

# Regulatory Genomics

**Charles Blatti**

*Research Scientist*

Based on the lecture of

**Saurabh Sinha**

*Professor*

*Biomedical Engineering*

*Georgia Tech*



**National Center for  
Supercomputing Applications**

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

# The Importance of Gene Regulation



*Image Credit: Nick Youngson / Alpha Stock Images*

# DNA, RNA, Proteins

Gene: a piece of DNA, has the “code” to make a protein

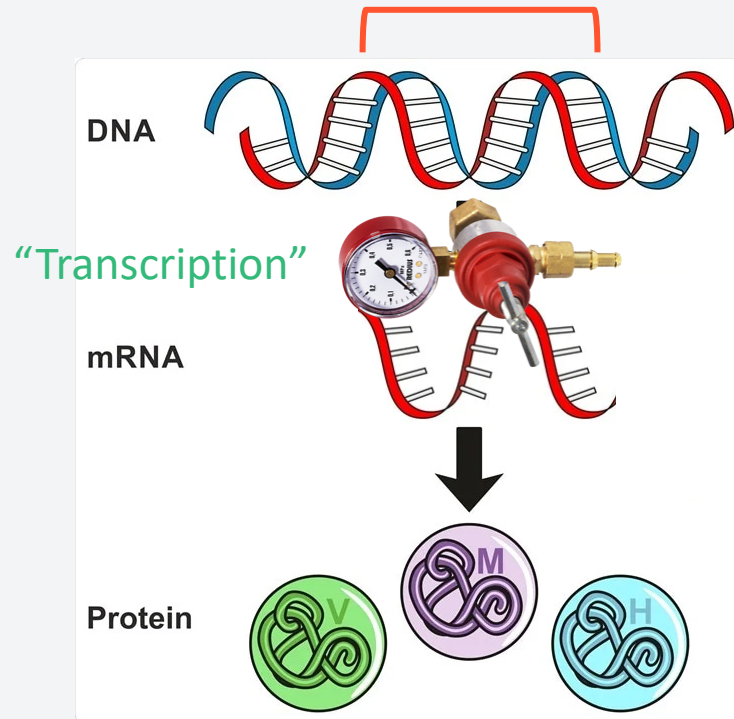


Image Credit: udaix / Shutterstock.com

DNA: a long sequence of nucleotides (a,c,g,t)

## GENE EXPRESSION

mRNA: a physical “copy” of gene

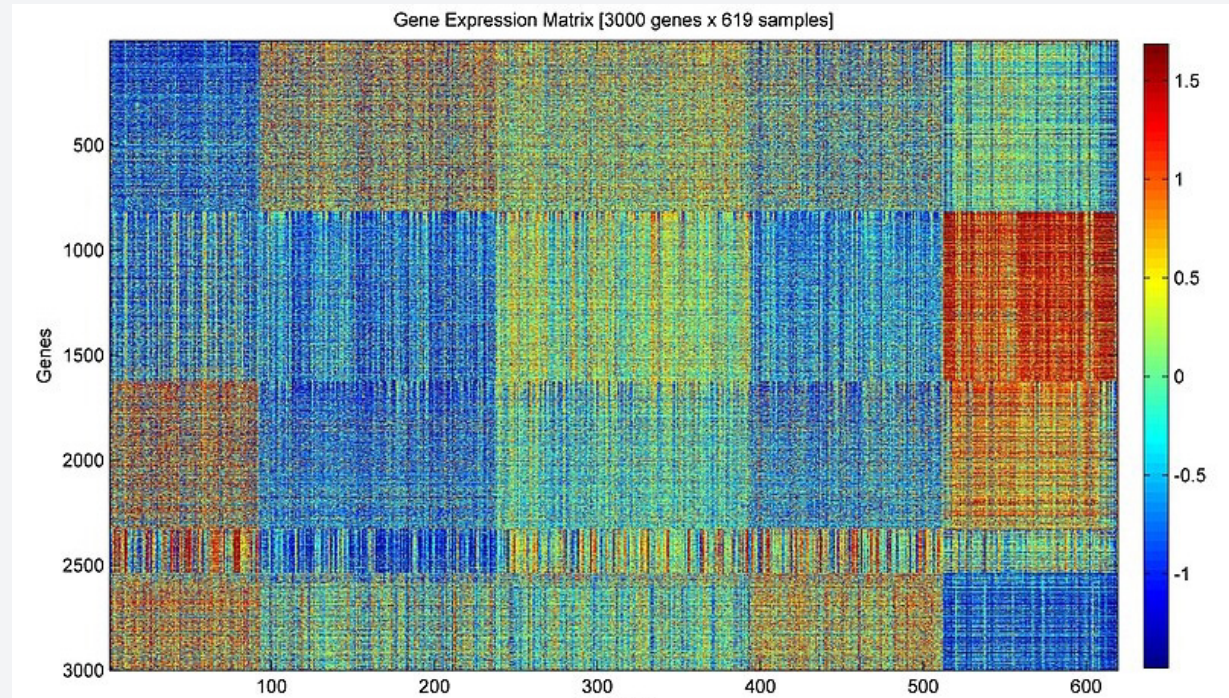
## CAN BE REGULATED

protein: molecule with important functions in cell



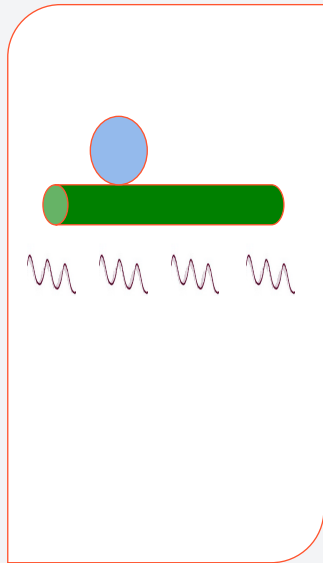
# Gene Regulation

- Gene regulation is the process of turning genes on and off.
- Gene regulation ensures that the appropriate genes are expressed in the right cells at the proper times.

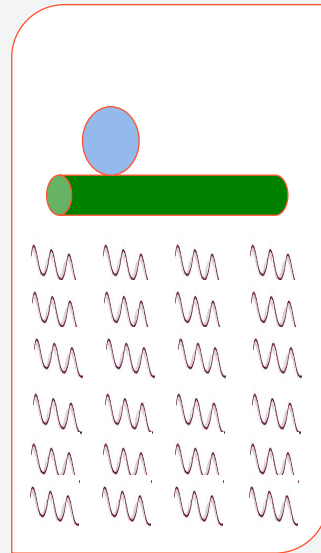


*Image Credit: Wikimedia Commons*

# Gene Regulation: fast and slow transcription



Low gene  
expression



High gene  
expression

Machinery for  
transcribing gene



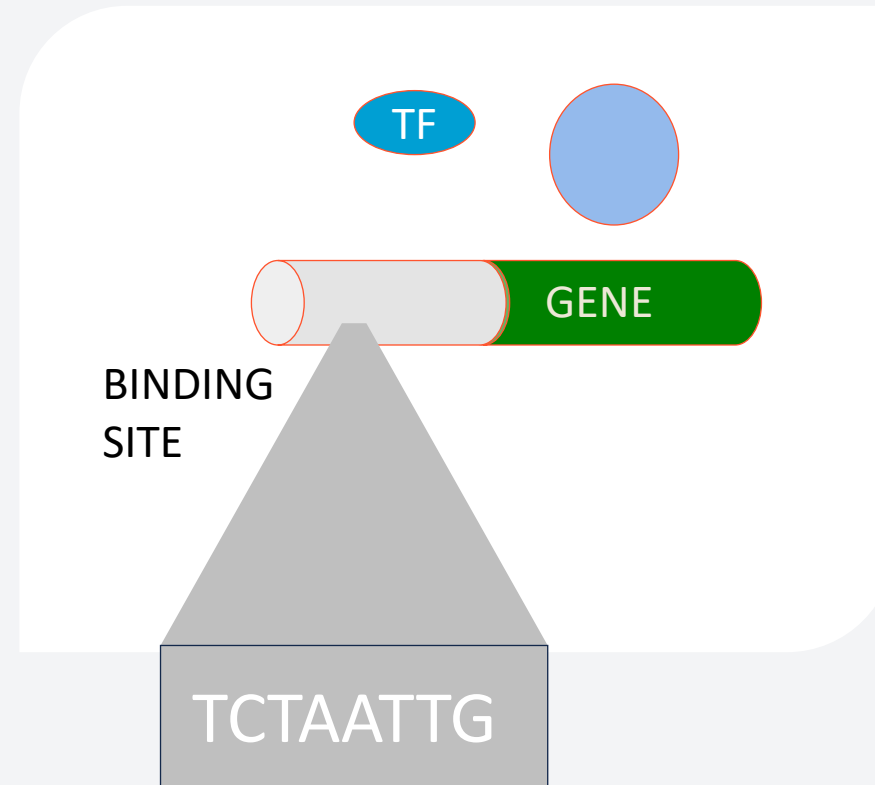
Gene



mRNA



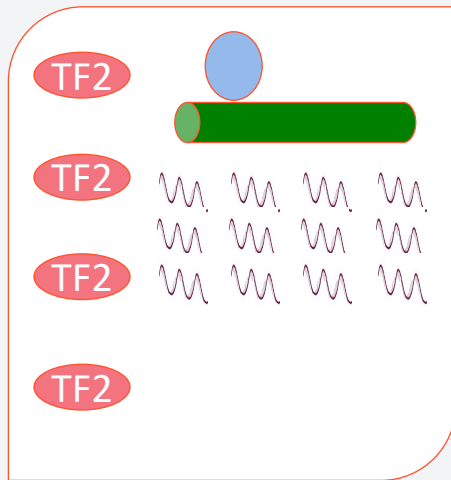
# Regulation by Proteins called Transcription Factors (TFs)



Humans have ~2000 TFs

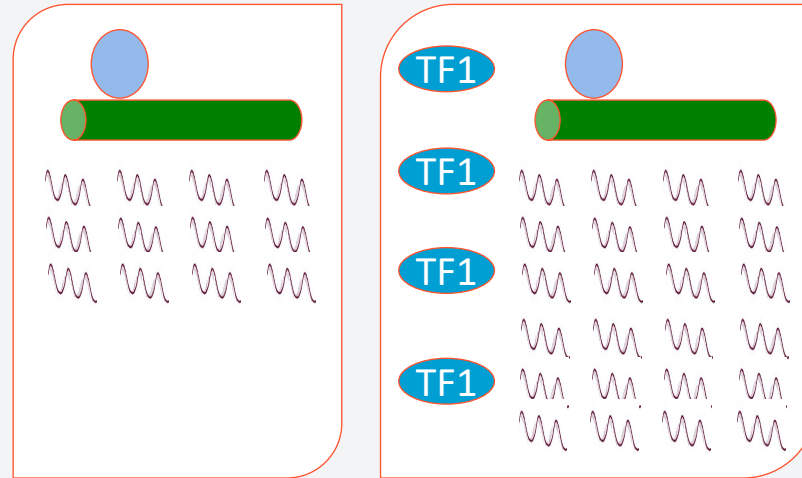
# Different cells may have different TFs

TF2 represses gene.  
Low gene expression



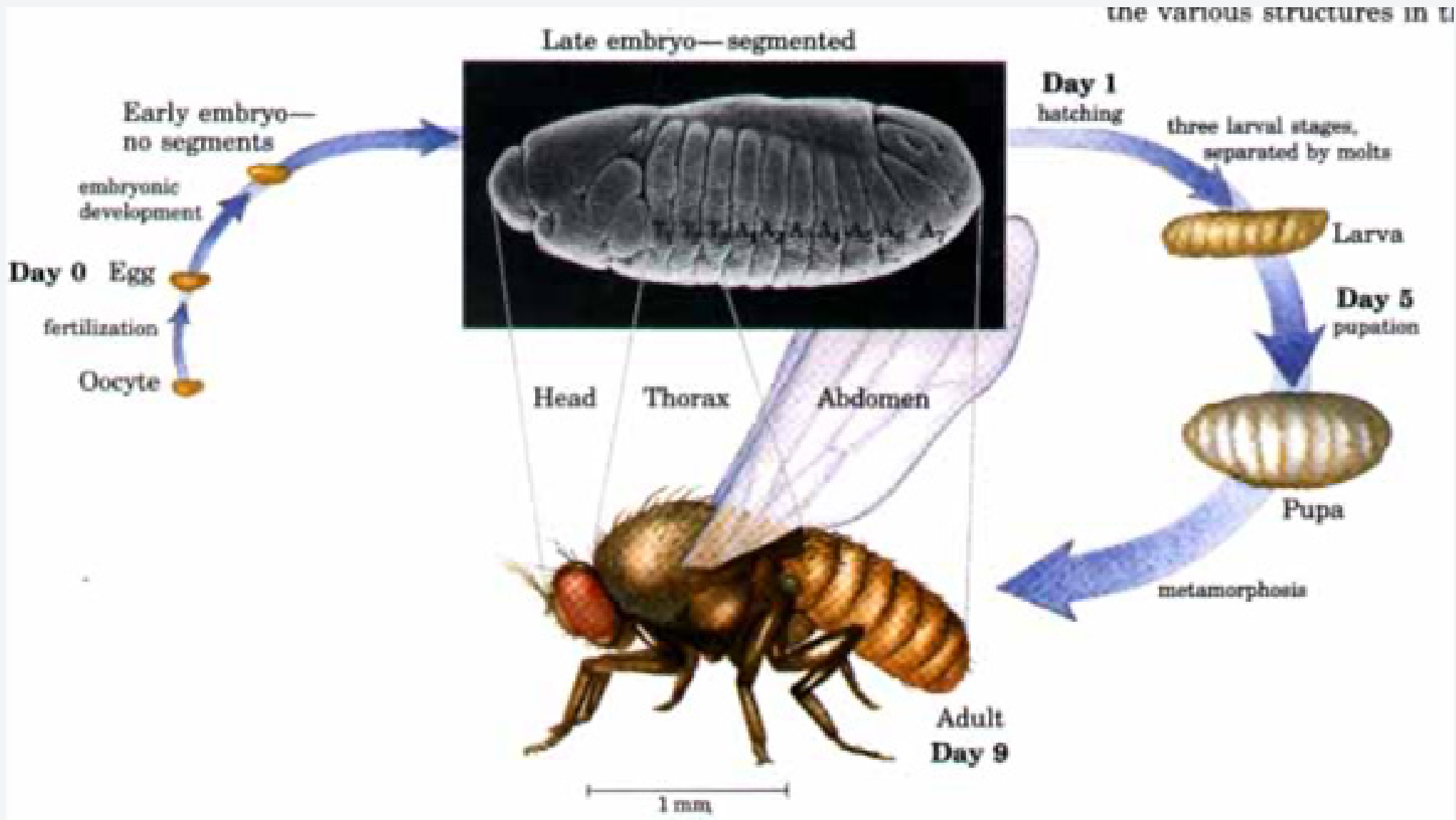
Liver cell

TF1 activates gene.  
High gene expression



Heart cell

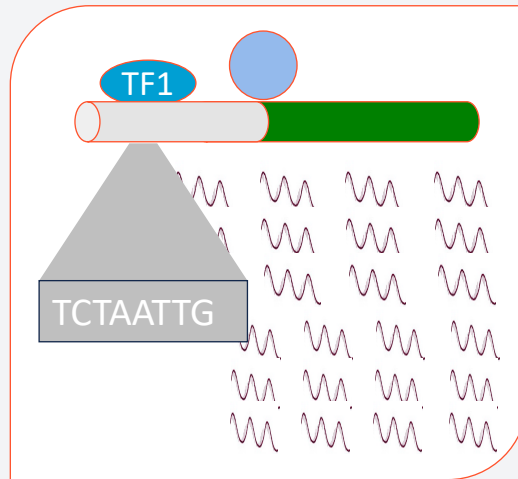
# Gene regulation builds bodies



line various structures in U

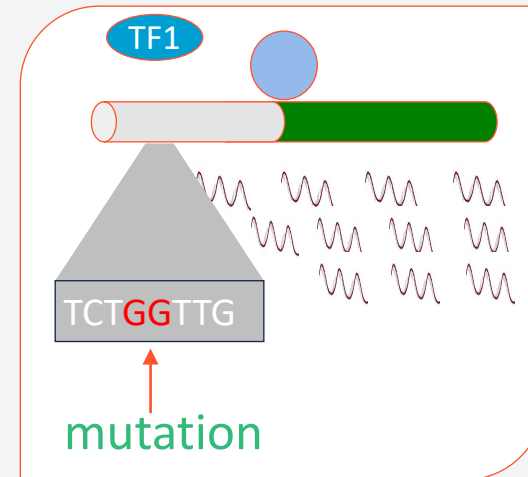
# Different cells occasionally have different DNA

