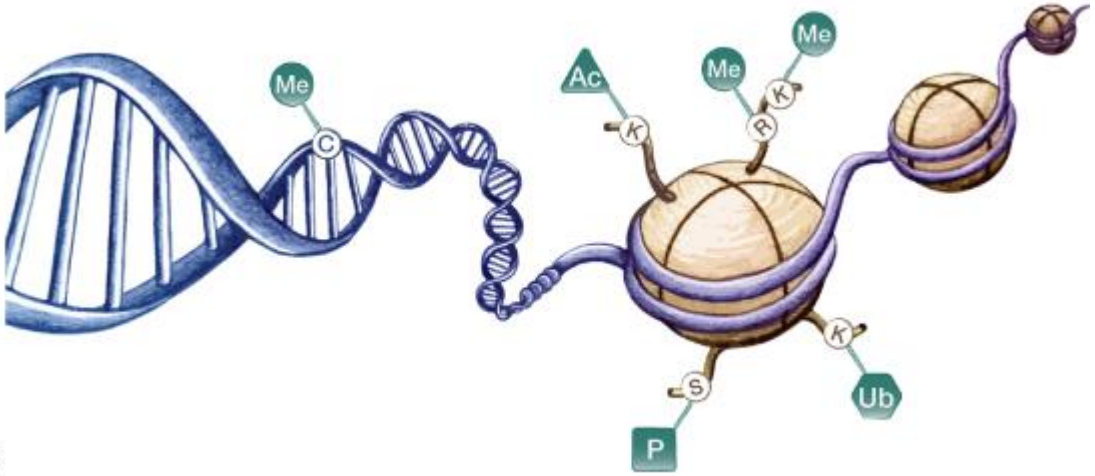


Transcriptional Regulation by Protein Factors and Epigenetic Mechanisms

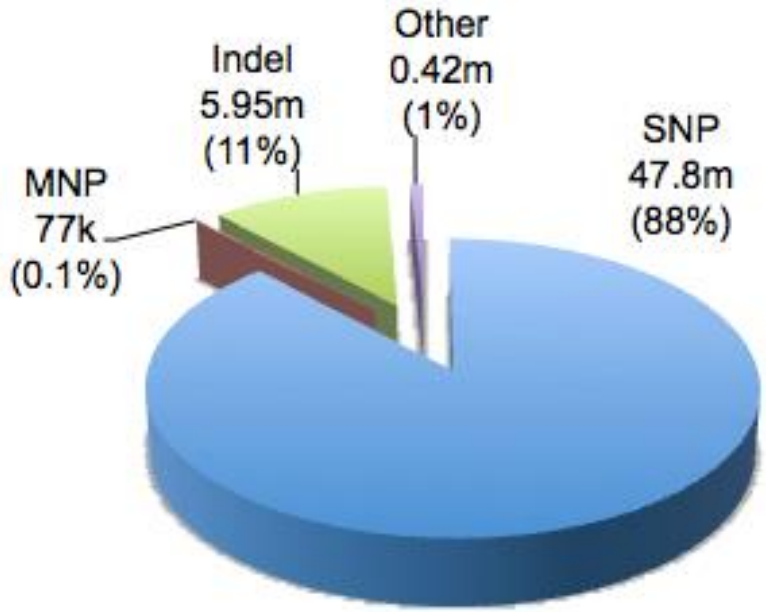
Introduction to epigenetics, transcription factors, and related next-generation sequencing methods

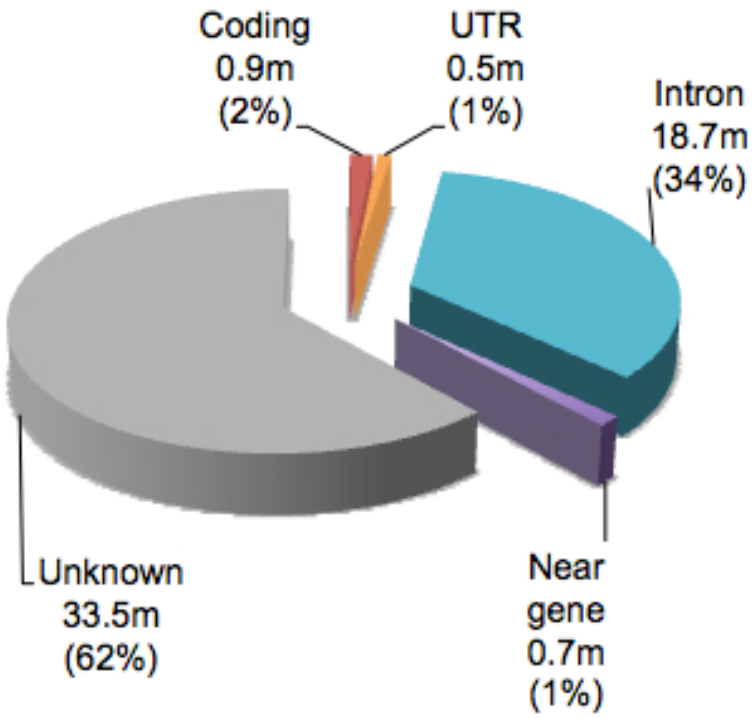
Alexander Predeus

Bioinformatics Institute, Saint Petersburg

Why study regulation?

- Evolution mainly works on regulatory, not protein-coding, DNA
- If we know regulation, we can find key spots in large pathways
- Next-generation sequencing! (i.e. because we can)

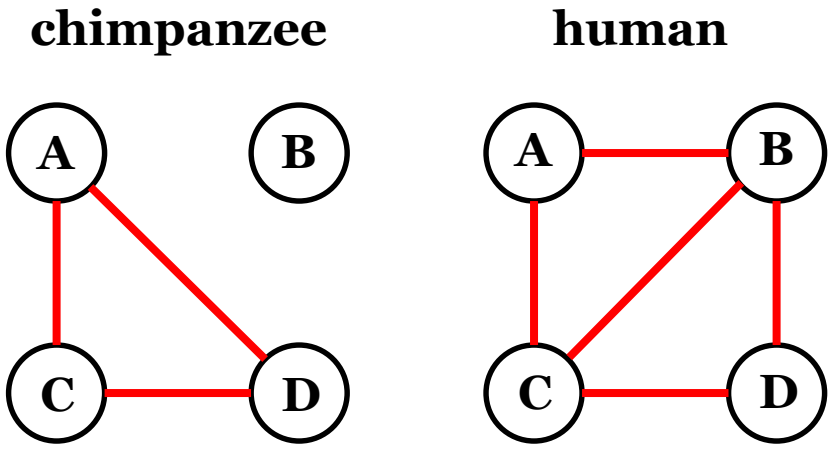


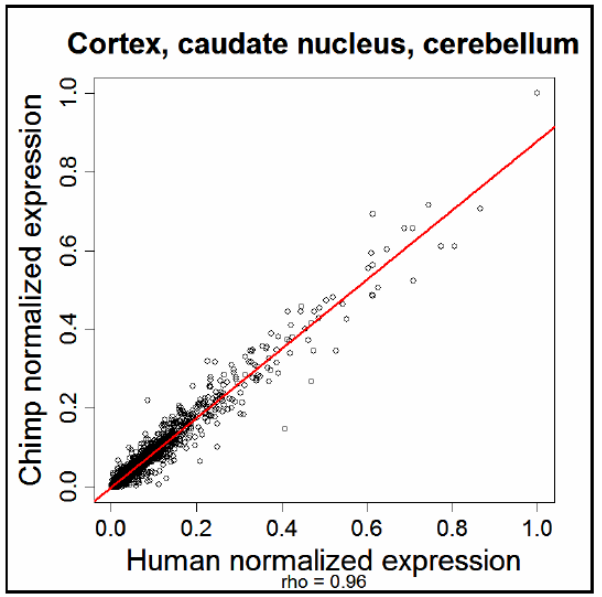


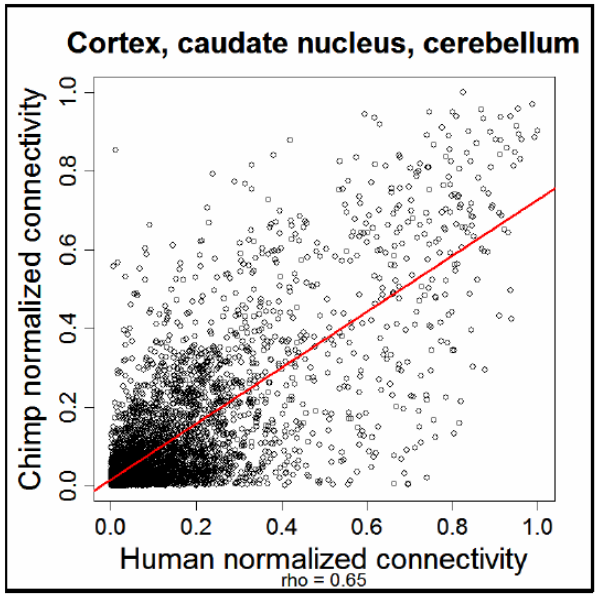
http://massgenomics.org/2012/01/the-current-state-of-dbsnp.html

Primate to human: it's in the regulation

- gene structure and expression are well conserved
- however, gene coexpression is not
- the difference lies in gene regulation





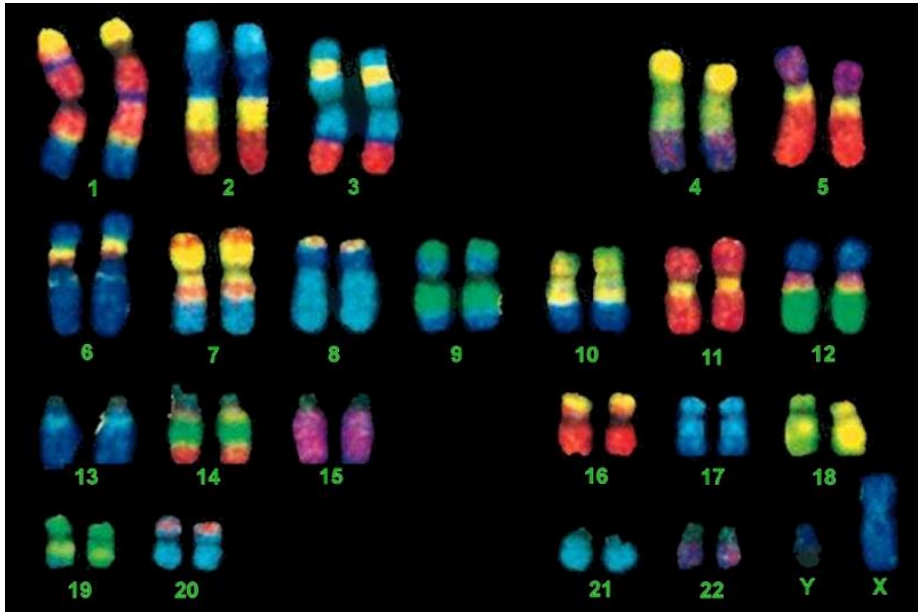


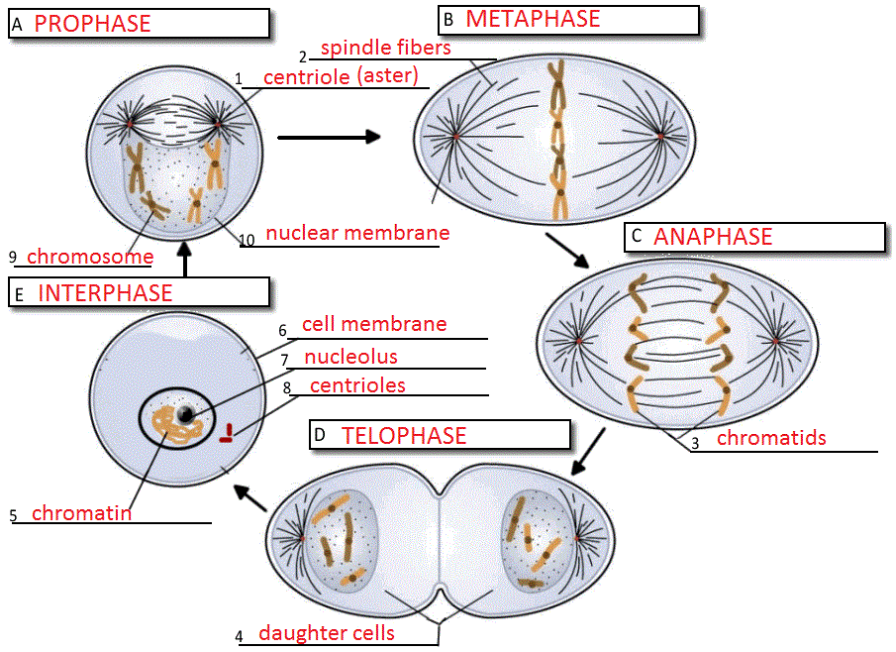
Proc Natl Acad Sci U S A. 2006 Nov 21;103(47):17973-8.

Chromosomes and chromatin

- chromosomes are dense complexes of DNA and proteins
- each human chromosome contains on average 5 cm of DNA
- this is about 2 m of DNA overall – too much!
- chromatin = euchromatin + heterochromatin

0.2-20 µm

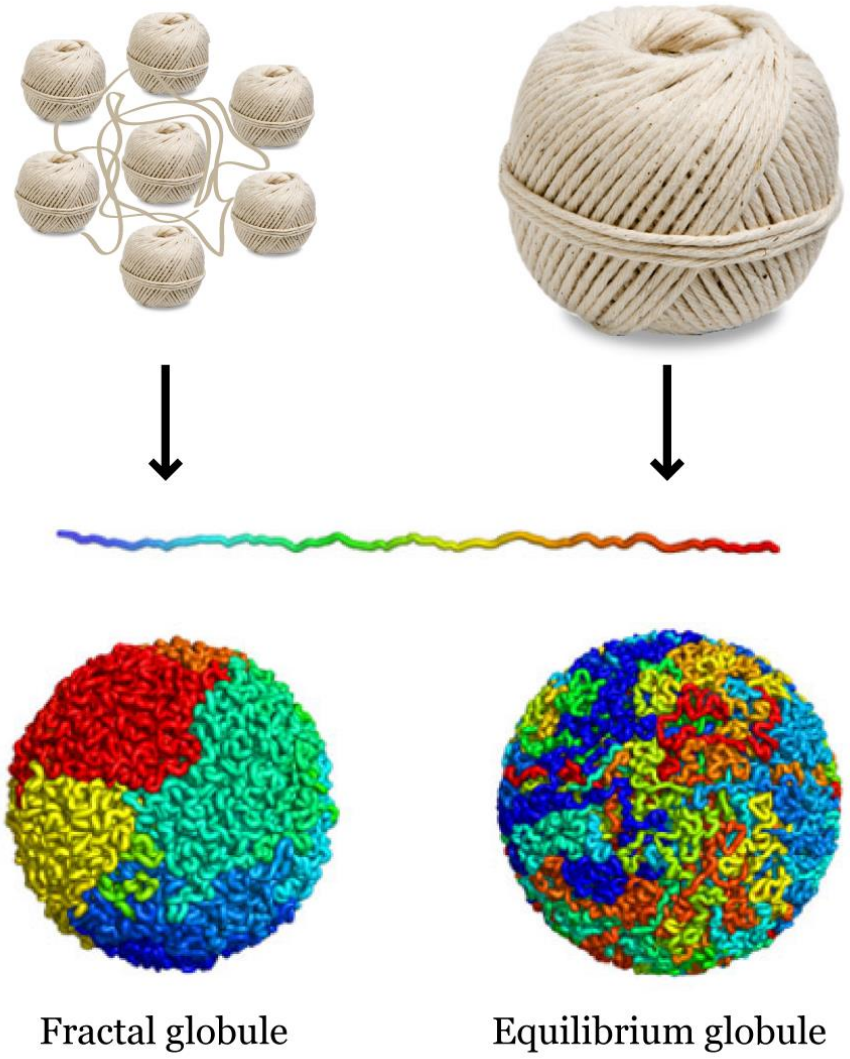




http://www.biologycorner.com/worksheets/cellcycle_labelme_key.html
http://www.mun.ca/biology/scarr/FISH_chromosome_painting.html

Chromatin is a fractal globule

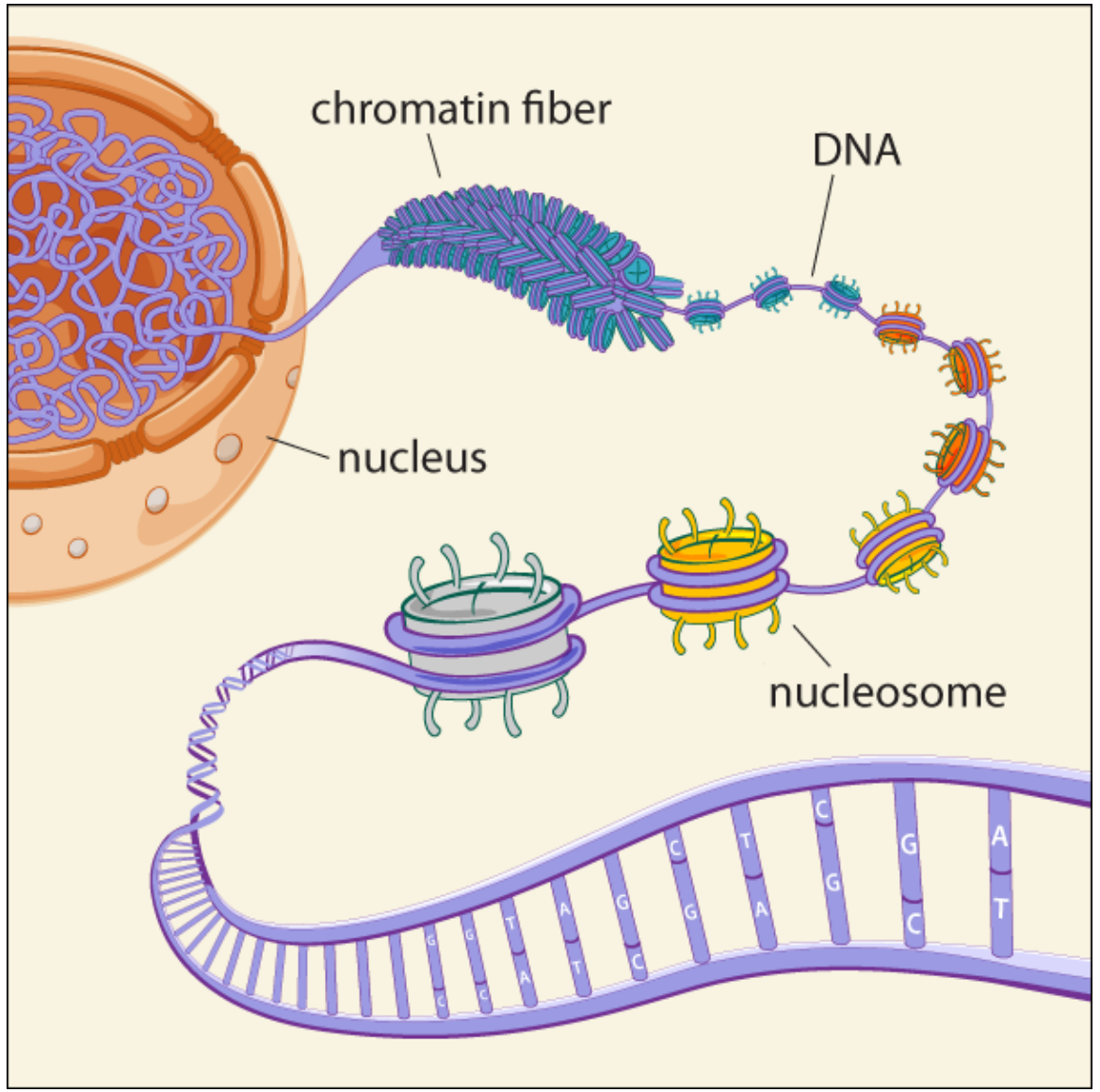
- equilibrium ~ roll of twine
- fractal globule ~ many rolls of twine loosely put together
- easier access, relative independence of regions



Mirny, L.A., Chromosome Res 19, pp 37–51 (2011)

Chromatin organization

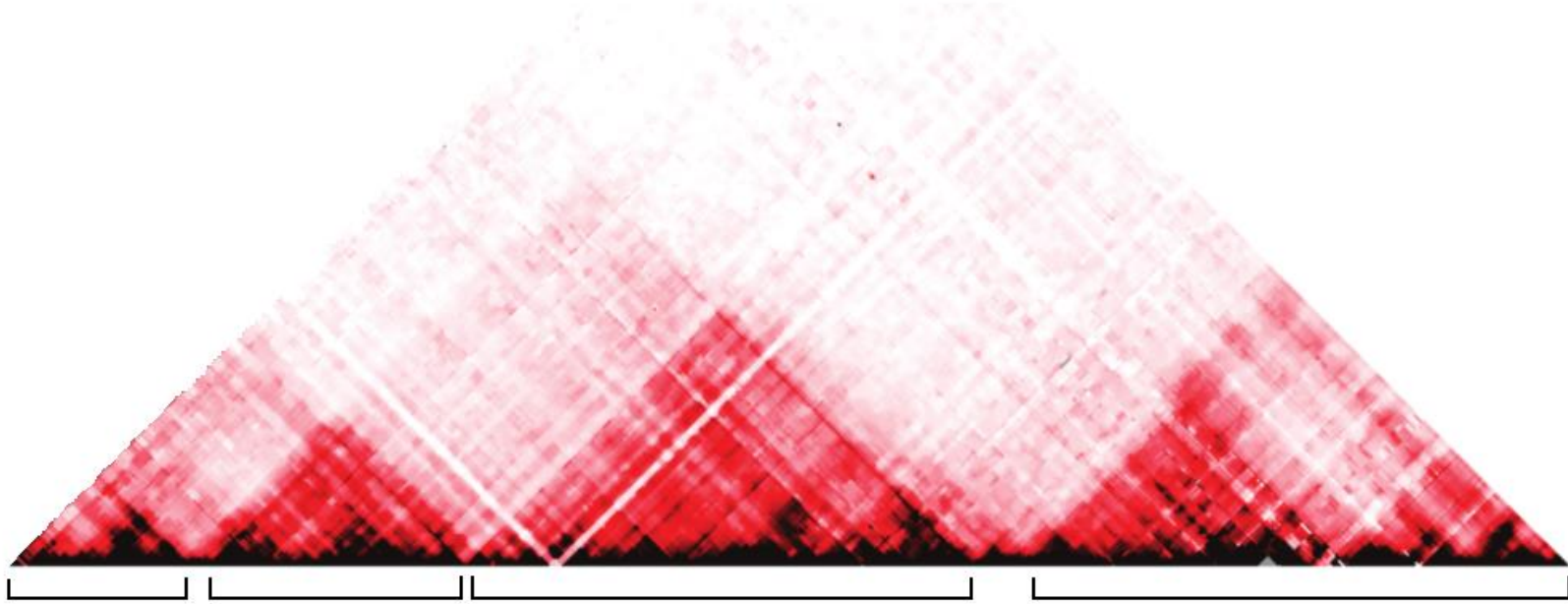
- there are three major levels of DNA packaging:
  - "beads on a string" nucleosomes
  - 30 nm fibre (still euchromatin!)
  - densely packed structures (heterochromatin)
- nucleosome beads include histone proteins

