REGULATORY GENOMICS

Saurabh Sinha, Dept. of Computer Science & Carl R. Woese Institute of Genomic Biology, University of Illinois.

The importance of gene regulation

DNA, RNA, Proteins

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

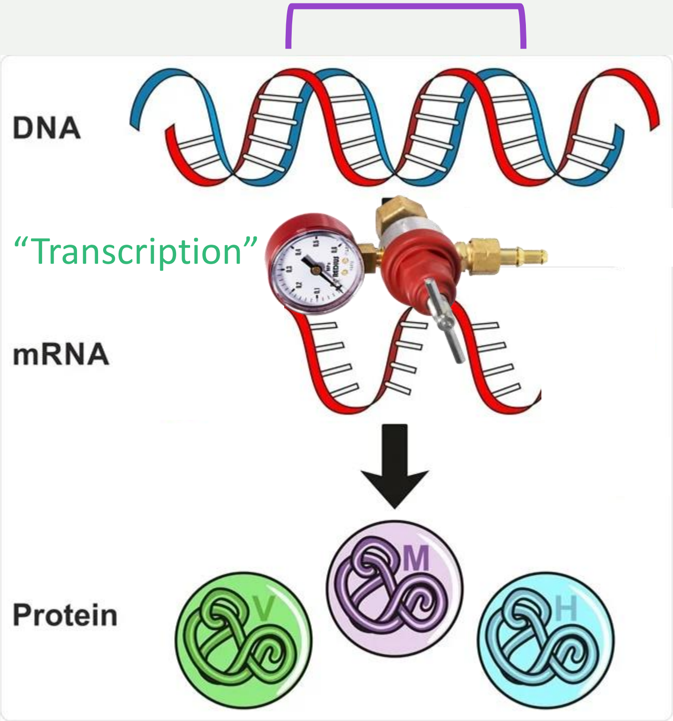


Image Credit: udaix/Shutterstock.com

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

GENE EXPRESSION

mRNA: a physical “copy” of gene

CAN BE REGULATED

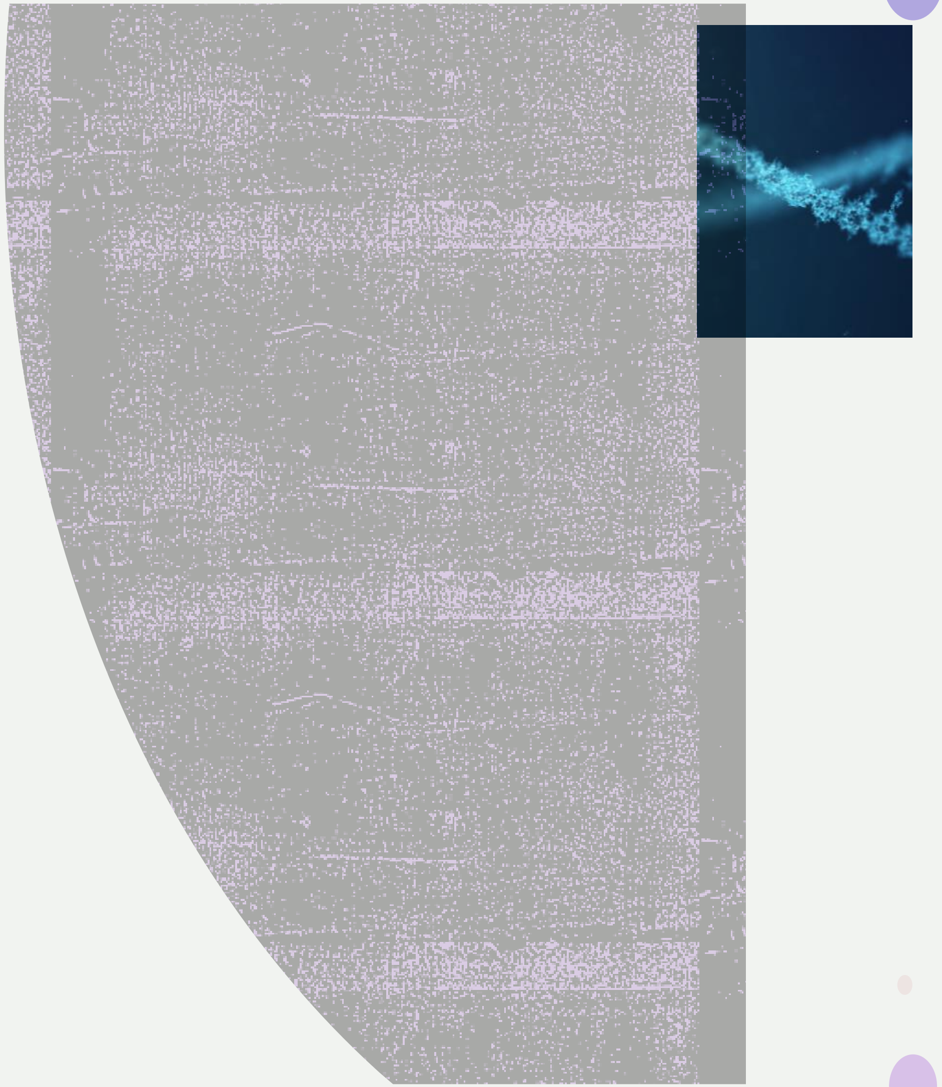
protein: molecule with important functions in cell

Gene regulation

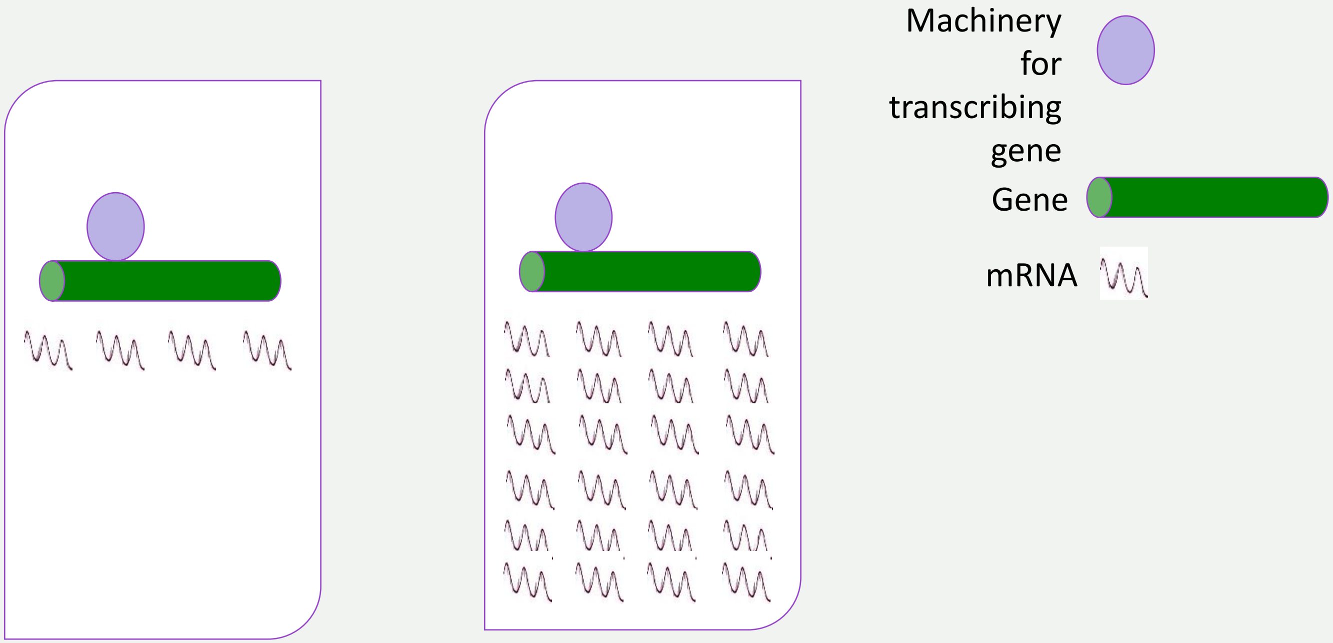
- Gene regulation is the process of turning genes on and off.
- Gene regulation ensures that the appropriate genes are expressed in the right cells at the proper times.

Source:

NIH National Human Genome Research Institute



Gene Regulation:
fast and slow transcription



Machinery for transcribing gene

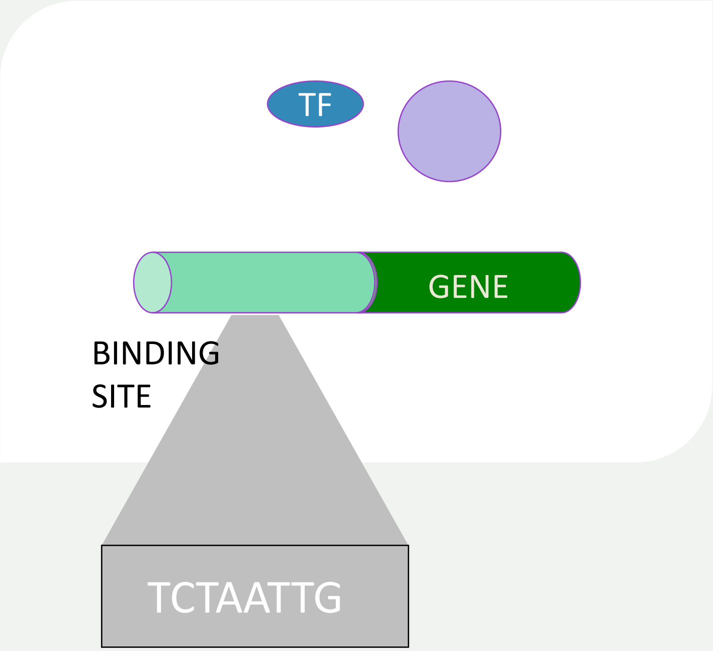
Gene

mRNA

Low gene expression

High gene expression

Transcription can be regulated by Proteins called Transcription Factors (TFs)

