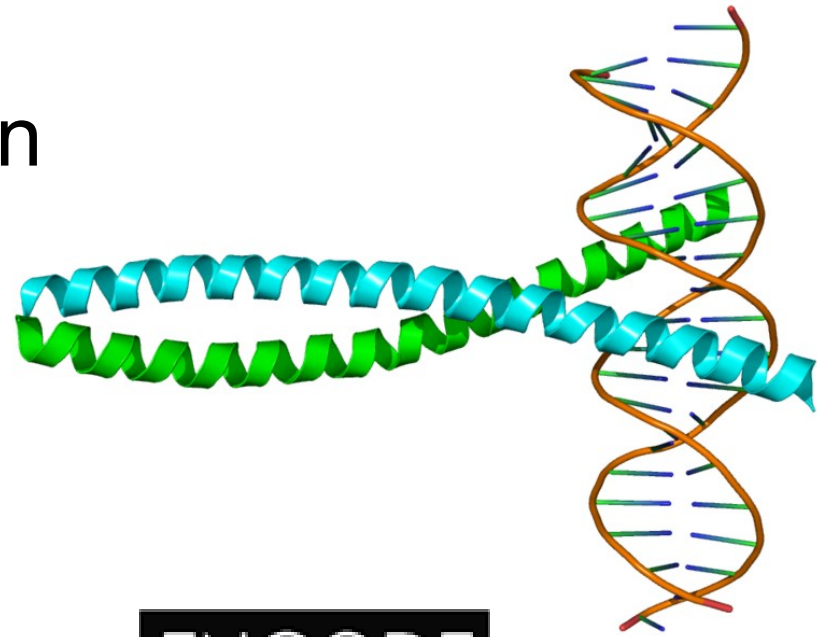
Future Work: Organizing principles of the genome

- Systems biology approach to develop biophysical models



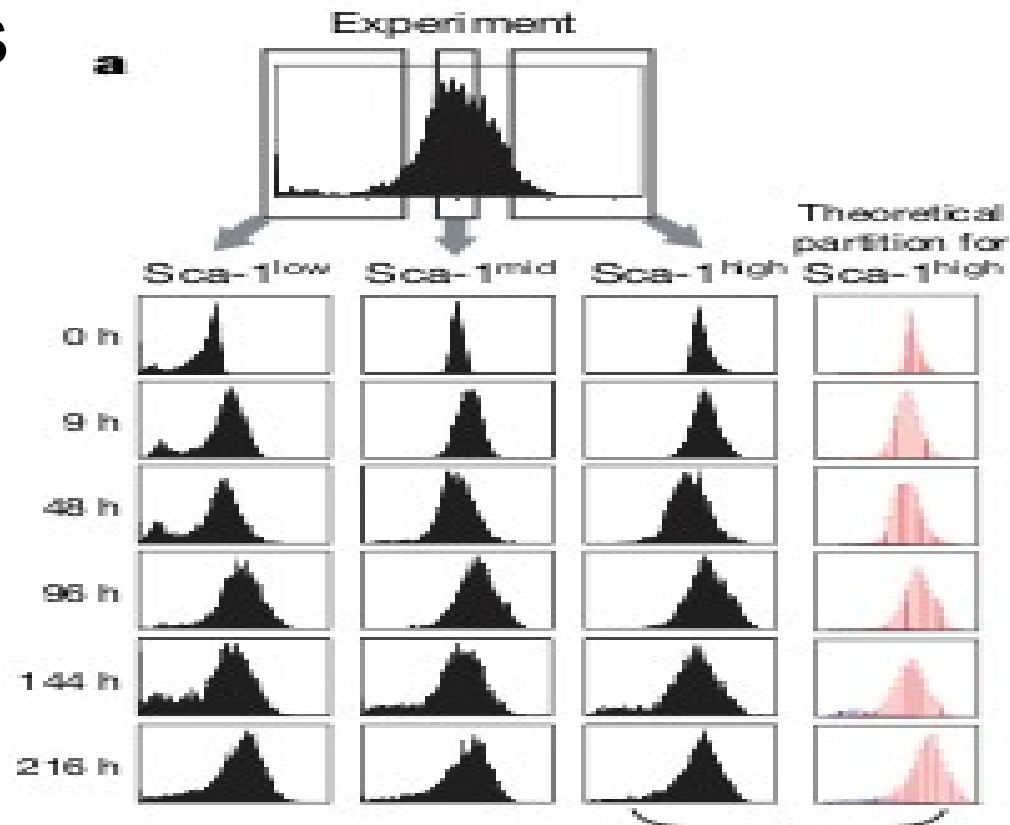
What determines the level of 'epigenomic modifications' and how are they read out?

- How can histone modifications be read and written?
- What determines transcription factor binding?
- What determines the level of transcription?



What is the impact on the phenotype from gene expression noise?

- RNA-Seq for single cells
- Global view of noise in gene expression
 - Pathways
 - Proximity
 - Cell-types
 - Propagation

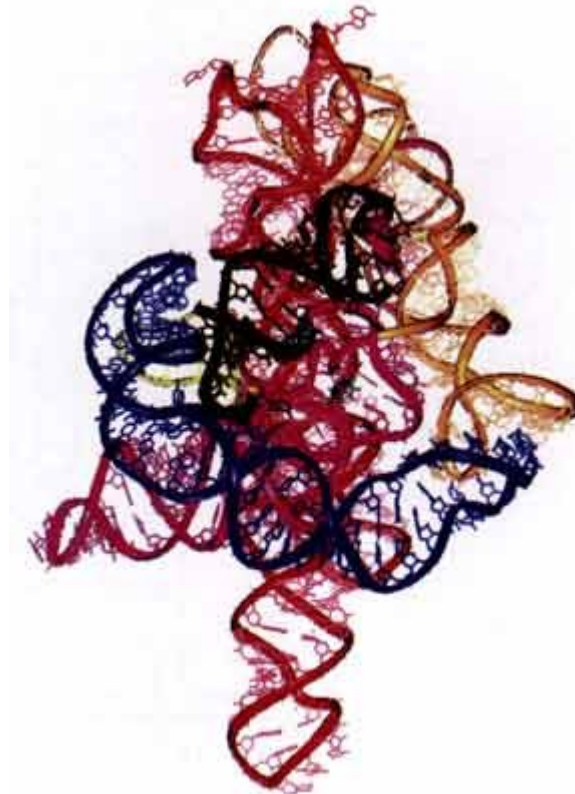


Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis

Fuchou Tang,^{1,3} Catalin Barbacioru,² Siqin Bao,¹ Caroline Lee,¹ Ellen Nordman,² Xiaohui Wang,² Kaiqin Lao,^{2,*} and M. Azim Surani^{1,*}

Is there a non-coding genetic code for determining the structure of RNAs?

.....ACGUCCAAAUUCCCUAGGCUCAAGGCAUUCGAUCGGGAUUUAUA.....