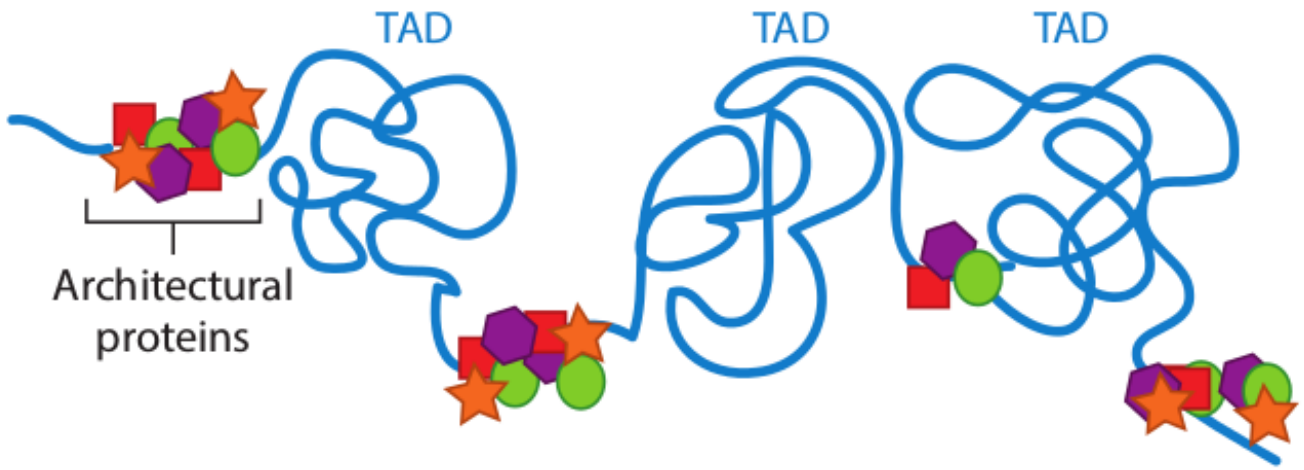


http://www.broadinstitute.org/news/1504

Topologically Associated Domains

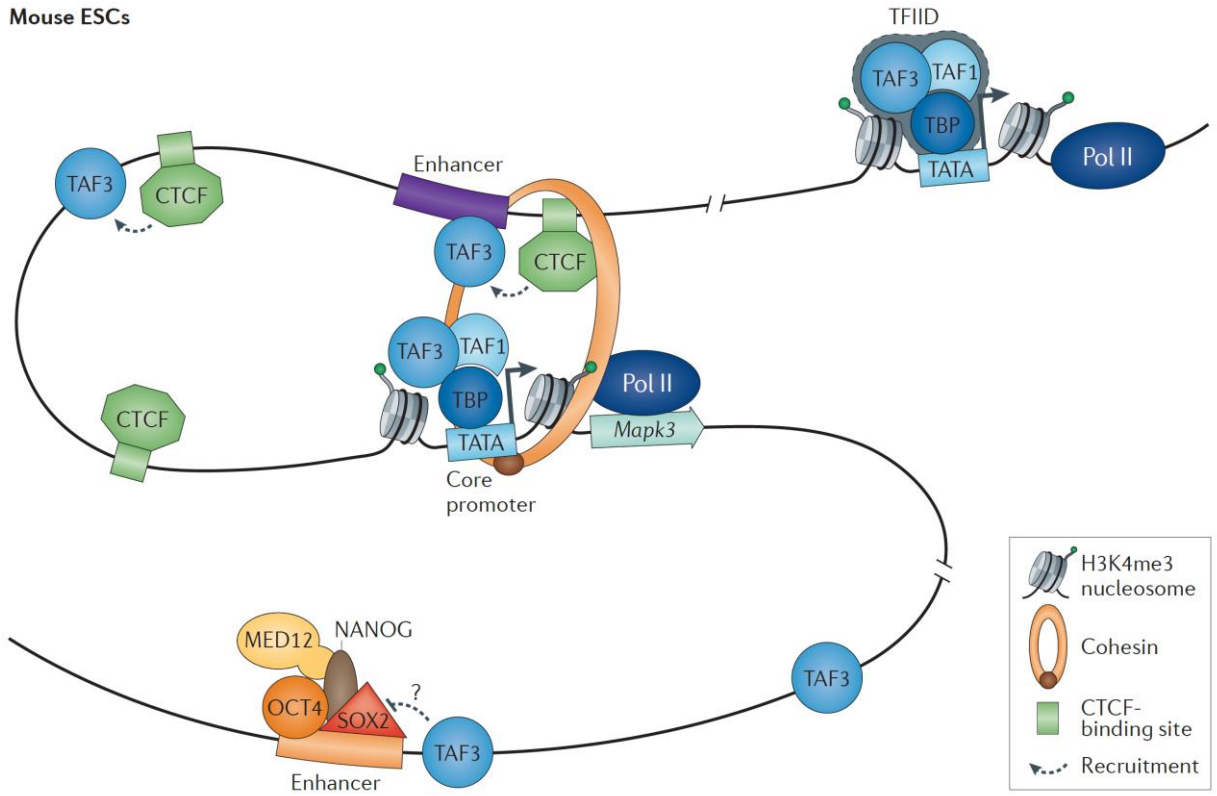
- TADs are regions that keep most physical interactions within
- separated by architectural proteins
- relatively conserved
- experimentally determined by Hi-C method

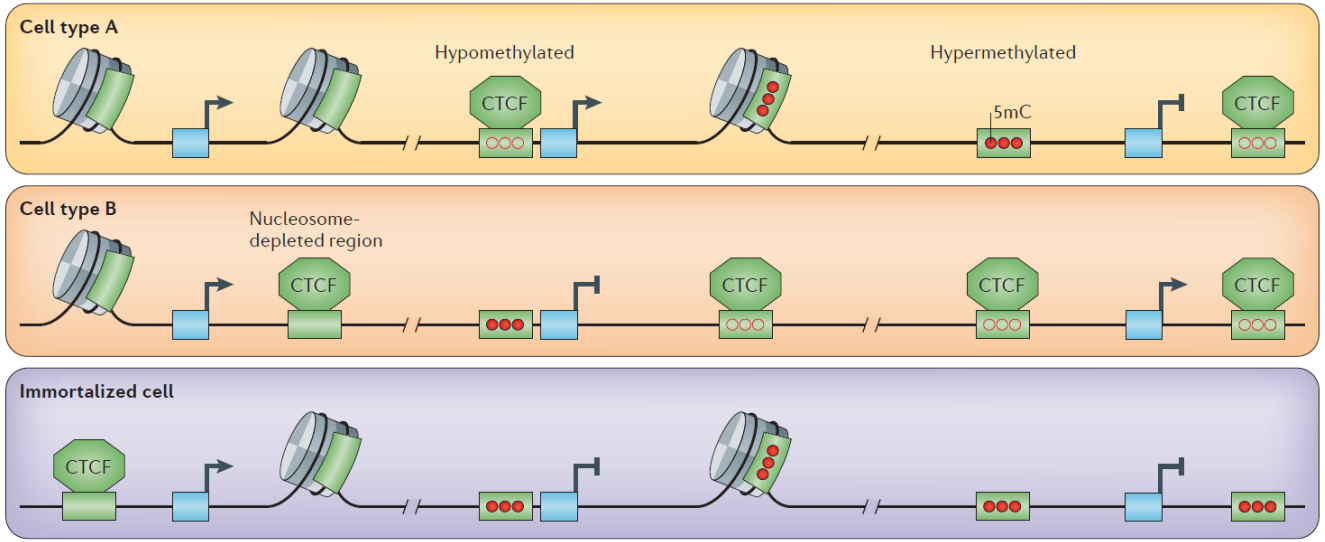




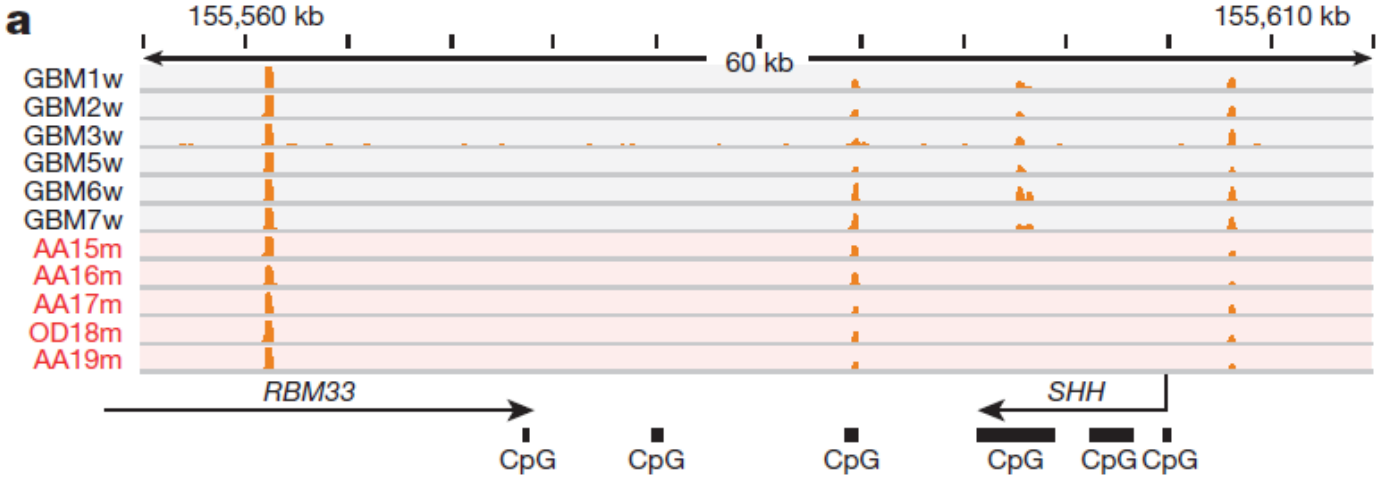
CTCF and cohesin

- CTCF and cohesin are major architectural proteins
- enhancer-promoter contacts usually involves both
- CTCF binding is often controlled by DNA methylation

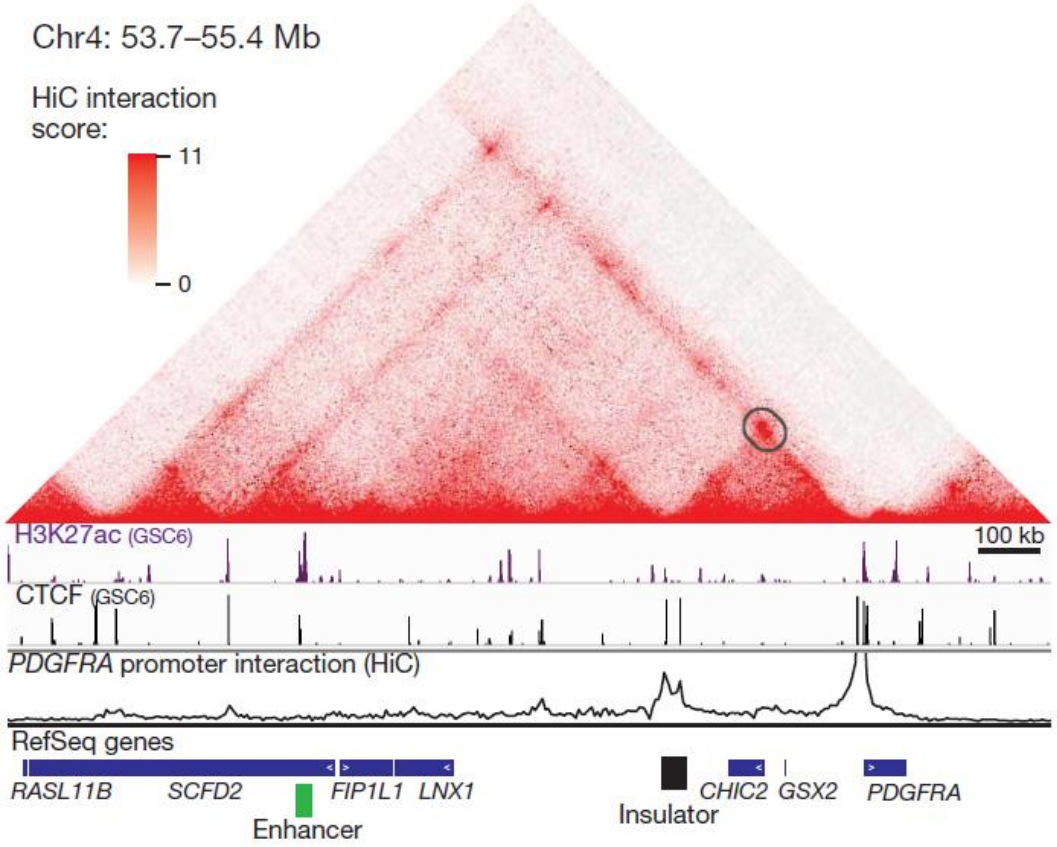




TADs and glioma

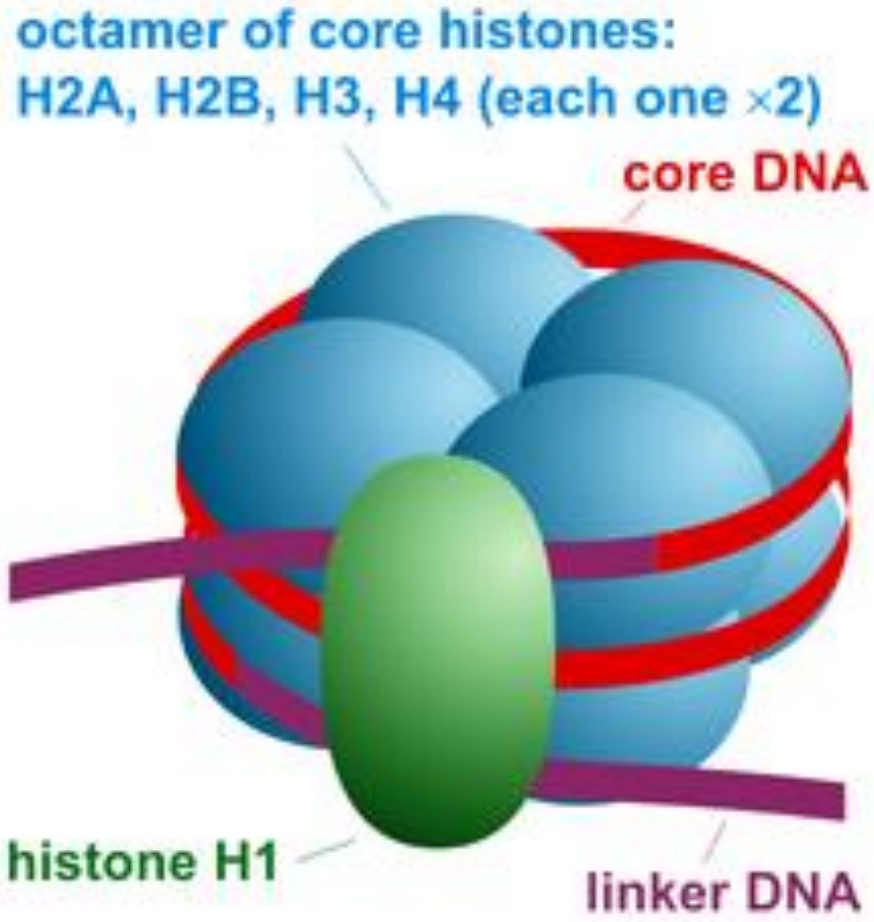


- IDH gain-of-function mutations are known to be driver in gliomas
- Mutant IDH makes new metabolite (2-hydroxyglutarate) that inhibits TET enzymes that demethylate DNA
- methylation disrupts CTCF binding
- enhancers drive up PDGFRA oncogene



Nucleosome

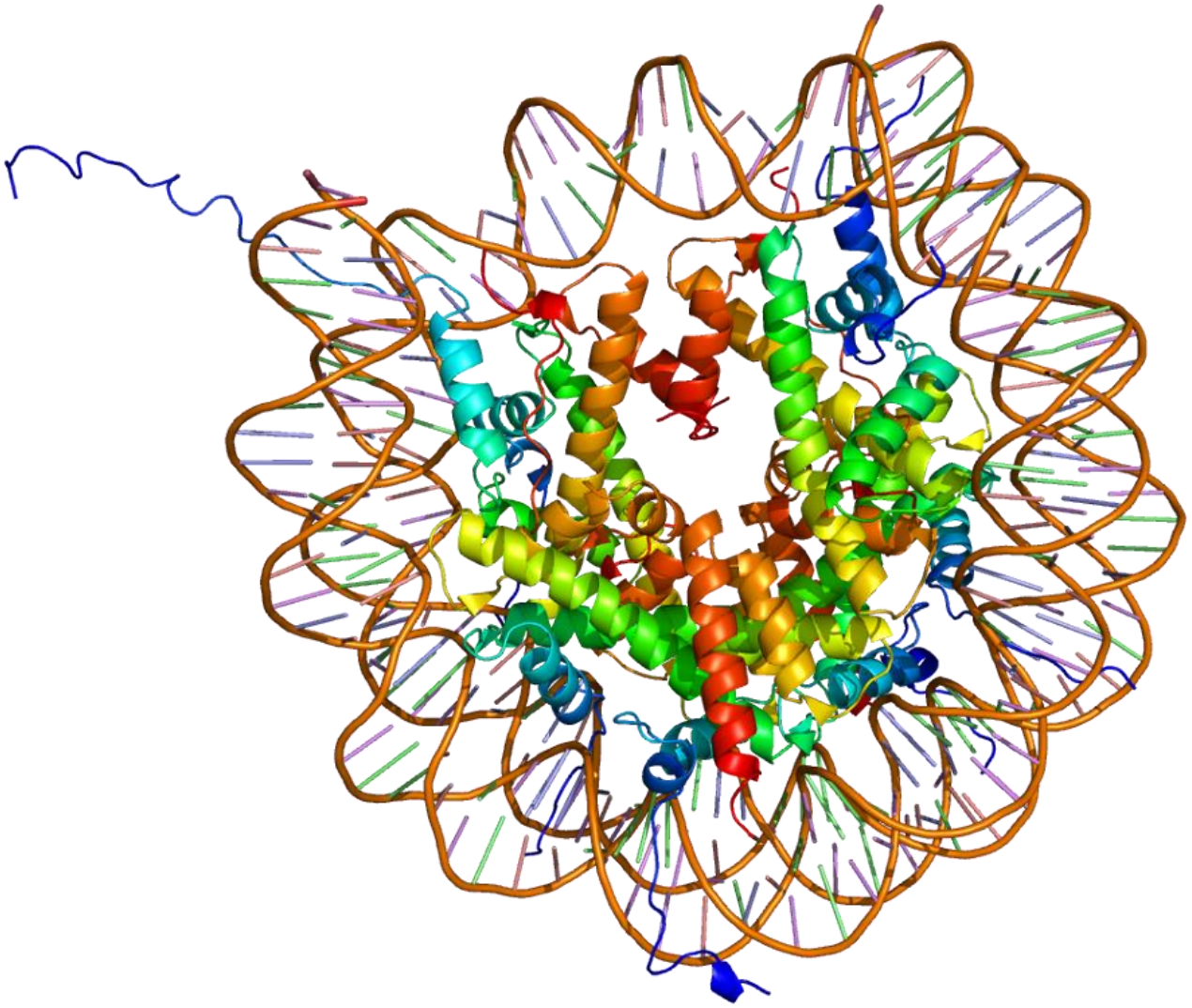
- nucleosomes are made of 8 histone proteins (2x H2A, H2B, H3, and H4) and include 147 nt of DNA
- H1 is a linker histone
- DNA is almost never histone-free, but can be nucleosome-poor (easy access, DNase I digestion)



http://en.wikipedia.org/wiki/Nucleosome

Nucleosome

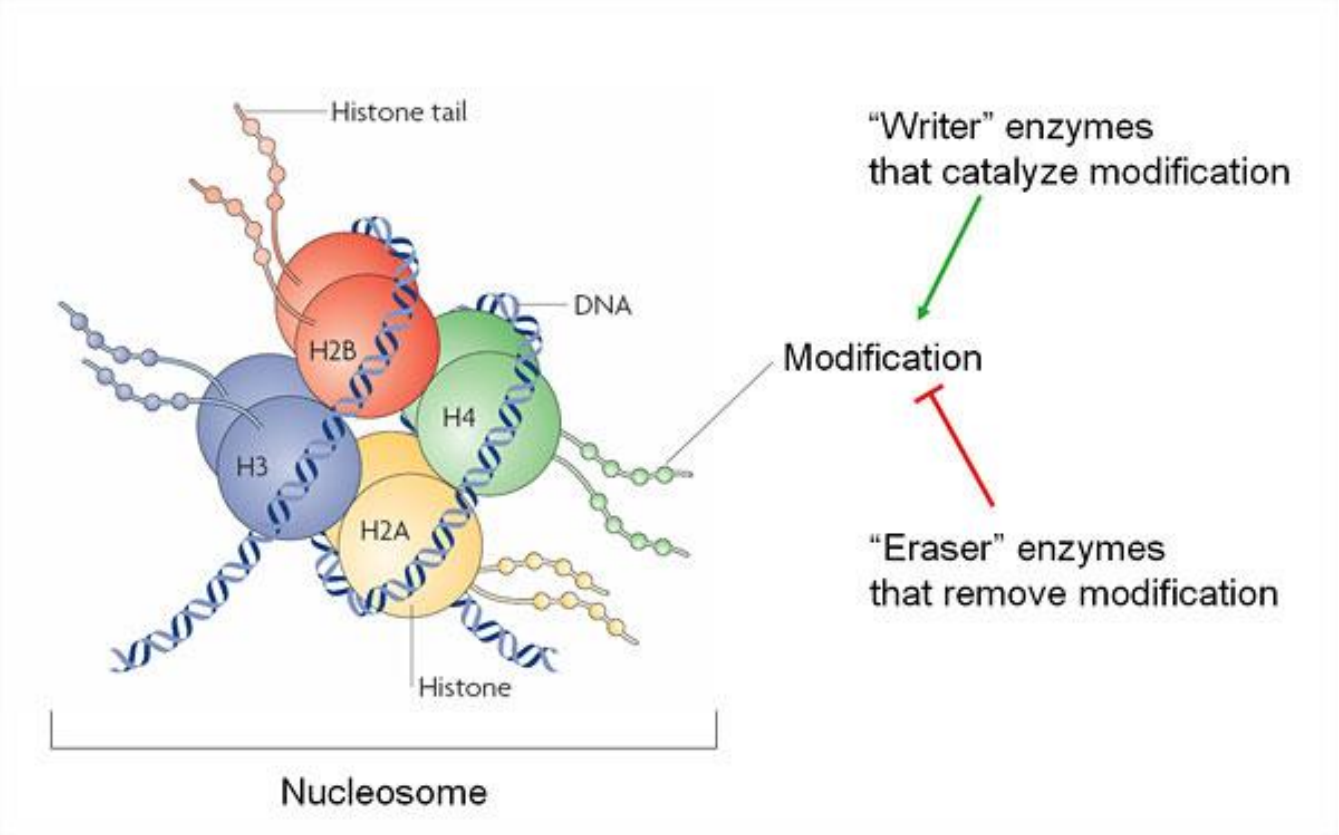
- DNA is negatively charged, so histones have positive charge to achieve neutrality - important for packing!