TF1 binds DNA and activates gene.  
High gene expression



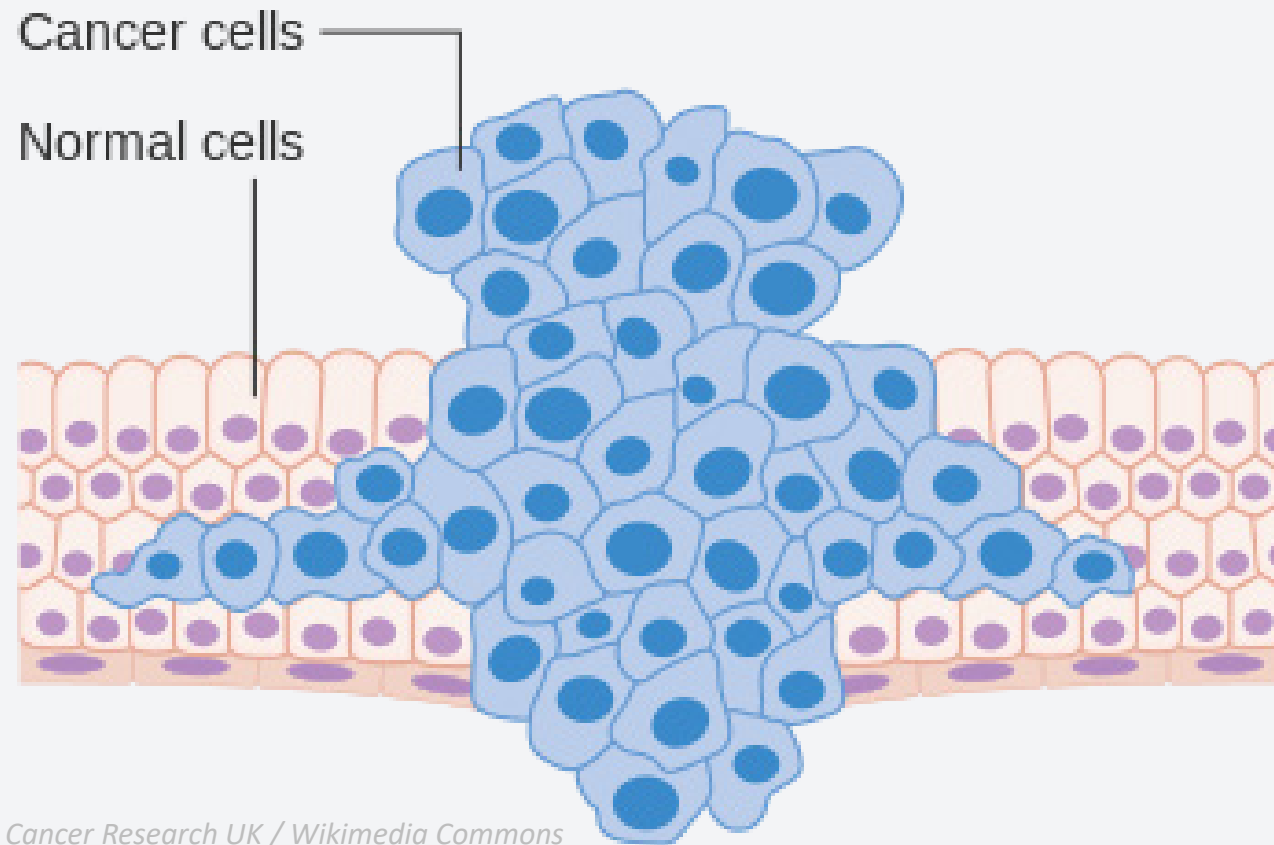
Normal cell

TF1 cannot bind DNA, doesn't activate gene.  
Low gene expression



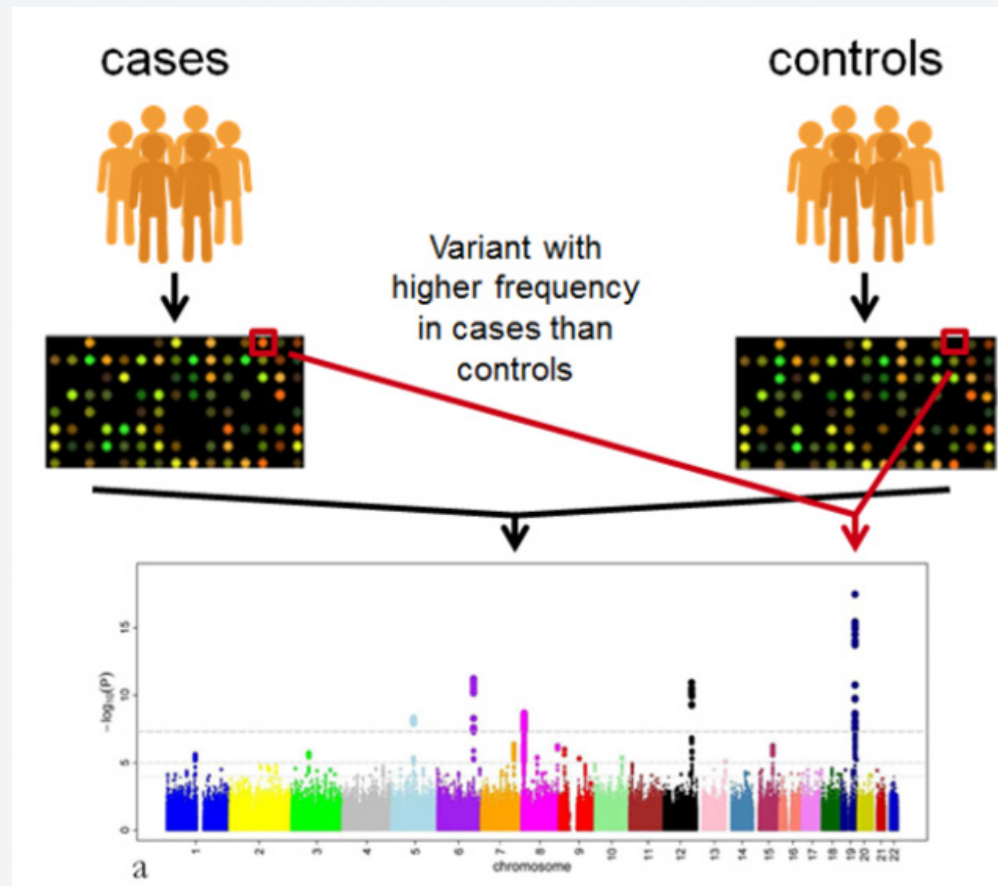
Tumor cell

# Gene Regulation is disrupted in cancer



Source: Cancer Research UK / Wikimedia Commons

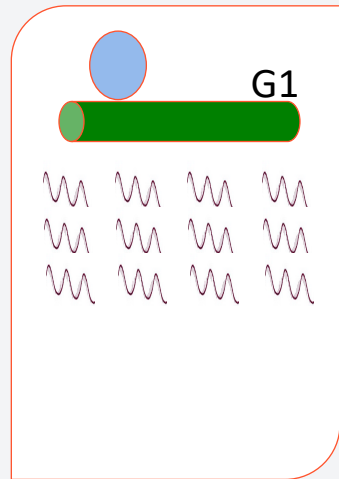
# Most disease-related mutations are outside of genes



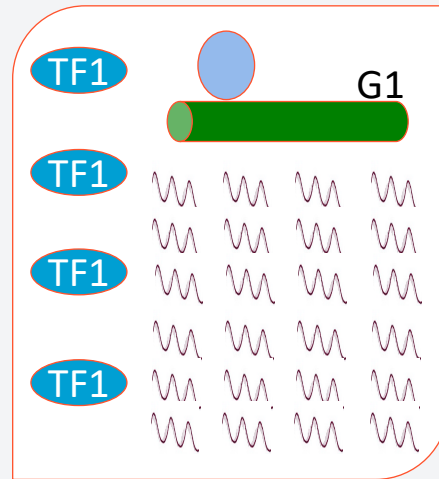
(impact gene regulation)



# Gene Regulatory Networks: TF-gene relationships



Healthy  
sample



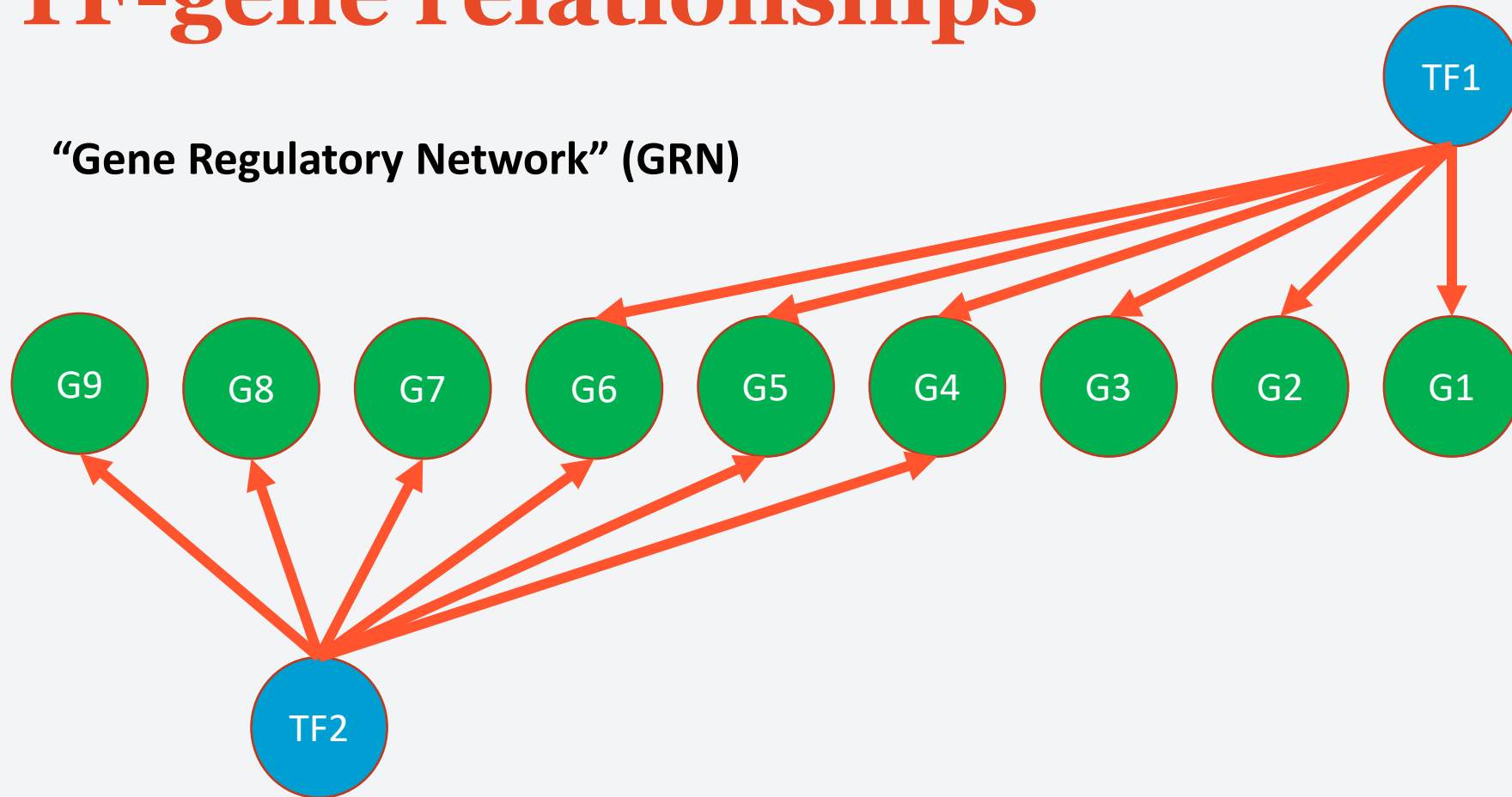
Tumor  
sample

TF1 activates gene.  
High gene expression



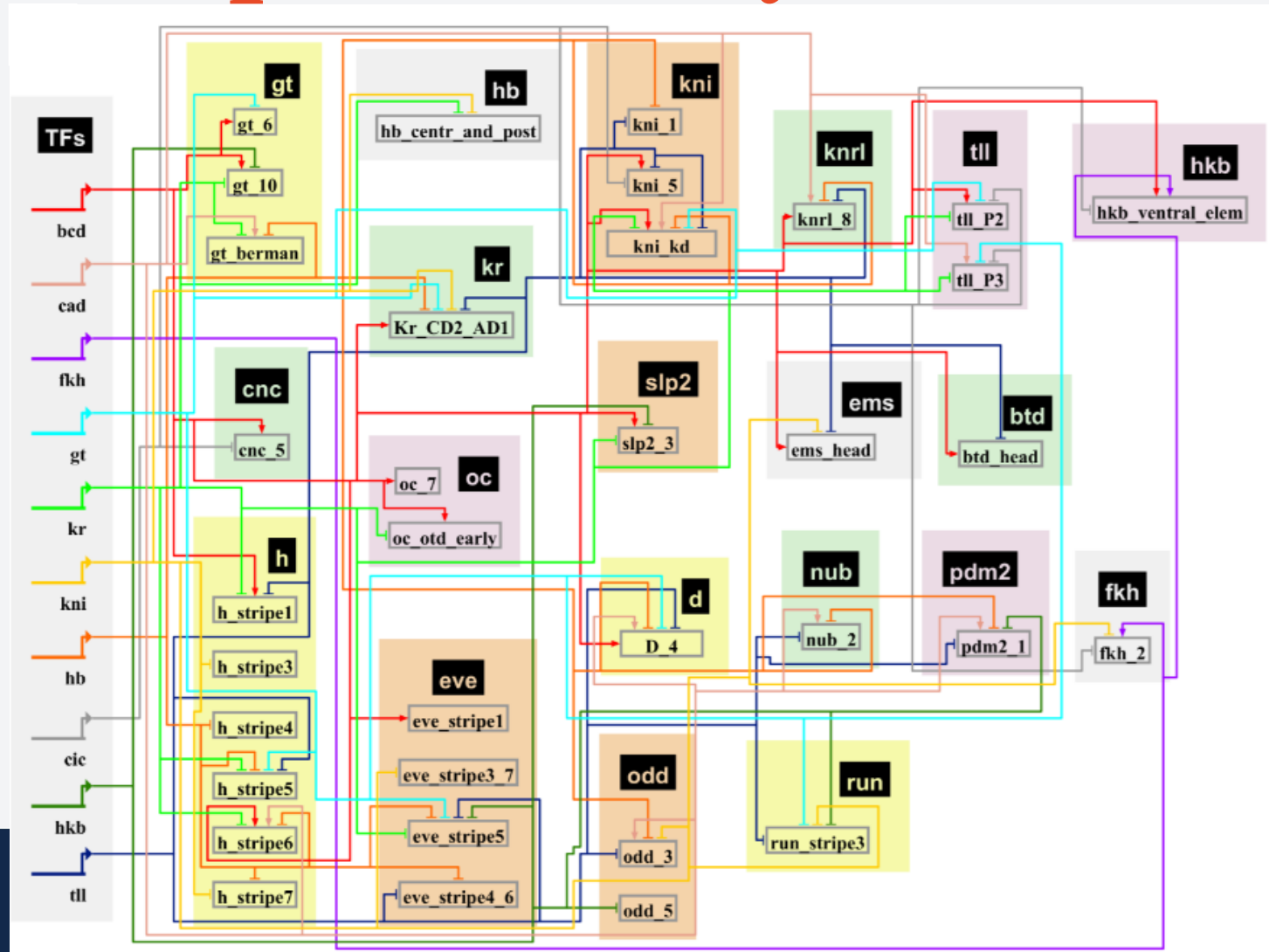
# Gene Regulatory Networks: TF-gene relationships

“Gene Regulatory Network” (GRN)









