CS273A

Lecture 7: Gene Regulation II

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

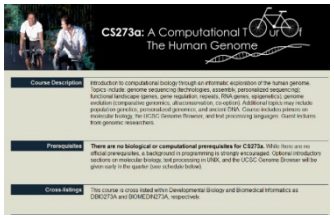
Profs: Serafim Batzoglou & Gill Bejerano

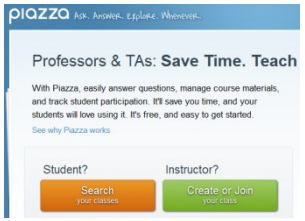
CAs: Karthik Jagadeesh & Johannes Birgmeier

* Mostly: track on website/piazza

Announcements

- http://cs273a.stanford.edu/
  - Lecture slides, problem sets, etc.
- Course communications via Piazza
  - Auditors please sign up too

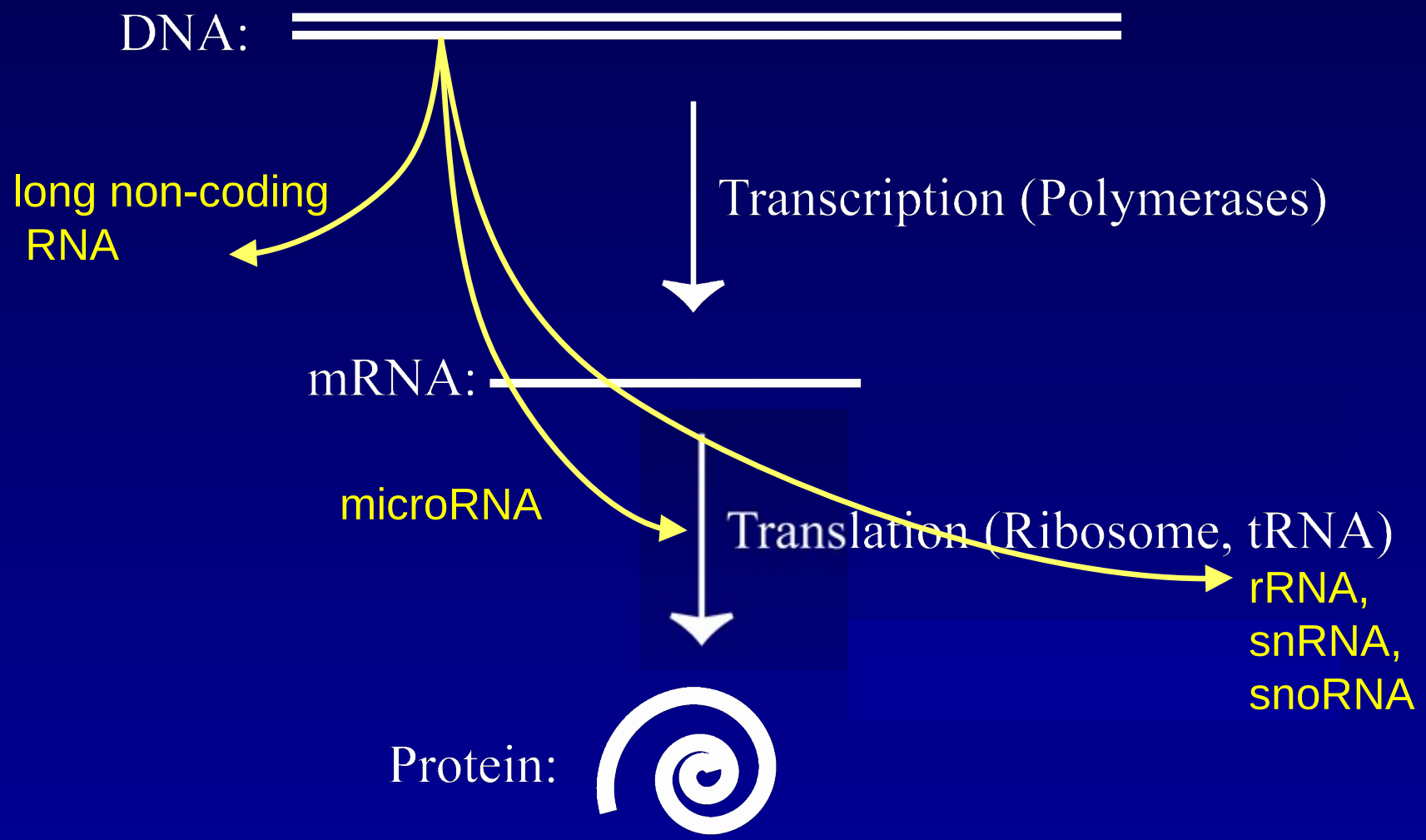




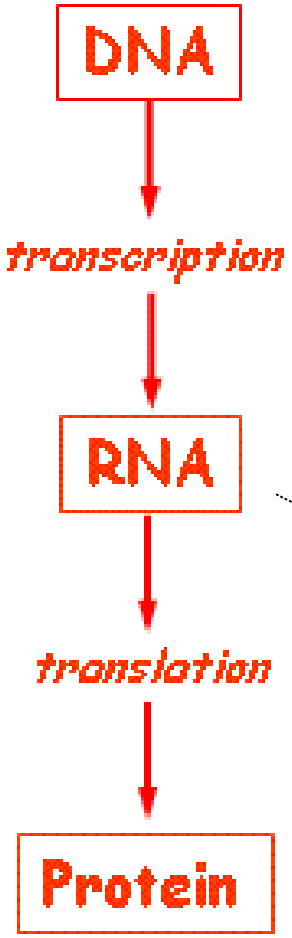
- PS1 due this Friday.


Genome Content

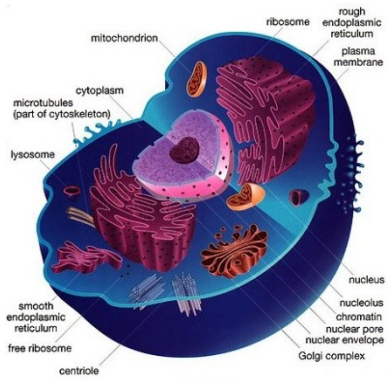
Genes = coding + “non-coding”



Coding and non-coding gene production



To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.



The cell is constantly making new proteins and ncRNAs.

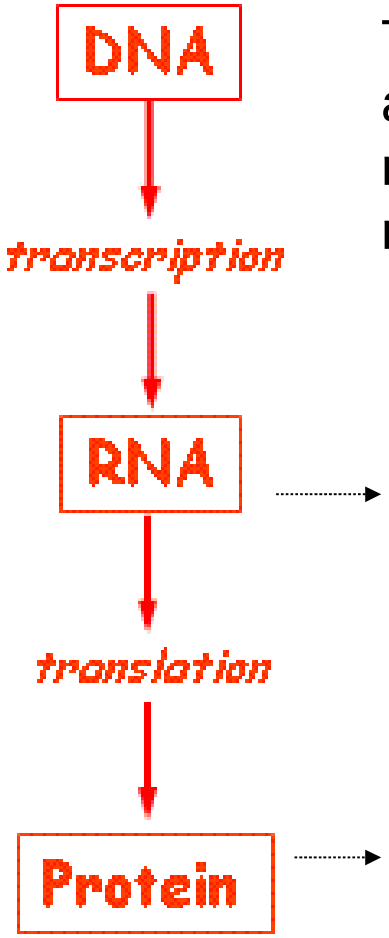
These perform their function for a while,

And are then degraded.

Newly made coding and non coding gene products take their place.

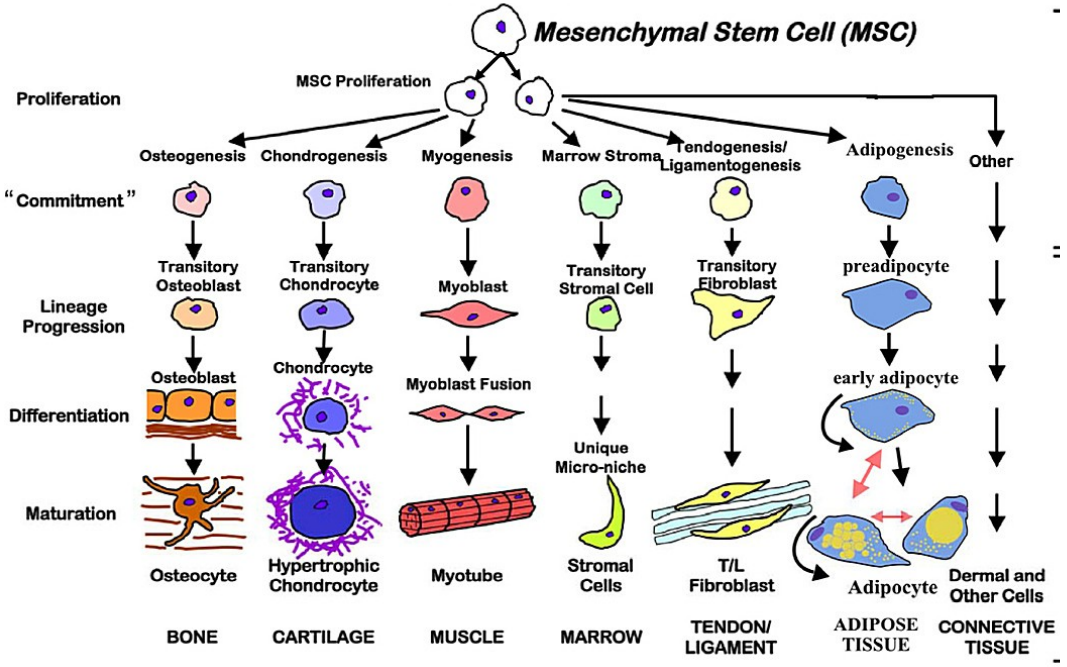
The picture within a cell is constantly “refreshing”.

Cell differentiation

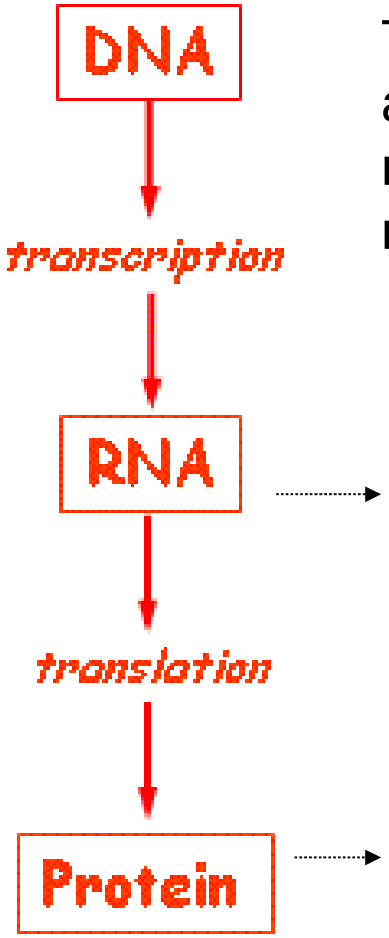


To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.

That is exactly what happens when cells differentiate during development from stem cells to their different final fates.

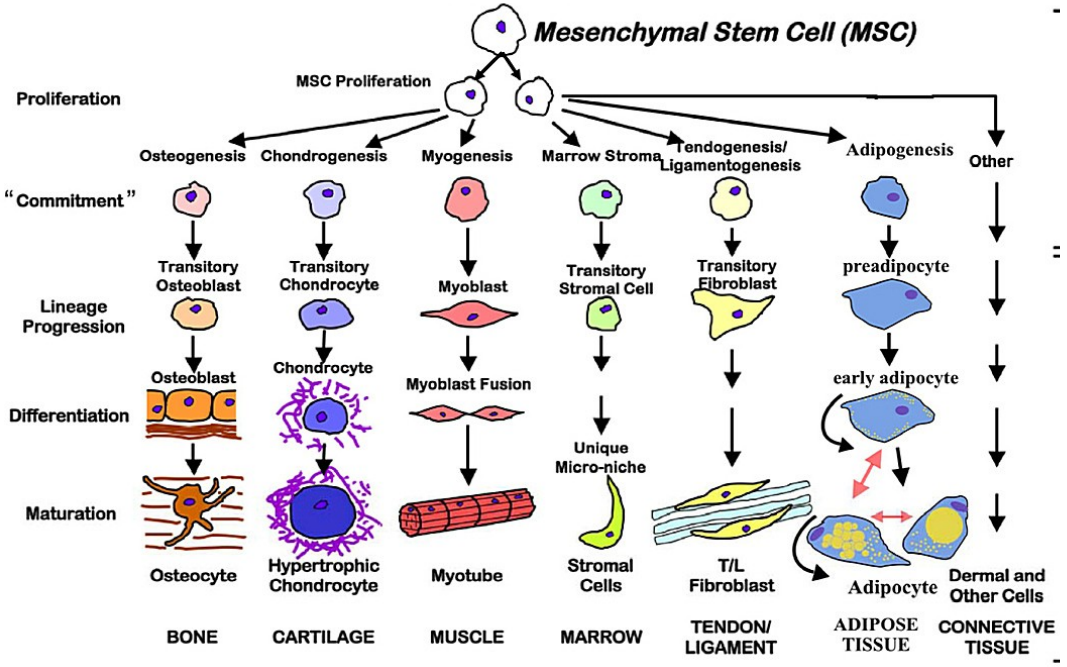


Cell differentiation



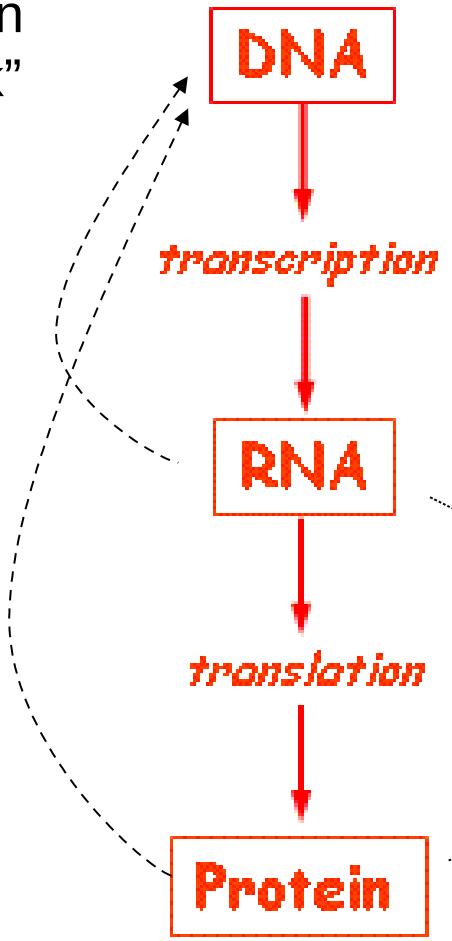
To change its behavior a cell can change the repertoire of genes and ncRNAs it makes.

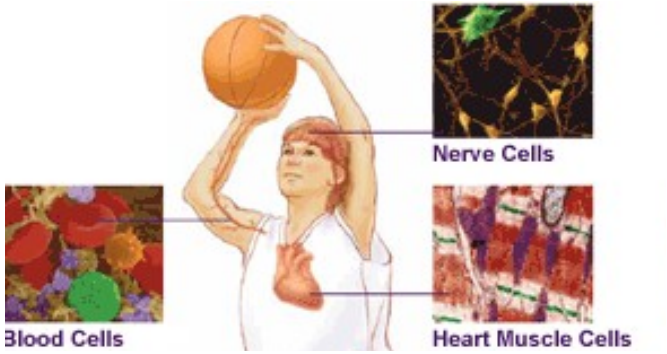
But how?

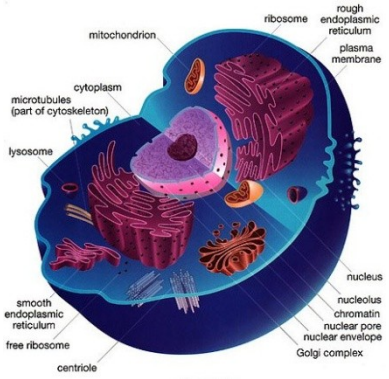


Closing the loop

Some proteins and non coding RNAs go “back” to bind DNA near genes, turning these genes on and off.

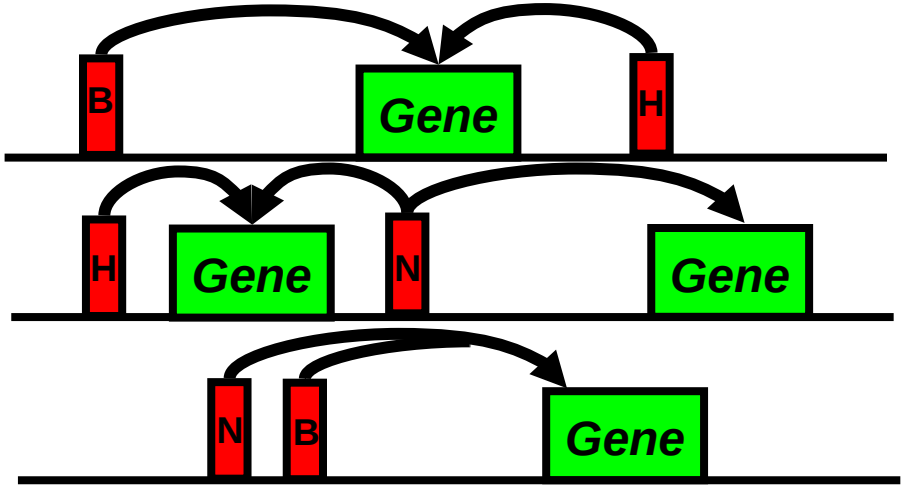


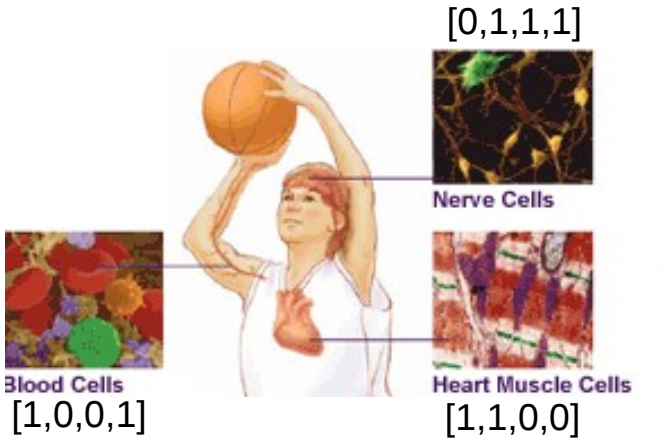




Genes & Gene Regulation

- Gene = genomic substring that encodes HOW to make a protein (or ncRNA).
- Genomic switch = genomic substring that encodes WHEN, WHERE & HOW MUCH of a protein to make.