http://en.wikipedia.org/wiki/Nucleosome

Nucleosome

- histones tails stick outside and can be recognized
- chemical modifications of histones influence DNA accessibility
- histone modifications can be read, erased, and recognized



http://en.wikipedia.org/wiki/Nucleosome

Major protein components of transcription machinery

- transcription of regular protein-coding genes is done by RNA polymerase II
- proteins that attach to DNA are called transcription factors (TFs)

https://www.boundless.com/biology/gene-expression/eukaryotic-transcription-gene-regulation/transcriptional-enhancers-and-repressors/

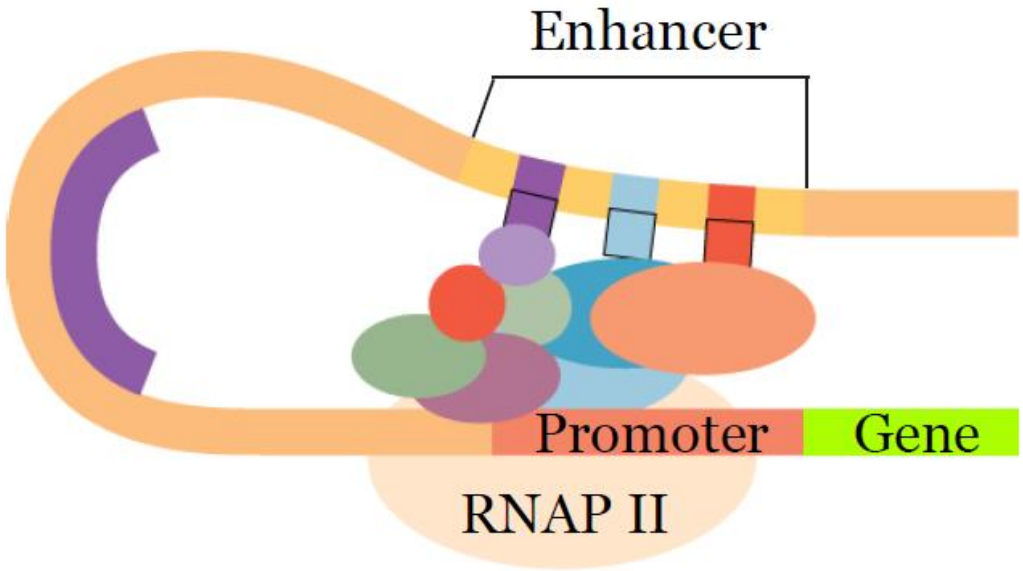
Major protein components of transcription machinery

- transcription of regular protein-coding genes is done by RNA polymerase II
- proteins that attach to DNA are called transcription factors (TFs)
- there are common general transcription factors (GTFs - TFIIA-H) necessary for pre-initiation complex formation
- mediator complex, GTFs, and RNAPII combine into holoenzyme (100+ protein subunits!)

http://en.wikipedia.org/wiki/RNA_polymerase_II_holoenzyme

Major genomic regulators of transcription

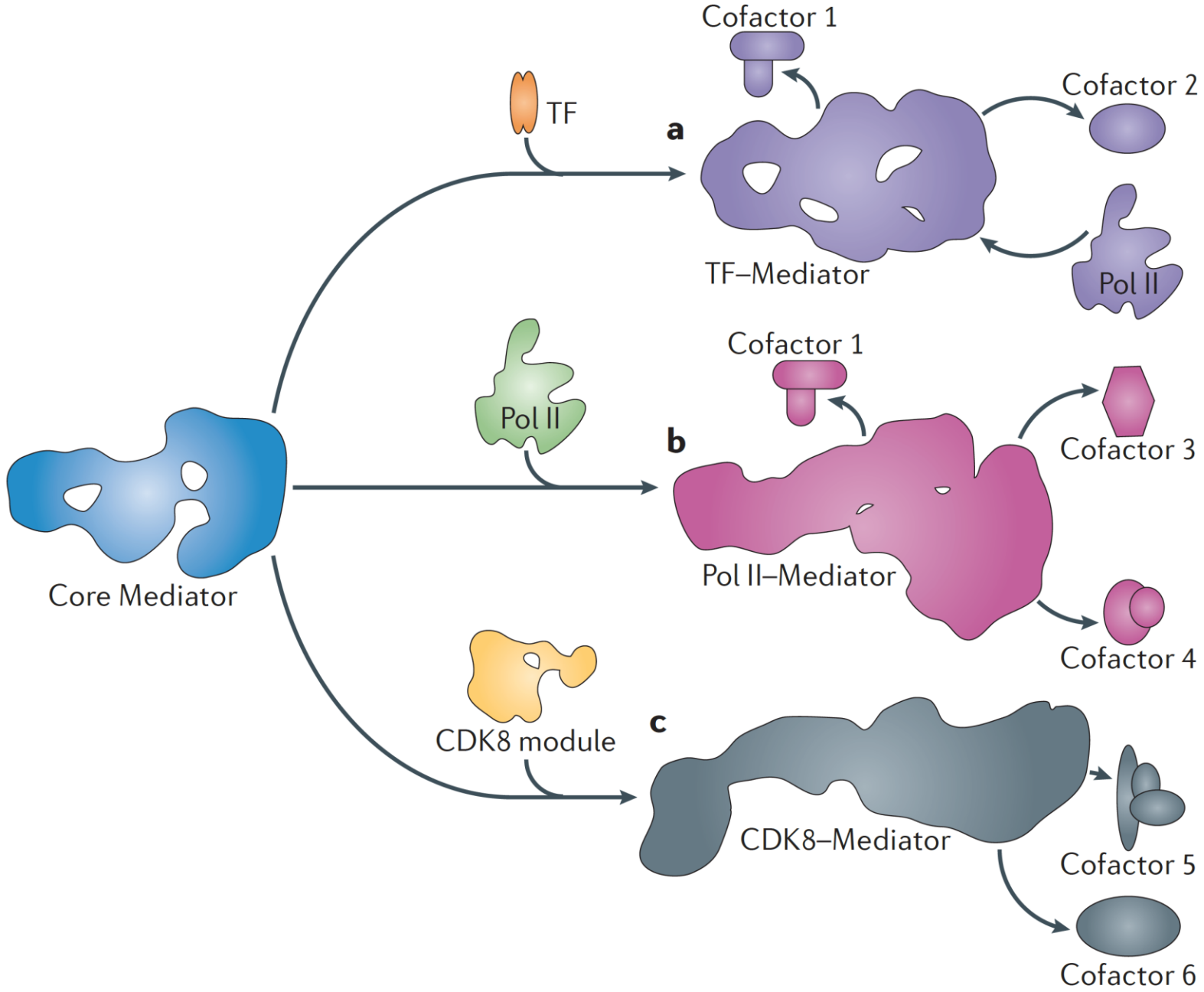
- basal transcription: general transcription factors bind the promoter and RNA polymerase II
- activator proteins bind DNA spots named enhancers
- enhancers are often located far and have to loop



https://www.boundless.com/biology/gene-expression/eukaryotic-transcription-gene-regulation/transcriptional-enhancers-and-repressors/

Mediator complex

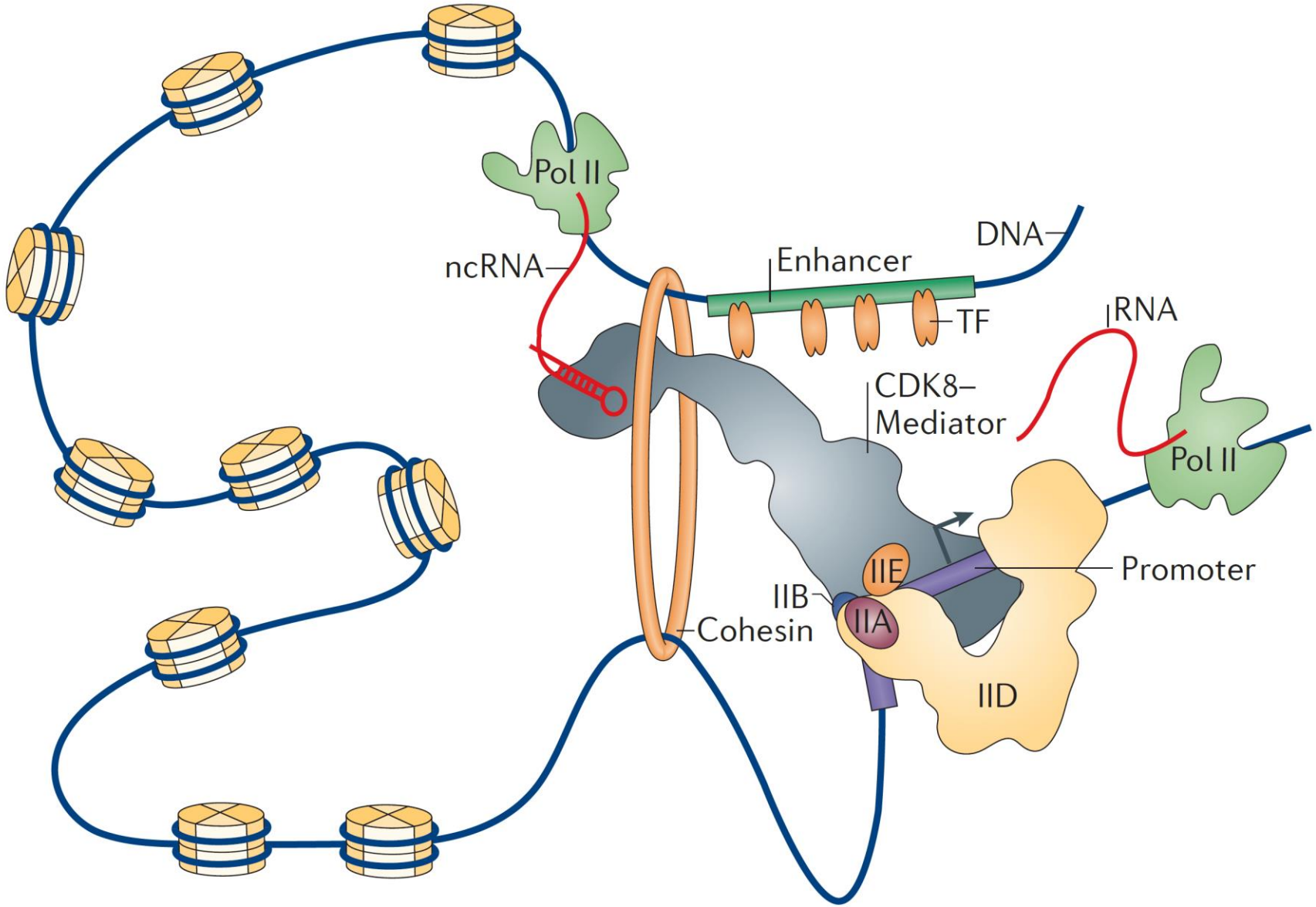
- many intrinsically disordered regions
- complexity is thought to be an evolutionary advantage
- mediator is involved in most active transcription



https://www.ncbi.nlm.nih.gov/pubmed/25693131

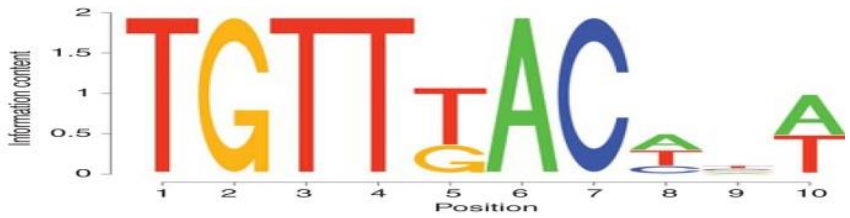
Mediator and loops

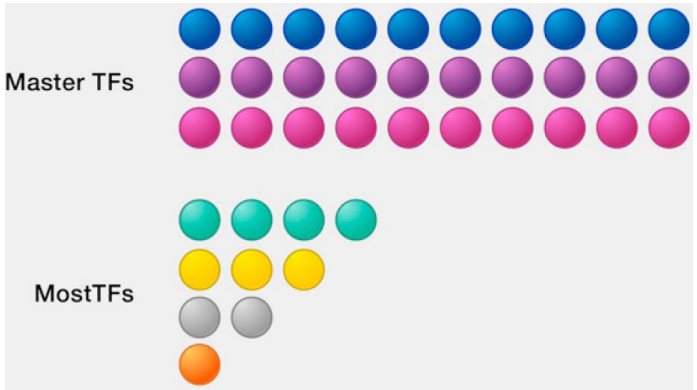
- mediator does not facilitate long range enhancer-promoter loops in yeast

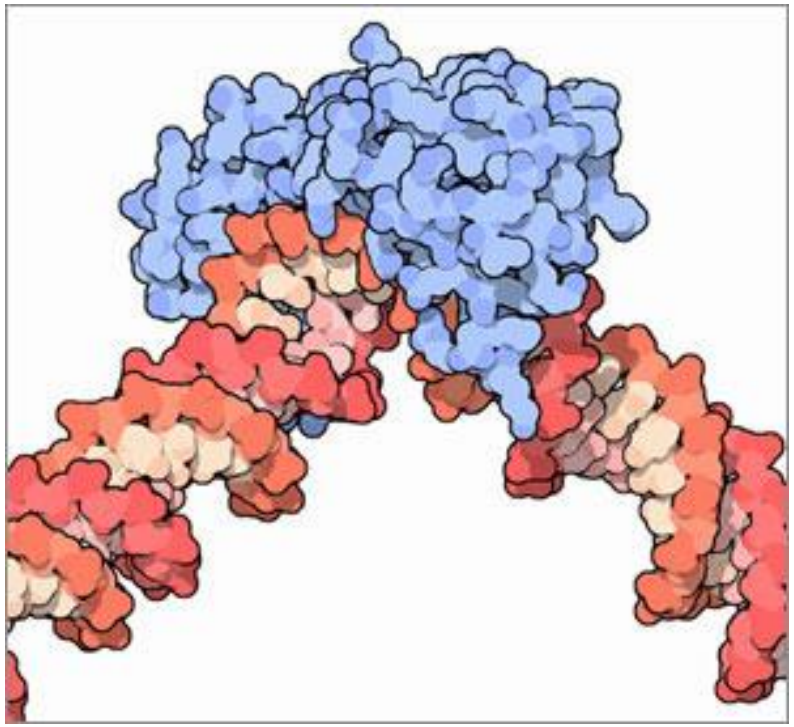


Transcription factors

- there are estimated 1,500 transcription factors in humans
- binding motif represented by consensus sequence



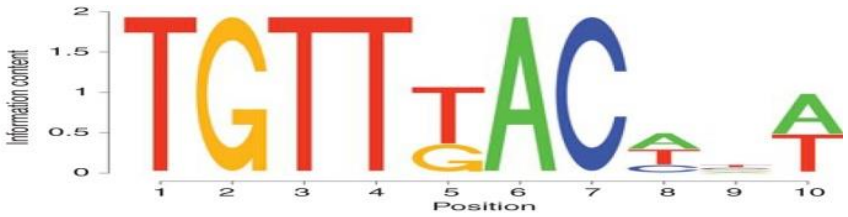


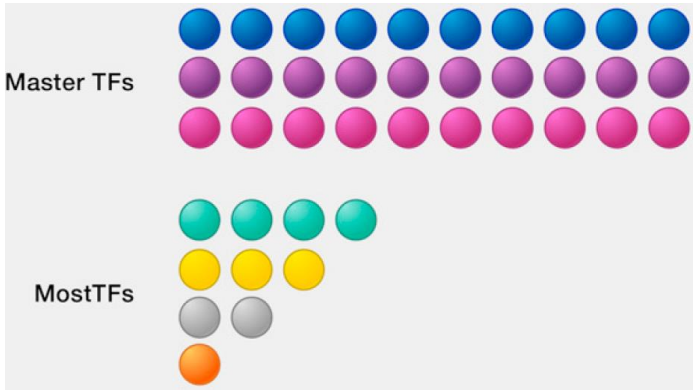


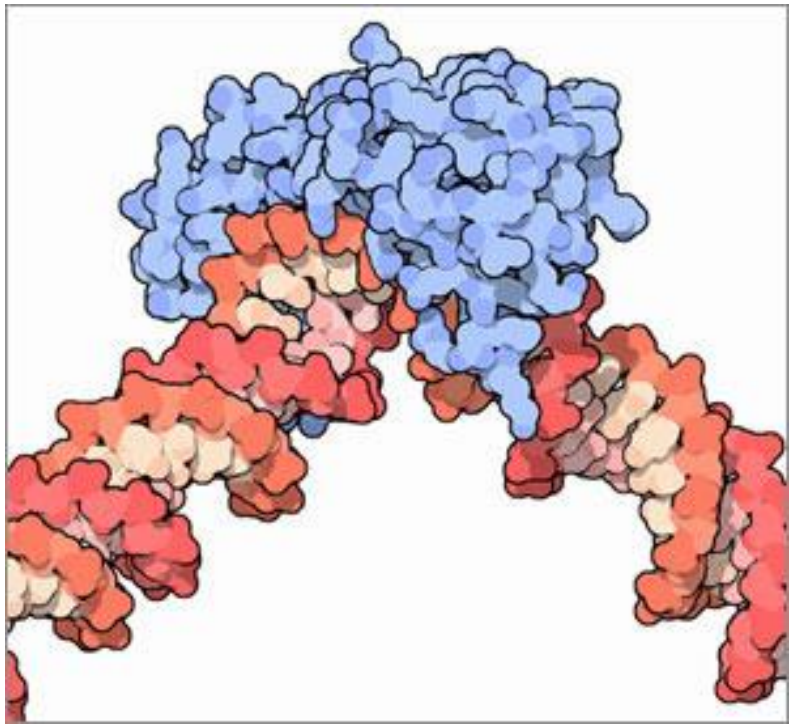
http://www.broadinstitute.org/education/glossary/transcription-factor

Transcription factors

- there are estimated 1,500 transcription factors in humans
- binding motif represented by consensus sequence
- master regulators exist but are not always known
- mechanisms of expression modulation involve:
  - stabilize/block RNAP II binding to DNA
  - catalyze histone acetylation or deacetylation
  - recruit coactivator or corepressor



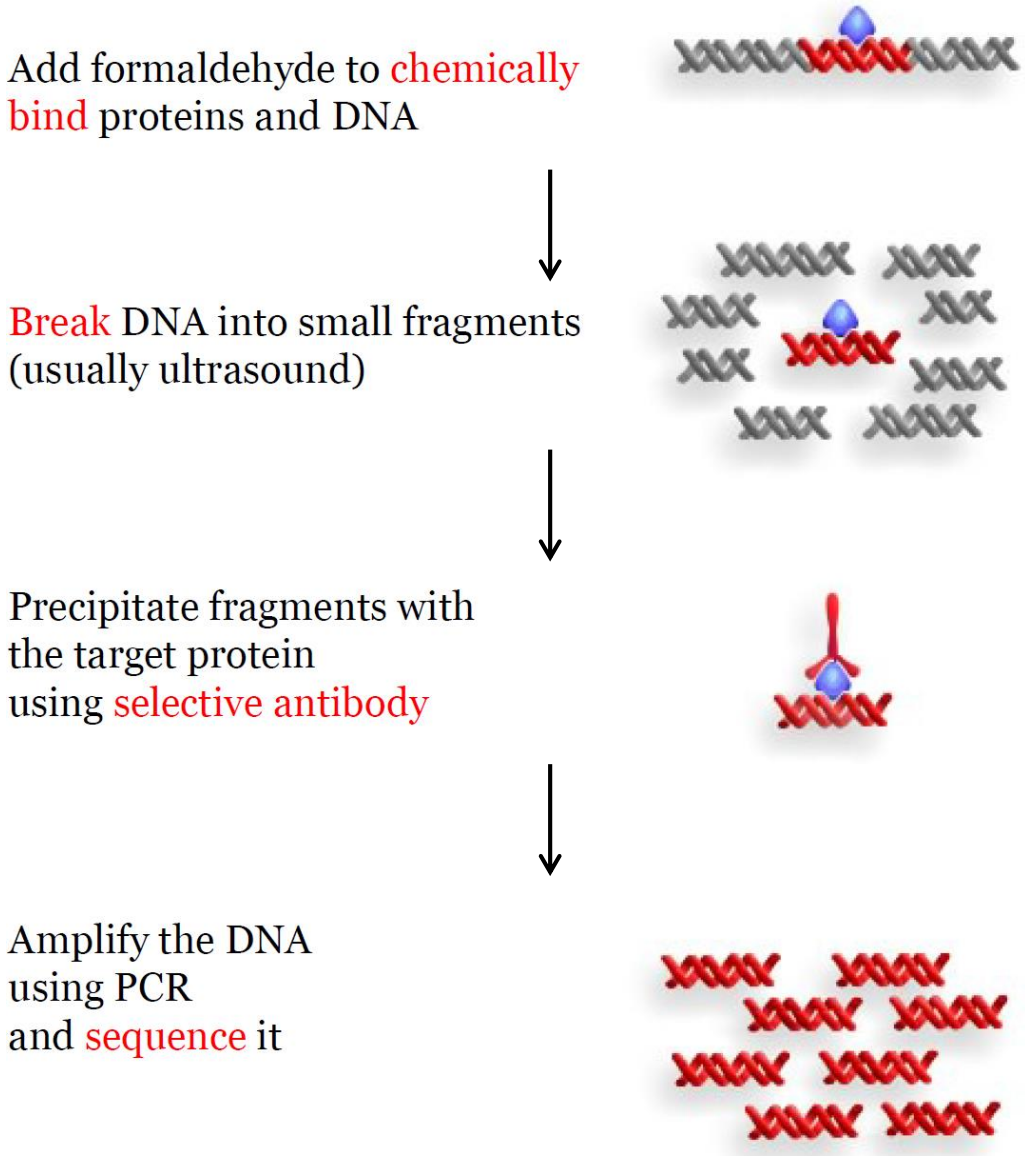




http://www.broadinstitute.org/education/glossary/transcription-factor

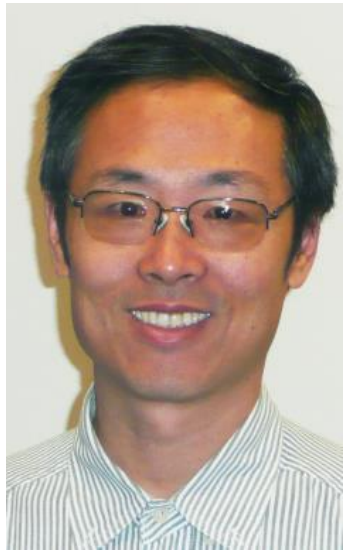
Chromatin Immunoprecipitation (ChIP)

- simple but effective: reversible chemical cross-linking of proteins to DNA



http://www.bio.brandeis.edu/haberlab/jehsite/chIP.html

Who came first?