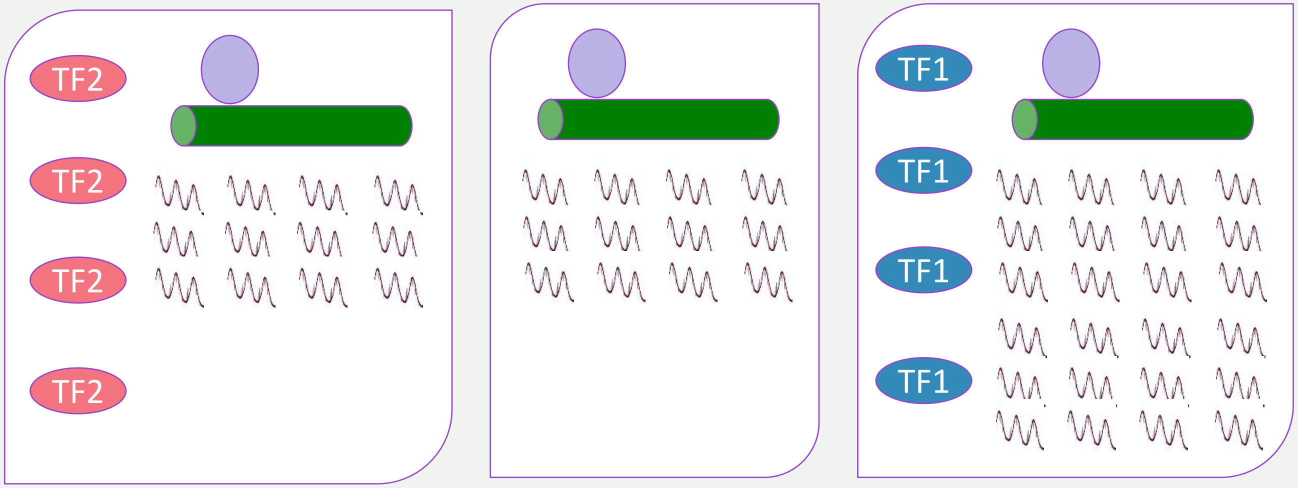


Humans have ~2000 TFs

Different cells may have different TFs

TF2 represses gene.
Low gene expression

TF1 activates gene.
High gene expression

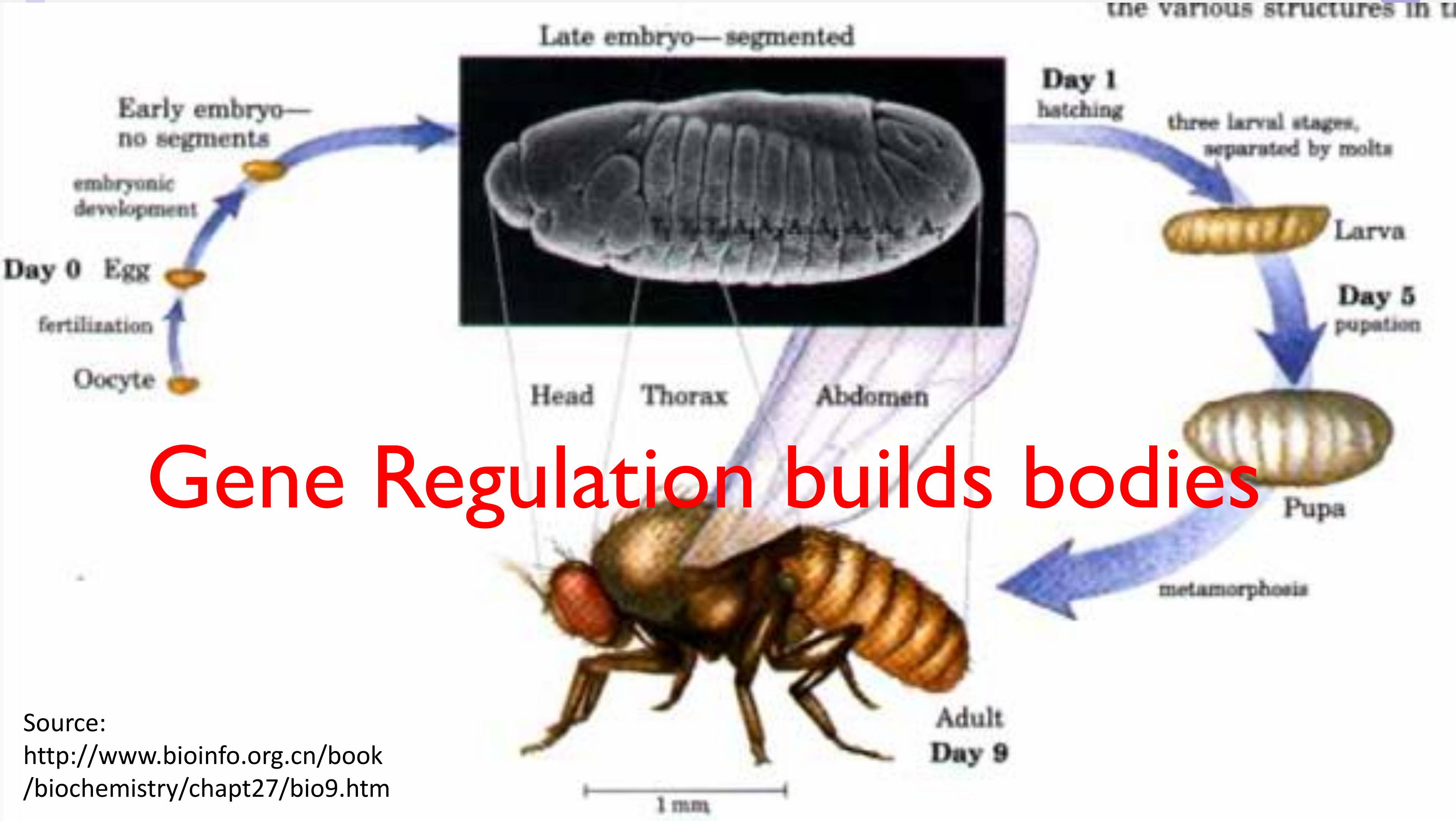


Liver cell

Skin cell

Heart cell

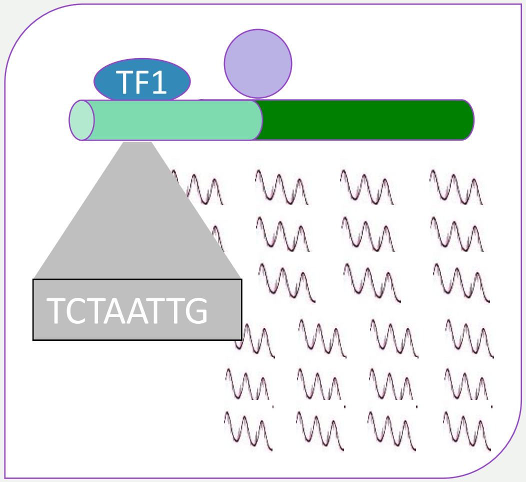
Gene Regulation builds bodies



Source:
http://www.bioinfo.org.cn/book/biochemistry/chapt27/bio9.htm

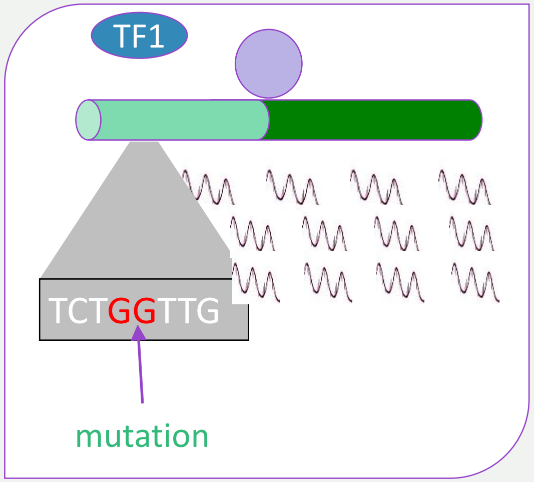
Different cells occasionally have different DNA

TF1 binds DNA and activates gene.
High gene expression



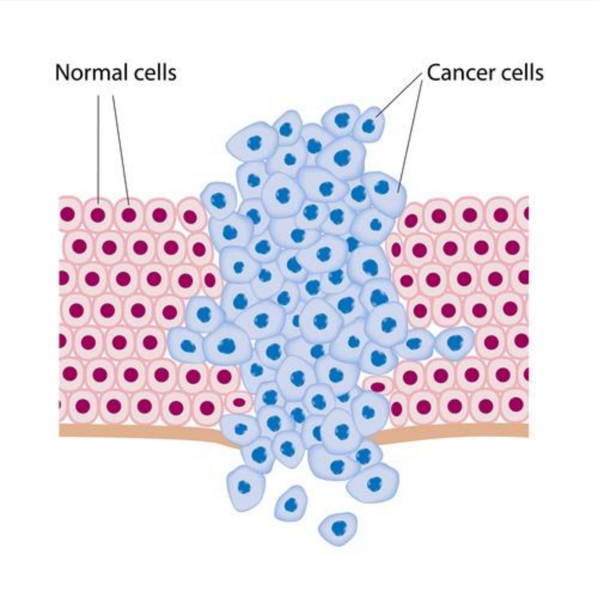
Normal cell

TF1 cannot binds DNA, does not activate gene.
Low gene expression



Tumor cell

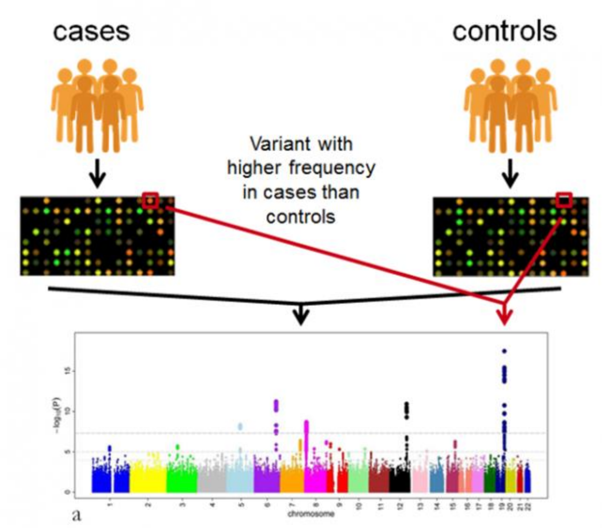
Gene Regulation is disrupted in cancer



Source: https://www.ck12.org/c/biology/gene-regulation-and-cancer/lesson/Gene-Regulation-and-Cancer-Advanced-BIO-ADV/

Most disease-related mutations are outside of genes

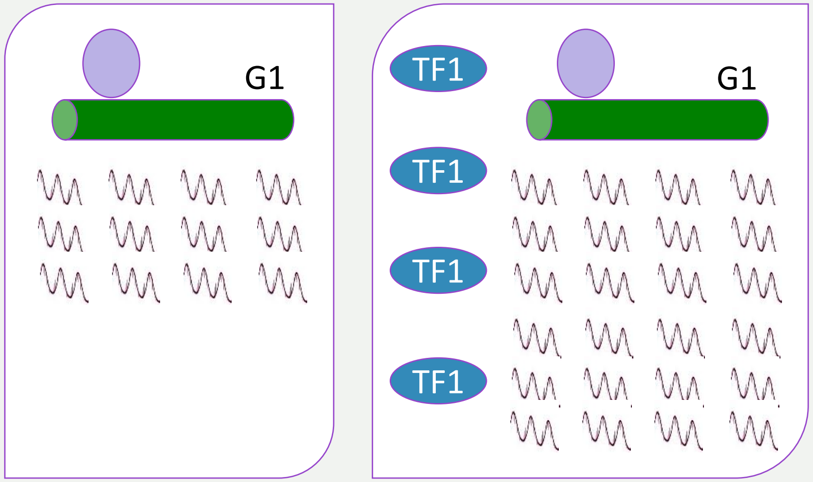
(impact gene regulation)



Source: https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

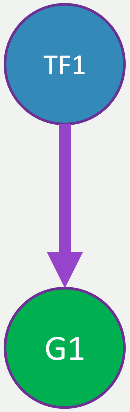
Gene Regulatory Networks: TF-gene relationships