Acknowledgements

Gabriel Kreiman, **Children's Hospital Boston**

Wui Ip

Enrique Tobis

Michael Greenberg, **Harvard Medical School**

Tae-Kyung Kim

Jesse Gray

Allen Costa

Daniel Bear

David Harmin

Mike Laptewicz

Eirene Markenscoff-Papadimitriou

**Molecular Genetics Core
Children's Hospital Boston**

Kellie Haley

Josh Davis

Hal Schneider

Life Technologies

Rob David

Jingwei Ni

Scott Kuersten

Gina Costa

Kevin McKernan

**Harvard Medical School
Biopolymer facility**

Kristin Waraska

Robert Steen

Johns Hopkins

Jing Wu, Paul Worley Lab

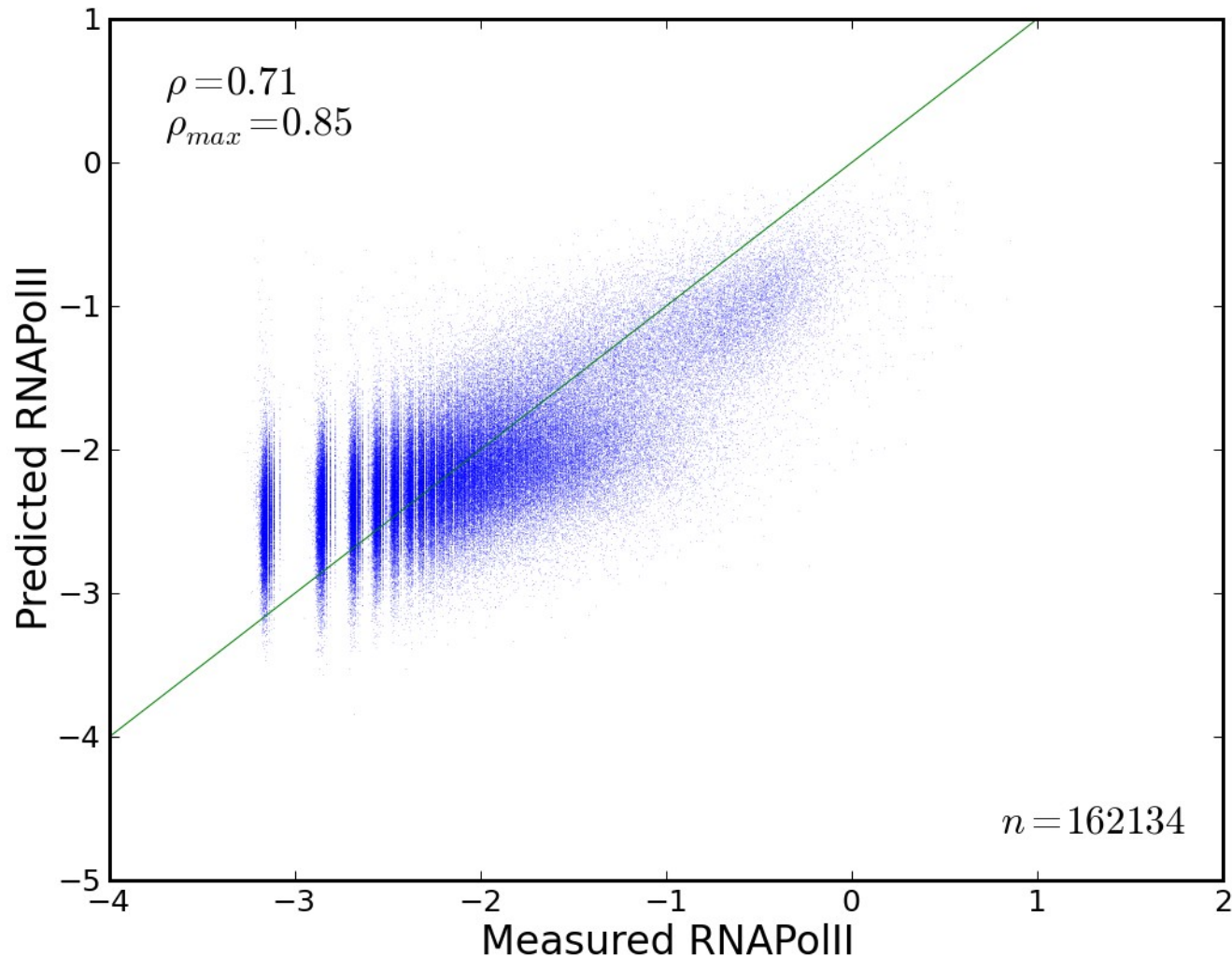


Thank You

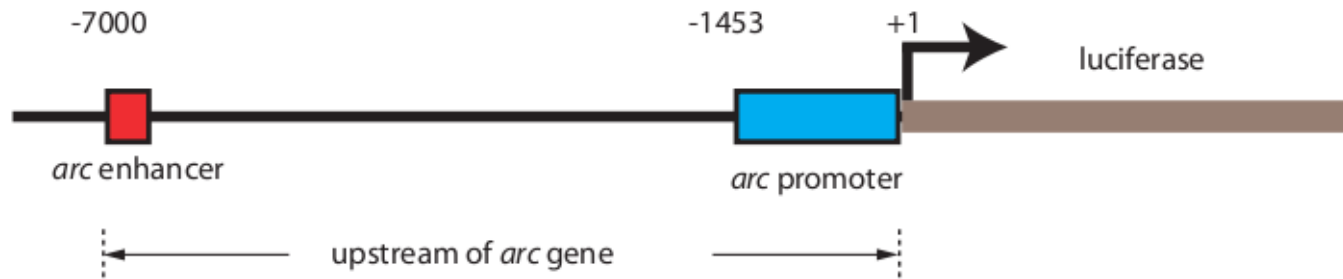


?

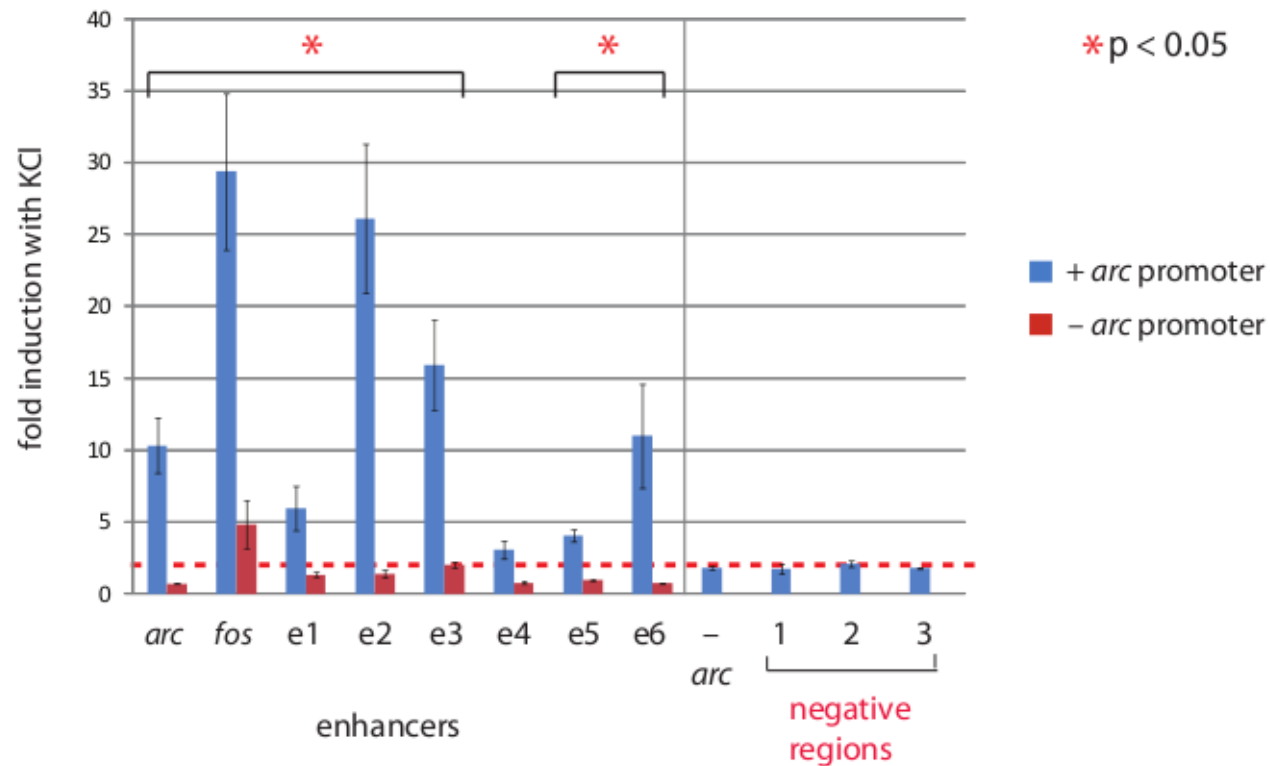
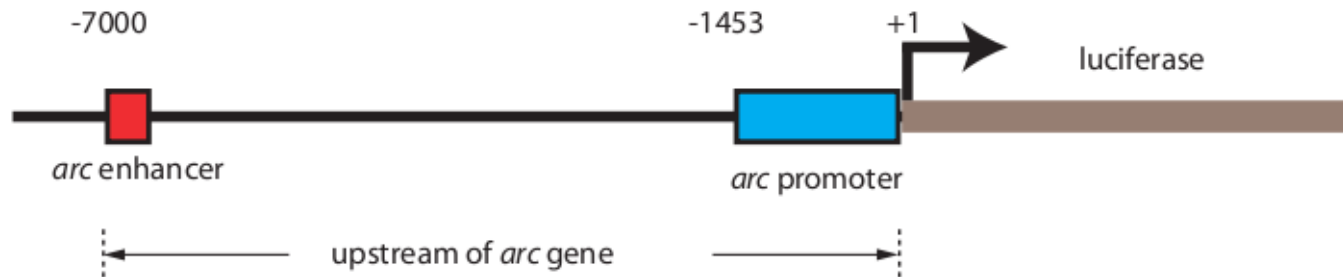
Is there an epigenetic code to determine the cell-type specific function of the sequence?