Science. 2007 Jun 8;316(5830):1497-502. Epub 2007 May 31.

Genome-wide mapping of in vivo protein-DNA interactions.

Johnson DS[1], Mortazavi A, Myers RM, Wold B.

Author information

Abstract

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element-1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [+/-50 base pairs (bp)], which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity [ROC (receiver operator characteristic) area >/= 0.96] and statistical confidence (P <10(-4)), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

Cell

Volume 129, Issue 4, 18 May 2007, Pages 823–837

Resource

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski[1, 3], Suresh Cuddapah[1, 3], Kairong Cui[1, 3], Tae-Young Roh[1, 3], Dustin E. Schones[1, 3], Zhibin Wang[1, 3], Gang Wei[1, 3], Iouri Chepelev[2], Keji Zhao[1].

nature

ARTICLES

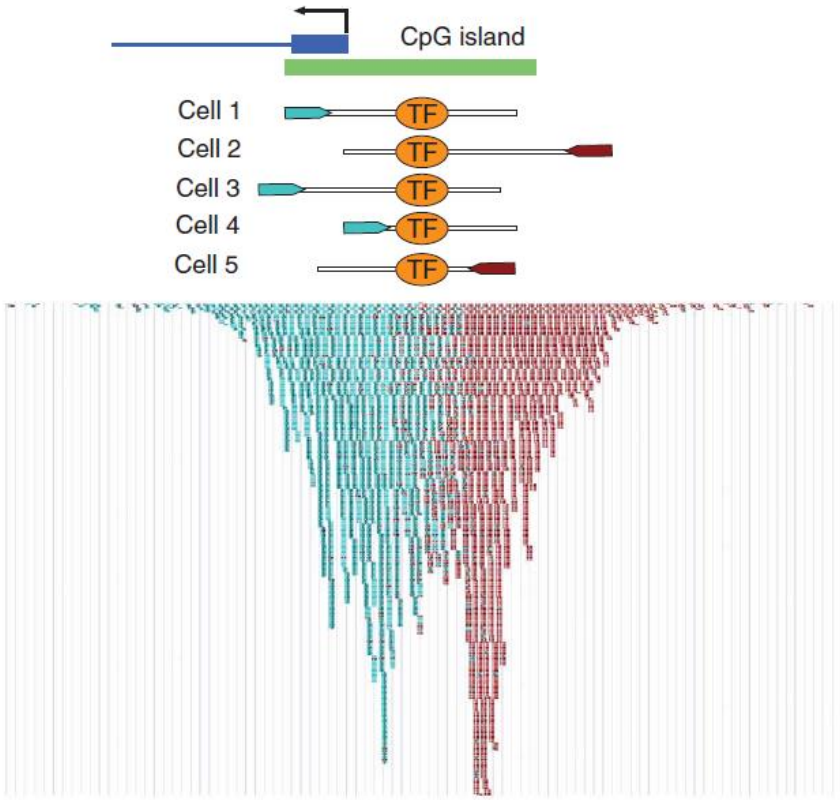
Genome-wide maps of chromatin state in pluripotent and lineage-committed cells

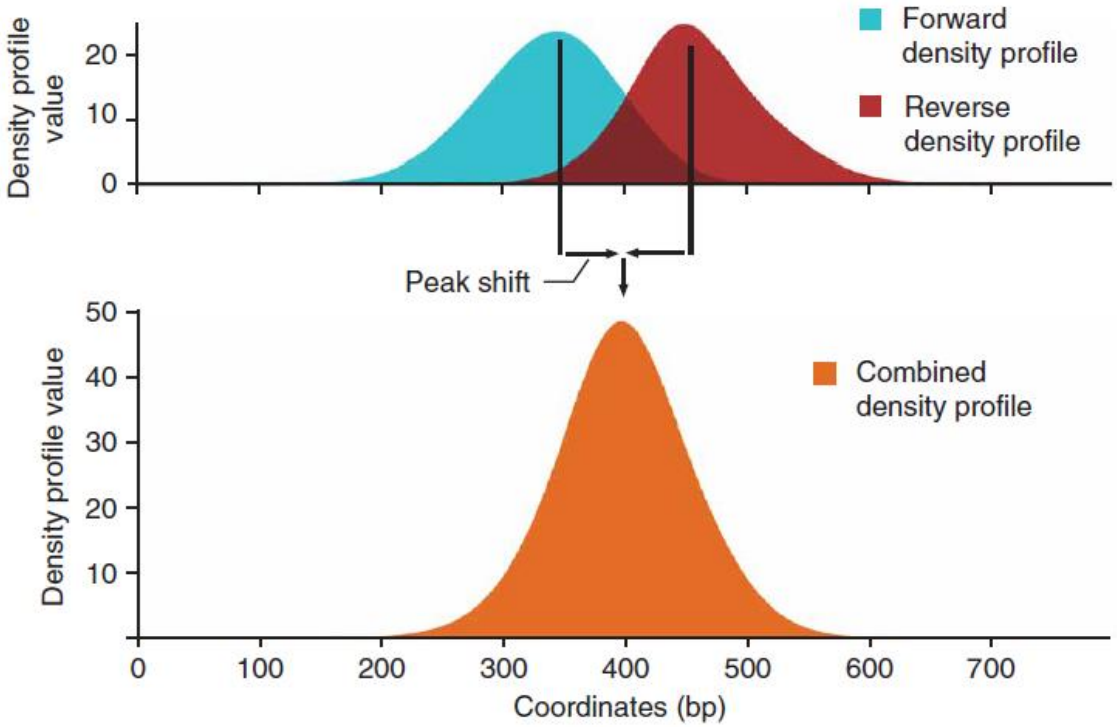
Tarjei S. Mikkelsen[1,2], Manching Ku[1,4], David B. Jaffe[1], Biju Issac[1,4], Erez Lieberman[1,2], Georgia Giannoukos[1], Pablo Alvarez[1], William Brockman[1], Tae-Kyung Kim[5], Richard P. Koche[1,2,4], William Lee[1], Eric Mendenhall[1,4], Aisling O'Donovan[4], Aviva Presser[1], Carsten Russ[1], Xiaohui Xie[1], Alexander Meissner[3], Marius Wernig[3], Rudolf Jaenisch[3], Chad Nusbaum[1], Eric S. Lander[1,3*] & Bradley E. Bernstein[1,4,6*]

http://www.snarkyscientist.com/2013/06/19/the-history-of-the-biggest-technique-of-2009-who-invented-chip-seq/

ChIP-Seq

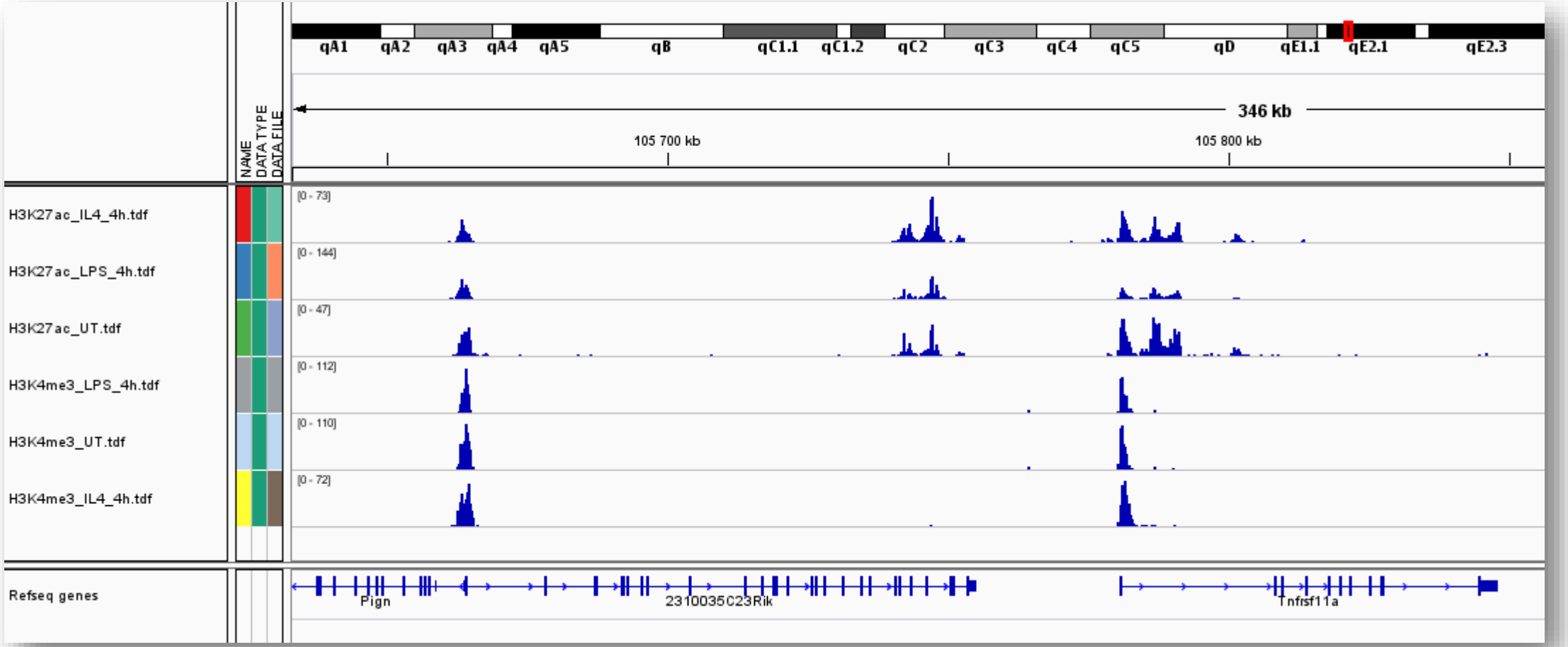
- DNA library obtained after ChIP can be amplified and sequenced
- ChIP-Seq can be used for both transcription factors and histone modifications, like H3K4me3





http://www.avadis-ngs.com/features/chip_seq

A typical ChIP-seq experiment

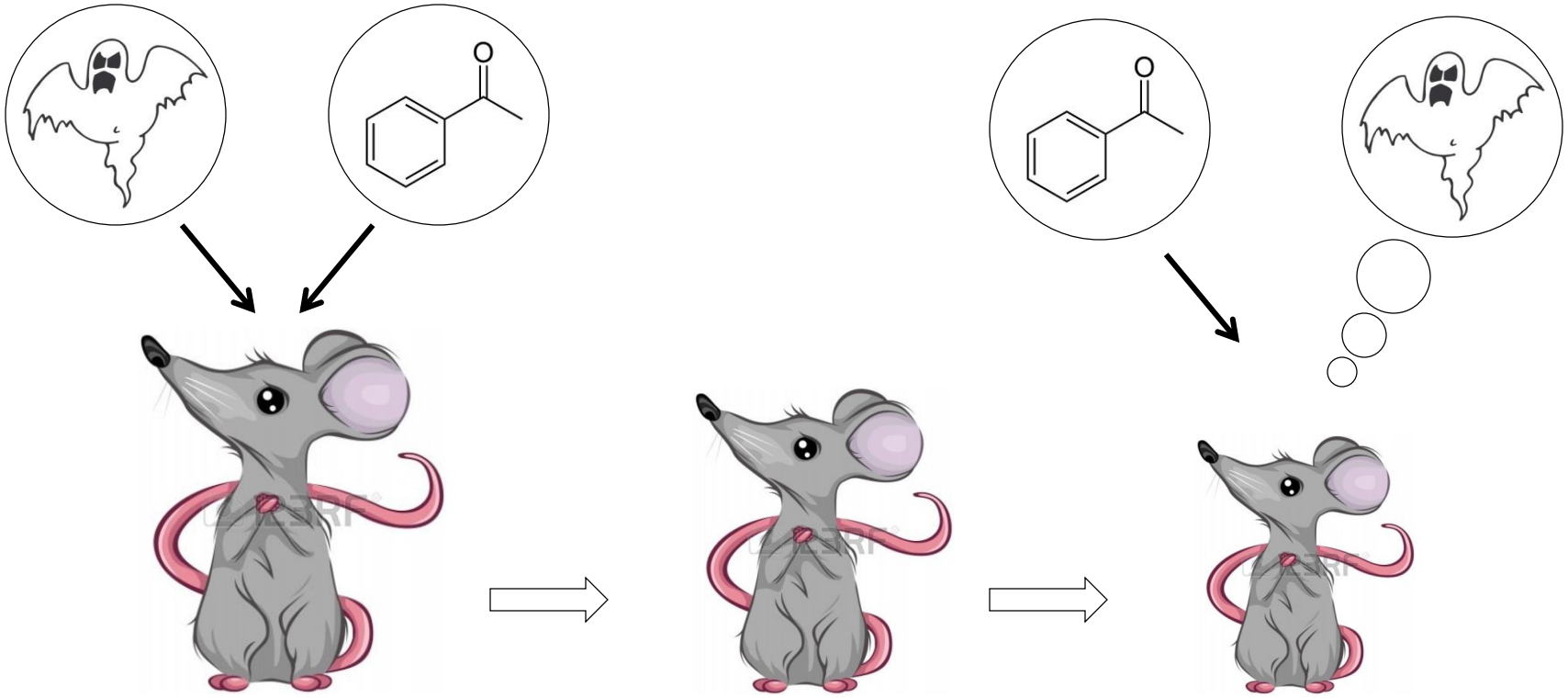


Epigenetics

- Two major definitions that mean very different things:
  - Heritable changes that do not involve DNA modification (still heavily debated)
  - DNA modification and packaging changes that influence expression (well established)
- Very different mechanisms and meaning in
  - differentiating cells
  - pluripotent cells
  - differentiated cells

Fear and loathing in Mouse-Vegas

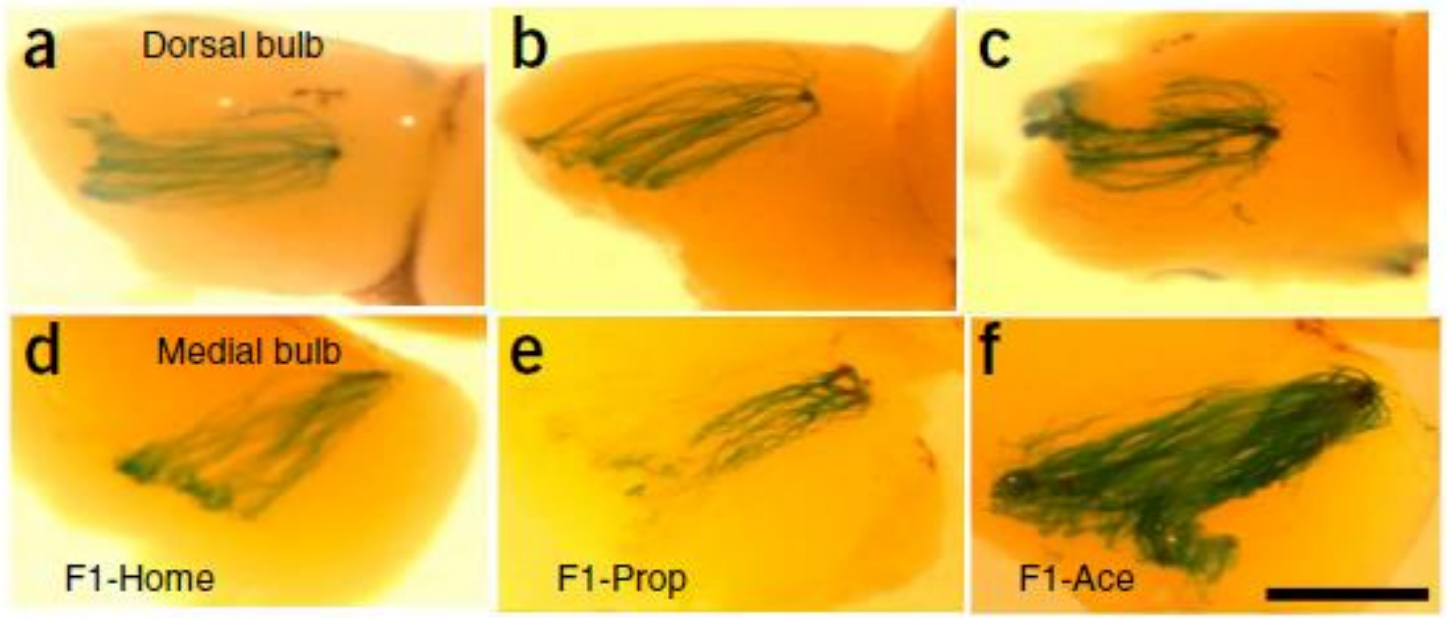
- odor fear conditioning (acetophenone)
- the effect lasted 2 generations

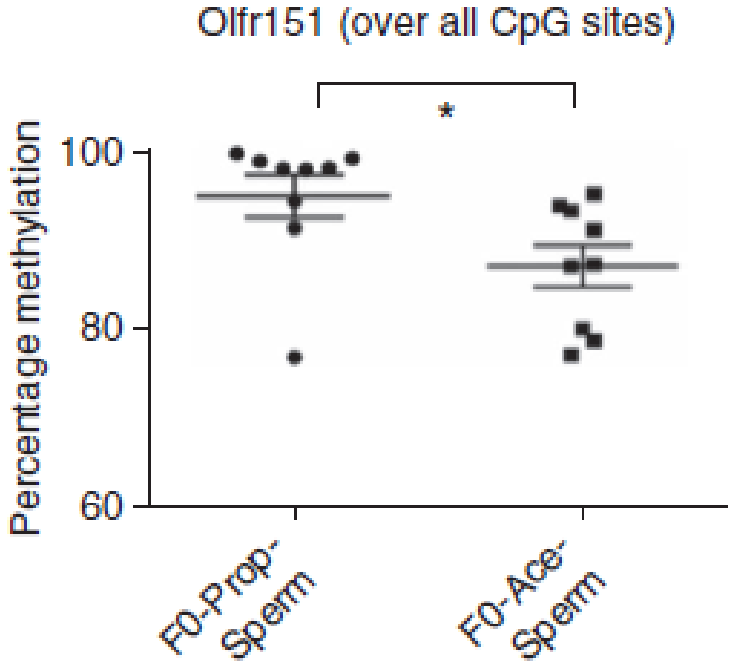


http://www.123rf.com/photo_7535422_scared-mouse-looking-up.html
Dias B.G.; Ressler K.J., Nature Neuroscience, 17(1), pp 89-96 (2014)

Mechanism: *Olf151* promoter hypomethylation

- activates *Olf151* that expresses olfactory glomerulus M71
- bisulfite sequencing revealed that Olf151 was hypomethylated at CpG promoter