TF1 activates gene.
High gene expression



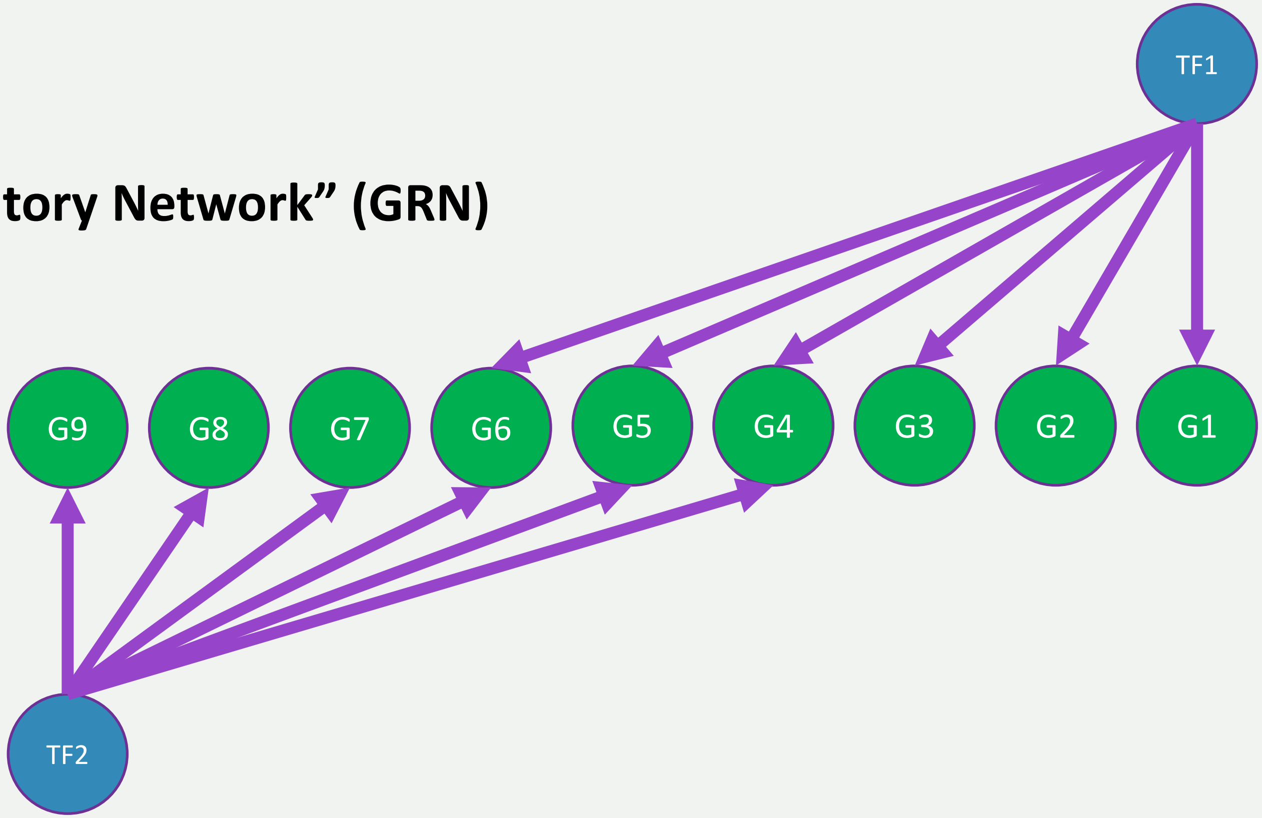
Healthy sample

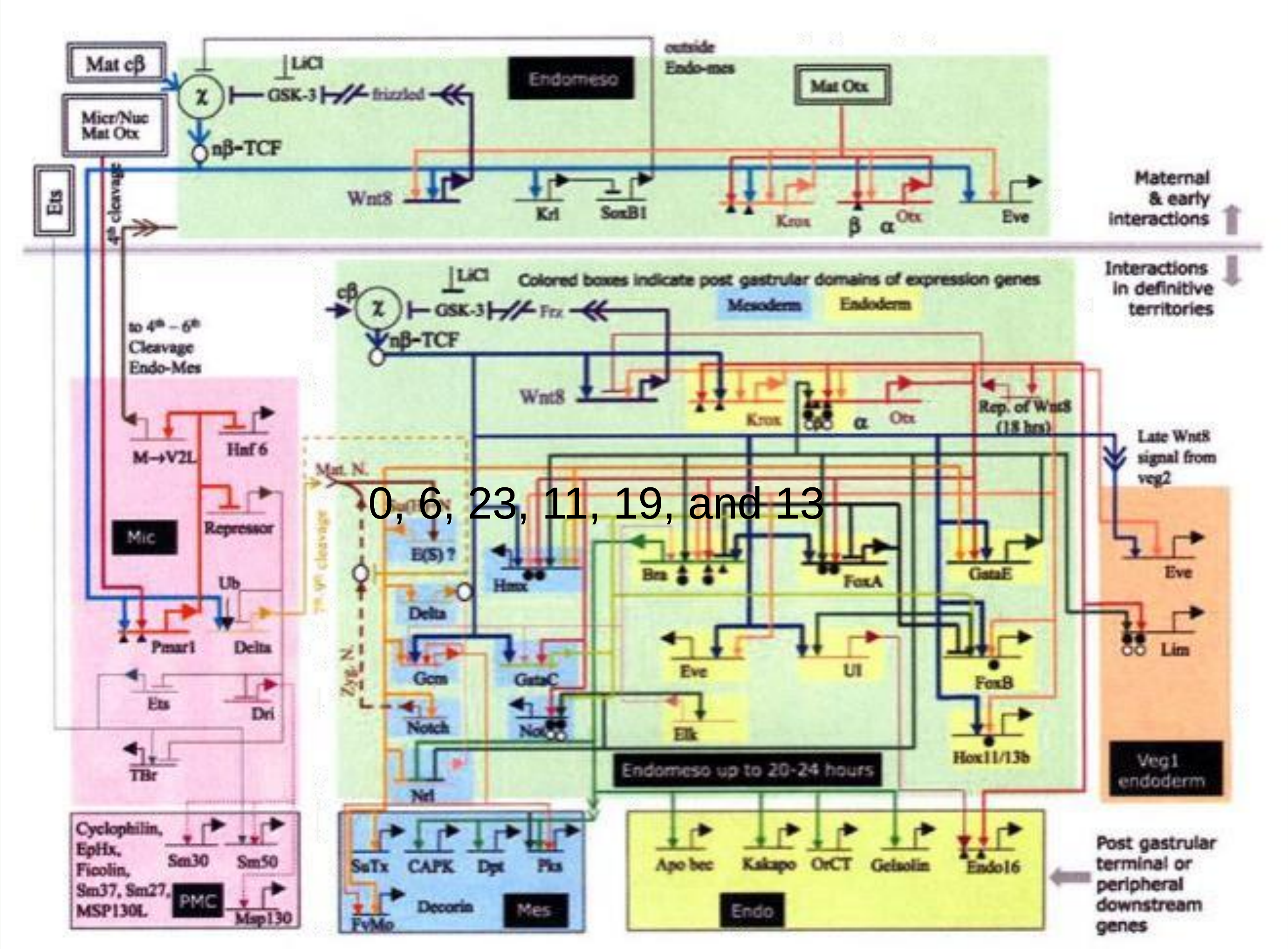
Tumor sample



Gene Regulatory Networks: TF-gene relationships

“Gene Regulatory Network” (GRN)

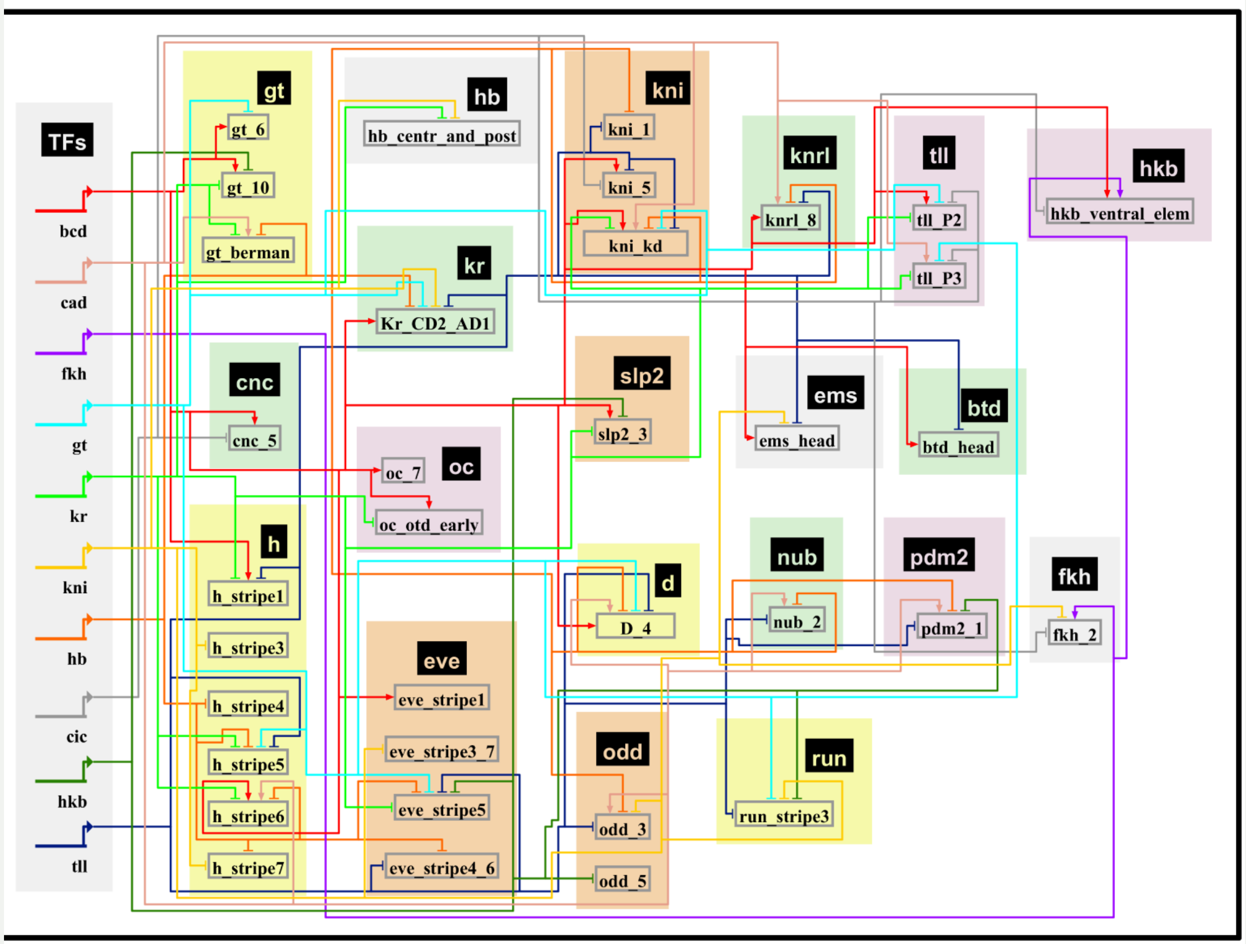


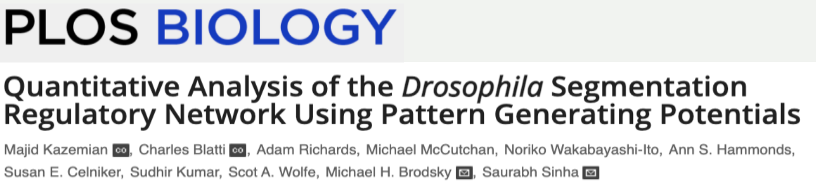


0, 6, 23, 11, 19, and 13

Genetic regulatory network controlling the development of the body plan of the sea urchin embryo. Davidson et al., Science, 295(5560):1669-1678

GRNs can be reconstructed computationally



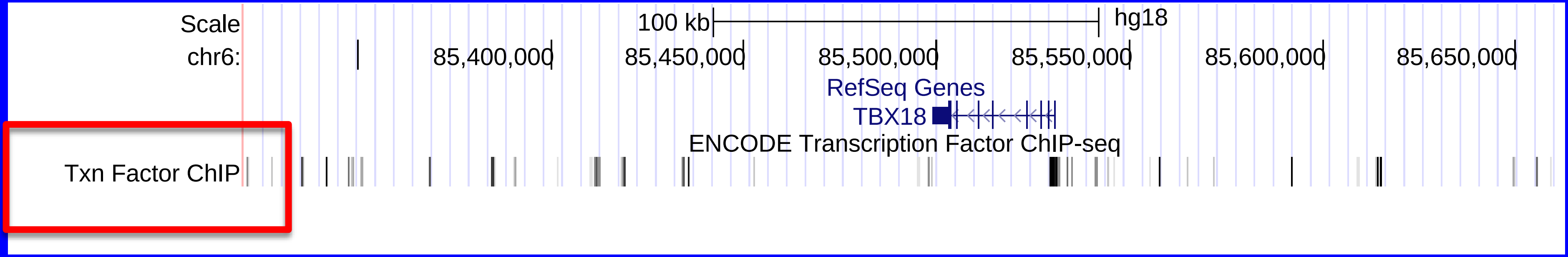

PLOS BIOLOGY

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



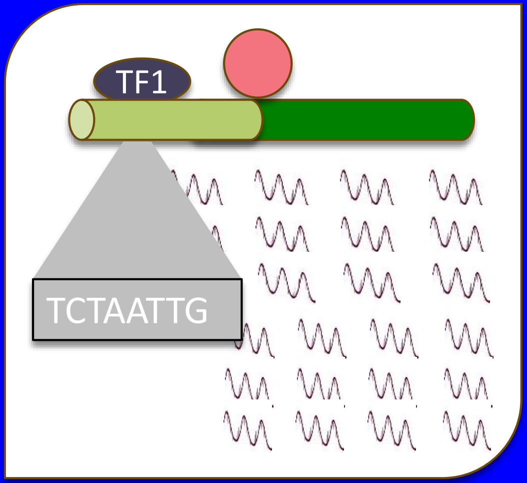
One experiment per cell type AND PER TF

… 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

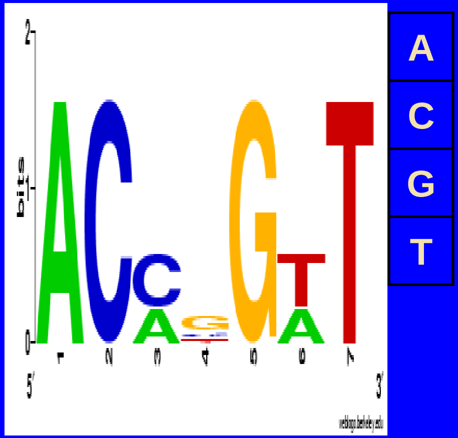
Motifs and DNA sequence analysis

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

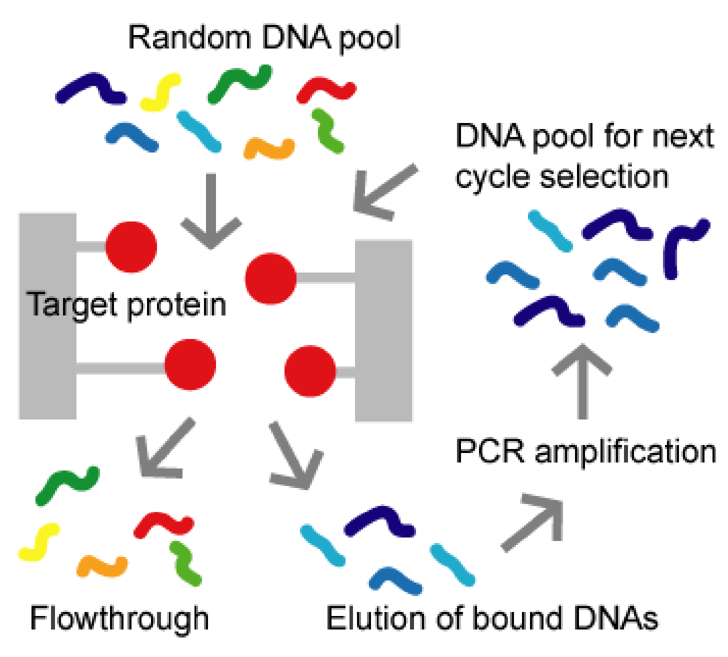
ACCCGTT
ACCGGTT
ACAGGAT
ACCGGTT
ACATGAT



"MOTIF"

How?