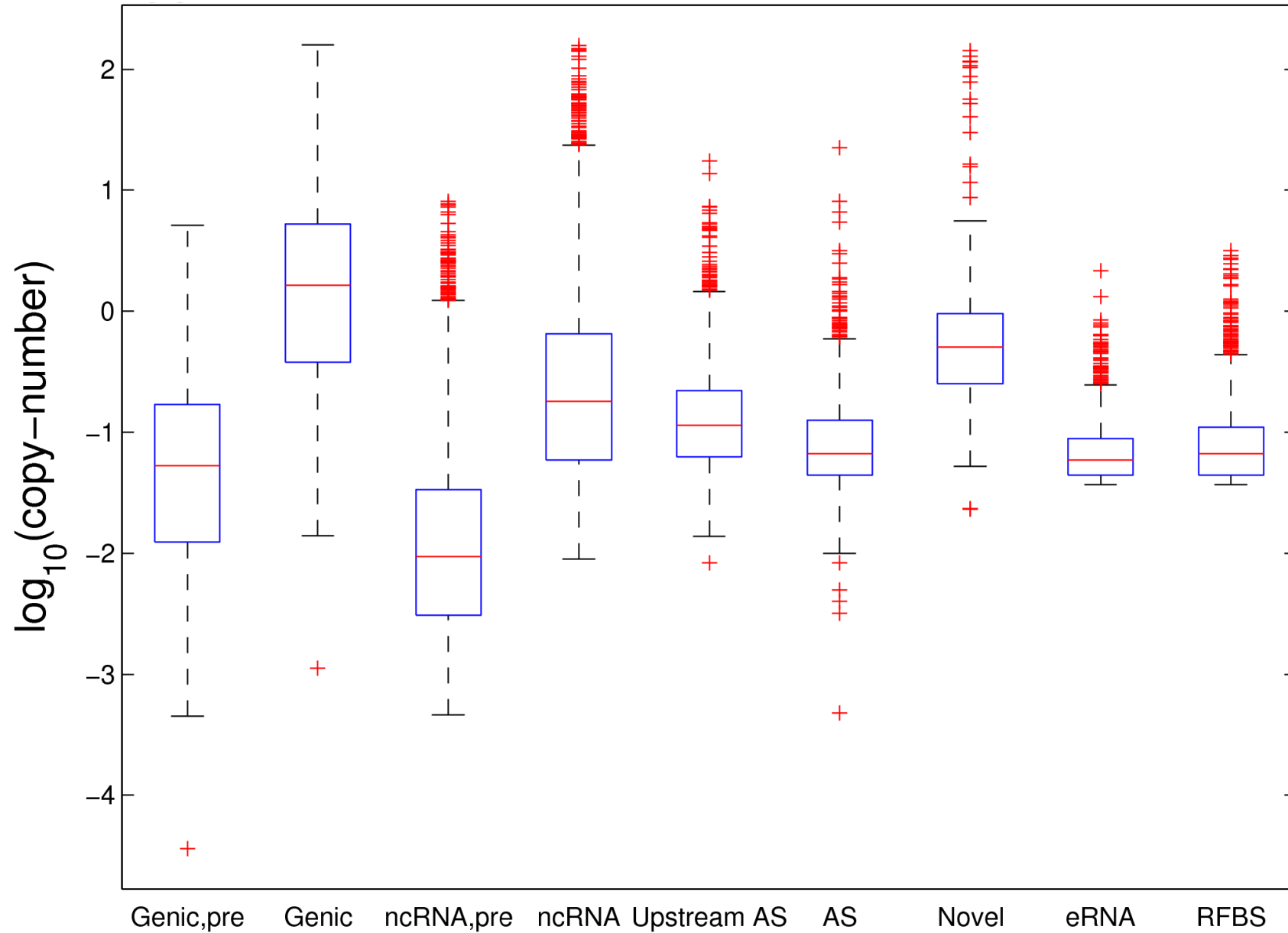
Experimental validation of 8 enhancers using a luciferase assay



Experimental validation of 8 enhancers using a luciferase assay



Copy numbers for different categories



Non-coding sequences are important

- Conserved
 - ~60% of highly conserved are non-coding
- Regulatory
 - Enhancers: distal regulatory binding sites
 - ~1/2 million estimated for human and mouse
 - Insulators
 - Silencers

Vol 444 | 23 November 2006 | doi:10.1038/nature05295

nature

LETTERS

***In vivo* enhancer analysis of human conserved non-coding sequences**

Len A. Pennacchio^{1,2}, Nadav Ahituv², Alan M. Moses², Shyam Prabhakar², Marcelo A. Nobrega^{2†}, Malak Shoukry², Simon Minovitsky², Inna Dubchak^{1,2}, Amy Holt², Keith D. Lewis², Ingrid Plajzer-Frick², Jennifer Akiyama², Sarah De Val⁴, Veena Afzal², Brian L. Black⁴, Olivier Couronne^{1,2}, Michael B. Eisen^{2,3}, Axel Visel² & Edward M. Rubin^{1,2}

Non-coding sequences are important

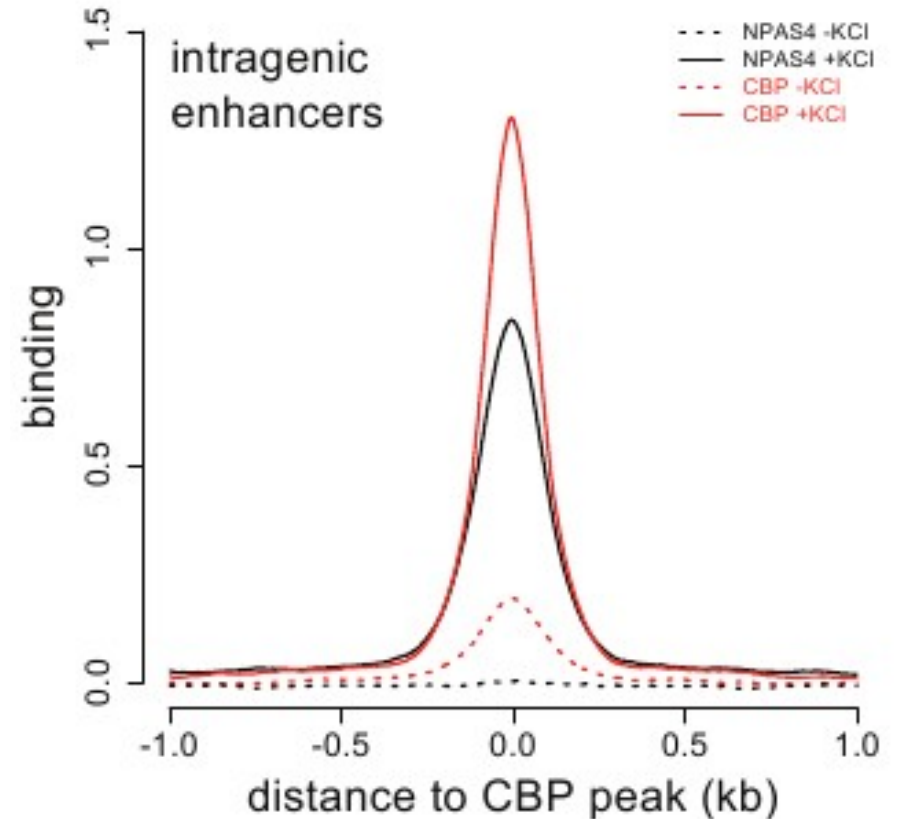
- Conserved
 - ~60% of highly conserved are non-coding
- Regulatory
 - Enhancers: distal regulatory binding sites
- Transcribed
 - Long non-coding RNAs (lncRNAs)
 - Most have unknown function

Functional Demarcation of Active and Silent Chromatin Domains in Human *HOX* Loci by Noncoding RNAs

John L. Rinn,¹ Michael Kertesz,^{3,5} Jordon K. Wang,^{1,5} Sharon L. Squazzo,⁴ Xiao Xu,¹ Samantha A. Brugmann,² L. Henry Goodnough,² Jill A. Helms,² Peggy J. Farnham,⁴ Eran Segal,^{3,*} and Howard Y. Chang^{1,*}