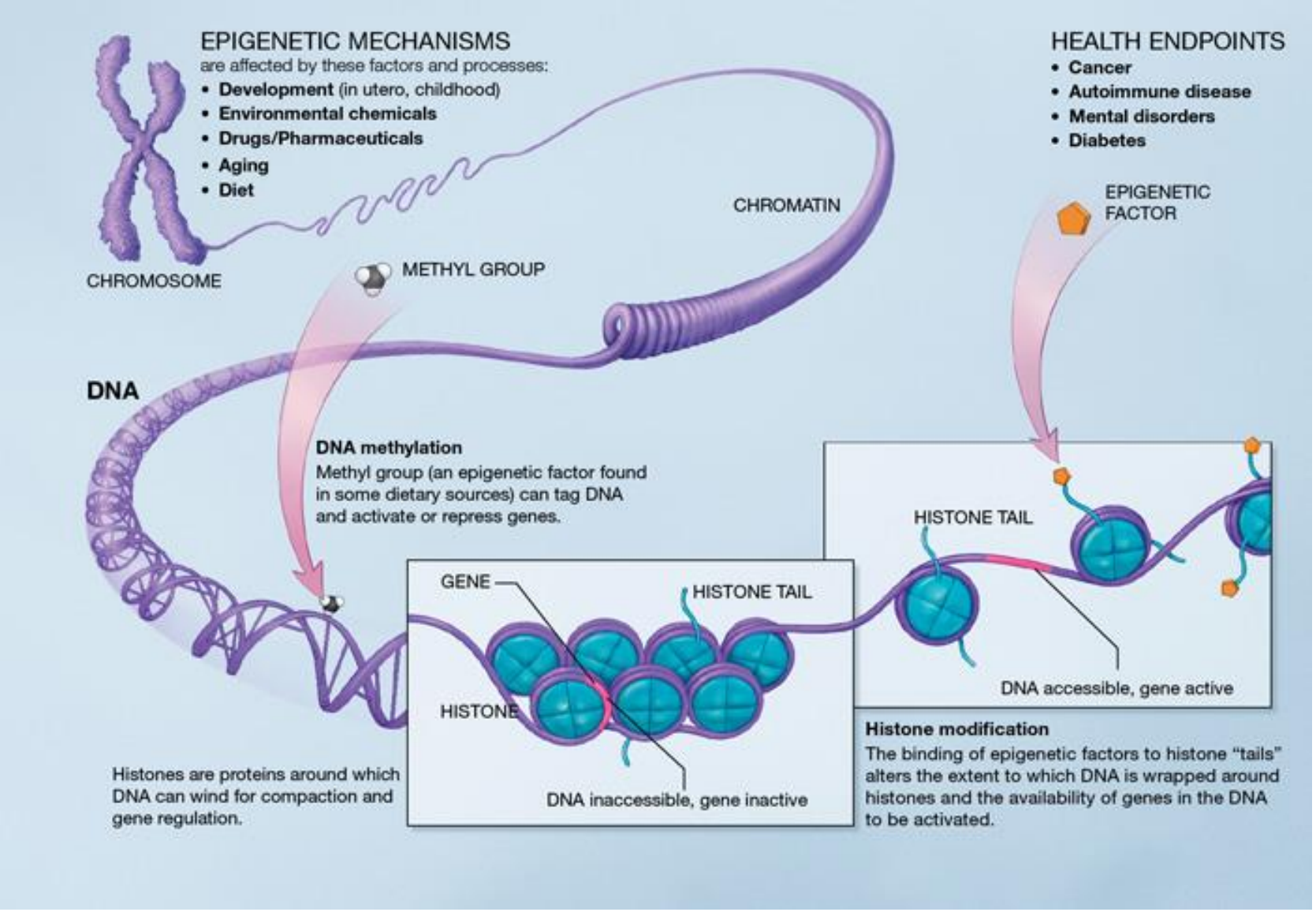




Dias B.G.; Ressler K.J., Nature Neuroscience, 17(1), pp 89-96 (2014)

Epigenetic regulation

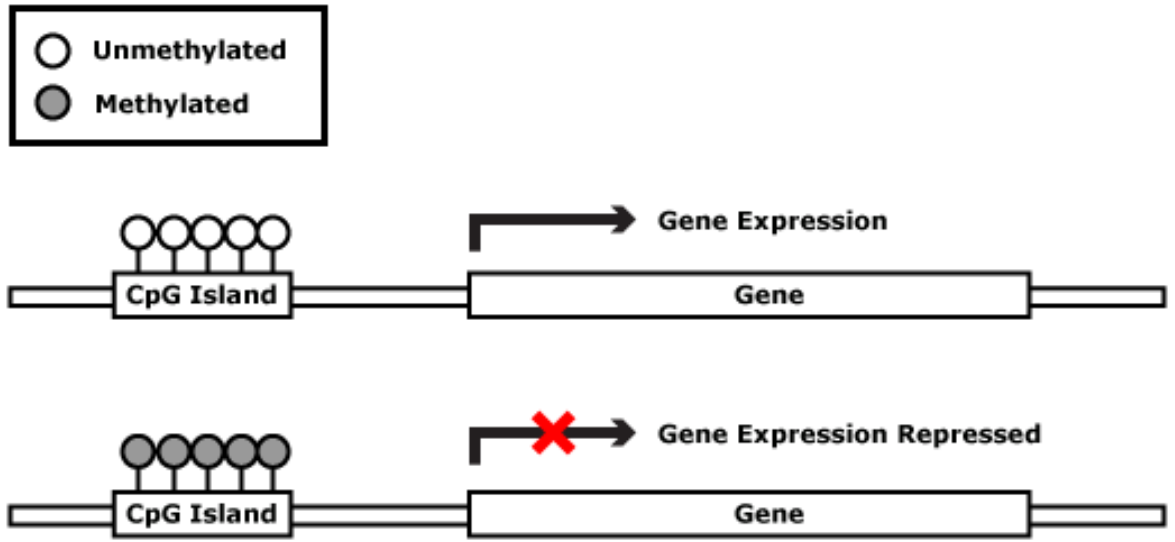
- histone modifications
- DNA methylation
- noncoding RNA



http://en.wikipedia.org/wiki/Epigenetics

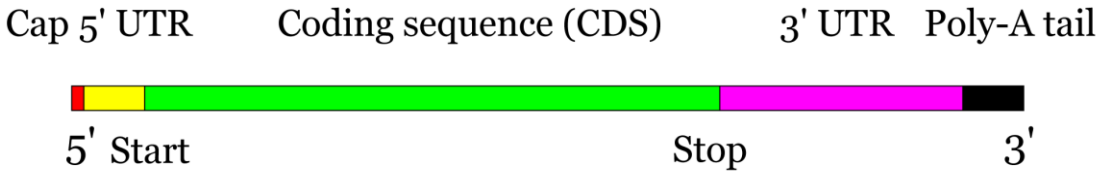
Promoters and CpG islands

- many human promoters are CpG rich (two types of promoters?)
- in the rest of the genome GC mutates to GT
- methylation of the promoter CpG silences the gene

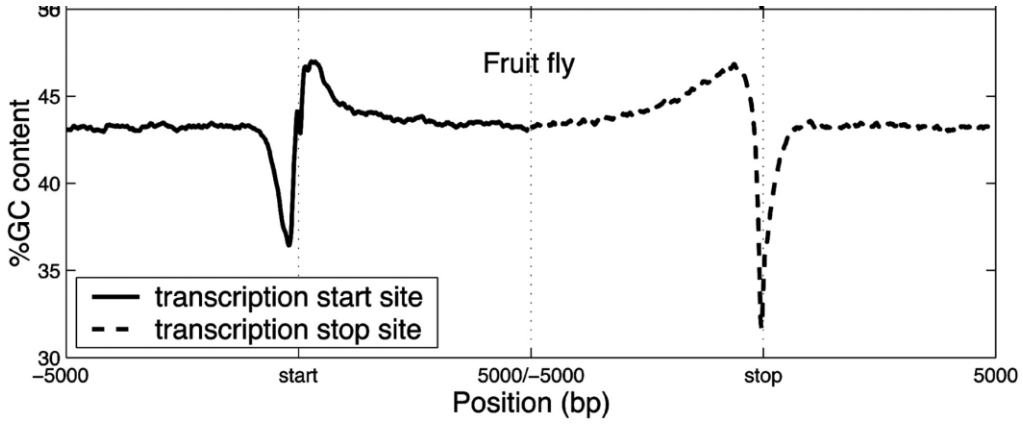


http://missinglink.ucsf.edu/lm/genes_and_genomes/methylation.html

Evolution: Human vs. Fruit Fly



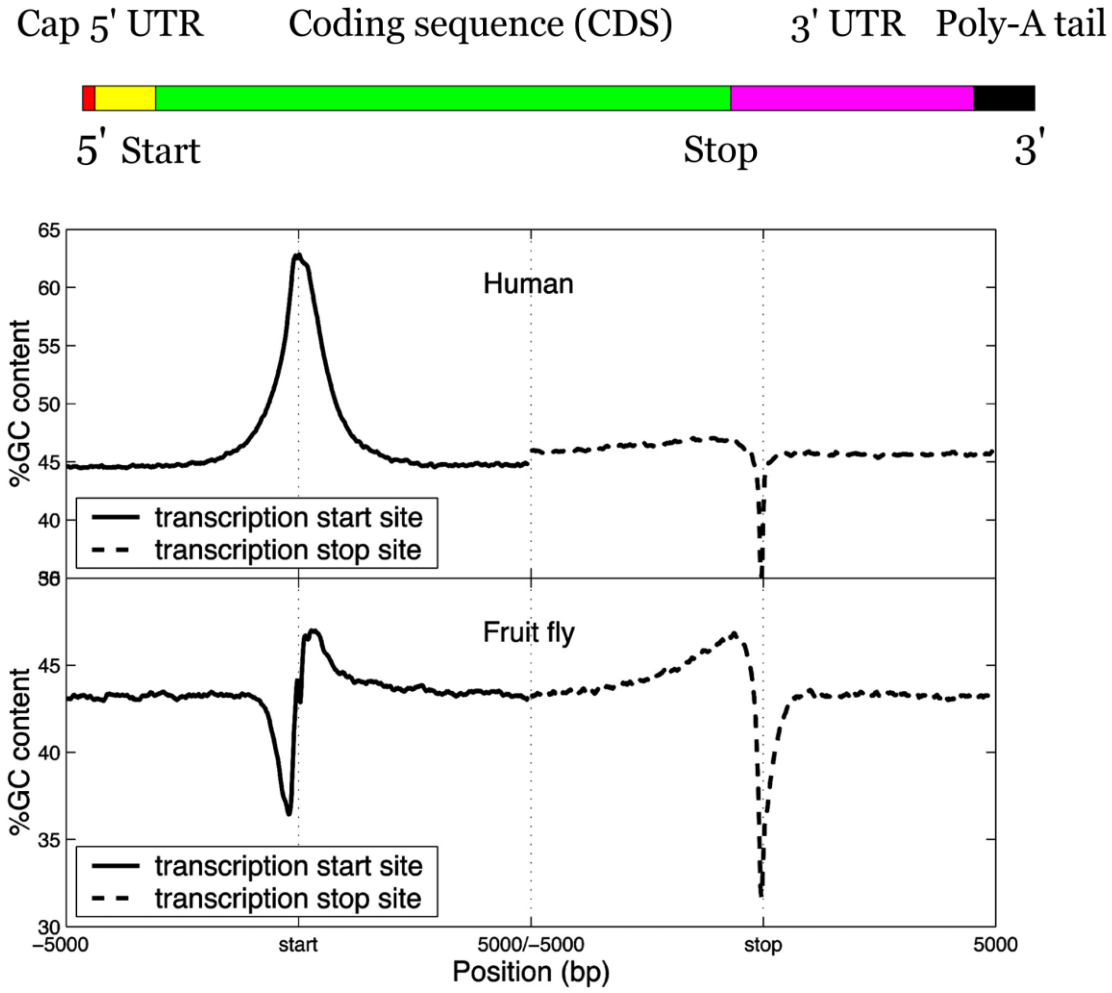
- Drosophila is a "Dnmt2 only" organism
- no Dnmt1 and Dnmt3
- no functional homologs of known 5-methylcytosine reader proteins



http://en.wikipedia.org/wiki/Messenger_RNA
Zhang L. et al, Proc Nat Acad Sci USA, 101, pp 16855-16860 (2004)

Evolution: Human vs. Fruit Fly

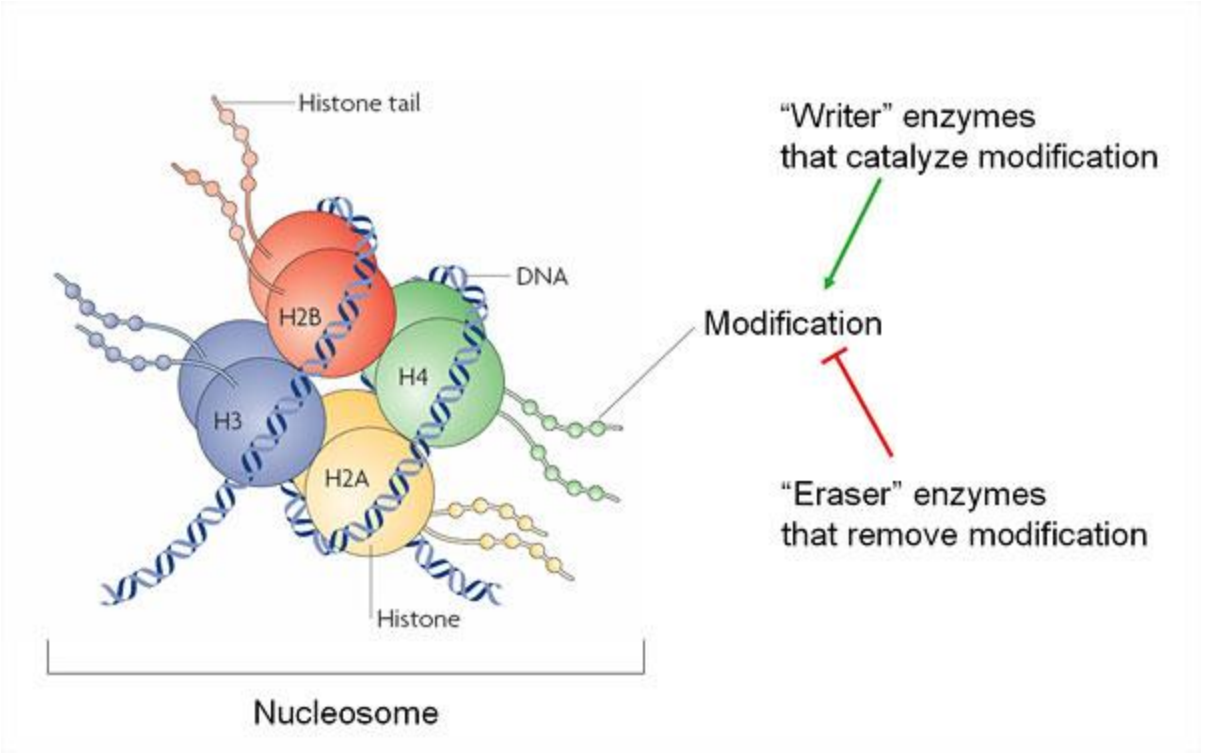
- Drosophila is a "Dnmt2 only" organism
- no Dnmt1 and Dnmt3
- no functional homologs of known 5-methylcytosine reader proteins
- human promoter GC% indicates pronounced conservation

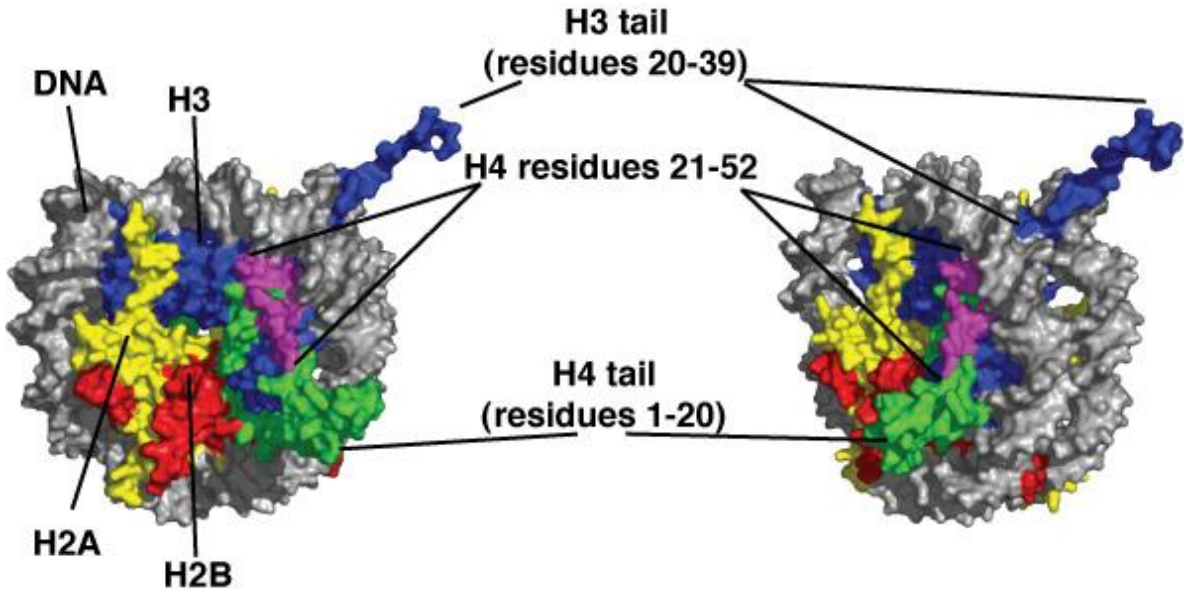


http://en.wikipedia.org/wiki/Messenger_RNA
Zhang L. et al, Proc Nat Acad Sci USA, 101, pp 16855-16860 (2004)

Histone tail modifications

- histone tail can be chemically modified
- are in fact normal post-translational modifications
- histone modifications change transcription

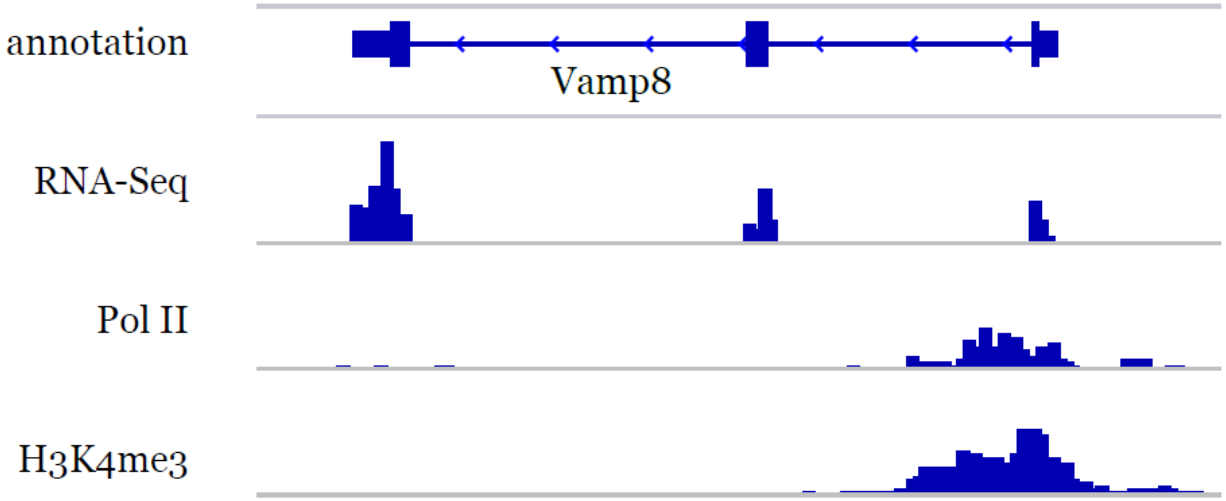


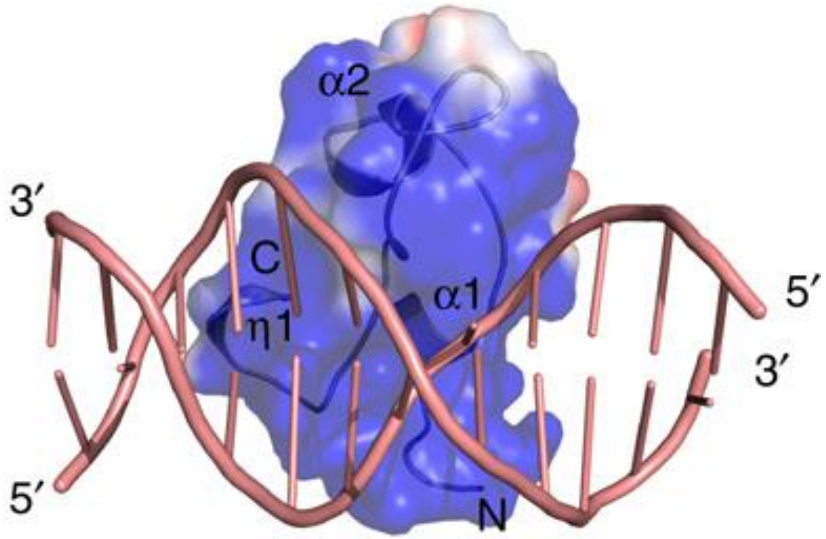


Promoter histone marks

- narrow peaks of H3K4me3 mark promoters.
- enzyme that methylates K4 binds only to non-CpG-methylated promoters!