- SELEX

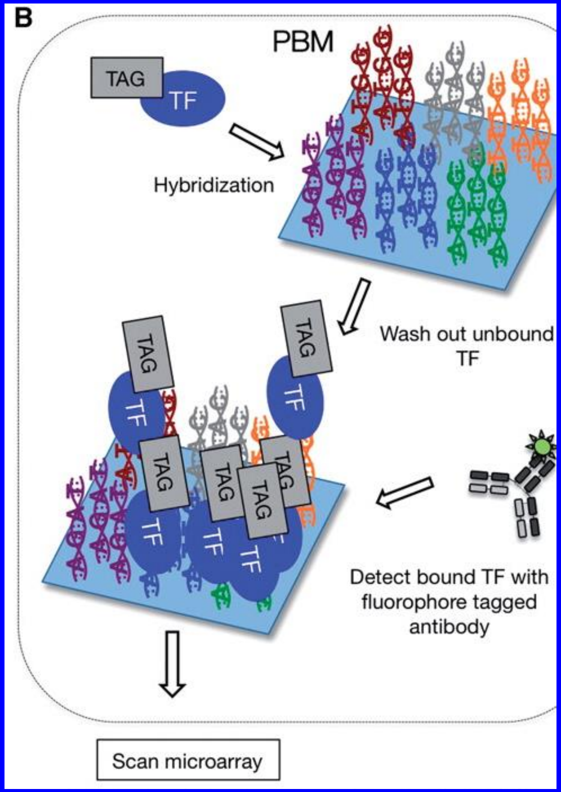


http://altair.sci.hokudai.ac.jp/g6/Projects/Selex-e.html

TAACCCGTTC
GTACCGGTTG
ACACAGGATT
AACCGGTTA
GGACATGAT

How?

- Protein binding microarrays

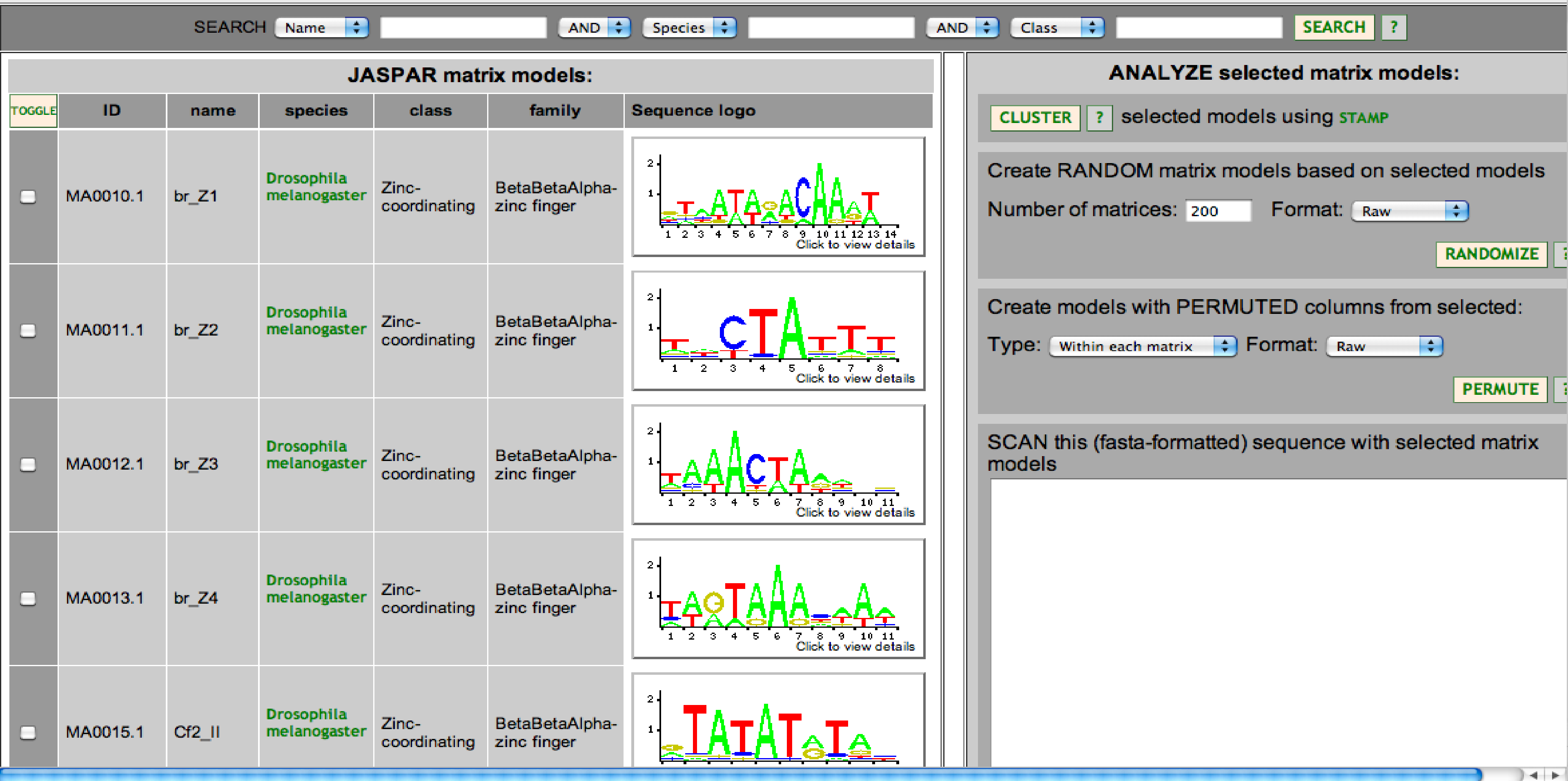


http://bfg.oxfordjournals.org/content/9/5-6/362/F2.large.jpg

```
TAACCCGTTC
GTACCGGTTG
ACACAGGATT
 AACCGGTTA
GGACATGAT
```

Motif Databases

- JASPAR: http://jaspar.genereg.net/



SEARCH Name AND Species AND Class SEARCH ?

JASPAR matrix models:

TOGGLE	ID	name	species	class	family	Sequence logo
□	MA0010.1	br_Z1	Drosophila melanogaster	Zinc-coordinating	BetaBetaAlpha-zinc finger	Click to view details
□	MA0011.1	br_Z2	Drosophila melanogaster	Zinc-coordinating	BetaBetaAlpha-zinc finger	Click to view details
□	MA0012.1	br_Z3	Drosophila melanogaster	Zinc-coordinating	BetaBetaAlpha-zinc finger	Click to view details
□	MA0013.1	br_Z4	Drosophila melanogaster	Zinc-coordinating	BetaBetaAlpha-zinc finger	Click to view details
□	MA0015.1	Cf2_II	Drosophila melanogaster	Zinc-coordinating	BetaBetaAlpha-zinc finger	

ANALYZE selected matrix models:

CLUSTER ? selected models using STAMP

Create RANDOM matrix models based on selected models

Number of matrices: 200 Format: Raw

RANDOMIZE

Create models with PERMUTED columns from selected:

Type: Within each matrix Format: Raw

PERMUTE

SCAN this (fasta-formatted) sequence with selected matrix models

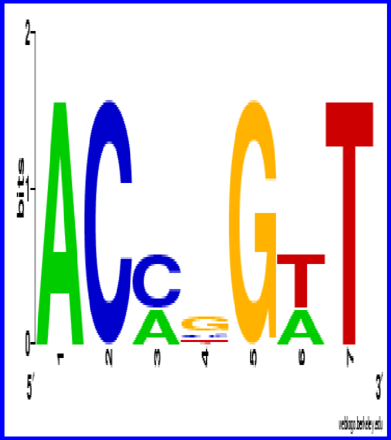
Motif Databases

- TRANSFAC
  - Public version and License version
- Cis-BP http://cisbp.ccbr.utoronto.ca/
  - Experimentally determined as well as computationally inferred motifs
- Hocomoco: http://hocomoco.autosome.ru/
  - Human and mouse motifs
- UniProbe: http://thebrain.bwh.harvard.edu/uniprobe/
  - variety of organisms, mostly mouse and human
- Fly Factor Survey: http://pgfe.umassmed.edu/TFDBS/
  - 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

$$\Pr(s \mid W)$$

What is Pr (s | W)?

5	0	2	0	0	2	0	**A**
0	5	3	1	0	0	0	**C**
0	0	0	3	5	0	0	**G**
0	0	0	1	0	3	5	**T**

1	0	0.4	0	0	0.4	0	**A**
0	1	0.6	0.2	0	0	0	**C**
0	0	0	0.6	1	0	0	**G**
0	0	0	0.2	0	0.6	1	**T**

Now, say s =ACCGGTT (consensus)
Pr(s|W) = 1 x 1 x 0.6 x 0.6 x 1 x 0.6 x 1 = 0.216.

Then, say s = ACACGTT (two mismatches from consensus)
Pr(s|W) = 1 x 1 x 0.4 x 0.2 x 1 x 0.6 x 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* =

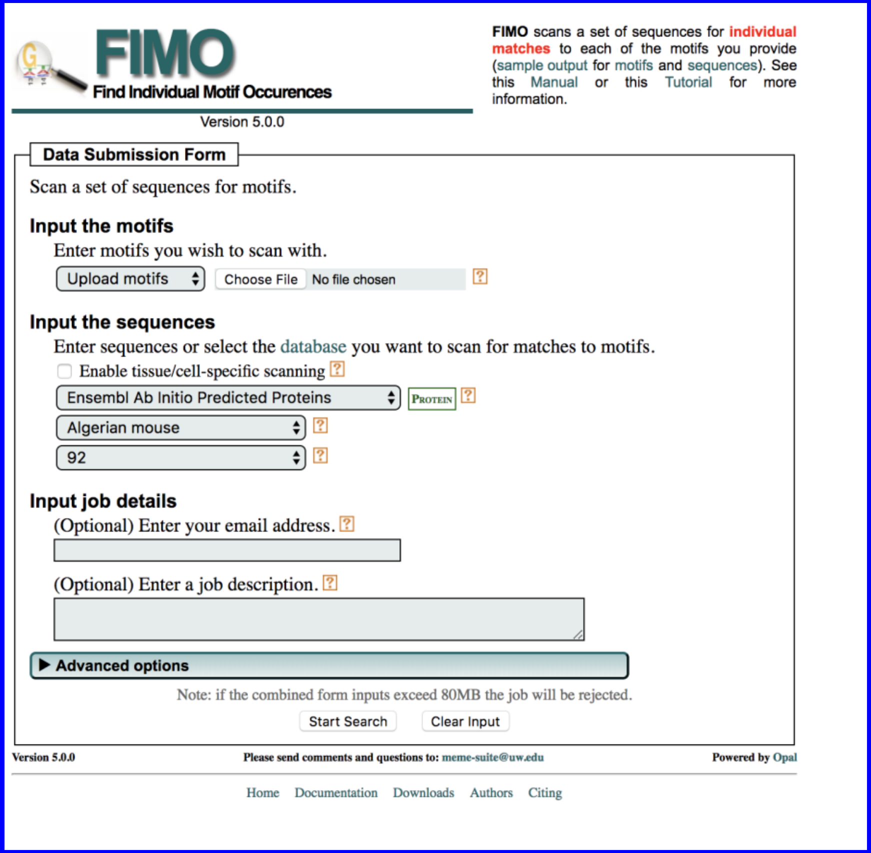
$$\boxed{\log \frac{\Pr(s \mid W)}{\Pr(s \mid W_b)}}$$

- Good scores > 0. Bad scores ≲ 0.

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

The FIMO program

- Grant, Bailey, Noble; *Bioinformatics* 2011.



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.

FIMO

Find Individual Motif Occurences

Version 5.0.0

Data Submission Form

Scan a set of sequences for motifs.

Input the motifs

Enter motifs you wish to scan with.

Upload motifs ⇕ | Choose File | 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 ⇕ | PROTEIN ?

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.

Start Search | Clear Input

Version 5.0.0 | Please send comments and questions to: meme-suite@uw.edu | Powered by Opal

Home Documentation Downloads Authors Citing

- Takes motif W, background W_b and a sequence S.