# GRNs can be reconstructed computationally



## PLOS BIOLOGY

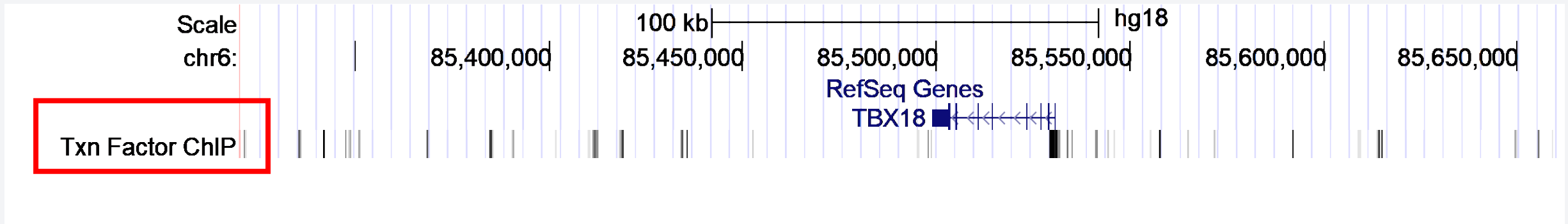
### Quantitative Analysis of the *Drosophila* Segmentation Regulatory Network Using Pattern Generating Potentials

Majid Kazemian , Charles Blatti , Adam Richards, Michael McCutchan, Noriko Wakabayashi-Ito, Ann S. Hammonds, Susan E. Celniker, Sudhir Kumar, Scot A. Wolfe, Michael H. Brodsky , Saurabh Sinha 

- *Goal: discover the gene regulatory network*
- *Sub-goal: discover the genes regulated by a transcription factor*



# Genome-wide assays



One experiment per cell type AND PER TF  
... tells us which TF might regulate a gene of interest

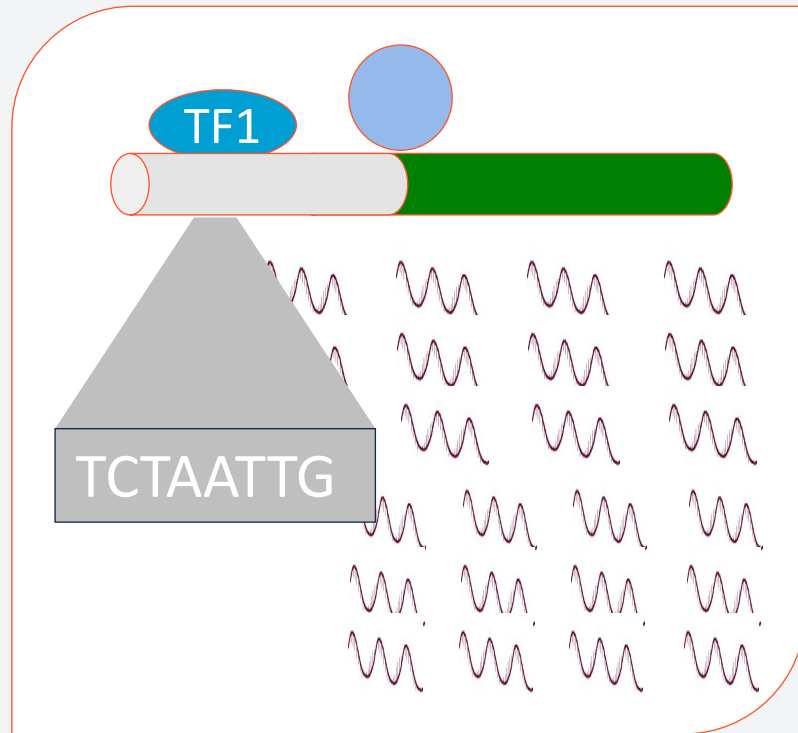
Expensive !



- *Goal: discover the gene regulatory network*
- *Sub-goal: discover the genes regulated by a transcription factor*
- *... by DNA sequence analysis*



# The regulatory network is encoded in the DNA



**It should be possible to predict where transcription factors bind, by reading the DNA sequence**

# Motifs and DNA sequence analysis

GCTCCTCAAG **GCTTGTTTACA** TAATCACCGT  
 TCCGCTTCTT **GCTTGTTTACA** GTTAGGAATG  
 GTTAGATGTG **GCTTGTTTACA** TAAGCGATAA  
 GCTCGCGTCC **GCTTGTTTACA** CAAACGCCCA  
 TTTAGTGGCG **GCTTGTTTACA** GGGTTGCAGC  
 CATGCCGATC **GCTTGTTTACA** CACAAGCATG  
 AGCAGTTTCA **GTTTGTTTACA** CACCGGAGTC  
 CACTTTAATG **GTTTGTTTACA** TCATAAATAA  
 TGCTGCCTTG **GTTTGTTTACA** CCAGAATCGA  
 TTCGCATTTT **GCTTGTTTACT** TGCGTCAAAC  
 CGGCAACCAC **GCTTGTTTATA** ACATACAAAC  
 CACTCACGCT **GTTTGTTTACT** TTCGATAAAG  
 TAGTCACTCT **CCTTGTTTACA** TTTTGAATGT  
 TGGAATTTTC **CCTTGTTTACA** GCTATTATGC  
 AAGAGTGCCT **CCTTGTTTACA** CTAACTATTT  
 CTATTTTTAA **GTTTGTTTATA** GGAAATAGGT  
 CTGCACATCT **GTTTGTTTATA** TTGTAATTGT  
 CGCCTTCCTT **CTTTGTTTACA** TTCGTTCTTT  
 CAACTTCTGC **TCTTGTTTACA** CTGACGAATG  
 GATTTGTTTCG **GGTTGTTTACT** GGGATCTCGA  
 ATCGCTTCTG **GCTTGTTTATT** TCTGGAGTAG  
 TCGTCGAGCT **GCTTGTTTATT** TGCTGCCGTC  
 CAGCGCCATT **GTTTGTTTACG** CAACTTTGAT  
 GTTTTCATGT **GCTTGTTTATT** TTTTCTGTGG  
 TCCGTAGCCA **GCTTGTTTATT** TGTTTGCCTT  
 GTTGCCGGCG **GTTTGTTTACG** GACCCGCGGC  
 TTTTGTCTGGT **TTTTGTTTACA** ATTCGCTTCC

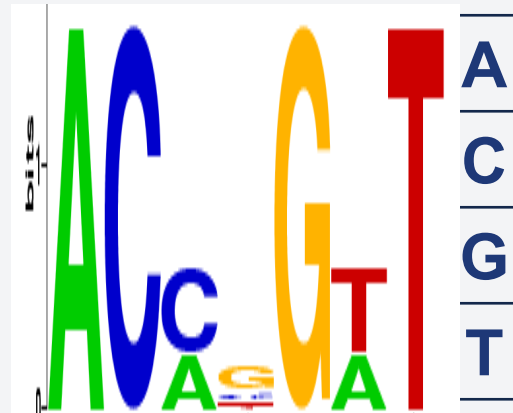
# Finding TF targets

- Step 1. Determine the binding specificity of a TF
- Step 2. Find motif matches in DNA
- Step 3. Designate nearby genes as TF targets



# Step 1. Determine the binding specificity of a TF

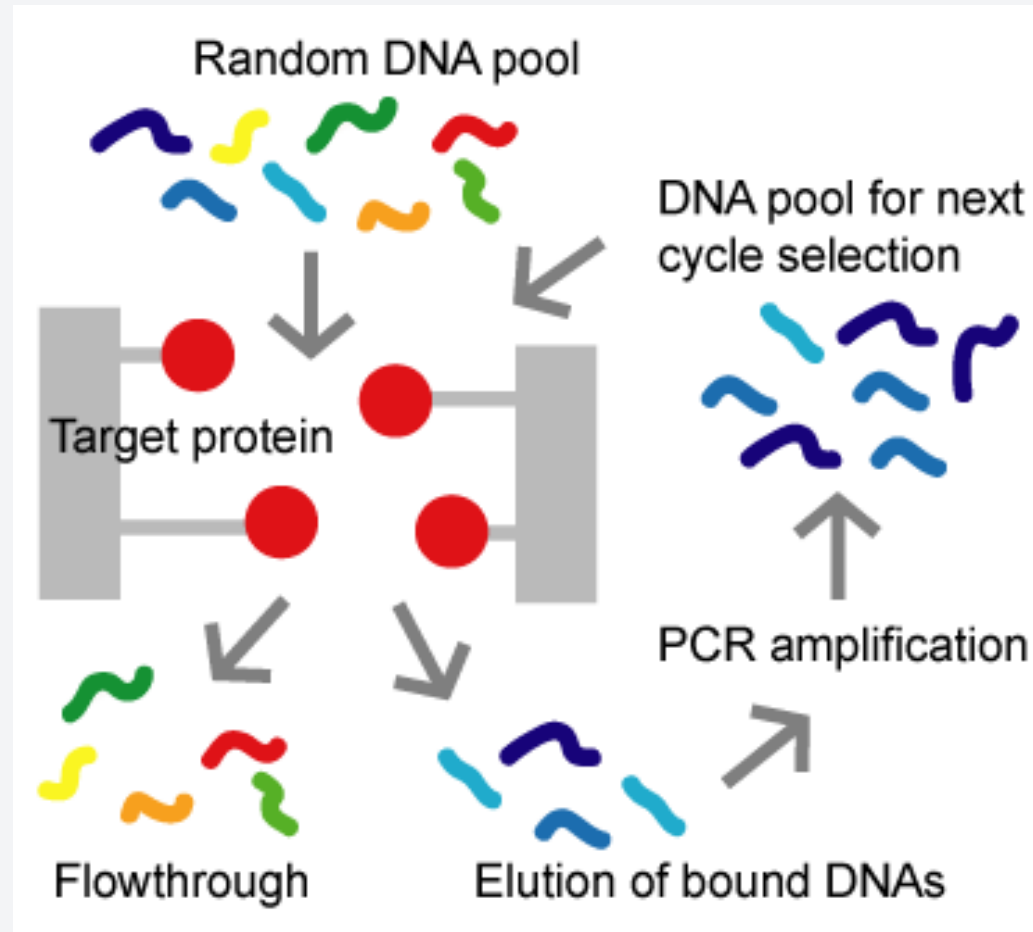
ACCCGTT  
 ACCGGTT  
 ACAGGAT  
 ACCGGTT  
 ACATGAT



“MOTIF”

# How?

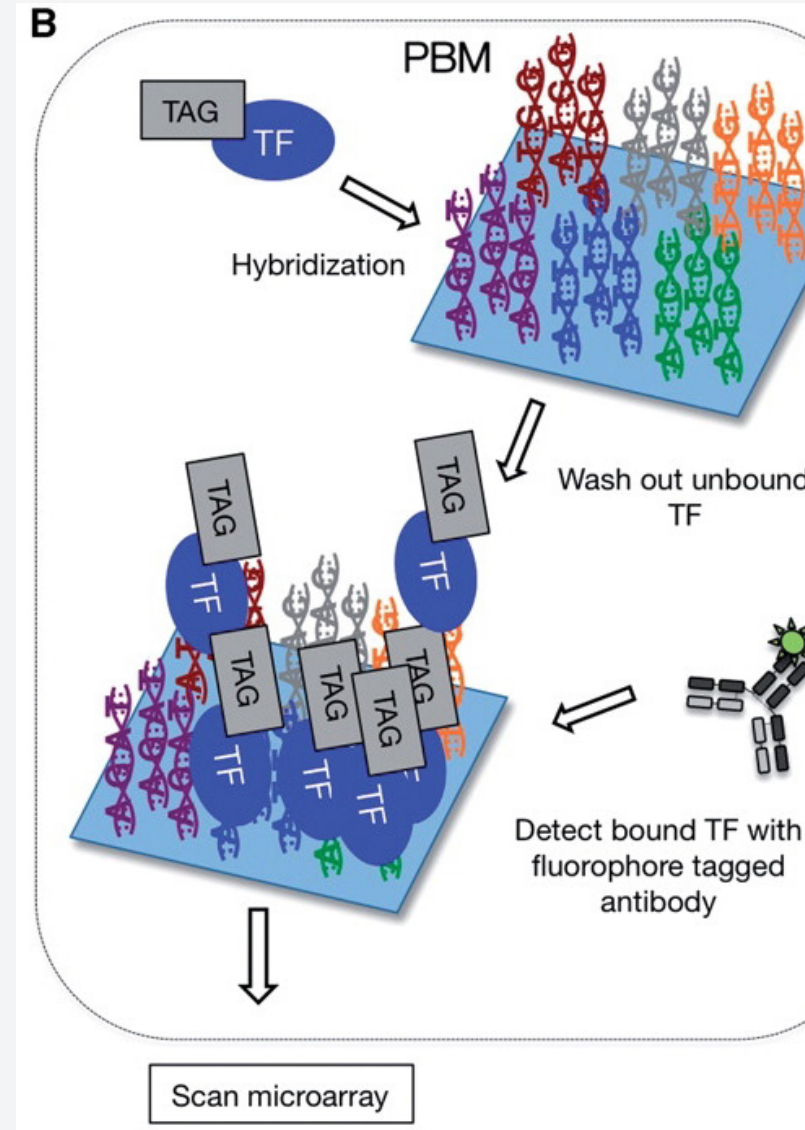
- SELEX



TAACCCG TTC  
 GTACCGG TTG  
 ACACAGG ATT  
 AACCGG TTA  
 GGACATG AT

# How?

- Protein binding microarrays



TAACCCG TTC  
 GTACCGG TTG  
 ACACAGG ATT  
 AACCGG TTA  
 GGACATG AT

Total 1964 profiles

Display  profiles

Filter:

<input type="checkbox"/>	ID ▲	Name ▼	Species ▼	Class ▼	Family ▼	Sequence logo
<input type="checkbox"/>	MA0001.1	AGL3	Arabidopsis thaliana	MADS box factors	MADS	
<input type="checkbox"/>	MA0001.2	AGL3	Arabidopsis thaliana	MADS box factors		
<input type="checkbox"/>	MA0002.1	RUNX1	Homo sapiens	Runt domain factors	Runt-related factors	
<input type="checkbox"/>	MA0002.2	RUNX1	Mus musculus	Runt domain factors	Runt-related factors	
<input type="checkbox"/>	MA0003.1	TFAP2A	Homo sapiens	Basic helix-span-helix factors (bHSH)	AP-2	
<input type="checkbox"/>	MA0003.2	TFAP2A	Homo sapiens	Basic helix-span-helix factors (bHSH)	AP-2	

# Motif Databases

- JASPAR:
- <https://jaspar2020.genereg.net/>

# Motif Databases

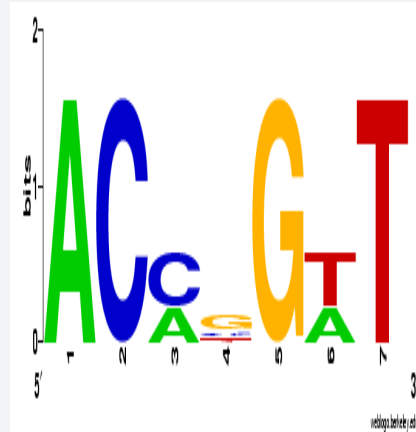
- TRANSFAC <https://genexplain.com/transfac/>
  - Public version and License version
- Cis-BP <http://cisbp.cabr.utoronto.ca/>
  - Experimentally determined as well as computationally inferred motifs
- Hocomoco: <https://hocomoco11.autosome.org/>
  - Human and mouse motifs
- UniProbe: <http://thebrain.bwh.harvard.edu/uniprobe/>
  - variety of organisms, mostly mouse and human
- Fly Factor Survey: <https://pgfe.umassmed.edu/ffs/>
  - Drosophila specific



## Step 2. Finding motif matches in DNA

- Basic idea:

Motif:



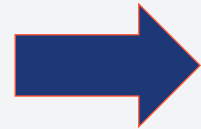
Match: ACCGGTT  
 Apprx. Match: ACACGTT

- To score a single site  $s$  for match to a motif  $W$ , we use

$$\Pr(s|W)$$

# What is $\Pr(s | W)$ ?

5	0	2	0	0	2	0	<b>A</b>
0	5	3	1	0	0	0	<b>C</b>
0	0	0	3	5	0	0	<b>G</b>
0	0	0	1	0	3	5	<b>T</b>



1	0	0.4	0	0	0.4	0	<b>A</b>
0	1	0.6	0.2	0	0	0	<b>C</b>
0	0	0	0.6	1	0	0	<b>G</b>
0	0	0	0.2	0	0.6	1	<b>T</b>

Now, say  $s = \text{ACCGGTT}$  (consensus)

$$\Pr(s | W) = 1 \times 1 \times 0.6 \times 0.6 \times 1 \times 0.6 \times 1 = 0.216.$$

Then, say  $s = \text{ACACGTT}$  (two mismatches from consensus)

$$\Pr(s | W) = 1 \times 1 \times 0.4 \times 0.2 \times 1 \times 0.6 \times 1 = 0.048.$$

# Scoring motif matches with “LLR”

- $\Pr(s | W)$  is the key idea.
- However, some statistical massaging is done on this.
- Given a motif  $W$ , background nucleotide frequencies  $W_b$  and a site  $s$ ,
- LLR score of  $s$  =

$$\log \frac{\Pr(s|W)}{\Pr(s|W_b)}$$

- Good scores  $> 0$ . Bad scores  $\lesssim 0$ .



