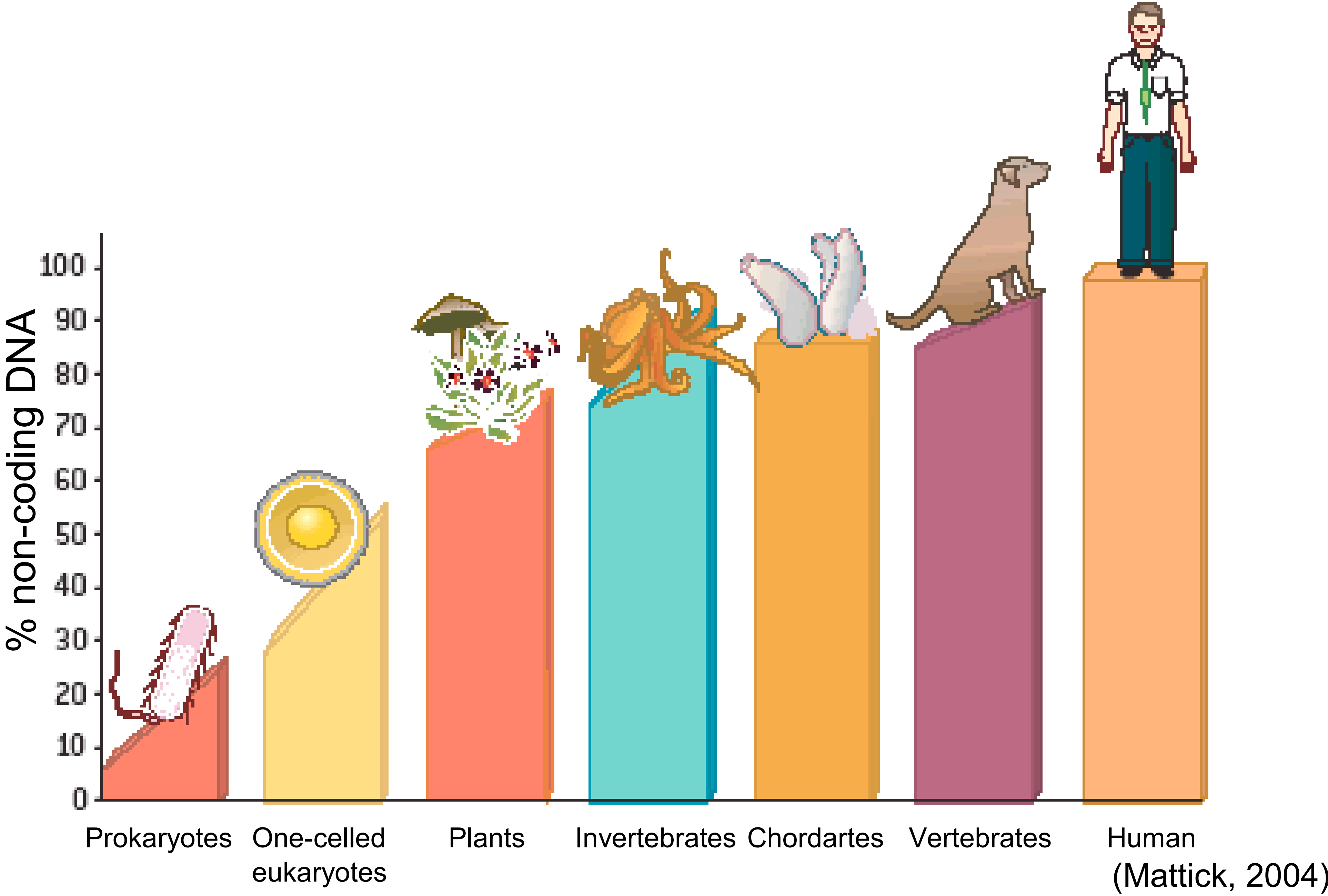
Probing the function of non-coding DNA using high throughput sequencing

Martin Hemberg

University of California, San Diego
April 25, 2011

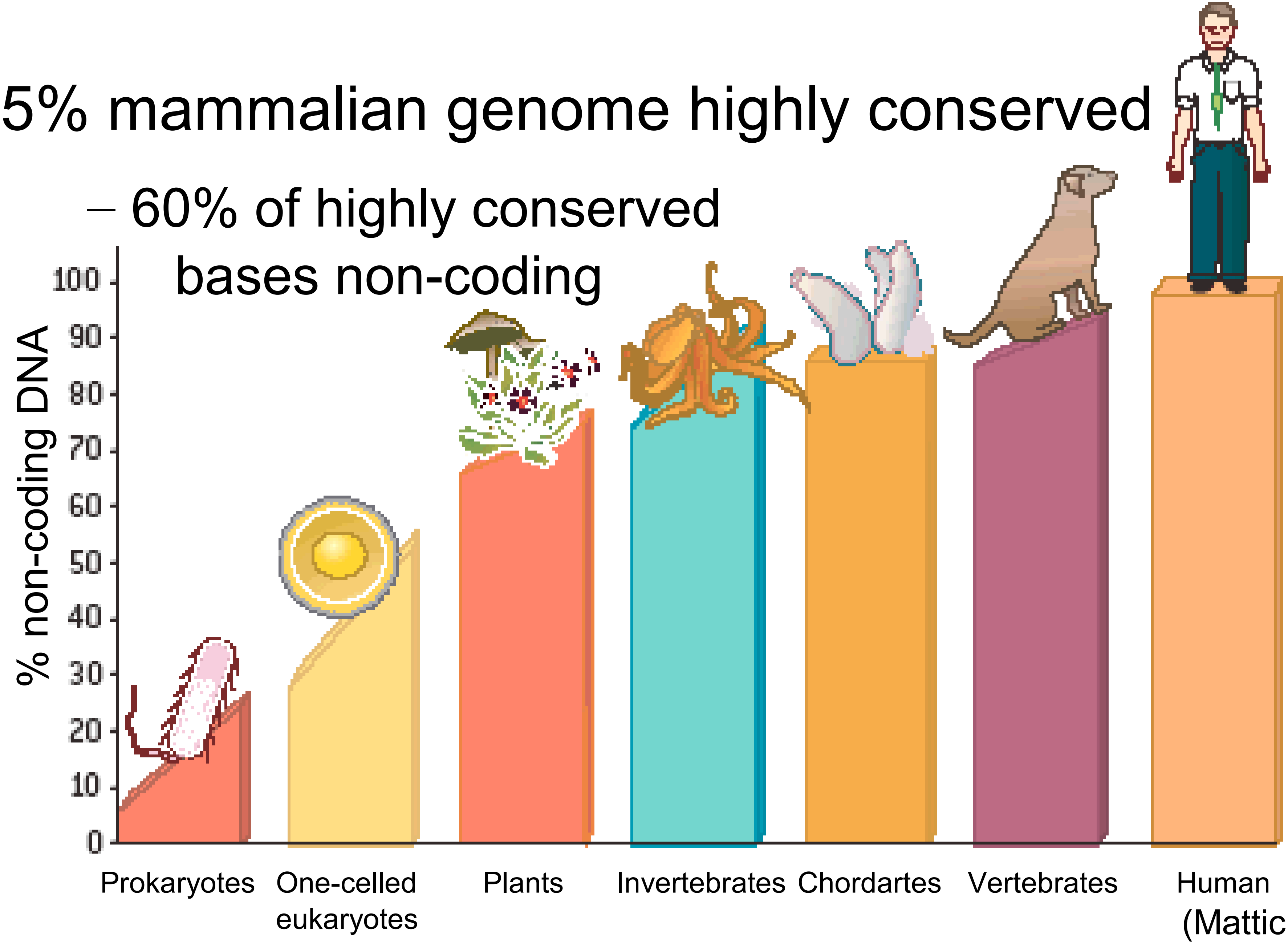
Most of the genome is **not** protein-coding



(Mattick, 2004)

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

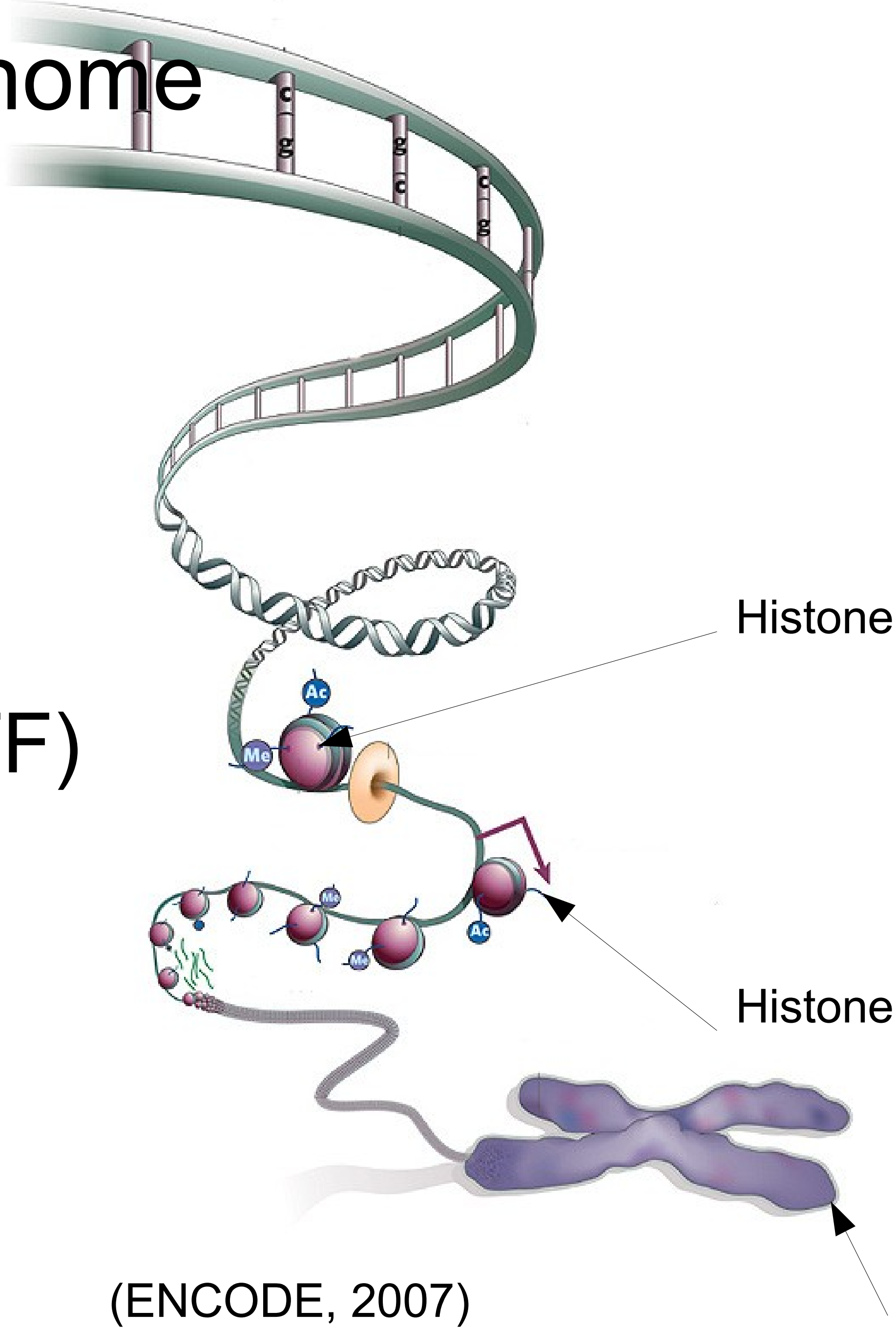
- 5% mammalian genome highly conserved
  - 60% of highly conserved bases non-coding



(Mattick, 2004)

Additional layers of modifications determine the function of the genome

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



Histone

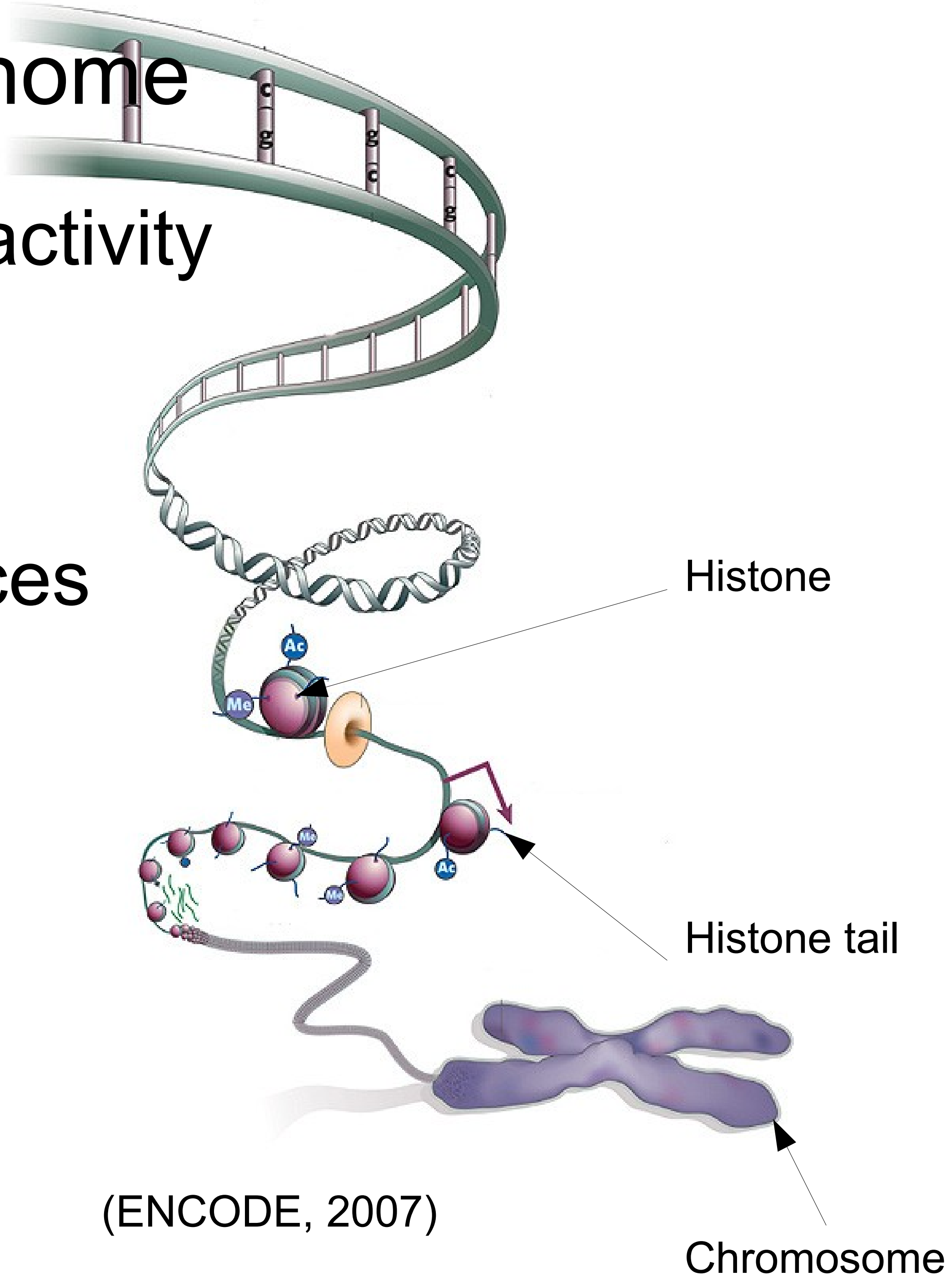
Histone tail

Chromosome

(ENCODE, 2007)

Additional layers of modifications determine the function of the genome

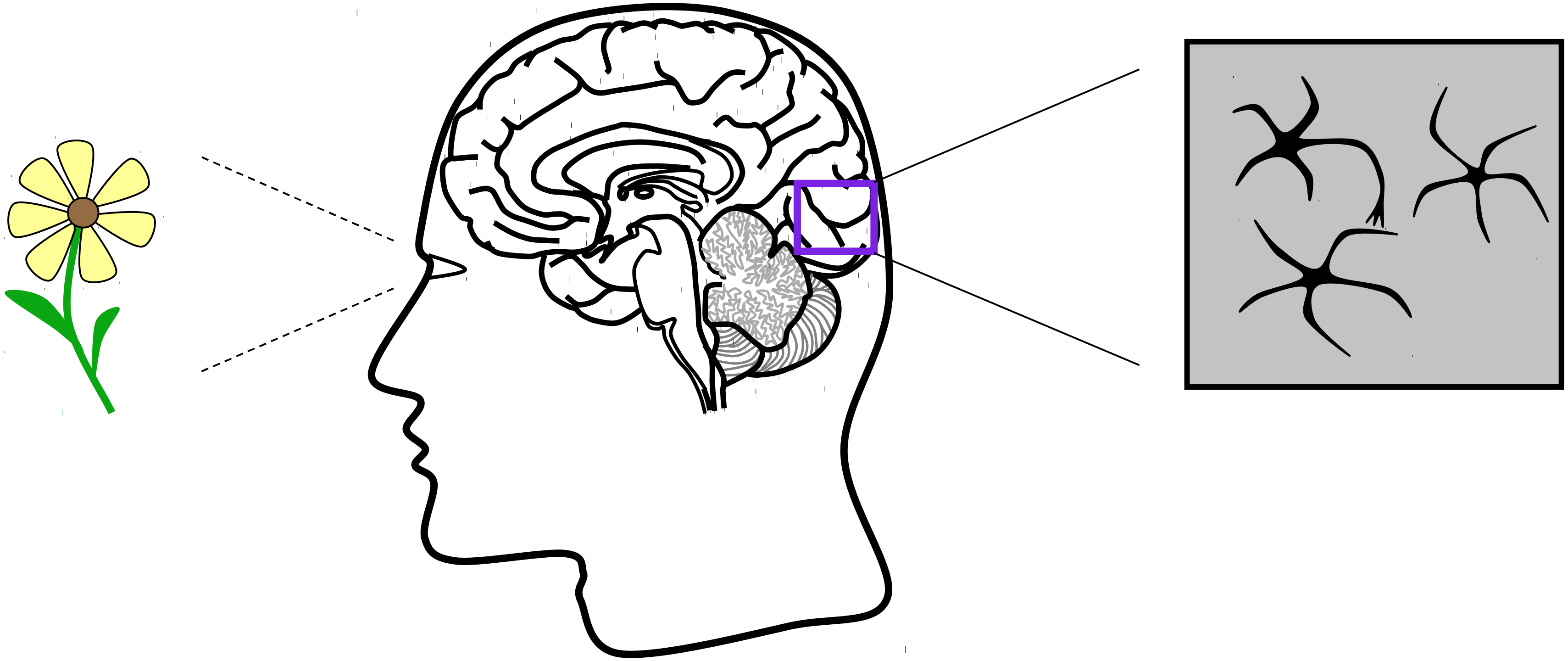
- Correlates with gene activity
  - Cell-type specificity
- Understand role of non-coding sequences



(ENCODE, 2007)

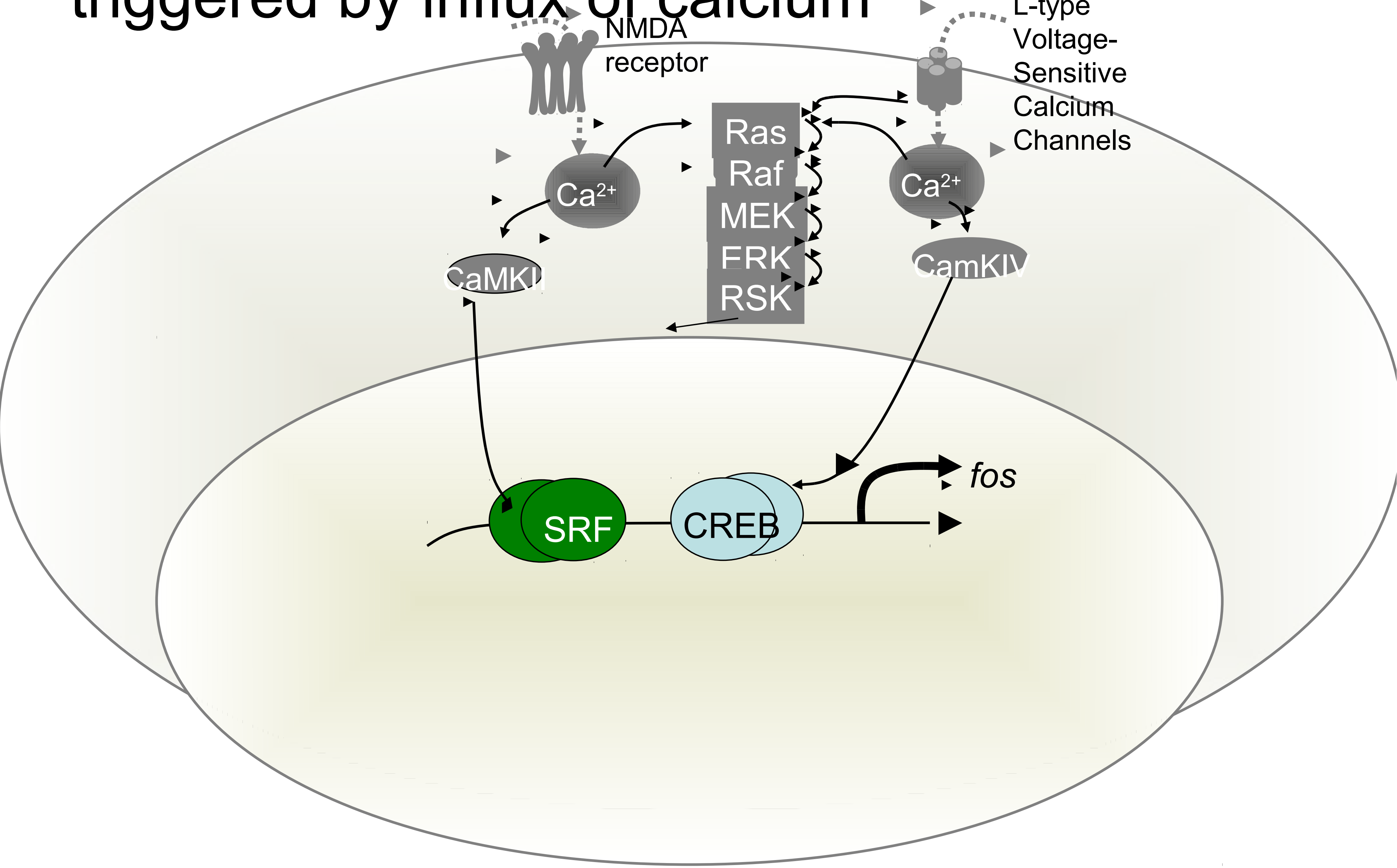
Activity-dependent gene expression

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

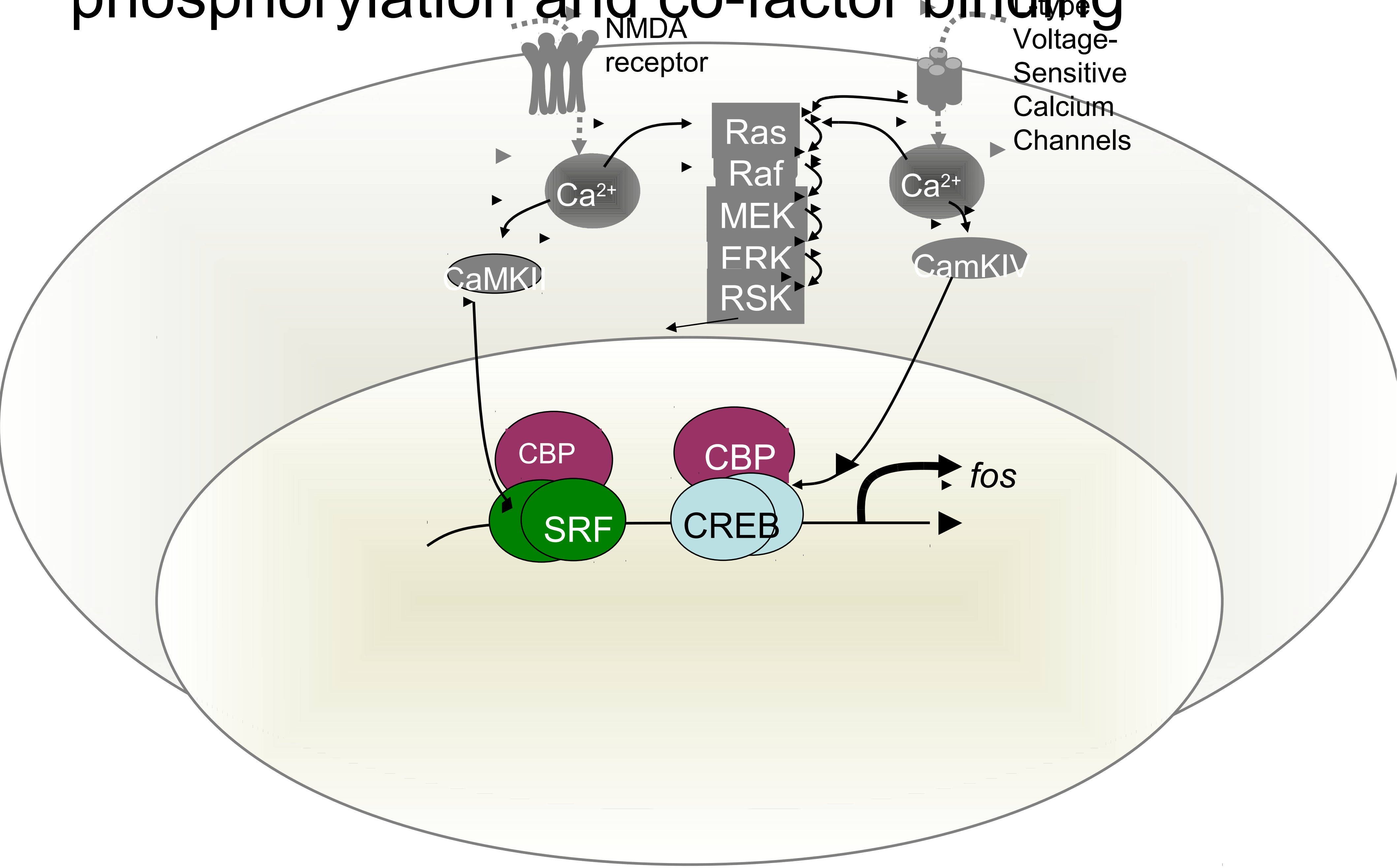


Hubel & Wiesel, 1970’s

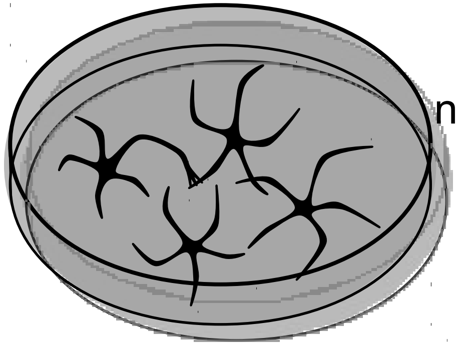
Activity-dependent gene expression is triggered by influx of calcium



Immediate-early genes are activated by phosphorylation and co-factor binding



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



mouse cortical neurons

neuronal activation via potassium chloride (KCl) depolarization

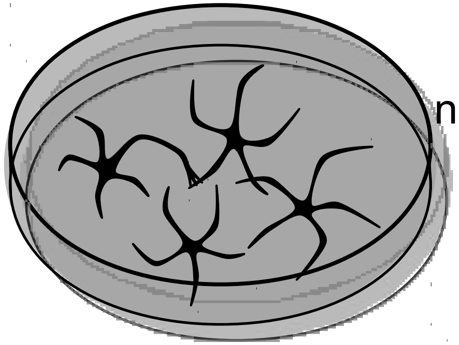
- KCl

ChIP-Seq
RNA-Seq

+ KCl

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

ChIP-Seq
RNA-Seq

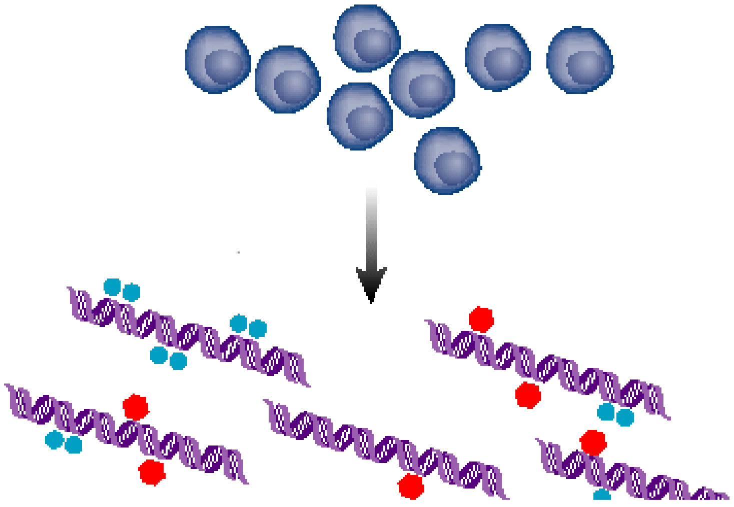
+ KCl

ChIP-Seq
RNA-Seq

Jesse Gray
Tae-Kyung Kim
Greenberg Lab

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

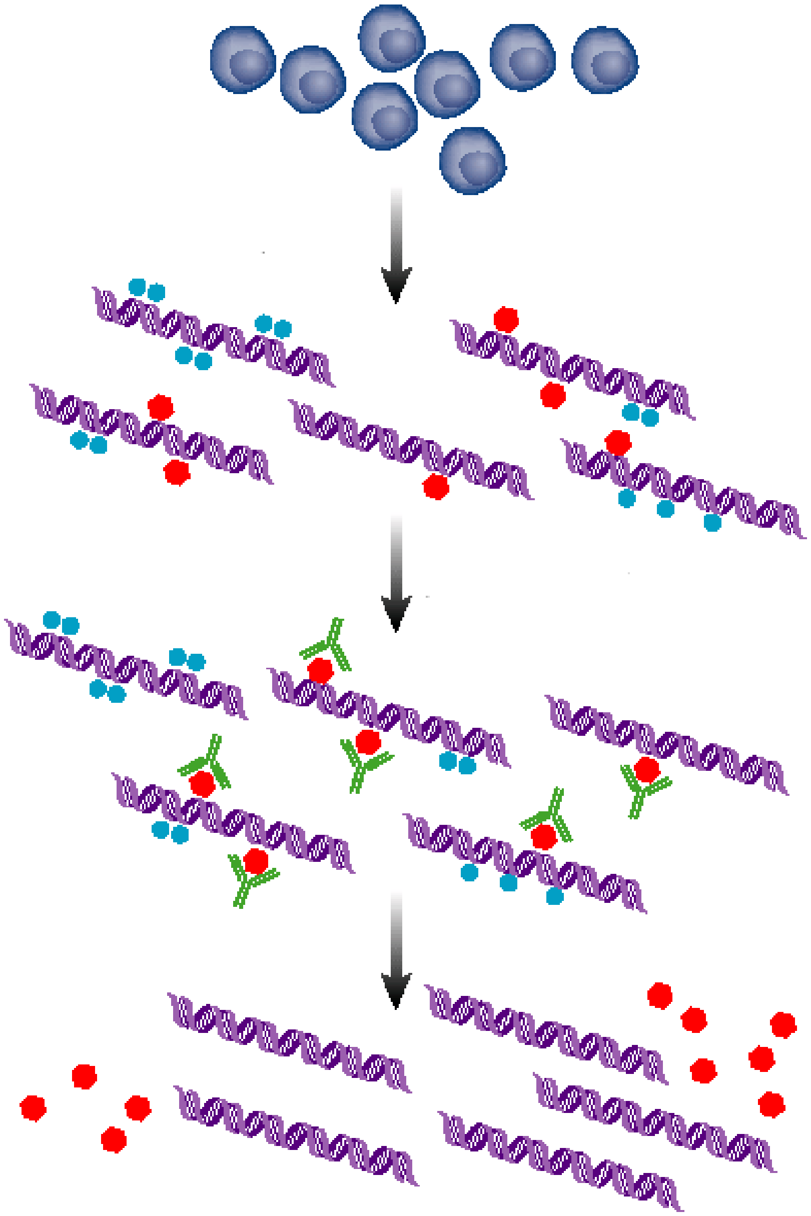
- Cross-link TF
- Fragment DNA



(Mardis, 2007)