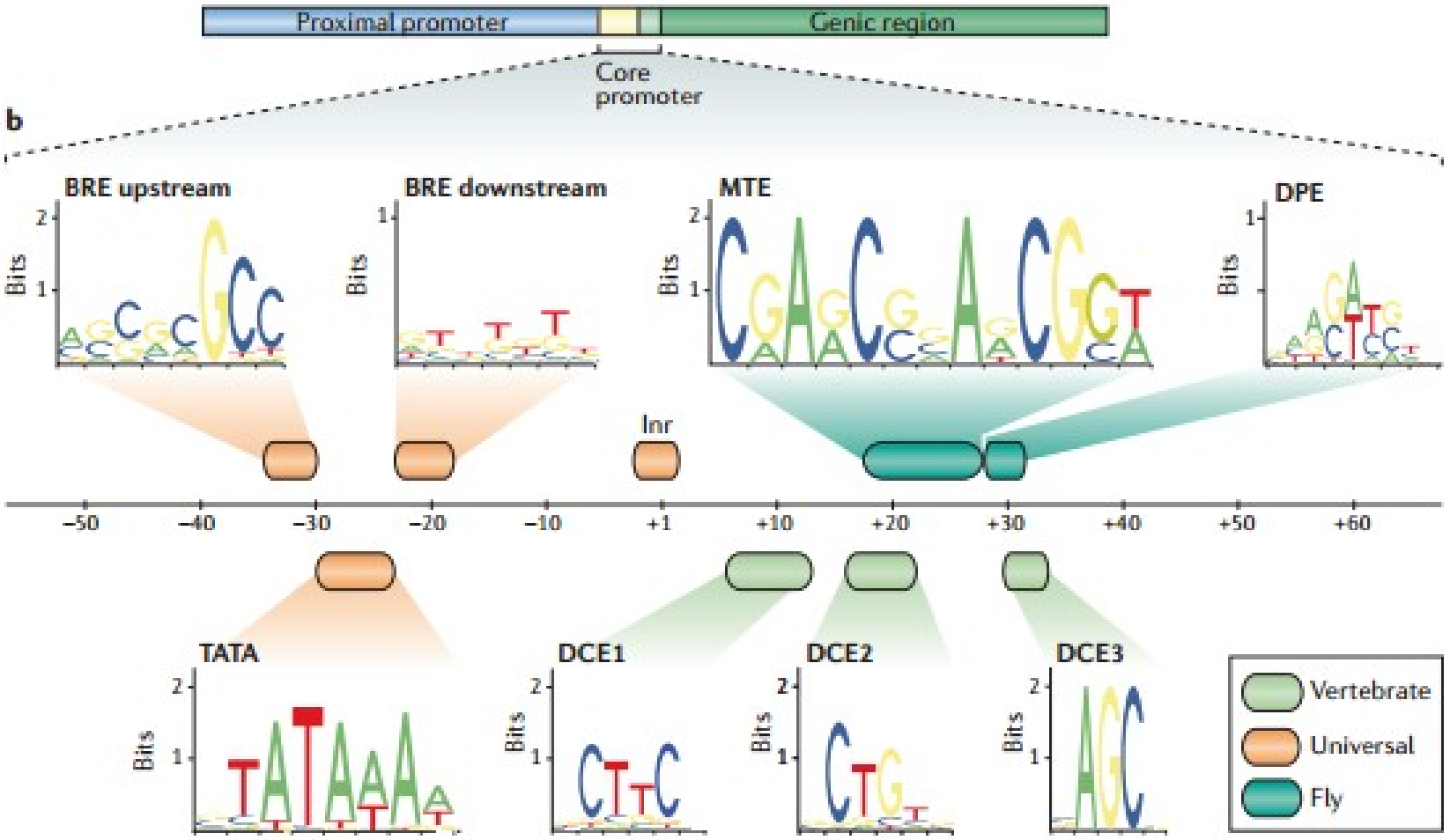
Transcription Regulation

Conceptually simple:

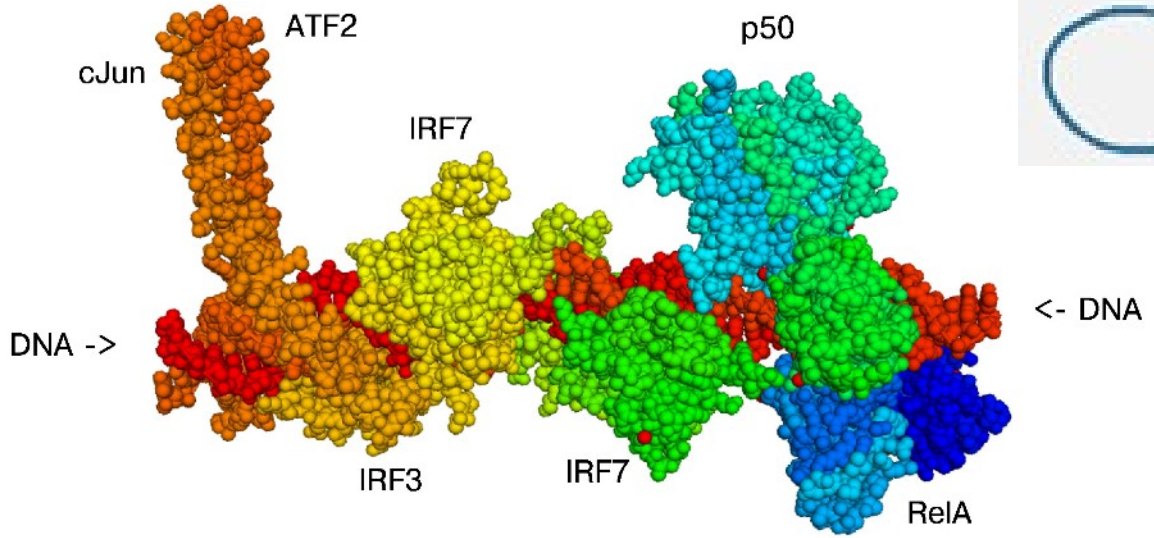
1. The machine that transcribes (“RNA polymerase”)
2. All kinds of proteins and ncRNAs that bind to DNA and to each other to attract or repel the RNA polymerase (“transcription associated factors”).
3. DNA accessibility – making DNA stretches in/accessible to the RNA polymerase and/or transcription associated factors by un/wrapping them around nucleosomes.

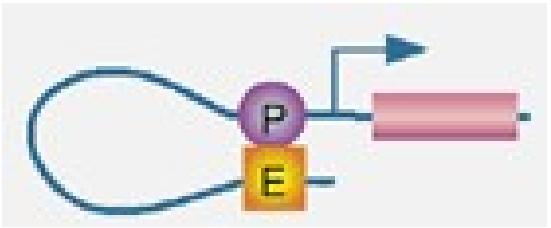
(Distinguish DNA patterns from proteins they interact with)

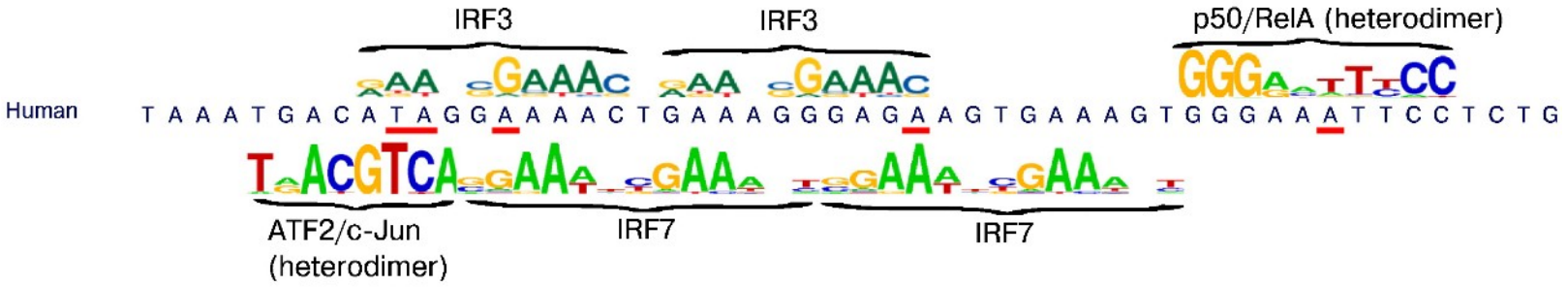
Promoters


Proximal promoter
Genic region
Core promoter
b
BRE upstream
BRE downstream
MTE
DPE
Inr
Bits
−50
−40
−30
−20
−10
+1
+10
+20
+30
+40
+50
+60
TATA
DCE1
DCE2
DCE3
Vertebrate
Universal
Fly

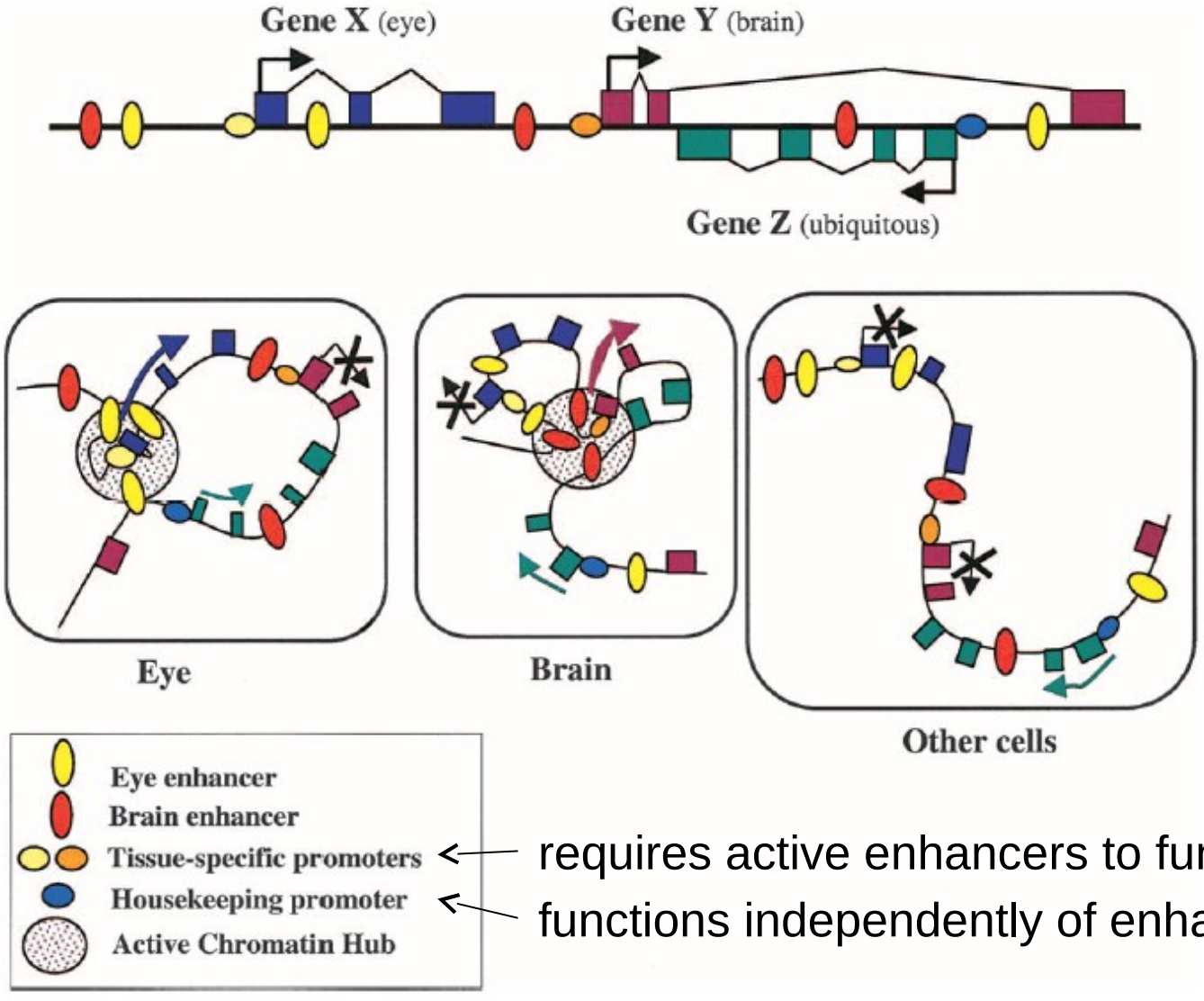
Enhancers


ATF2
cJun
IRF7
p50
DNA ->
<- DNA
IRF3
IRF7
RelA


P
E


IRF3
IRF3
p50/RelA (heterodimer)
Human
T A A A T G A C A T A G G A A A A C T G A A A G G G A G A A G T G A A A G T G G G A A A T T C C T C T G
ATF2/c-Jun
(heterodimer)
IRF7
IRF7

One nice hypothetical example

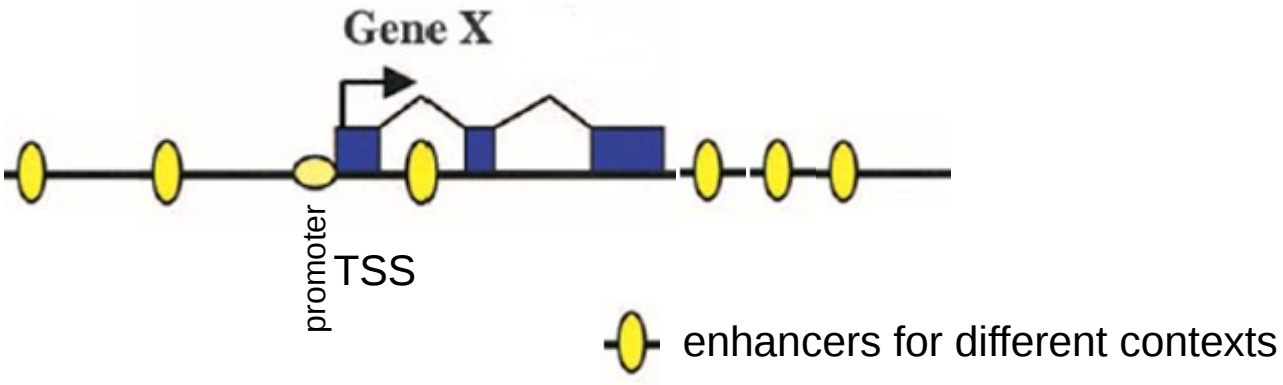


requires active enhancers to function

functions independently of enhancers

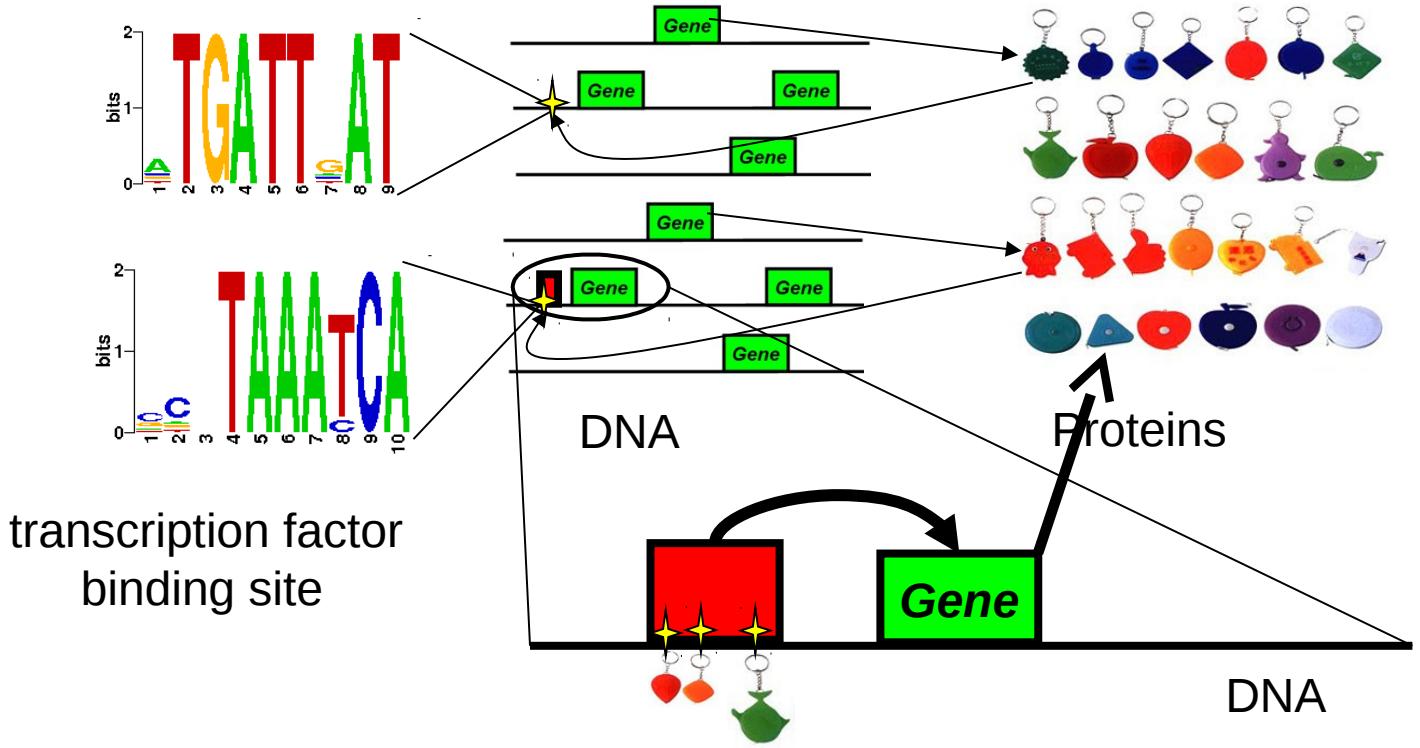
Terminology

- Gene regulatory domain: the full repertoire of enhancers that affect the expression of a (protein coding or non-coding) gene, at some cells under some condition.
  - Gene regulatory domains do not have to be contiguous in genome sequence.
  - Neither are they disjoint: One or more enhancers may well affect the expression of multiple genes (at the same or different times).