<https://meme-suite.org/meme/tools/fimo>



Version 5.0.0

FIMO scans a set of sequences for **individual matches** to each of the motifs you provide (sample output for motifs and sequences). See this [Manual](#) or this [Tutorial](#) for more information.

### Data Submission Form

Scan a set of sequences for motifs.

#### Input the motifs

Enter motifs you wish to scan with.

Upload motifs  No file chosen

#### Input the sequences

Enter sequences or select the [database](#) you want to scan for matches to motifs.

Enable tissue/cell-specific scanning

Ensembl Ab Initio Predicted Proteins

Algerian mouse

92

#### Input job details

(Optional) Enter your email address.

(Optional) Enter a job description.

#### Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.



Grant, Bailey, Noble; *Bioinformatics* 2011.

Version 5.0.0

Please send comments and questions to: [meme-suite@uw.edu](mailto:meme-suite@uw.edu)

Powered by Opal

[Home](#) [Documentation](#) [Downloads](#) [Authors](#) [Citing](#)

# FIMO program

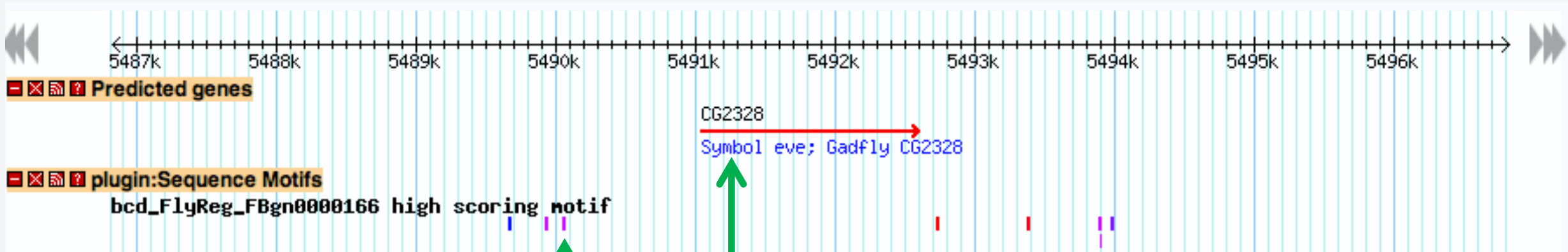
- Takes motif  $W$ , background  $W_b$  and a sequence  $S$ .
- Scans every site  $s$  in  $S$  and computes its LLR score.
- Uses sound statistics to deduce an appropriate (p-value) threshold on the LLR score. All sites above threshold are predicted as binding sites.

# Finding TF targets

- Step 1. Determine the binding specificity of a TF
- Step 2. Find motif matches in DNA
- Step 3. Designate nearby genes as TF targets



# Step 3: Designating genes as targets



Predicted binding sites for motif of TF called “bcd”

Designate this gene as a target of the TF



*Sub-goal: discover the genes regulated by a transcription factor ... by DNA sequence analysis*

# Computational motif discovery



*Image Credit: Nick Youngson / Alpha Stock Images*

# Why?

- We assumed that we have experimental characterization of a transcription factor's binding specificity (motif)
- What if we don't?
- There's a couple of options ...



# Option 1

- Suppose a TF regulates five different genes
- Each of the five genes should have binding sites for TF in their promoter region



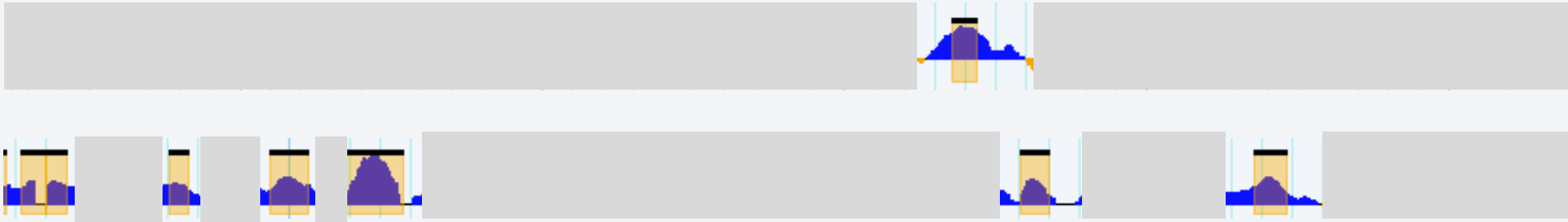
# Option 1

- Now suppose we are given the promoter regions of the five genes
  - G1, G2, ... G5
- Can we find binding sites of a TF, without knowing them *a priori* ?
- This is the computational ***motif discovery problem***
- Find a motif that represents binding sites of an unknown TF



# Option 2

- Suppose we have ChIP-Seq data on binding locations of a transcription factor.



- Collect sequences at the peaks
- Computationally find the motif from these sequences
- This is another version of the motif discovery problem

# Motif discovery algorithms

- **Version 1:** Given promoter regions of co-regulated genes, find the motif
- **Version 2:** Given bound sequences (ChIP peaks) of a transcription factor, find the motif
- **Idea:** Find a motif with many (surprisingly many) matches in the given sequences



# Motif discovery algorithms

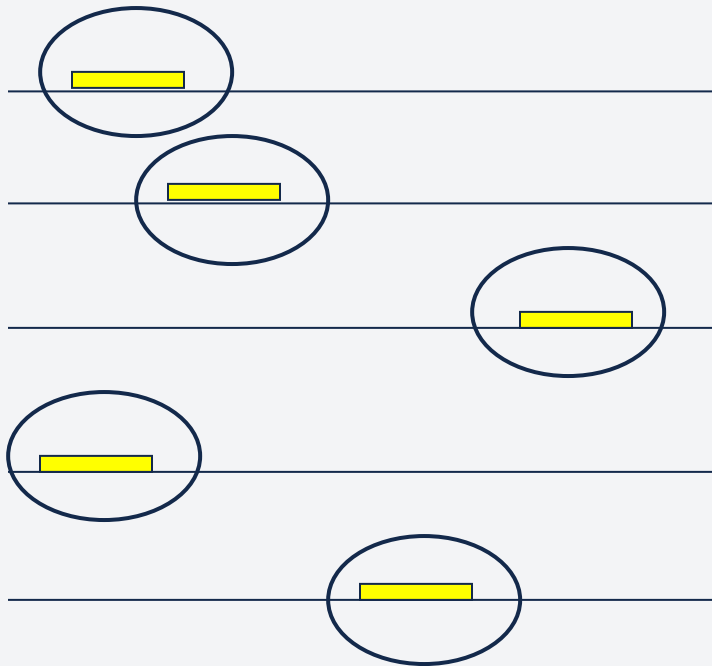
- Gibbs sampling (**MCMC**) : Lawrence et al. 1993
- **MEME** (Expectation-Maximization) : Bailey & Elkan 94.  
(Very popular, visited in today's lab.)
- **CONSENSUS** (Greedy search) : Stormo lab.
- **Priority** (Gibbs sampling, but allows for additional prior information to be incorporated): Hartemink lab.
- Many many others ...



# Examining one such algorithm



# The “CONSENSUS” algorithm



Final goal: Find a set of “substrings” (sites), one in each input sequence

Set of substrings define a motif.  
Goal: This motif should have high “information content”.

High information content means that sites are identical or similar to each other

# The “CONSENSUS” algorithm

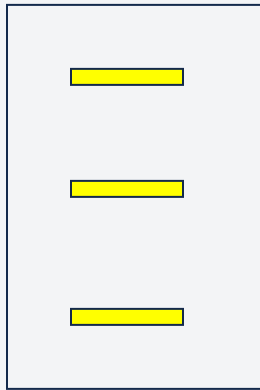


Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

# The “CONSENSUS” algorithm



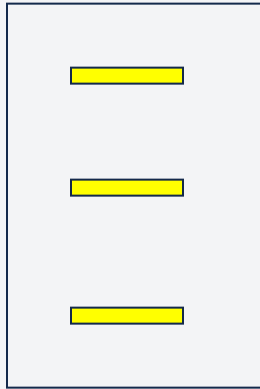
Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

# The “CONSENSUS” algorithm



Start with a substring in one input sequence

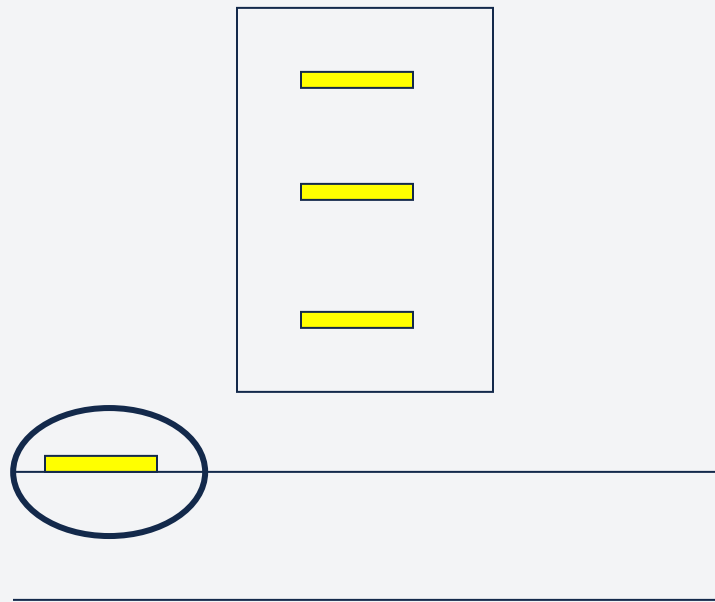
Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

Consider every substring in the next sequence, try adding it to current motif and scoring resulting motif's information content

# The “CONSENSUS” algorithm



Start with a substring in one input sequence

Build the set of substrings incrementally, adding one substring at a time

The current set of substrings.

The current motif.

Pick the best one ....


# The “CONSENSUS” algorithm



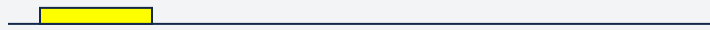
Start with a substring in one input sequence



Build the set of substrings incrementally, adding one substring at a time



The current set of substrings.



The current motif.



Pick the best one ....

... and repeat



# Summary so far

- To find genes regulated by a TF
  - Determine its motif experimentally
  - Scan genome for matches (e.g., with FIMO & the LLR score)
- Motif can also be determined computationally
  - From promoters of co-expressed genes
  - From TF-bound sequences determined by ChIP assays
  - MEME, CONSENSUS, etc.



# Further reading

- Introduction to theory of motif discovery
  - Moses & Sinha. Regulatory Motif Analysis.  
[http://www.moseslab.csb.utoronto.ca/Moses\\_Sinha\\_Bioinf\\_Tools\\_apps\\_2009.pdf](http://www.moseslab.csb.utoronto.ca/Moses_Sinha_Bioinf_Tools_apps_2009.pdf)
  - Das & Dai. A survey of DNA motif discovery algorithms.  
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2099490/pdf/1471-2105-8-S7-S21.pdf>



# Motif discovery tools

- **MEME:** <https://meme-suite.org/>
- **RSAT:** <https://rsat.eead.csic.es/plants/>

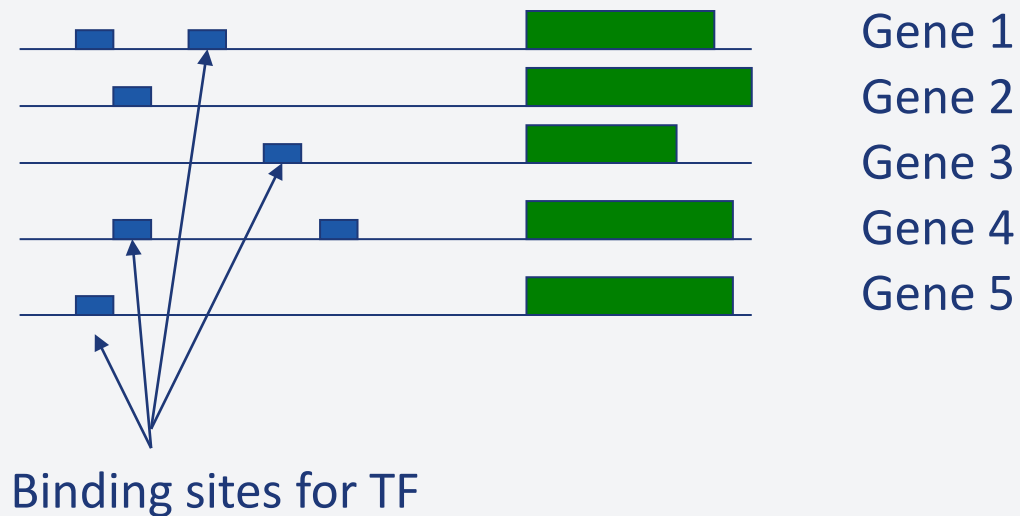
# Associating sequence analysis and expression data