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

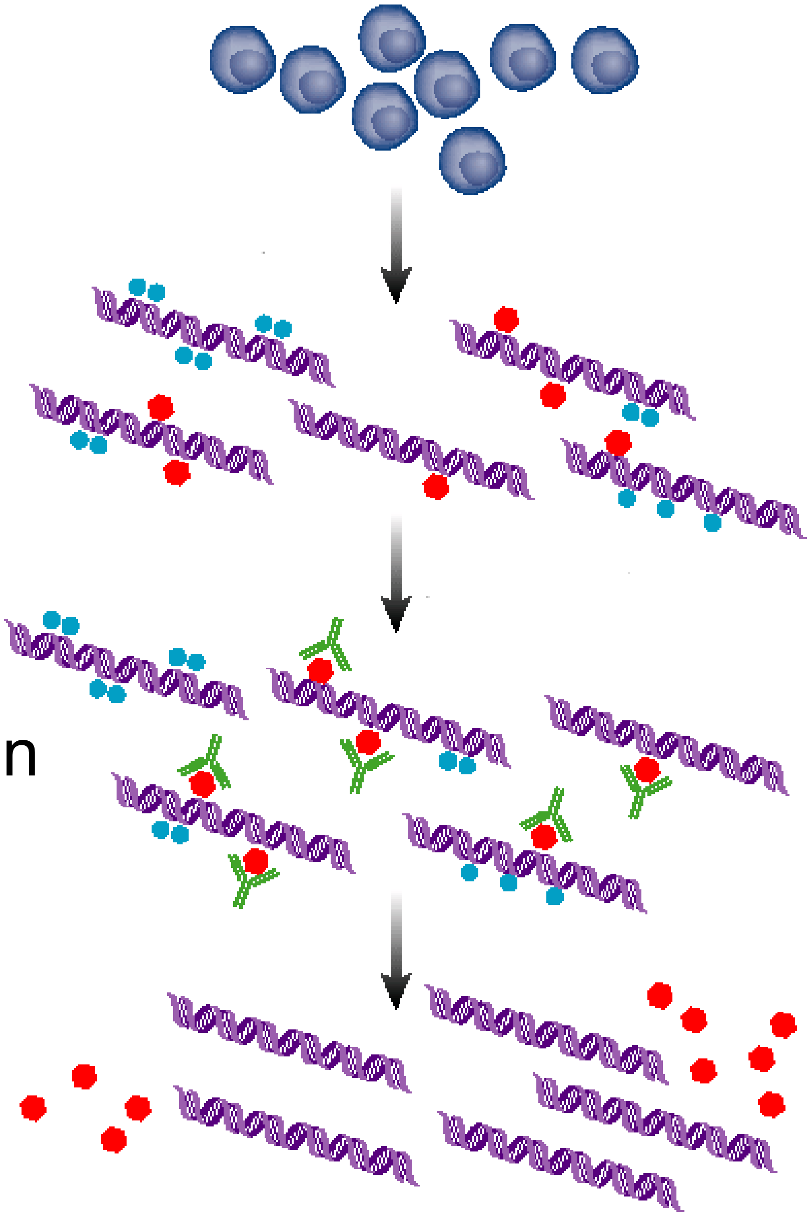
- Cross-link TF
- Fragment DNA
- Extract with antibody
- Reverse crosslink
- Sequence fragments



(Mardis, 2007)

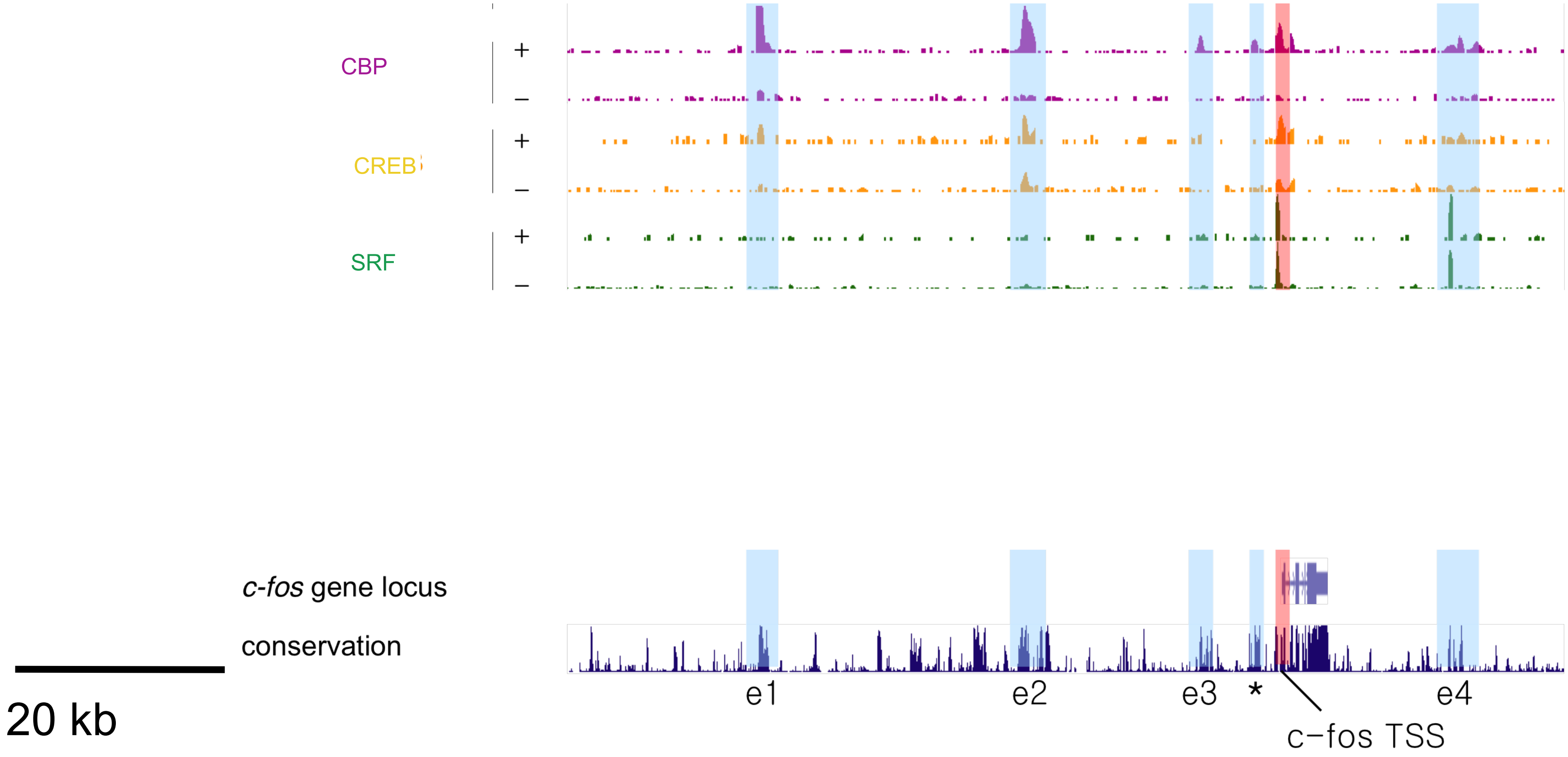
Chromatin immunoprecipitation and sequencing (ChIP-Seq) finds protein binding sites *in vivo*

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

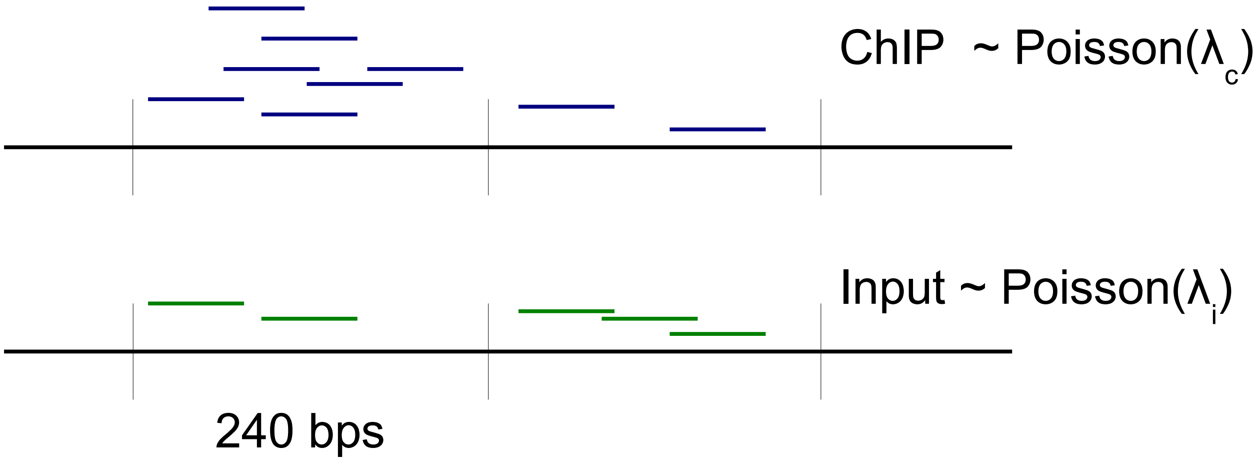


(Mardis, 2007)

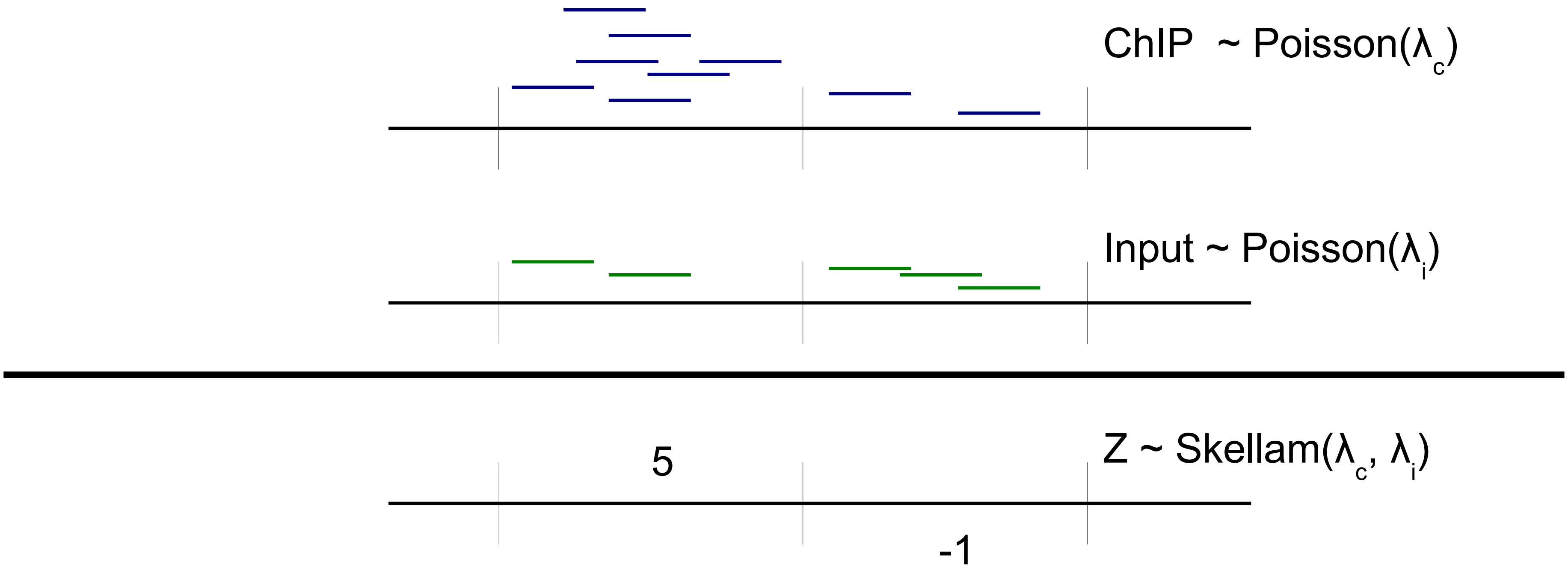
CBP binding depends strongly on activity at the *fos* promoter and flanking loci



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



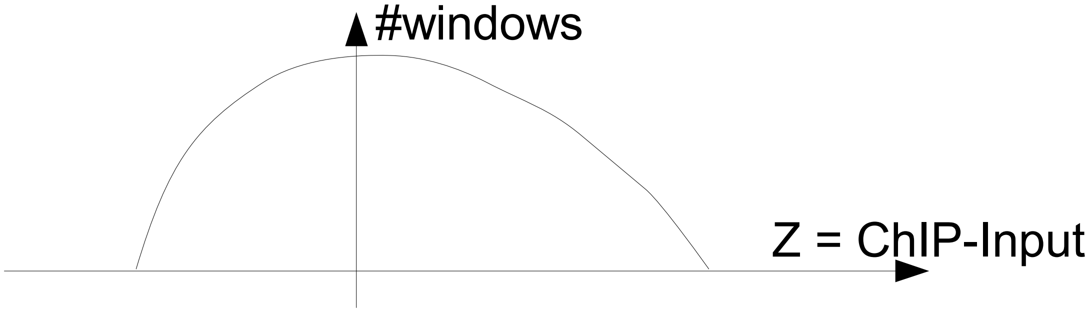
Identifying regions with larger than expected number of ChIP-Seq reads



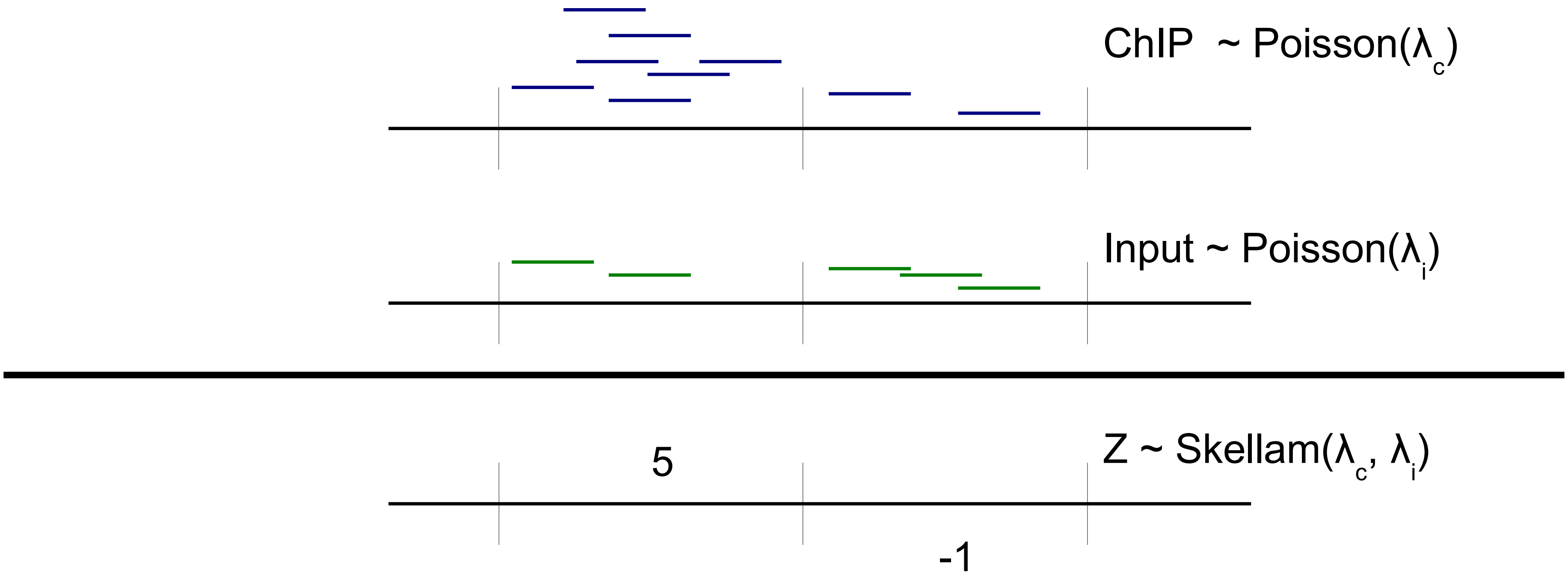
ChIP ~ Poisson(λ_c)

Input ~ Poisson(λ_i)

Z ~ Skellam(λ_c, λ_i)



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

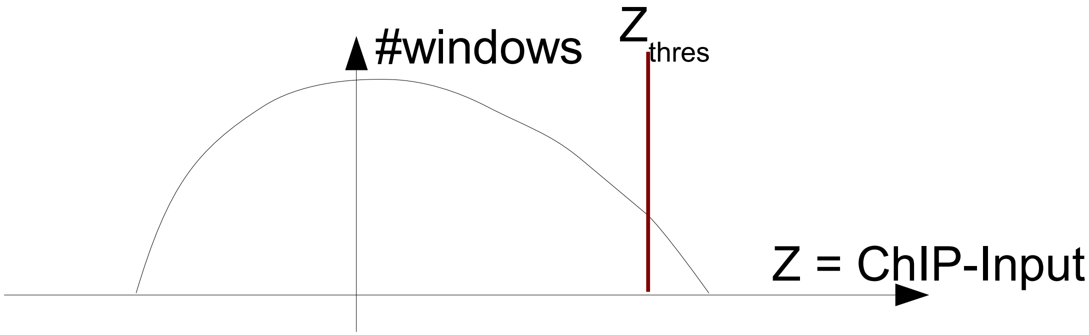


ChIP ~ Poisson(λ_c)

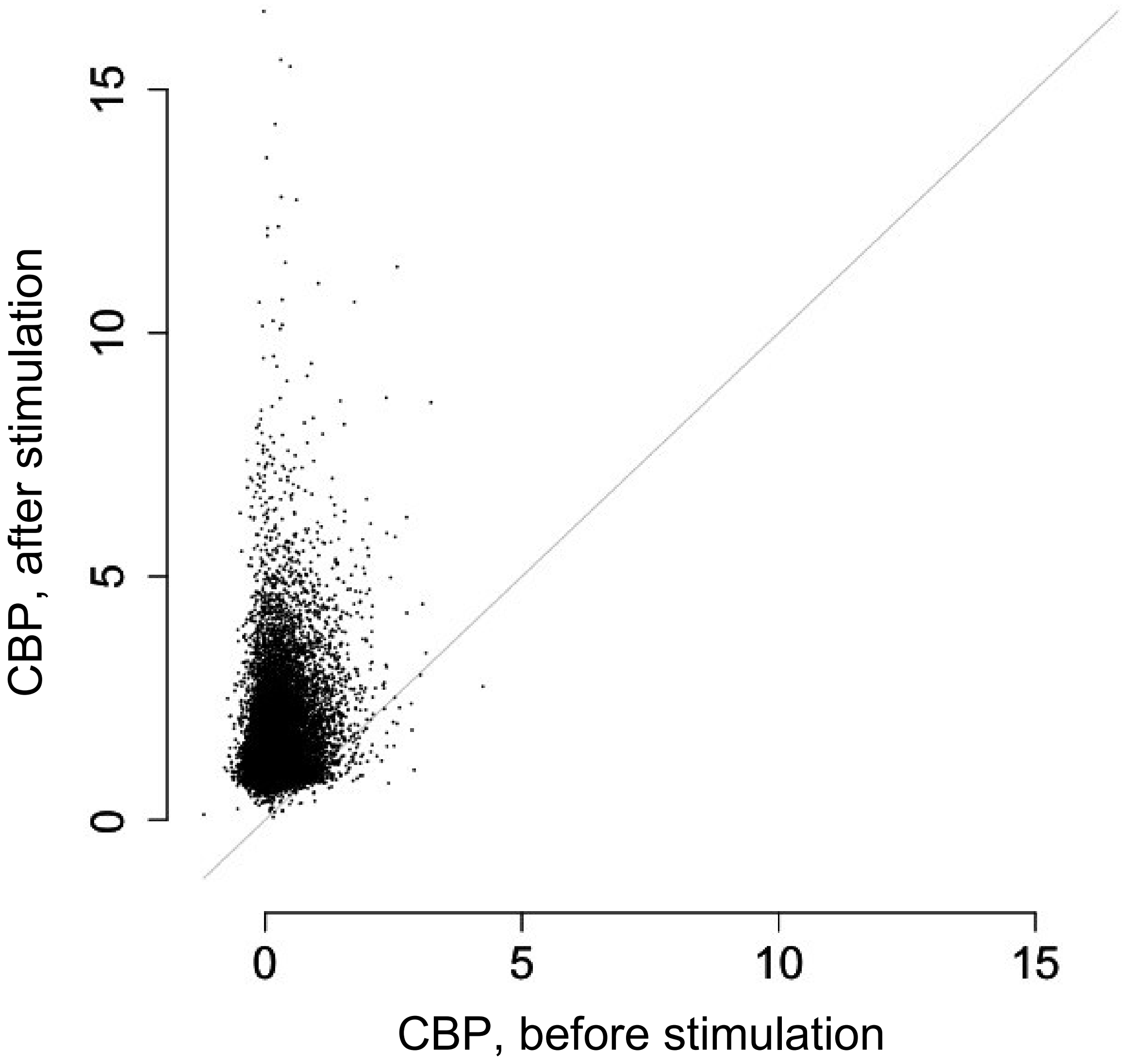
Input ~ Poisson(λ_i)

Z ~ Skellam(λ_c, λ_i)

- False Detection Rate (FDR) determine threshold

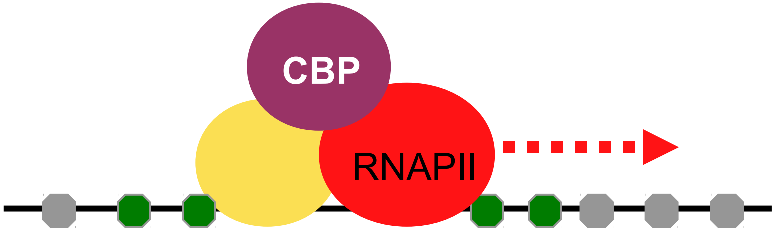


CBP binds in an activity regulated manner to ~28,000 sites throughout the genome



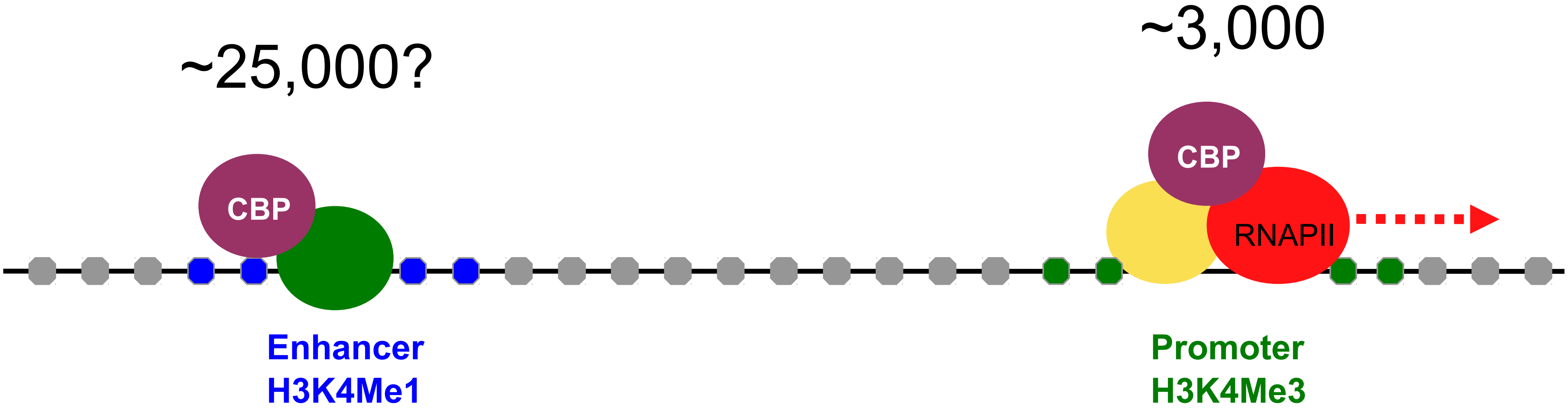
Only ~3000 CBP peaks at promoters

~3,000



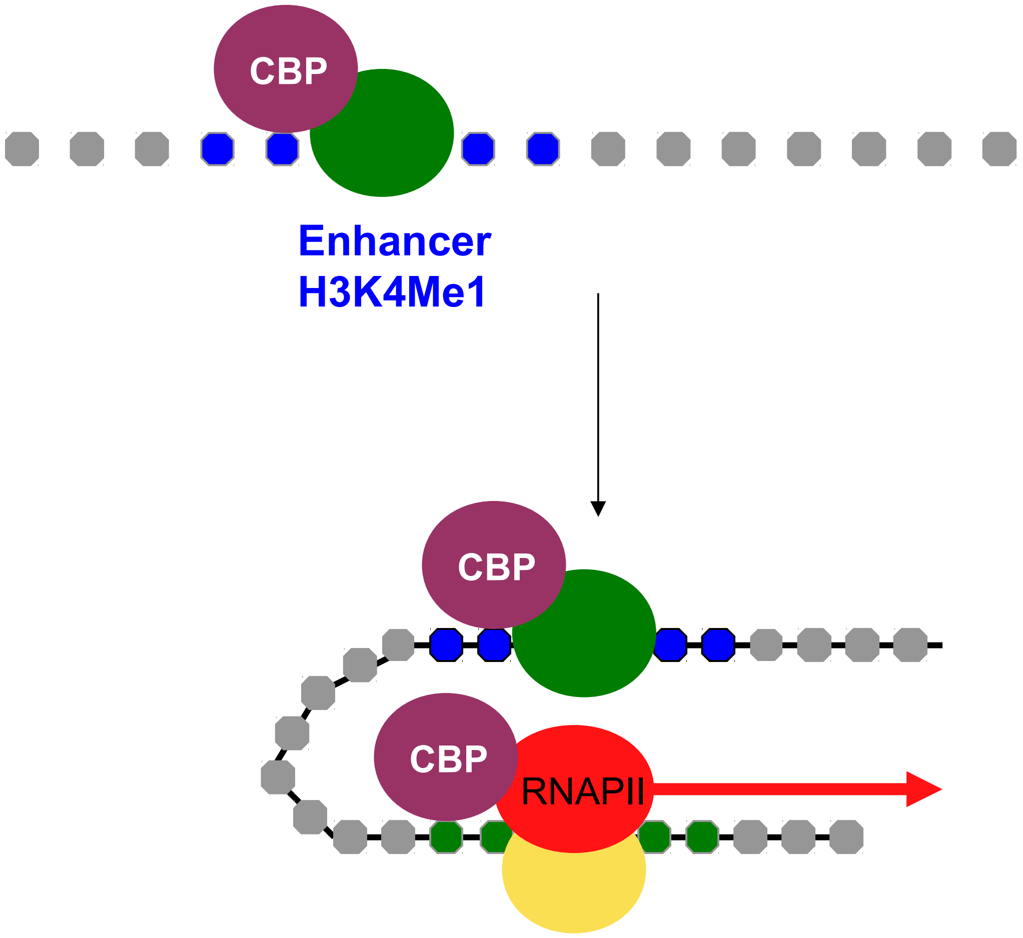
Promoter
H3K4Me3

CBP hypothesized to bind at enhancers



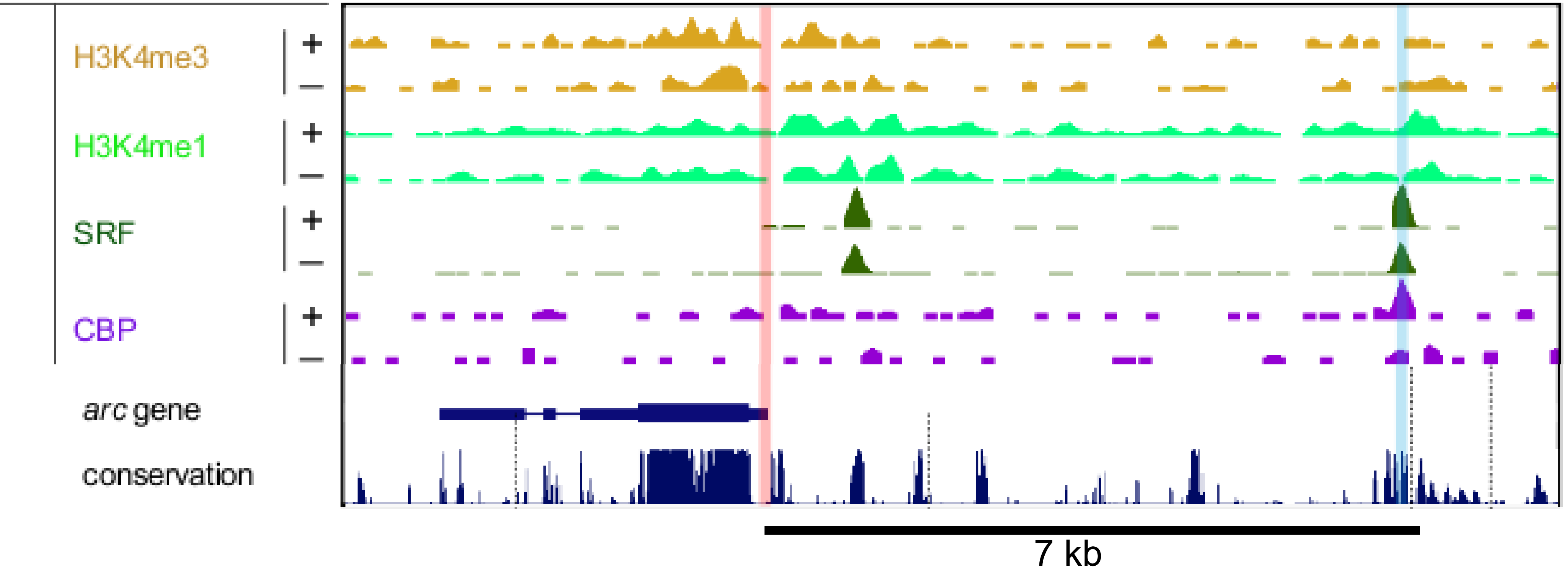
Enhancers are distal TF binding sites

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

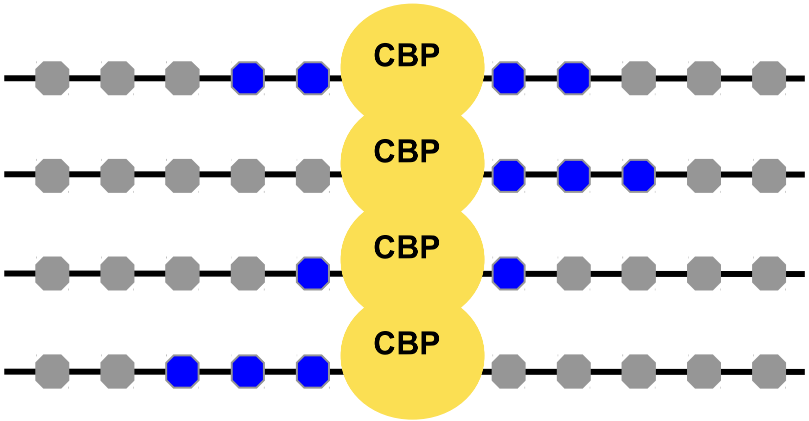


ENCODE, 2007
Heintzman et al, 2007
Roh et al, 2005
Visel et al, 2009

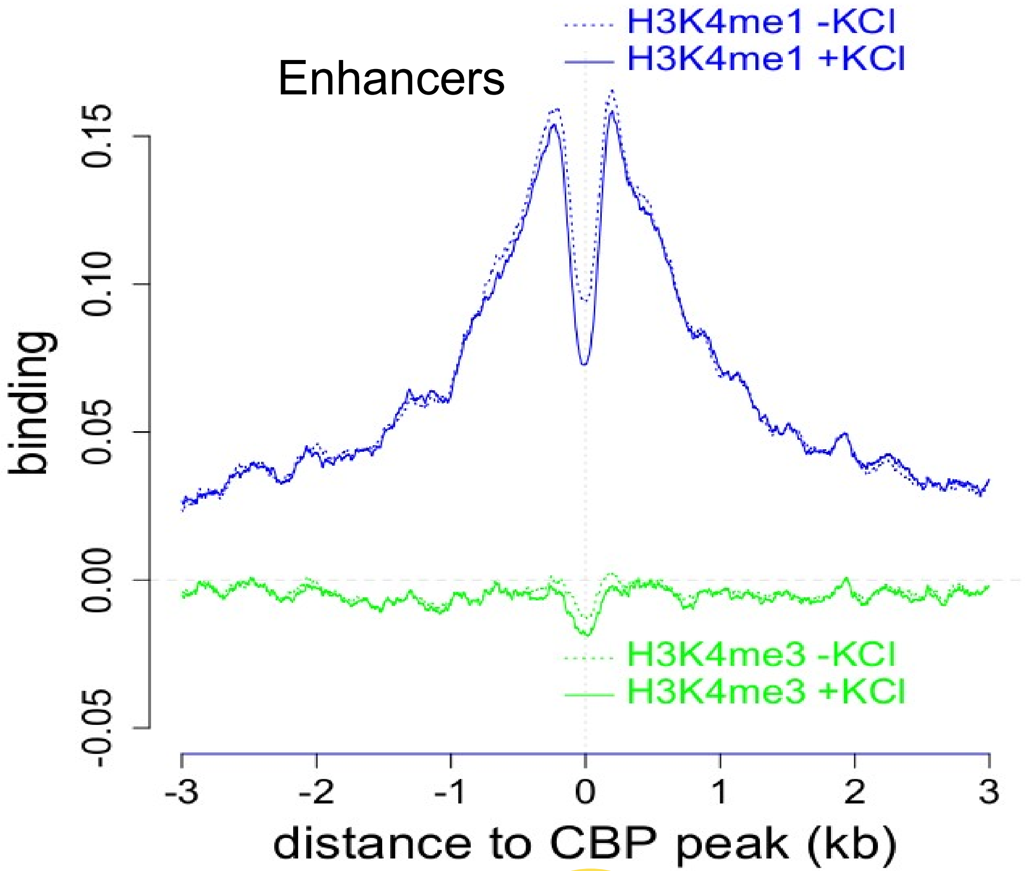
Distal CBP peaks have high levels of H3K4me1 but not H3K4me3