Mouse heart



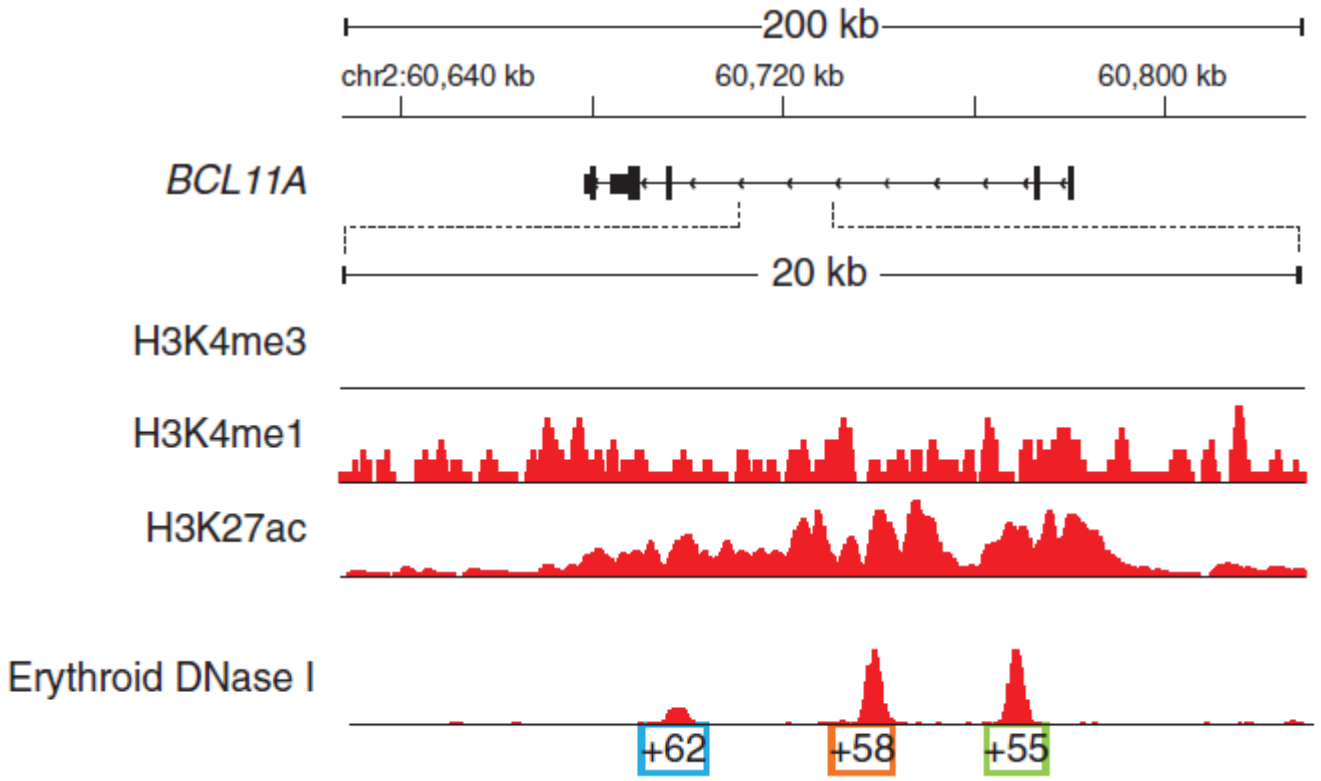


The EMBO Journal, 31, pp 3130–3146 (2012)
C. Xu et al, Nature Communications , 2(227),pp 1-8 (2011)

Enhancer histone marks

- enhancers are associated with H3K4me1 and H3K27ac
- H3K27ac is thought to distinguish active enhancers from poised

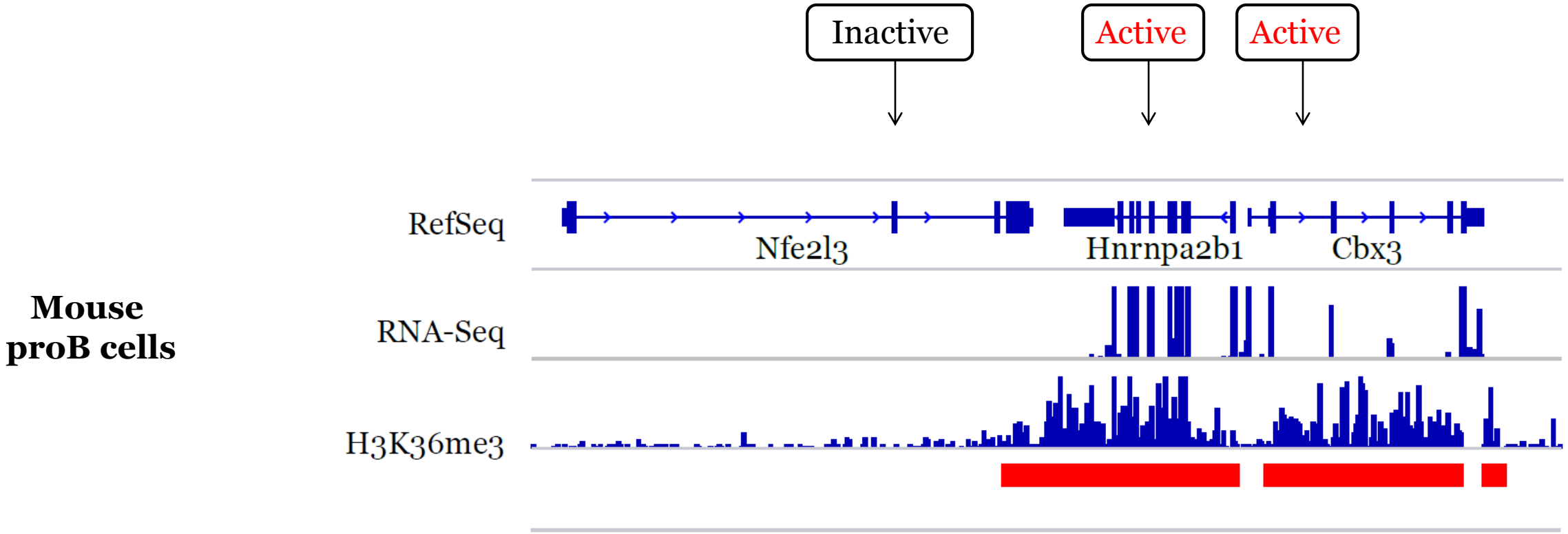
Human erythroblasts



Bauer, D.E.; Kamran, S.C. et al, Science, 342(6155), pp 253-7 (2013)

Marks of transcriptional elongation

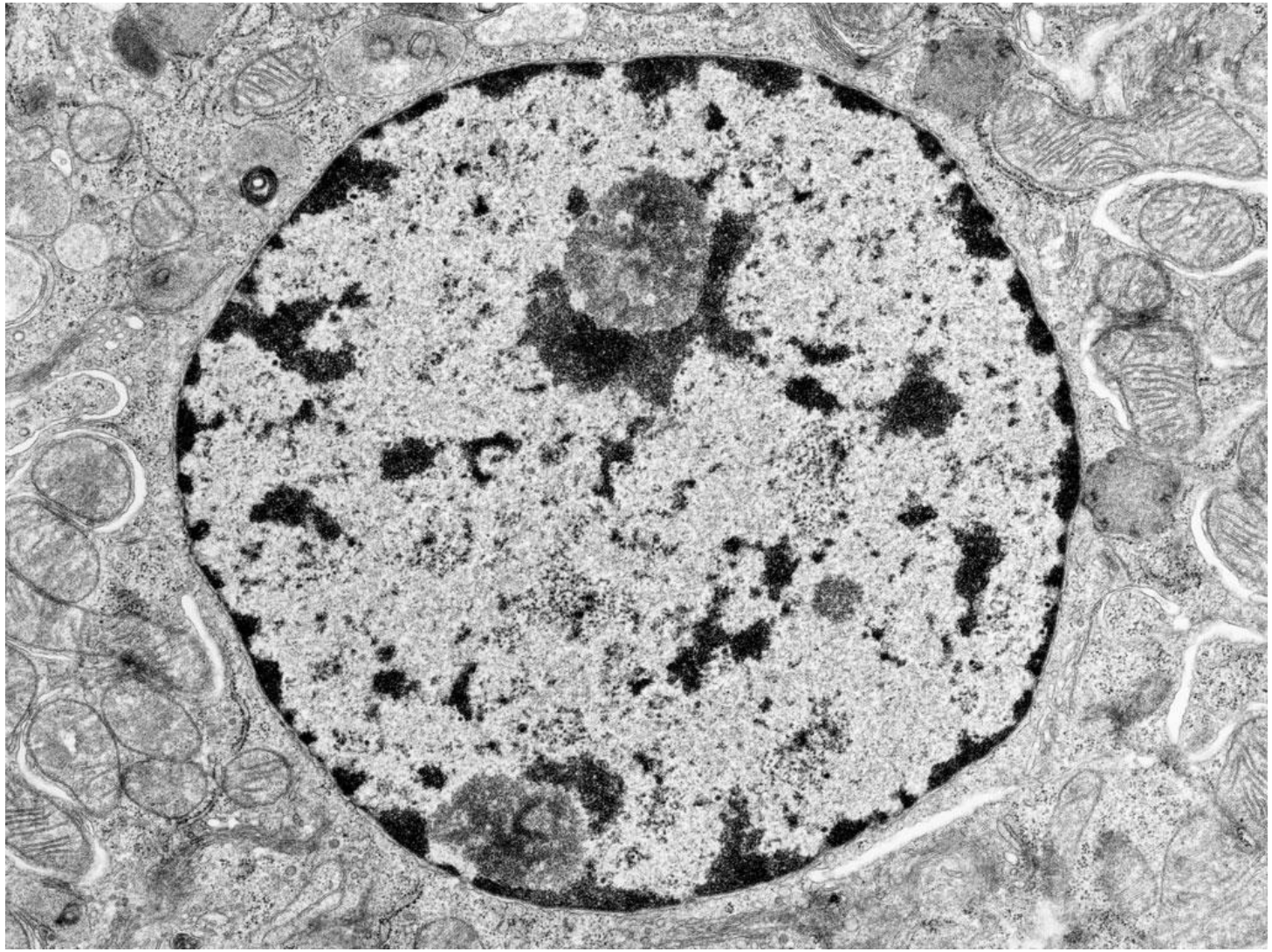
- Elongation is marked my H3K36me3 and H3K79me2



Proc Natl Acad Sci U S A, 107(50), pp 21931-6 (2010)

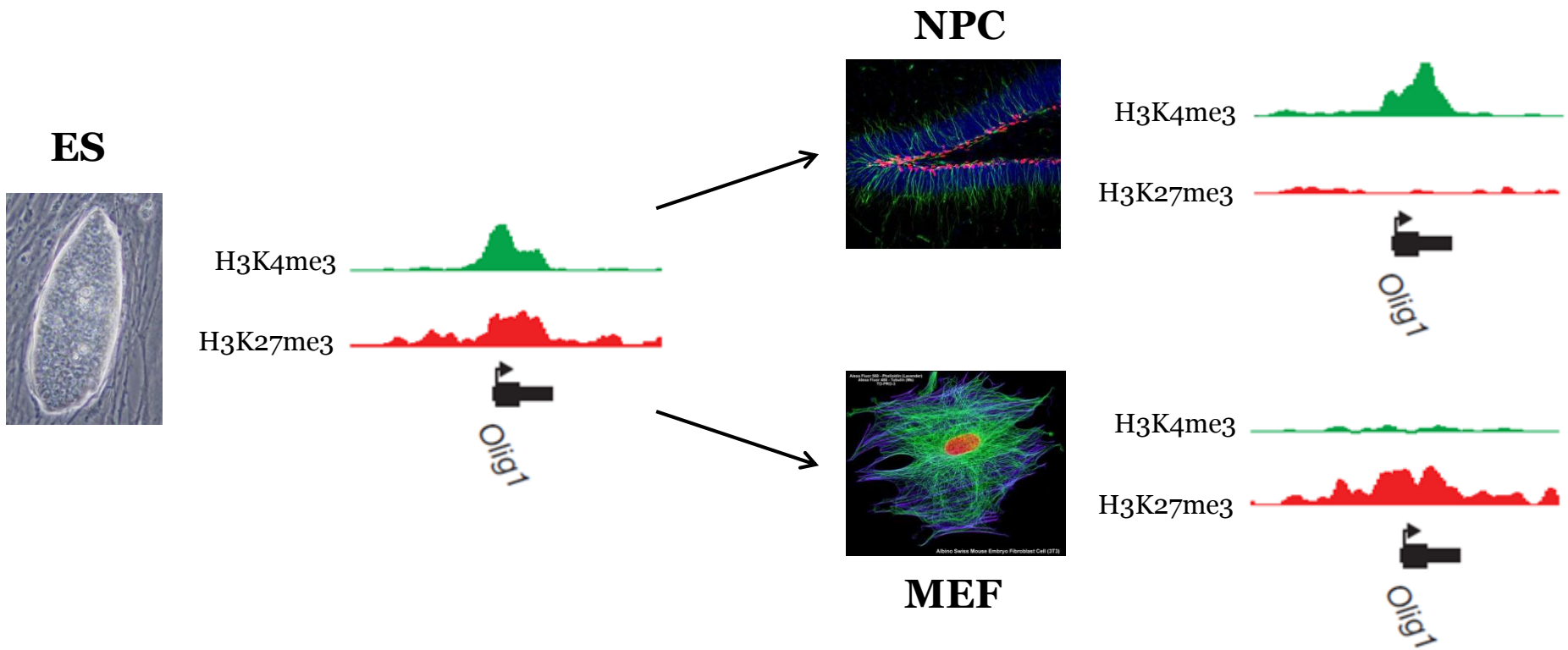
Gene repression by chromatin

- Gene silencing can be reversible and (relatively) irreversible
- irreversible = heterochromatin
- reversible can be removed in the process of cell differentiation



Reversible repression

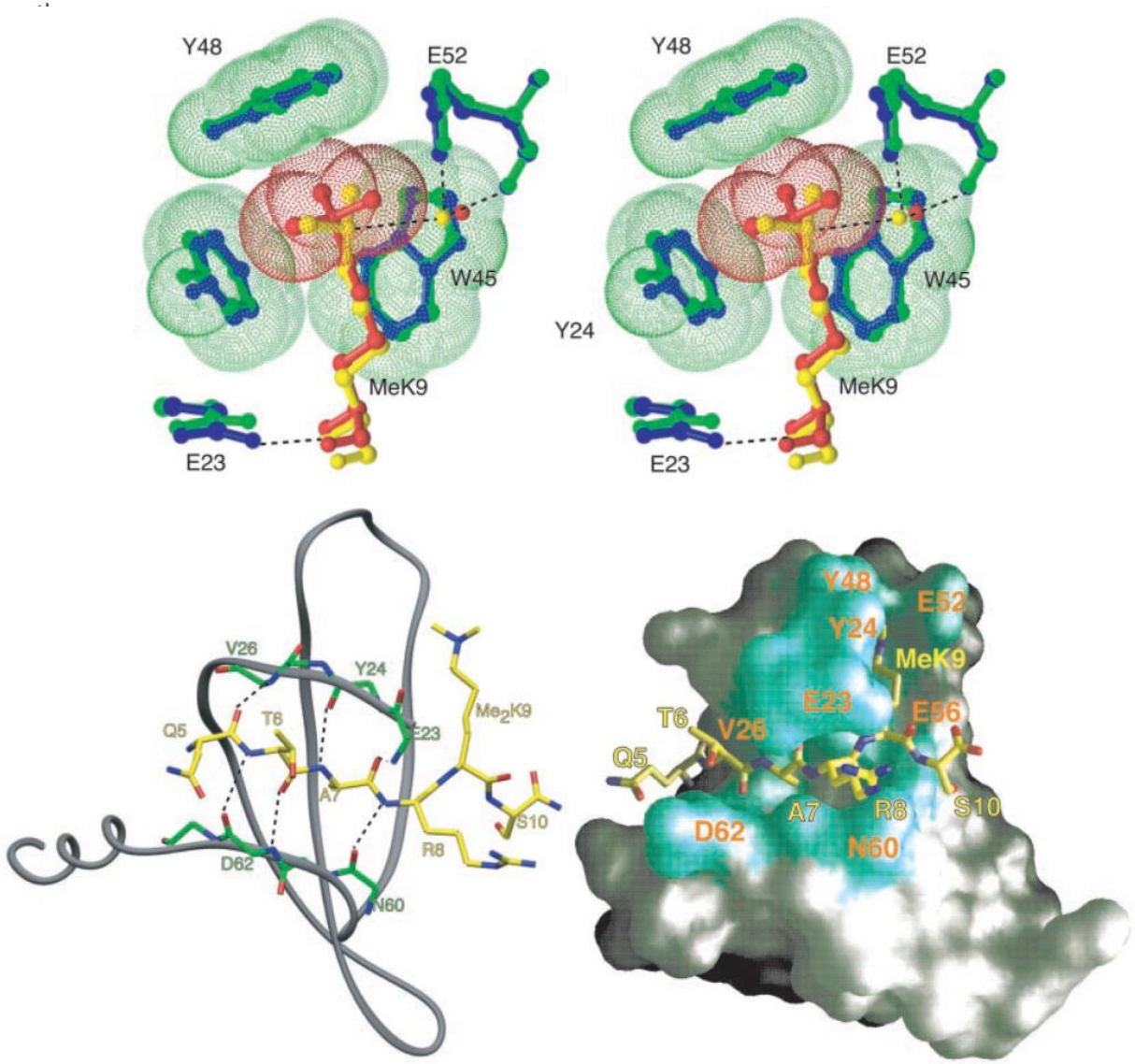
- H3K27me3 marks suppressed genes poised to be activated
- Stem cells can have both H3K4me3 and H3K27me3 – unique!



Mikkelsen, T.S.; Ku, M. et al, Nature 448, pp 553-560 (2007)
Vastenhouw N.L.; Zhang Y. et al, Nature 464, pp 922-6 (2010)

Heterochromatin marks

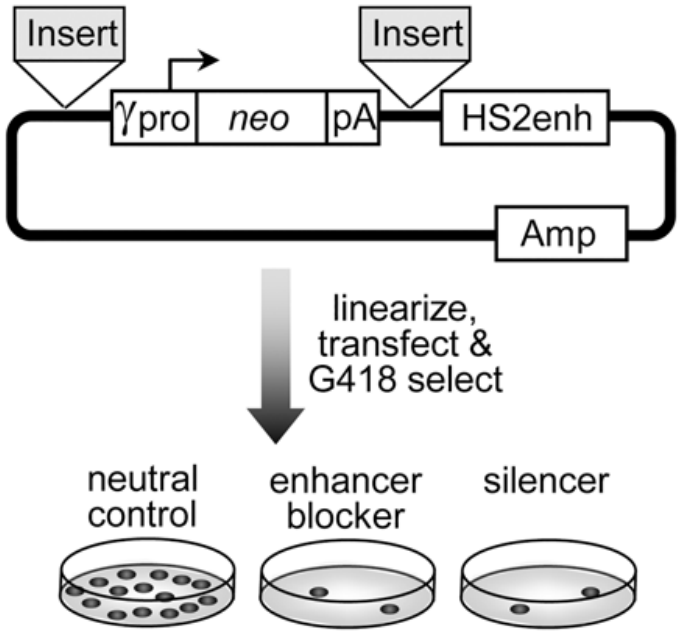
- Marks H3K9me2 and H3K9me3 are strongly associated with heterochromatin
- Binding with protein HP1

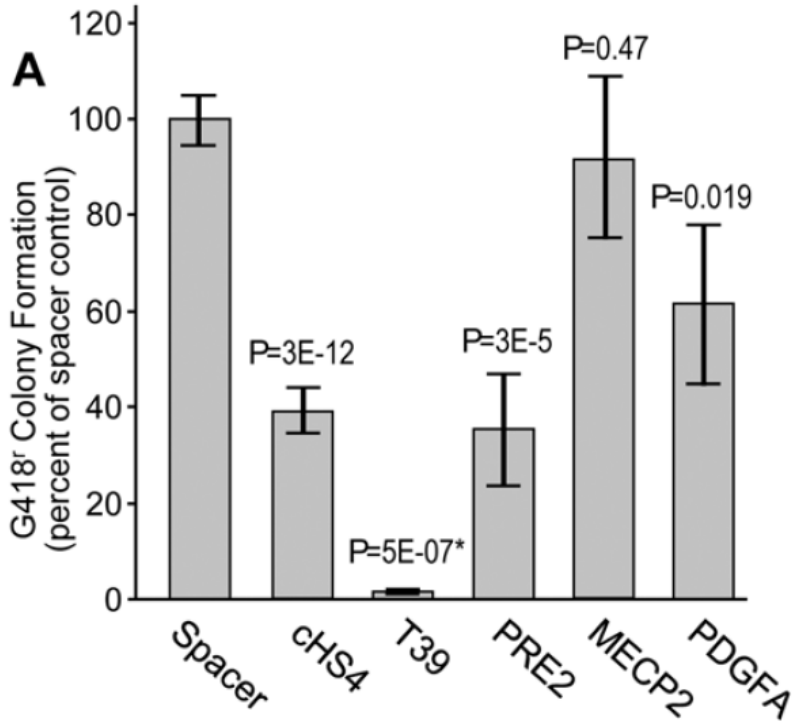


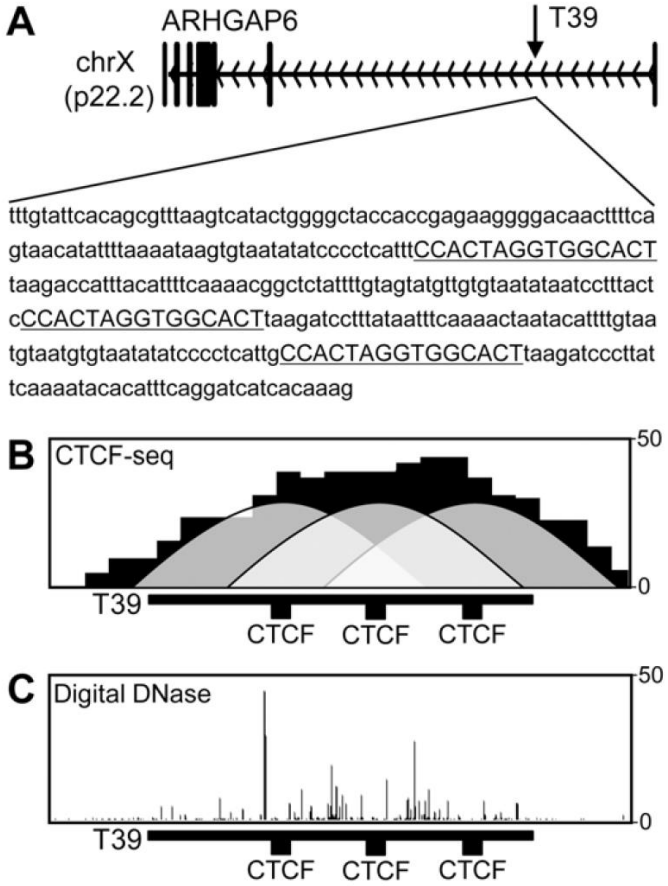
Jacobs, S.A.; Khorasanizadeh, S., Science, 295(5562), pp 2080-3 (2002)

Silencers in eukaryotes

- well-characterized silencers in eukaryotes are extremely rare
- most repression is thought to be done via chromatin