Imagine a giant state machine

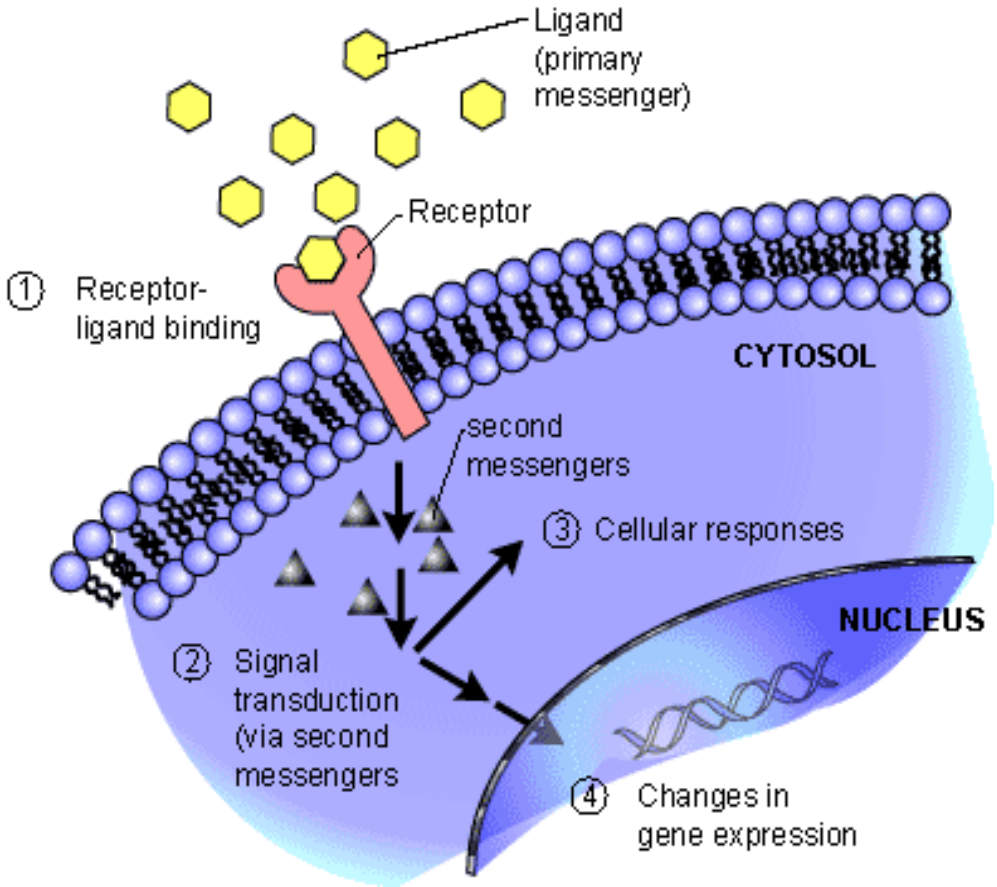
Transcription factors bind DNA, turn on or off different promoters and enhancers, which in-turn turn on or off different genes, some of which may themselves be transcription factors, which again changes the presence of TFs in the cell, the state of active promoters/enhancers etc.



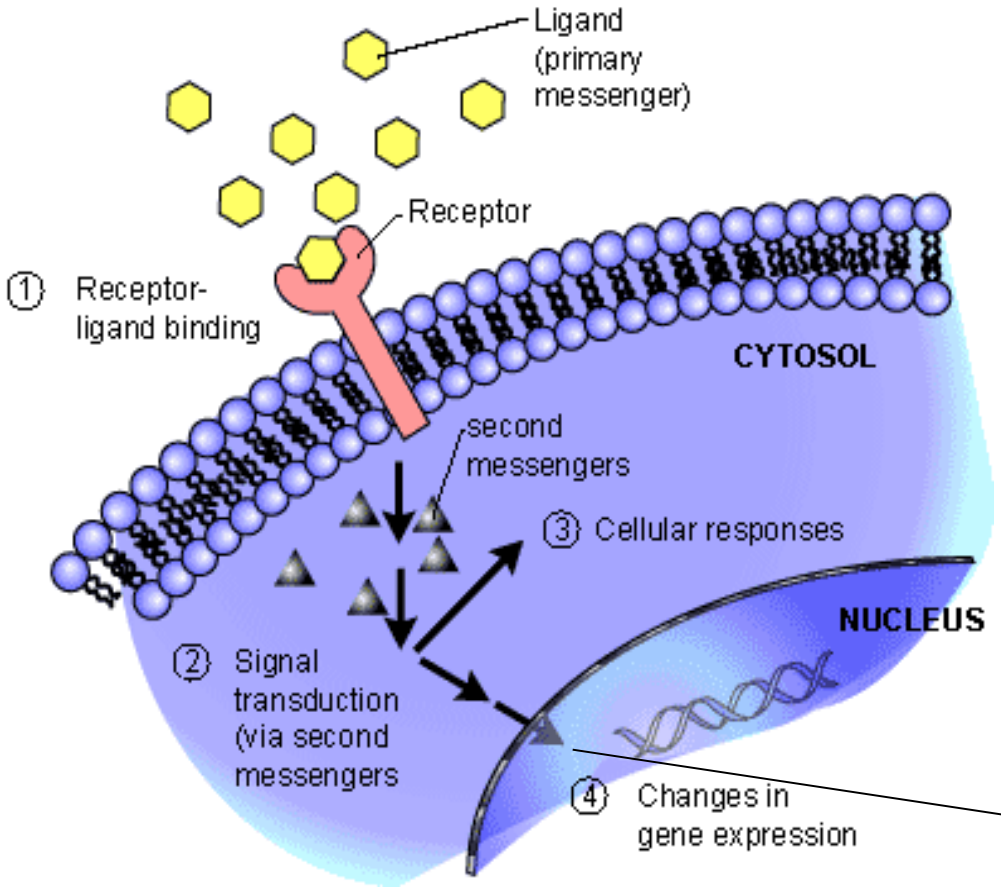
Signal Transduction: distributed computing

Everything we discussed so far happens within the cell.

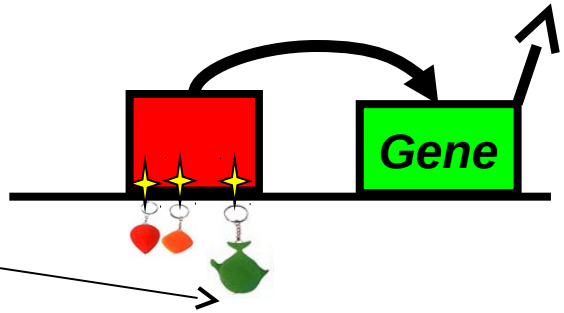
But cells talk to each other, copiously.



Enhancers as Integrators



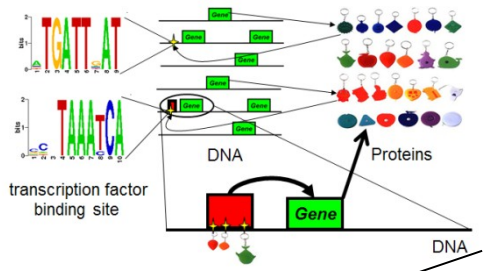
IF the cell is
part of a certain tissue
AND
receives a certain signa
THEN turn Gene ON

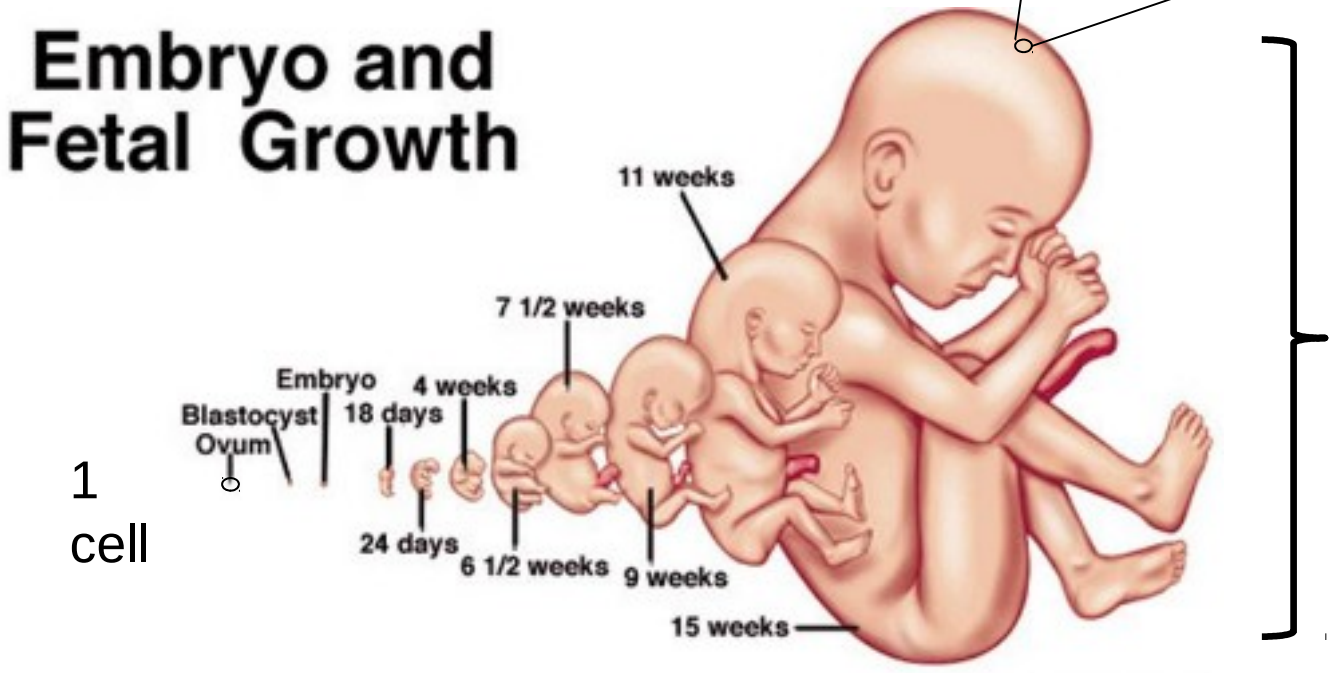


The State Space

Discrete, but very very large.

All states served by same genome(!)



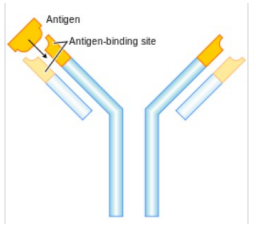


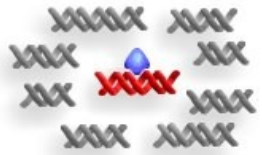
Transcription Activation:

Some measurements and observations

Transcription Factor Binding Sites (TFBS)

- An antibody is a large Y-shaped protein used by the immune system to identify and neutralize foreign objects such as bacteria.
- Antibodies can be raised that instead recognize specific transcription factors.





- Chromatin Immunoprecipitation followed by deep sequencing (ChIP-seq): Take DNA (region or whole genome) bound by TFs, crosslink DNA-TFs, shear DNA, select DNA fragments bound by TF of interest using antibody, get rid of TF and antibody, sequence pool of DNA.

⇒ Obtain genomic regions bound by TF.

ChIP-seq Position Weight Matrix

Sites

ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGGCGTT
ATGACATG
ATGGCATG
ACTGGATG

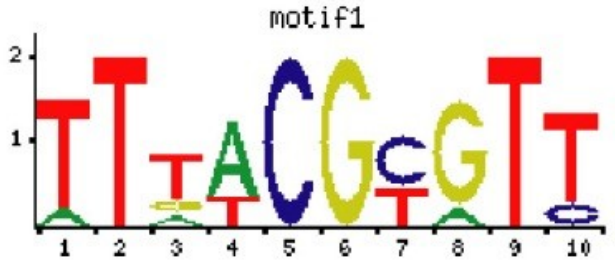
Alignment Matrix

Pos	A	C	G	T
1	9	0	0	1
2	0	1	2	7
3	0	1	7	2
4	1	1	8	0
5	0	7	1	2
6	8	0	2	0
7	0	3	0	7
8	0	0	8	2

Frequency weight Matrix

Pos	A	C	G	T	Con
1	0.9	0	0	0.1	A
2	0	0.1	0.2	0.7	T
3	0	0.1	0.7	0.2	G
4	0.1	0.1	0.8	0	G
5	0	0.7	0.1	0.2	C
6	0.8	0	0.2	0	A
7	0	0.3	0	0.7	T
8	0	0	0.8	0.2	G

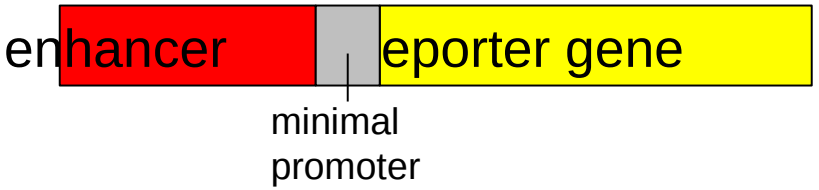
Computational challenge:
The sequenced DNA fragments are 200-500bp.
In each is one or more instance of the 6-20bp motif.
Find it…



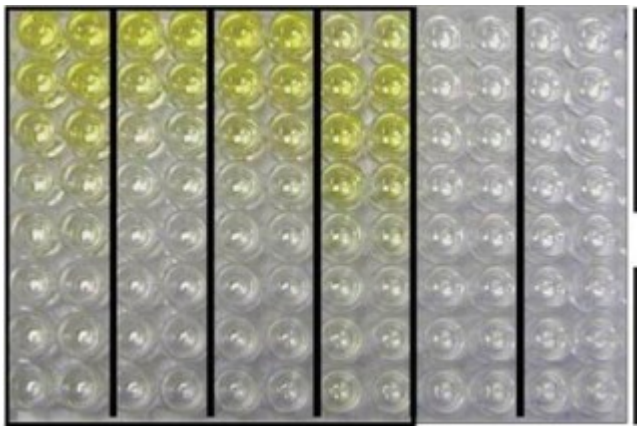

```
A  [ 2  0  2 11  0  0  0  2  0  0 ]
C  [ 0  0  0  0 15  0  9  0  0  3 ]
G  [ 0  0  3  0  0 15  0 13  0  0 ]
T  [13 15 10  4  0  0  6  0 15 12 ]
```

Transfections

As far as we've seen, enhancers work "the same" irrespective of distance (or orientation) to TSS, or identity of target gene.

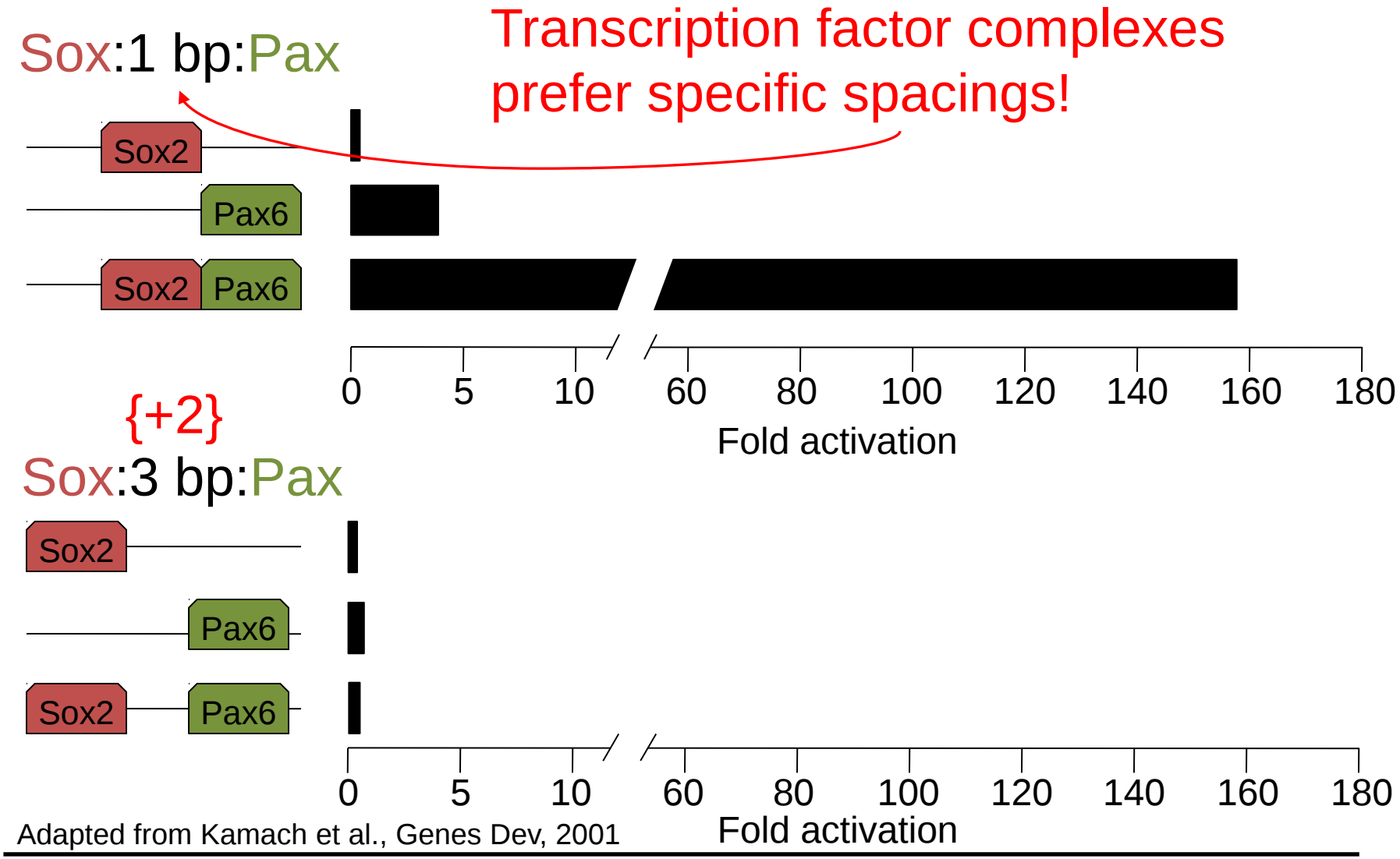


in cellular context of choice



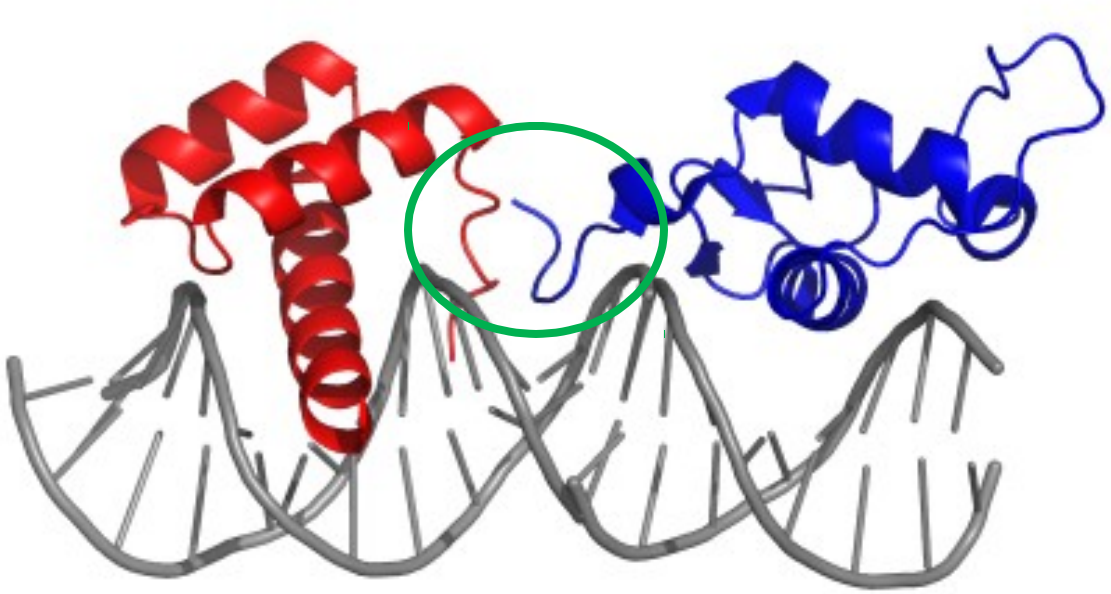
- Which enhancers work in what contexts?
- What if you mutate enhancer bases (disrupt or introduce binding sites) and run the experiment again?
- What if you co-transfect a TF you think binds to this enhancer?
- What if you instead add siRNA for that TF?