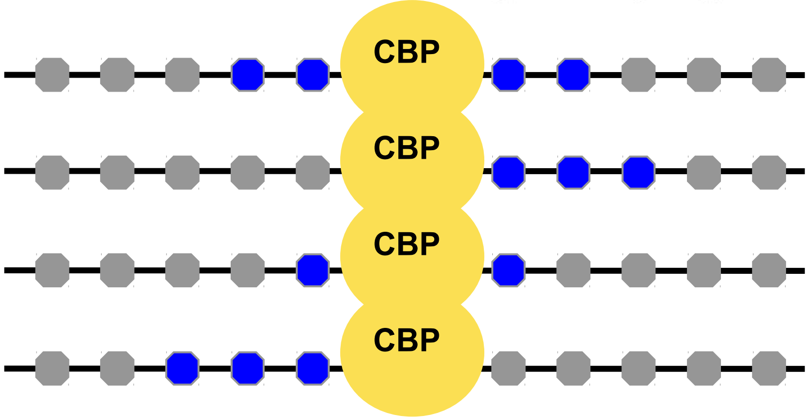


Aligning CBP peaks to calculate average binding profiles

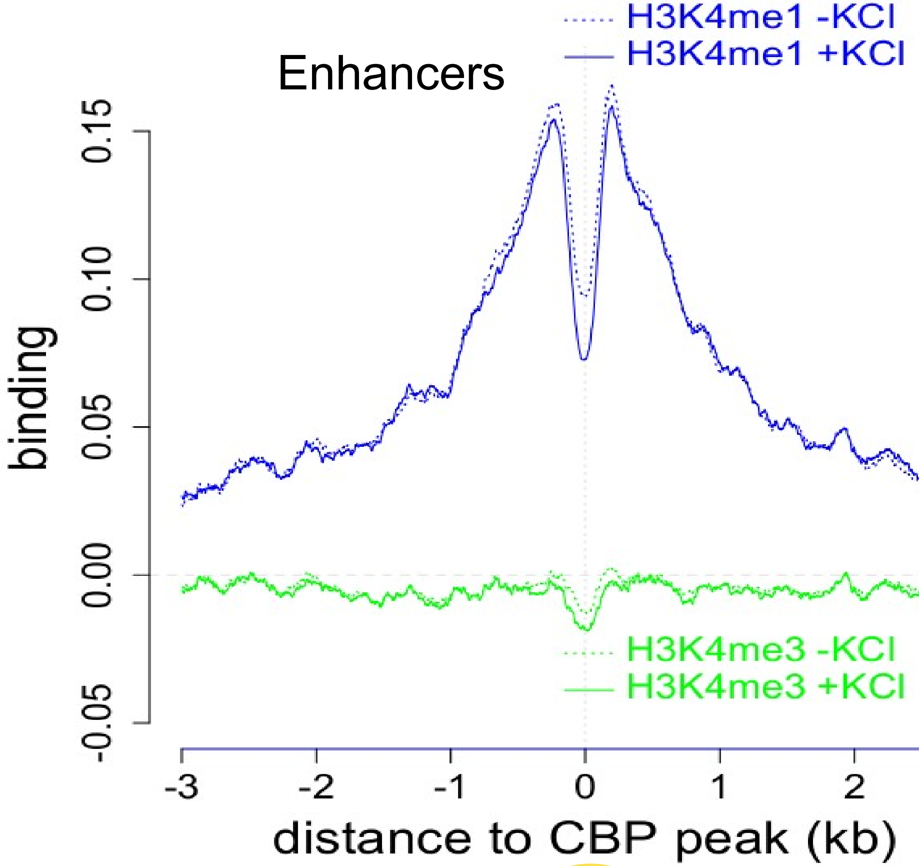


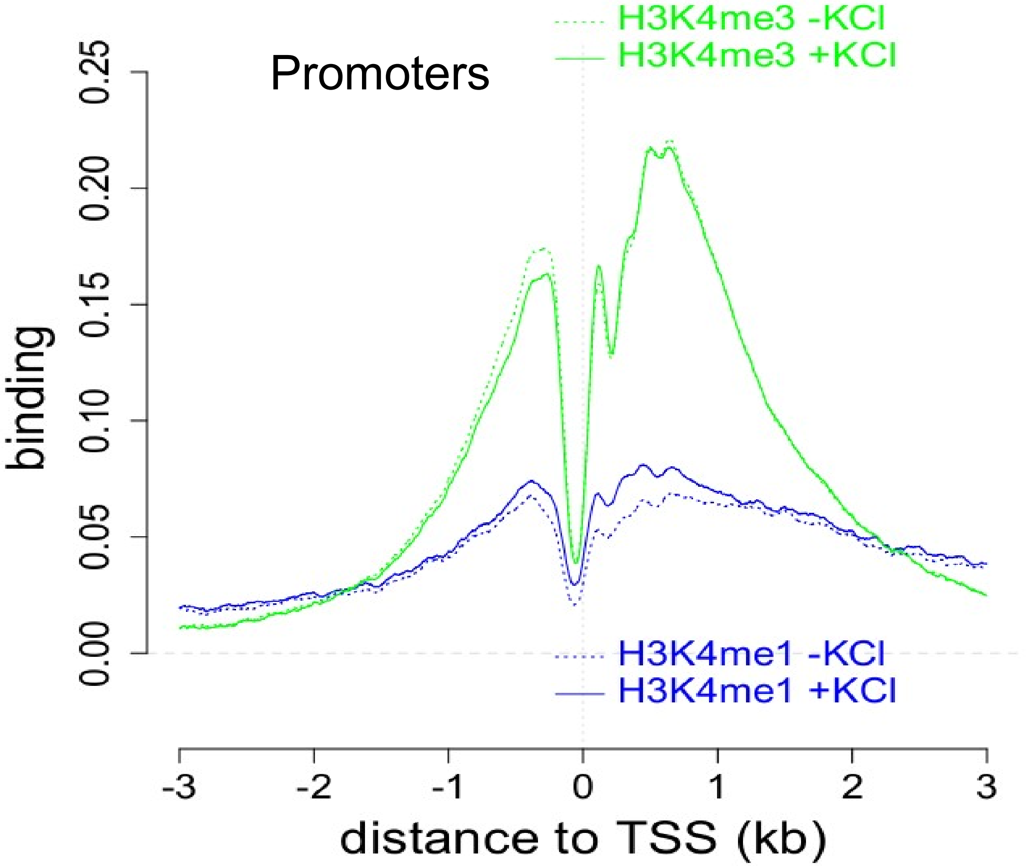
Most CBP peaks have high levels of H3K4me1 but not H3K4me3

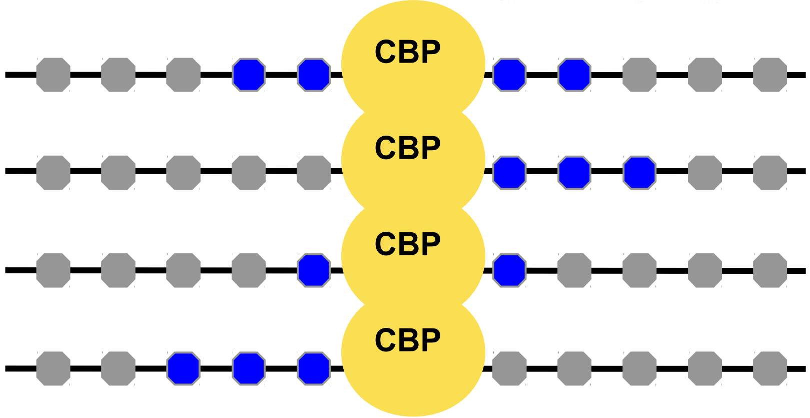




Transcription start sites (TSSs) have high levels of H3K4me1 and H3K4me3





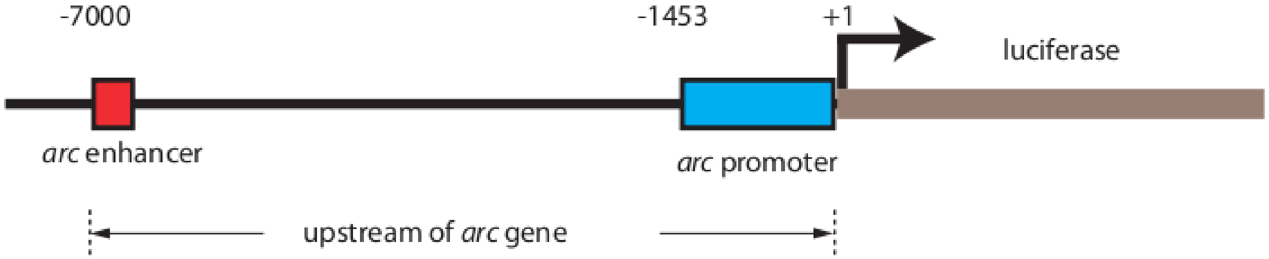


Identifying 5130 activity regulated enhancers

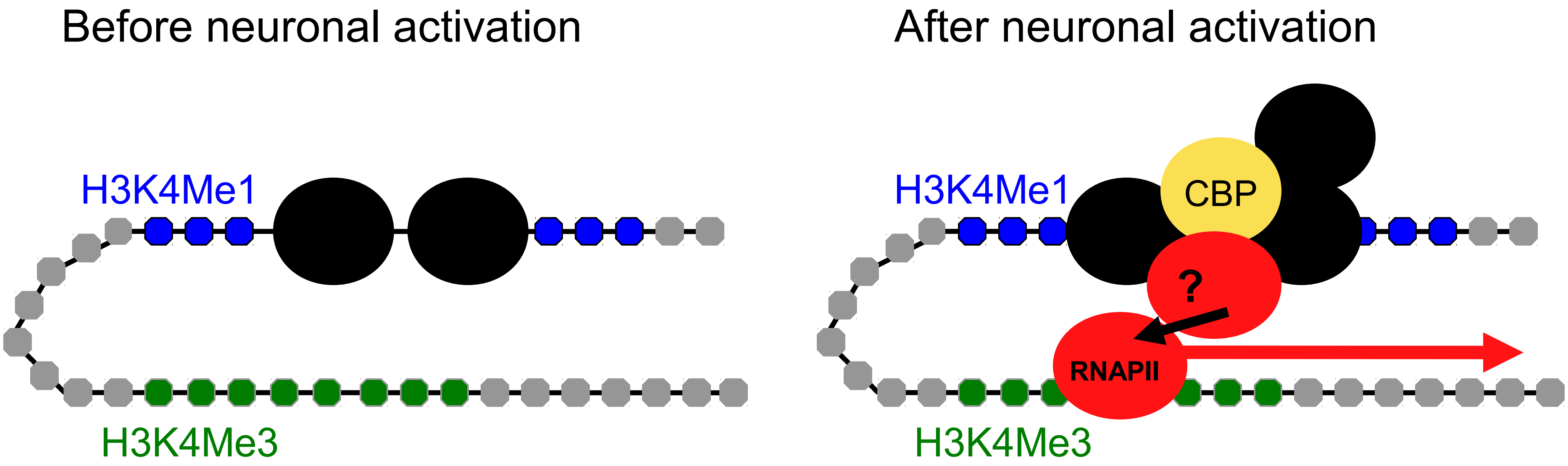
- CBP peak
- High levels of flanking H3K4me1
- Low levels of H3K4me3
- >1 kb from annotated promoter

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

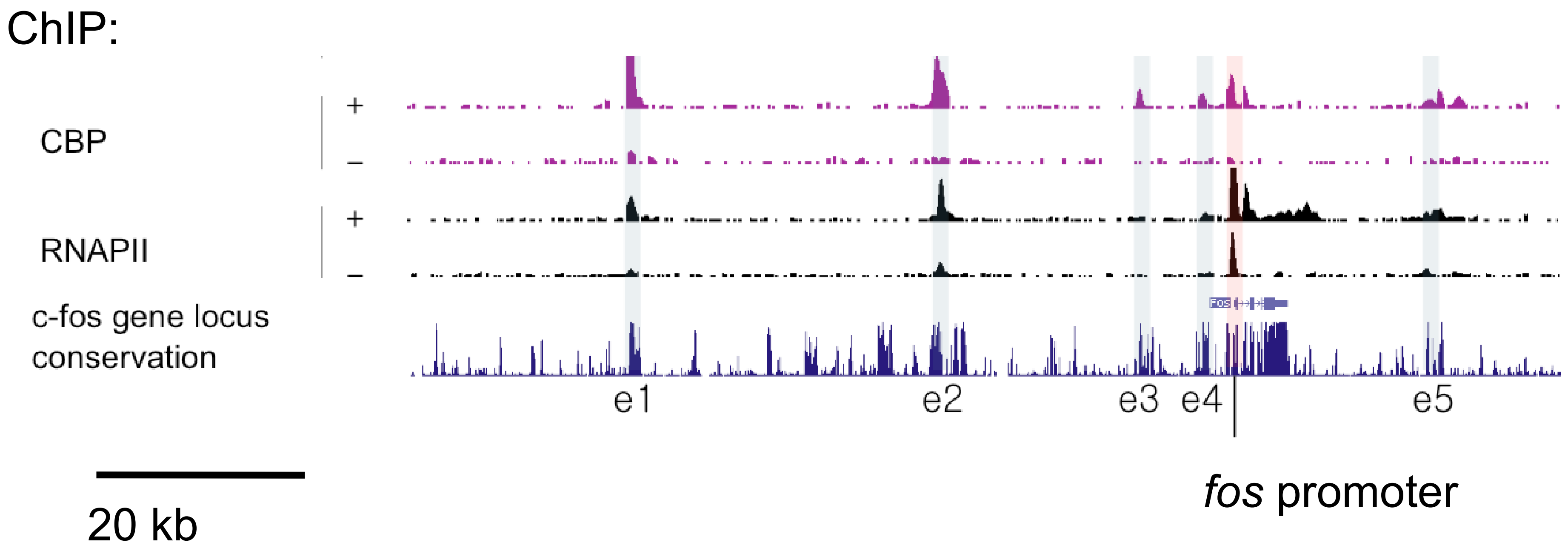


Before neuronal activation

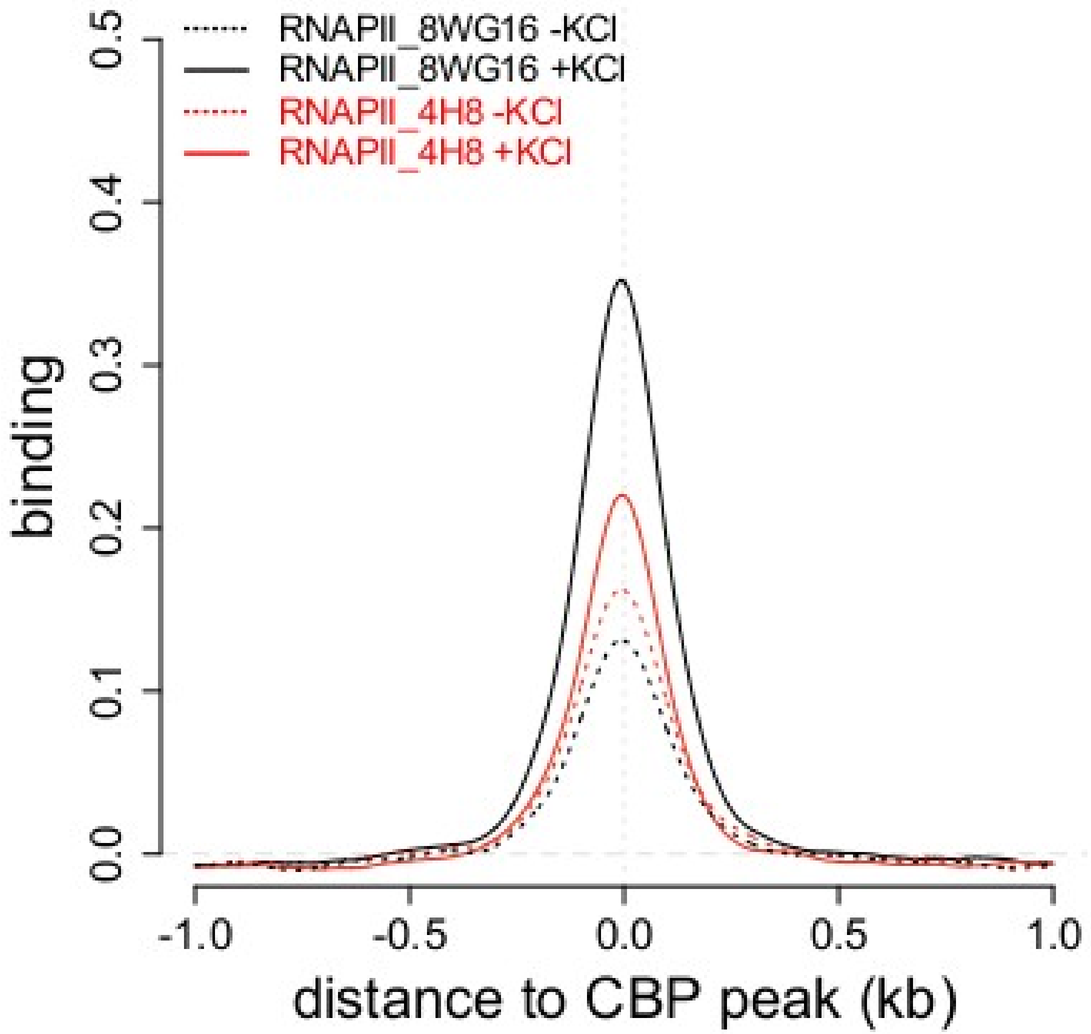
After neuronal activation

- Does RNAPII bind at enhancers?

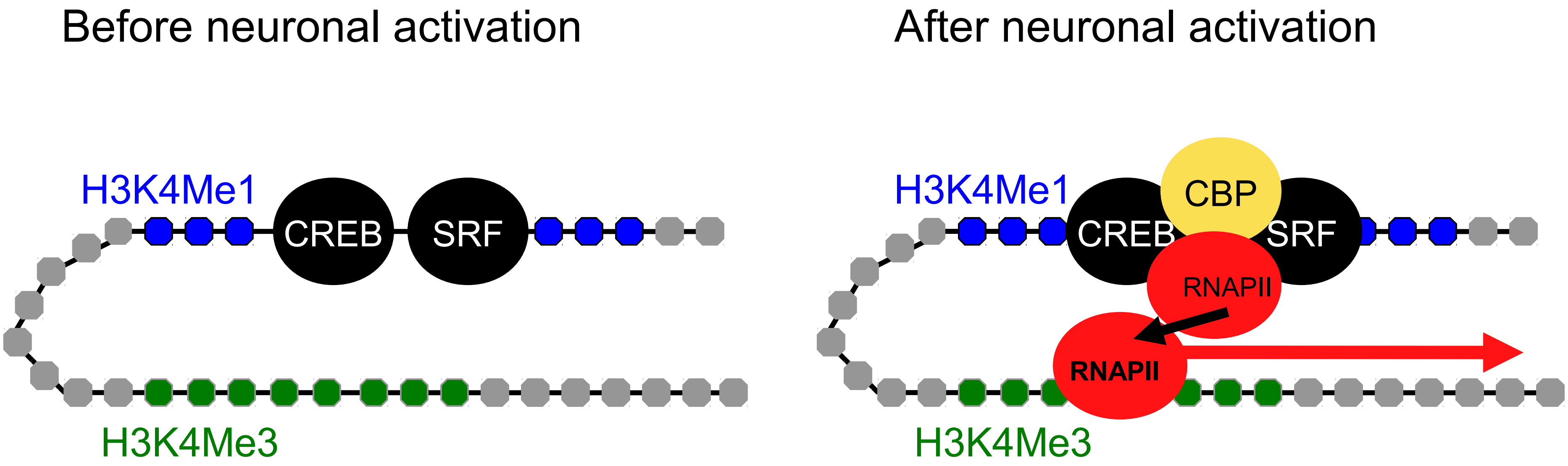
RNAPII is recruited to CBP binding sites at the *fos* locus



RNAPII is recruited at all enhancers



Properties of activity regulated enhancers

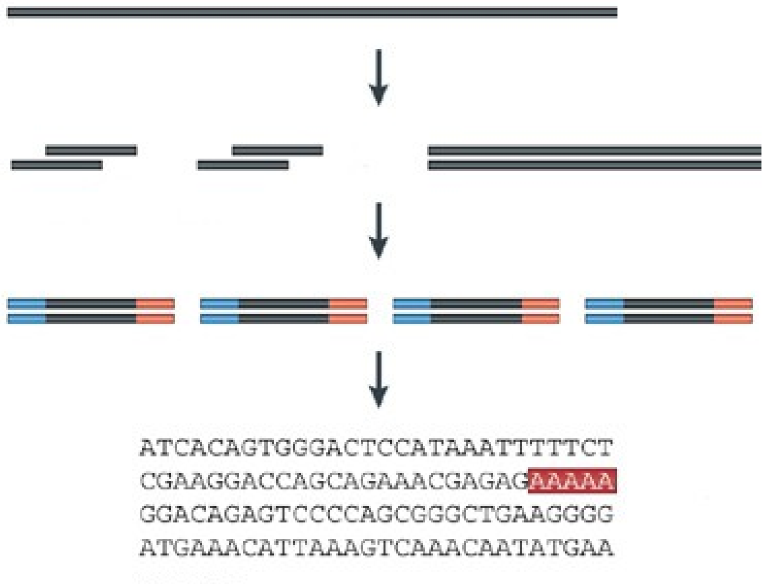


Before neuronal activation

After neuronal activation

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

RNA-Seq reveals which parts of the genome are transcribed

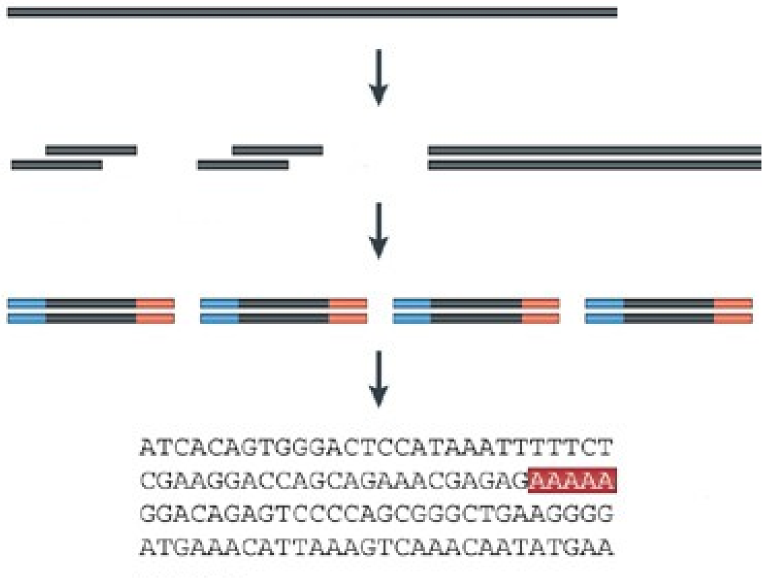

ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA

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

(Wang et al, 2009)

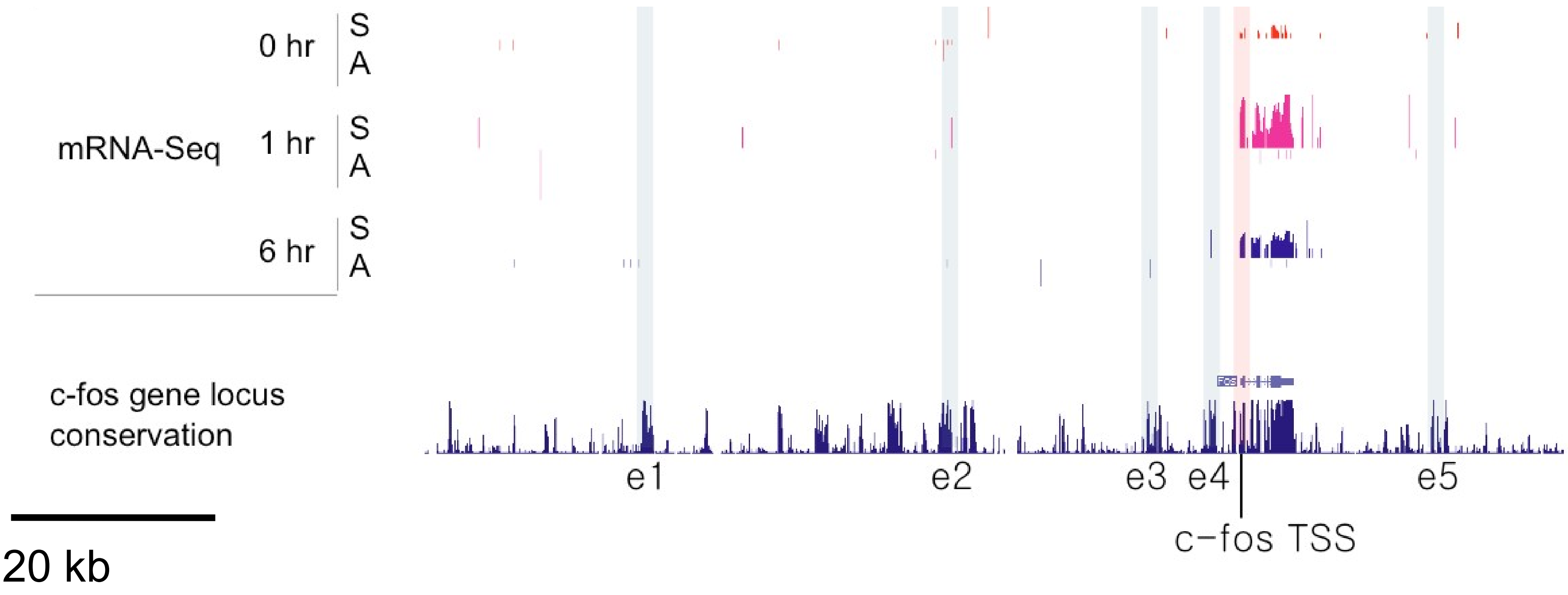
RNA-Seq reveals which parts of the genome are transcribed

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

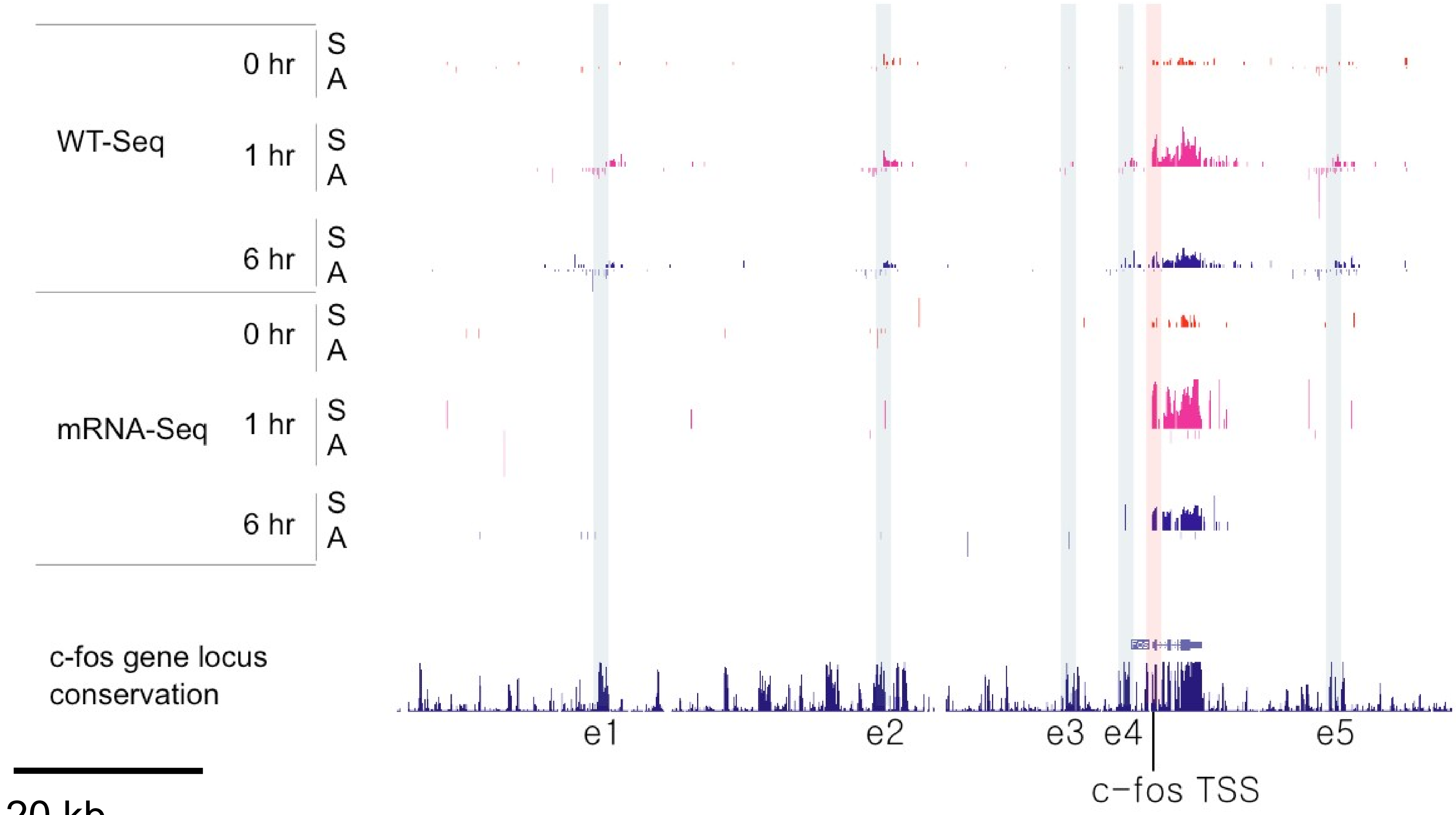

ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA

(Wang et al, 2009)