*Image Credit: Nick Youngson / Alpha Stock Images*

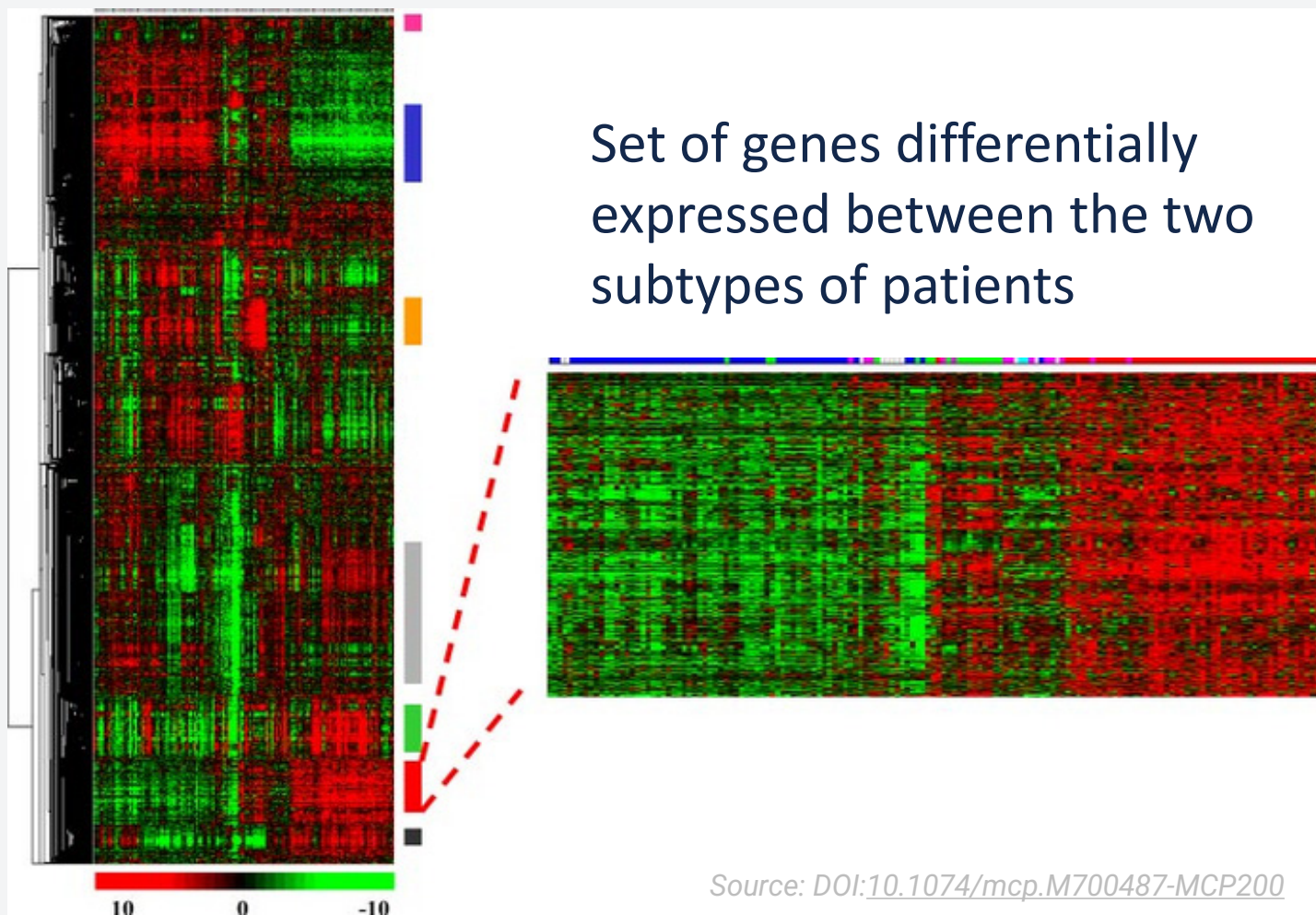
# 1. Predict regulatory targets of a TF

**Motif module:** a set of genes predicted to be regulated by a TF (motif)

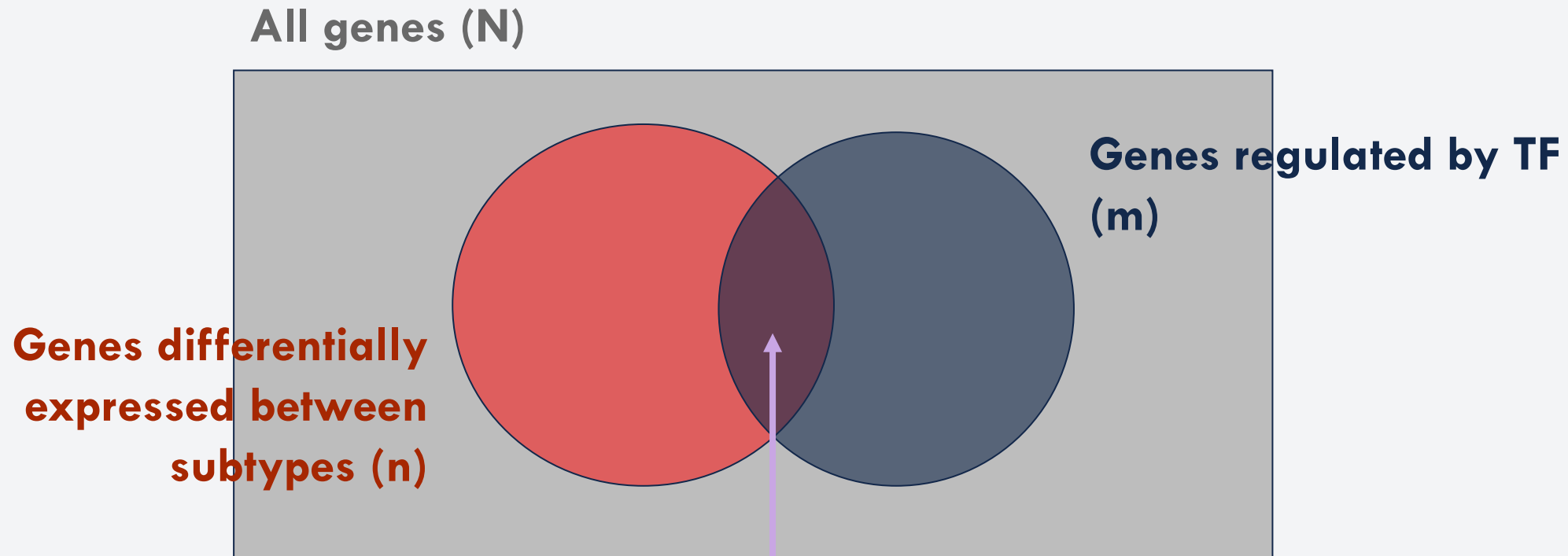


## 2. Identify dysregulated genes in phenotype of interest

Two “subtypes” of patients



# 3. Combine motif analysis and gene expression data

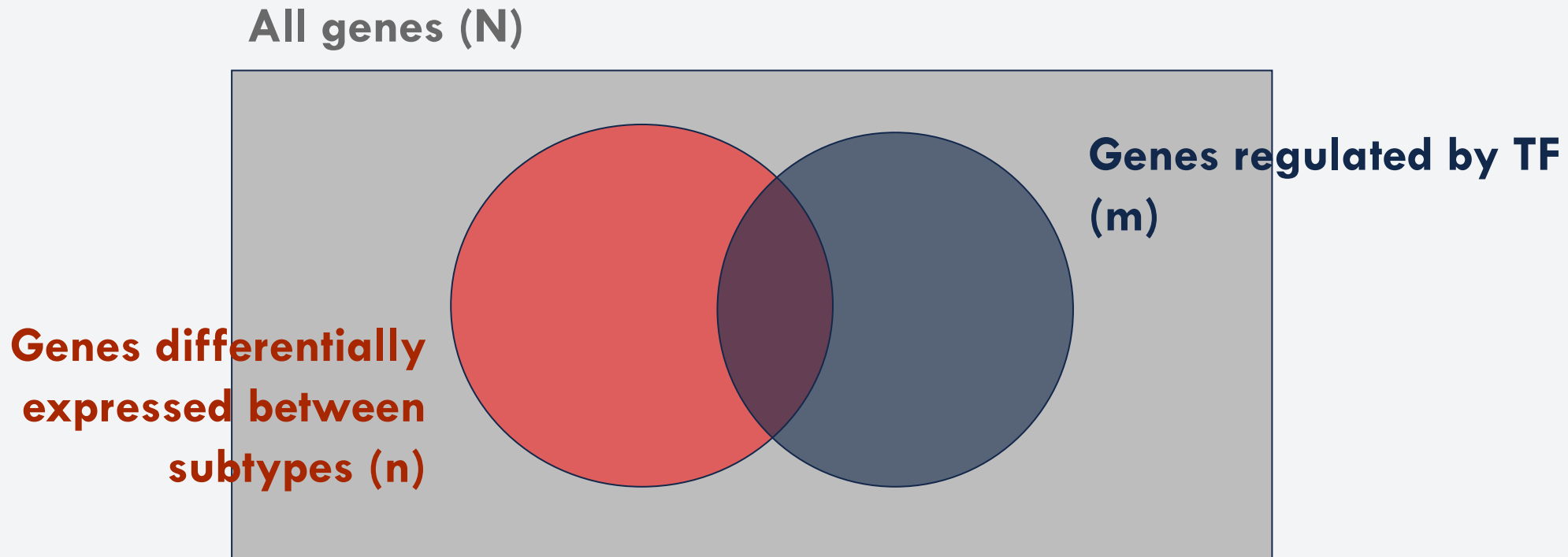


Is the intersection (size “k”) significantly large, given N, m, n?



Use Hypergeometric test to obtain “p-value”

# 3. Combine motif analysis and gene expression data



Infer TF may drive subtypes from “association” between motif and condition

# Useful tools

- **GREAT:** <http://great.stanford.edu/public/html/>
  - Input a set of genomic segments (e.g., ChIP peaks)
  - Obtain what annotations enriched in nearby genes
  - only for human and mouse
- **DAVID:** <https://davidbioinformatics.nih.gov/workspace.html>
  - Input a set of genes
  - Obtain what annotations enriched in those genes
  - Many different species



# Quick Break



# Epigenomics

- Where do TFs bind?
- Which genomic segments actively regulate gene expression?

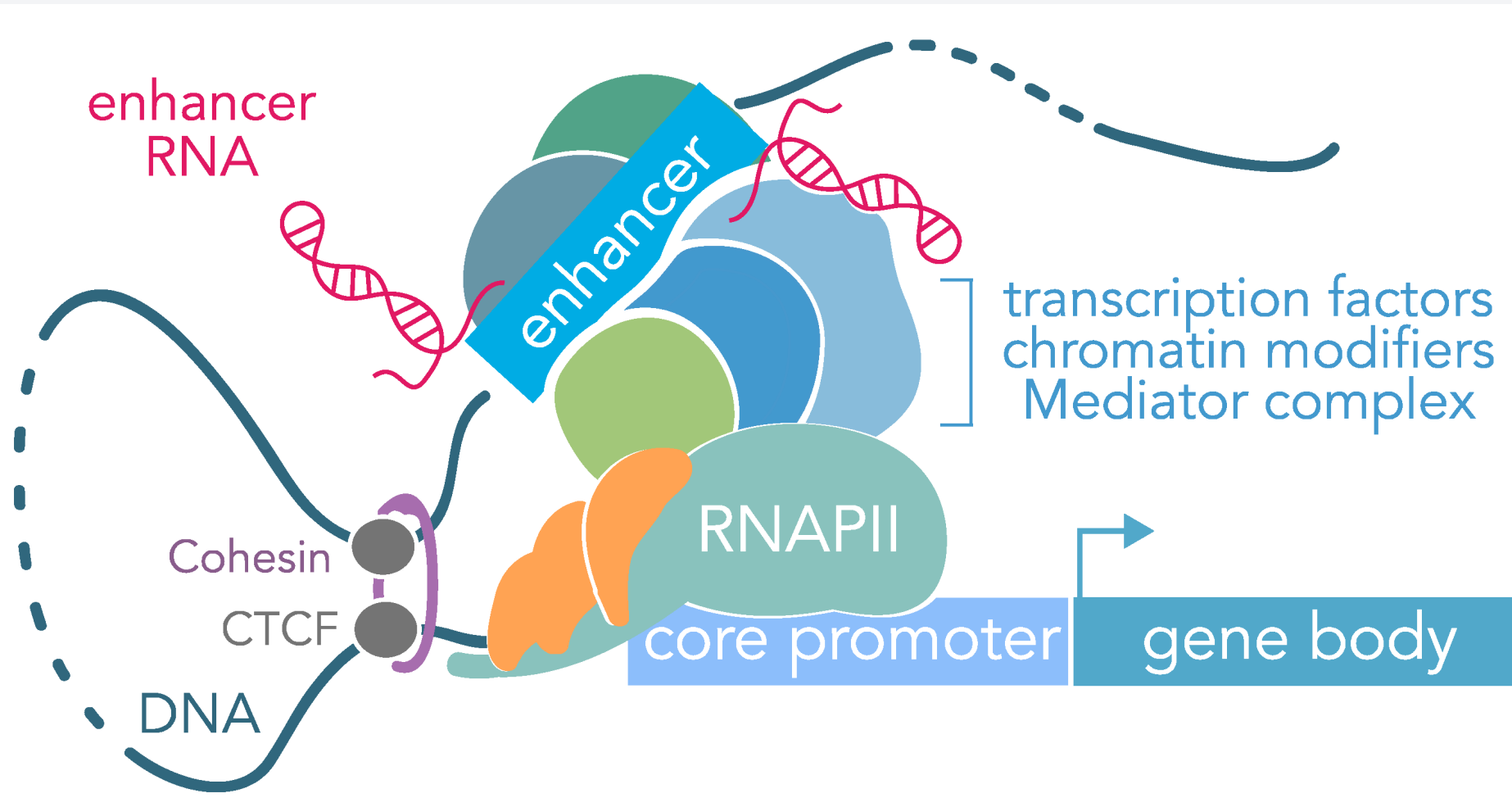


# Outline

- Decorations on the genome
- Experimental assays to profile the decorated genome
- Insights from large scale epigenomics studies

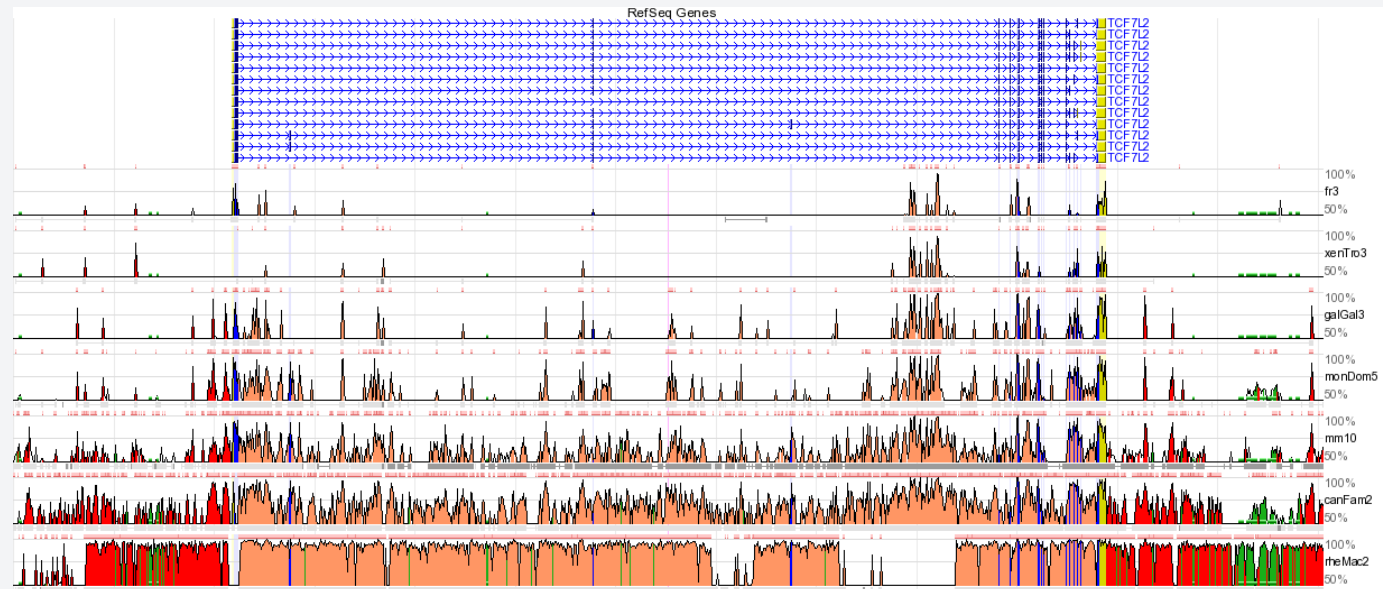


# The regulatory genome



# How to find enhancers?

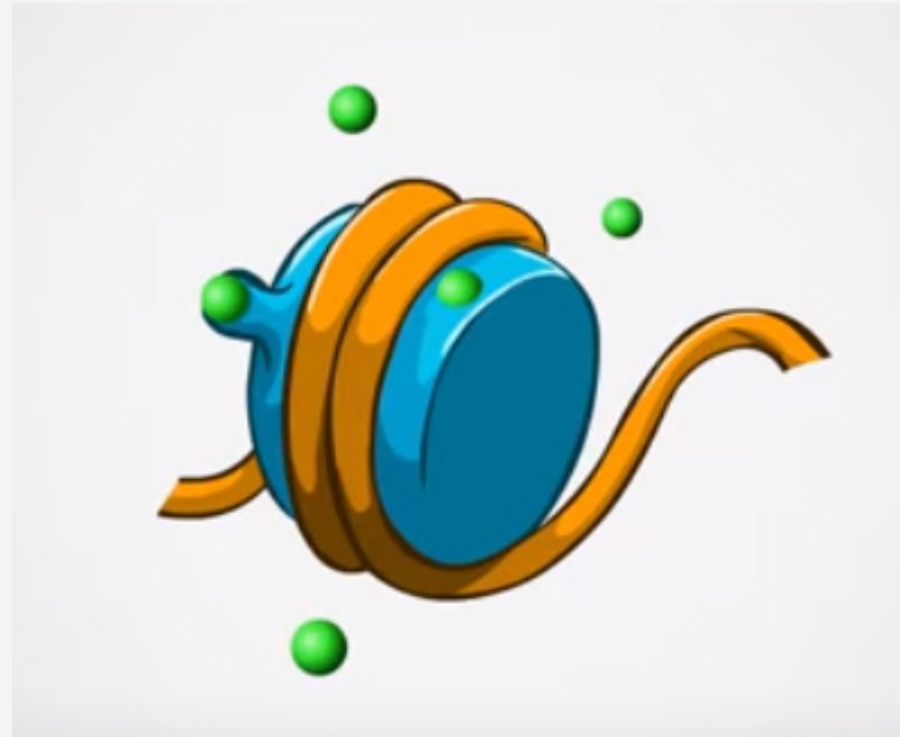
- Like finding needle in a haystack
- Evolutionary conservation is sometimes used to identify enhancers



- but not all functional elements are conserved at the level that DNA sequence alignments can detect. So how do we find regulatory elements?