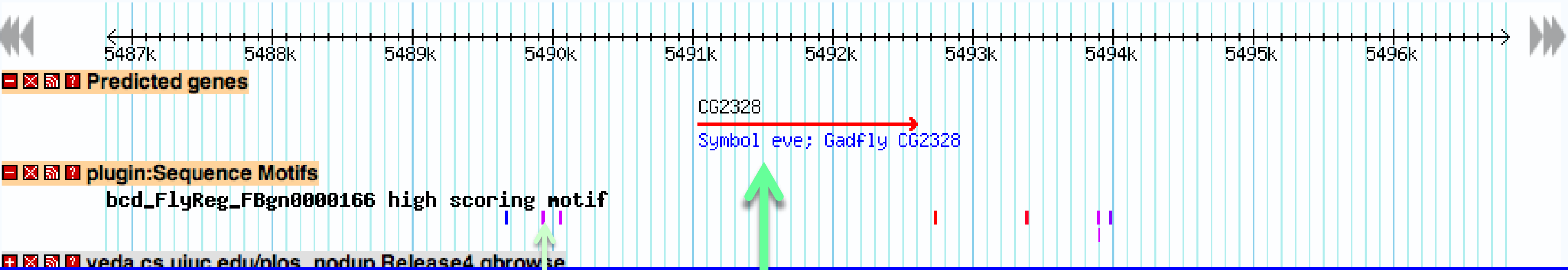
- Scans every site *s* in S, and computes its LLR score.

- 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



5487k
5488k
5489k
5490k
5491k
5492k
5493k
5494k
5495k
5496k
Predicted genes
CG2328
Symbol eve; Gadfly CG2328
plugin:Sequence Motifs
bcd_FlyReg_FBgn0000166 high scoring motif
veda.cs.uiuc.edu/plos_nodup Release4.gbrowse

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

Why?

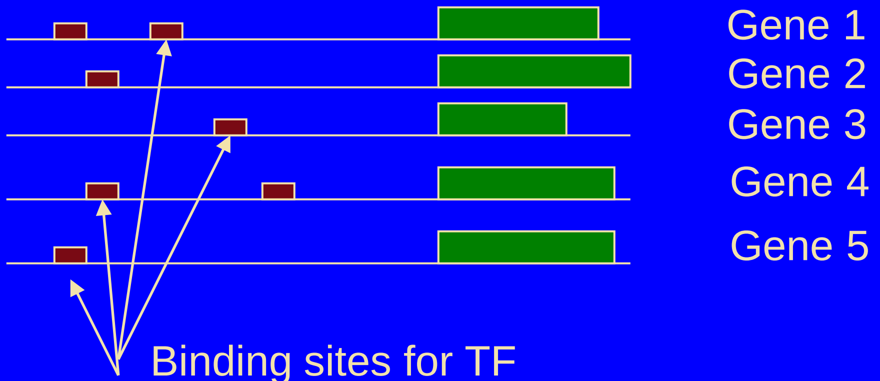
- We assumed that we have experimental characterization of a transcription factor's binding specificity (motif)

- What if we don't?

- There's a couple of options …

Option 1

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

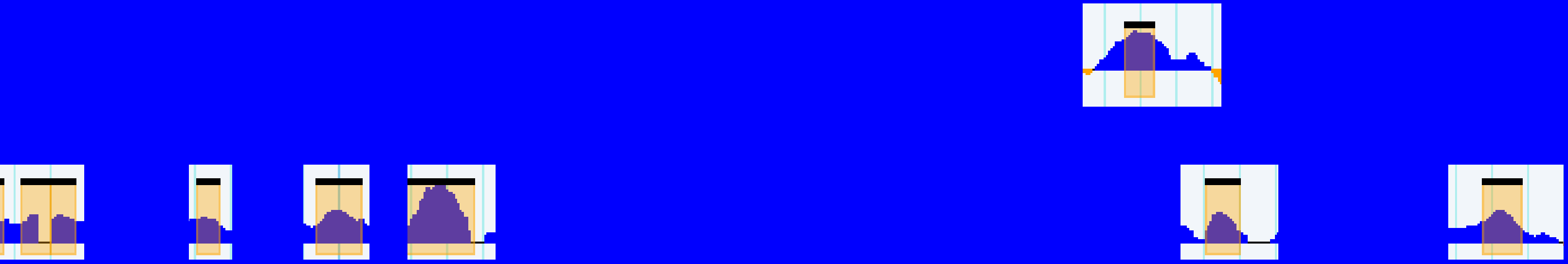


Option 1

- □ Now suppose we are given the promoter regions of the five genes G1, G2, … G5
- □ Can we find the binding sites of TF, without knowing about them *a priori* ?
- □ This is the computational motif finding problem
- □ To 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 finding problem

Motif finding 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 finding algorithms

- Gibbs sampling (MCMC) : Lawrence et al. 1993

- MEME (Expectation-Maximization) : Bailey & Elkan 94. (Very popular, visited in today's lab.)

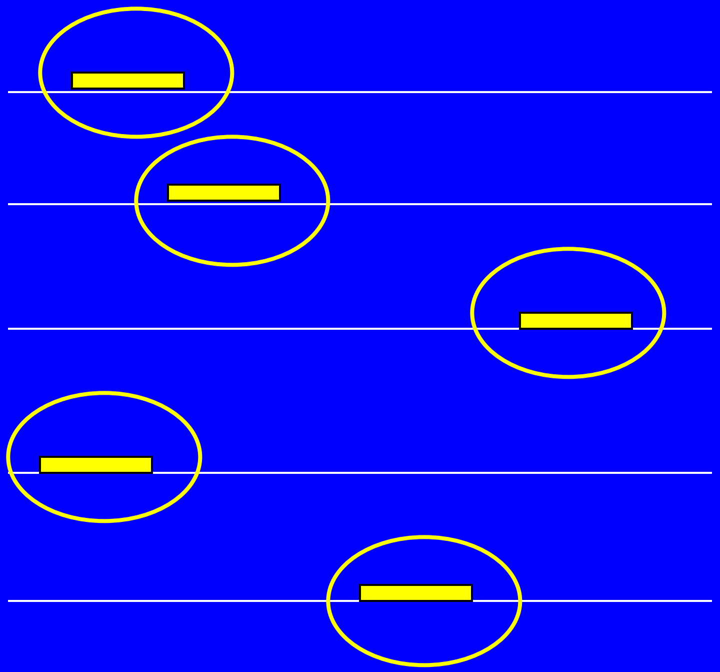
- CONSENSUS (Greedy search) : Stormo lab.

- Priority (Gibbs sampling, but allows for additional prior information to be incorporated): Hartemink lab.

- Many many others …

Examining one such algorithm

The “CONSENSUS” algorithm

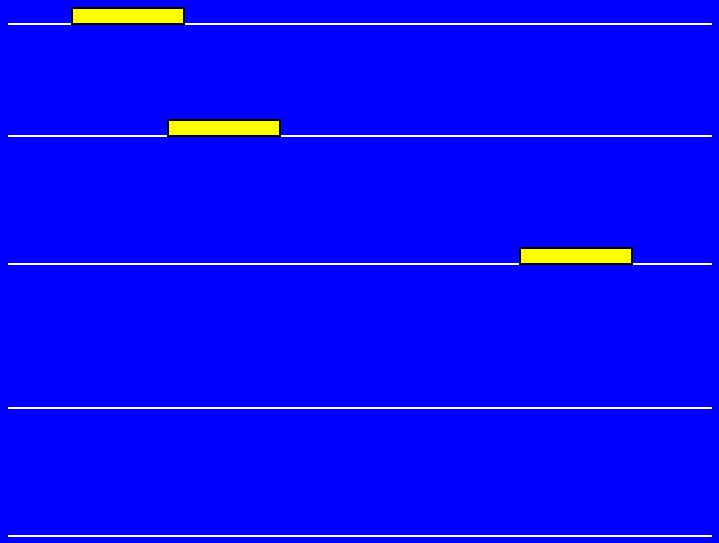


Final goal: Find a set of “substrings” (sites), one in each input sequence

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

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

The “CONSENSUS” algorithm

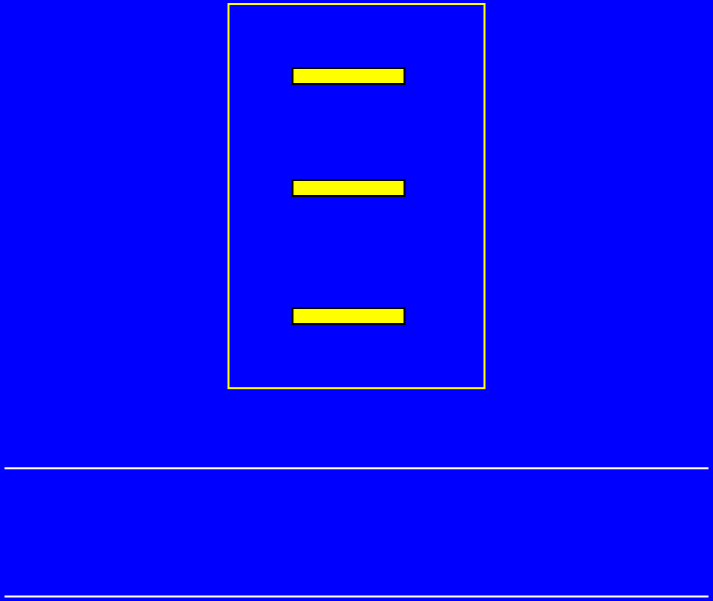


Start with a substring in one input sequence

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

The current set of substrings.

The “CONSENSUS” algorithm



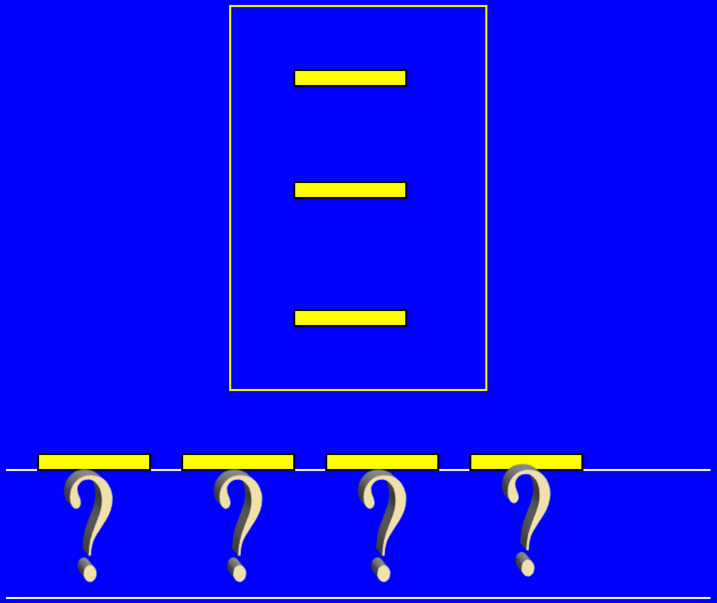
Start with a substring in one input sequence

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

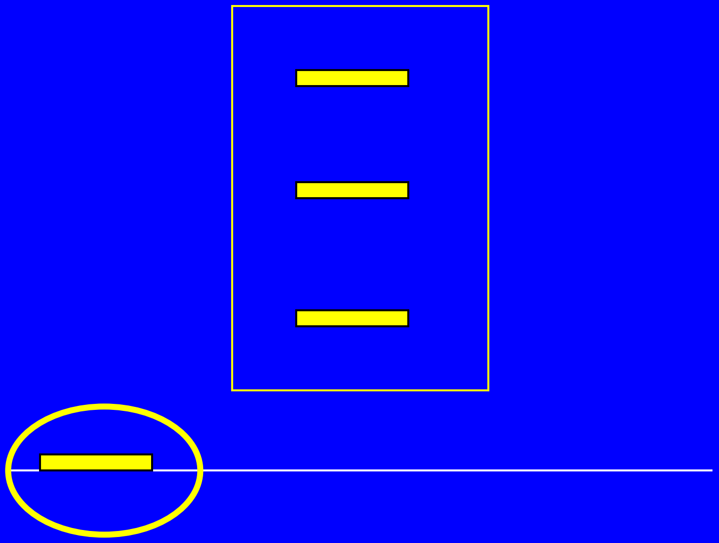
Build the set of substrings incrementally, adding one substring at a time