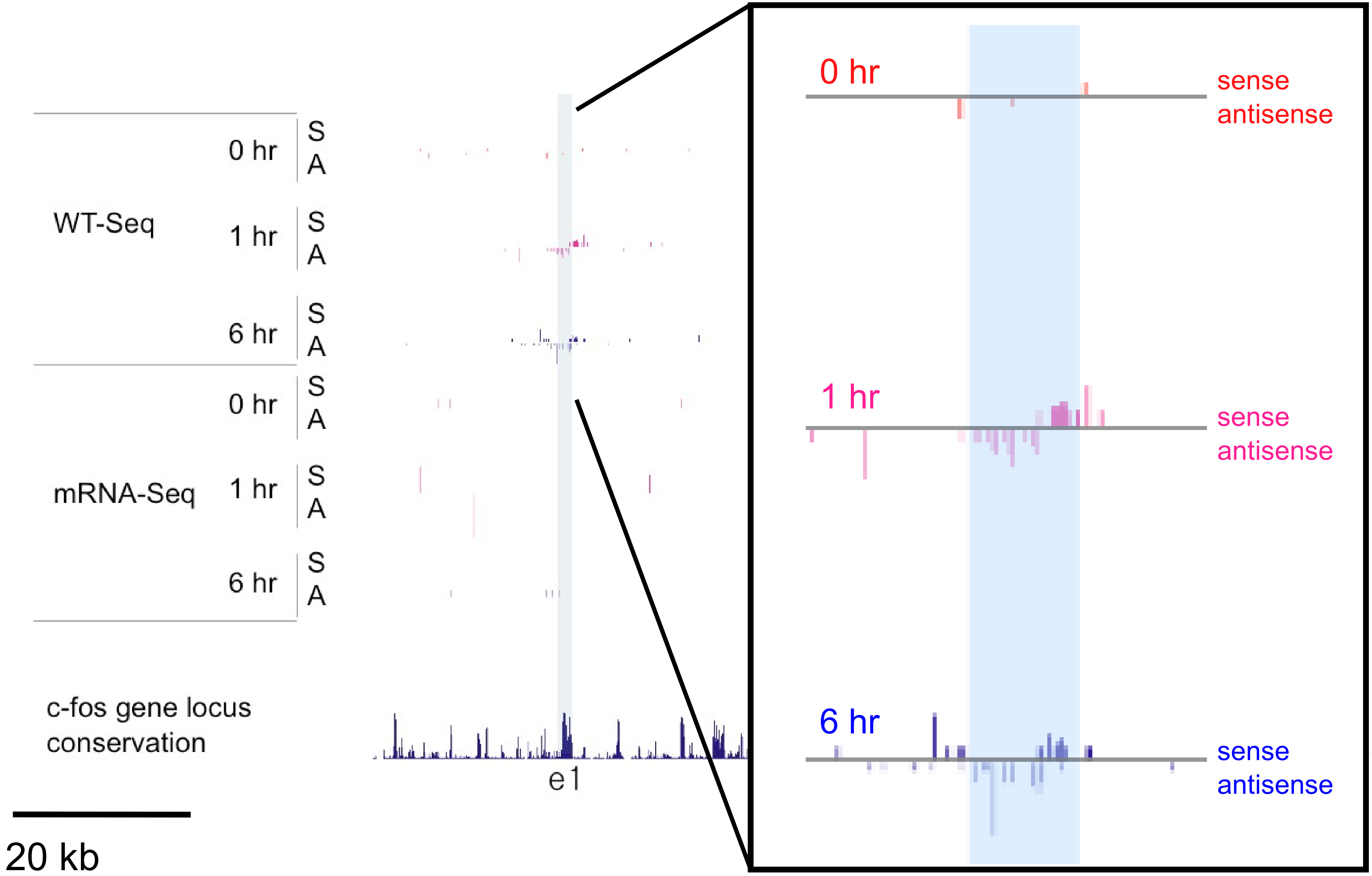
Transcription of enhancer RNA (eRNA) at the *fos* locus



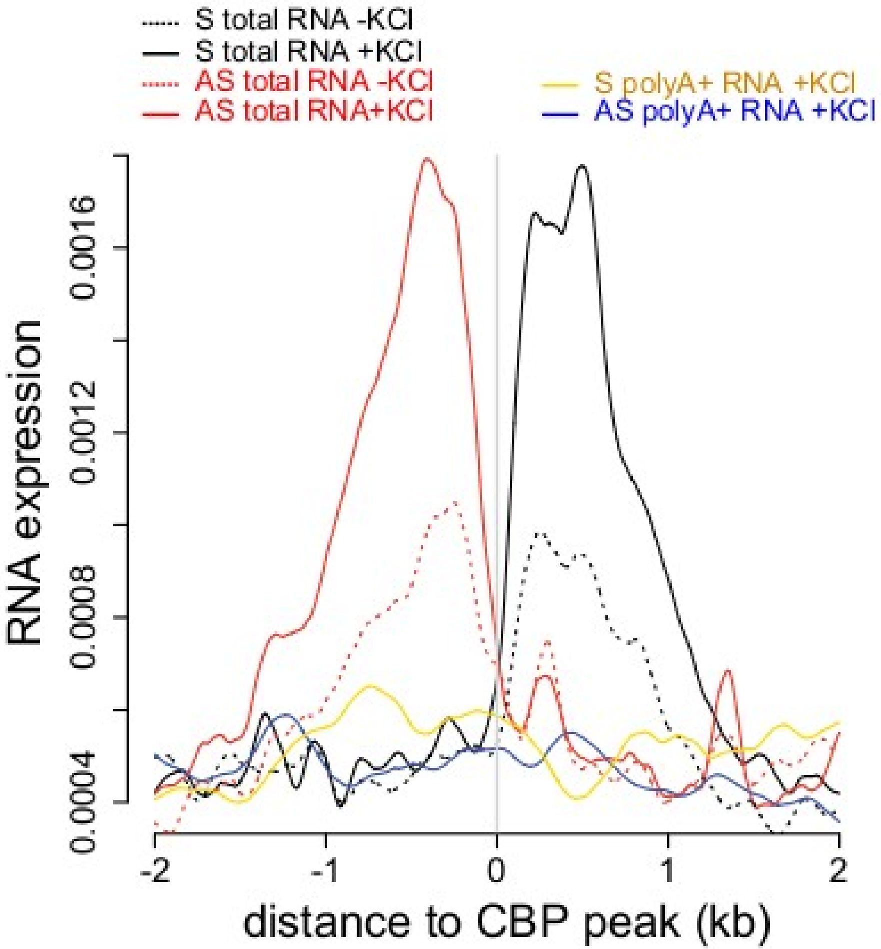
Transcription of enhancer RNA (eRNA) at the *fos* locus



Transcription of eRNA is activity-dependent

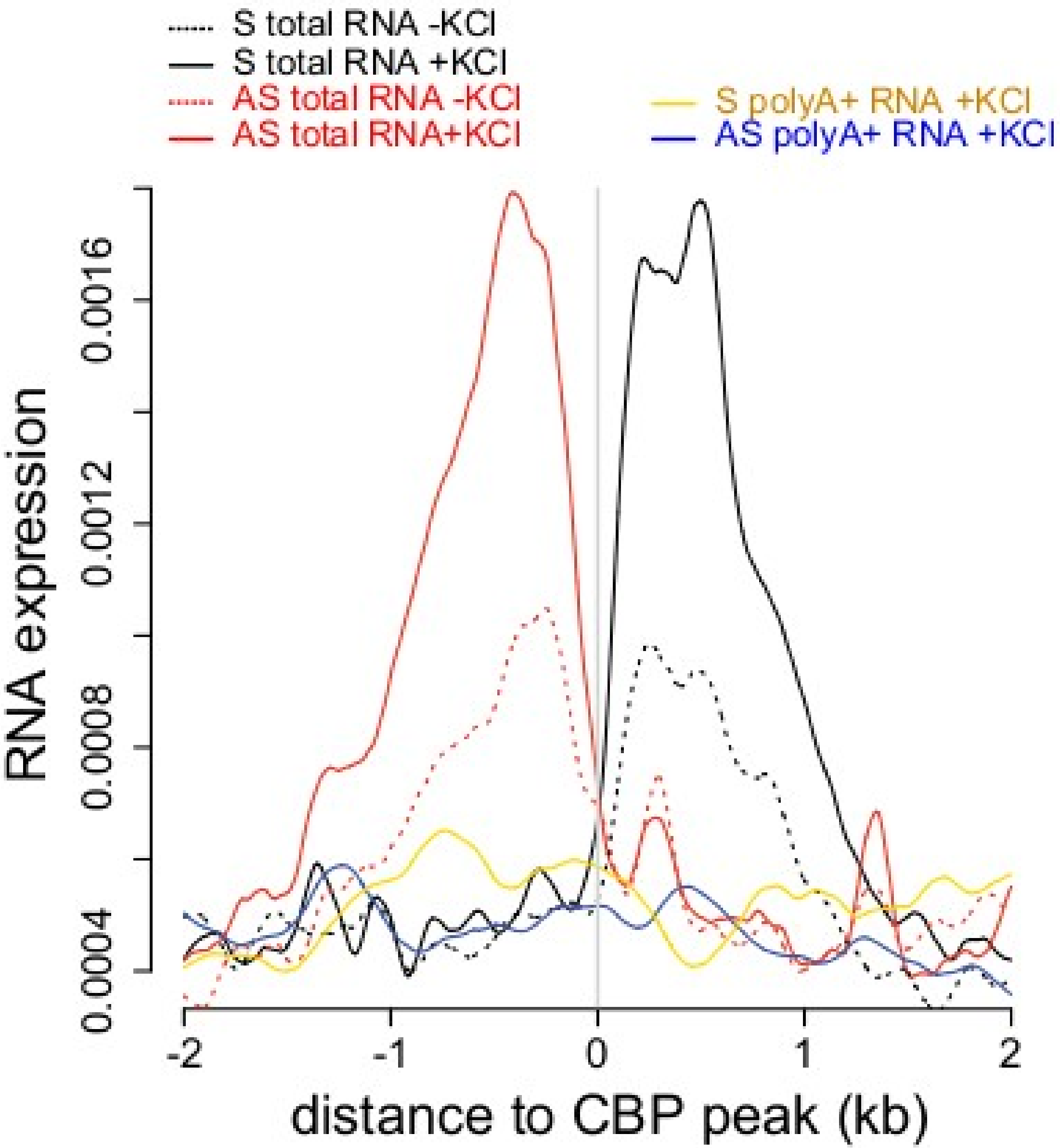


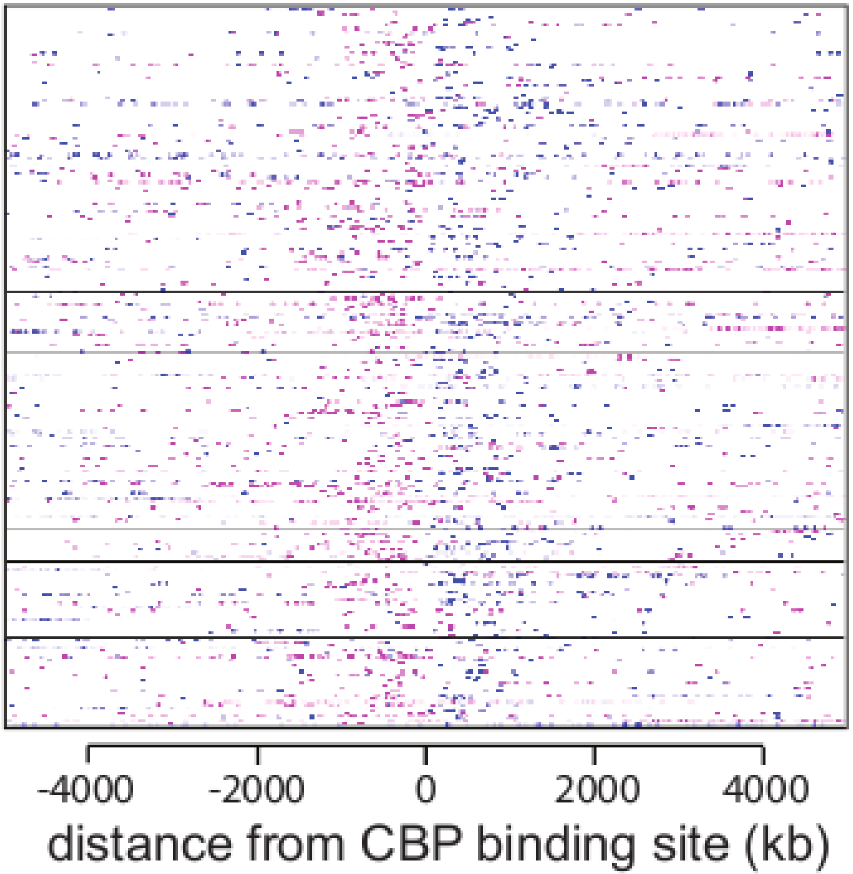
Genome-wide profile of transcription at enhancers



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

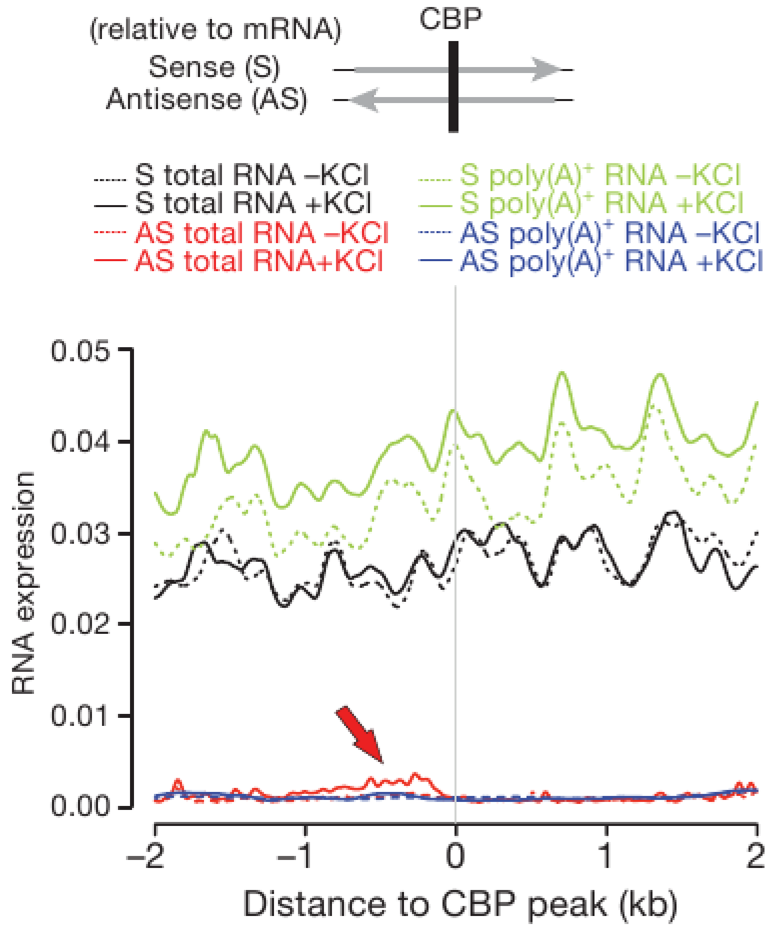
Genome-wide profile of transcription at enhancers



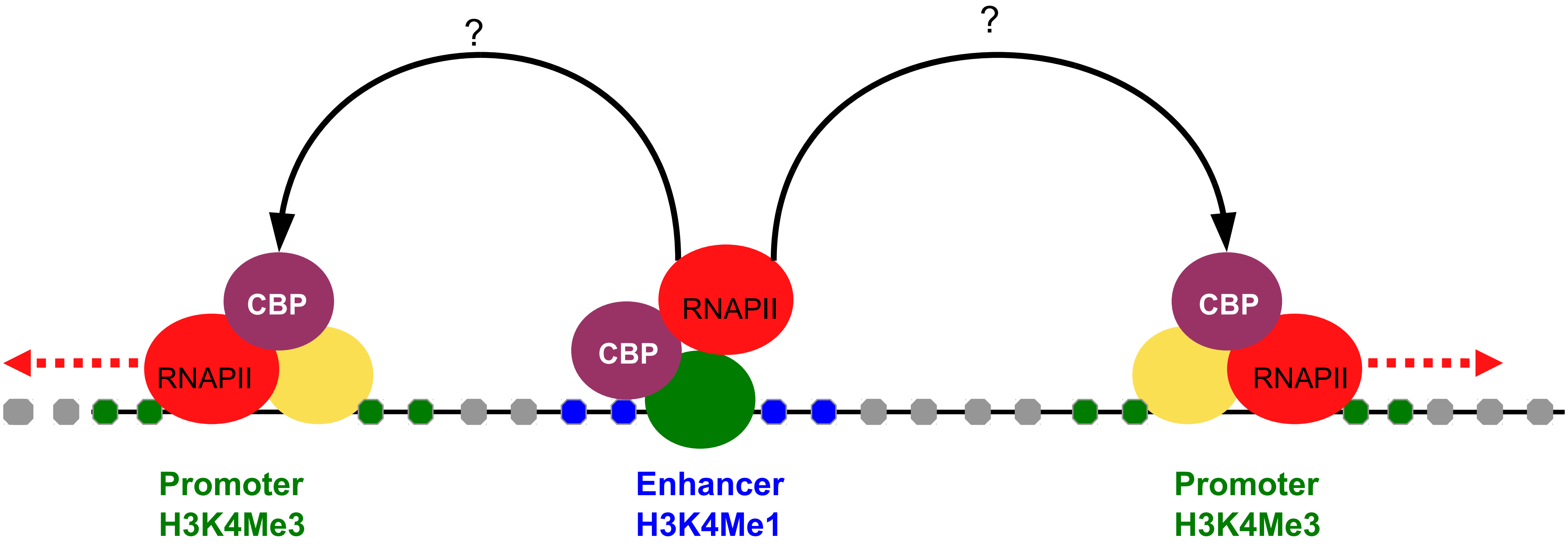


Intragenic enhancers are also transcribed

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

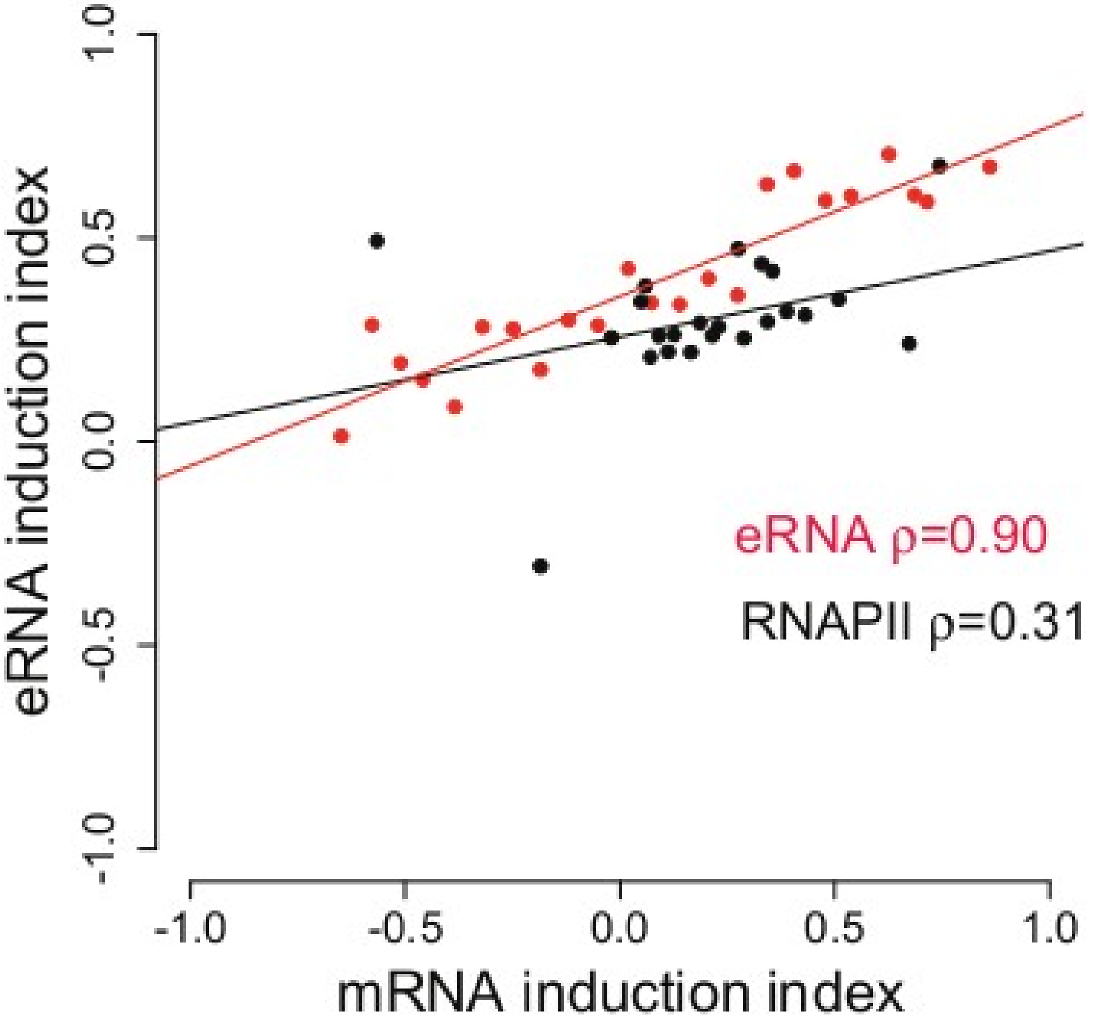


How do eRNA levels relate to mRNA levels?

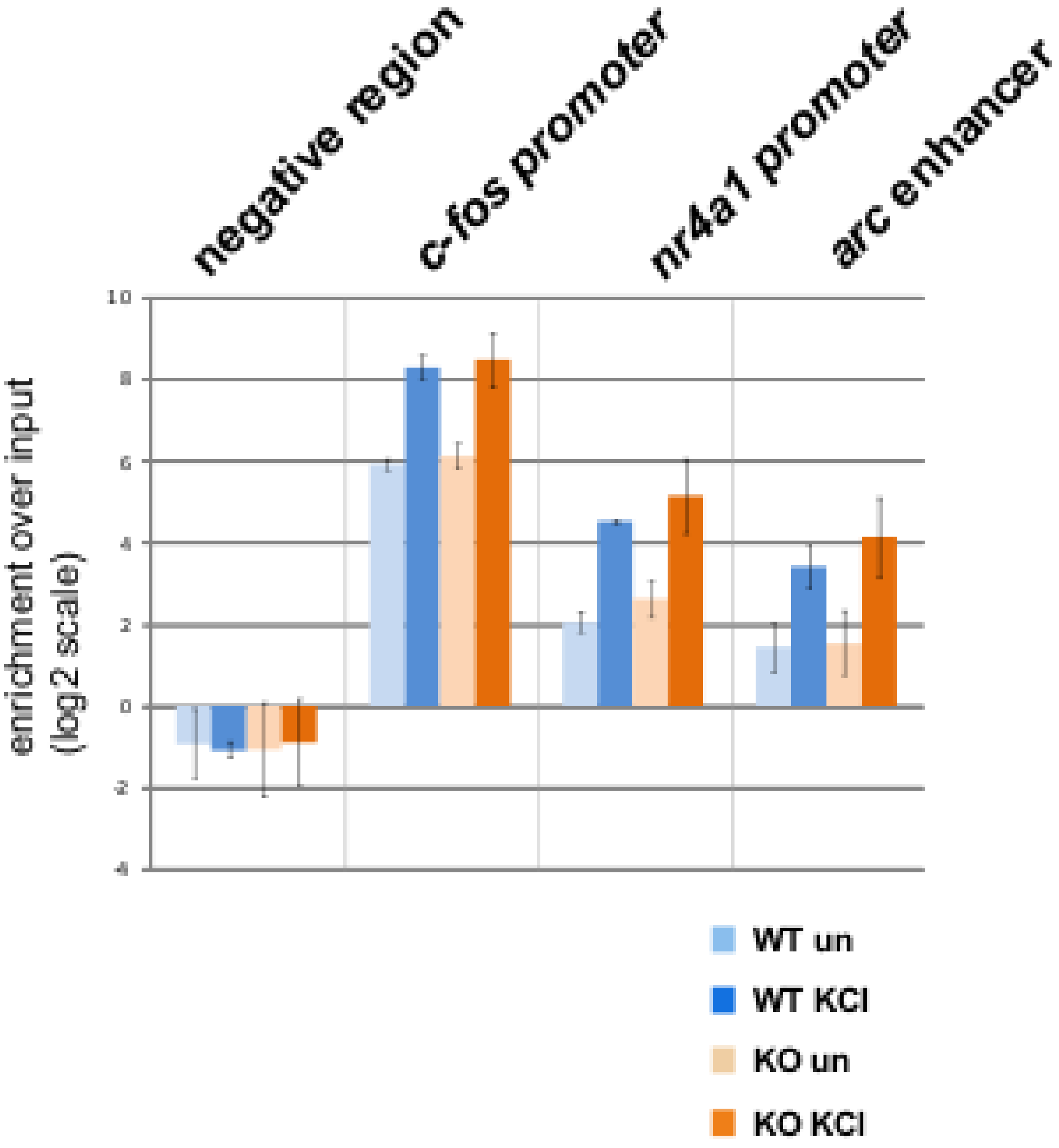


eRNA induction is correlated with induction of nearby mRNAs

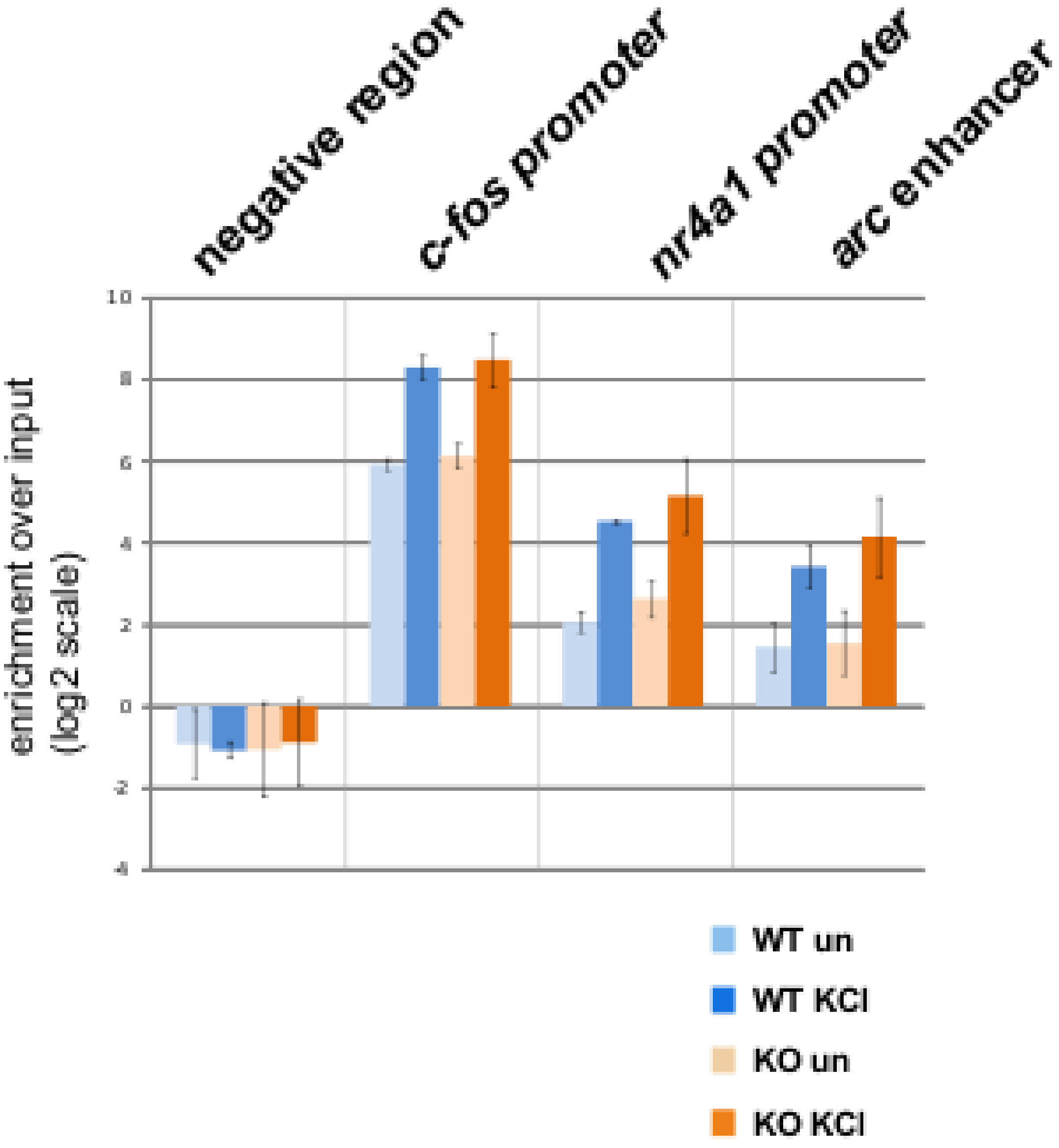
induction index = (KCl⁺ – KCl⁻)/(KCl⁺ + KCl⁻)

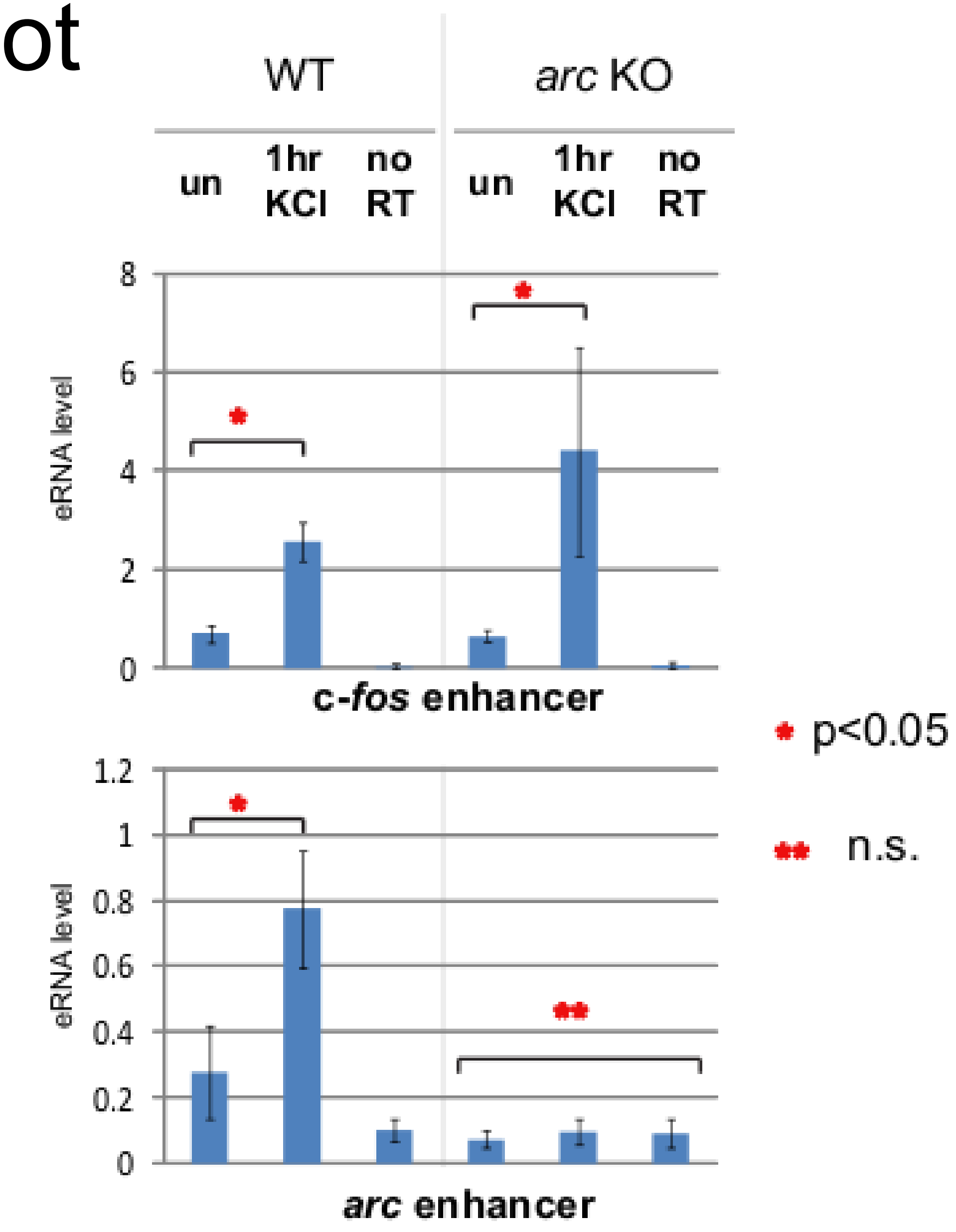


Knock-out experiment confirms that RNAPII recruitment is independent of the promoter

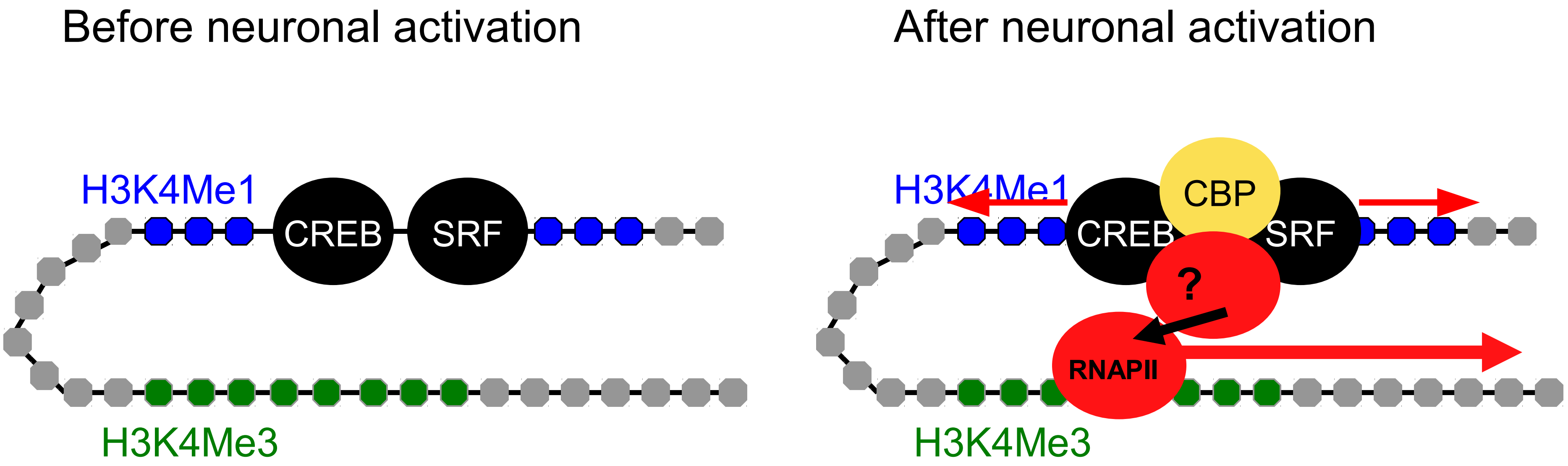


Knock-out experiment confirms that RNAPII recruitment is independent of the promoter but eRNA synthesis is not