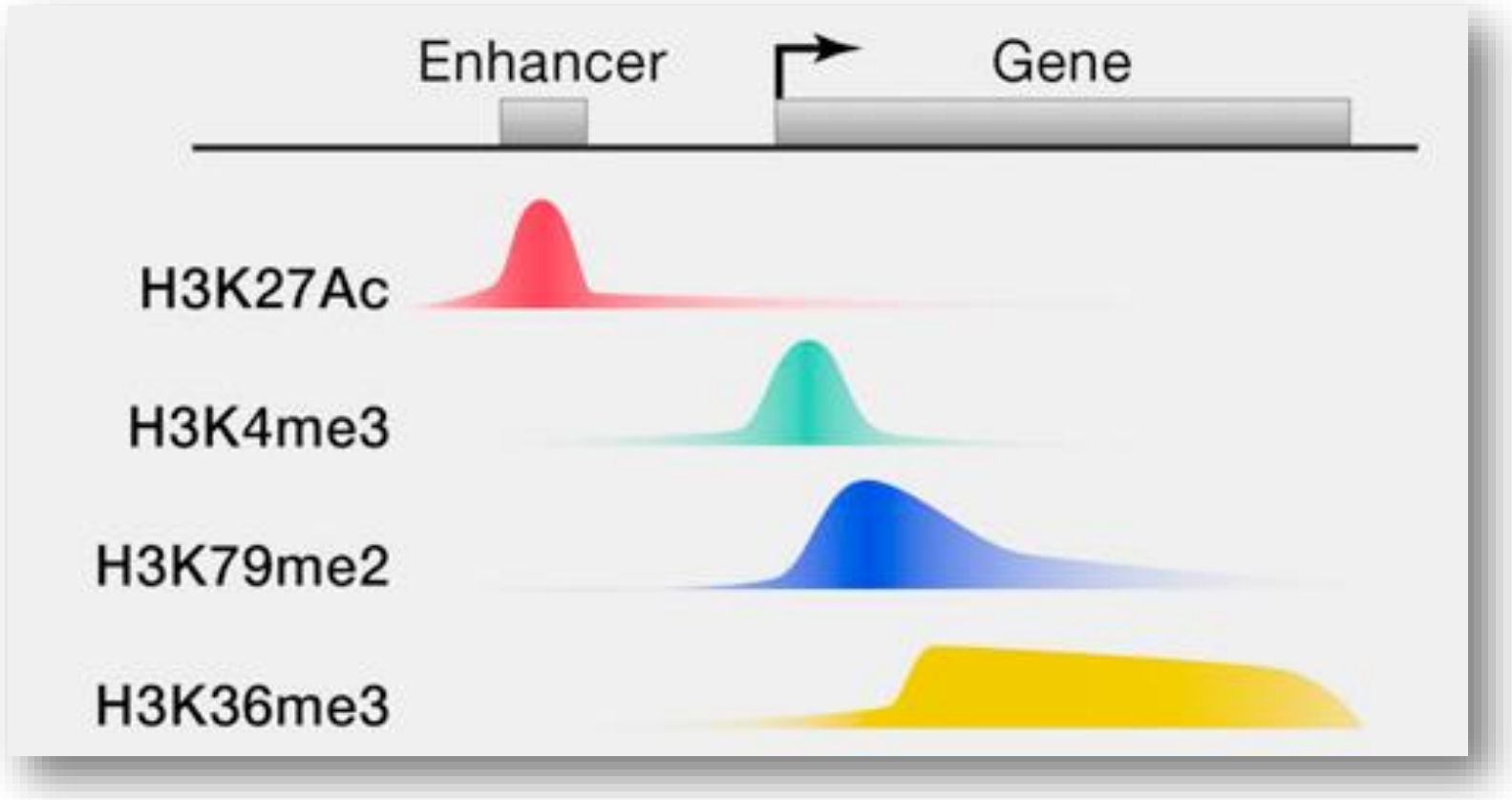




www.ncbi.nlm.nih.gov/pubmed/25910277

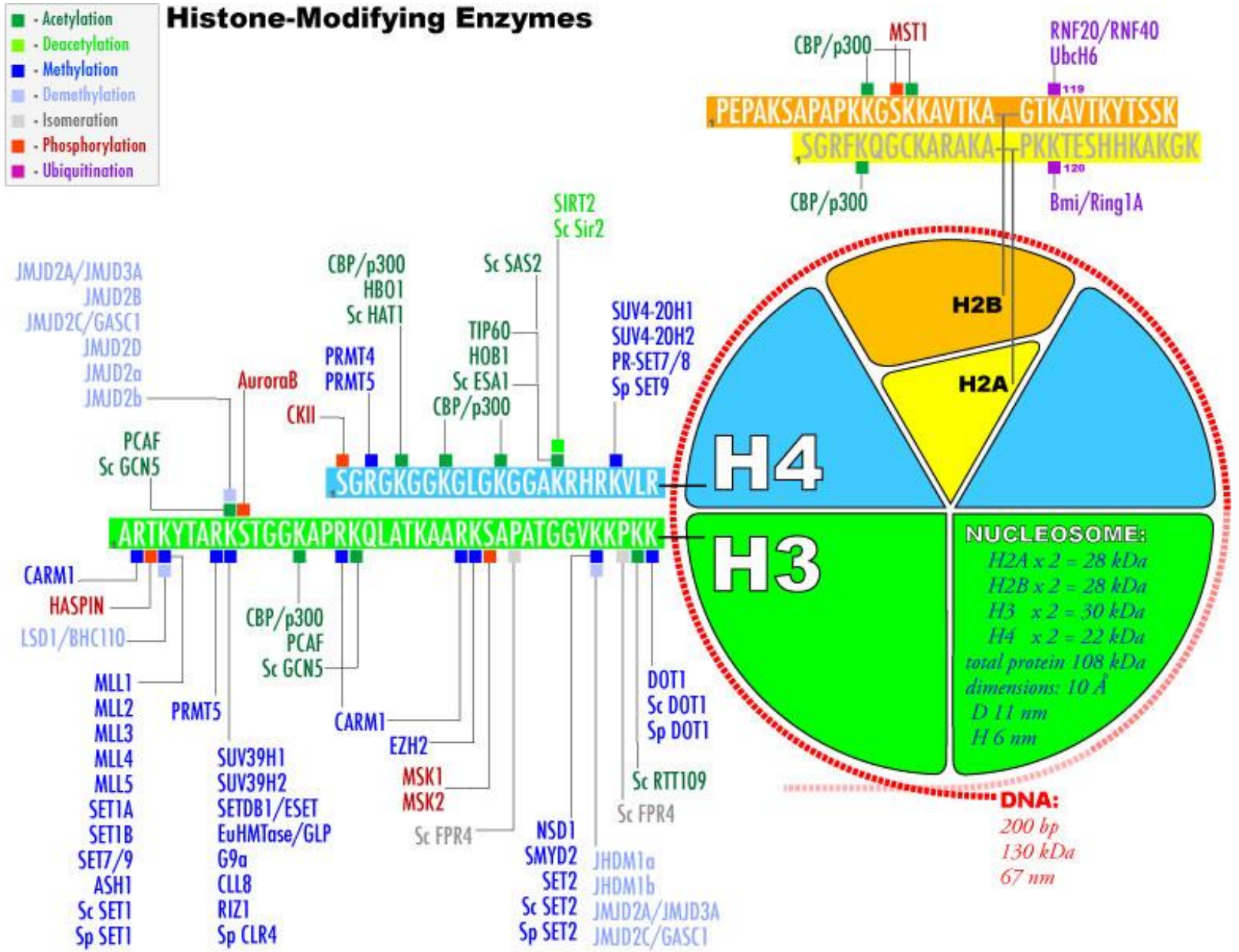
Easy, right?

- Histone modifications can be our guide in the genome!



Lee, T.I.; Young, R.A., Cell, 152(6), pp 1237-51 (2013)

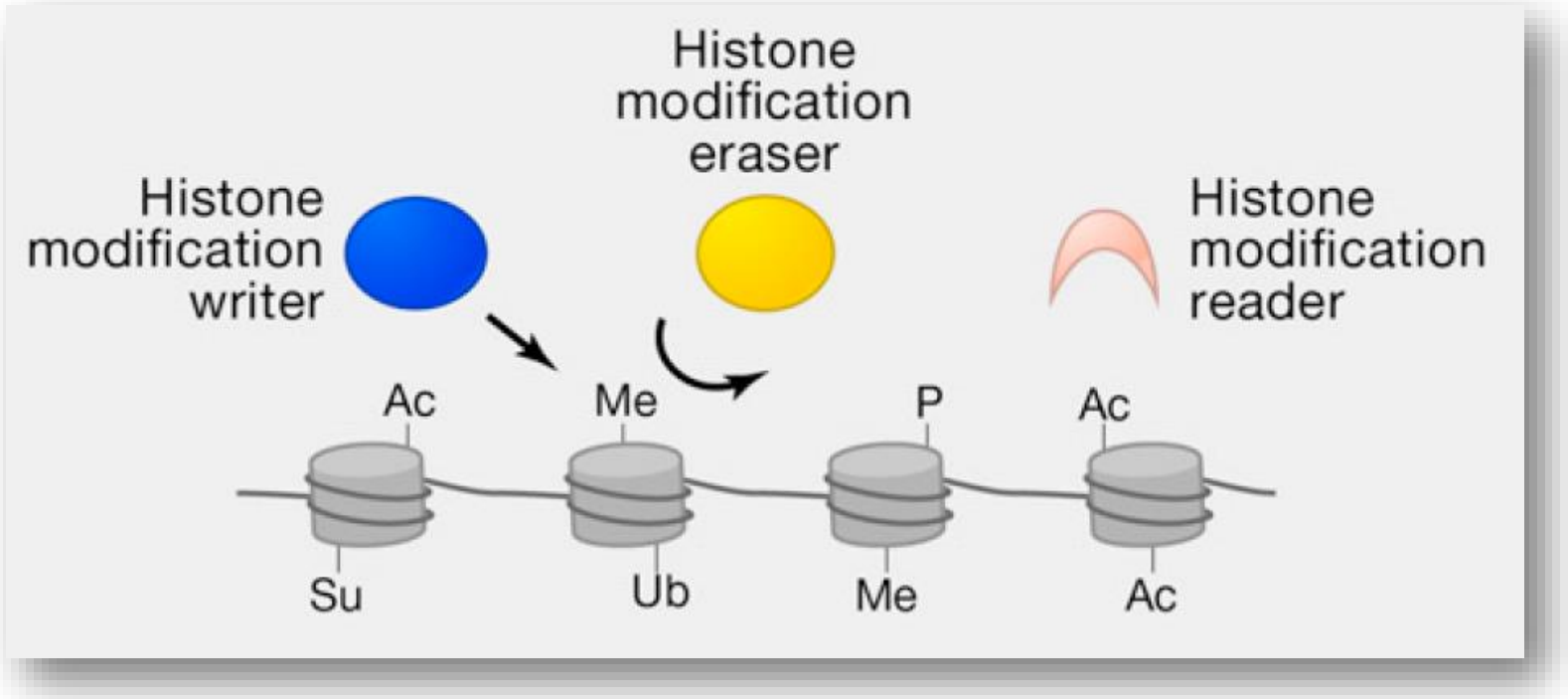
Wrong…



http://en.wikipedia.org/wiki/Histone-modifying_enzymes

Histone modification summary

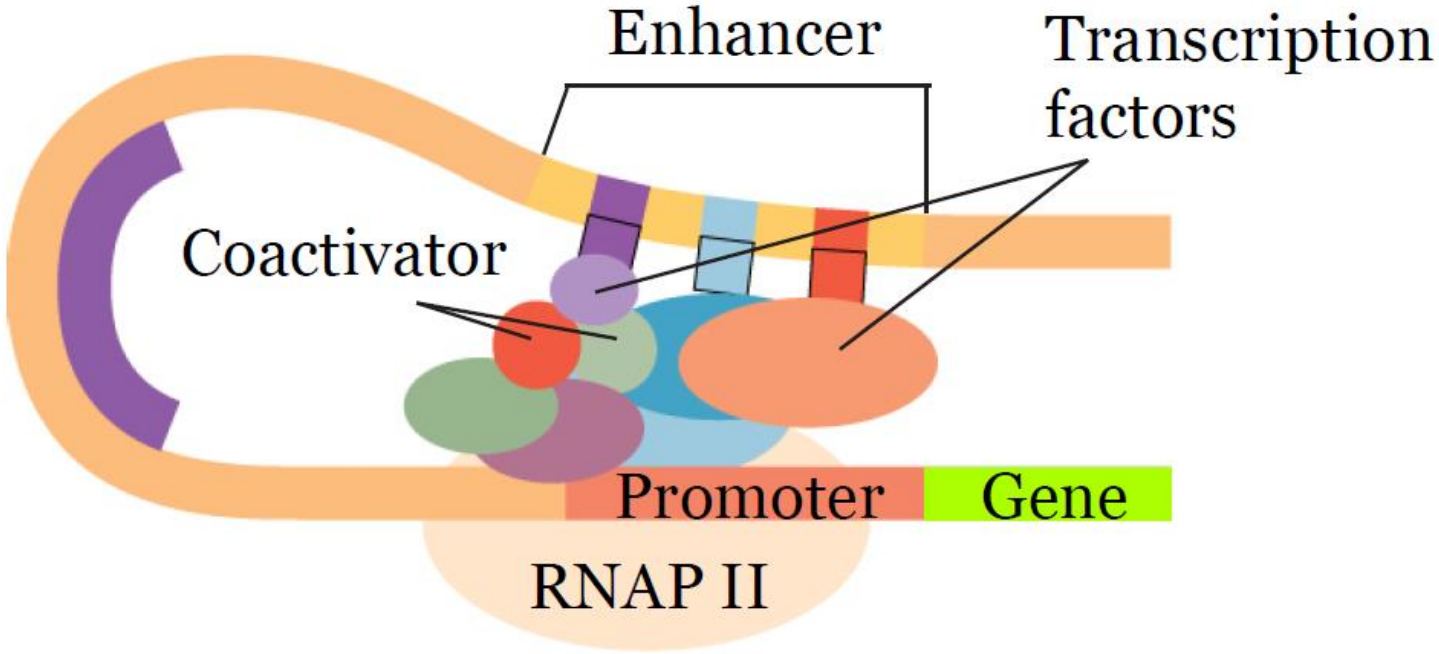
- there are protein domains that add, remove, and recognize various histone modifications
- histone additions are basically post-translational modifications
- main modifications are methylation and acetylation, main histone is H3



Lee, T.I.; Young, R.A., Cell, 152(6), pp 1237-51 (2013)

Extended model of transcription

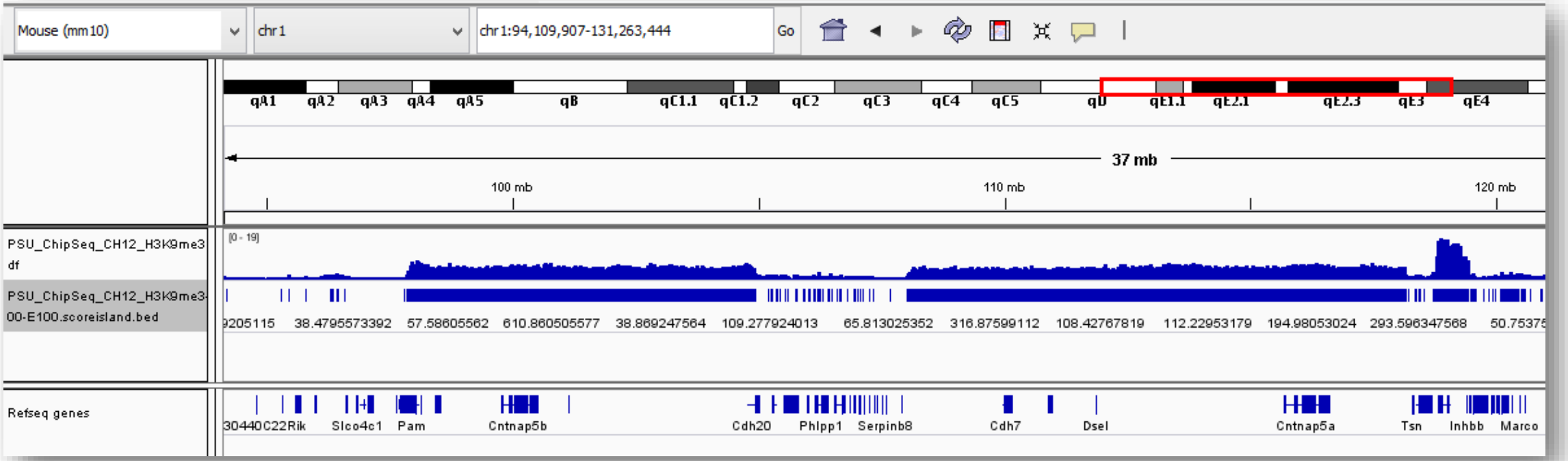
- transcriptional regulation is very complex and is influenced by (at least):
  - DNA methylation
  - histone modifications
  - transcription factors (10-20 per gene!)
- transcription factors recruit proteins that often also modify histones
- regulation can happen at transcriptional pre-initiation, initiation, or elongation



Lee, T.I.; Young, R.A., Cell, 152(6), pp 1237-51 (2013)

TFs vs. Histone Modifications in ChIP-Seq

- TFs and histone modifications are quite different:
  - TFs have narrow binding motifs (5-20 bps) and are very localized
  - histone marks are often limited only by insulators (e.g. CTCF) and show broad signal patterns
- broad signal marks require higher library complexity (20M reads vs 3-10M for TFs)

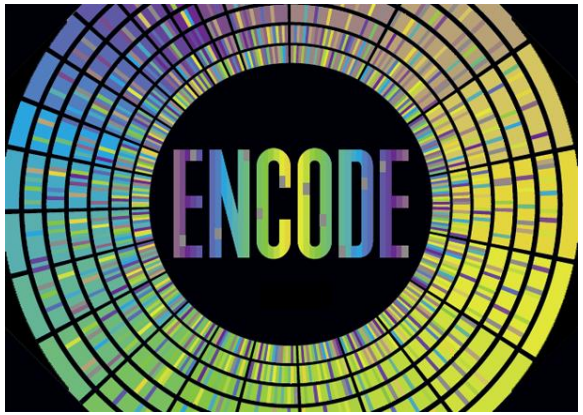


ENCODE project

- ENCODE = ENCyclopedia Of DNA Elements
- pilot cost (2007): $55M, up to date: ~$300M
- do RNA-Seq, ChIP-seq of major TFs and histone modifications, DNA methylation
- series of publications in the Fall of 2012 (6 Nature papers, 30 papers overall)

National Human Genome Research Institute

ENCODE



ENCODE By the Numbers

147 cell types studied
80% functional portion of human genome
20,687 protein-coding genes
18,400 RNA genes
1640 data sets
30 papers published this week
442 researchers
$288 million funding for pilot, technology, model organism, and current project

http://www.sciencemag.org/content/337/6099/1159/F2.expansion.html

ENCODE ChIP-seq Experiment Matrix *hg19*

search for: tracks files

Antibody Targets

Cell Types	Histone Modification	H2AFZ	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me2	H3K4me3	H3K79me2	H3K9ac	H3K9me1	H3K9me3	H4K20me1	Transcription Factor	ARID3A	ATF1	ATF2	ATF3	BACH1	BATF	BCL11A	BCL3	BCLAF1
Tier 1																							
GM12878		1	1	2	2	1	1	2	1	1		1	1				1	1		1	1	1	1
H1-hESC		1	1	1	1	1	1	1	1	1		1	1				1	1	1		2		
K562		1	1	3	2	2	1	8	1	2	1	1	1		1	1		2	1			1	1
Tier 2																							
A549		2	2	2	2	2	2	3	2	1		1	1					1				1	
CD20+		1					1																
CD20+_RO01778								1															
CD20+_RO01794			1					1					1										
H1-neurons																							
HeLa-S3		1	1	2	2	1	1	2	1	1		1	1										
HepG2		1	1	2	2	1	1	2	1	1		1	1		1			1					
HUVEC		1	1	2	2	1	1	2	1	1	1	1	1										