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

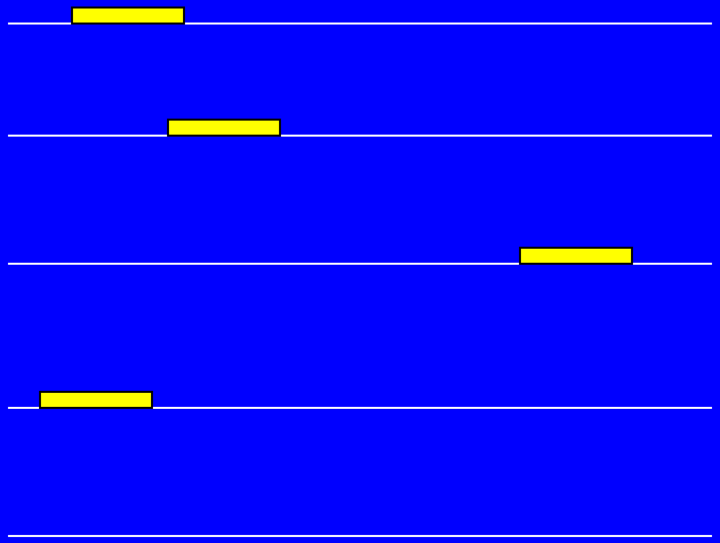
Build the set of substrings incrementally, adding one substring at a time

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

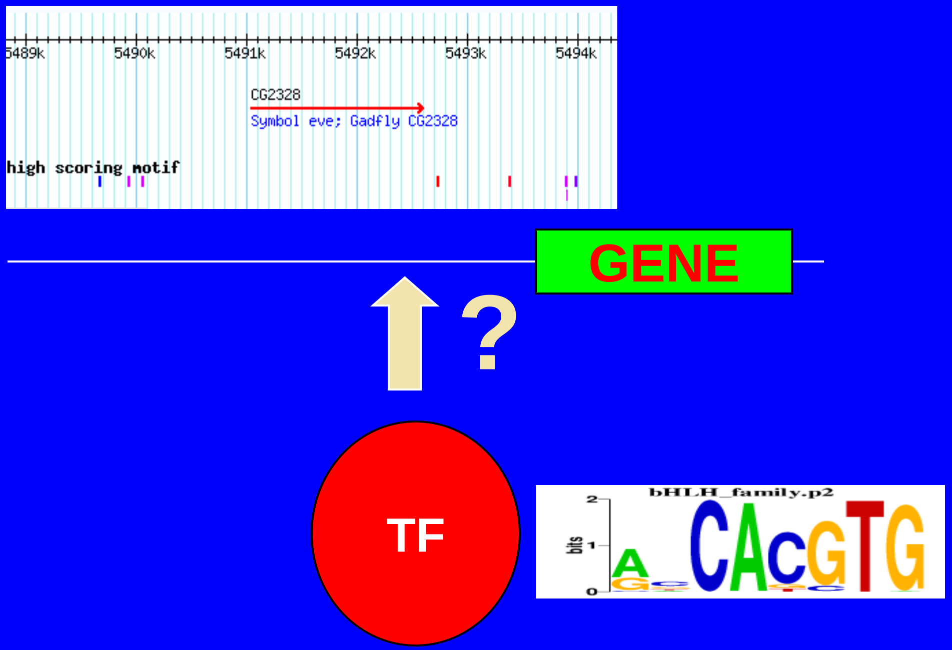
- Introduction to theory of motif finding
  - Moses & Sinha: http://www.moseslab.csb.utoronto.ca/Moses_Sinha_Bioinf_Tools_apps_2009.pdf
  - Das & Dai: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2099490/pdf/1471-2105-8-S7-S21.pdf

Motif finding tools

- MEME: http://meme-suite.org/
- RSAT: http://rsat.sb-roscoff.fr/

Integrating sequence analysis and expression data

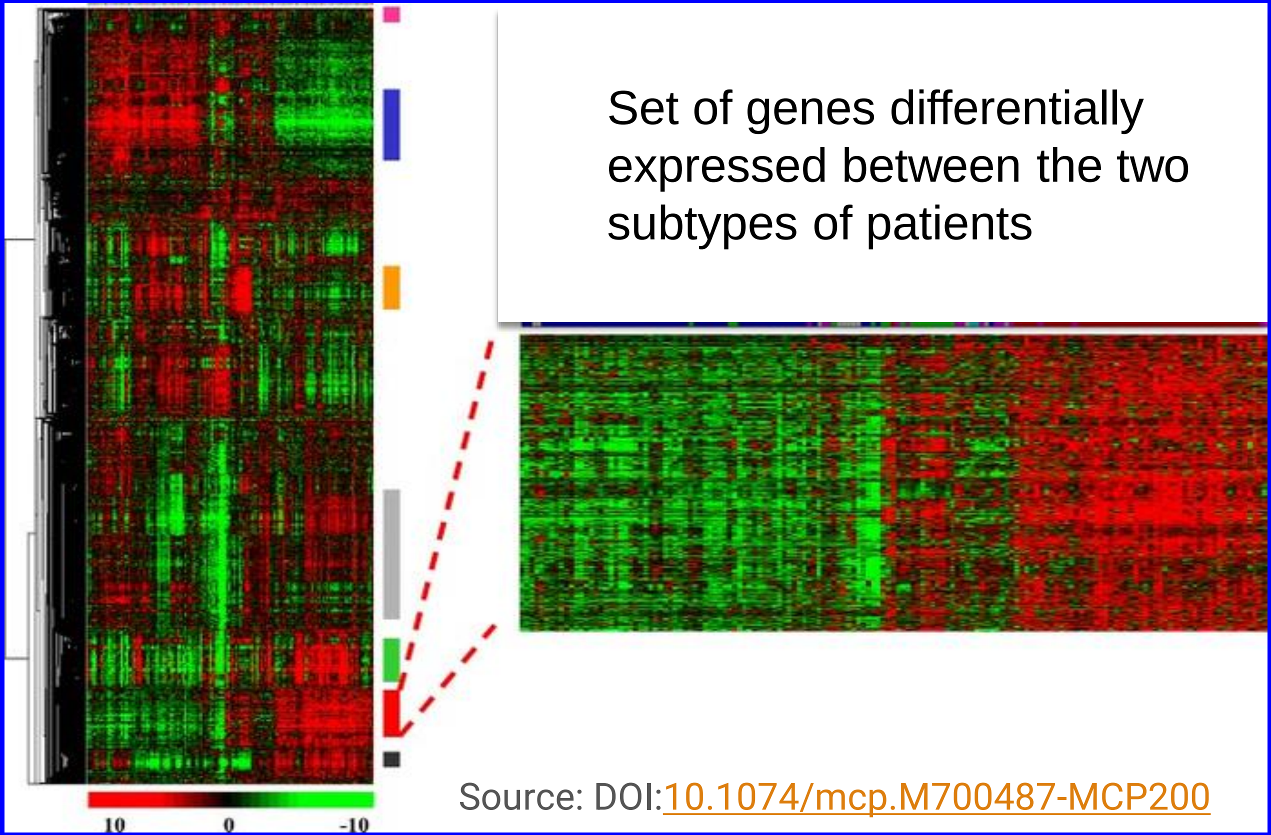
1. Predict regulatory targets of a TF



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

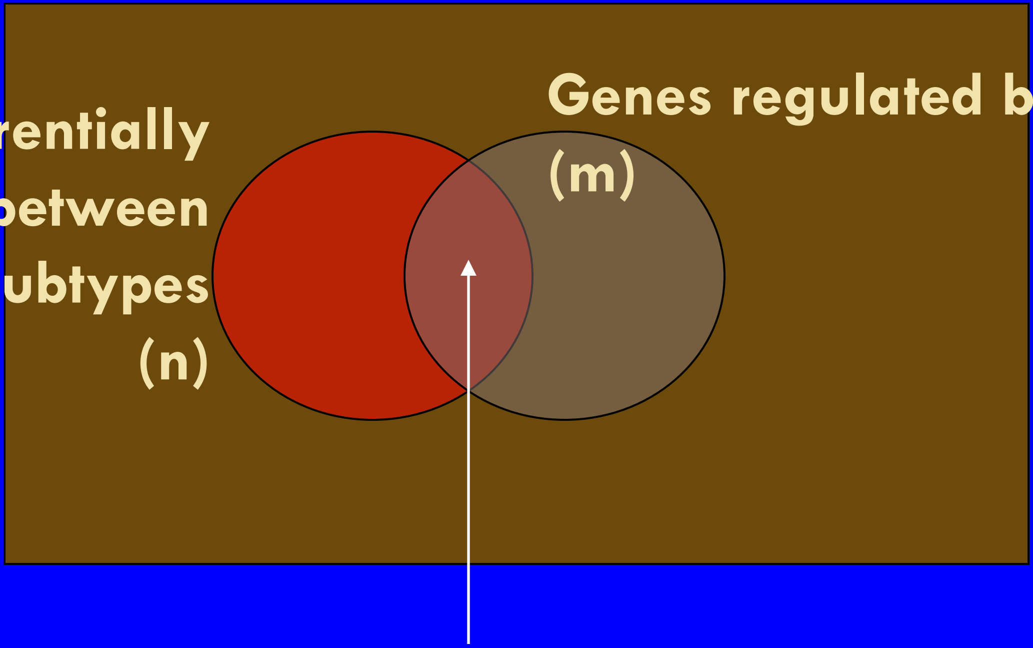


Set of genes differentially expressed between the two subtypes of patients

Source: DOI:10.1074/mcp.M700487-MCP200

3. Combine motif analysis and gene expression data

All genes (N)



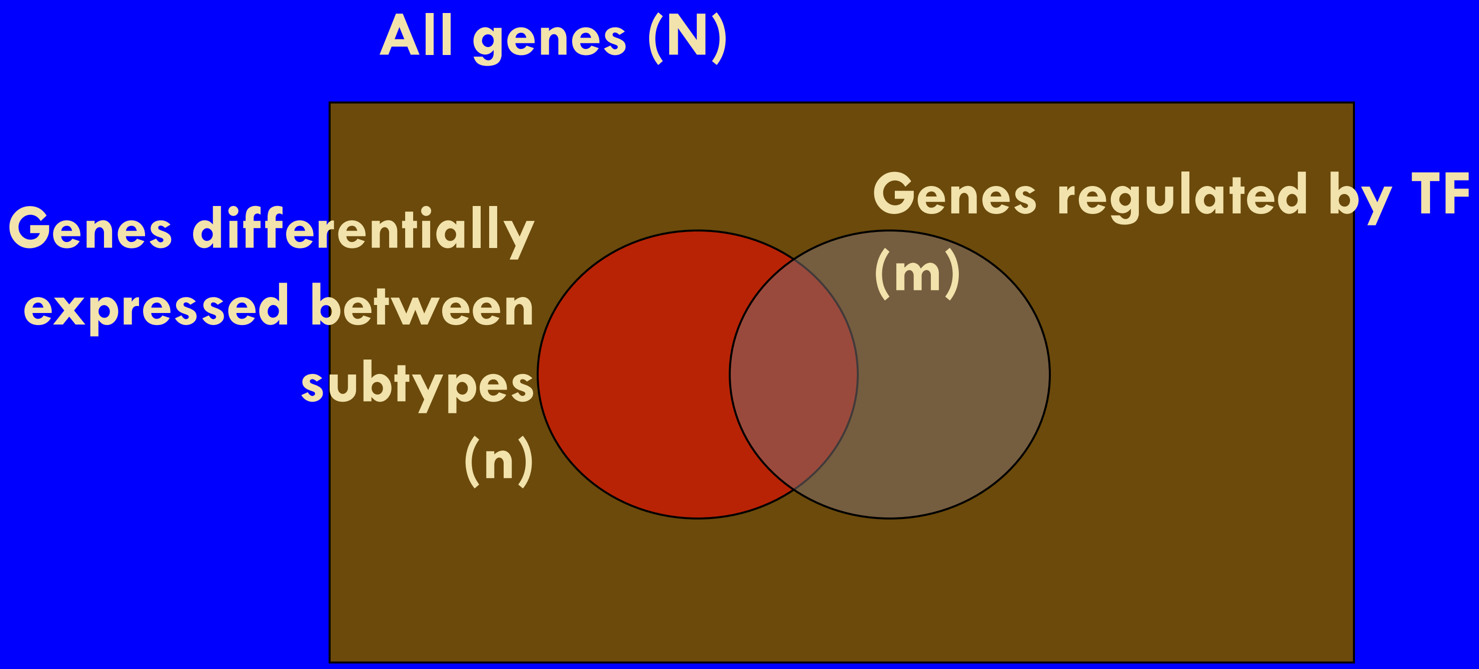
Genes differentially expressed between subtypes (n)

Genes regulated by TF (m)

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



All genes (N)

Genes differentially expressed between subtypes (n)

Genes regulated by TF (m)

Infer: TF may be determining cancer subtypes.