**I** • More important question is: which enhancers are *active* in a particular cell type?

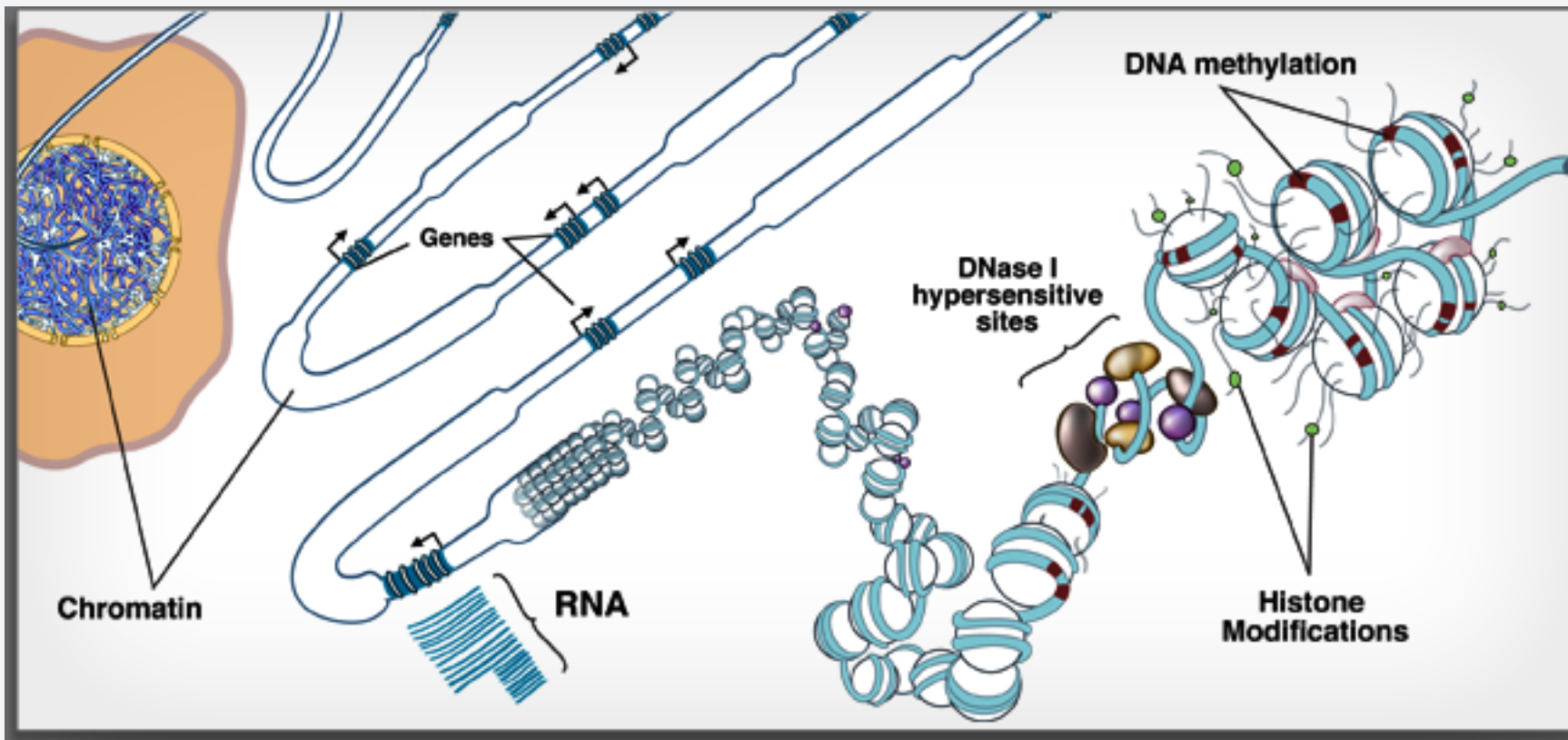
# Regulatory activity leaves its “mark” on the genome: epigenomics



*Image Credit: Ahmed.yosri / Wikimedia Commons*

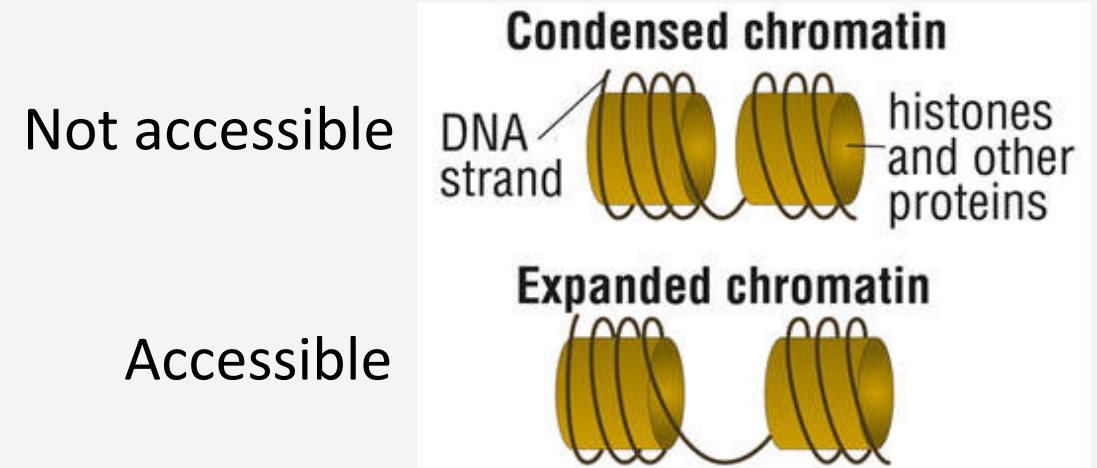
# Genomes are complex 3D structures

- Comprised of modified and unmodified DNA, RNA and many types of interacting proteins
- Most DNA is wrapped around a “**histone core**”. Such wrapped-around DNA is relatively “**inaccessible**” to other molecules such as TFs. But there are “**accessible regions**” as well, can be detected as “**DNase I hypersensitive sites**” (DHS)
- **TFs bind** to their preferred sites (especially in **accessible** regions), or not
- Histone proteins are ‘**marked**’ (like flags), or not
- CpG dinucleotides in DNA are **methylated**, or not



# Epigenomic clues into regulatory activity

- Look for **accessible regions** of DNA, that's where active regulatory elements might lie
- Also: specific **histone modifications** and **DNA methylation** mark regulatory activity
- If you know a particular **TF** that is important for regulation, look for its **binding sites**

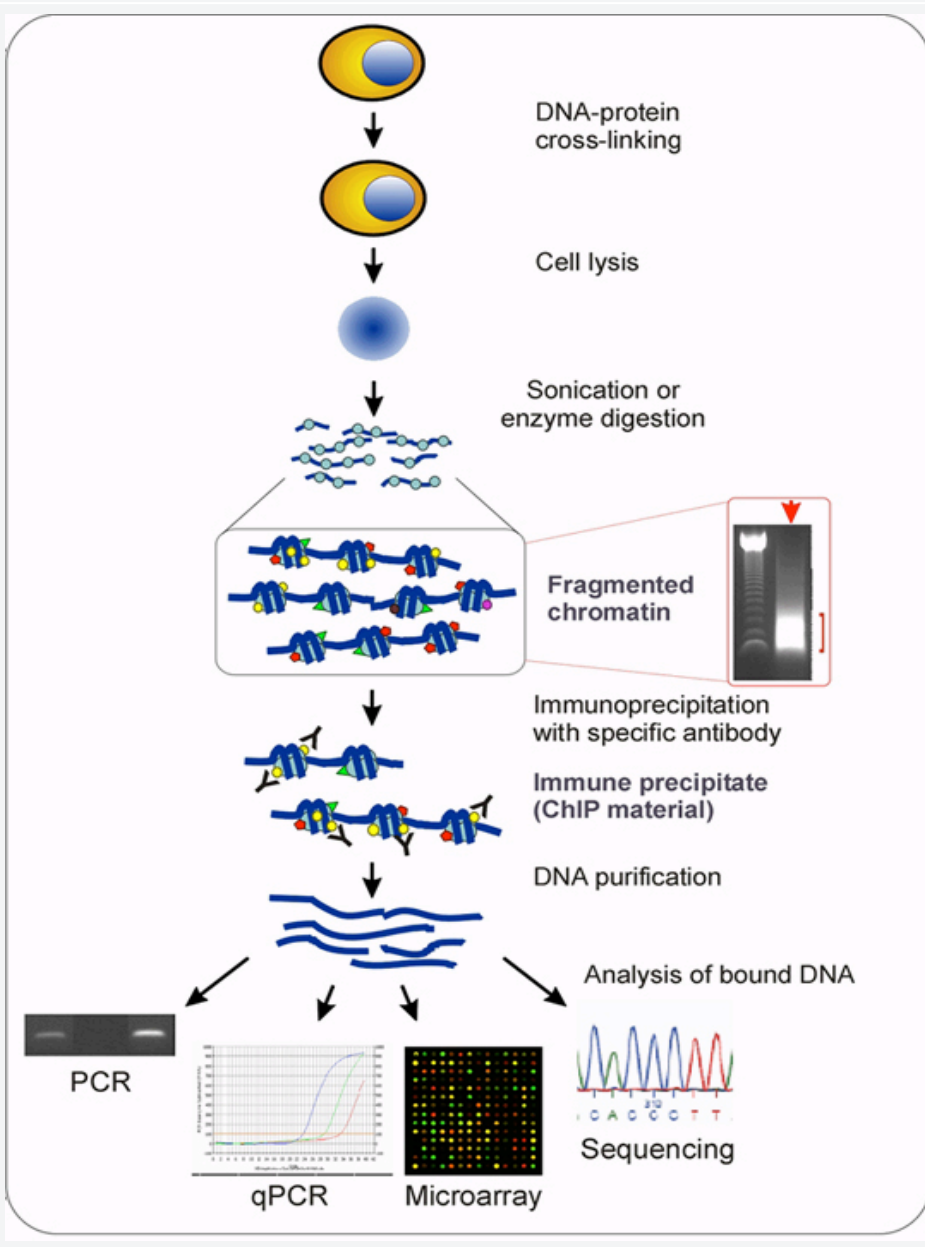


# Experimental assays



*Image Credit: Nick Youngson / Alpha Stock Images*

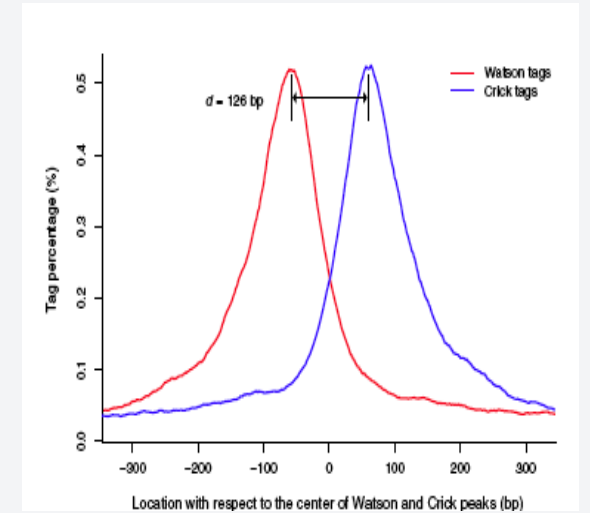
# Chromatin Immunoprecipitation (ChIP)



- Antibody to a DNA binding protein is used to “fish out” DNA bound to the protein in a living cell
  - DNA and protein are crosslinked in the cell using formaldehyde
  - Crosslinked chromatin is sheared, usually by sonication, to yield short fragments of DNA+protein complexes
  - Antibody to a TF or other binding protein used to fish out fragments containing that DNA binding protein
  - DNA is then “released” and can be analyzed by sequencing
- Creates a pool of sequences enriched in binding sites for a particular protein
- Requires availability of excellent **antibodies** that can detect the protein *in vivo*

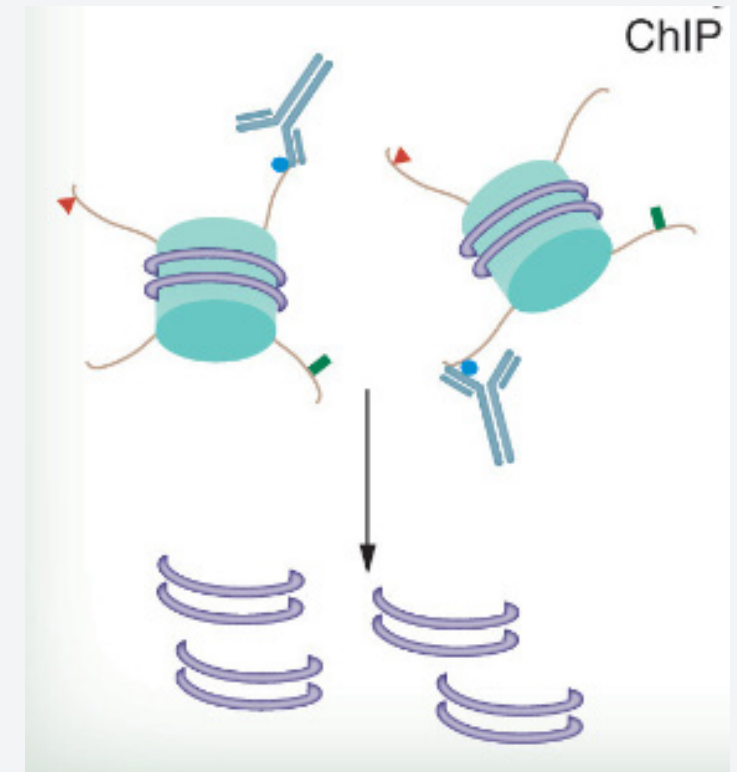
# ChIP computational issues

- First step is to map reads: BOWTIE, Noalign, BWA or other
- ChIP-seq reads **surround but may not contain** the DNA binding site
  - Sequence is generated from the *ends* of *randomly sheared* fragments, which overlap at the protein binding site
- Gives rise to two adjacent sets of read peaks
- Programs like MACS and HOMER automatically subtract your control (genomic input) from sample reads to define a final set of peaks