Transcription factors bind synergistically, often with preferred spacing

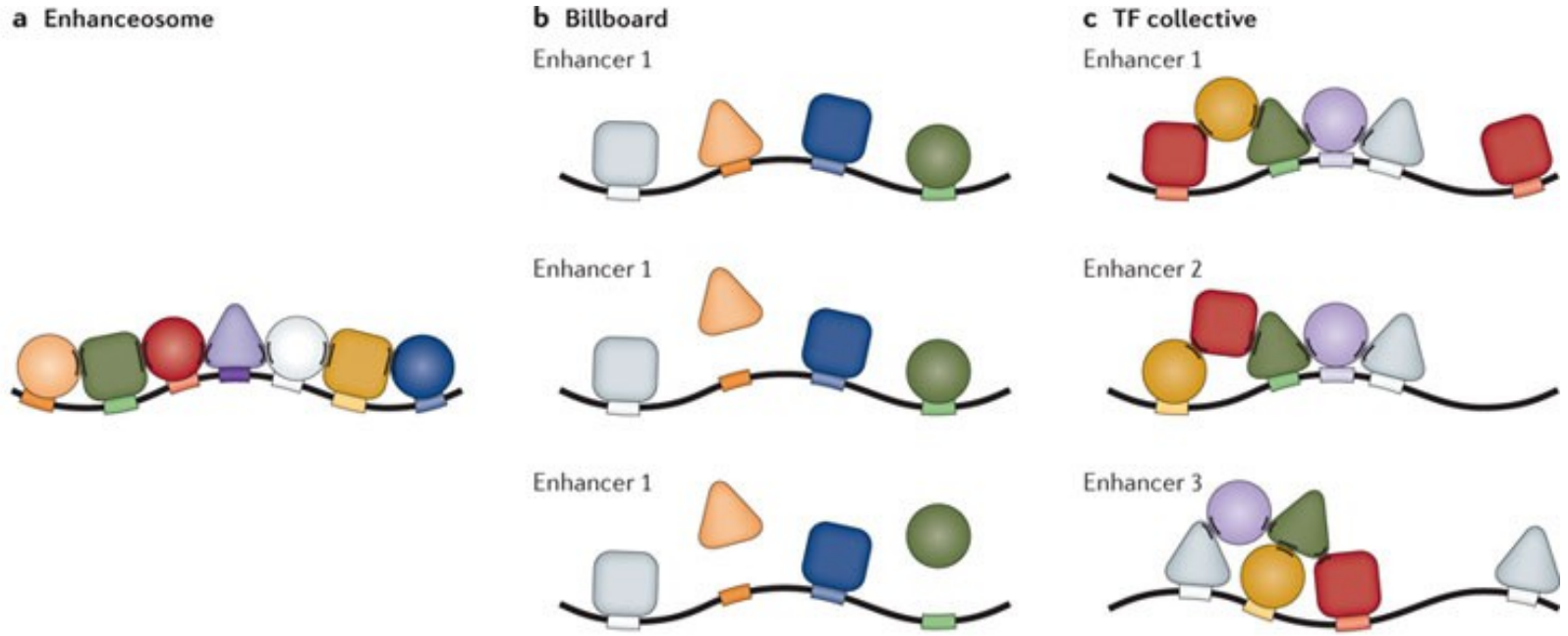


Adapted from Kamach et al., Genes Dev, 2001

Strict spacing between binding sites is important for structural interactions



Different Enhancer Structures



	a Enhanceosome	b Billboard	c TF collective
Protein DNA	• Highly cooperative DNA binding • DNA sequence acts as scaffold	• Cooperative and additive binding	• Cooperative DNA binding • Both DNA and protein may act as scaffold
Protein interface	• Fixed (formed from a higher-order TF–DNA complex)	• Variable	• Variable
Motif	• Fixed motif composition (sites for all factors must be there) • Fixed motif positioning (grammar)	• Fixed motif composition • Flexible motif grammar	• Flexible motif composition (as different TFs directly bind to DNA) • Flexible motif grammar

Massively parallel reporter assays

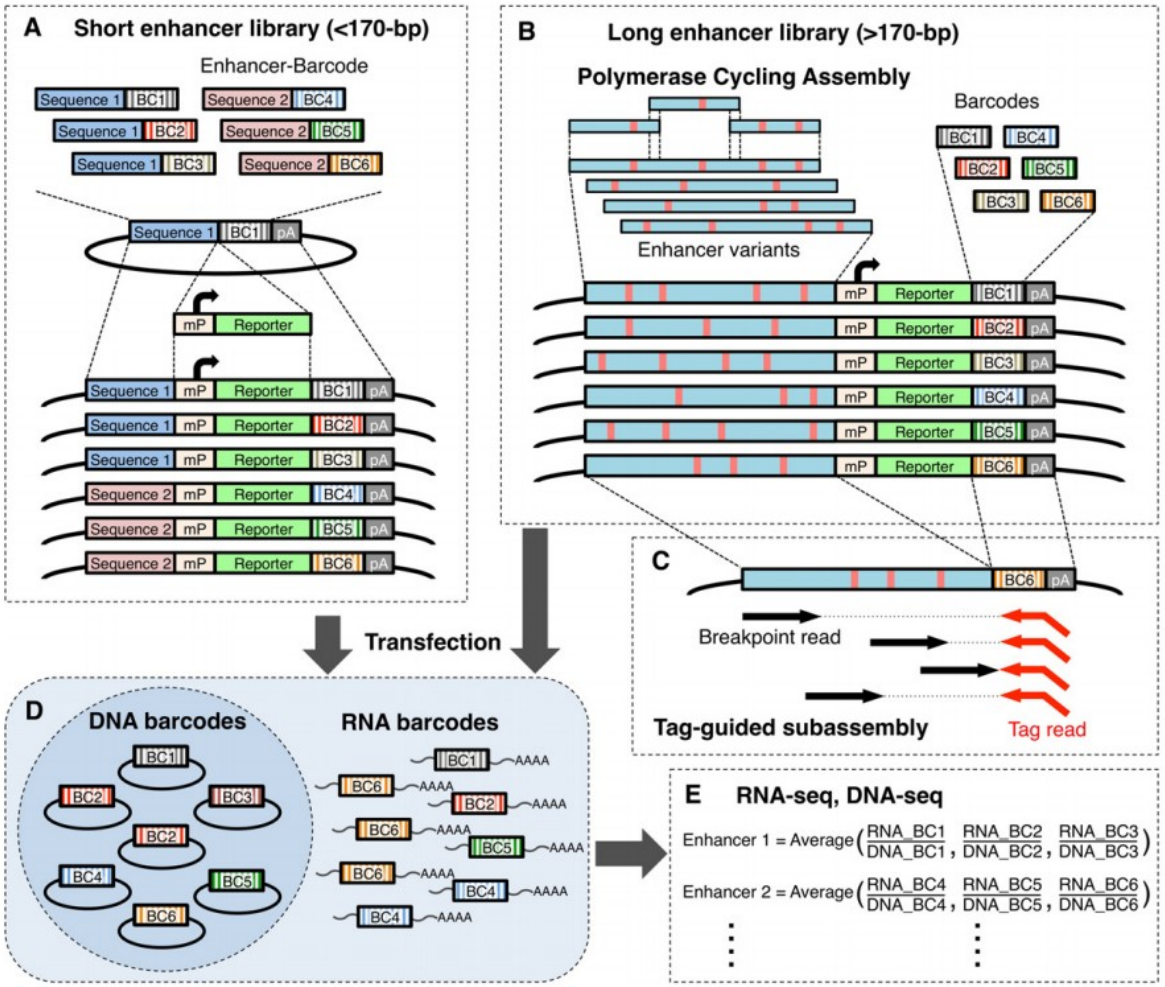


Fig. 2. Overview of the massively parallel reporter assay. (A) To generate a short enhancer MPRA library, candidate enhancer sequences and barcode sequences are synthesized on a programmable microarray and cloned into a plasmid vector. Minimal promoter (mP) and EGFP reporter gene are inserted between the enhancer and barcode. (B) Polymerase cycling assembly is used to generate a long enhancer library. Long enhancer variants (red vertical lines) and barcodes are separately inserted upstream of mP or within the 3′ UTR of EGFP gene, respectively. (C) To associate enhancer and barcodes, the downstream end of the enhancer and upstream end of the barcode are digested and re-ligated to bring them adjacent to one another, followed by sequencing with tag-guided subassembly. (D) The MPRA library is introduced into cell lines or tissues. Transcribed barcode RNAs are extracted from the cells. (E) The relative transcriptional activities of distinct enhancers are measured by sequencing and counting their corresponding barcode RNAs. The counts of transcribed RNA barcodes are normalized by DNA barcode counts that are from the library. BC, barcode; mP, minimal promoter; pA, polyadenylation signal.

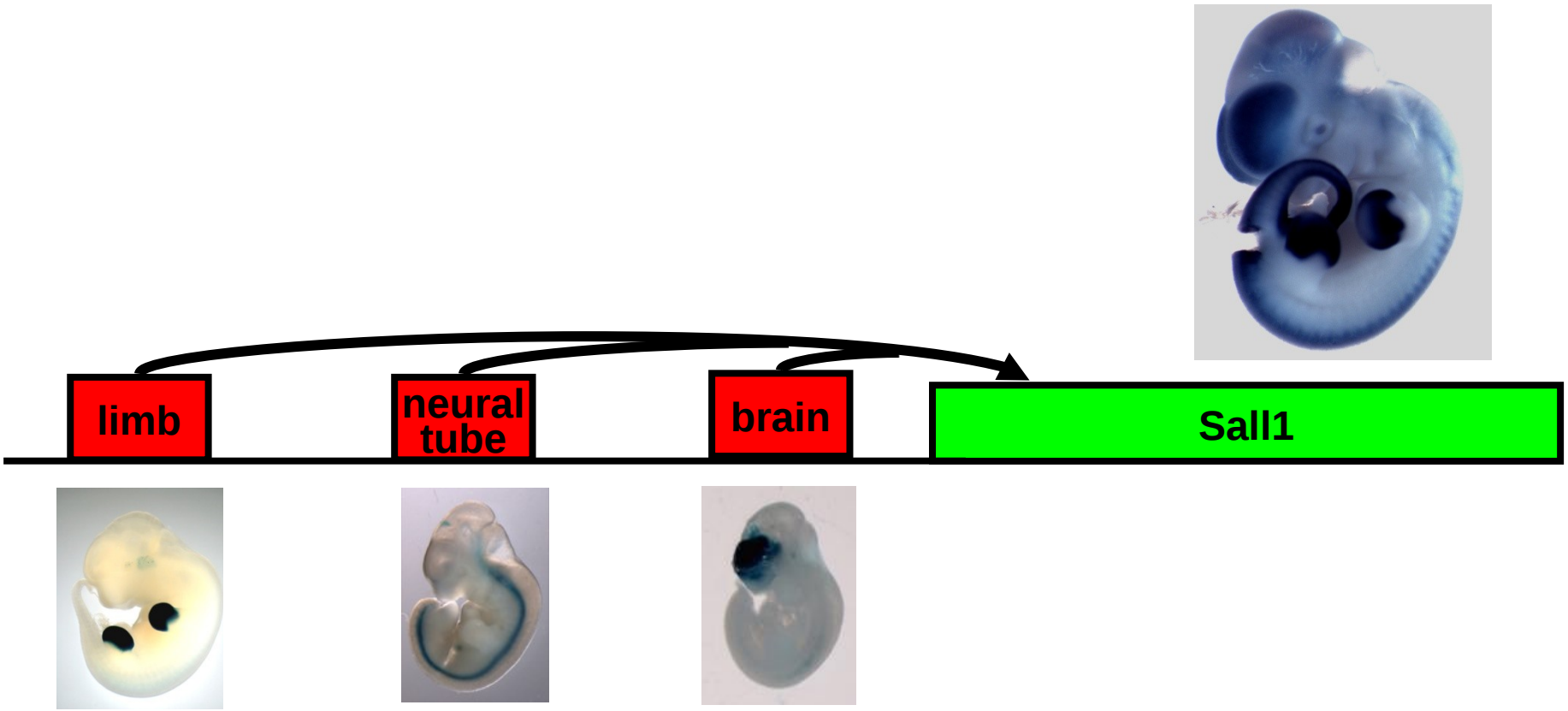
Transgenics

Observe enhancer behavior in vivo.

Qualitative (not quantitative) assay.

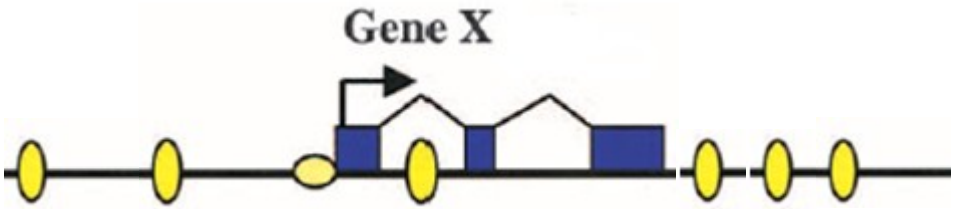
Can section and stain to obtain more specific cell-type information.

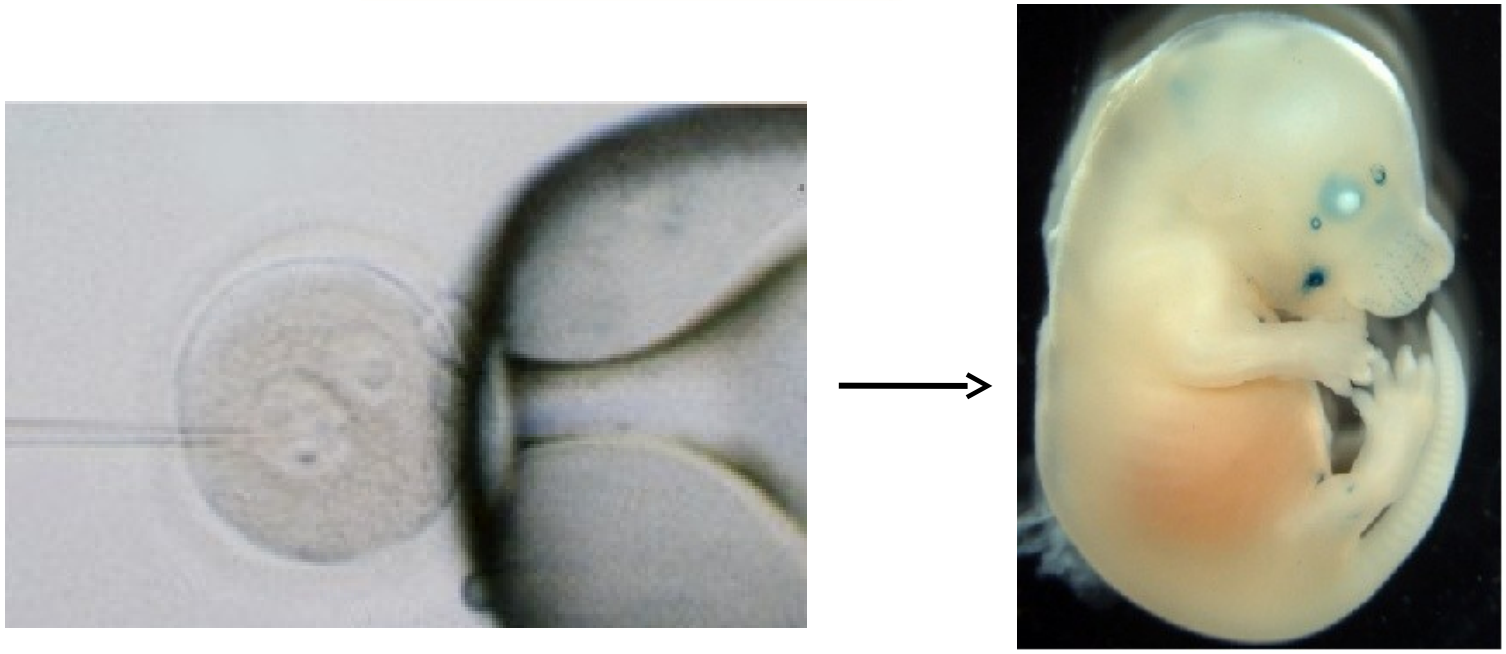
Gene Regulation: Enhancers are modular and additive



Temporal gene expression pattern “equals”
sum of promoter and enhancers expression patterns.

BAC transgenics: necessity vs sufficiency





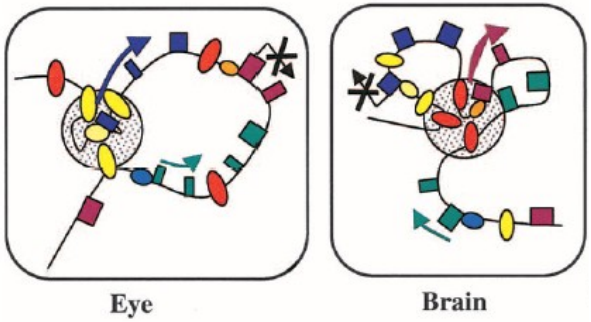
You can take 100-200kb segments out of the genome, insert a reporter gene in place of gene X, and measure regulatory domain expression.