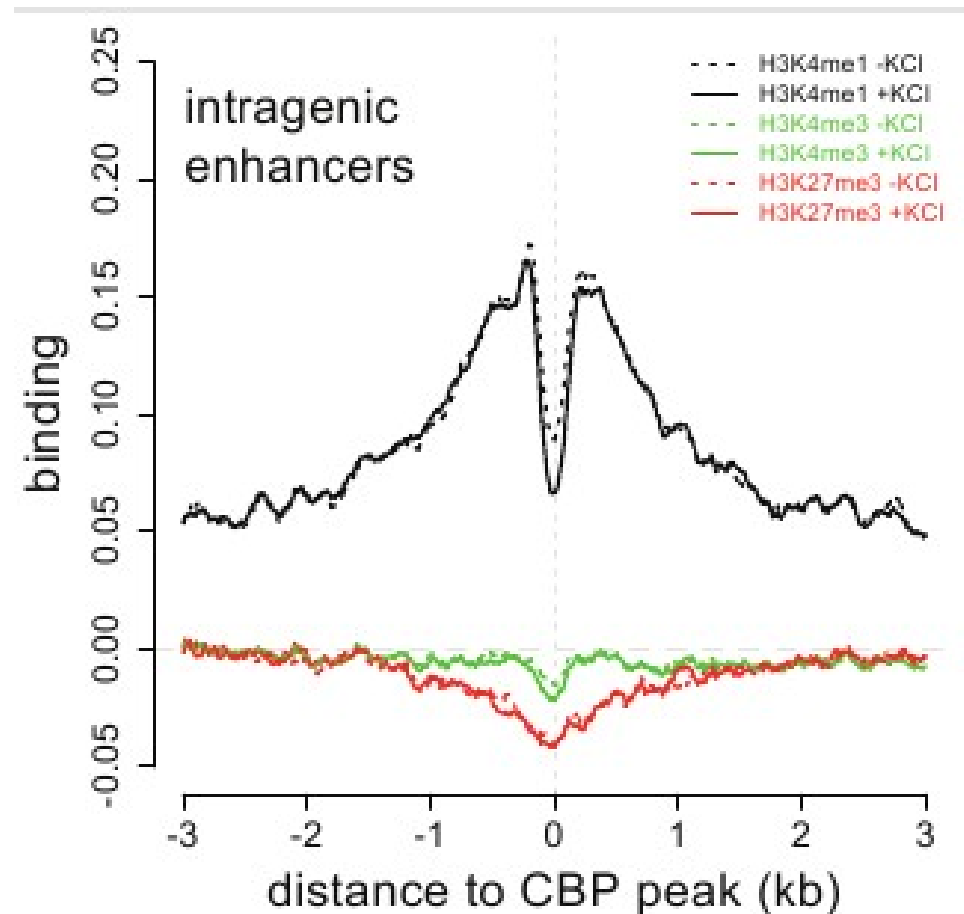
Intragenic enhancers

- ~7,000 enhancers overlapping introns
 - Inducible CBP



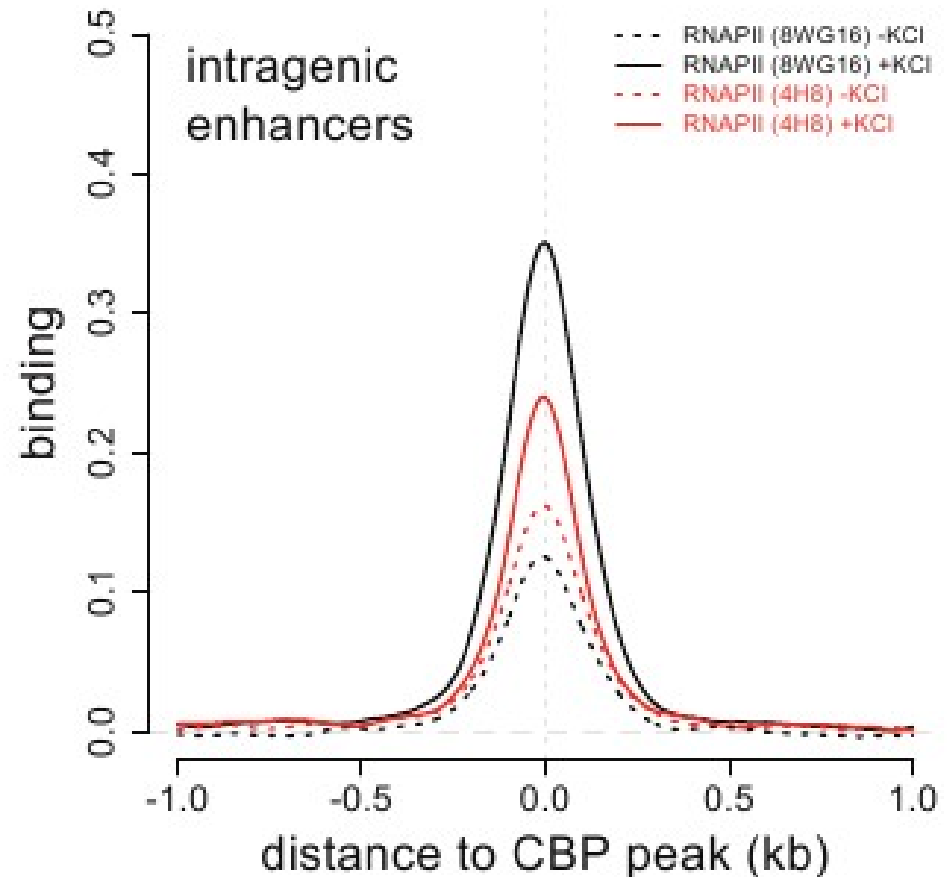
Intragenic enhancers

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



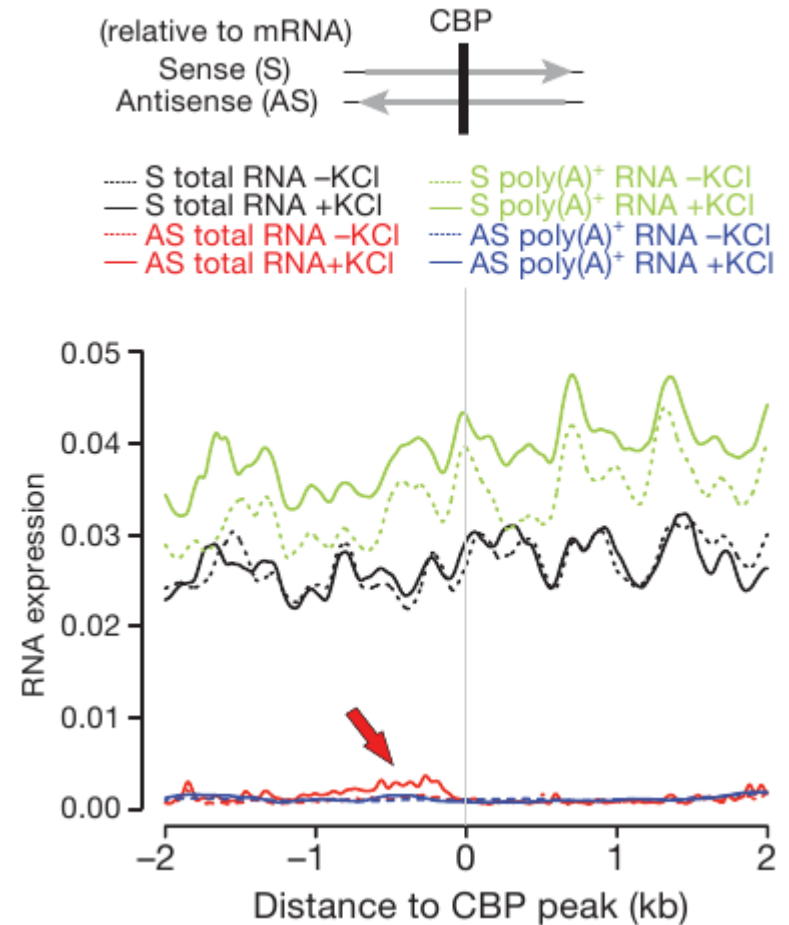
Intragenic enhancers

- ~7,000 enhancers overlapping introns
 - Inducible RNAPII



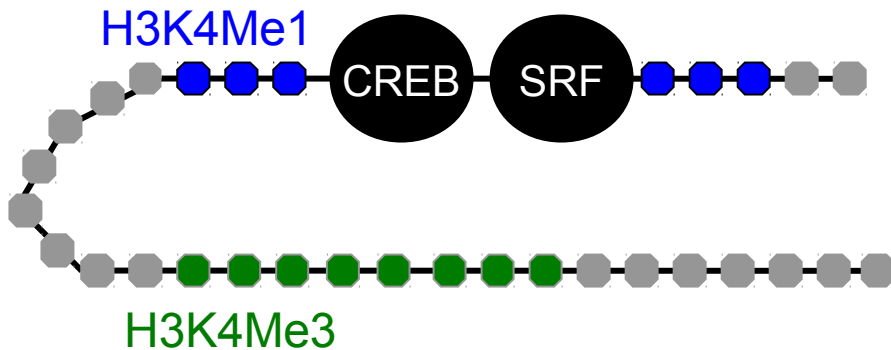
Intragenic enhancers