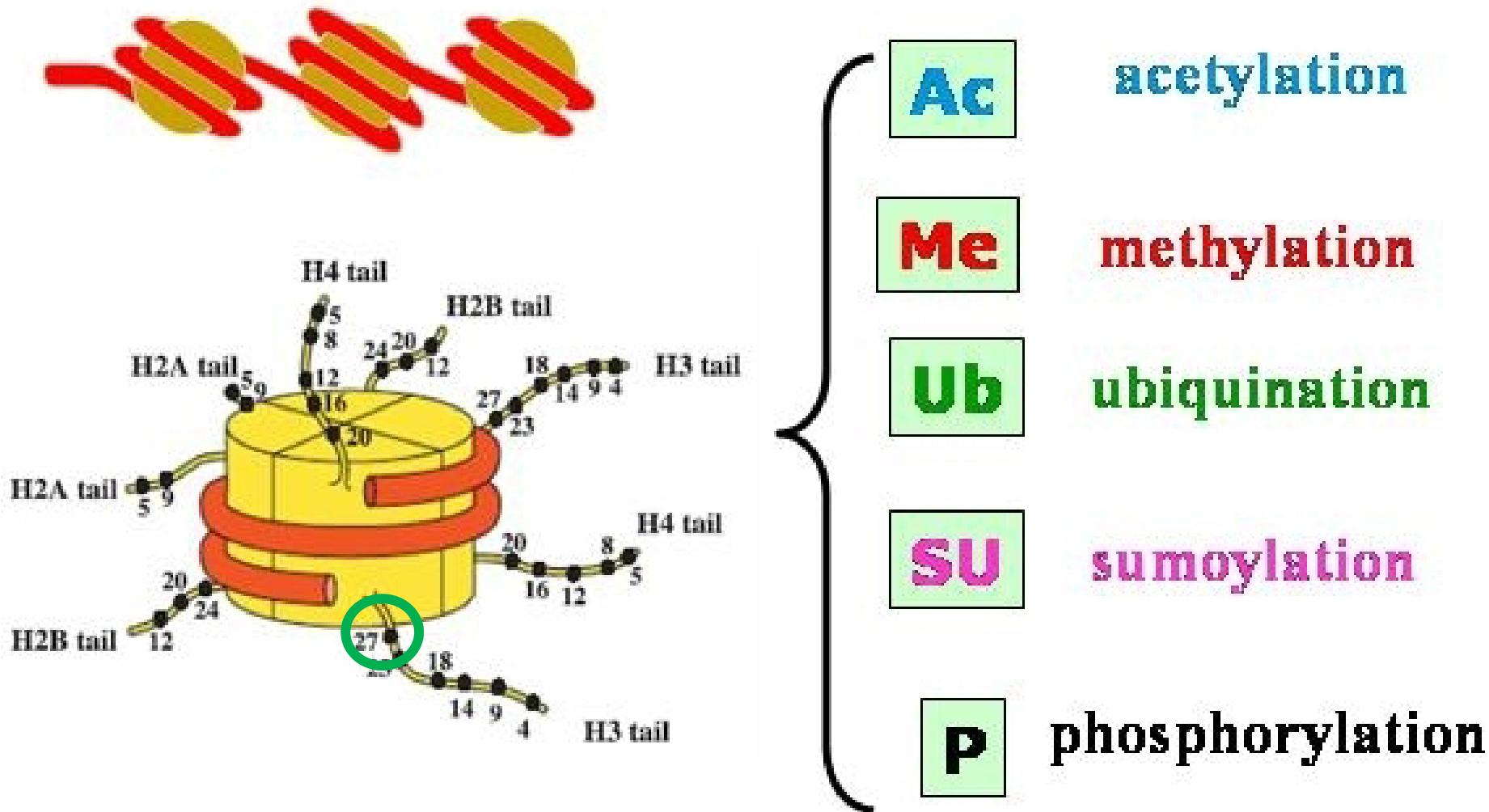


# ChIP for histone marks

- ChIP-seq can be used to profile not only TF binding sites but also histone modifications.
- Data/peak characteristics are different depending on what is profiled.
  - TFs are typically sharp peaks; chromatin marks are more diffuse

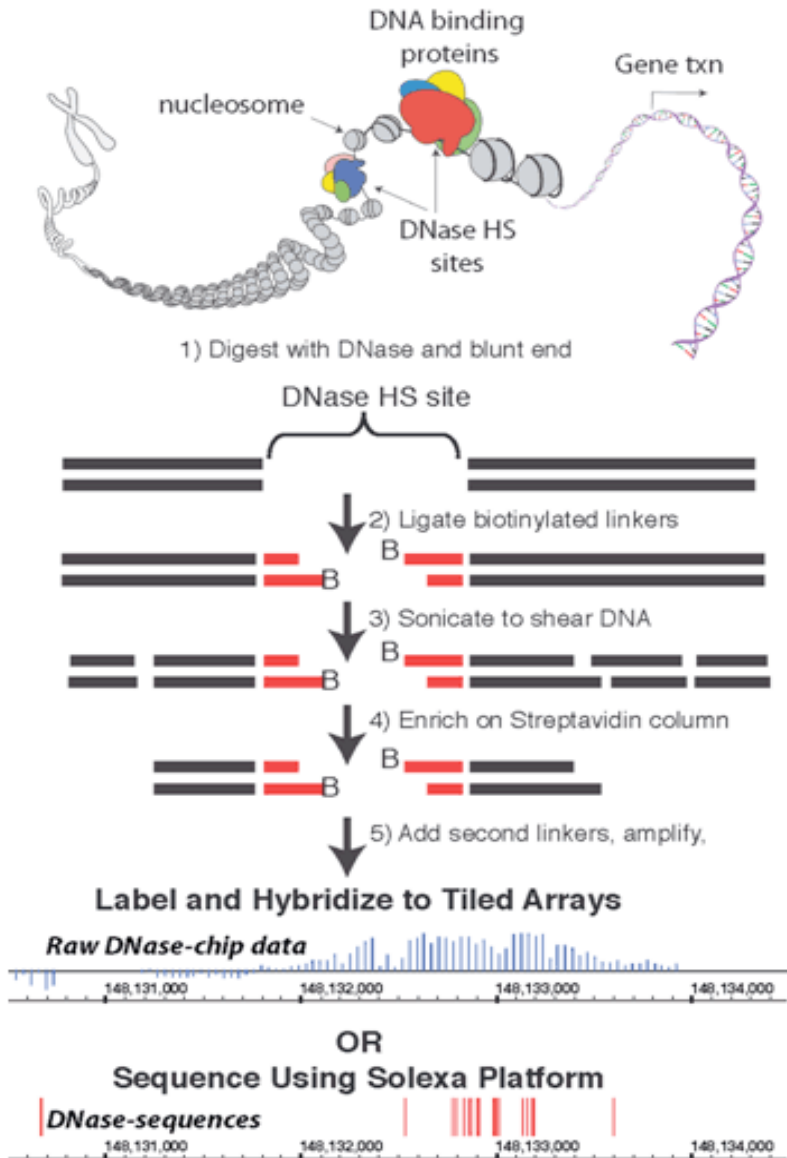


# All histones in the tetramer have “tails” that can be modified in various ways



Methylation or acetylation of **Lysines (K)** in histone H3 have a known effect on transcriptional activity

# How to find accessible DNA?



The first approach:

Crawford et al., *Genome Research* 16:123, 2006 (Francis Collins' laboratory)

Genome-wide identification of DNase I Hypersensitive sites (DHS)

Later variants also based on DNase I treatment, but different protocol and different philosophy.

ChIP-exo, FAIRE-seq, Mnase-seq, **ATAC-seq** etc.

2019 Review covering different methods:

## Chromatin accessibility and the regulatory epigenome

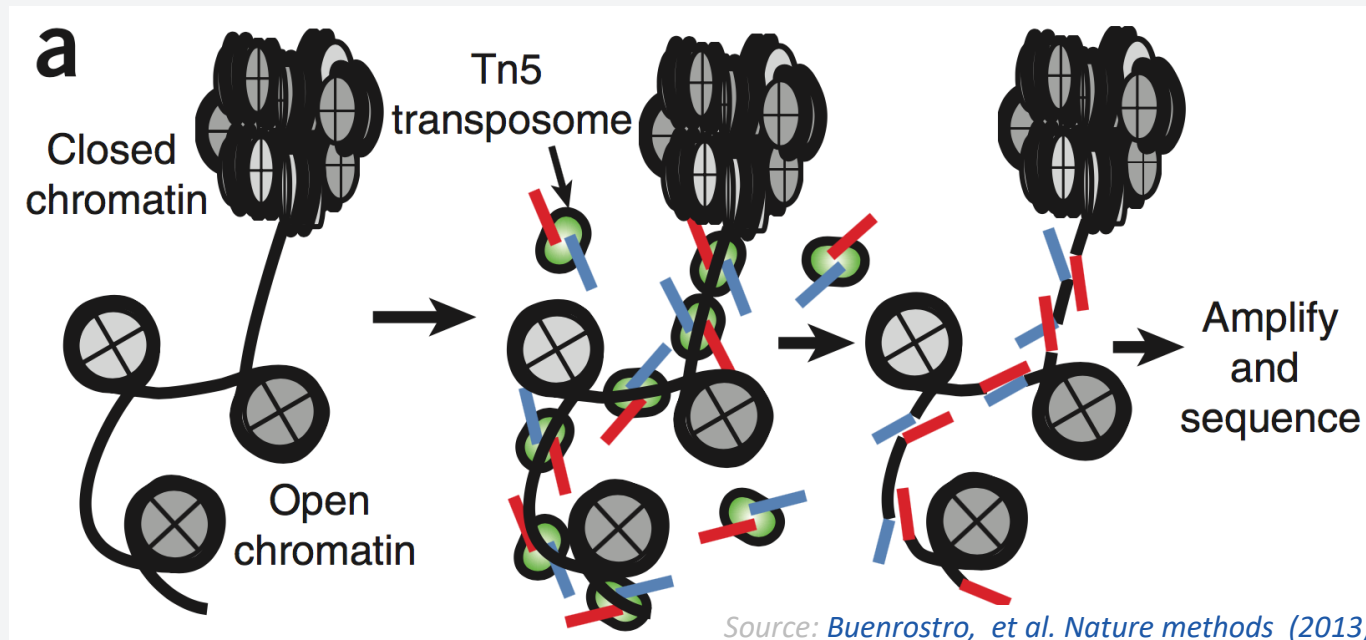
[Sandy L. Klemm, Zohar Shipony & William J. Greenleaf](#)

[Nature Reviews Genetics](#) 20, 207–220 (2019) | [Cite this article](#)

# ATAC-seq: approach to open chromatin

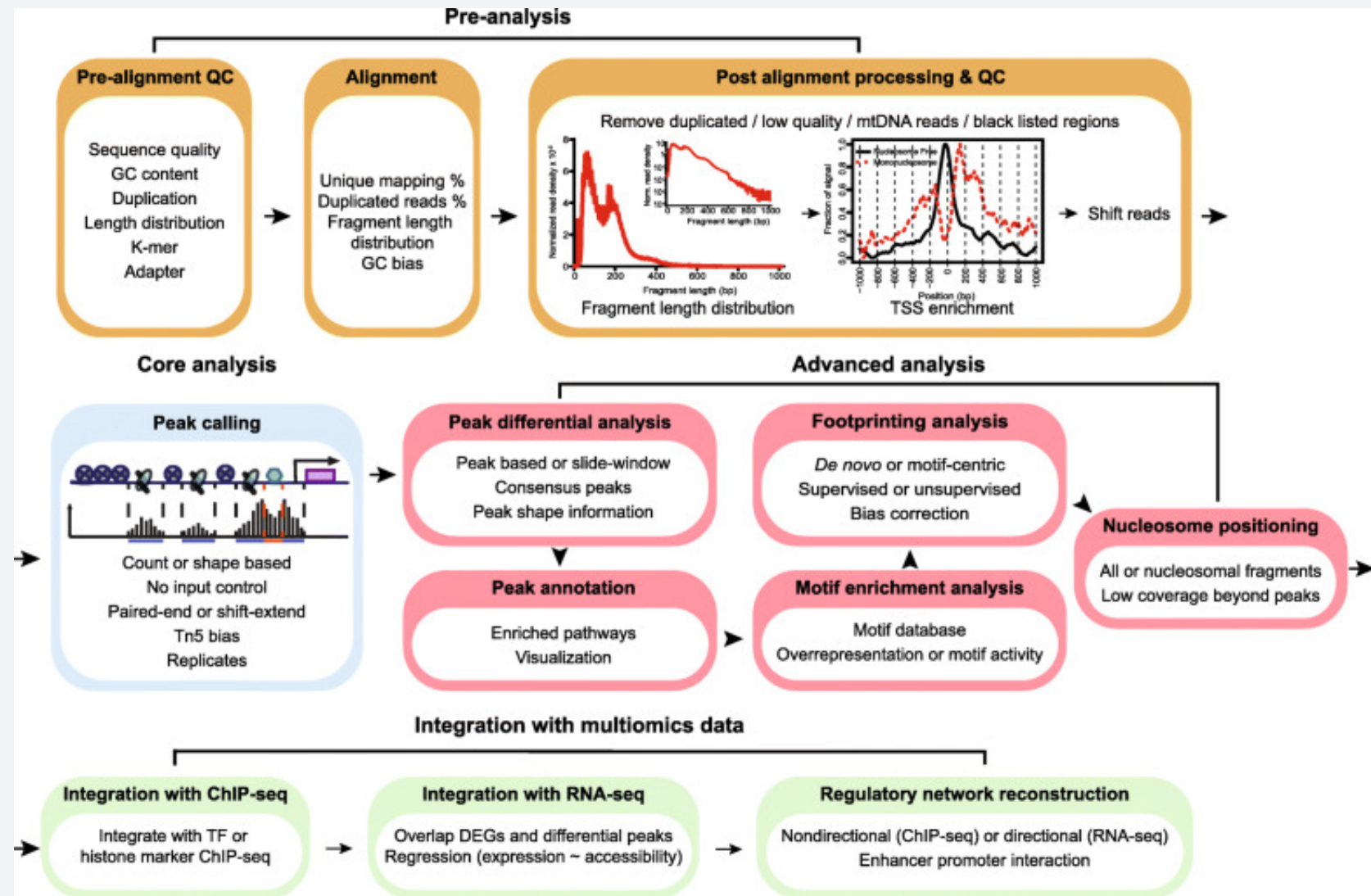
(Assay of Transposase Accessible Chromatin sequencing)

- Uses hyperactive Tn5 transposase to cut and tag accessible DNA
- Transposase “jumps” preferentially (and randomly) into accessible chromatin
- Because of the design the transposase breaks DNA where it jumps in, tagging the site with the primer
- Two insertions close together yield fragments of the size amenable for Illumina sequencing
- PCR amplification between primers is all you need to make a library
- Since it skips library-making steps, it can be done with small amounts of chromatin – e.g. 50K vs 1M cells



# General analysis workflow

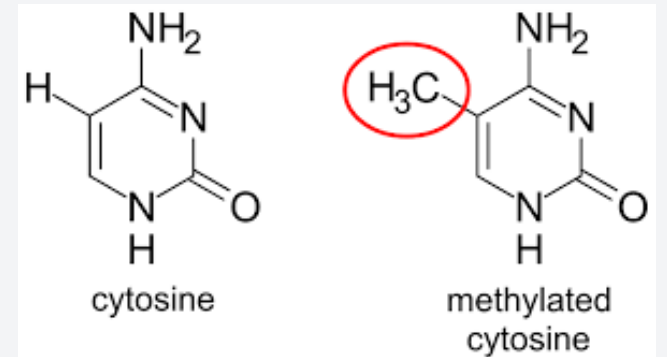
- **Upstream Analysis:**
- Alignment, Quality Control, and Peak Calling
- **Downstream Analysis:**
- Mapping peaks to nearby genes (esp. differentially expressed genes)
- Identifying enriched motifs
- Overlapping with multi-omics genome features