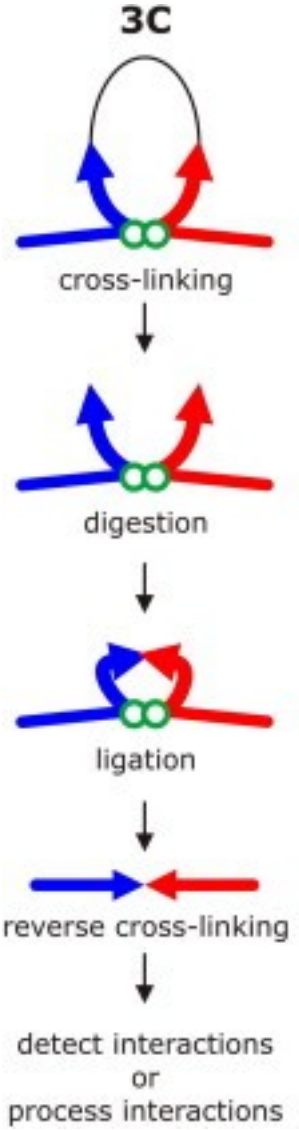
You can then continue to delete or mutate individual enhancers.

Chromosome conformation capture (3C)

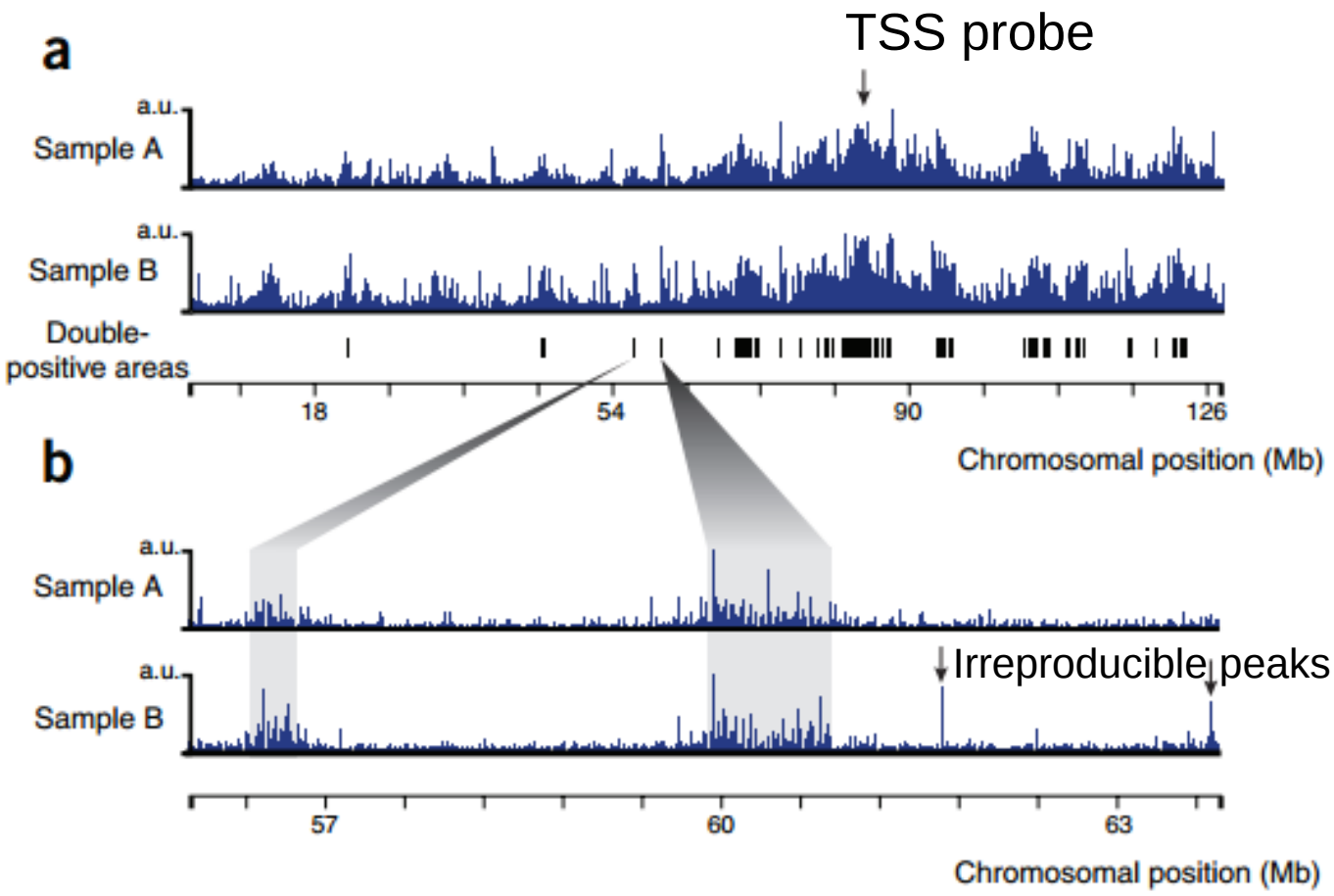
People are also developing methods to detect when two genomic regions far in sequence are in fact interacting in space.

Ultimately this will allow to determine experimentally the regulatory domain of each gene (likely condition dependent).





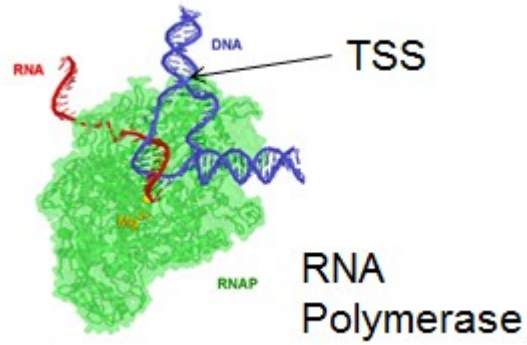
4C example result (in a single biological context)

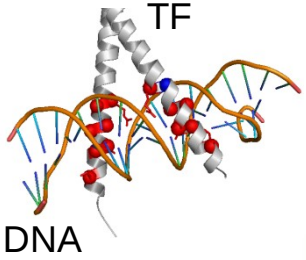


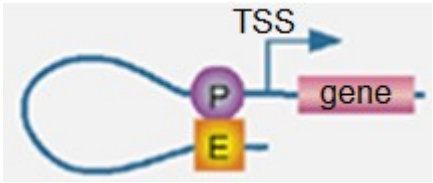
Transcription Activation

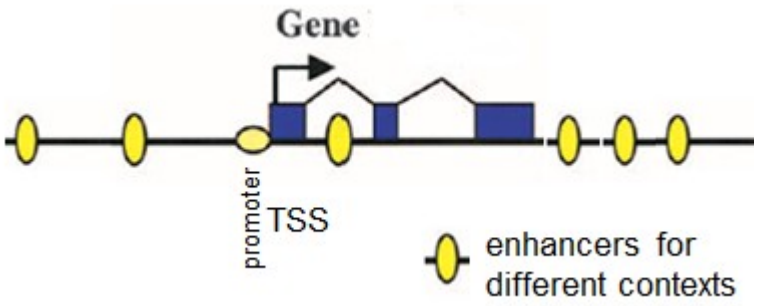
Terminology:

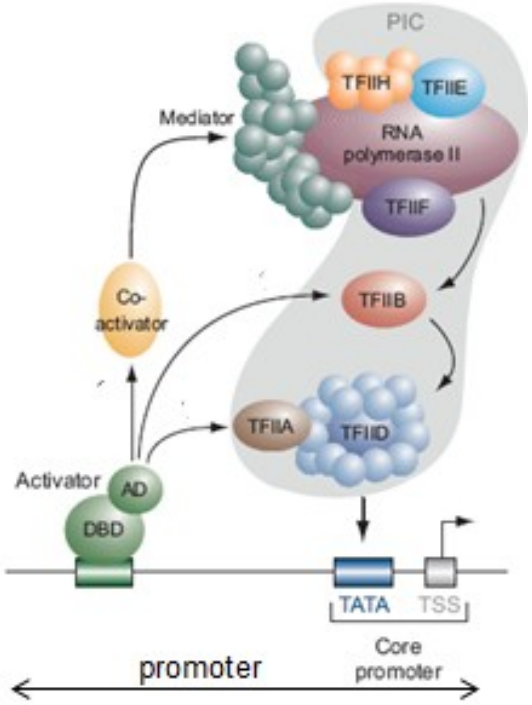
- RNA polymerase
- Transcription Factor
- Transcription Factor Binding Site
- Promoter
- Enhancer
- Gene Regulatory Domain







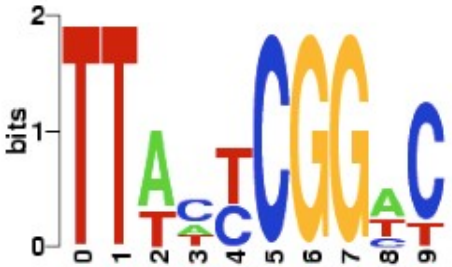


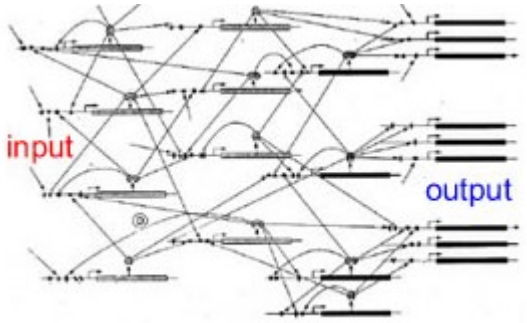


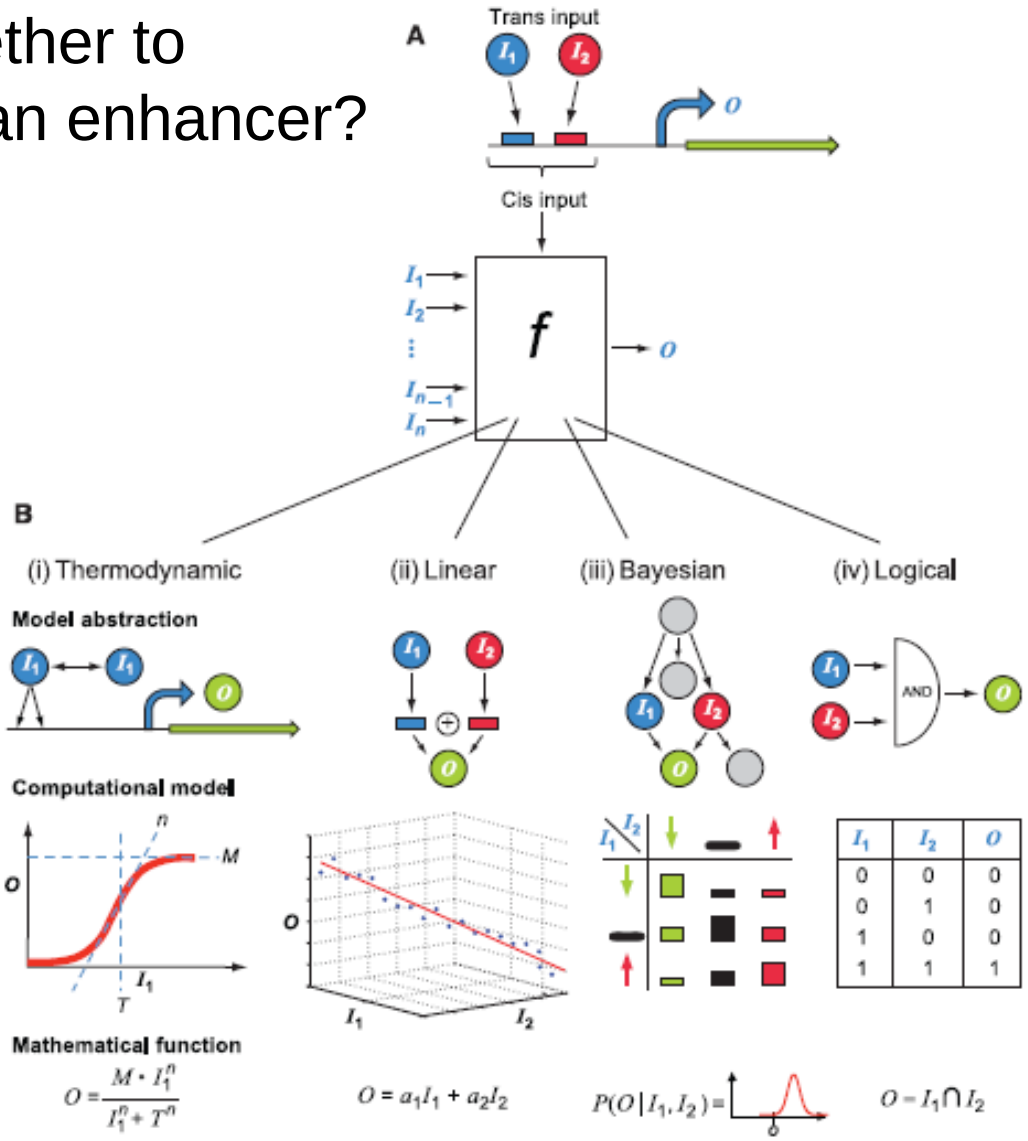
Enhancer Prediction

How do TFs “sum” together to provide the activity of an enhancer?

A network of genes?







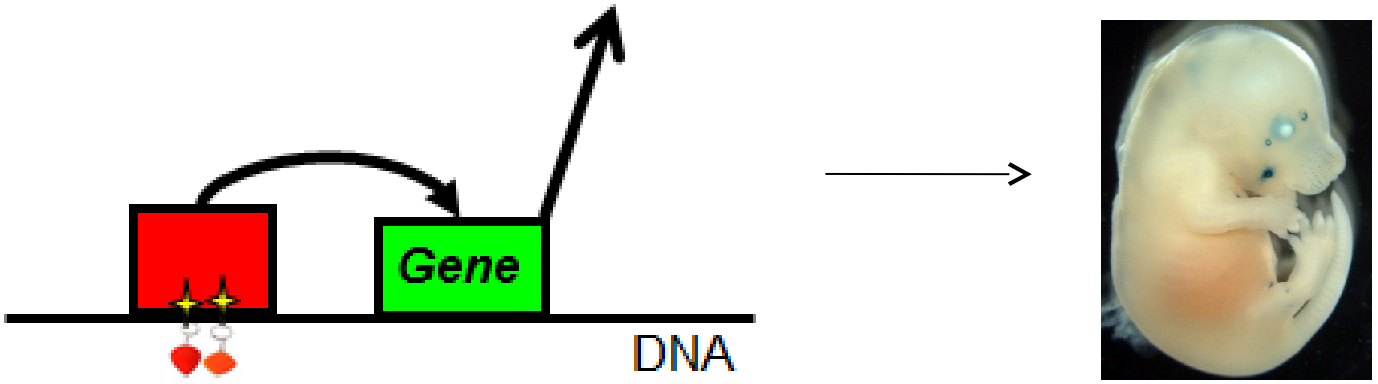
The cis-regulatory code

Given a sequence of DNA predict:

- Is it an enhancer? Ie, can it drive gene expression?
- If so, in which cells? At which times?
- Driven by which transcription factor binding sites?

Given a set of different enhancers driving expression in the same population of cells:

- Do they share any logic? If so what is it?
- Can you generalize this logic to find new enhancers?



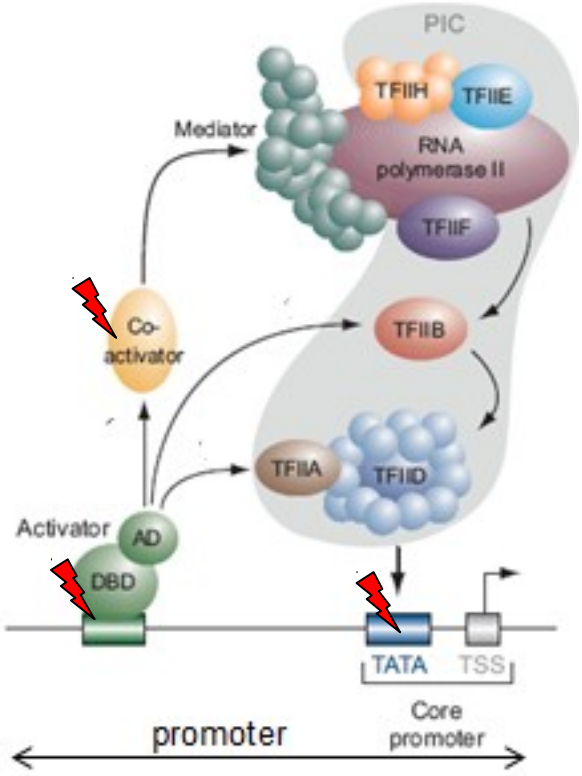
Transcription Regulation
is not just about activation

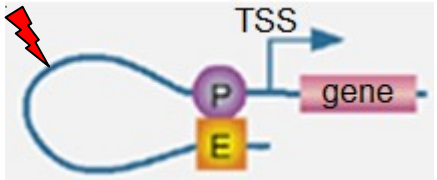
Transcriptional Repression

An equally important but less visible part of transcription (tx) regulation is transcriptional repression (that lowers/ablates tx output).

- Transcription factors can bind key genomic sites, preventing/repelling the binding of
  - The RNA polymerase machinery
  - Activating transcription factors (including via competitive binding)
- Some transcription factors have stereotypical roles as activators or repressors. Likely many can do both (in different contexts).
- DNA can be bent into 3D shape preventing enhancer – promoter interactions.
- Activator and co-activator proteins can be modified into inactive states.

Note: repressor thus can relate to specific DNA sequences or proteins.

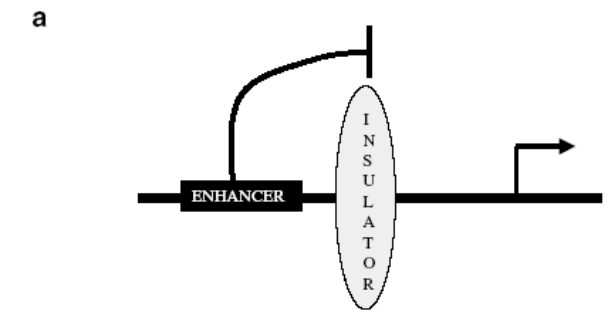


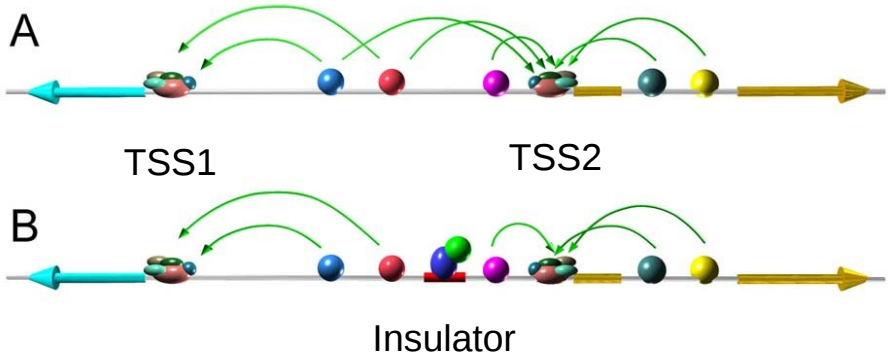


Insulators

Insulators are DNA sequences that when placed between target gene and enhancer prevent enhancer from acting on the gene.