An “association” between motif and condition

Useful tools

- GREAT: http://bejerano.stanford.edu/great/public/html/
  - Input a set of genomic segments (e.g., ChIP peaks)
  - Obtain what annotations enriched in nearby genes
  - only for human, mouse and zebrafish

- DAVID: https://david.ncifcrf.gov/
  - Input a set of genes
  - Obtain what annotations enriched in those genes
  - Many different species

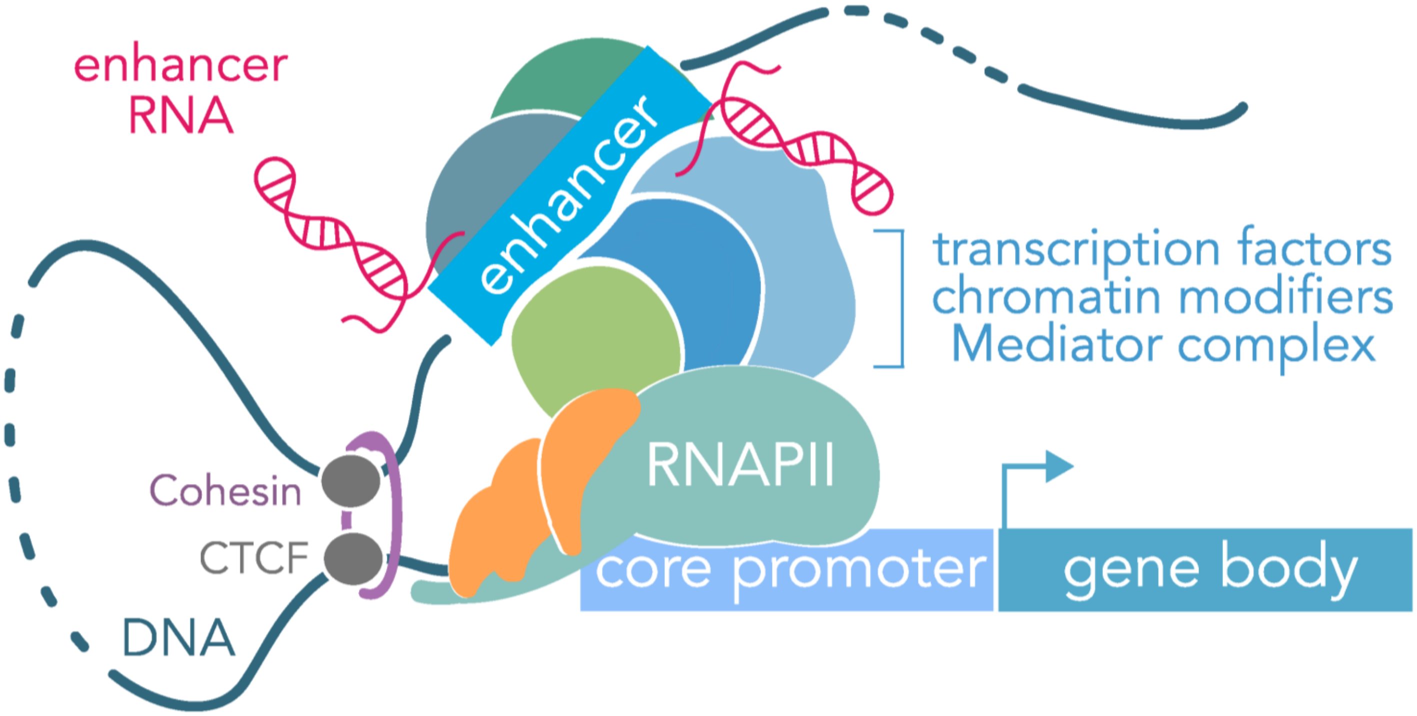
Epigenomics

- 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

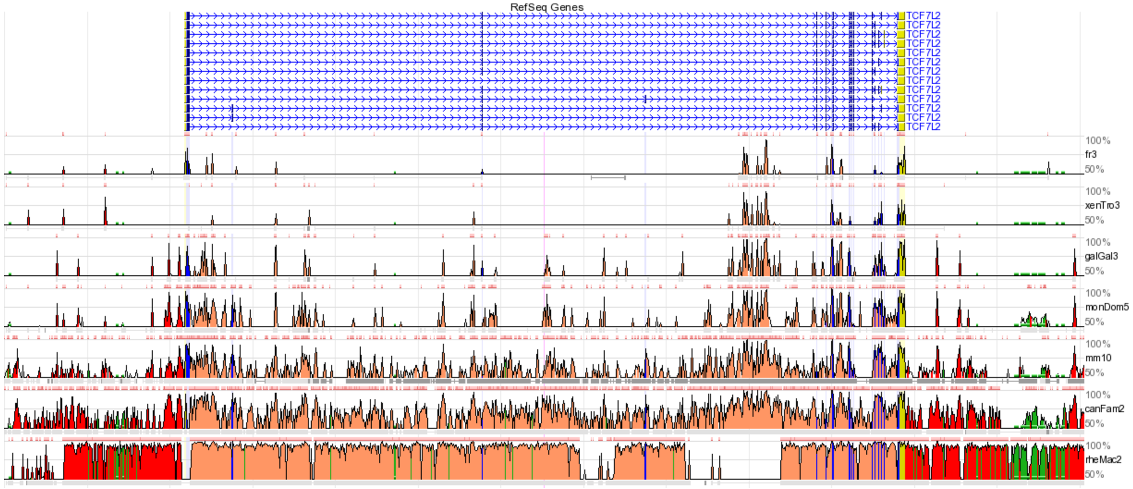


Source: Genomic Enhancers in Brain Health and Disease. Nancy V. N. Carullo and Jeremy J. Day. Genes 2019, 10(1), 43;

How to find enhancers?

- Like finding needle in a haystack

- Evolutionary conservation is sometimes used to identify enhancers

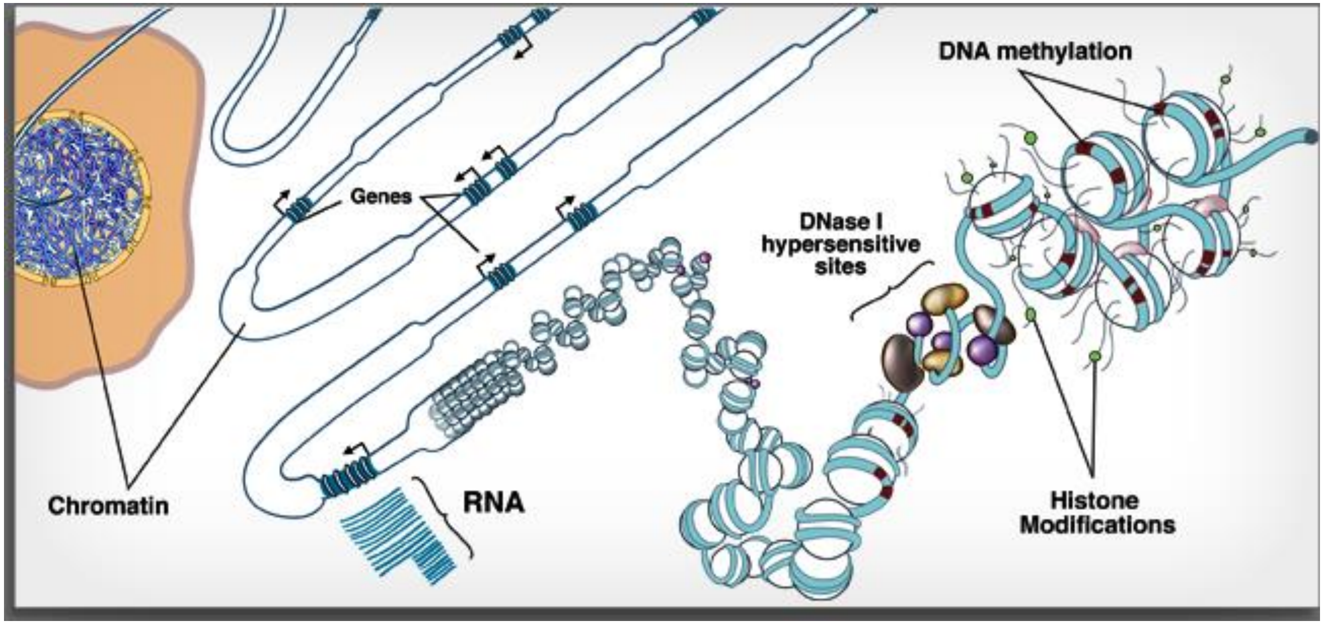


  - but not all functional elements are conserved at the level that DNA sequence alignments can detect. So how do we find regulatory elements?

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

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

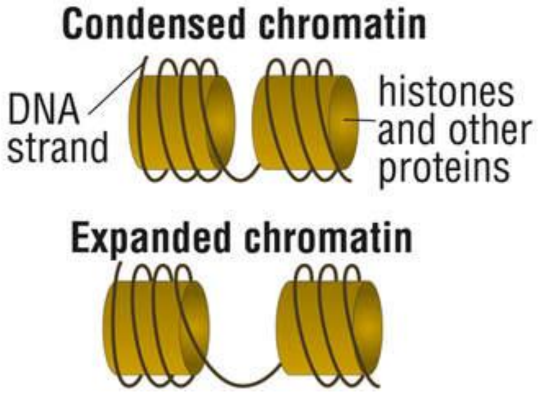
Eukaryotic genomes are complex 3D structures comprised of modified and unmodified DNA, RNA and many types of interacting proteins



- 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

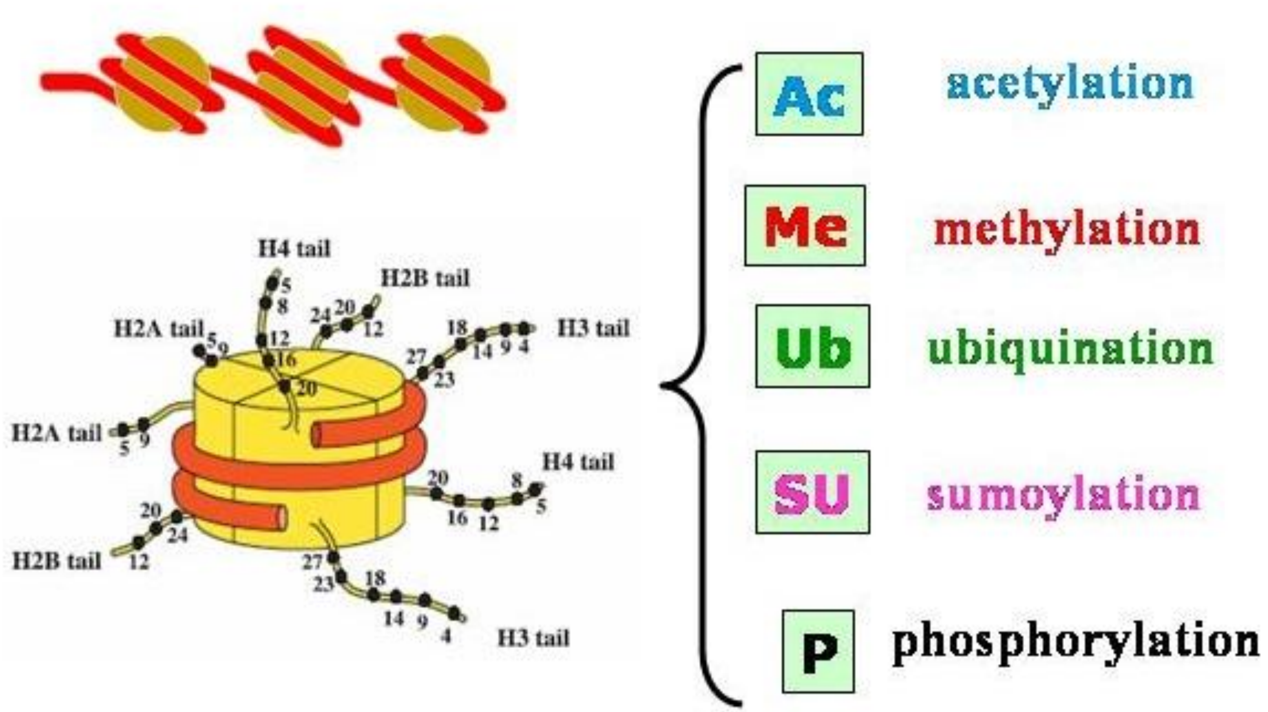


Not accessible

accessible

- 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

All four histones in the tetramer have "tails" that can be modified in various ways