Enhancers bind RNAPII independently, but the transcription is promoter-related

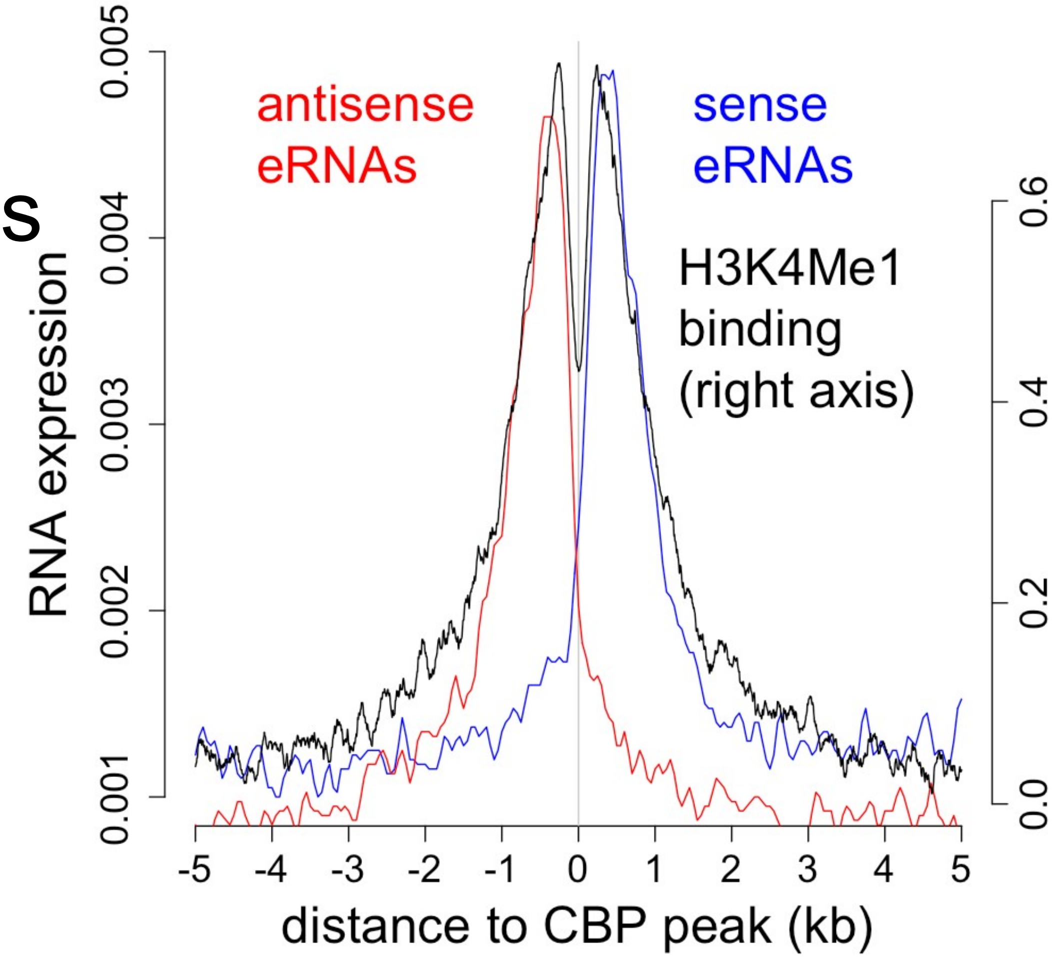


- Does RNAPII bind at enhancers? YES
- Are transcripts produced at enhancers? YES
- Is RNAPII recruitment independent? YES
- Is eRNA production independent? NO

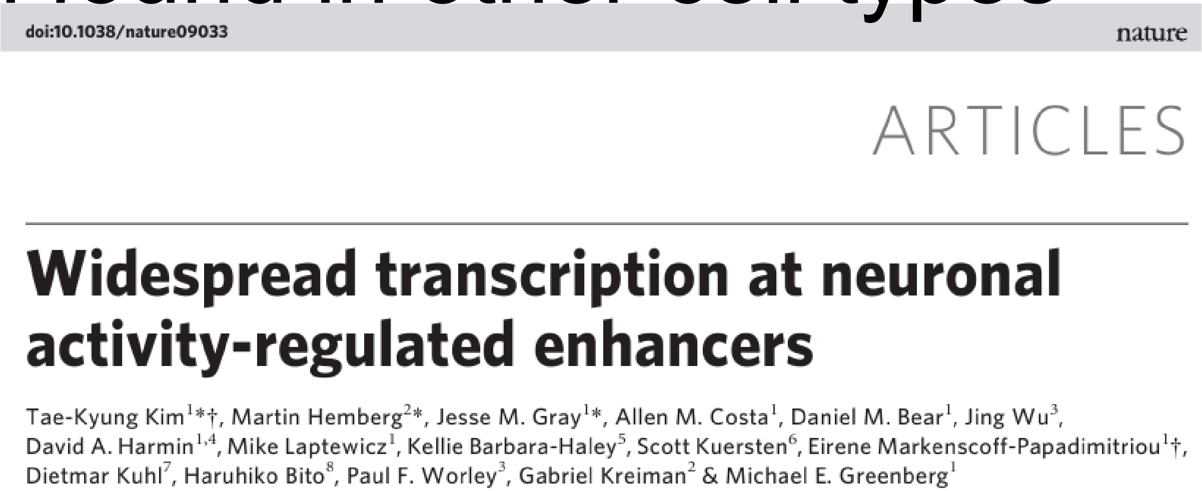
We have not yet been able to determine the function of eRNAs

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

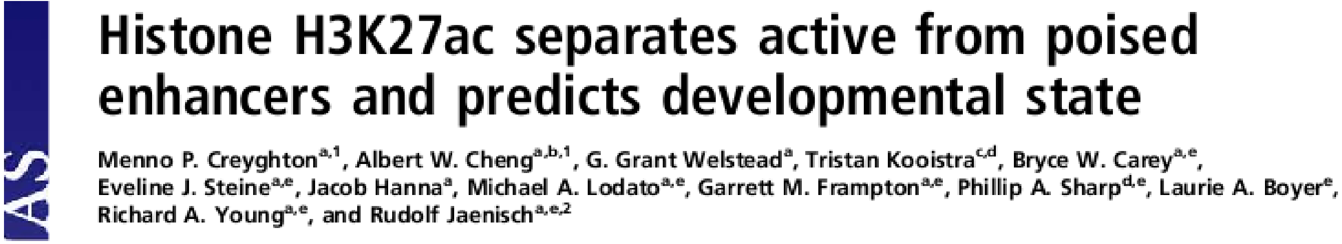


doi:10.1038/nature09033 nature

ARTICLES

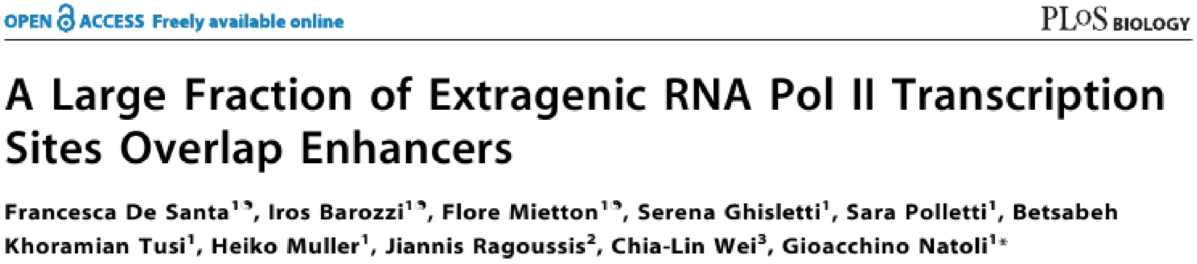
Widespread transcription at neuronal activity-regulated enhancers

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



Histone H3K27ac separates active from poised enhancers and predicts developmental state

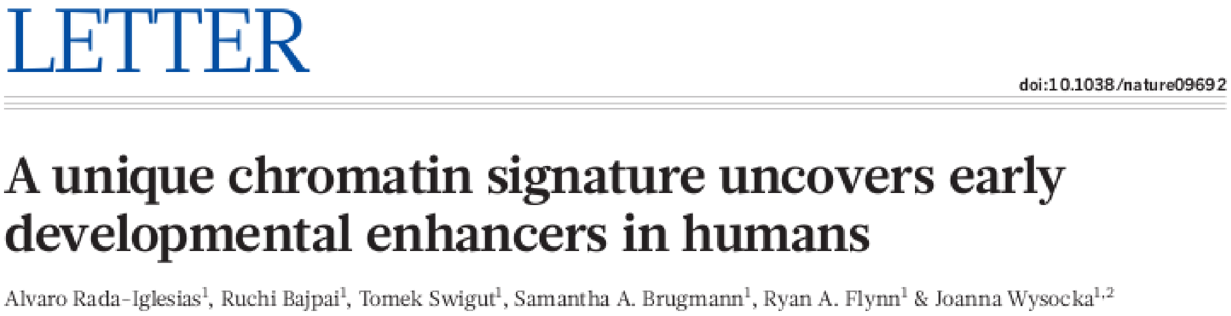
Menno P. Creyghton[a,1], Albert W. Cheng[a,b,1], G. Grant Welstead[a], Tristan Kooistra[c,d], Bryce W. Carey[a,e], Eveline J. Steine[a,e], Jacob Hanna[a], Michael A. Lodato[a,e], Garrett M. Frampton[a,e], Phillip A. Sharp[d,e], Laurie A. Boyer[e], Richard A. Young[a,e], and Rudolf Jaenisch[a,e,2]



OPEN ACCESS Freely available online PLoS BIOLOGY

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

Francesca De Santa[1¶], Iros Barozzi[1¶], Flore Mietton[1¶], Serena Ghisletti[1], Sara Polletti[1], Betsabeh Khoramian Tusi[1], Heiko Muller[1], Jiannis Ragoussis[2], Chia-Lin Wei[3], Gioacchino Natoli[1*]



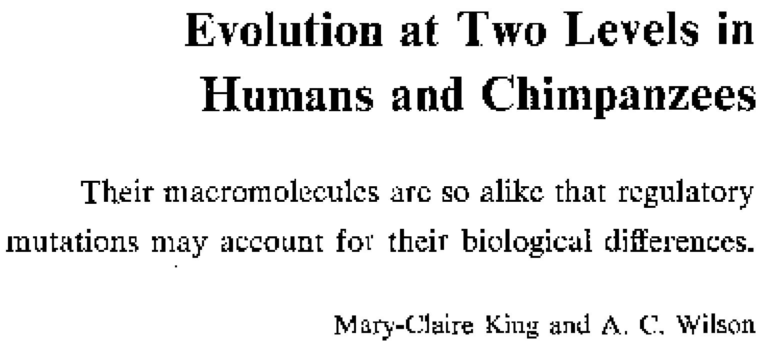
LETTER

doi:10.1038/nature09692

A unique chromatin signature uncovers early developmental enhancers in humans

Alvaro Rada-Iglesias[1], Ruchi Bajpai[1], Tomek Swigut[1], Samantha A. Brugmann[1], Ryan A. Flynn[1] & Joanna Wysocka[1,2]

What is the function of conserved non-coding sequences?

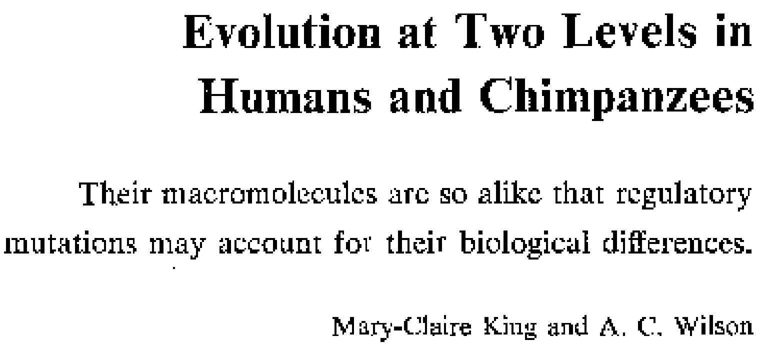


Evolution at Two Levels in Humans and Chimpanzees

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

Mary-Claire King and A. C. Wilson

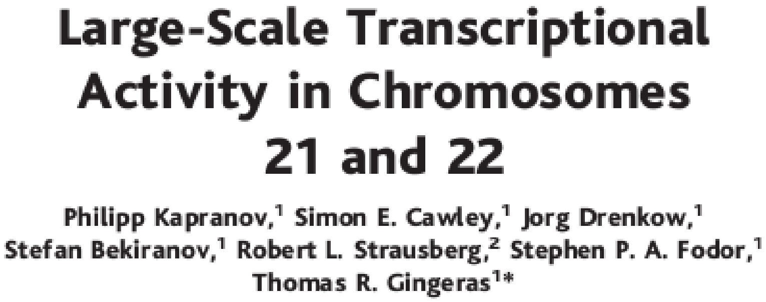
What is the function of conserved non-coding sequences?



Evolution at Two Levels in Humans and Chimpanzees

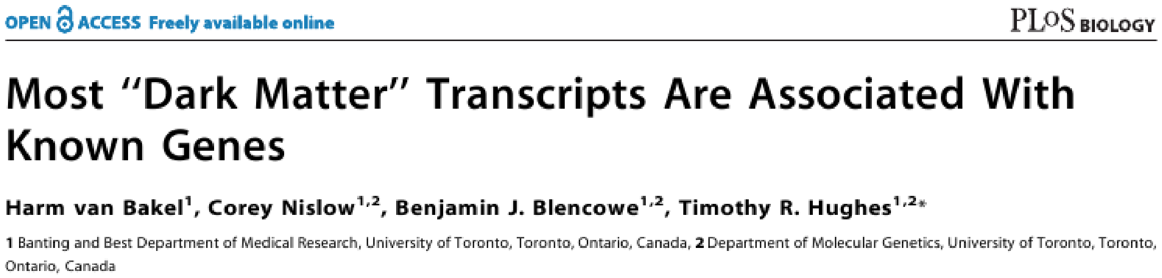
Their macromolecules are so alike that regulatory mutations may account for their biological differences.

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

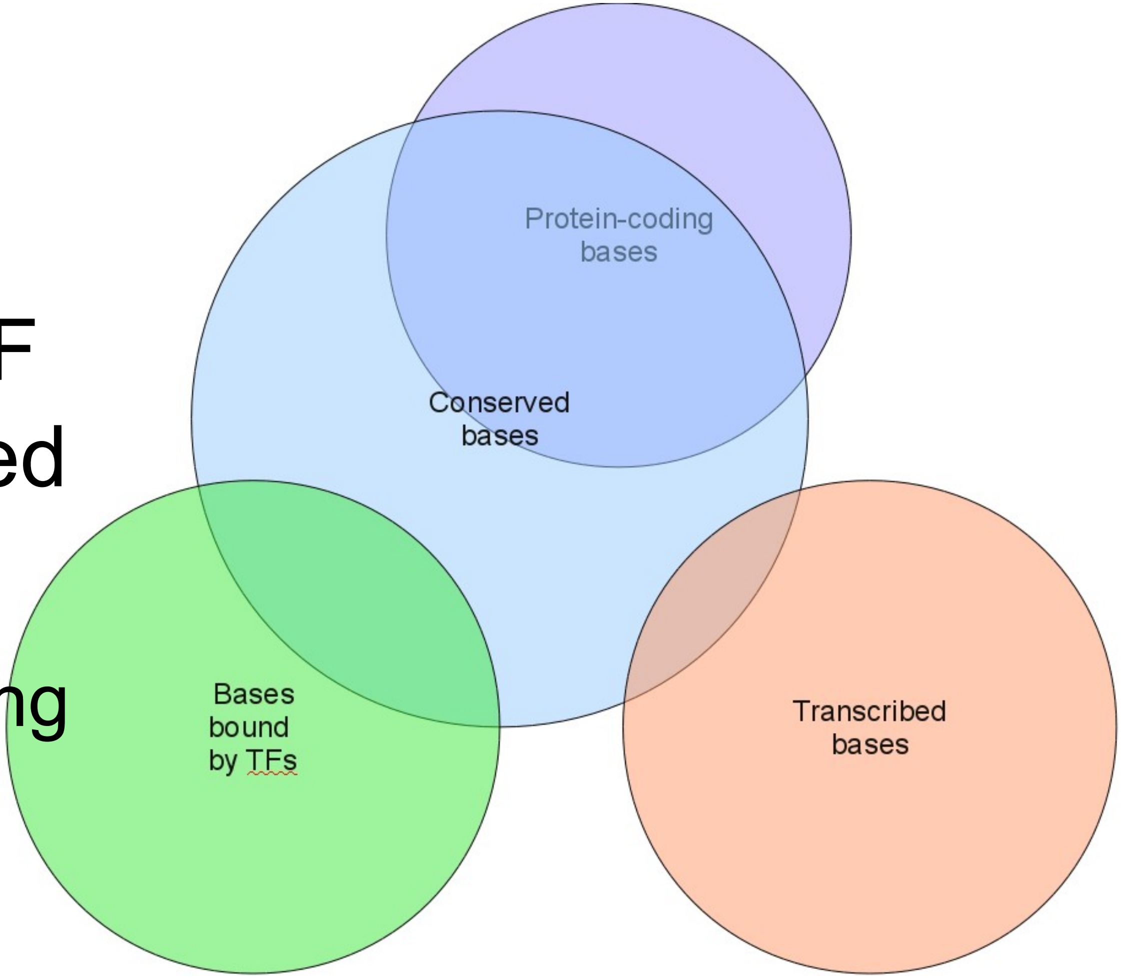
Most “Dark Matter” Transcripts Are Associated With Known Genes

Harm van Bakel[1], 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

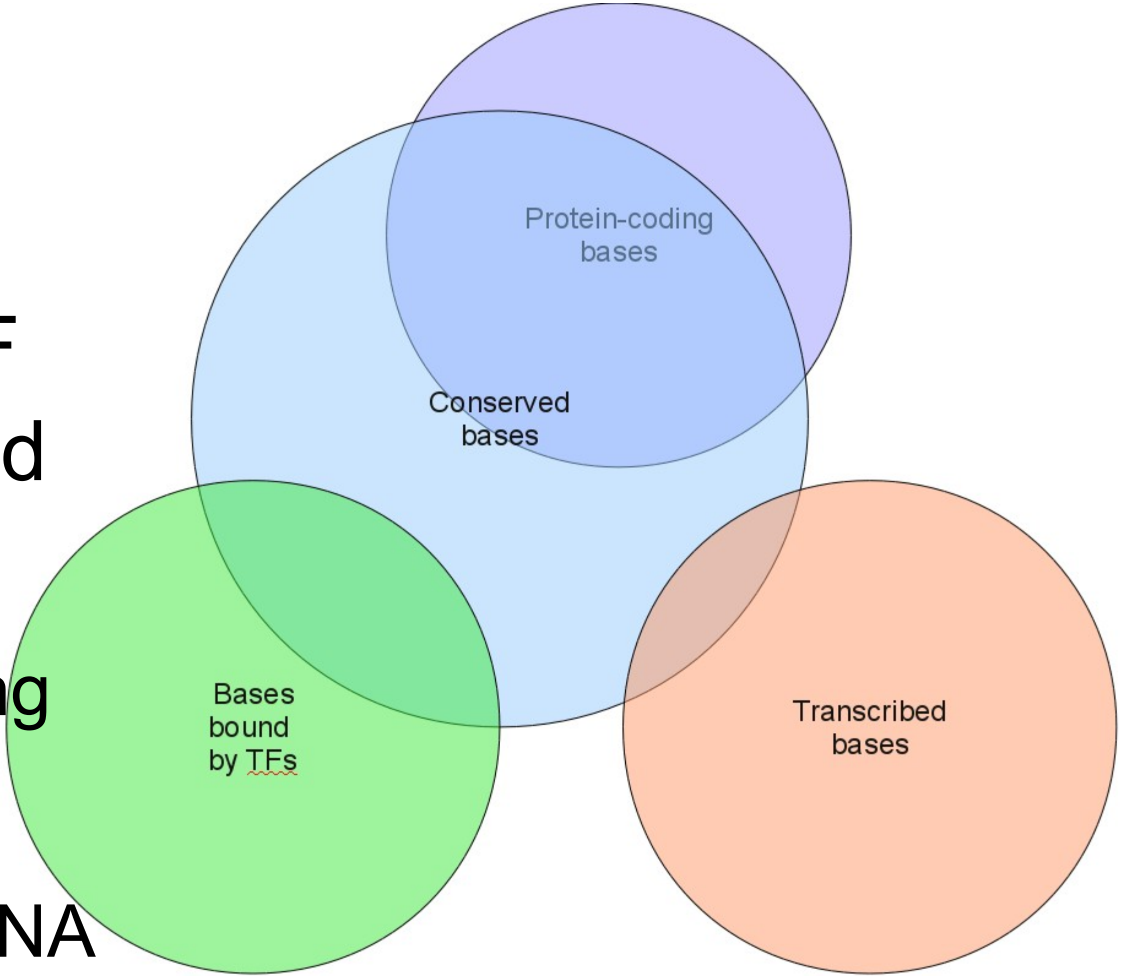
What determines the conservation of extragenic regions?

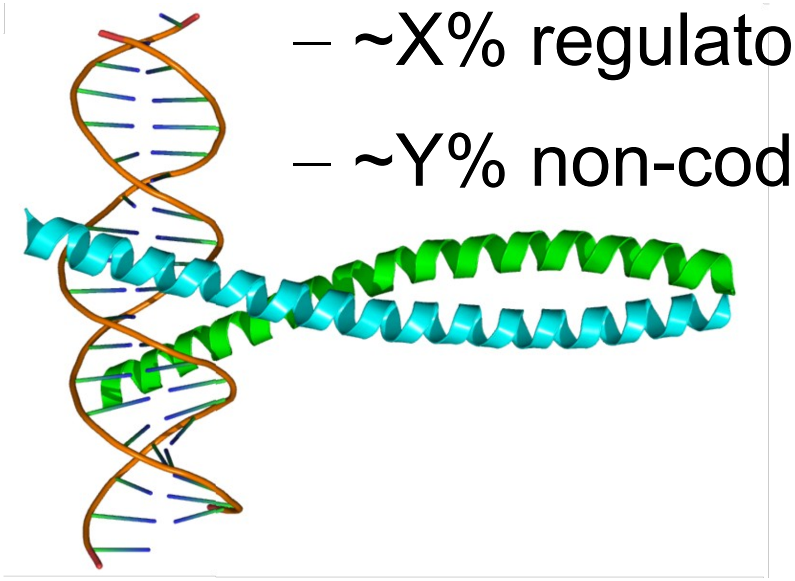
- Compare extragenic transcription and TF binding to conserved bases
  - ~40% protein coding

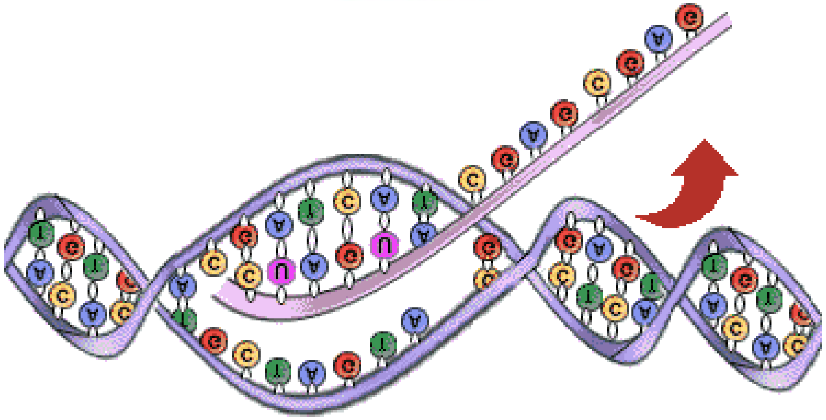


What determines the conservation of extragenic regions?

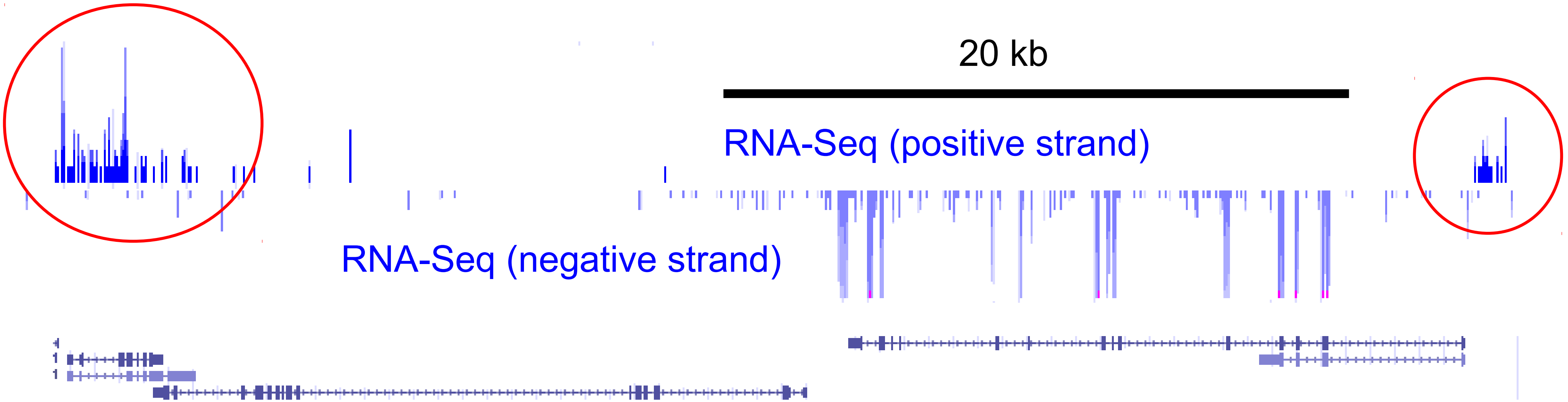
- Compare extragenic transcription and TF binding to conserved bases
  - ~40% protein coding
  - ~X% regulatory
  - ~Y% non-coding RNA





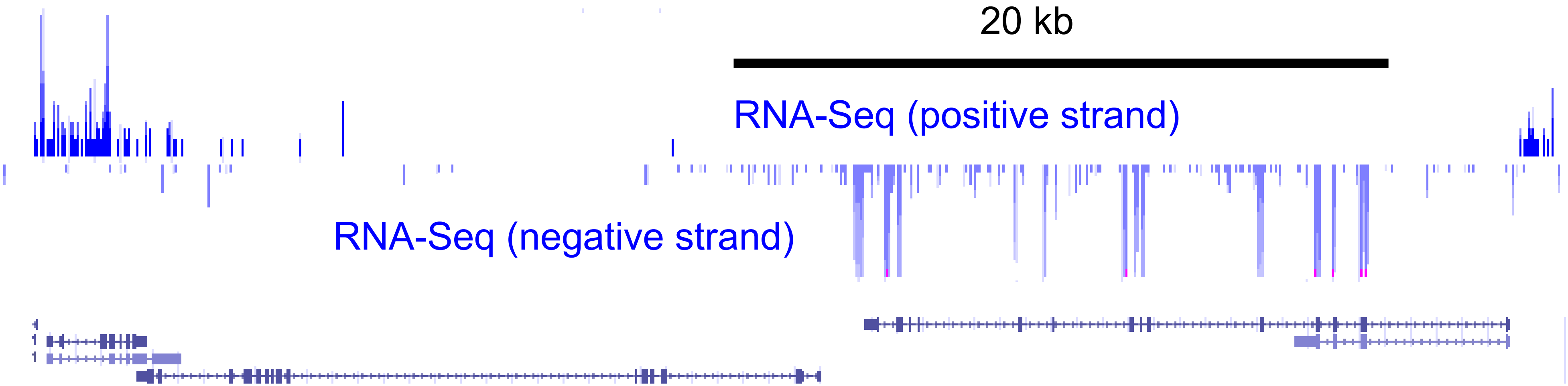


De novo identification of transcribed regions



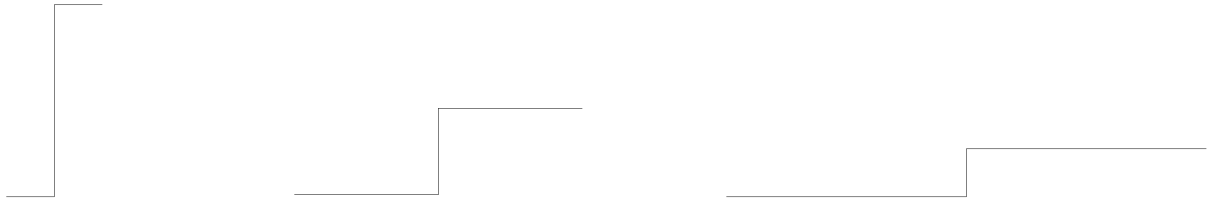
Using Haar-wavelets to identify transcribed regions (HaTriC)

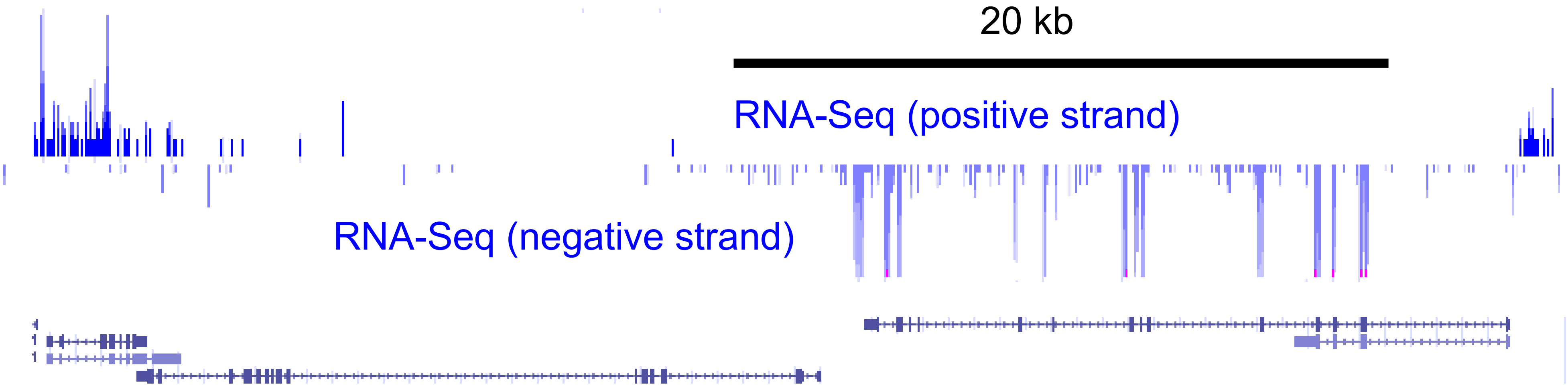
- Find where read-density changes abruptly



Using Haar-wavelets to identify transcribed regions (HaTriC)

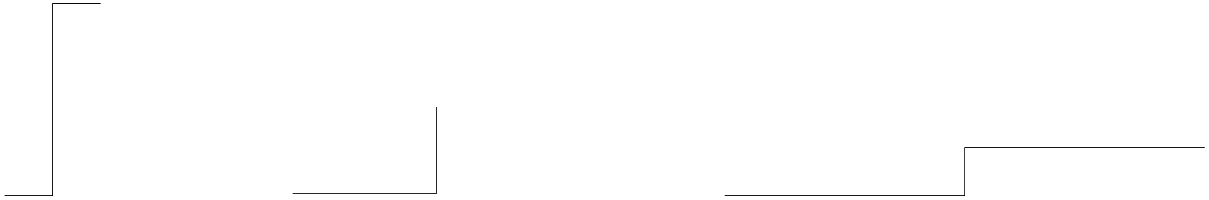
- Find where read-density changes abruptly
  - Consider multiple length scales