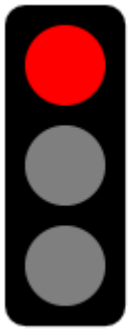
- The handful known insulators contain binding sites for a specific DNA binding protein (CTCF) that is involved in DNA 3D conformation.
- However, CTCF fulfills additional roles besides insulation. I.e, the presence of a CTCF site does not ensure that a genomic region acts as an insulator.



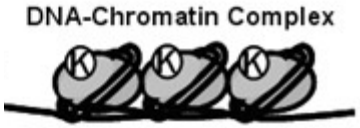


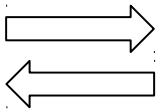
Transcription & its regulation
happen in open chromatin

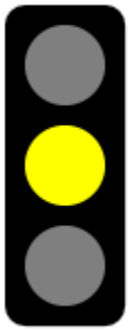
Nucleosomes, Histones, Transcription

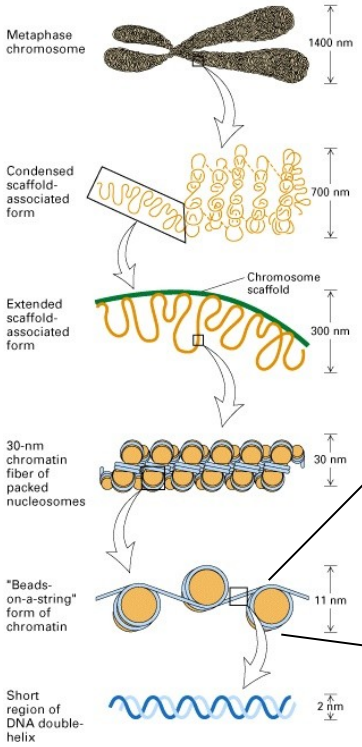


Chromatin / Proteins

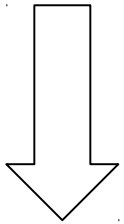




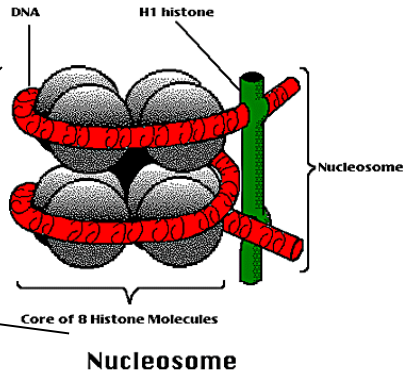


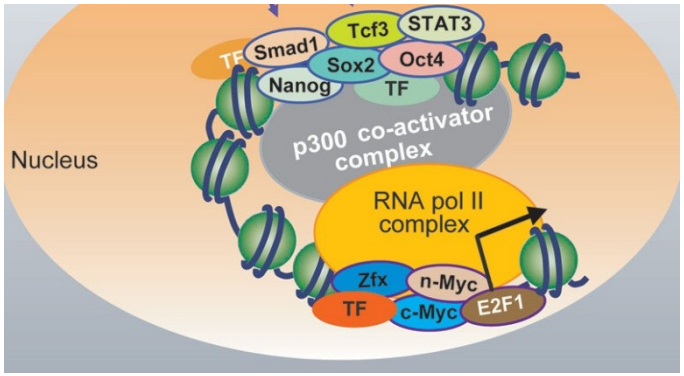


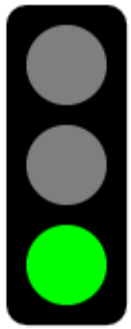
Genome packaging provides a critical layer of gene regulation.



DNA / Proteins



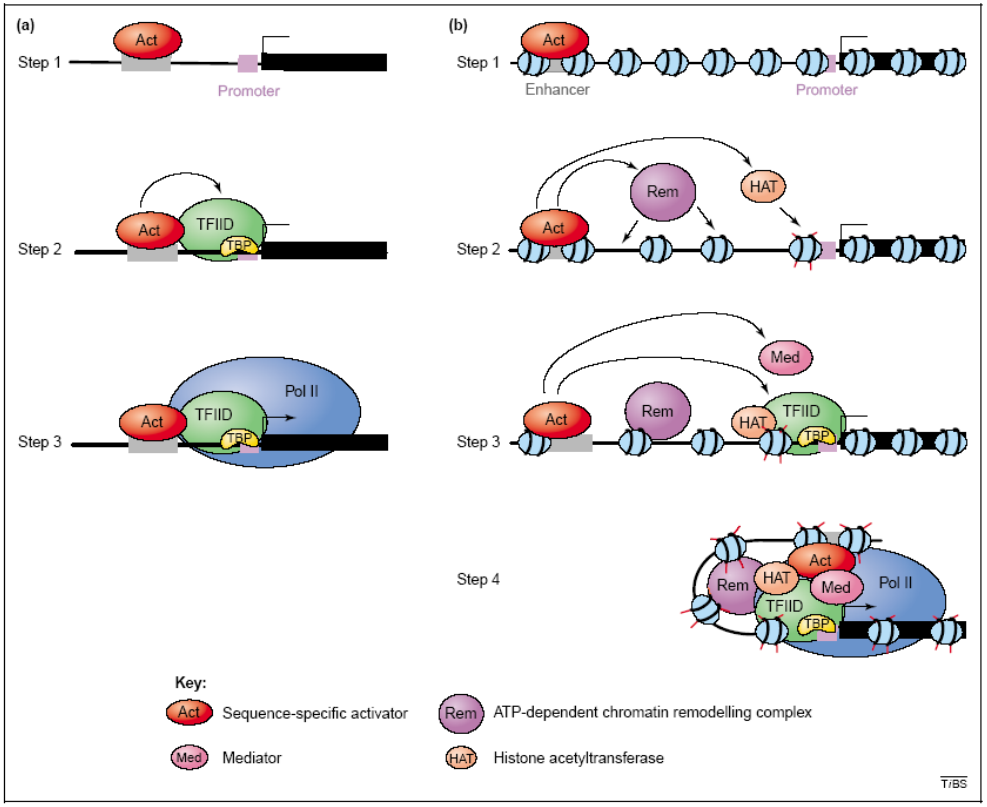




Gene Activation / Repression via Chromatin Remodeling

A dedicated machinery opens and closes chromatin.

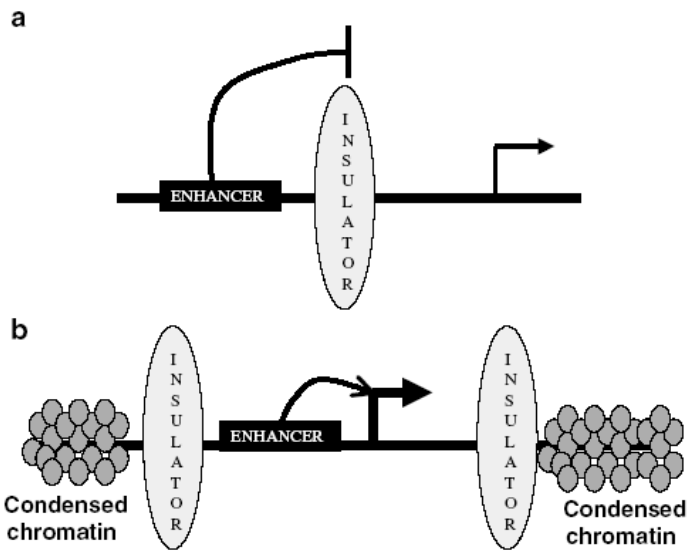
Interactions with this machinery turns genes and/or gene regulatory regions like enhancers and repressors on or off (by making the genomic DNA in/accessible)

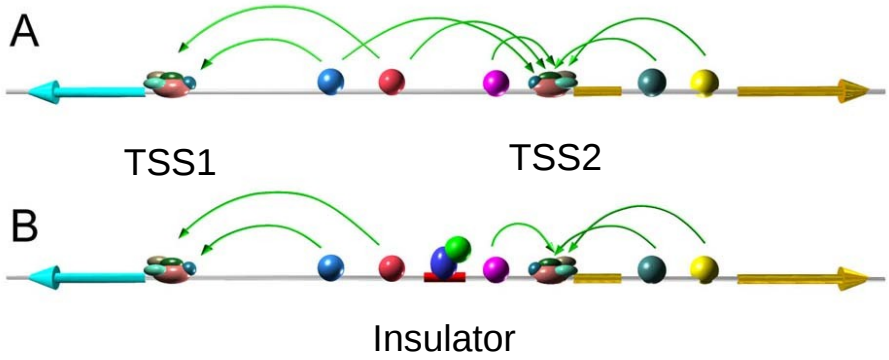


Insulators revisited

Insulators are DNA sequences that when placed between target gene and enhancer prevent enhancer from acting on the gene.

- Known insulators contain binding sites for a specific DNA binding protein (CTCF) that is involved in DNA 3D conformation.
- However, CTCF fulfills additional roles besides insulation. I.e, the presence of a CTCF site does not ensure that a genomic region acts as an insulator.

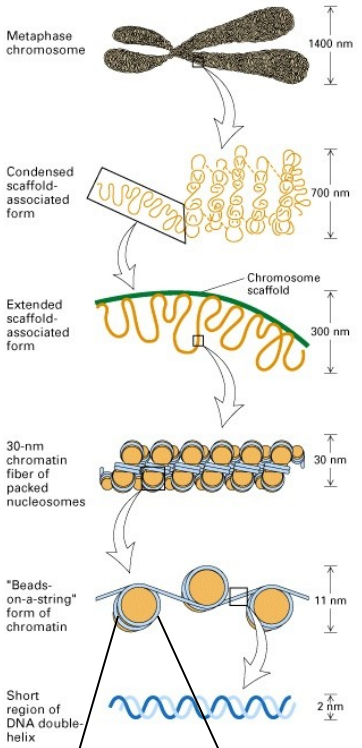


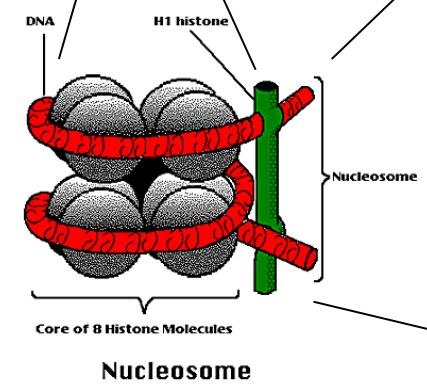


Epigenomics

The histone code

Histone Tails, Histone Marks



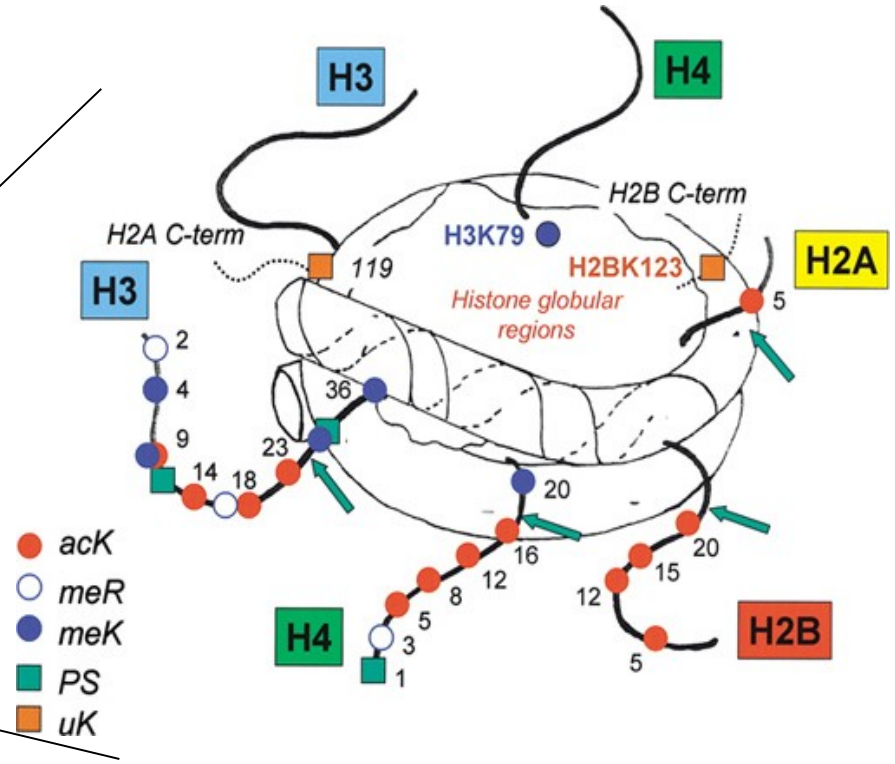


DNA is wrapped around nucleosomes.

Nucleosomes are made of histones.

Histones have free tails.

Residues in the tails are modified in specific patterns in conjunction with specific gene regulation activity.

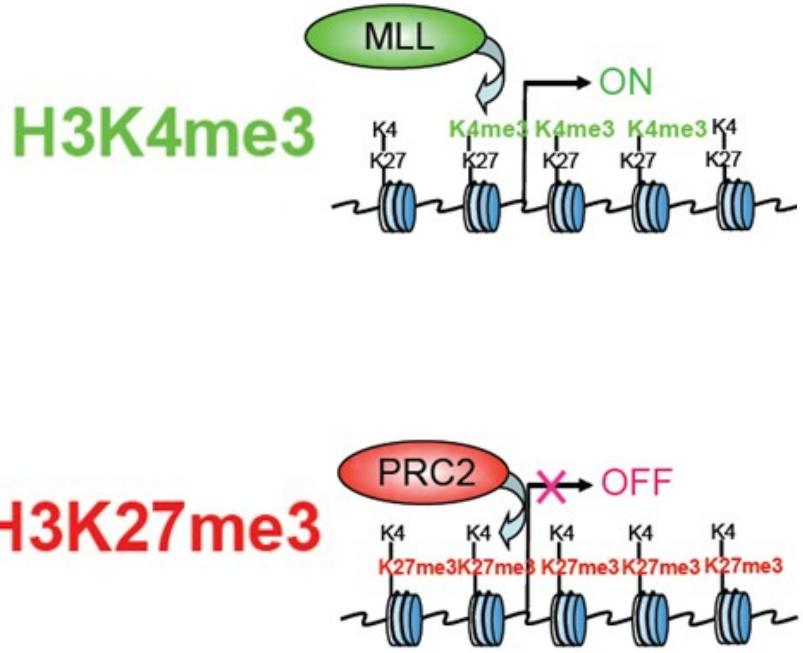


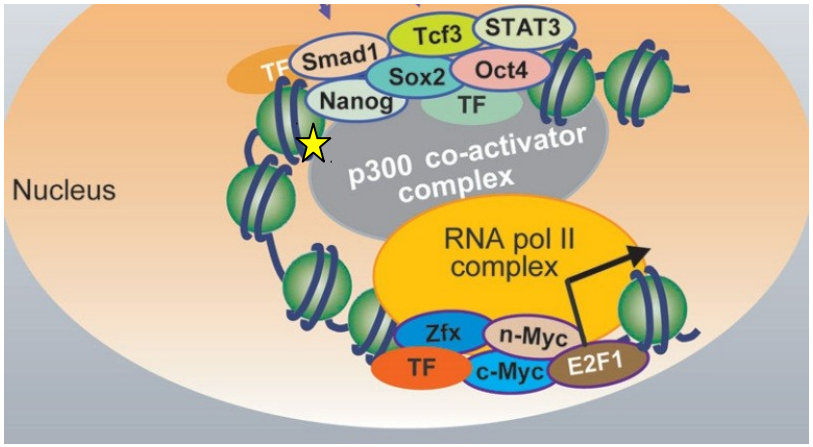
Histone Mark Correlation Examples

Active gene promoters are marked by H3K4me3

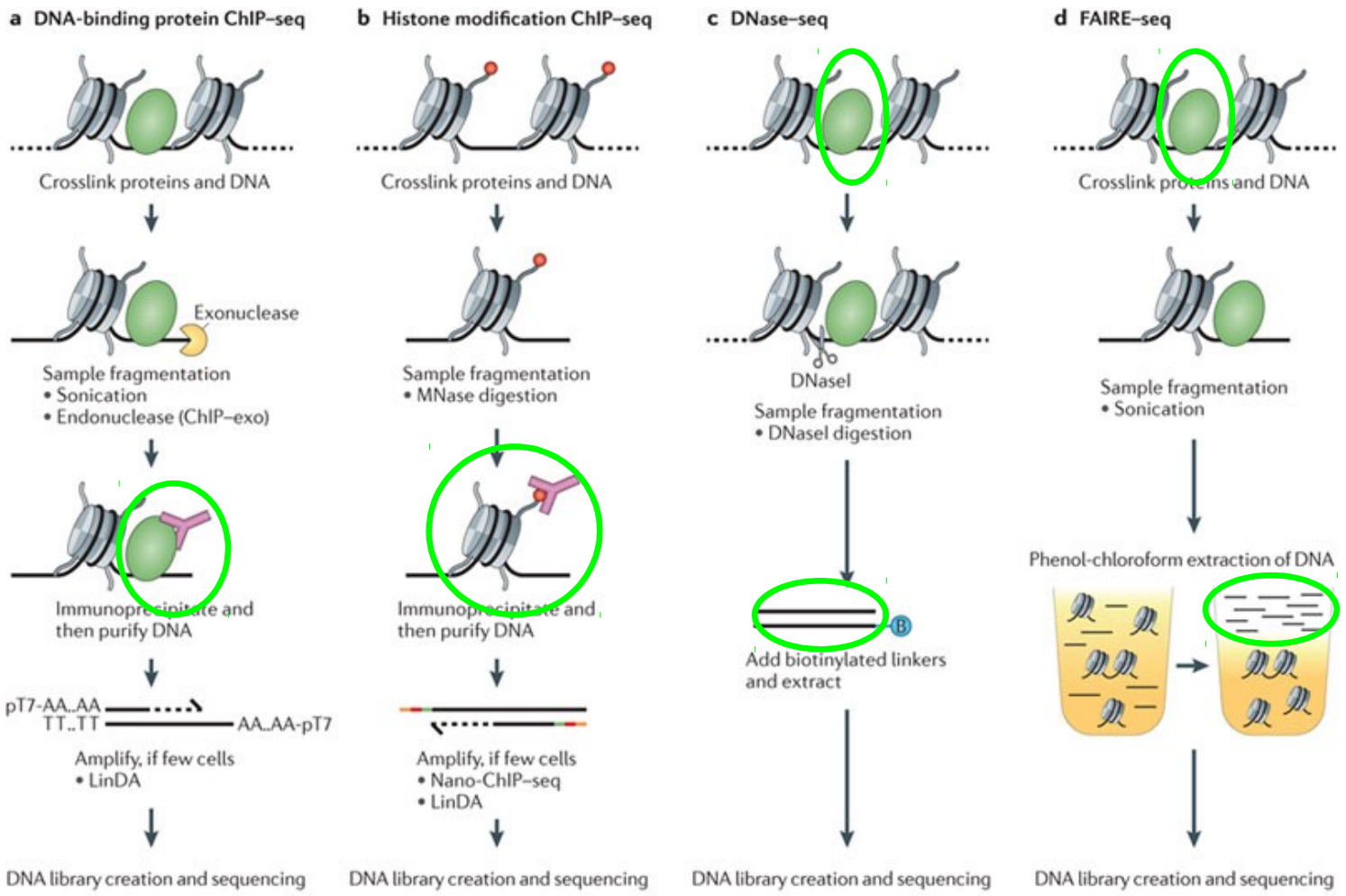
Silenced gene promoters are marked by H3K27me3

p300, a protein component of many active enhancers acetylates H3k27Ac.