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

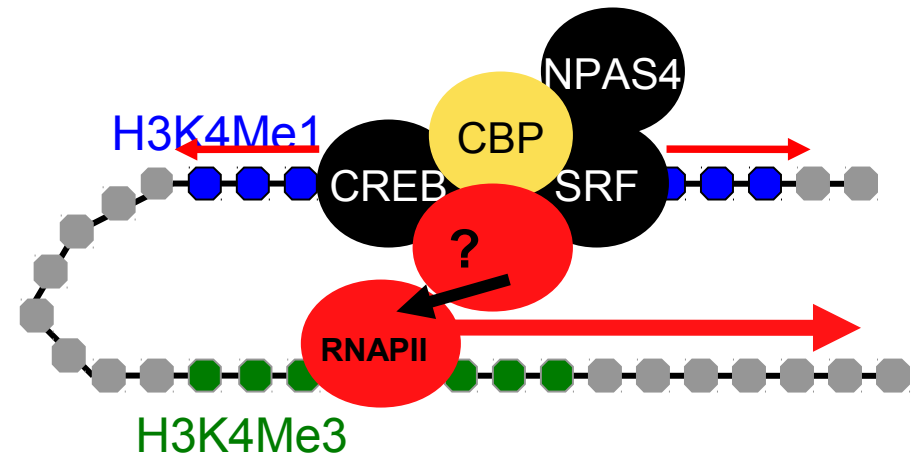


Properties of activity regulated enhancers

Before neuronal activation

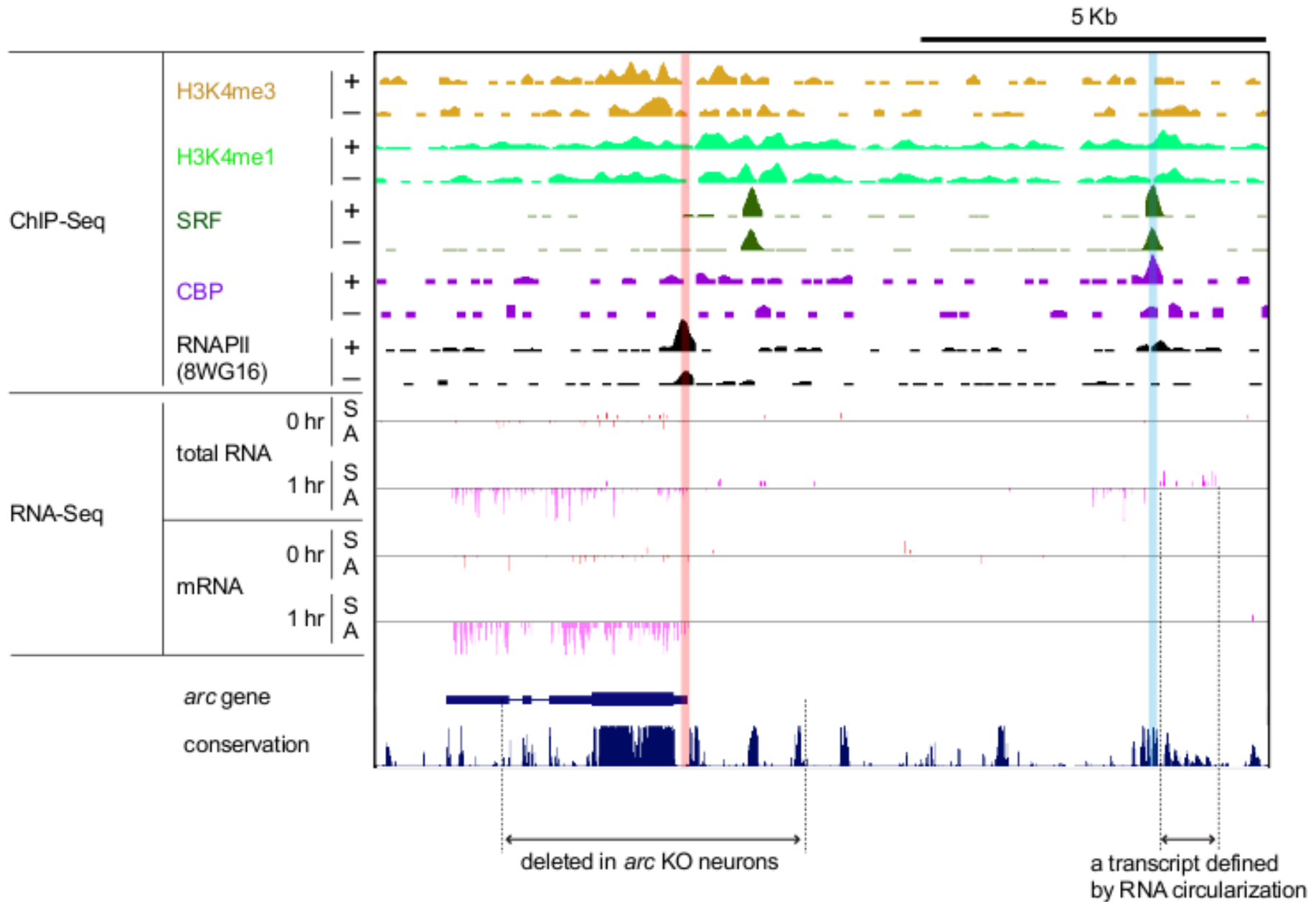


After neuronal activation

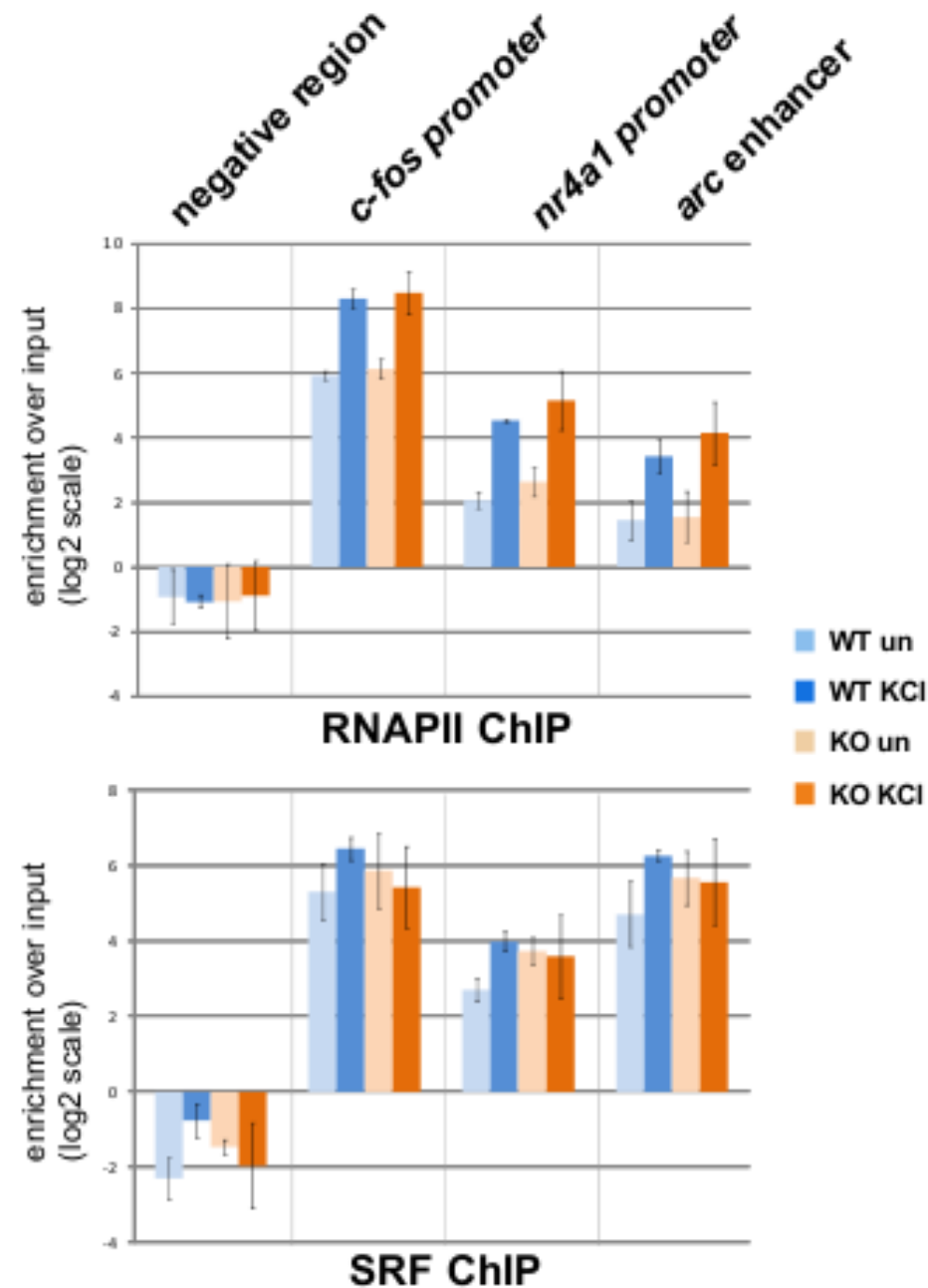


- Does RNAPII bind at enhancers? **YES**
- Are transcripts produced at enhancers? **YES**
- Is RNAPII recruitment independent?