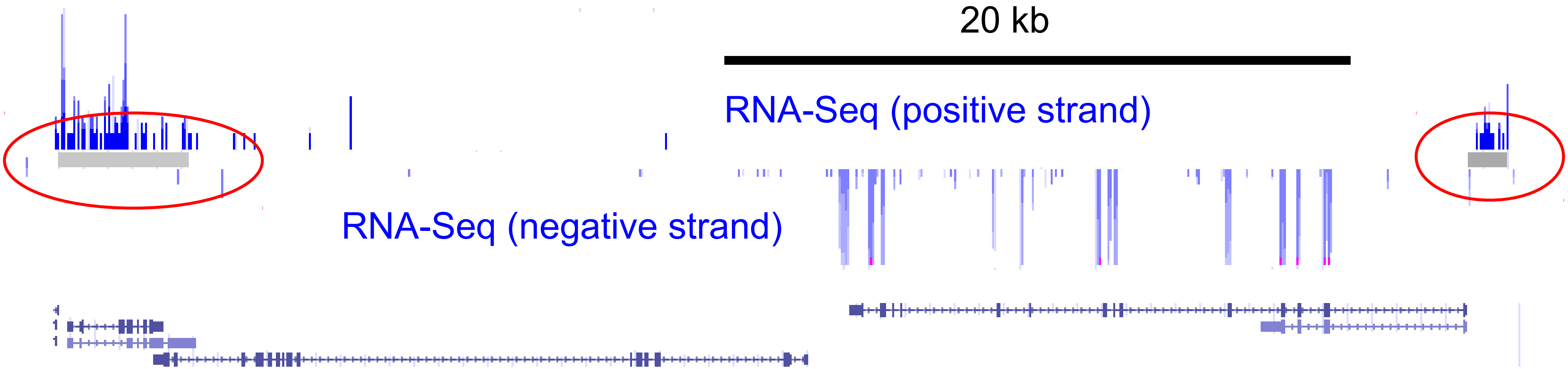


Using Haar-wavelets to identify transcribed regions (HaTriC)

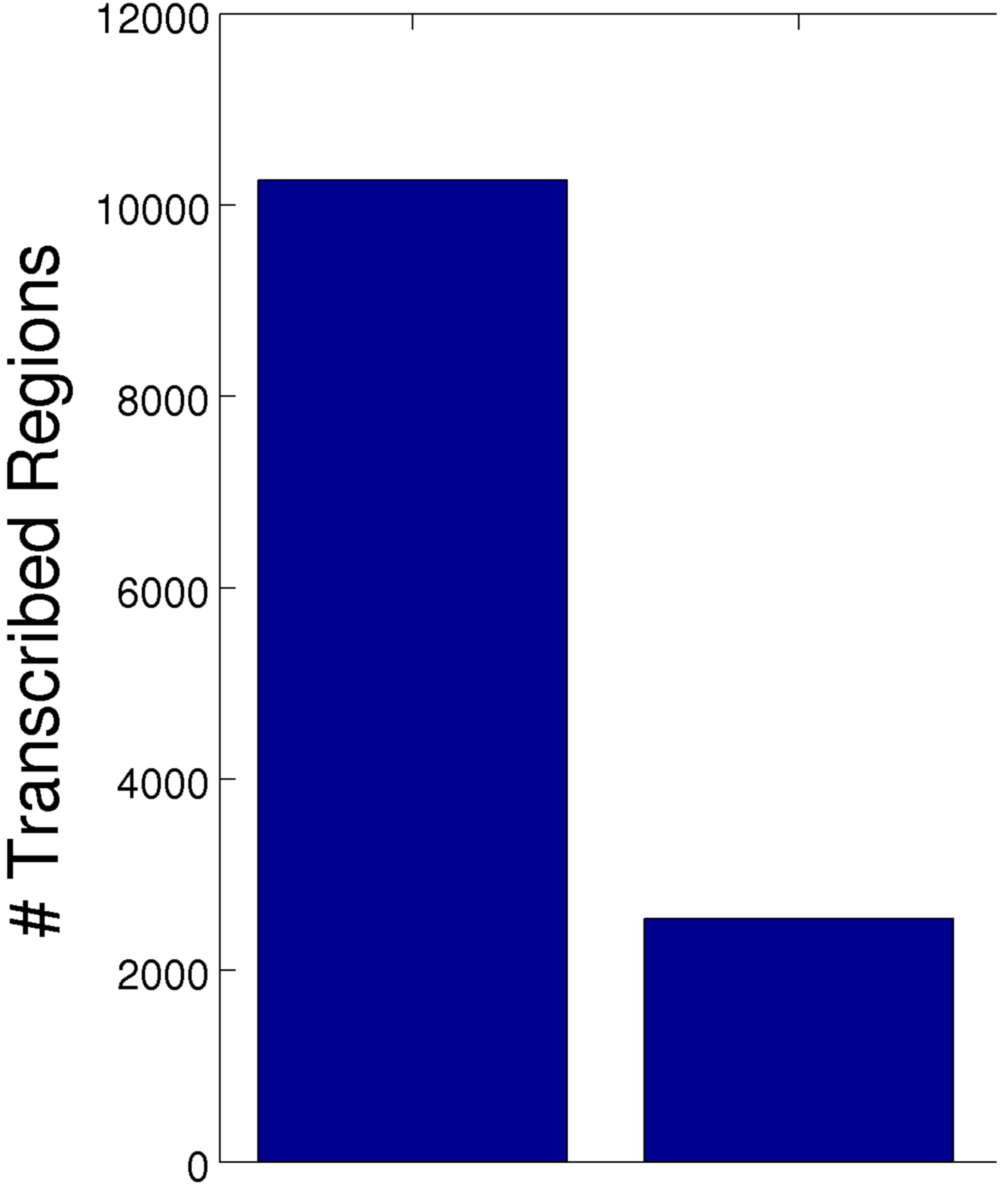
- Find where read-density changes abruptly
  - Consider multiple length scales



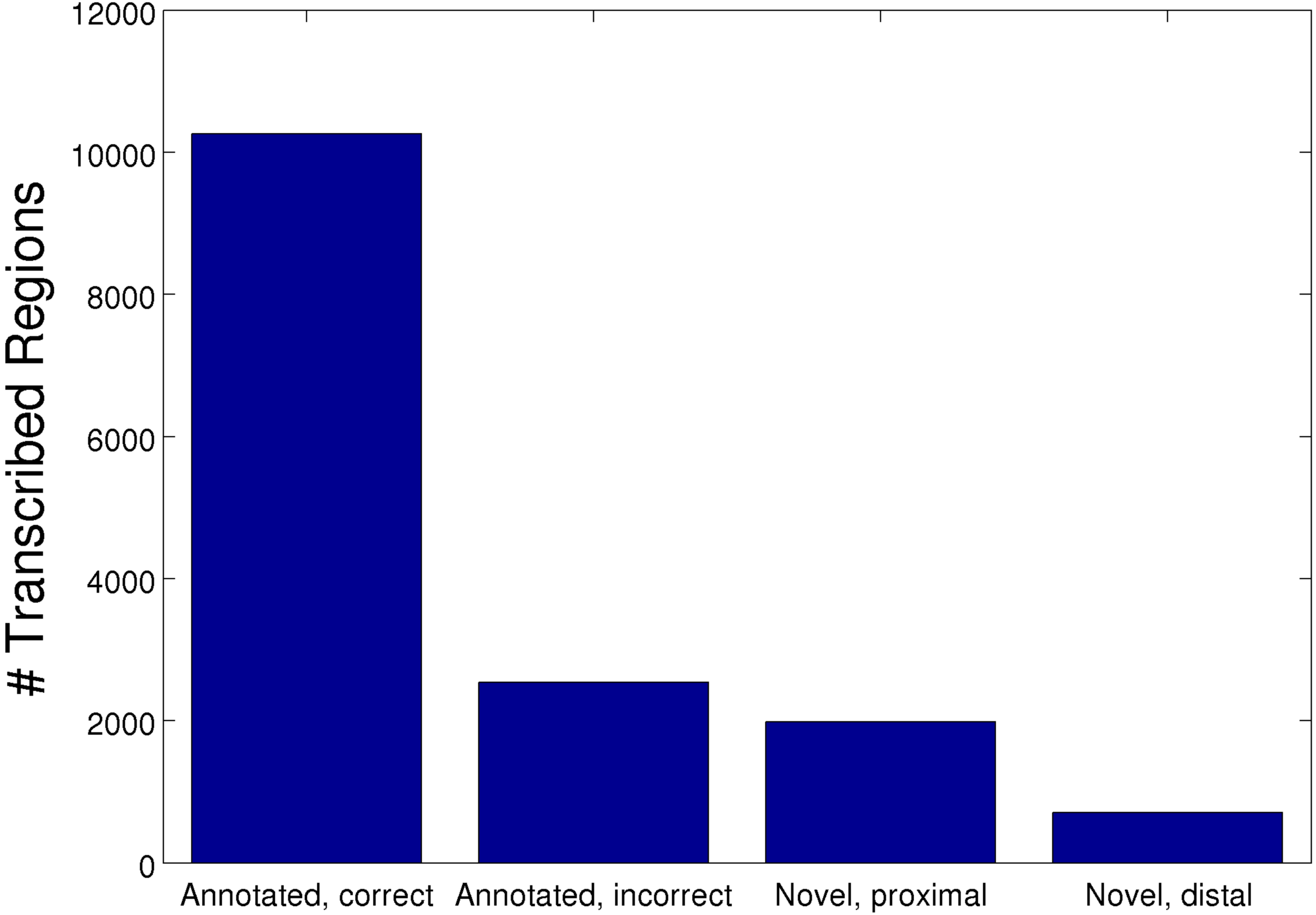
- Interleaving regions of high/low density



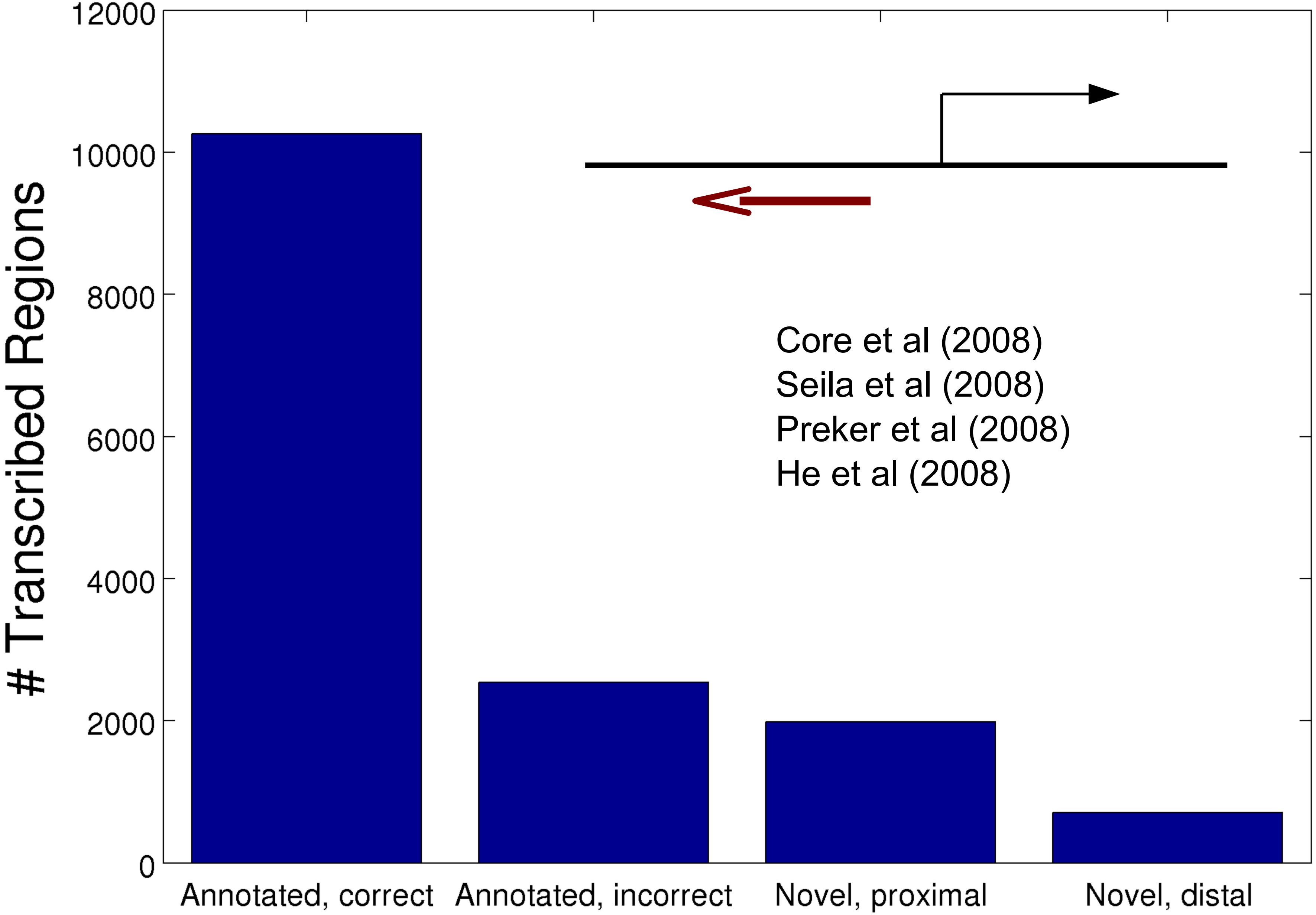
Most annotated genes and ncRNAs are correctly identified



HaTriC accounts for 92% of reads outside repeat regions

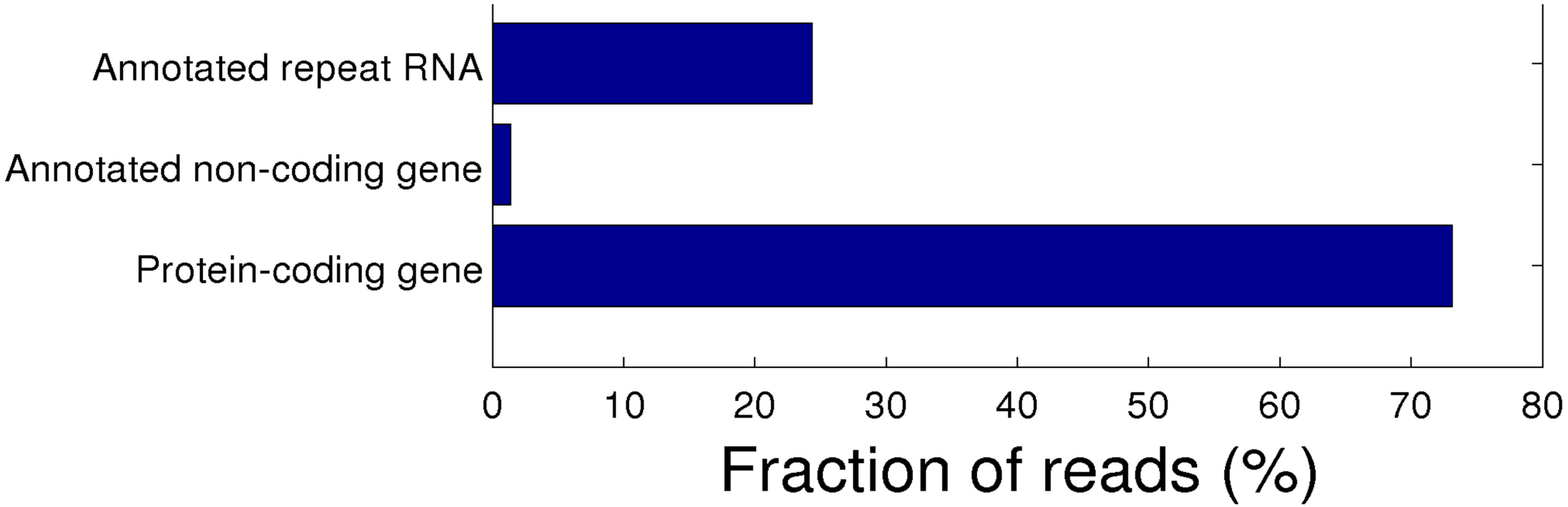


Most unannotated transcribed regions are promoter divergent anti-sense

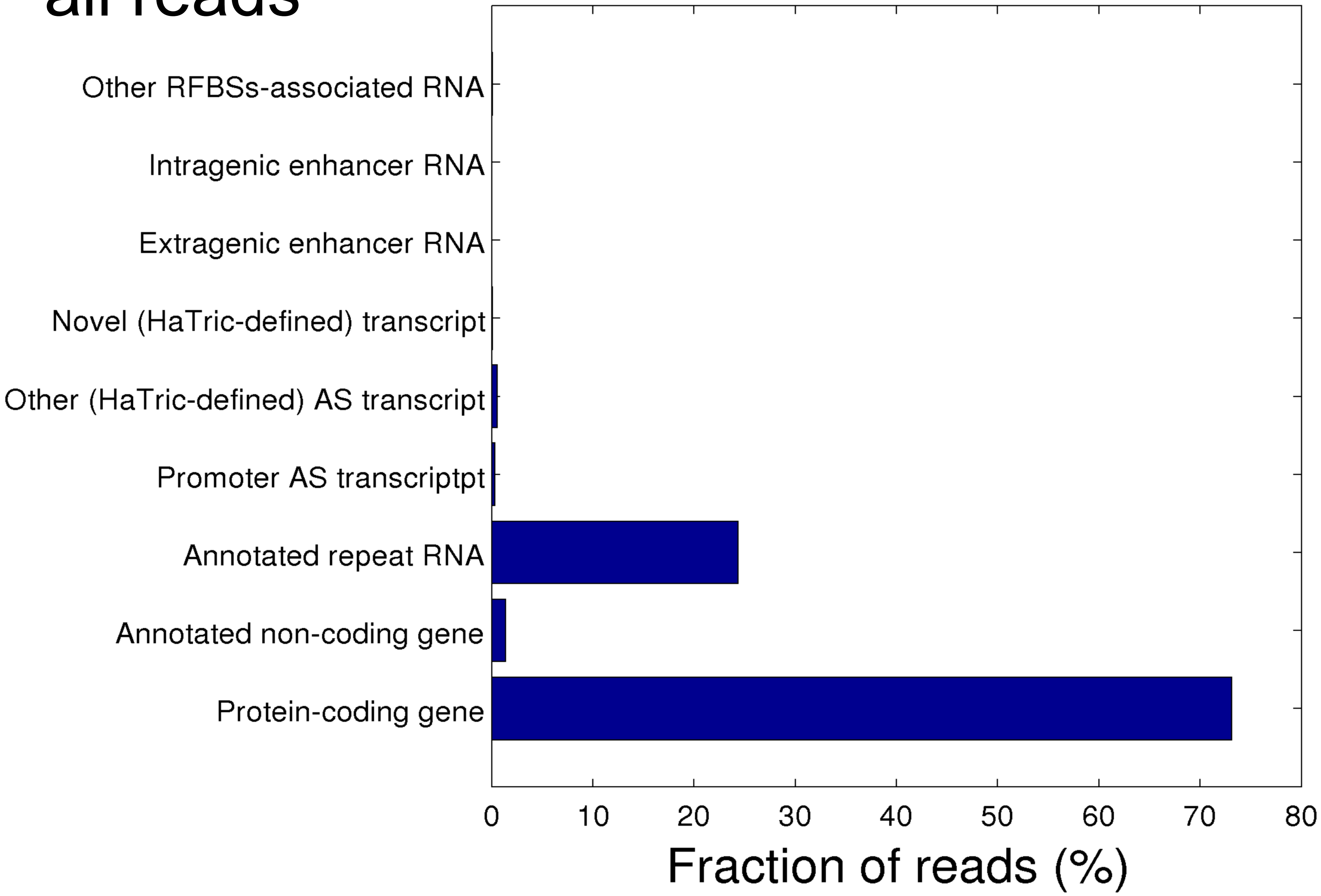


Core et al (2008)
Seila et al (2008)
Preker et al (2008)
He et al (2008)

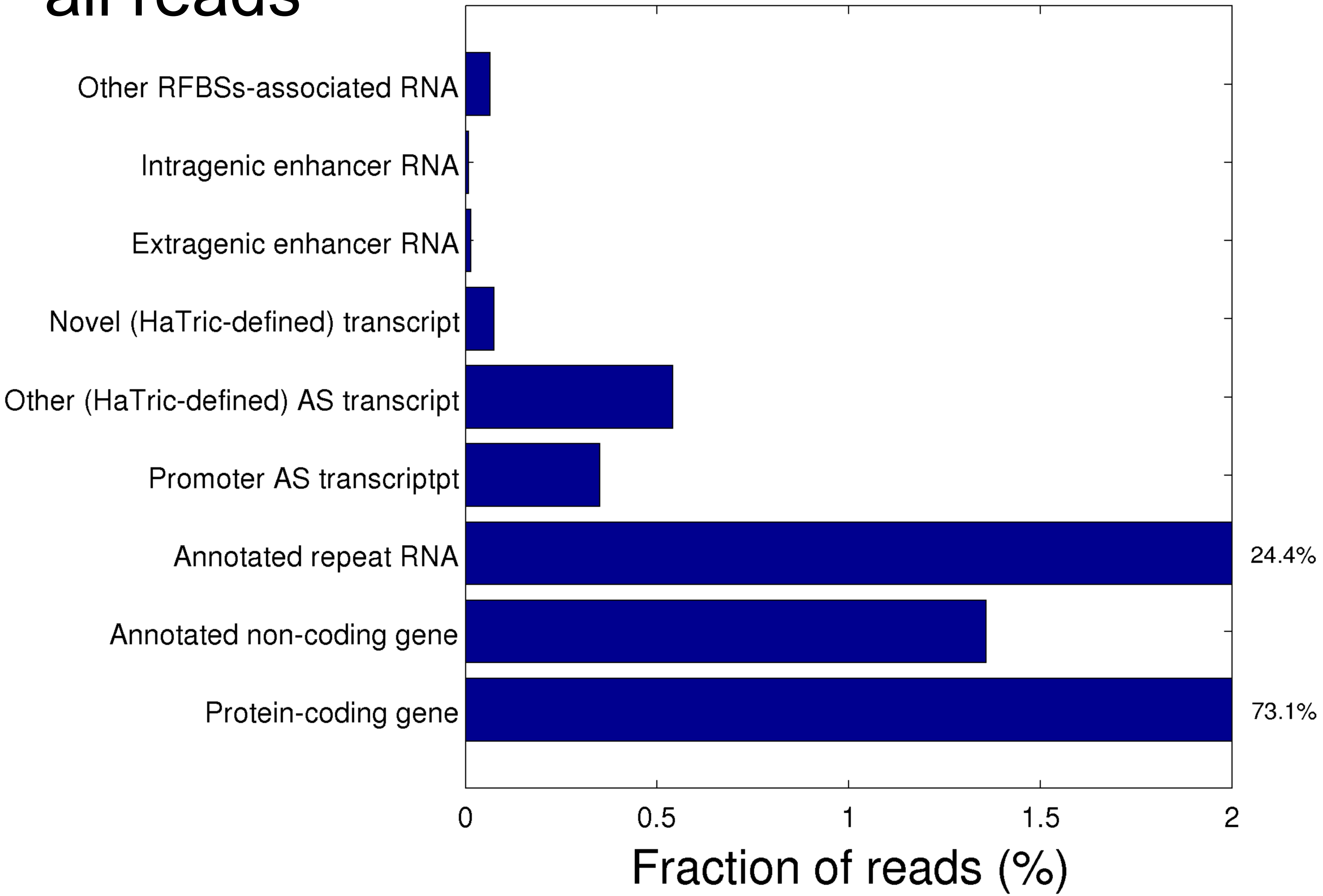
Most reads are found in annotated genes



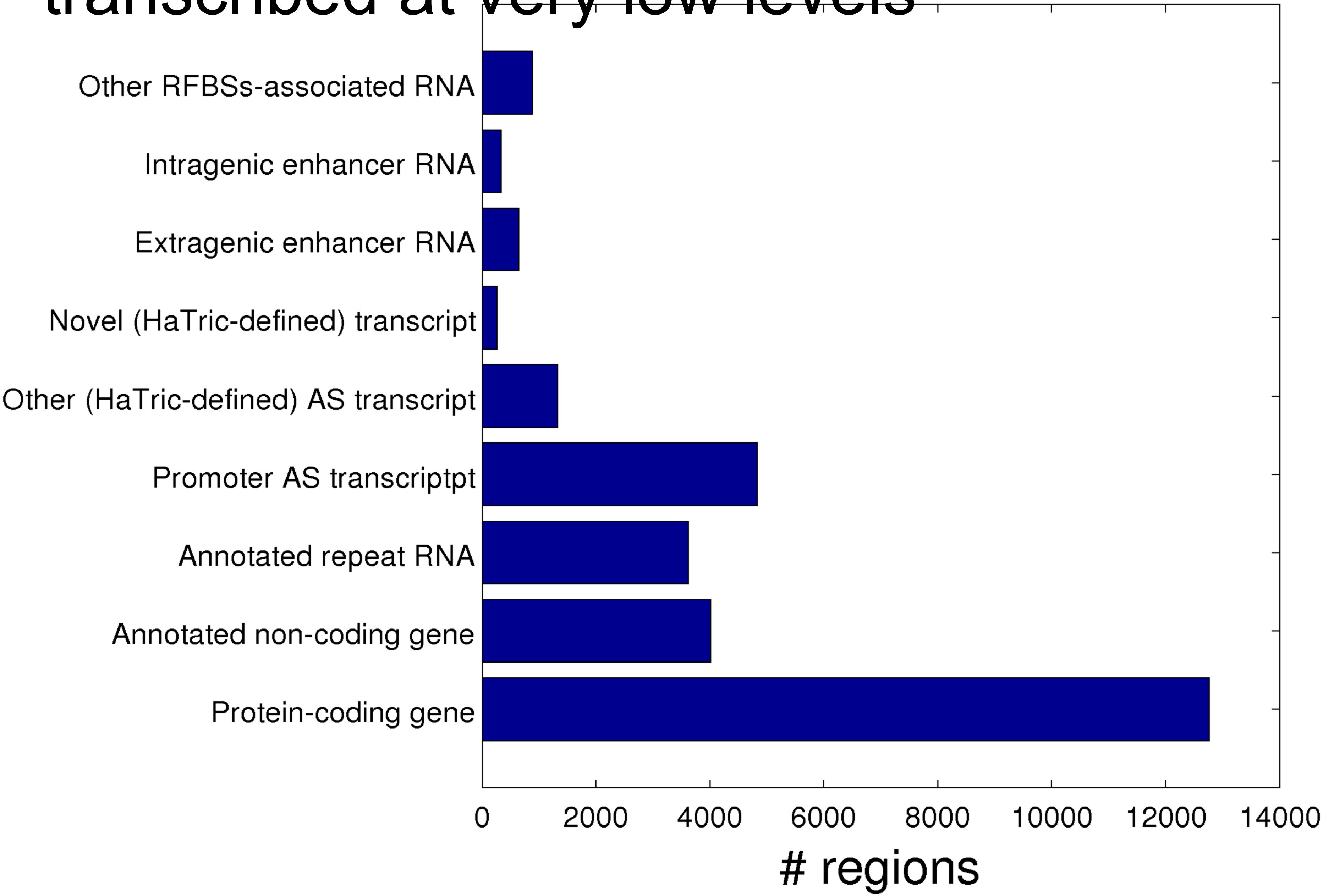
Transcribed regions account for 99.87% of all reads



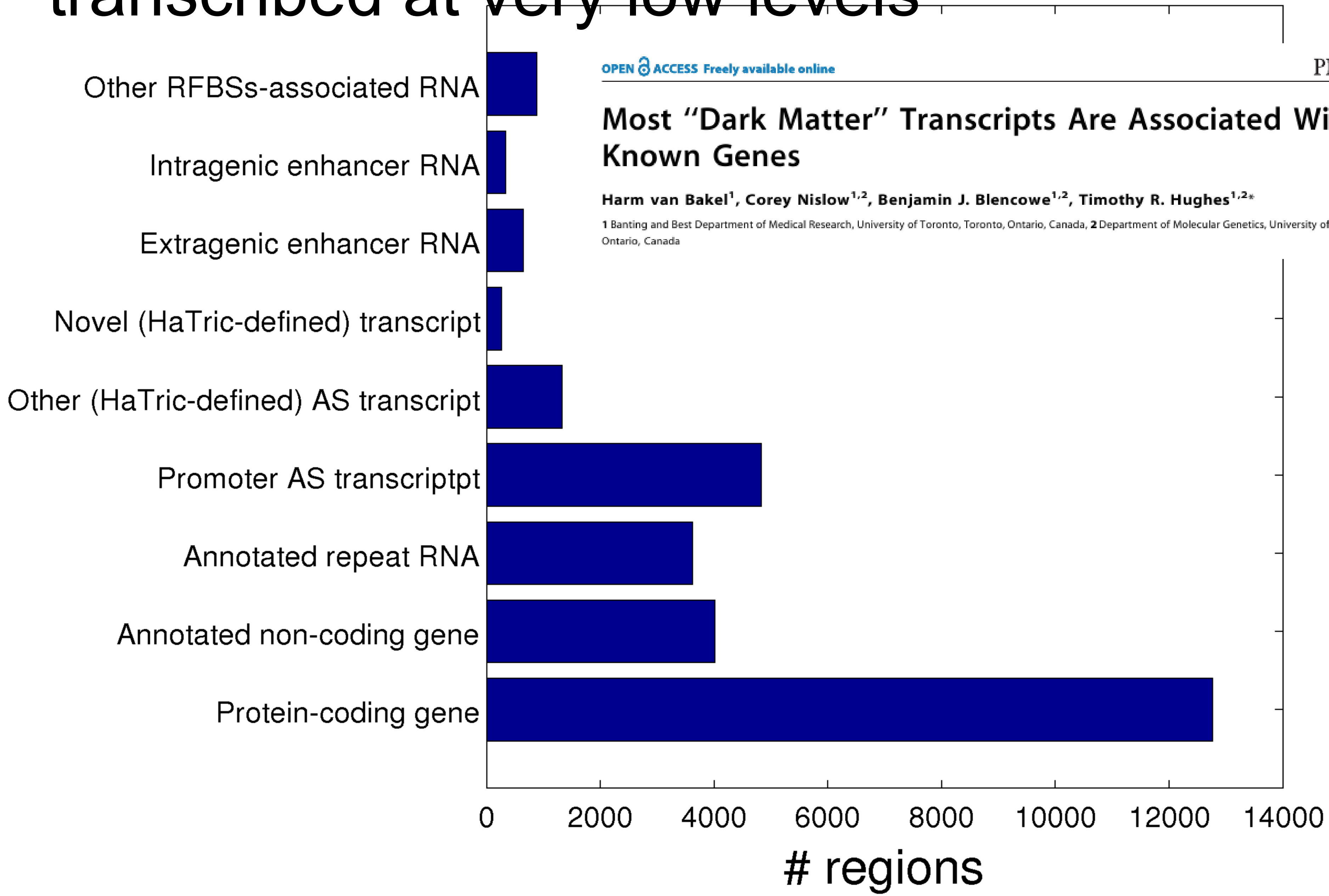
Transcribed regions account for 99.87% of all reads