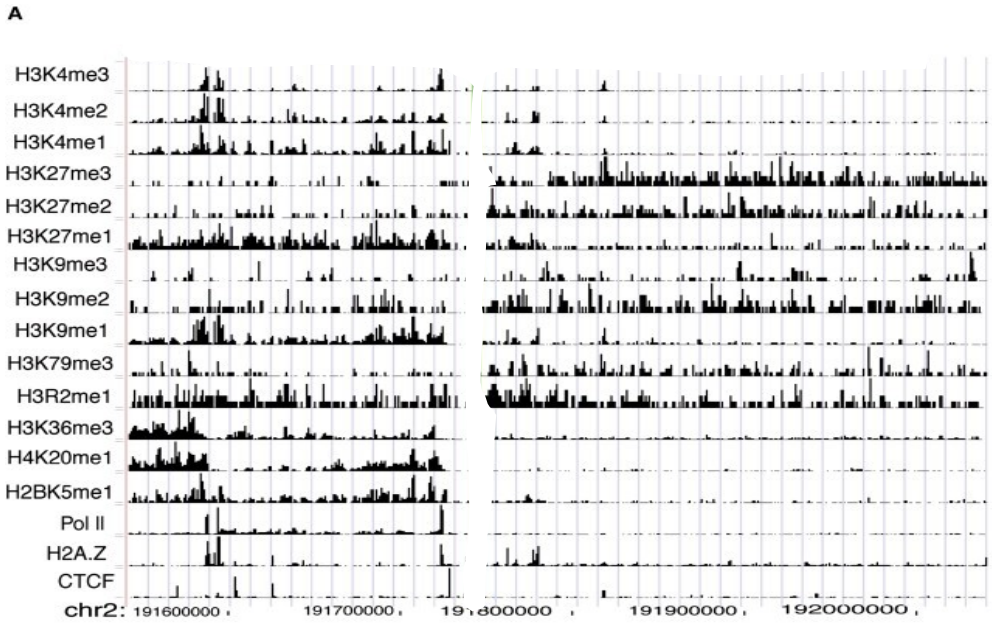


Measuring these different states

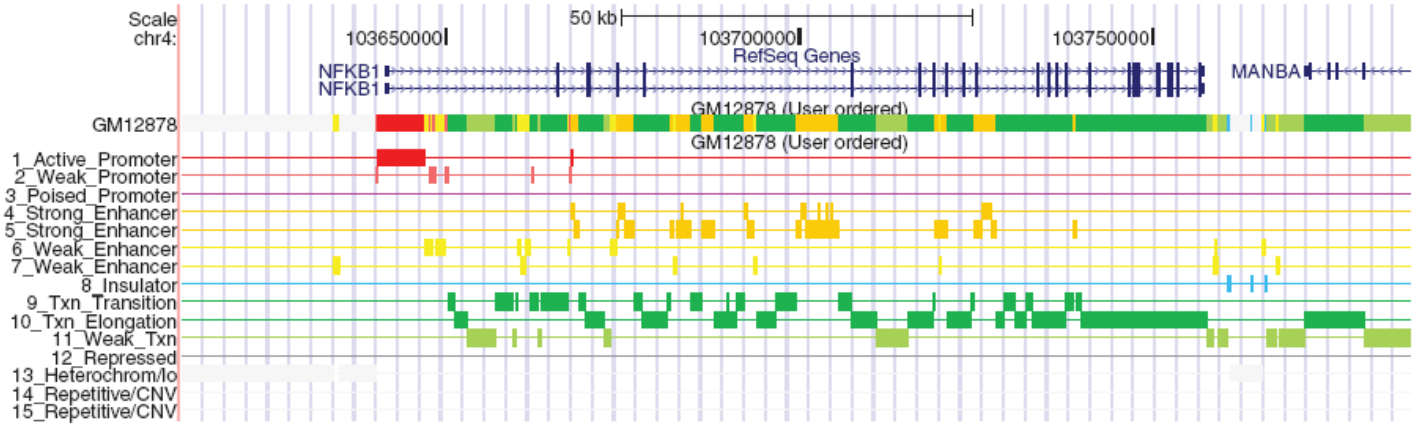


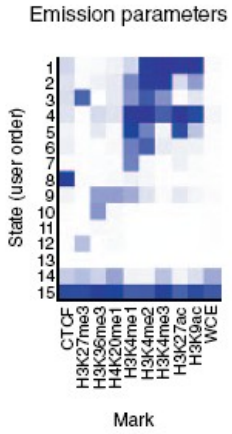
Note that the DNA itself doesn't change. We sequence different portions of it that are currently in different states (bound by a TF, wrapped around a nucleosome etc.)

Epigenomics: study all these marks genomewide



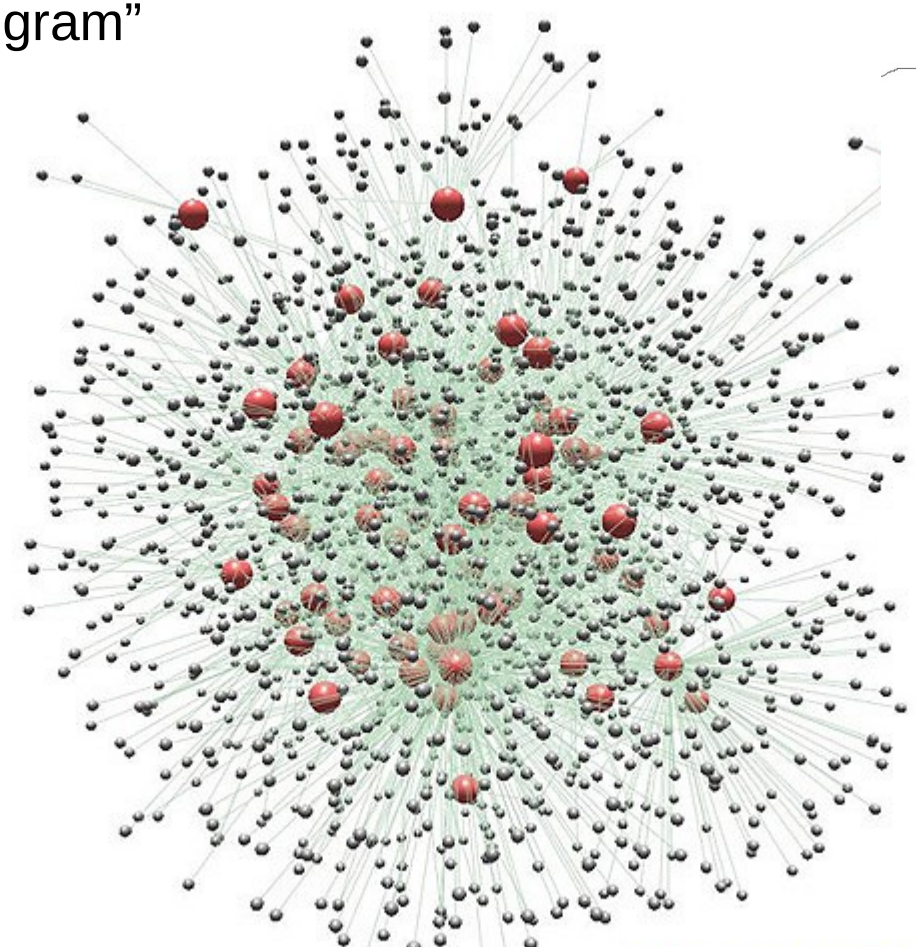
Translate observations into current genome state.

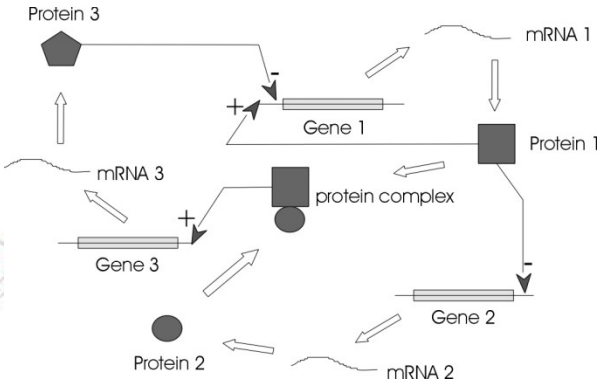




Obtain a network of all active genes & DNA

"Ridicilogram"

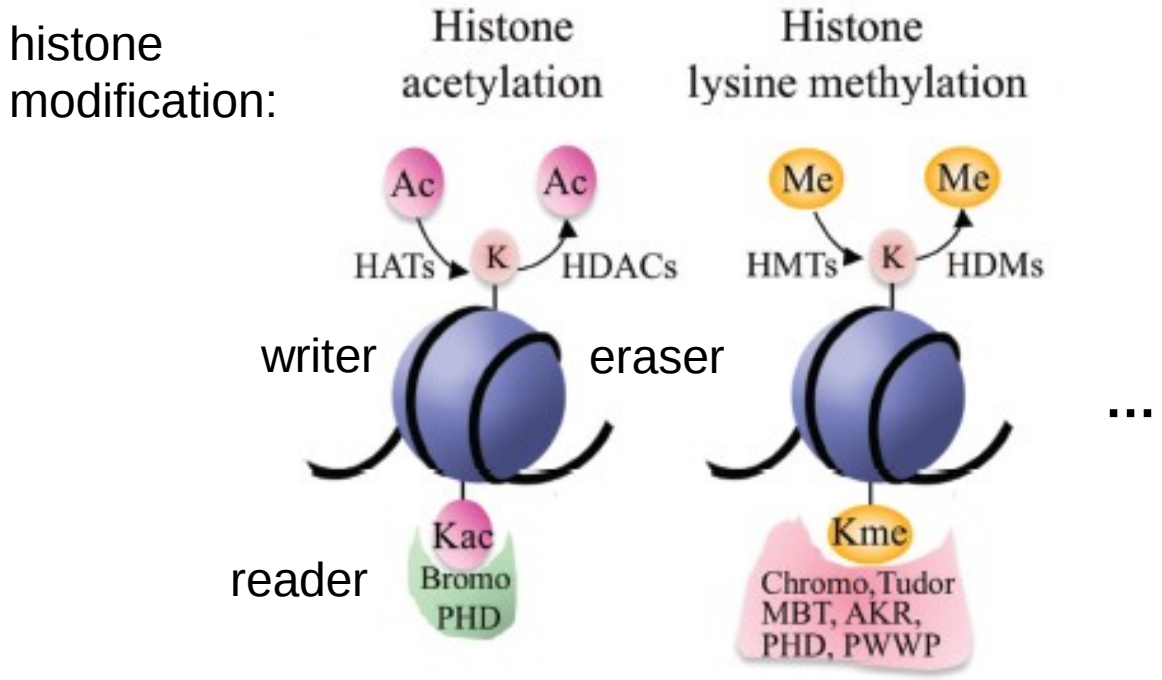




Now what?
(to be revisited)

Histone Code Hypothesis

Histone modifications serve to recruit other proteins by specific recognition of the modified histone via protein domains specialized for such purposes, rather than through simply stabilizing or destabilizing the interaction between histone and the underlying DNA.



Epigenomics is not Epigenetics

Epigenetics is the study of heritable changes in gene expression or cellular phenotype, caused by mechanisms other than changes in the underlying DNA sequence

There are objections to the use of the term epigenetic to describe chemical modification of histone, since it remains unknown whether or not these modifications are heritable.

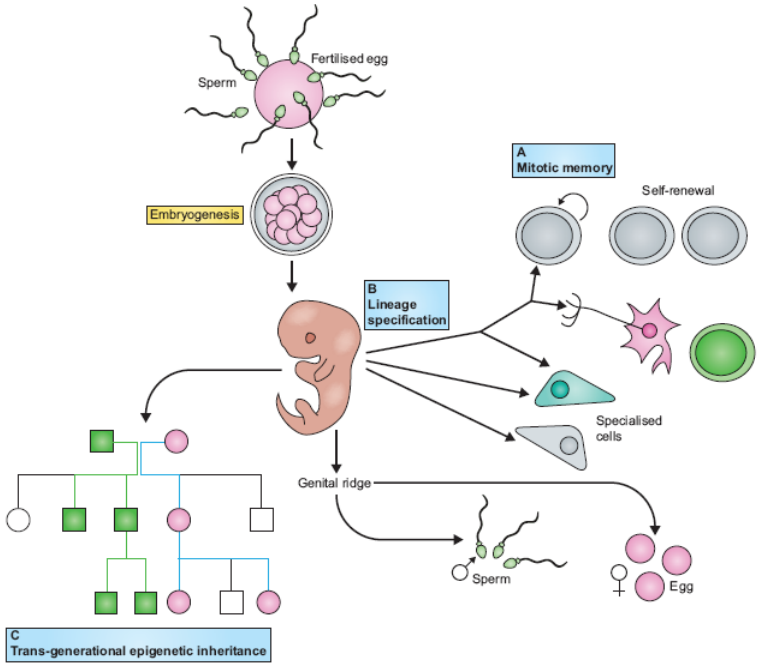
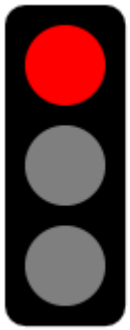
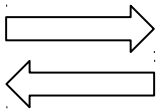


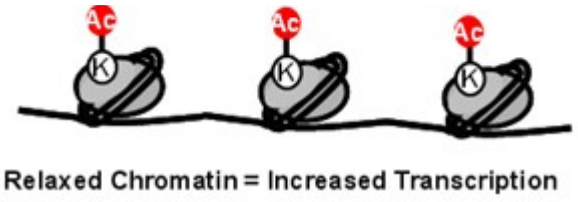
Fig. 2. Major questions in epigenetic memory. As well as understanding the mechanistic basis for 'mitotic memory', in which cellular identity is propagated through stem cell self renewal and somatic cell proliferation (**A**), meeting participants highlighted two other important epigenetic phenomena. The specification of different cell types (and transcriptional programs) from a fertilised egg (**B**) remains a classical epigenetic process in which the same genomic sequence is retained by the 200 or so different specialised cell types of the organism, and cell function is susceptible to experimental reprogramming. Specialised mammalian germ cells that originate from the genital ridge in developing embryos give rise to haploid sperm and egg. Understanding how epigenetics contributes to trans-generational epigenetic inheritance – in which parental experience is transmitted to successive generations of offspring (**C**), presumably through mechanisms that impact on their germ cells – remains a less well understood but important area of epigenetic research.

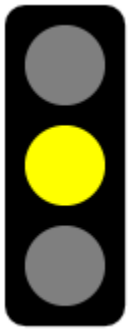
Gene Regulation

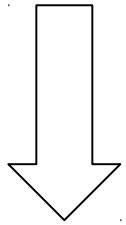


Chromatin / Proteins



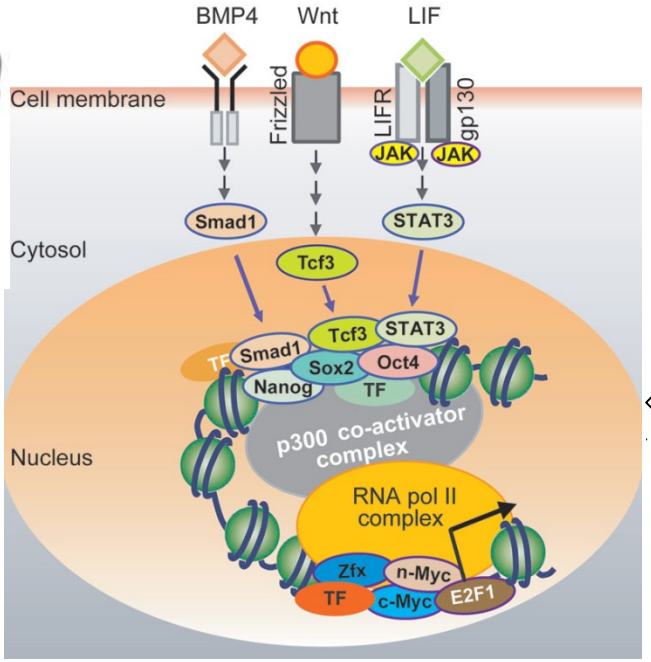


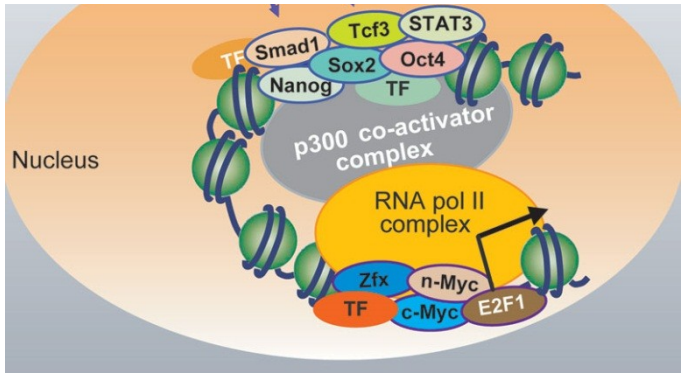


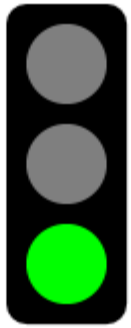


DNA / Proteins

Extracellular signals







Cis-Regulatory Components

Low level (“atoms”):