arc locus and enhancer

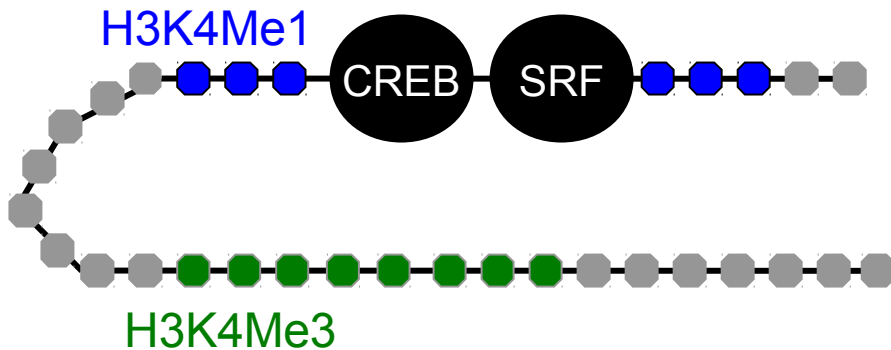


Recruitment of RNAPII and TFs without promoter

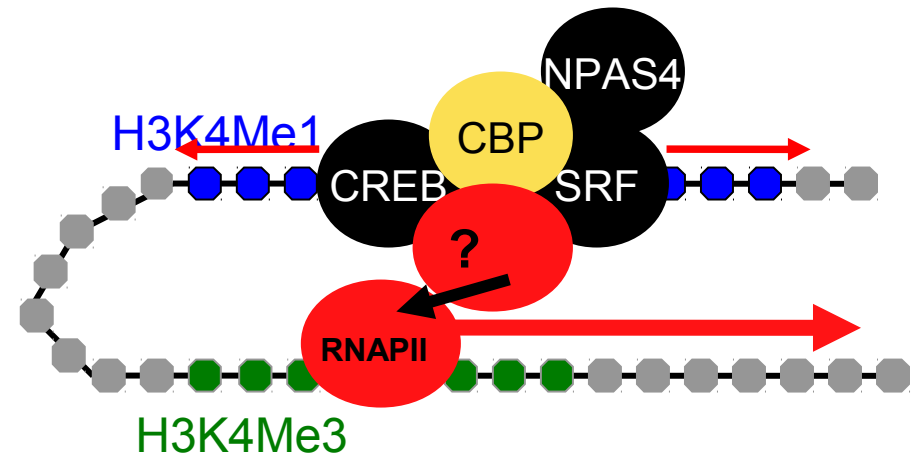


Properties of activity regulated enhancers

Before neuronal activation

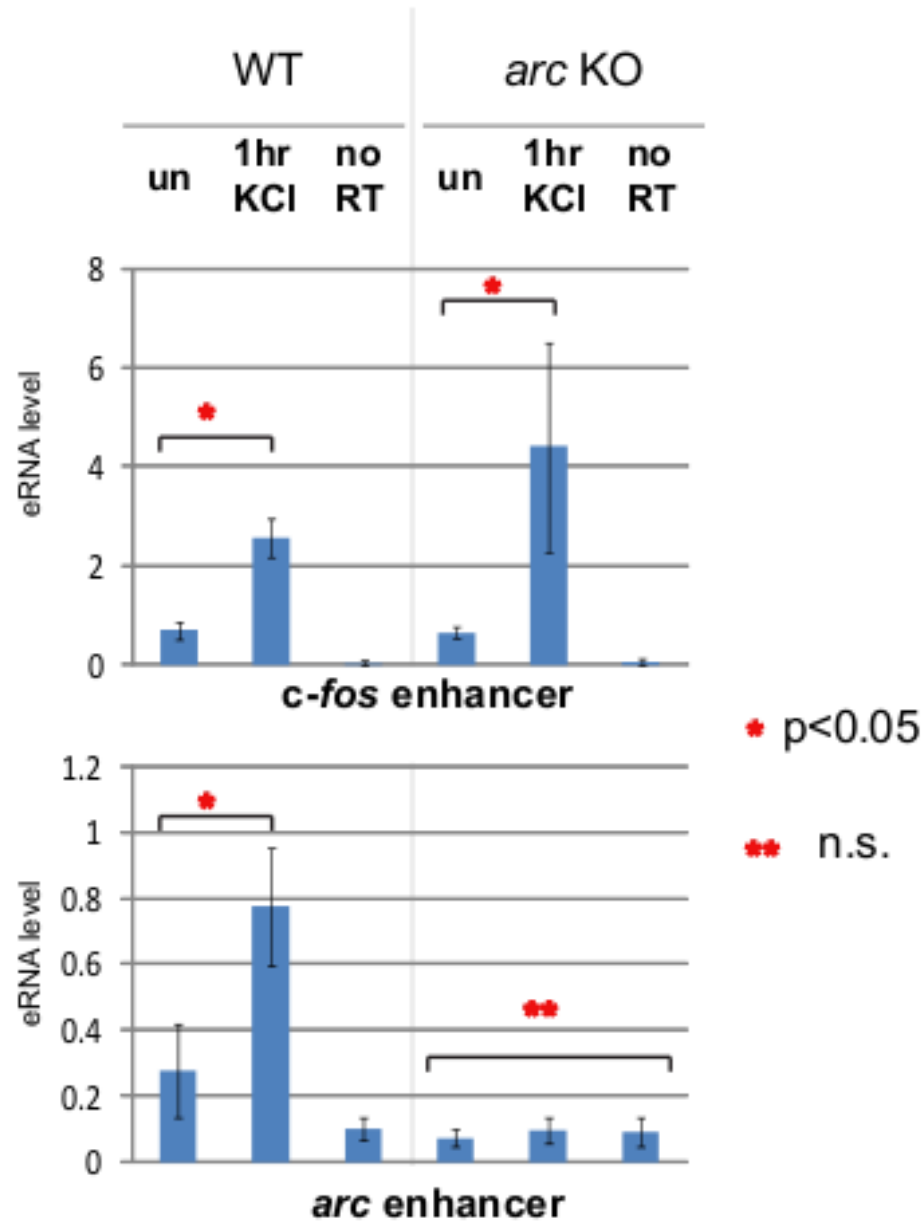


After neuronal activation



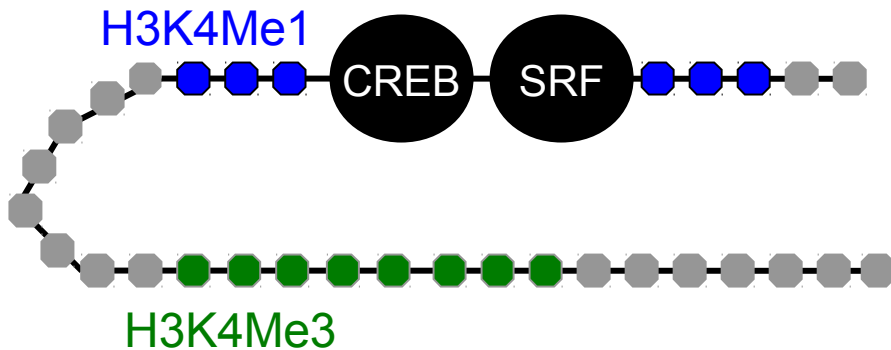
- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers? YES
- Is RNAPII recruitment independent? YES
- Is eRNA production independent?