There are many extragenic regions transcribed at very low levels



There are many extragenic regions transcribed at very low levels



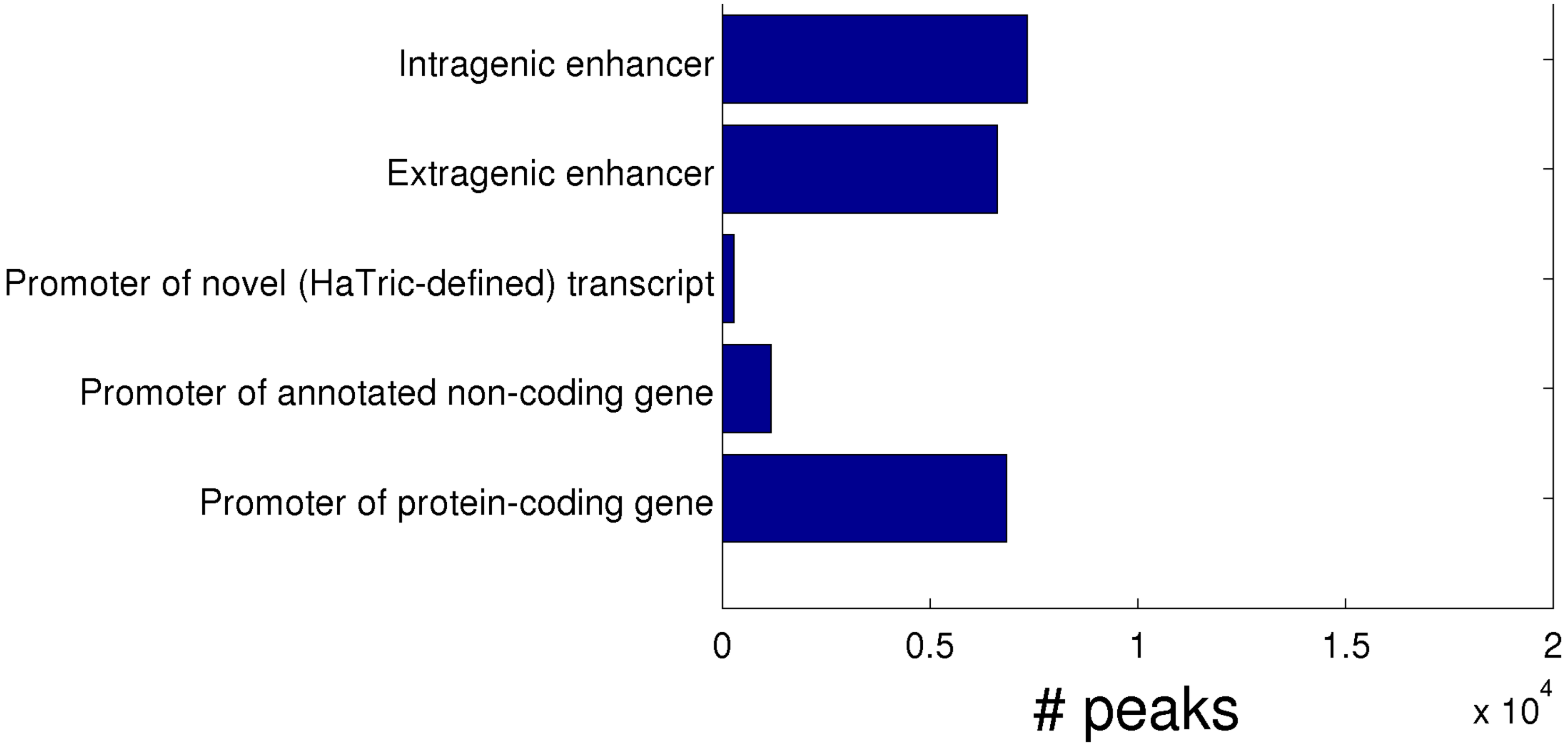
OPEN ACCESS Freely available online PLoS BIOLOGY

Most "Dark Matter" Transcripts Are Associated With Known Genes

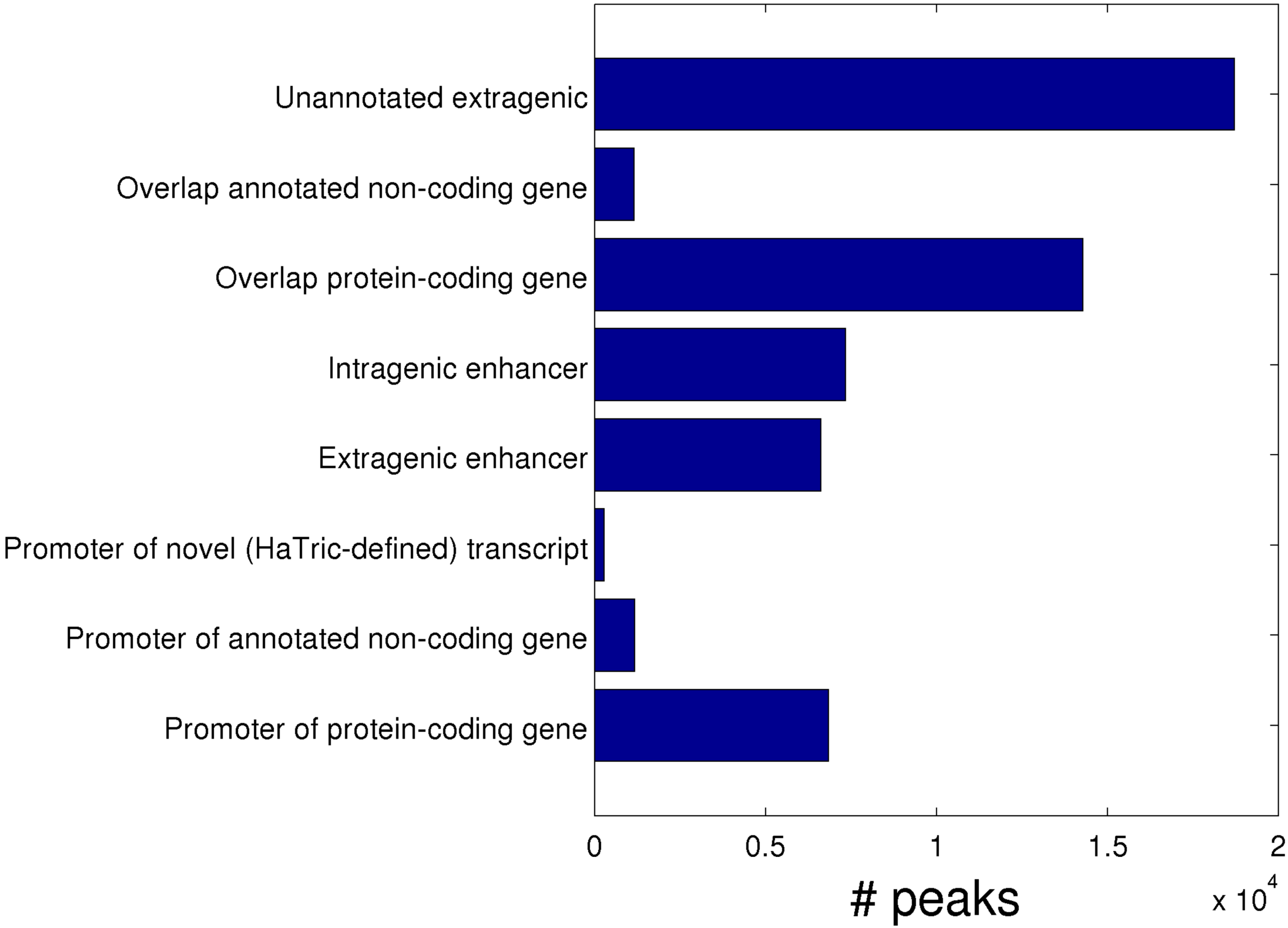
Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

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

Many TF binding sites are found at promoters or enhancers

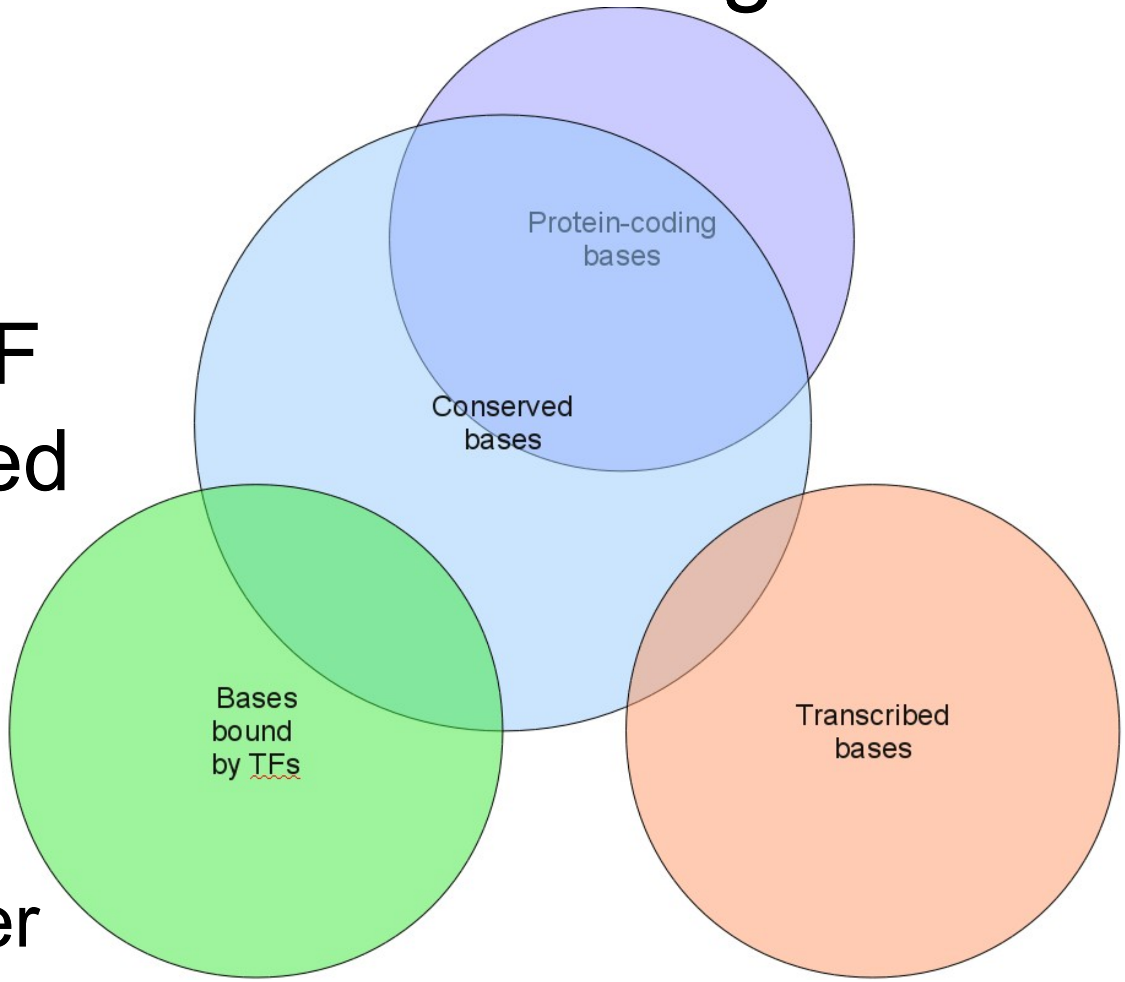


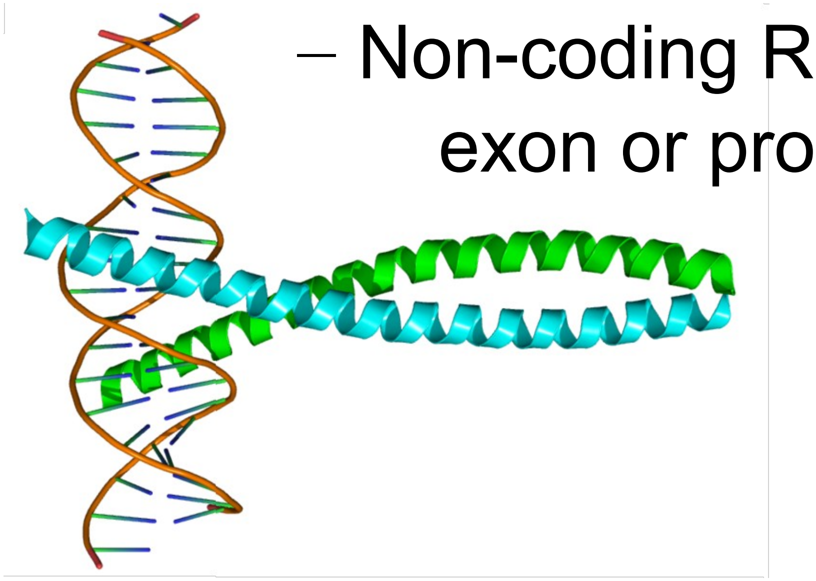
Most TFs bind in unannotated regions

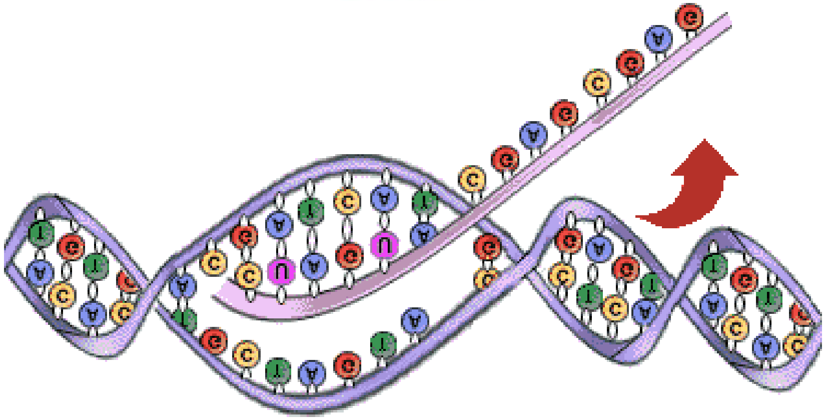


What drives the conservation of extragenic regions?

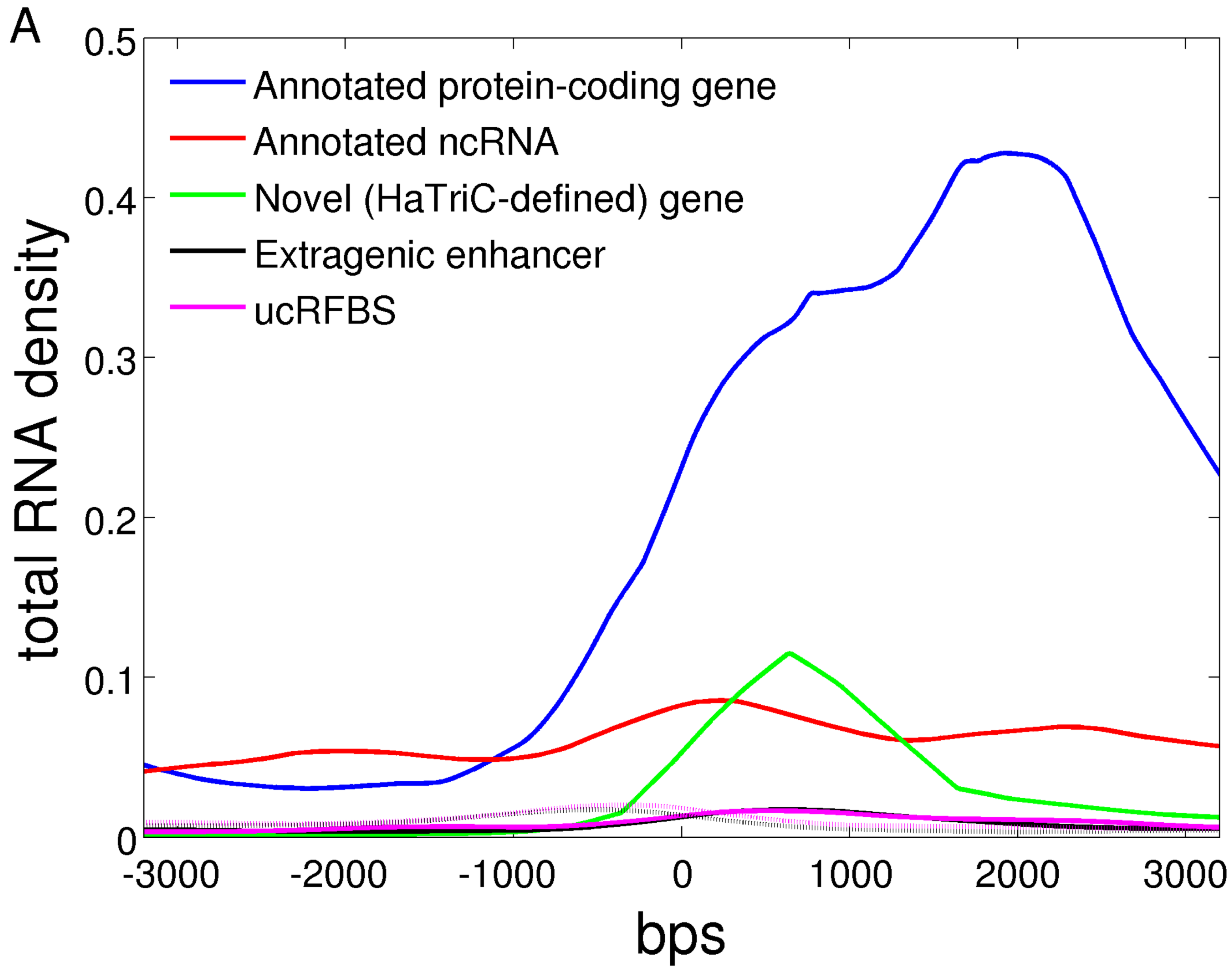
- Compare extragenic transcription and TF binding to conserved bases
  - TF binding sites
  - Non-coding RNA exon or promoter



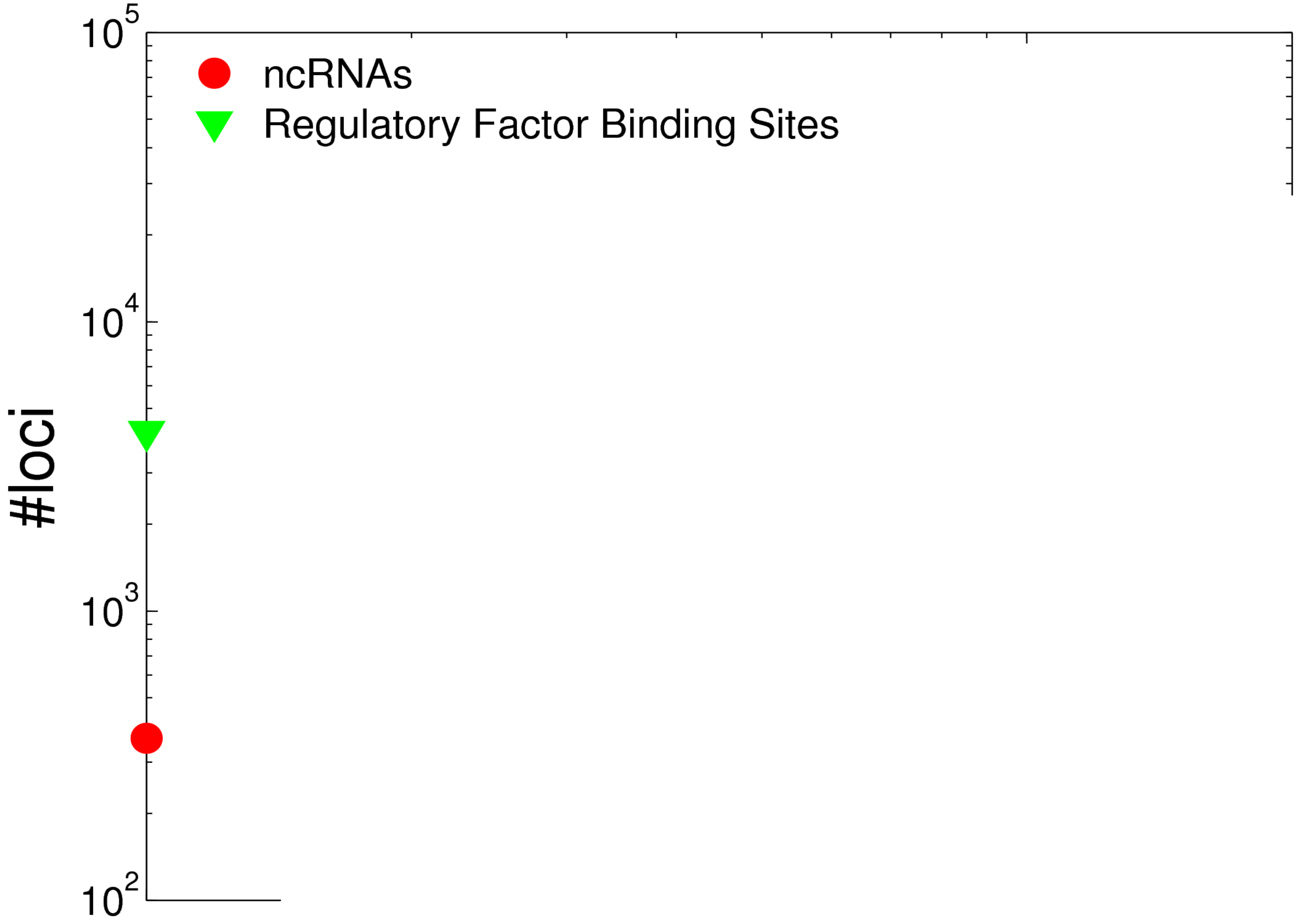




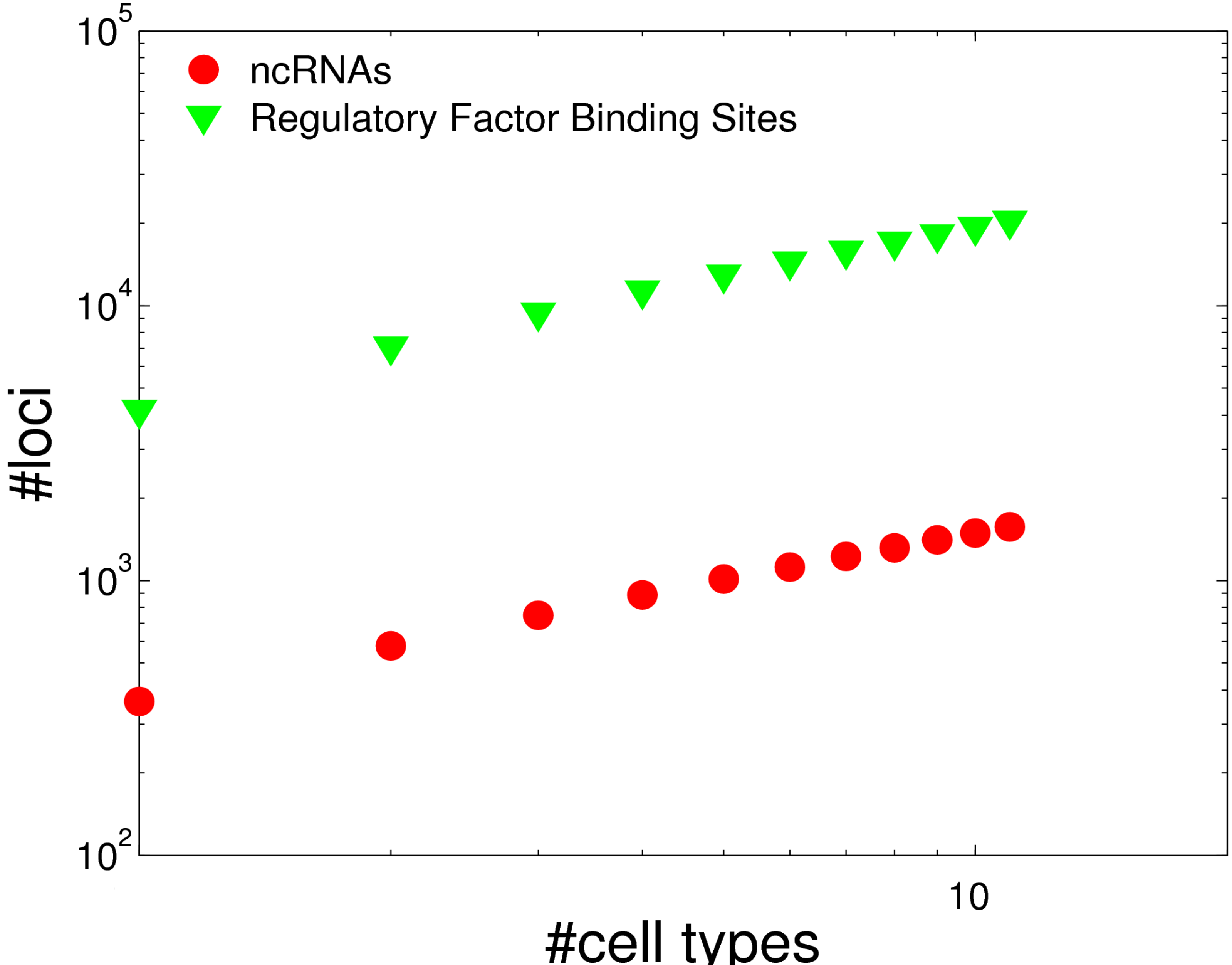
Easy to distinguish different types of transcription



About 80% of conserved bases are transcription factor binding sites

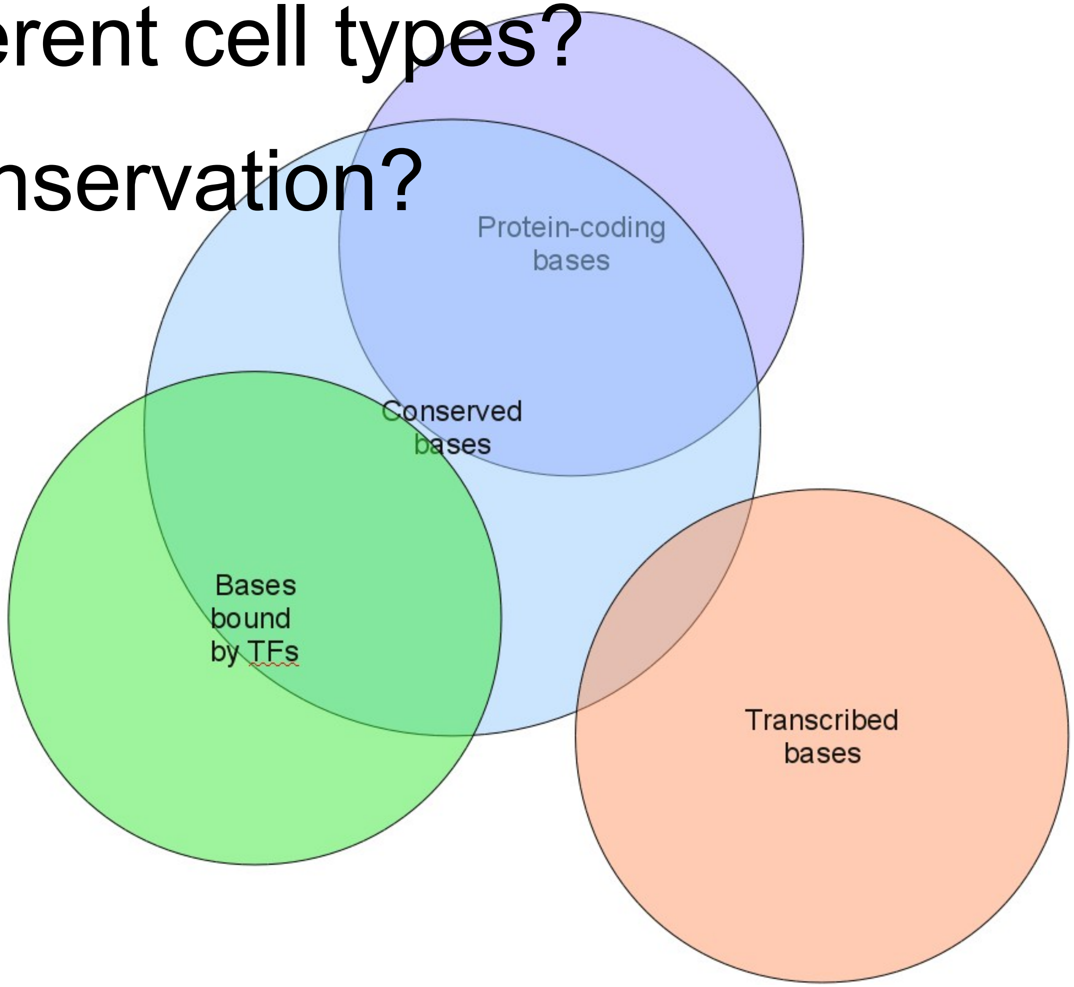


About 80% of conserved bases are transcription factor binding sites



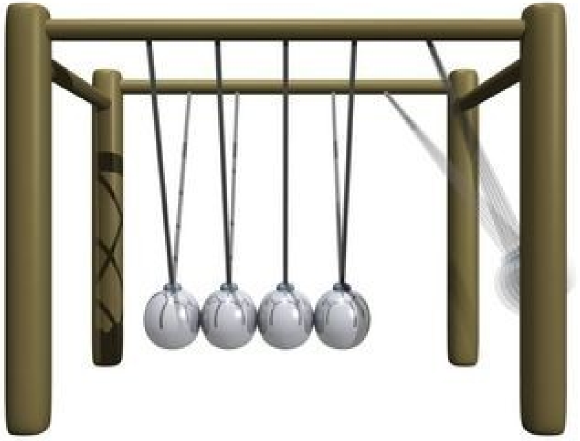
Summary II: *De novo* identification of transcribed regions suggests that most conservation is due to TF binding

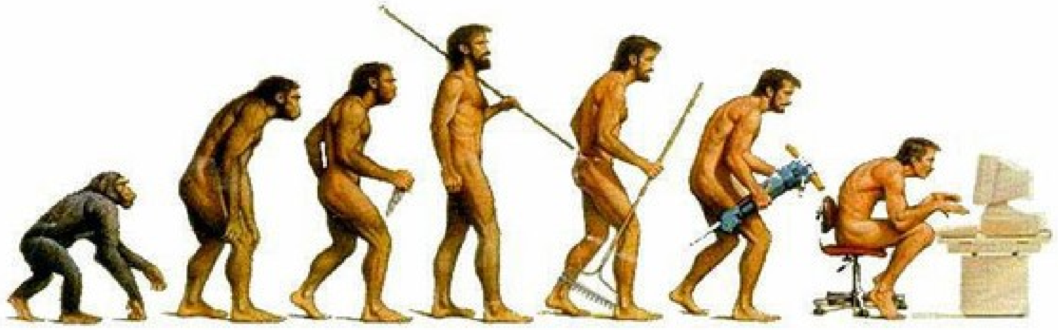
- Different roles in different cell types?
- Other reasons for conservation?



Future Work: Organizing principles of the genome

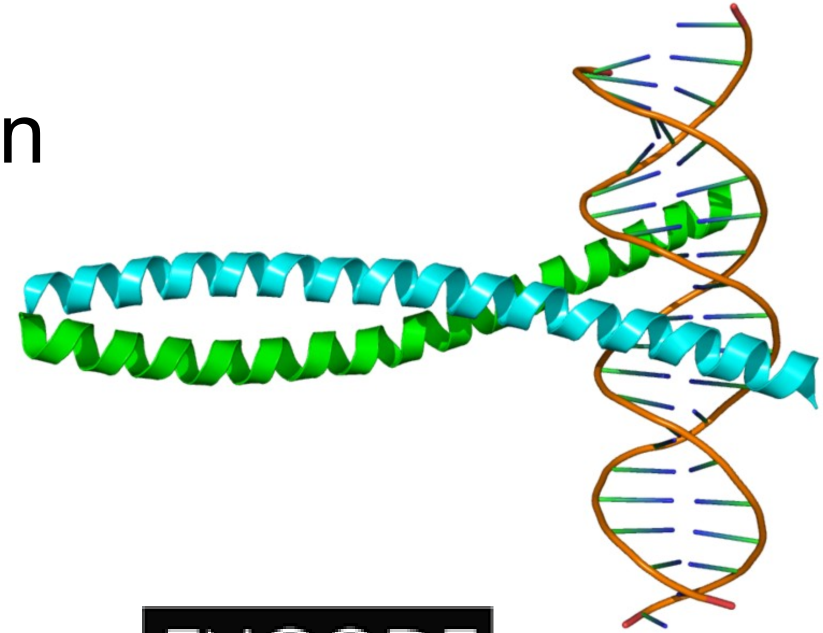
- Systems biology approach to develop biophysical models





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

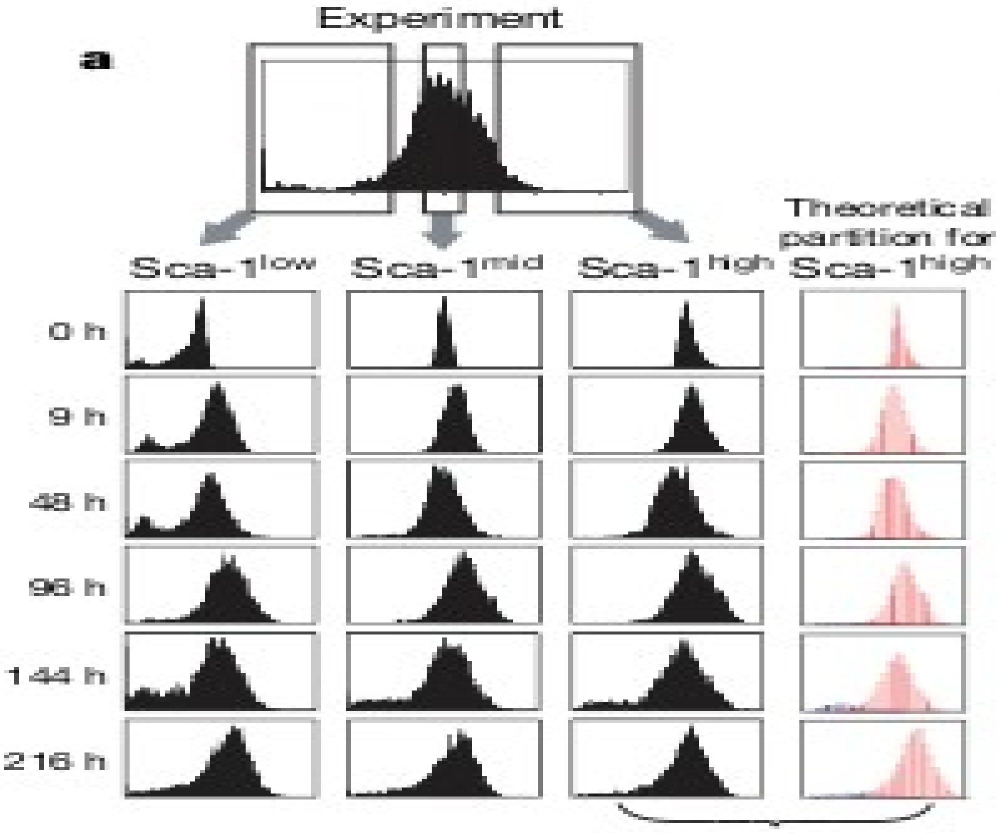
- How can histone modifications be read and written?
- What determines transcription factor binding?
- What determines the level of transcription?