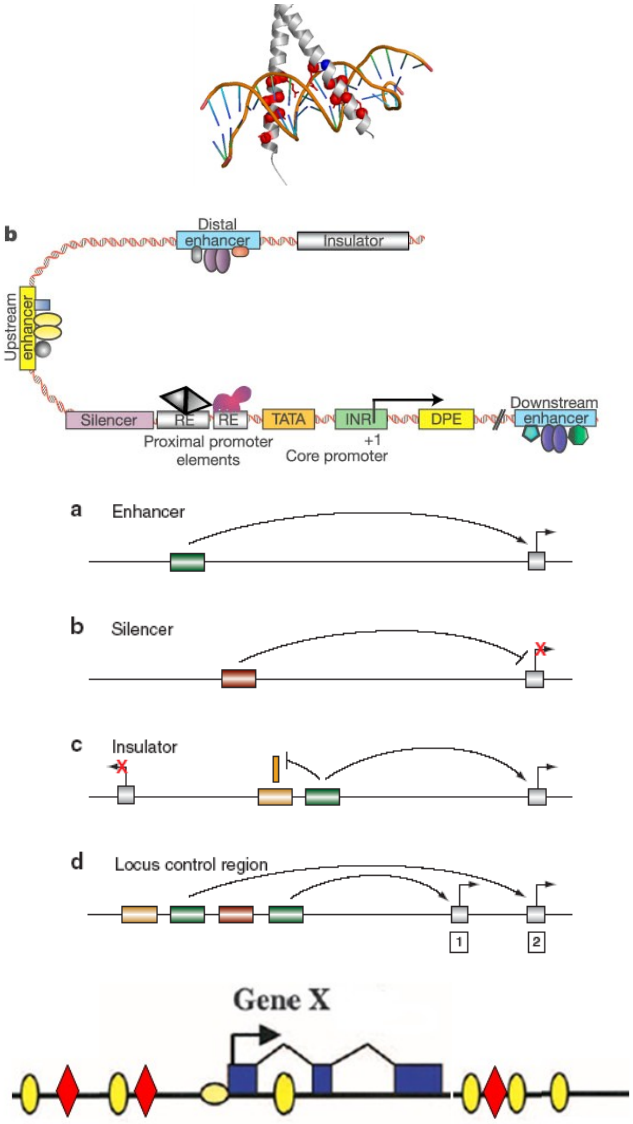
- Promoter motifs (TATA box, etc)
- Transcription factor binding sites (TFBS)

Mid Level:

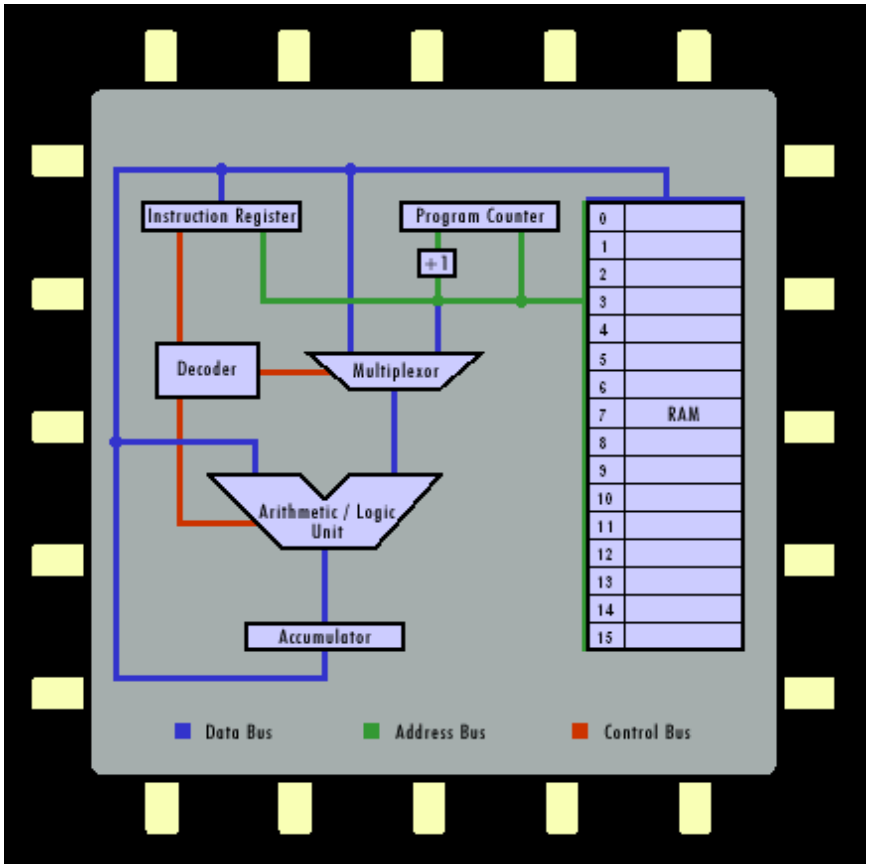
- Promoter
- Enhancers
- Repressors/silencers
- Insulators/boundary elements
- Locus control regions

High Level:

- Epigenomic domains / signatures
- Gene expression domains
- Gene regulatory networks



If you only measure gene expression



It's like only seeing the values change in RAM as a program is running.

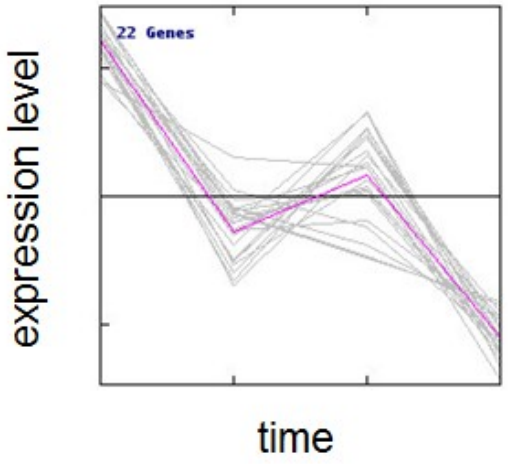
Inferring Gene Expression Causality

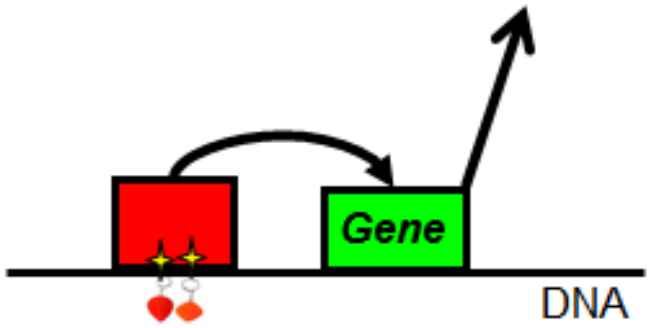
Measuring gene expression over time provides sets of genes that change their expression in synchrony.

- But who regulates whom?
- Some of the necessary regulators may not change their expression level when measured, and yet be essential.

“Reading” enhancers can provide gene regulatory logic:

- If present(TF A, TF B, TF C) then turn on nearby gene X





Gene Regulation is in Data Deluge mode

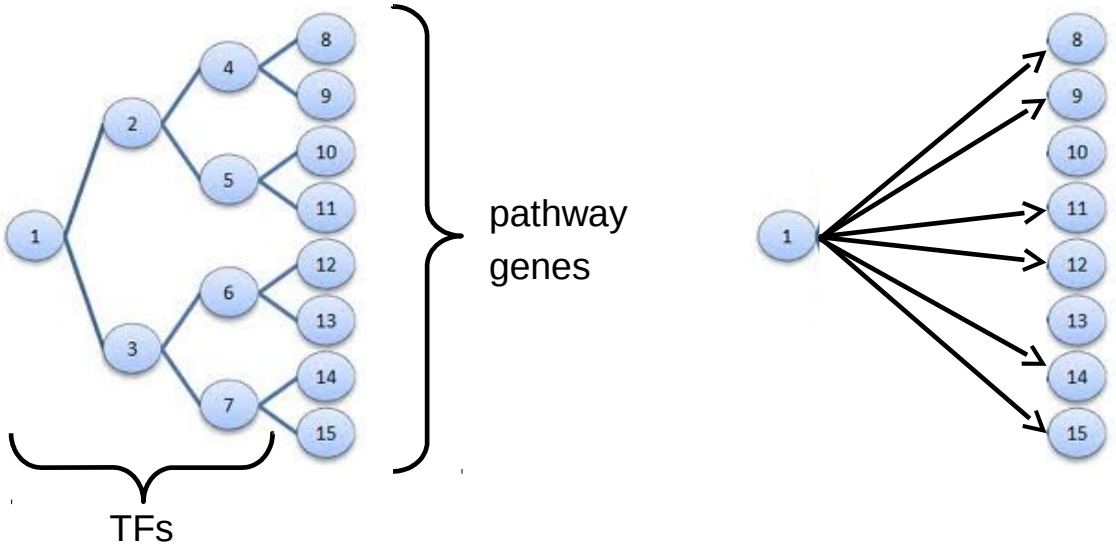
“Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom.”

Transcription Factors have Large “fan outs”

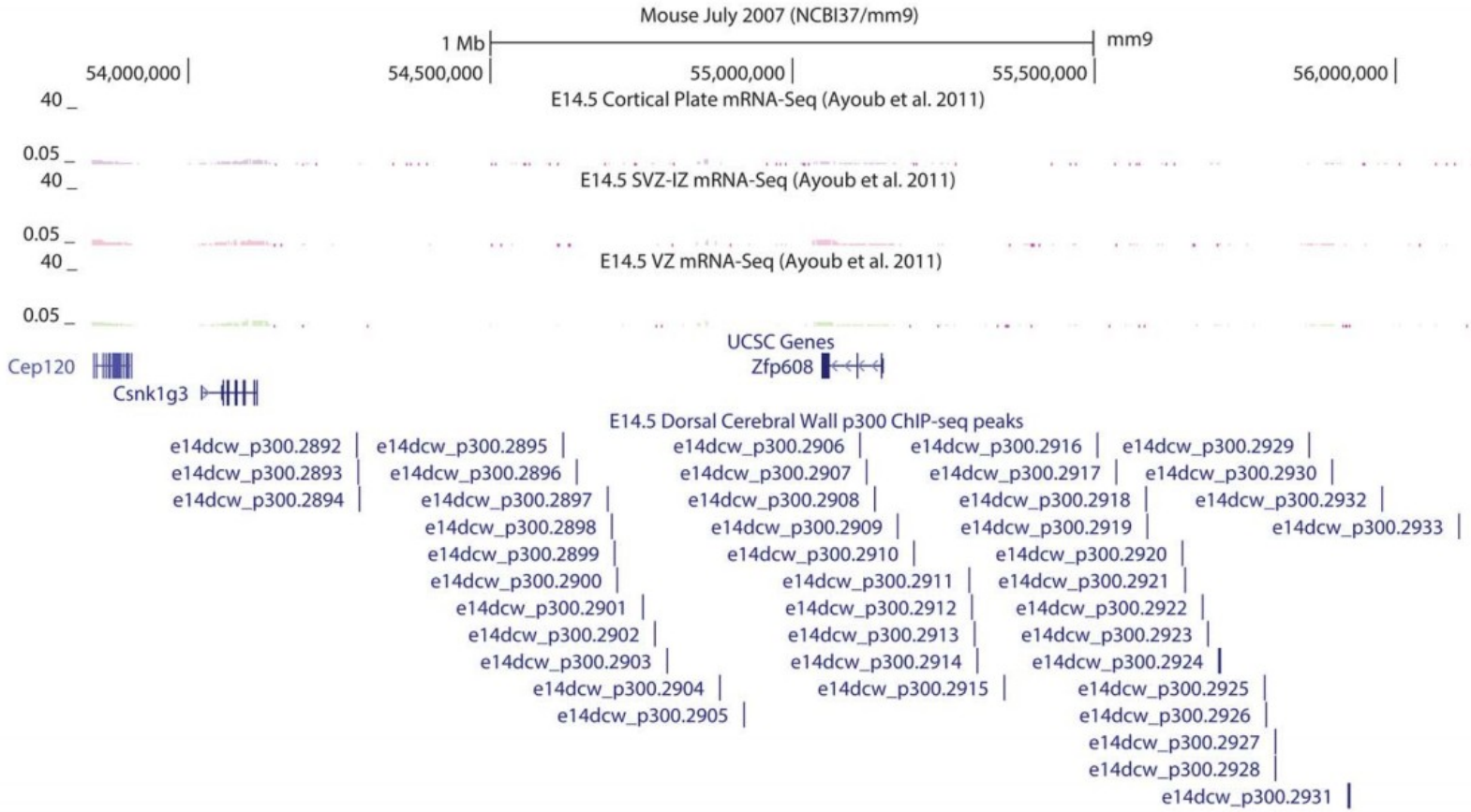
We could have had one TF regulate two TFS, each of which regulates two other TFs, etc. and each of those contributing to the regulation of a modest number of target genes (that do the real work).

Instead TFs reproducibly bind to thousands of genomic locations almost anywhere we’ve looked.

Gene regulation forms a dense network.



Some important genes have large “fan ins”



We are technically DONE with genome function

Biology – not that complicated!!

Functional part list

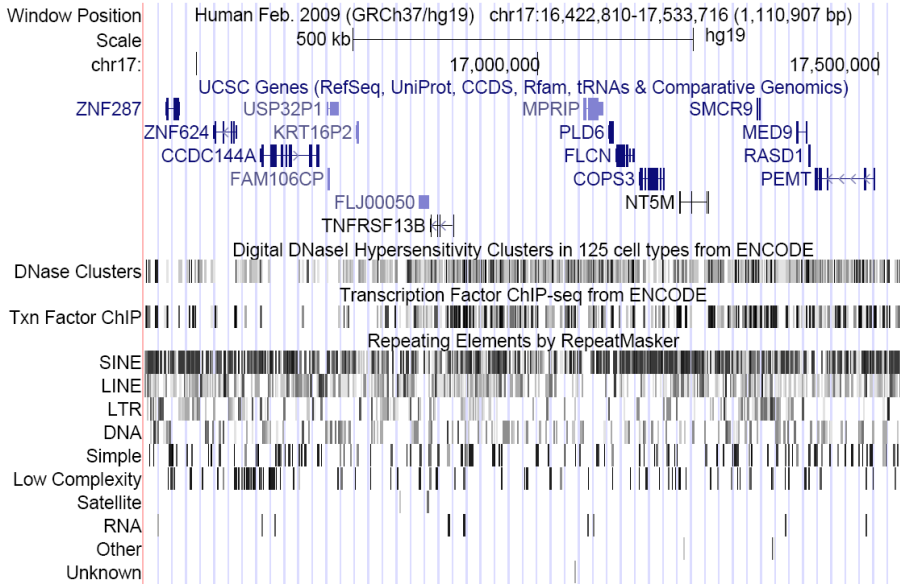
- In our genome:
  - Gene
    - Protein coding
    - Non coding / RNA genes
  - Gene regulatory elements
    - “Atomic” event: transcription factor binding site
    - Build up: promoters, enhancers, silencers, gene reg. domain
- “Around” our genome
  - Chromatin – open / closed
  - Epigenomic (and some epigenetic) marks

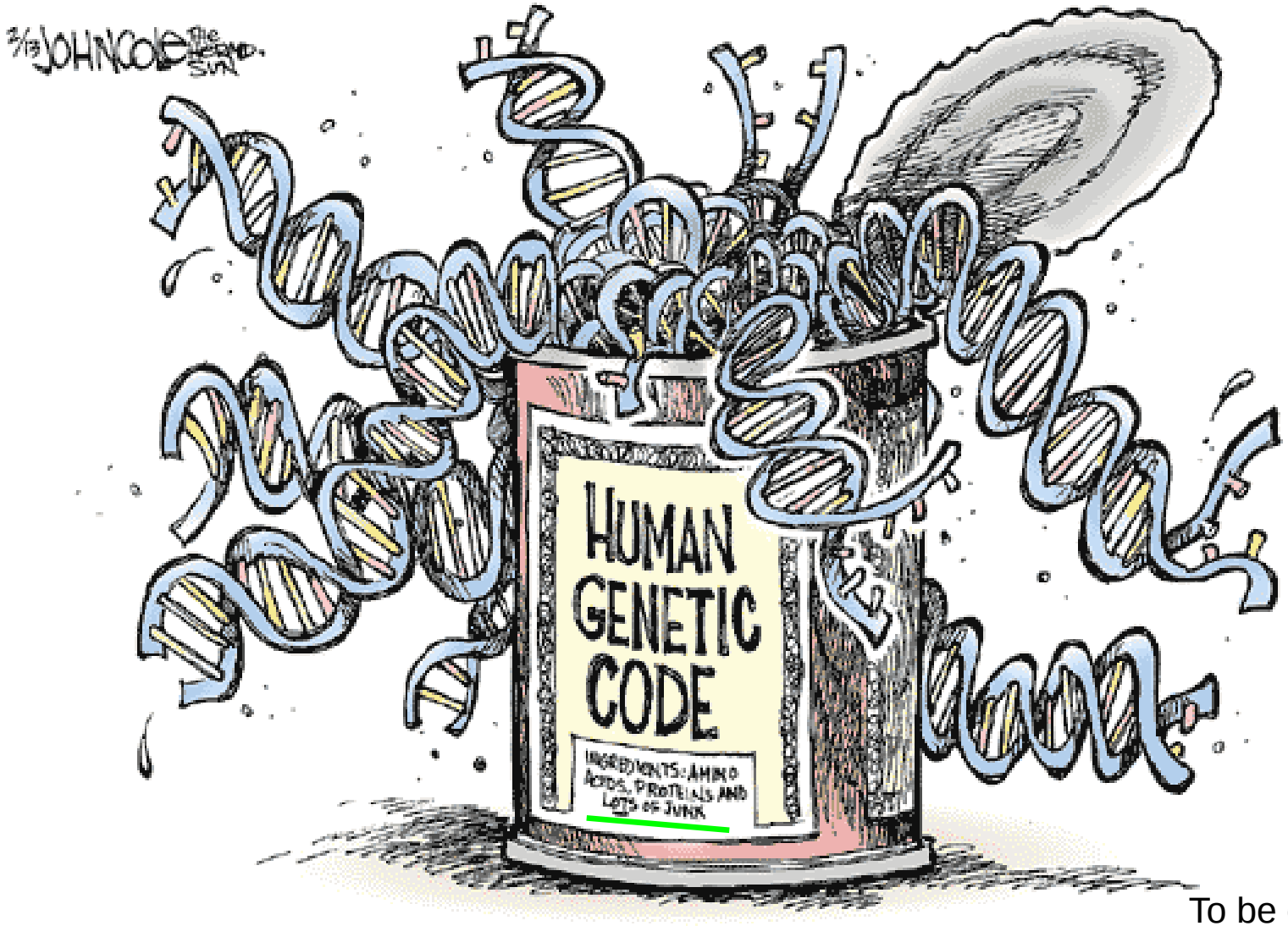
Actually almost done…

We've talked about transcripts and their regulation.

We're still ignoring most of the genome…

Type	# in genome	% of genome
genes	20,000	2%
ncRNA	20,000	2%
cis elements	1,000,000	>10%





To be continued