arc promoter required for eRNA

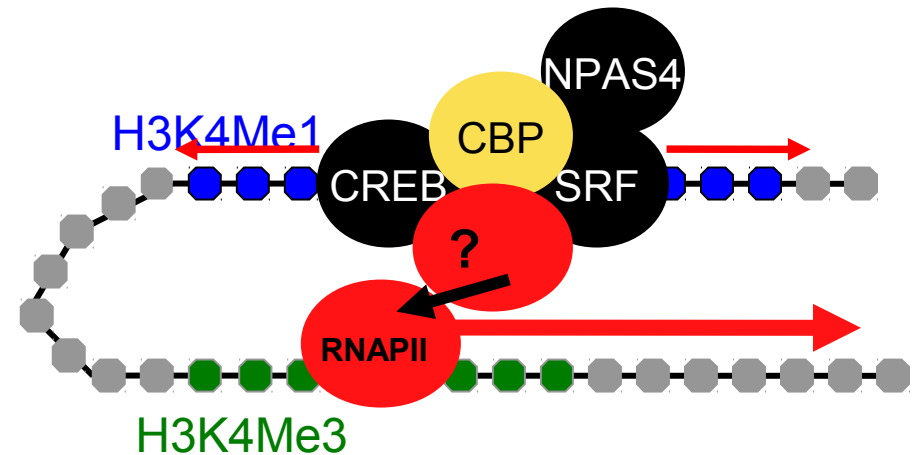


Properties of activity regulated enhancers

Before neuronal activation



After neuronal activation

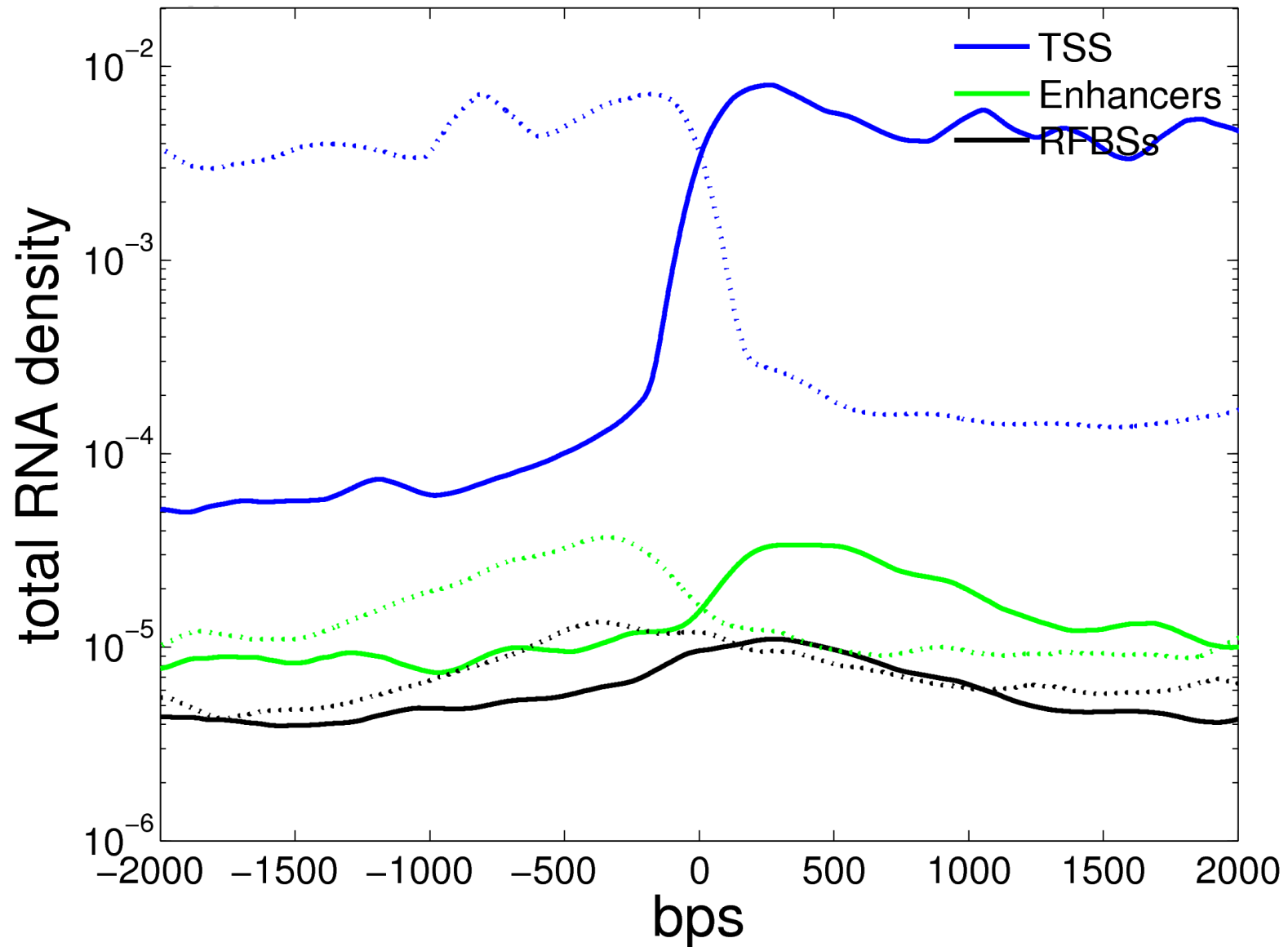


- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers? YES
- Is RNAPII recruitment independent? YES
- Is eRNA production independent? NO

Optimizing the parameters

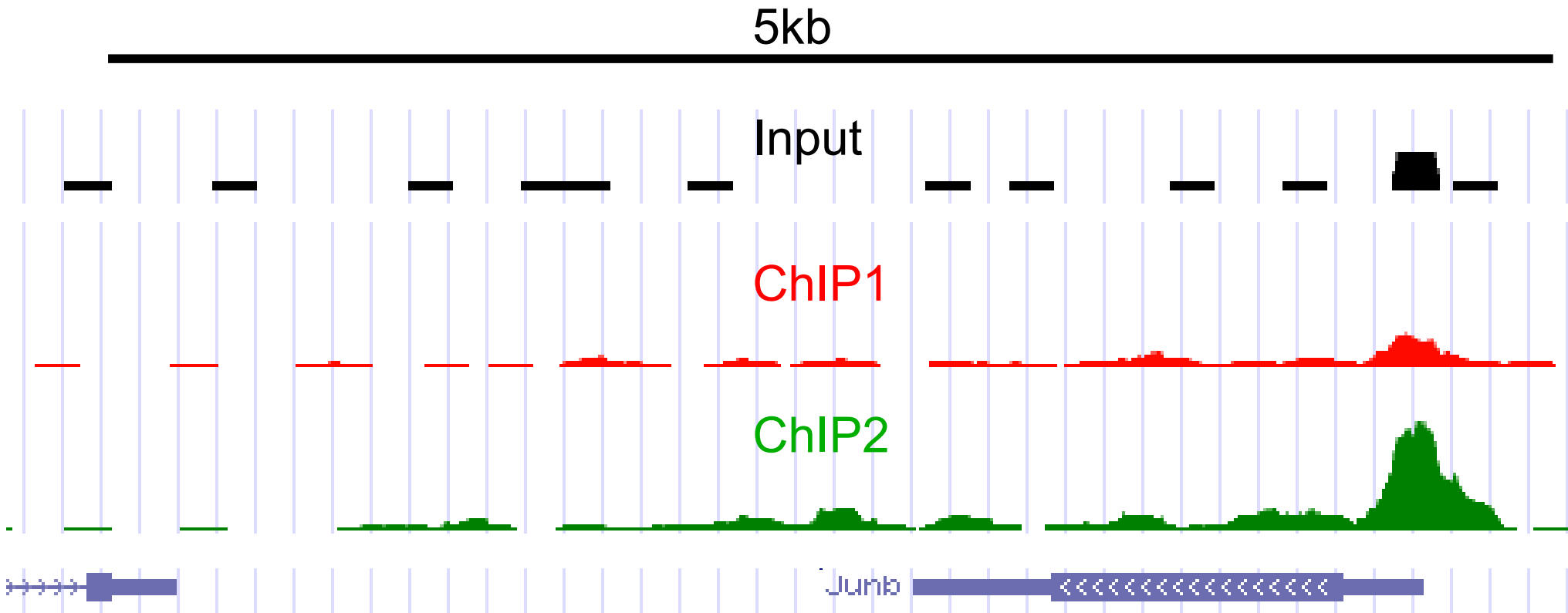
- Binning, minimum and maximum Haar-wavelet-length
- FDR for choosing break-points and transcribed regions
 - Sweep parameter space and maximize the fraction of regions that have a H3K4me3 peak at their start
 - Running HaTriC on one chr takes only a few minutes

mRNA much more abundant than eRNA



Subtract input to identify significant peaks

- $Z_i = \#ChIP \text{ reads} - \#input \text{ reads in window } i$
- $\sim 1 \text{ read}/100 \text{ bp}$



Assume ChIP and input Poisson distributed

- $Z_i = \text{\#ChIP reads} - \text{\#input reads in window } i$
- $\sim 1 \text{ read}/100 \text{ bp}$
 - Assume $\text{\#reads in window } P(k) = \lambda^k \exp(-\lambda)/k!$
 - Difference between two Poisson random variables
 - $Z_i \sim \text{Skellam}(z, \lambda_1, \lambda_2)$
$$p(x) = e^{-(\lambda_1 + \lambda_2)} (\lambda_1 / \lambda_2)^{x/2} I_x(2 \sqrt{\lambda_1 \lambda_2})$$

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

Haar-wavelet Transcript Calling (HaTriC) for *de novo* identification of transcribed regions

```
Calculate_RNA_density_for_128_bp_bins
```

```
do
```

```
    find_breakpoints
```

```
    calculate_region_densities
```

```
    determine_cutoff_density
```