ENCODE

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

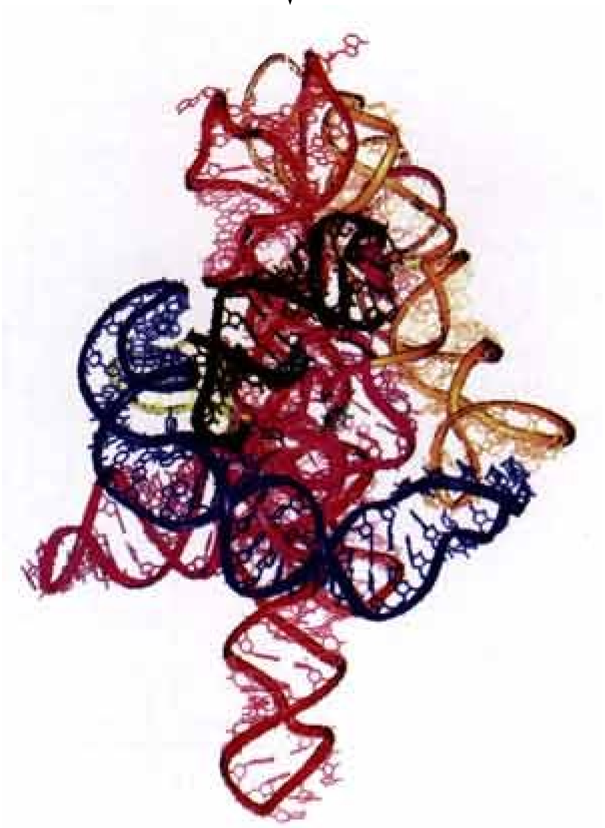


Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis

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?

......ACGUCCAAAUUCCCUAGGCUCAAGGCAUUCGAUCGGGAUUAUA.....



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

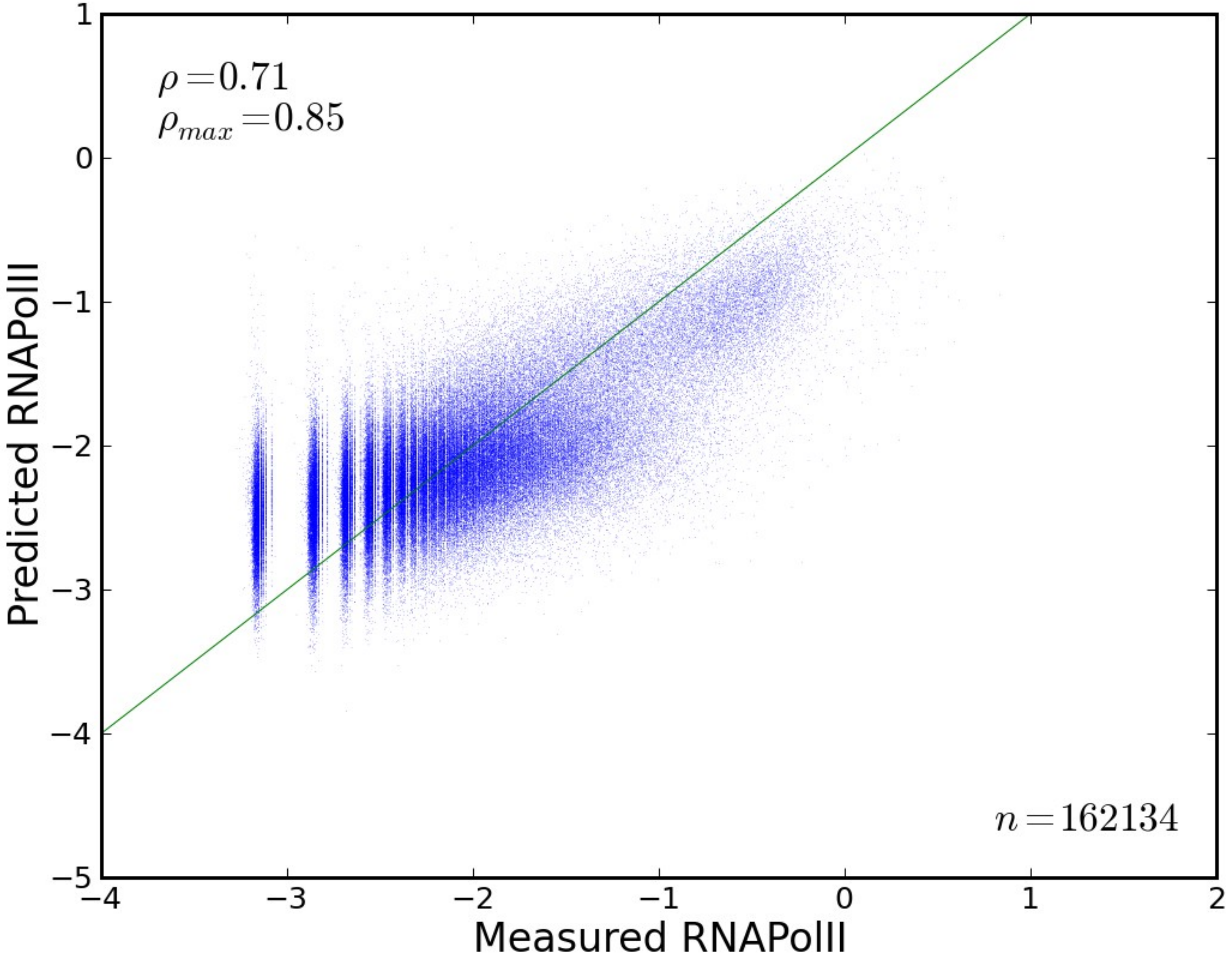
Harvard Medical School
Biopolymer facility
Kristin Waraska
Robert Steen

Johns Hopkins
Jing Wu, Paul Worley Lab

Thank You

?

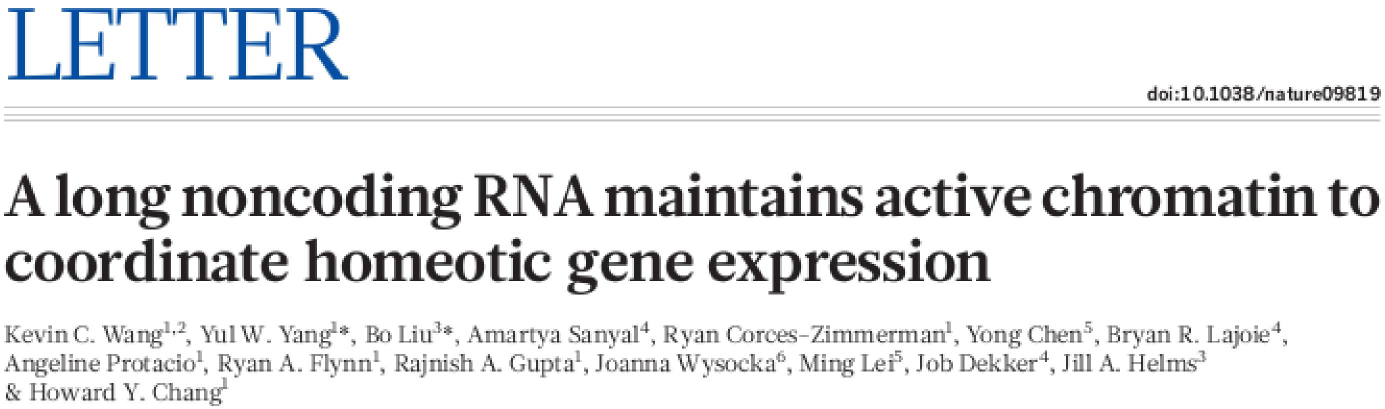
Is there an epigenetic code to determine the cell-type specific function of the sequence?



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
  - 3.8 kb, spliced, polyA+



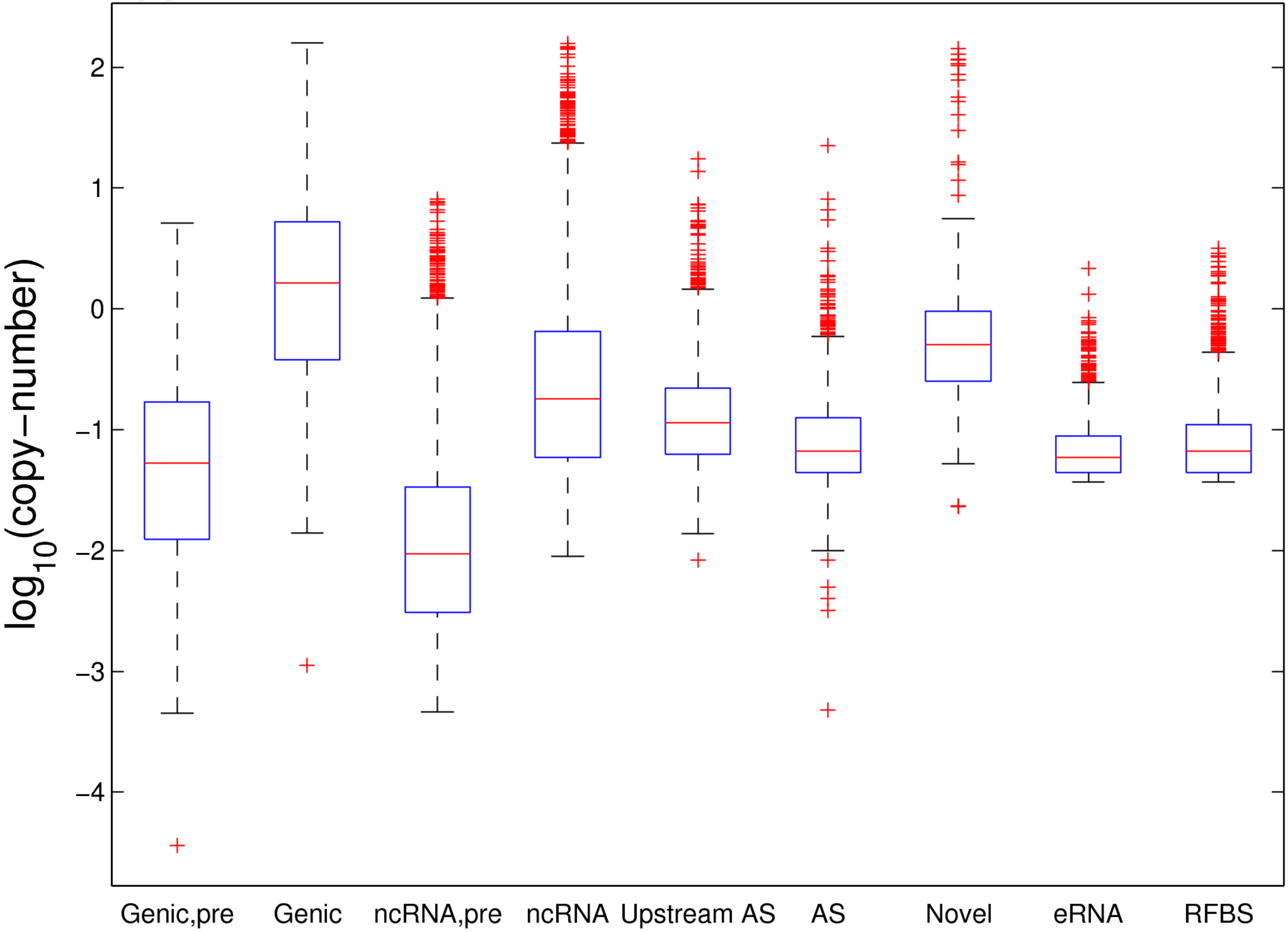
LETTER

doi:10.1038/nature09819

A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression

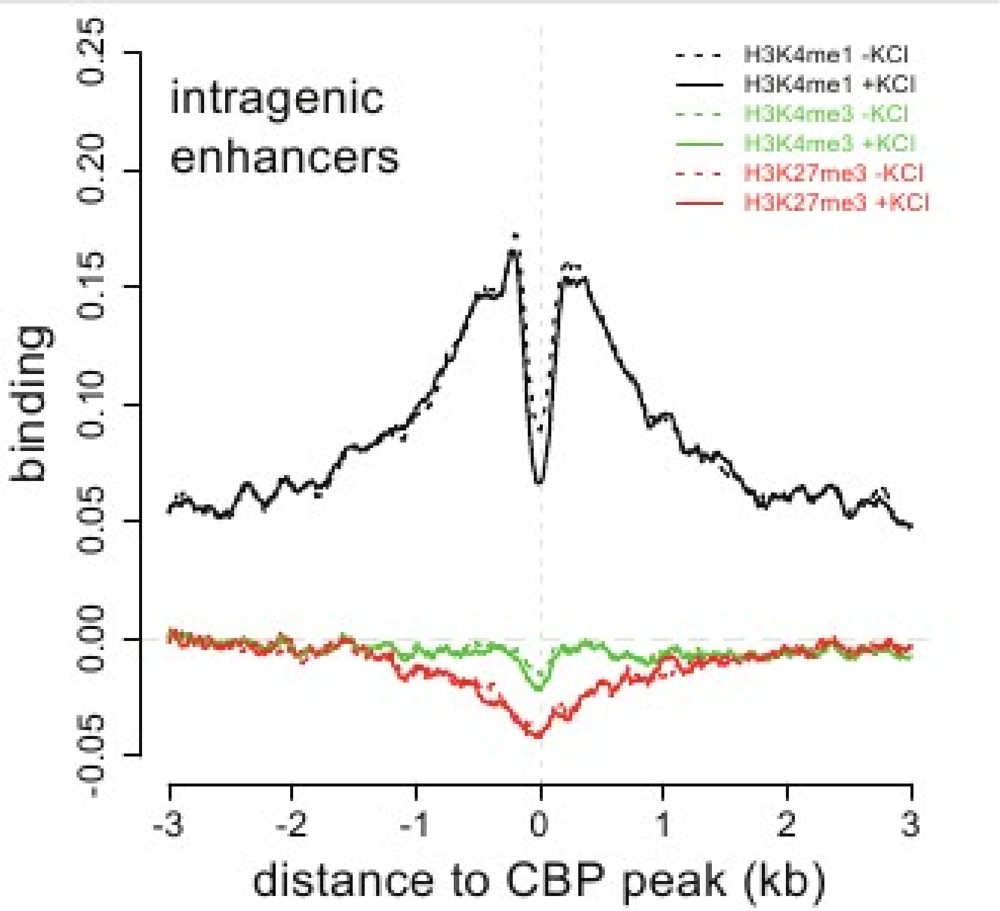
Kevin C. Wang[1,2], Yul W. Yang[1]*, Bo Liu[3]*, Amartya Sanyal[4], Ryan Corces-Zimmerman[1], Yong Chen[5], Bryan R. Lajoie[4], Angeline Protacio[1], Ryan A. Flynn[1], Rajnish A. Gupta[1], Joanna Wysocka[6], Ming Lei[5], Job Dekker[4], Jill A. Helms[3] & Howard Y. Chang[1]

Copy numbers for different categories



Intragenic enhancers

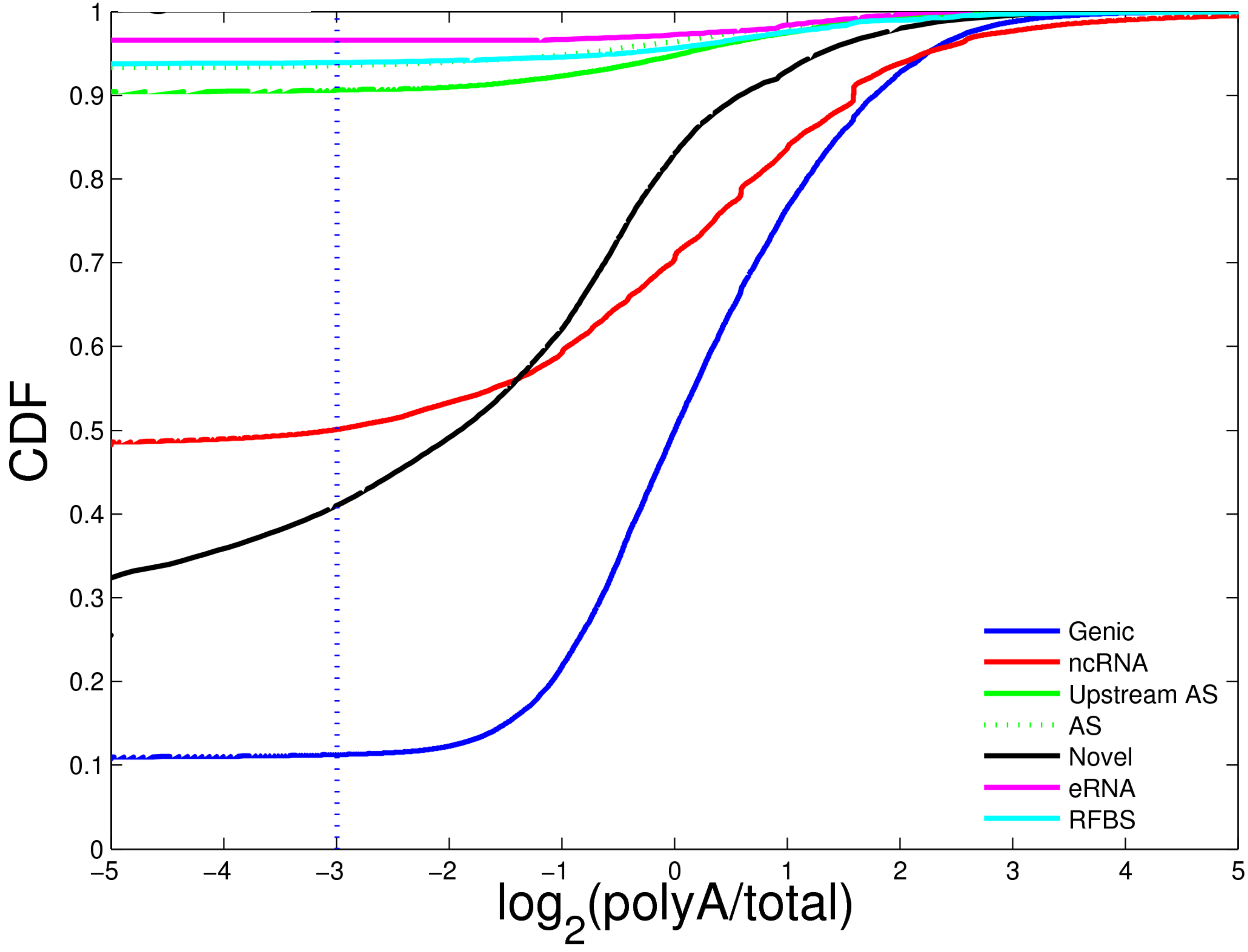
- ~7,000 enhancers overlapping introns
  - H3K4me1, but no H3K4me3



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

Most ncRNAs are not polyadenylated



Assume ChIP and input Poisson distributed

- Z_i = #ChIP reads - #input reads in window *i*
- ~1 read/100 bp
  - Assume #reads in window P(k) = λ^kexp(-λ)/k!
    - Difference between two Poisson random variables
    - Z_i ~ Skellam(z, λ_1, λ_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 = #ChIP reads - #input reads in window *i*
- ~1 read/100 bp
  - Assume #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

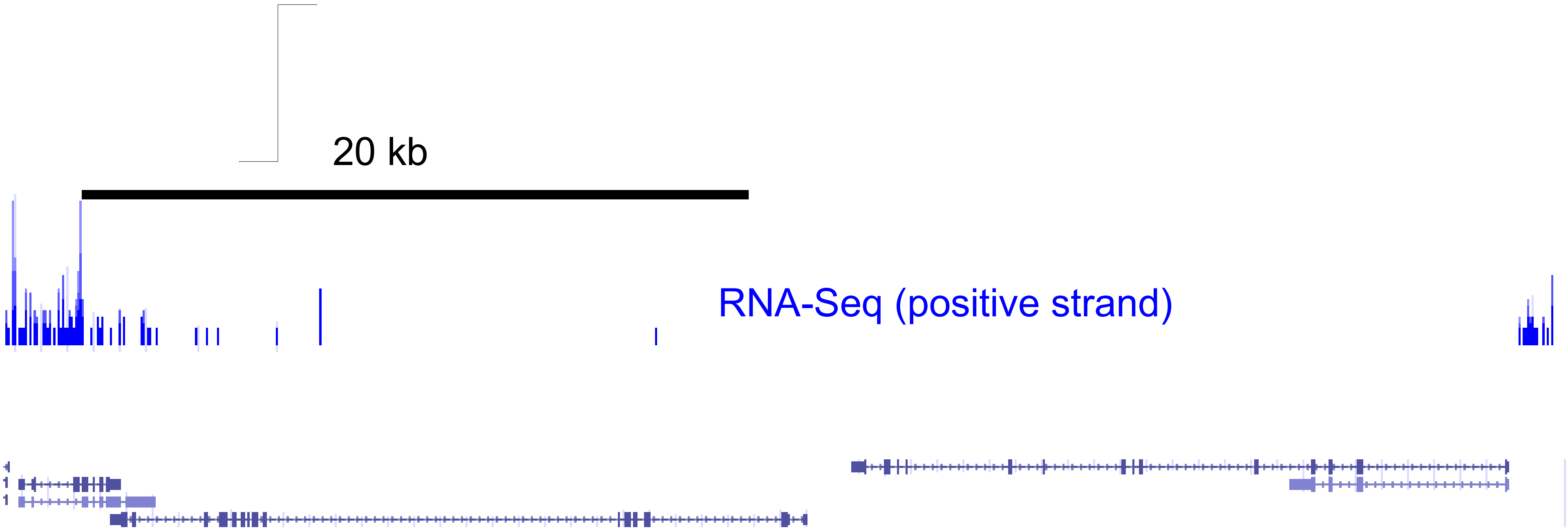
Haar-wavelet Transcript Calling (HaTriC) for *de novo* identification of transcribed regions

```
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

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

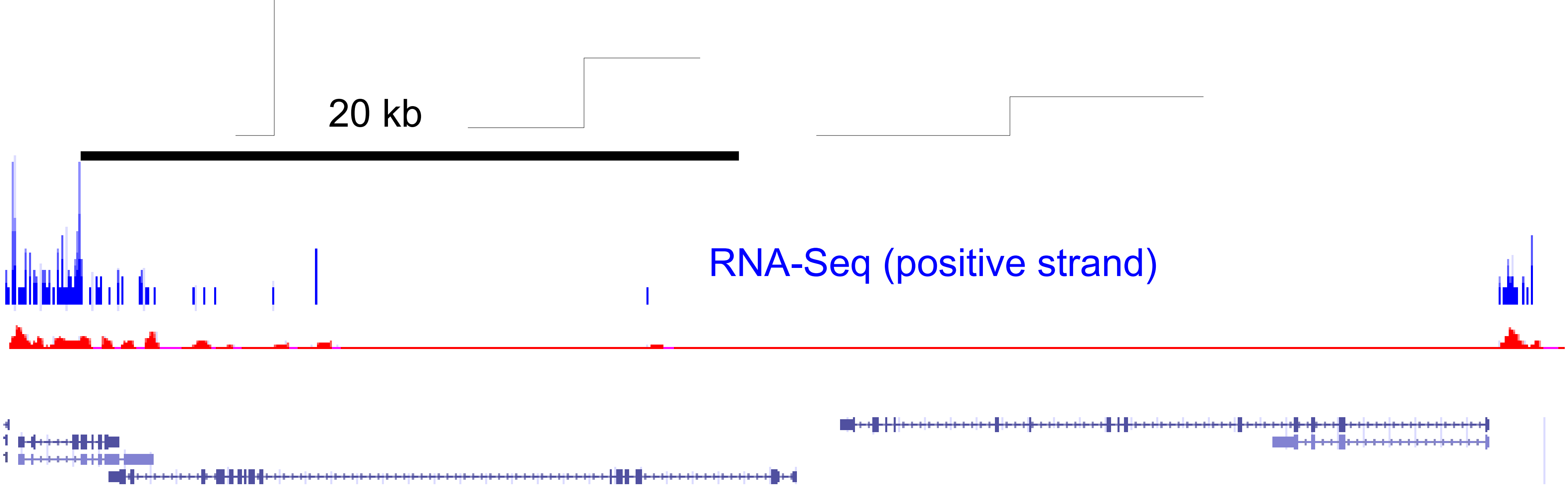


The Haar-wavelet can be scaled to analyze multiple length scales

- Break points correspond to sharp changes in read density

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

  - Use scales L from 8 to 20

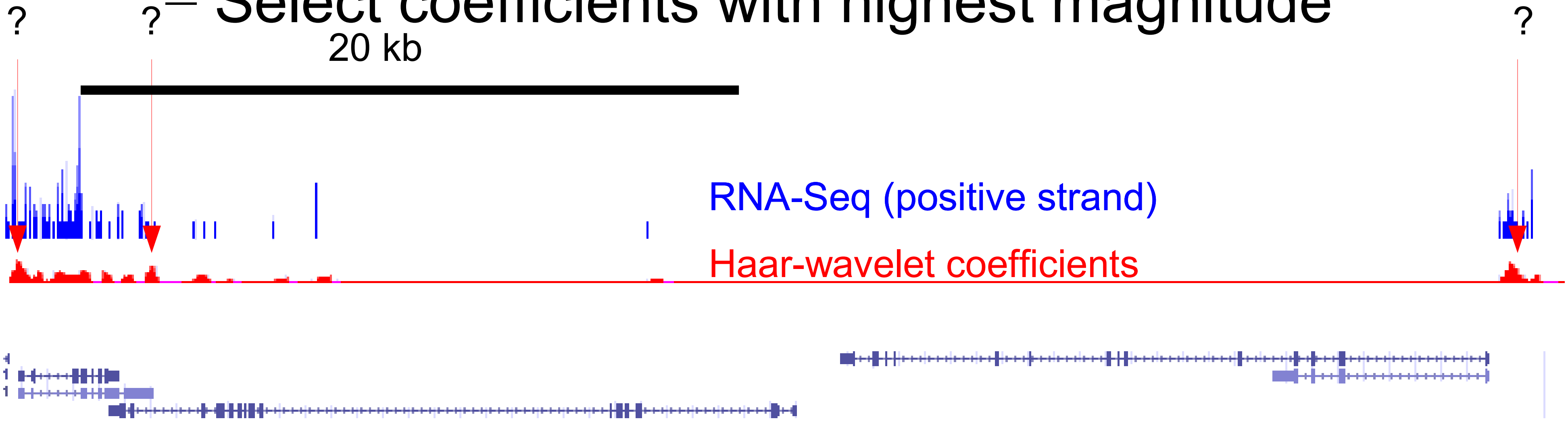


The coefficients with largest magnitude are selected as candidate break points

- Break points correspond to sharp changes in read density

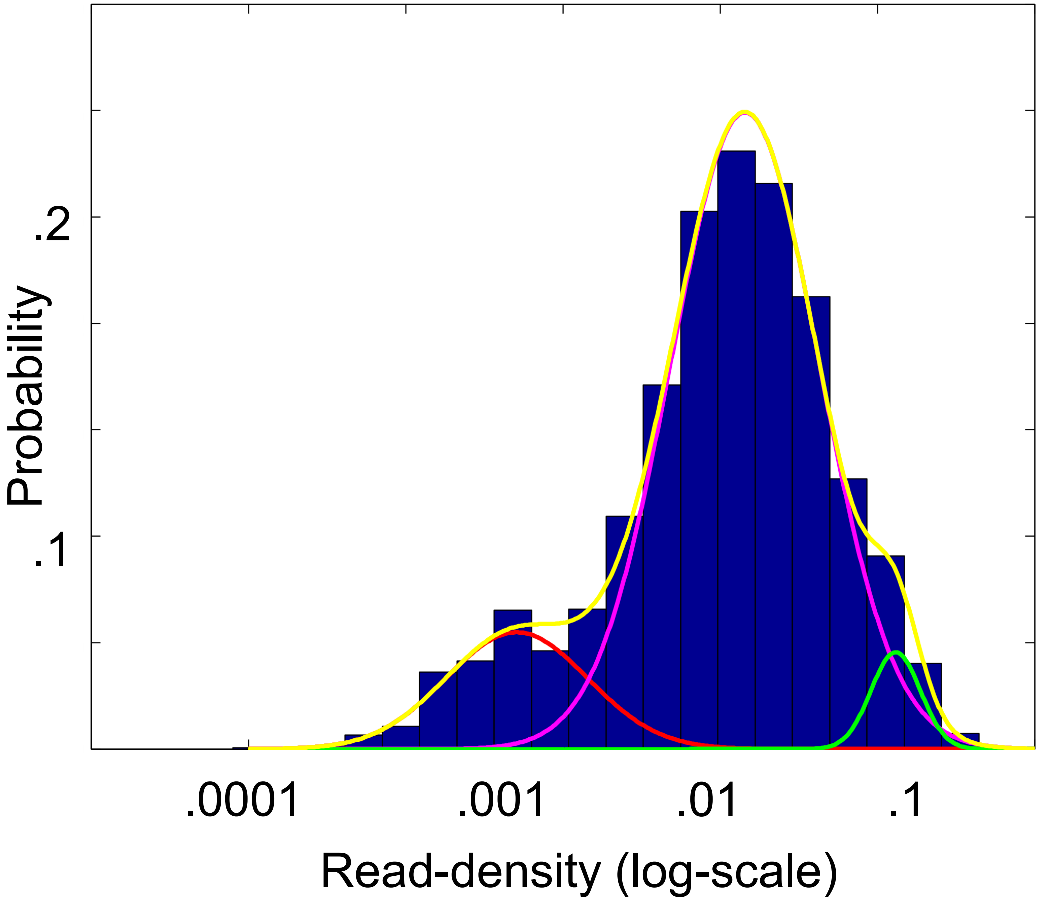
$$h_L(n) = \frac{1}{\sqrt{2^{L+1}}}\left(\sum_{i=n}^{n+2^L-1}\log(1+r_i) - \sum_{n-1}^{i=n-2^L}\log(1+r_i)\right)$$

  - Select coefficients with highest magnitude



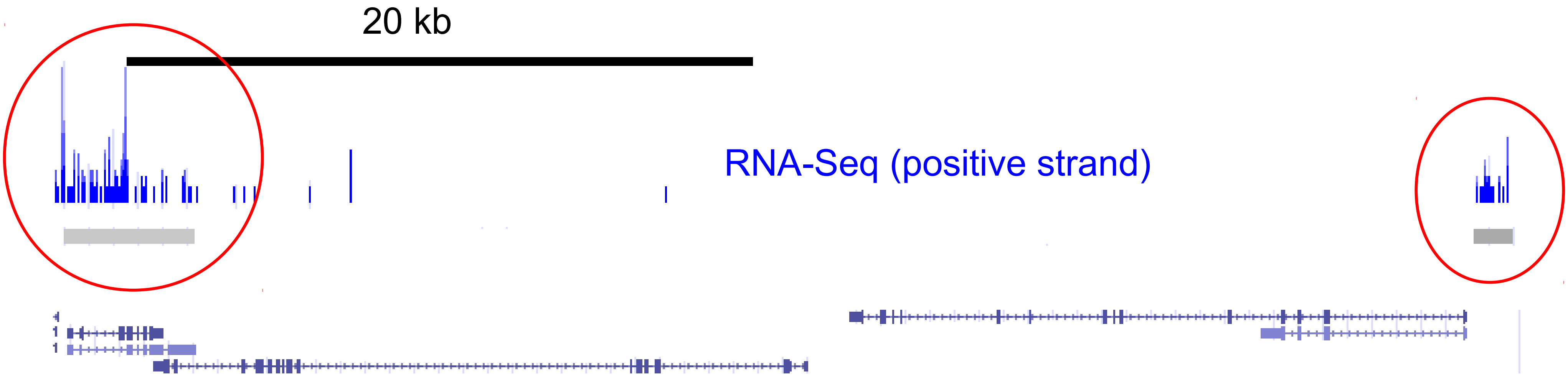
The density distribution for the regions determined by the break points is bimodal

- 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


20 kb
RNA-Seq (positive strand)