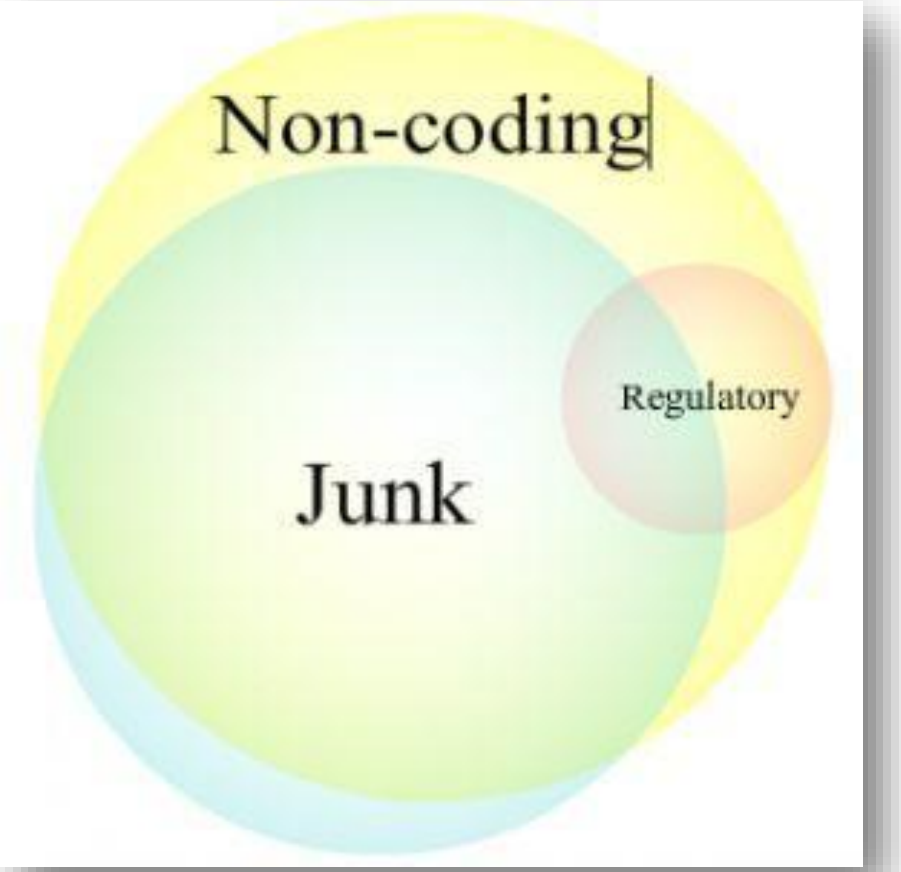
ENCODE project discoveries

- 400,000 enhancers and 70,000 promoters
- more than 90% of genomic variation are in noncoding areas
- DNase I footprint is not that big
- mRNAs are more abundant in cytosol, other RNAs – in the nucleus
- "*More than 80% of human genome is functionally active*"

http://www.evolutionnews.org/2012/09/the_demise_of_j_1064061.html

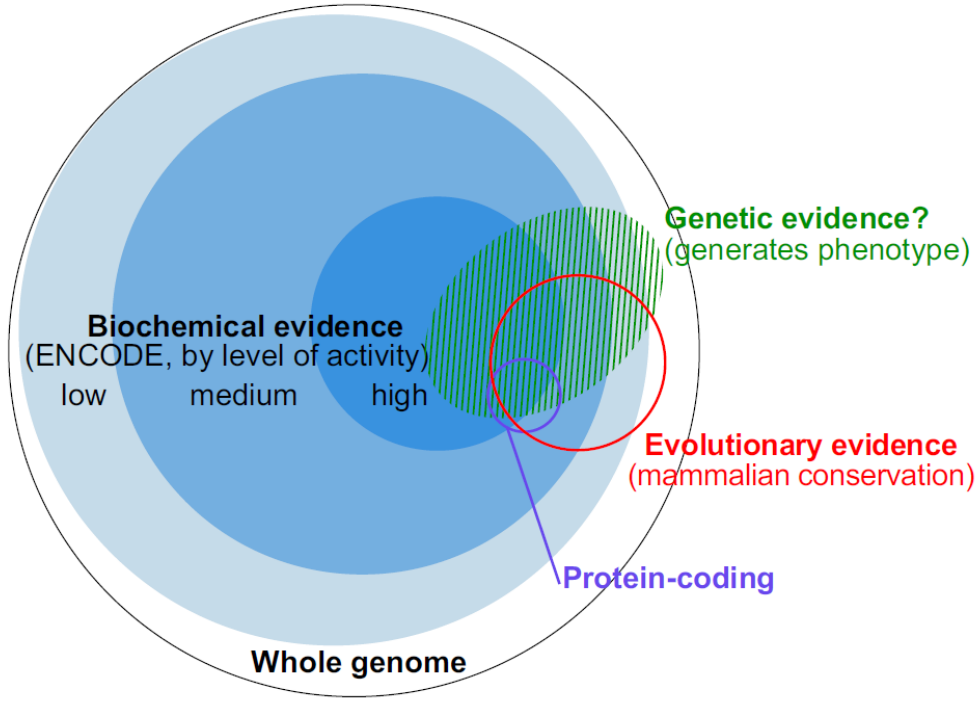
ENCODE analysis criticism

- 80% of DNA cannot be truly functional, since only about 10% (5-15%) is conserved
- This means ~70% of genome is either
  - impervious to deleterious mutations, or
  - does not mutate, or
  - does not have deleterious mutations



http://blogs.scientificamerican.com/guest-blog/2012/09/17/junk-dna-junky-pr/

ENCODE criticism



PERSPECTIVE

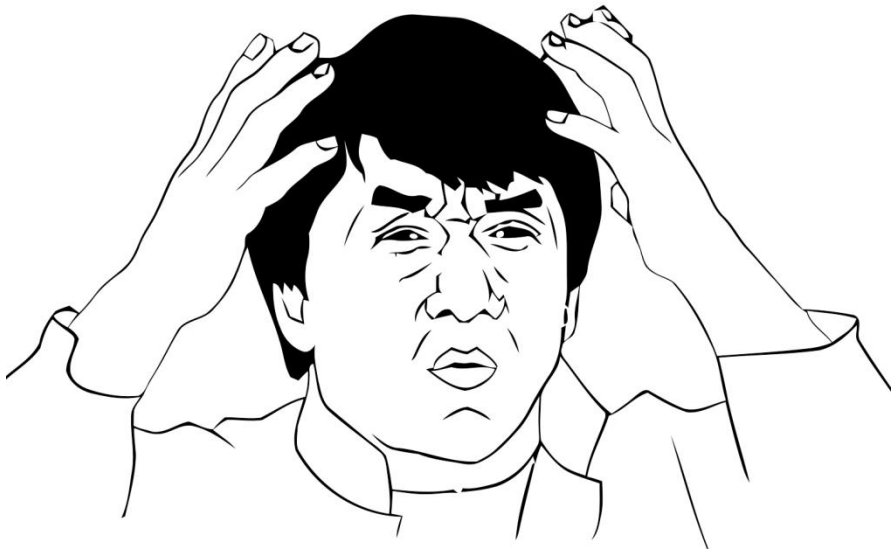
Defining functional DNA elements in human genome

Manolis Kellis[a,b,1,2], Barbara Wold[c,2], Michael P. Snyder[d,2], Bradley E. Bernstein[b,e,f,2], Anshul Kundaje[a,b,3], Georgi K. Marinov[c,3], Lucas D. Ward[a,b,3], Ewan Birney[g], Gregory E. Crawford[h], Job Dekker[i], Ian Dunham[g], Laura L. Elnitski[j], Peggy J. Farnham[k], Elise A. Feingold[j], Mark Gerstein[l], Morgan C. Giddings[m], David M. Gilbert[n], Thomas R. Gingeras[o], Eric D. Green[j], Roderic Guigo[p], Tim Hubbard[q], Jim Kent[r], Jason D. Lieb[s], Richard M. Myers[t], Michael J. Pazin[j], Bing Ren[u], John A. Stamatoyannopoulos[v], Zhiping Weng[i], Kevin P. White[w], and Ross C. Hardison[x,1,2]

http://blogs.scientificamerican.com/guest-blog/2012/09/17/junk-dna-junky-pr/

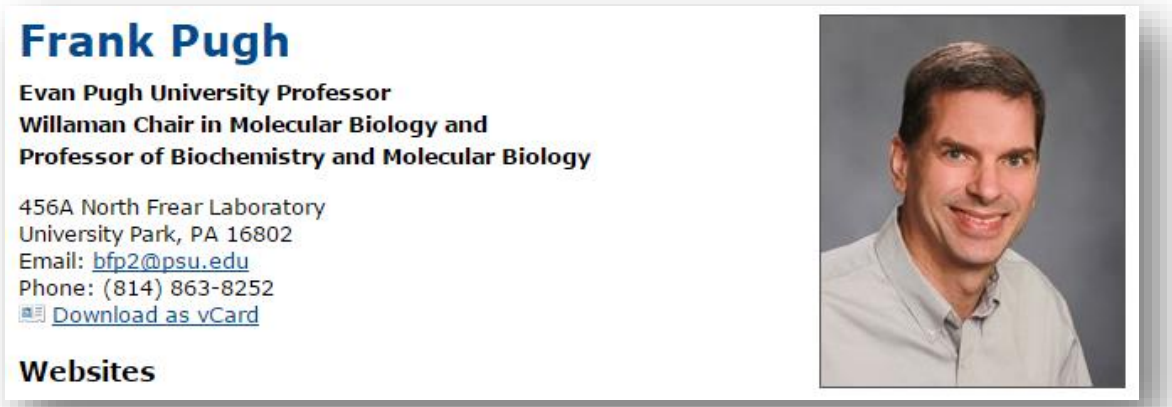
Flavors of DNA sequencing

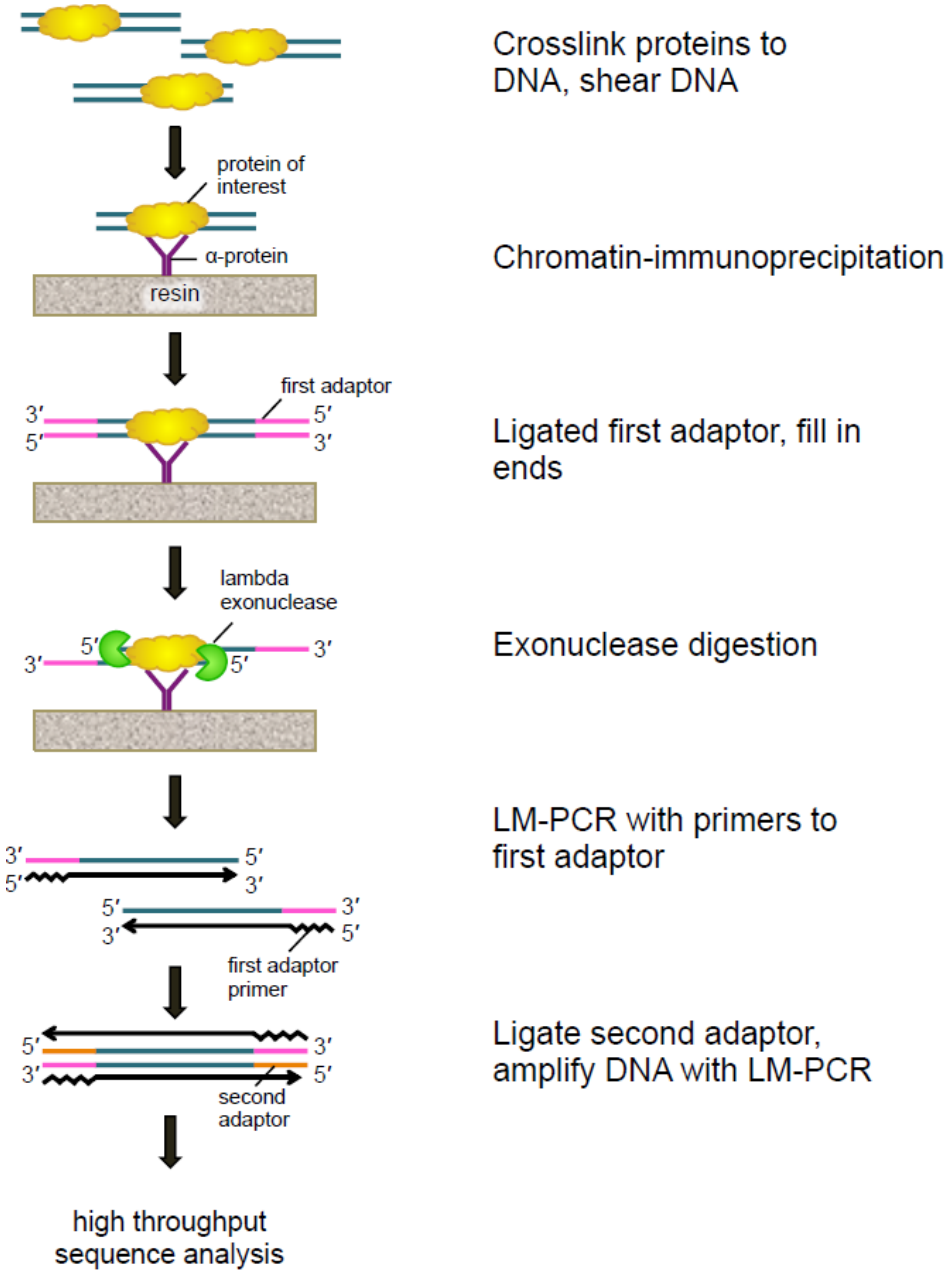
- Sono-Seq, identical to ChIP-Seq but skipping the IP step to see open regions
- HITS-CLIP, PAR-CLIP, etc – for RNA-interacting proteins
- ChiRP-Seq, to measure RNA-bound DNA and proteins
- ChIP-exo uses exonuclease treatment to achieve up to single base-pair resolution
- Hundreds more!



Example: ChIP-Exo

- ChIP-Exo allows to greatly increase precision of TF positioning on the DNA
- It uses endonuclease and ligation-mediated PCR
- Drawbacks – requires a lot of cells, generates libraries of poor complexity

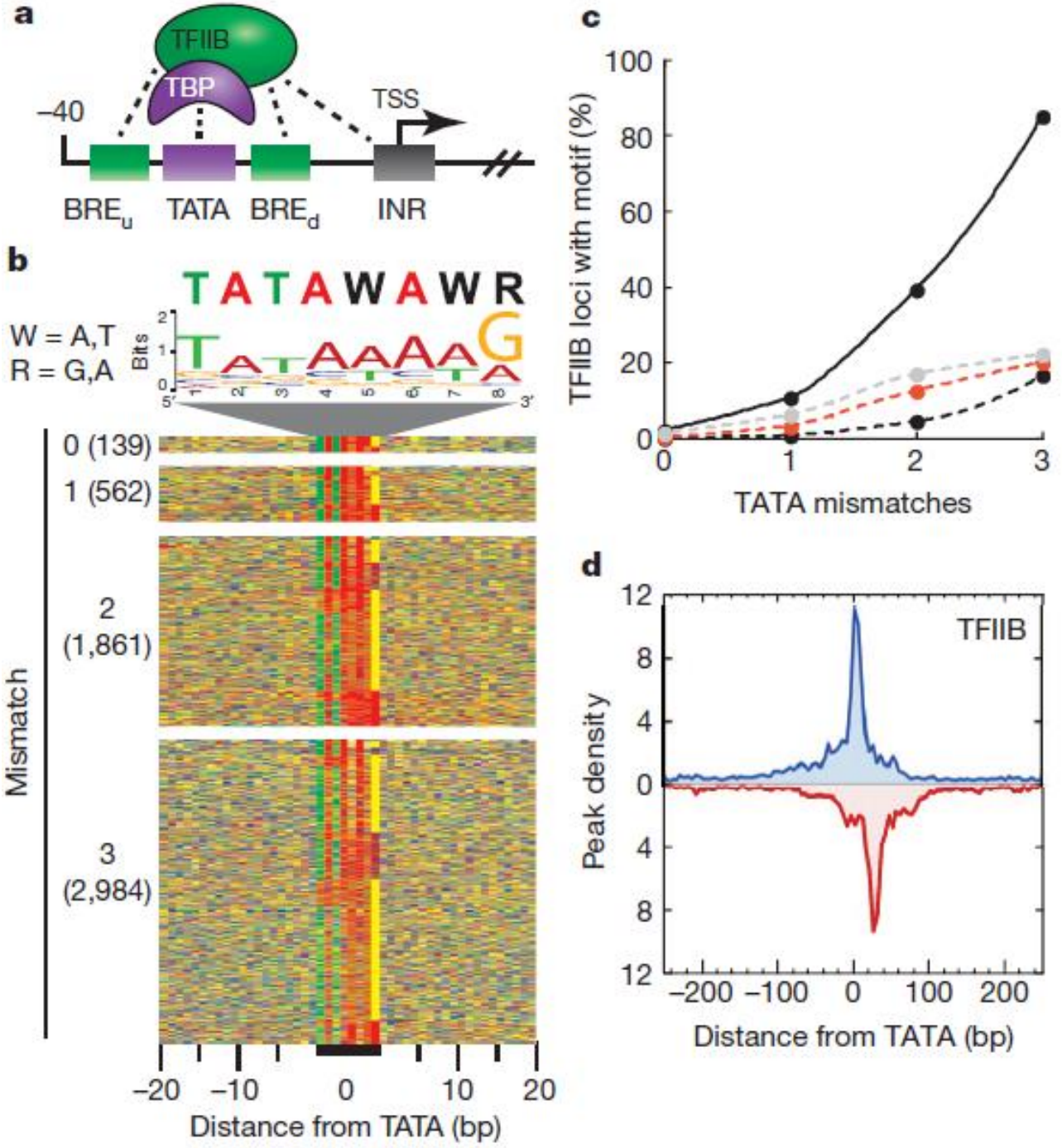




Venters, B.J.; Pugh, B.F., Nature., 502, pp 53-8 (2013)

ChIP-Exo

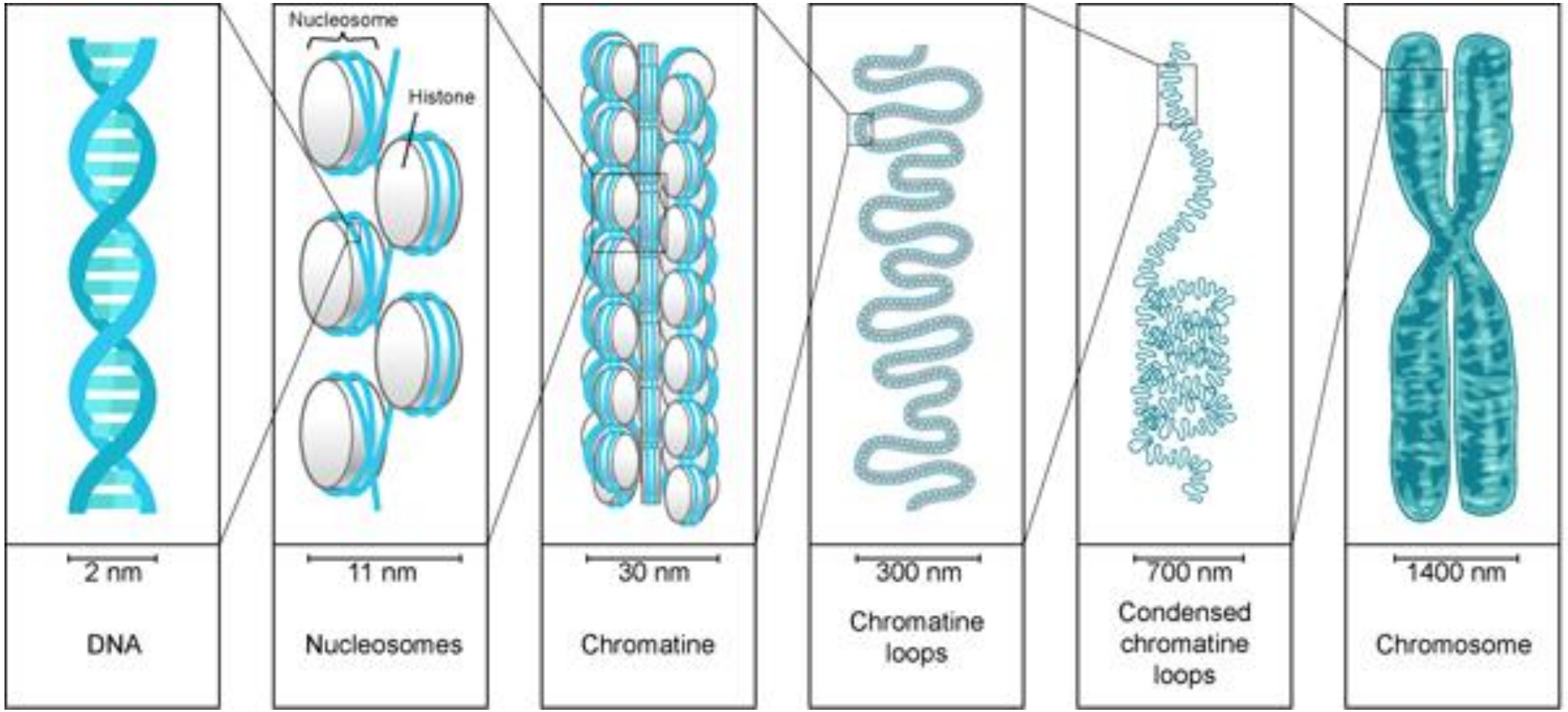
- virtually all gene promoters have the same general structure, including (often degenerate) TATA box!



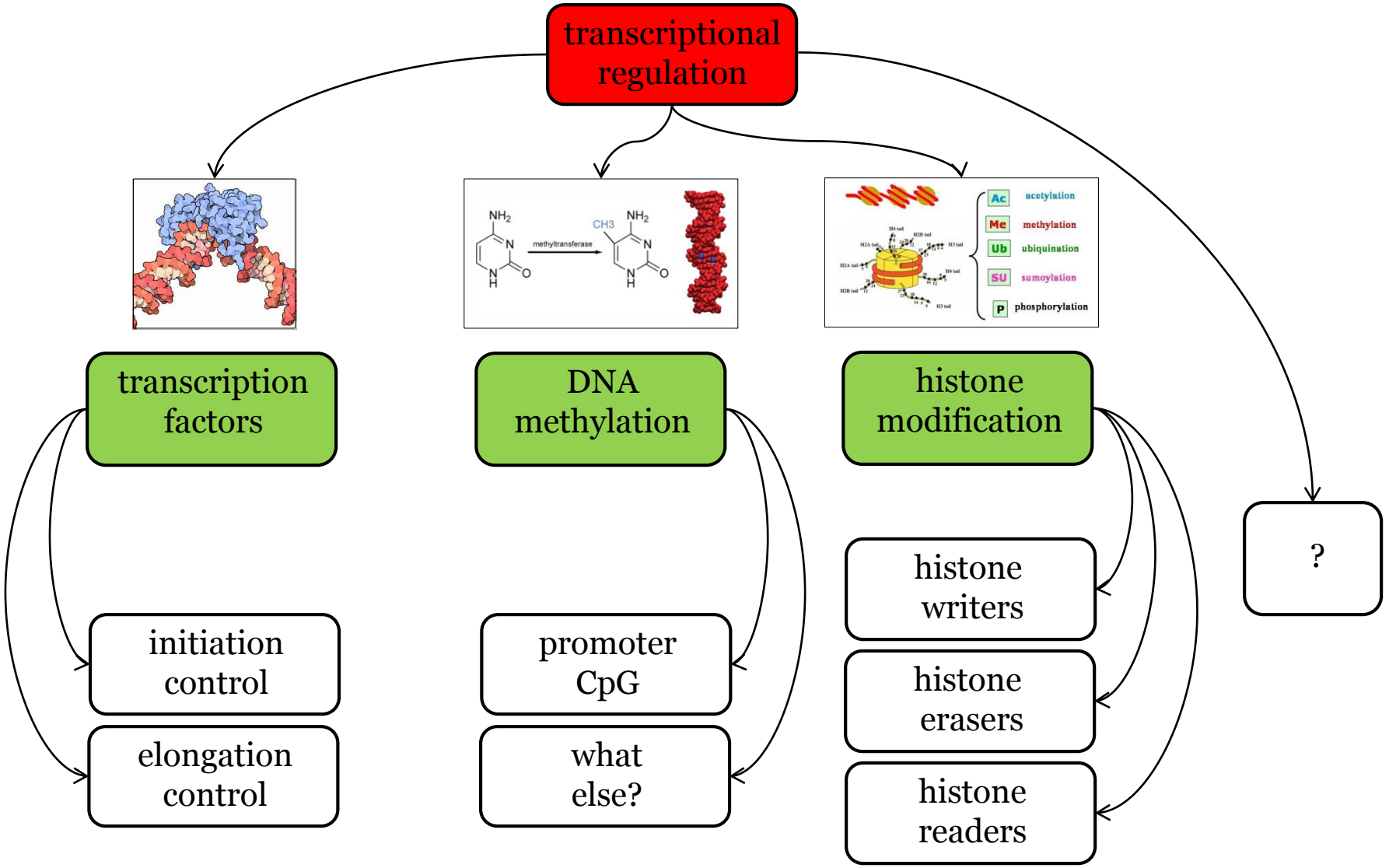
Venters, B.J.; Pugh, B.F., Nature., 502, pp 53-8 (2013)

Summary

- DNA packaging in eukaryotes is complicated and participates in regulation



Summary



http://www.pvri.info/content/epigenetic-drivers-and-genetic-passengers-road-cancer

Tired yet?