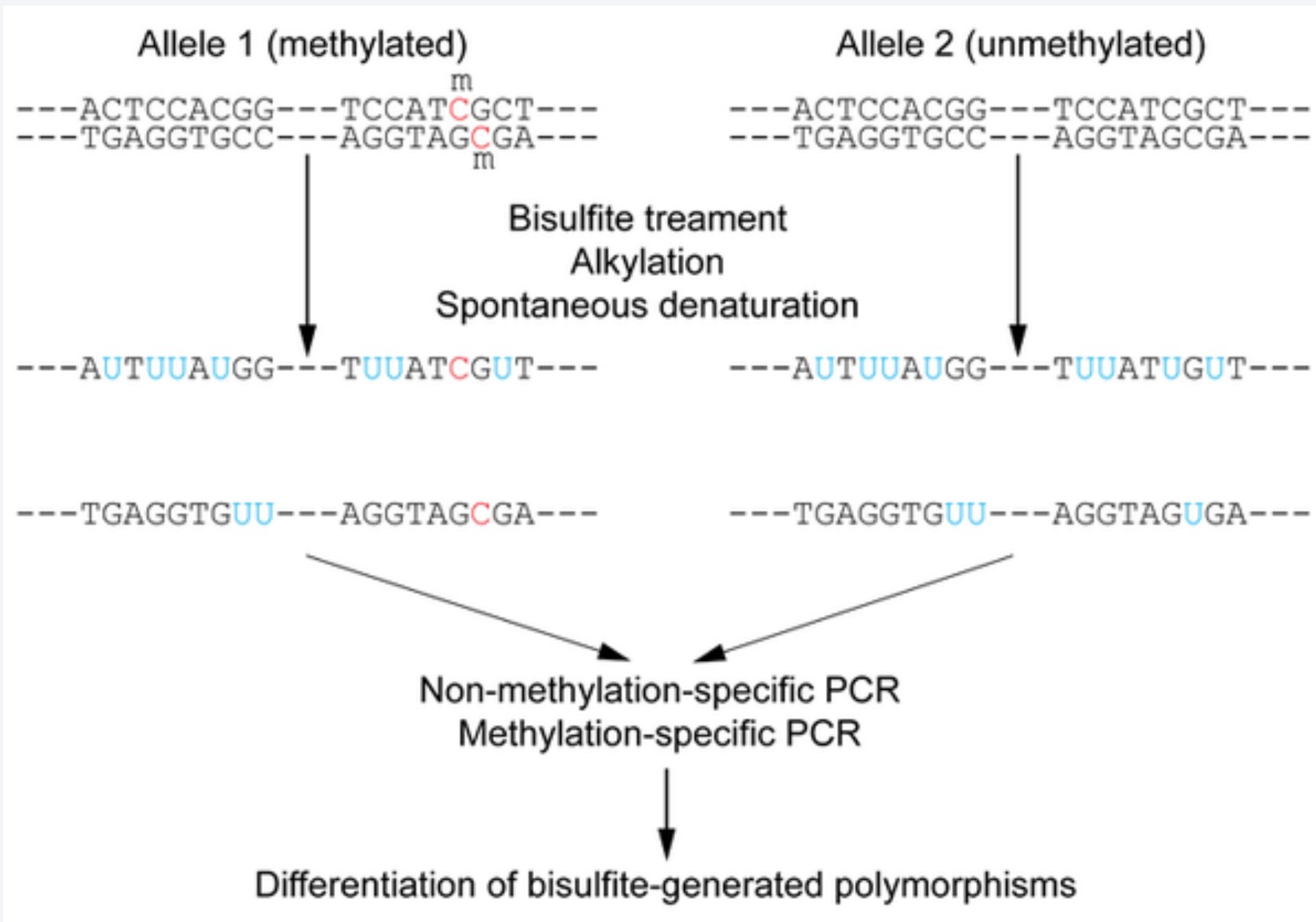
Source: [Yan, et al. Genome biology \(2020\).](#)

# DNA Methylation

- Methyl (-CH<sub>3</sub>) group added to Cytosine ('C')
- CpG (CG dinucleotide) is often methylated
- Methylated CpG may hinder transcription factor binding to DNA at that site
- Methylated CpG may recruit proteins that render local chromatin less accessible
- Roughly speaking, DNA methylation is repressive for gene expression



# CpG Methylation profiling

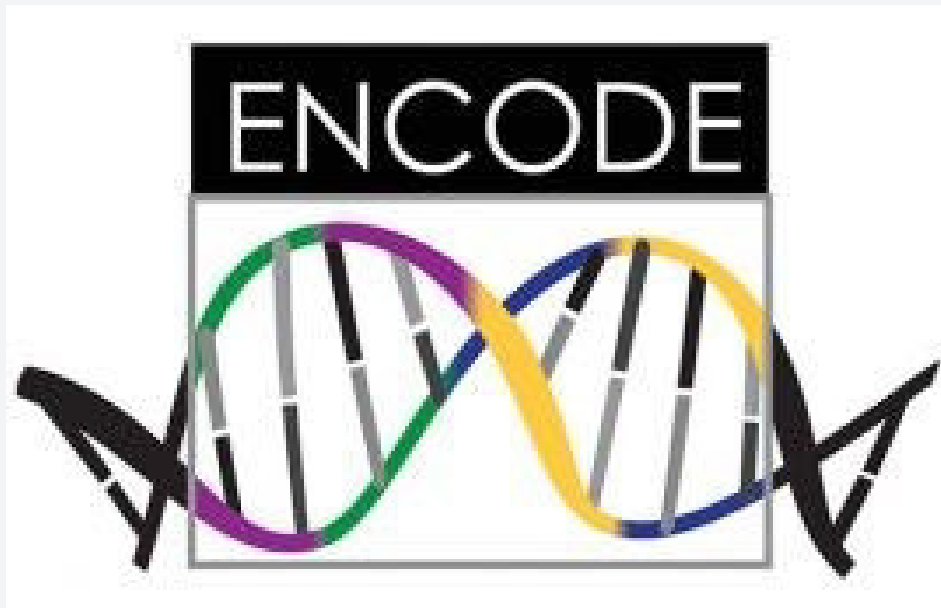


- Bisulfite sequencing

Other methods:

- DNA cleavage by methylation-sensitive restriction enzymes
- Immunoprecipitation with methyl-binding protein

# Insights from large-scale epigenomics studies



# Lessons from epigenomics assays

- Massive deep-sequencing of multiple chromatin features in cell lines (**ENCODE**), primary cell types and tissues (**Epigenetics Roadmap**)
  - Histone H3 modifications: highlight on H3K4me1, H3K4me3, H3K27Ac, H3K27me3.
  - TFs and other chromatin proteins: e.g. P300 (acetyltransferase)
- **H3K4me3** marks are enriched at **active promoters**
  - H3K4me3 marks are largely the same in all cell lines, with a small fraction of marks being cell-specific
- **P300**, and **H3K4me1** is enriched at **enhancers**
  - Most P300 peaks also contain H3K4me1
  - P300, H3K4me1 marks are highly cell-type specific
  - Most P300 marks are enhancers, but not all enhancers have P300
  - Most enhancers have an H3K4me1 mark, not all H3K4me1 marks are in enhancers
- Other marks: **H3K27Ac** or **H3K27me3**
  - Mutually exclusive marks for open (Ac) versus closed (Me3) chromatin regions
  - H3K27Ac may be most general open chromatin mark: promoters and enhancers
  - H3K27Ac often found in combination with H3K4 me1/me3



# Application 2: DNA Methylation profiles in cancer and aging

- DNA Methylation levels can be condition-dependent
  - Aberrant methylation patterns in cancer (e.g., hypermethylation of tumor suppressors and hypomethylation of oncogenes)
  - Progressive increase in global methylation levels with age. Also aging-correlated hypomethylation at some genes.

[Front Bioinform.](#) 2022; 2: 847629.

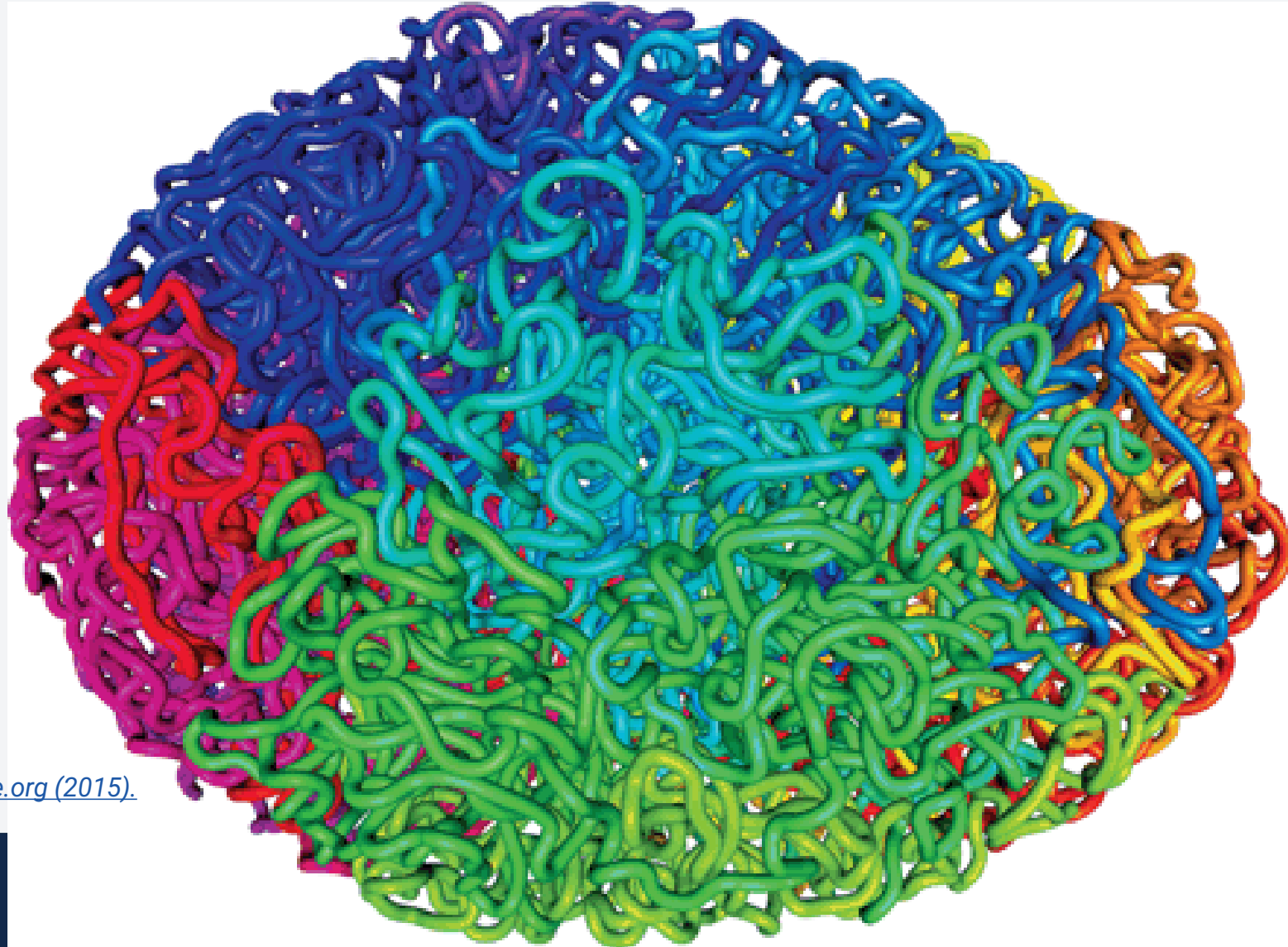
Published online 2022 Jun 2. doi: [10.3389/fbinf.2022.847629](https://doi.org/10.3389/fbinf.2022.847629)

DNA Methylation, Aging, and Cancer Risk: A Mini-Review

[Larry Chen](#),<sup>1</sup> [Patricia A. Ganz](#),<sup>2,3</sup> and [Mary E. Sehl](#)<sup>2,4,\*</sup>

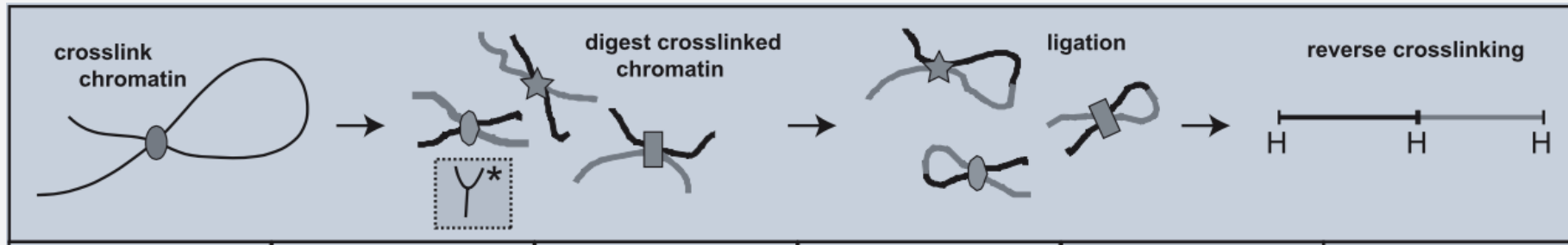


# 3D genome



Source: [Pennisi. Science.org](https://www.pennisi.com) (2015).

# Probing 3-dimensional chromatin structure with conformation capture

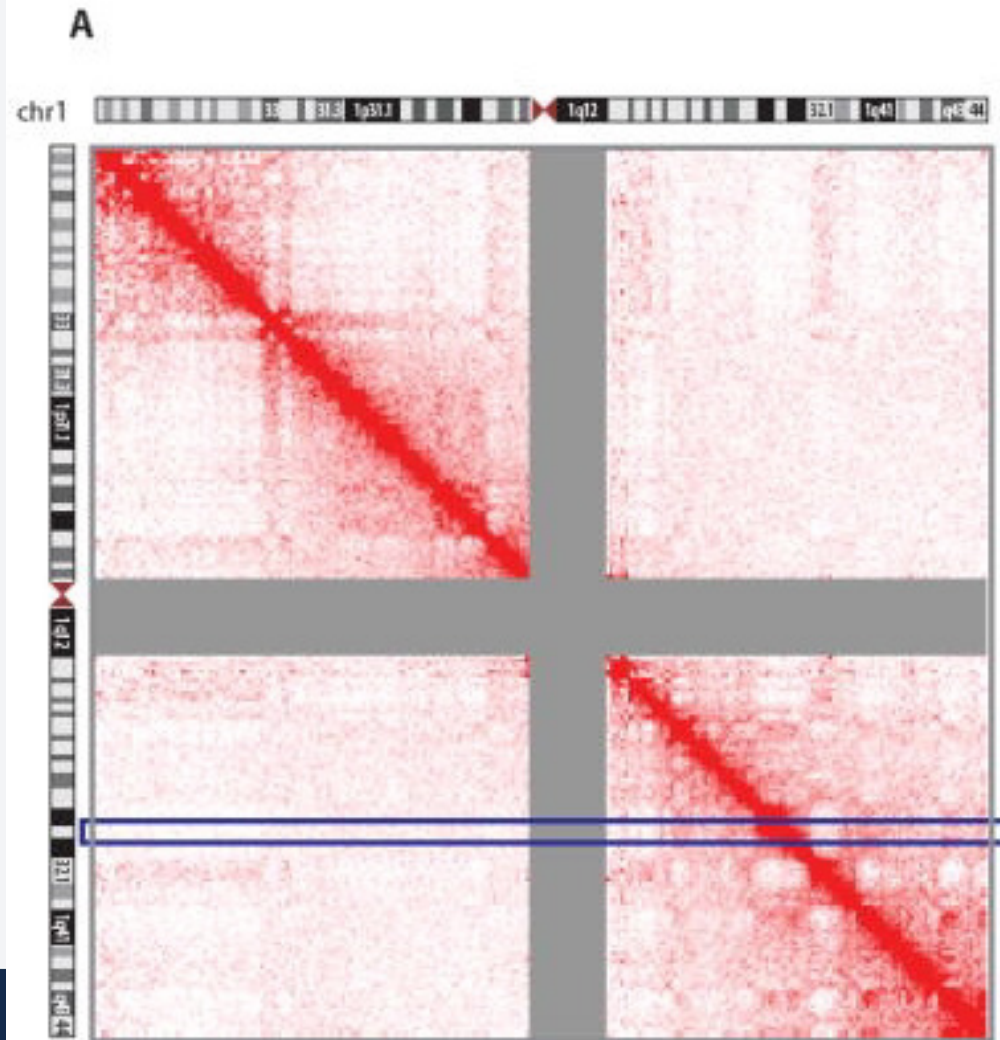


Source: Wit & Laat, 2012

# Hi-C “output”

Hi-C: A comprehensive technique to capture the conformation of genomes

[Jon-Matthew Belton](#),<sup>1</sup> [Rachel Patton McCord](#),<sup>1</sup> [Johan Gibcus](#),<sup>1</sup> [Natalia Naumova](#),<sup>1</sup> [Ye Zhan](#),<sup>1</sup> and [Job Dekker](#)<sup>1,\*</sup>



Heatmap of interactions between all 1 MB bins along chr1 for GM06990 cells.

The intensity of red color corresponds to the number of Hi-C interactions.

# Why is 3D information useful?

- The issue is finding out “who is talking to whom?”
  - Enhancers can be shared by multiple genes
  - Alternative promoters for the same gene can have very different regulatory partners
  - Position relative to the TSS is not a reliable indicator in large vertebrate genomes
  - 3D methods are necessary to tie enhancers and promoters (genes) together



# Summary (epigenomics)

- Transcription factor binding sites genome-wide
- Histone modification profiles (different marks or combinations of marks can point to different classes of regulatory elements)
- DNA accessibility profiles
- CpG methylation profiles
- Epigenomic profiles are informative about gene expression and regulatory mechanisms



# Questions ?



# Regulatory Genomics Lab

1. See Pythonic way to process single cell data on sample with both scRNA-seq and scATAC-seq
2. Look at normalization and signatures for scATAC-seq data
3. Identify differentially accessible peak intervals
4. Search for DNA sequence motifs under peaks