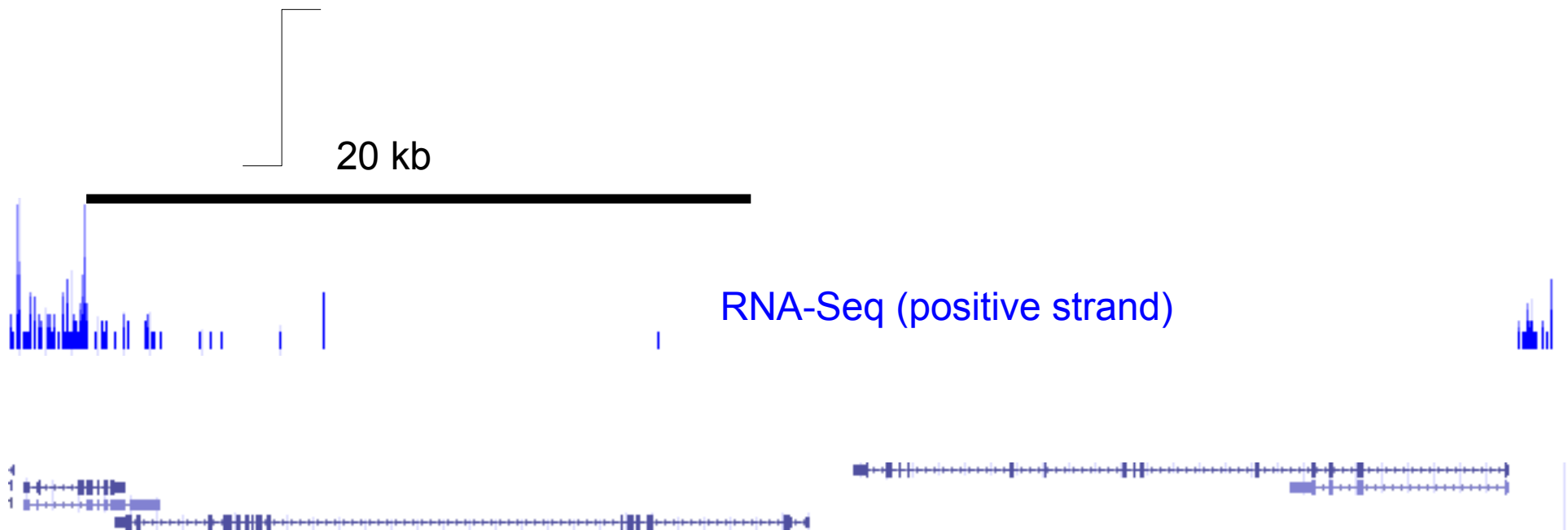
```
    remove_transcribed_regions
```

```
while new_regions_found
```

The Haar-wavelet picks out regions with sharp changes in read density

- Break points correspond to sharp changes in read density

$$h_L(n) = \frac{1}{\sqrt{2^{L+1}}} \left(\sum_{i=n}^{n+2^L-1} \log(1 + r_i) - \sum_{i=n-2^L}^{n-1} \log(1 + r_i) \right)$$



The Haar-wavelet can be scaled to analyze multiple length scales

- Break points correspond to sharp changes in read density

$$h_L(n) = \frac{1}{\sqrt{2^{L+1}}} \left(\sum_{i=n}^{n+2^L-1} \log(1 + r_i) - \sum_{i=n-2^L}^{n-1} \log(1 + r_i) \right)$$

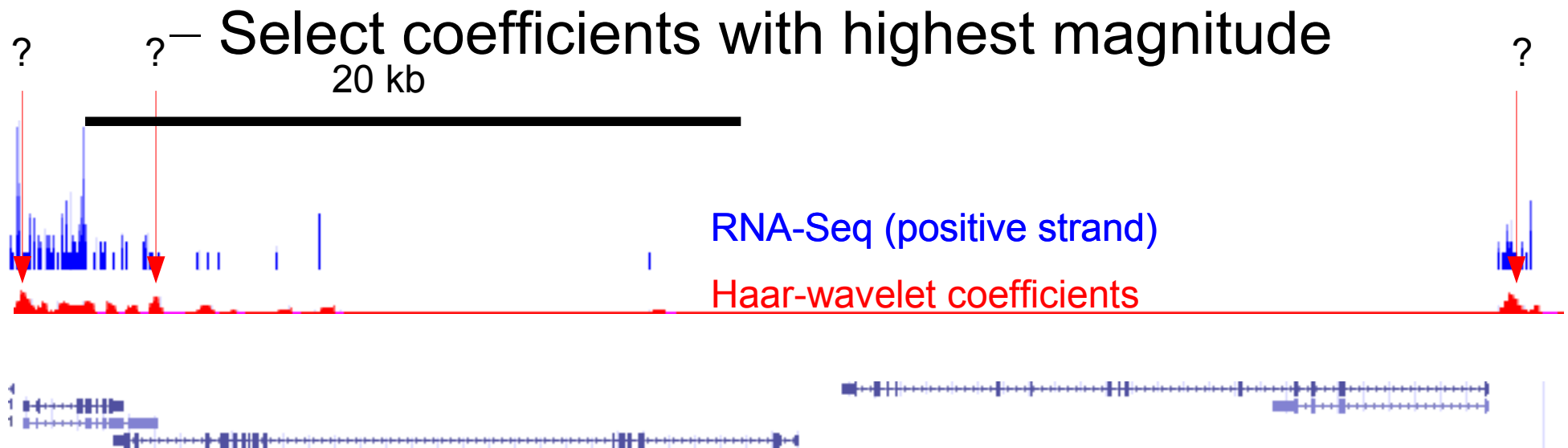
- Use scales L from 8 to 20



The coefficients with largest magnitude are selected as candidate break points

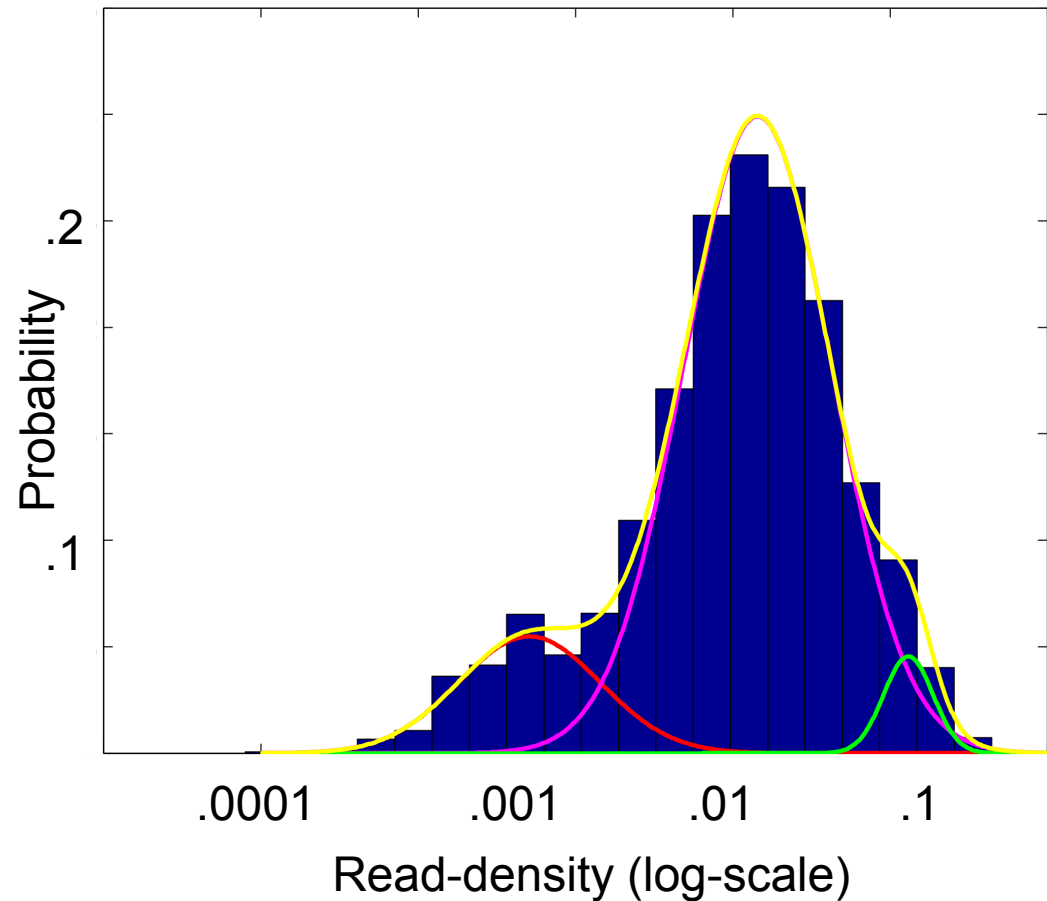
- Break points correspond to sharp changes in read density

$$h_L(n) = \frac{1}{\sqrt{2^{L+1}}} \left(\sum_{i=n}^{n+2^L-1} \log(1 + r_i) - \sum_{i=n-2^L}^{n-1} \log(1 + r_i) \right)$$



The density distribution for the regions determined by the break points is bimodal

- Average density between breakpoints
- Keep regions belonging to higher mode



Remove transcribed regions, iterate the process is until no new regions are found

- Allows us to find regions with lower expression levels