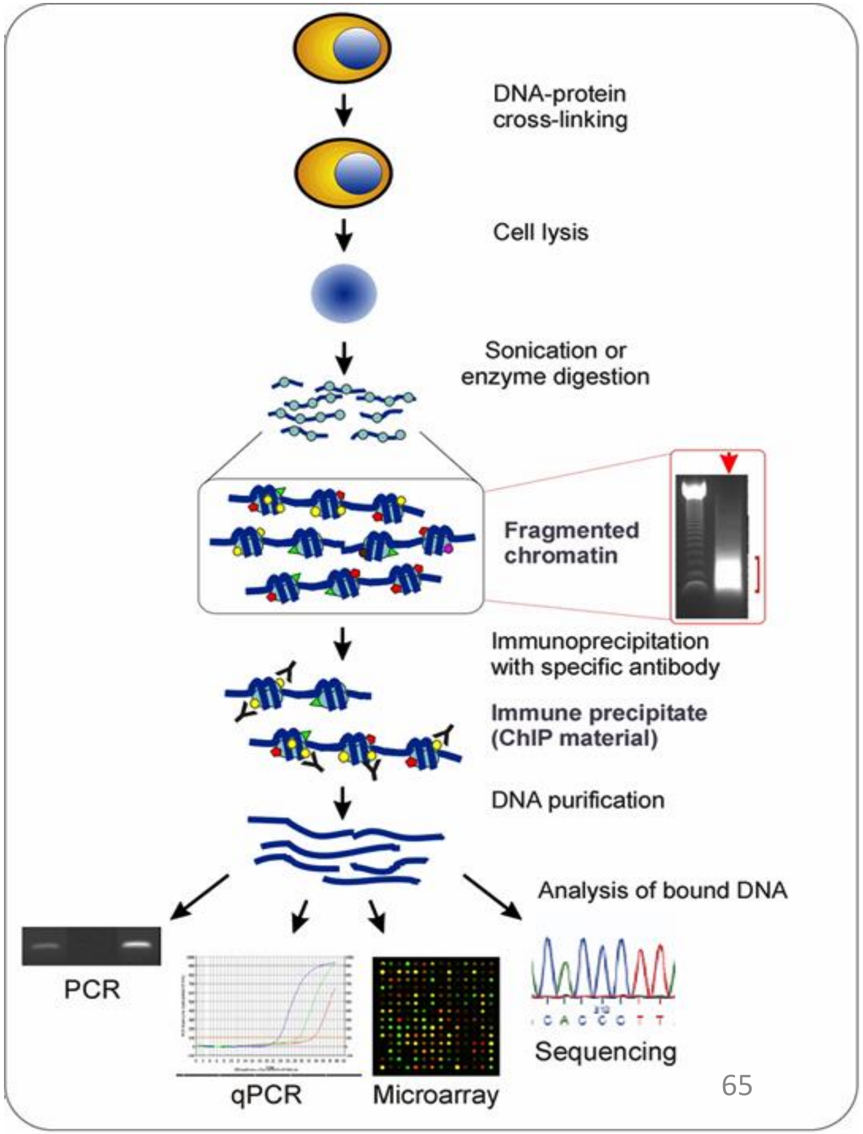


The most consequential modifications, with respect to transcriptional activity, appear to involve methylation or acetylation of Lysines (K) in histone H3

Experimental assays

How to find TF binding sites? 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 highly enriched in binding sites for a particular protein
- Requires availability of excellent **antibodies** that can detect the protein in its *in vivo* context


DNA-protein cross-linking
Cell lysis
Sonication or enzyme digestion
Fragmented chromatin
Immunoprecipitation with specific antibody
Immune precipitate (ChIP material)
DNA purification
Analysis of bound DNA
PCR
qPCR
Microarray
Sequencing

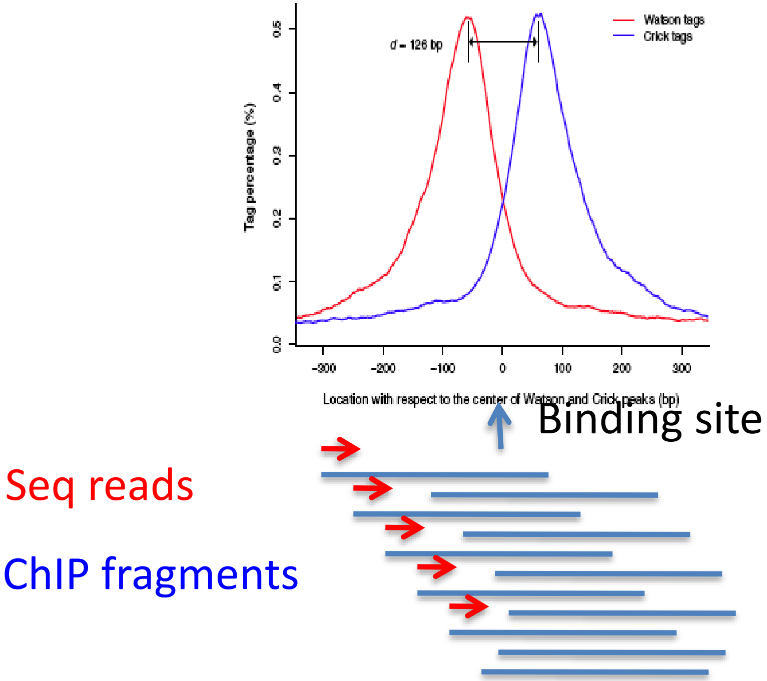
ChIP computational issues

- First step is to map reads: BOWTIE, Novalign, 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



Binding site

Seq reads

ChIP fragments

ChIP Analytical challenges

- Genomic neighborhoods
  - Shear efficiency is not really “random”
    - Some genomic regions are fragile and sensitive
    - Some regions are protected from shear or degradation
  - Other artifacts
    - Centromeres, polymorphic regions, repeats in general: most programs cannot manage sequence reads that are not mapped uniquely

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

Analyzing ChIP data

- User-friendly tools
  - MACS:
    - Zhang et al, *Genome Biology* 2008, Feng et al. 2012, *Nat Procols* PMID: 22936215 (Xiaole Liu lab);
    - **MACS1** is best for sharp peaks (TFs); will break diffuse peaks into smaller regions
    - **MACS2** is designed to allow broad- or sharp-peak detection
  - HOMER (http://homer.salk.edu/homer)
    - Can be easily tweaked for more sensitive peak detection
    - Comes packaged wiith a rich set of peak annotation tools
    - Tools for DNAse-seq, Hi-C, differential ChIP analysis and many more
  - Both tools permit generation of "wiggle files" or similar that can be viewed in the UCSC browser
    - Looking at your data is a very important step! Peak finders can miss peaks that you can easily see by eye!

ChIP analysis workflow

FASTQC -> BOWTIE -> Peak finder (MACS or HOMER)
This same workflow and tools can be used for a variety of assays
e.g., ATAC-seq, DNase seq, etc.

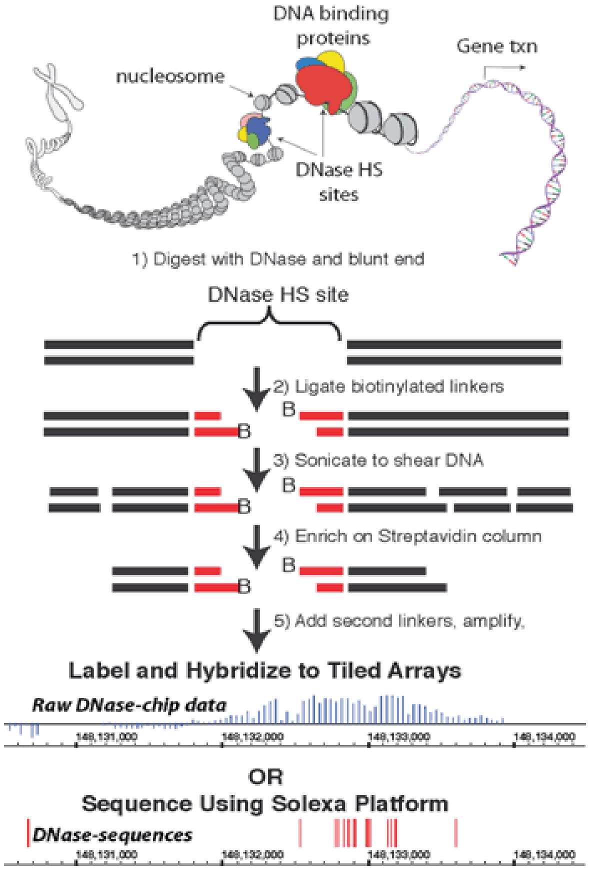
Downstream analysis:
Mapping peaks to nearby genes (and perhaps, differentially expressed genes)

Identifying enriched motifs
For your factor
For co-binding factors

Overlapping with other genome features
e.g., open chromatin, known binding sites, etc.

How to find accessible DNA?

High-throughput methods to identify DNaseI HS sites.



The first approach:

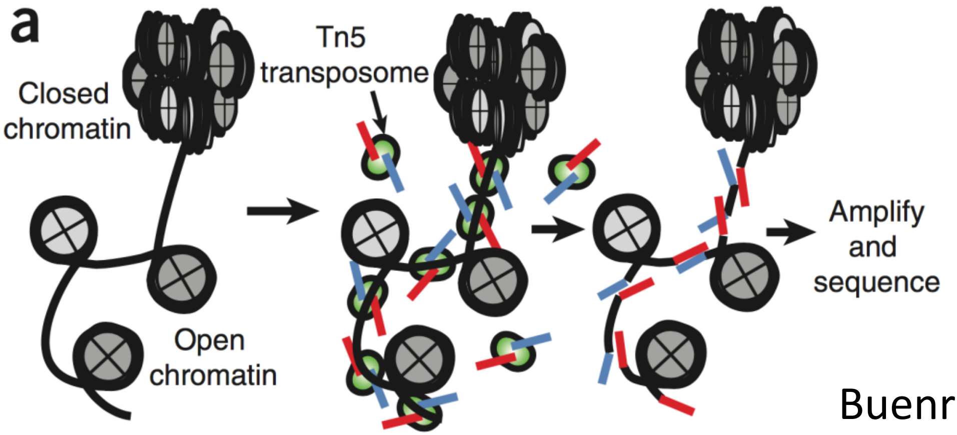
Crawford et al., Genome Research 16:123, 2006 (Francis Collins' laboratory)

Genome-wide identification of Dnase I Hypersensititive sites (DHS)

Later variants also based on DNase I treatment, but different protocol and different philosophy. See http://homer.ucsd.edu/homer/ngs/dnase/index.html

Many later methods: ChIP-exo, FAIRE, ATAC-seq etc. (see Furey et al., 2012 for older review)

An economical approach to open chromatin: ATAC-seq



Buenrostro et al., 2013, 2015

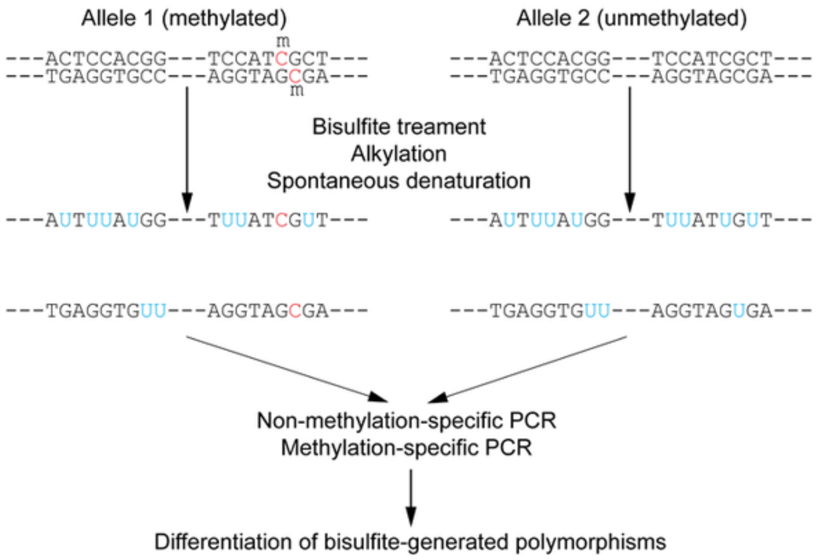
- Uses Tn5 transposase and a Transposon modified to contain Illumina primers at each end
- Transposon “jumps” preferentially (and randomly) into accessible chromatin
- Because of the design the transposon breaks DNA where it jumps in, tagging the site with the primer
- Two insertions close together yield 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** (ligation etc), can be done with small amounts of input chromatin – e.g. 50,000 vs 1,000,000 cells

DNA Methylation

- Methyl ($-\mathrm{CH3}$) 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


Allele 1 (methylated)
Allele 2 (unmethylated)
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
m
---ACTCCACGG---TCCATCGCT---
---TGAGGTGCC---AGGTAGCGA---
Bisulfite treament
Alkylation
Spontaneous denaturation
---AUTUUAUGG---TUUATCGUT---
---AUTUUAUGG---TUUATUGUT---
---TGAGGTGUU---AGGTAGCGA---
---TGAGGTGUU---AGGTAGUGA---
Non-methylation-specific PCR
Methylation-specific PCR
Differentiation of bisulfite-generated polymorphisms

Other methods:

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

Insights from large scale epigenomics studies

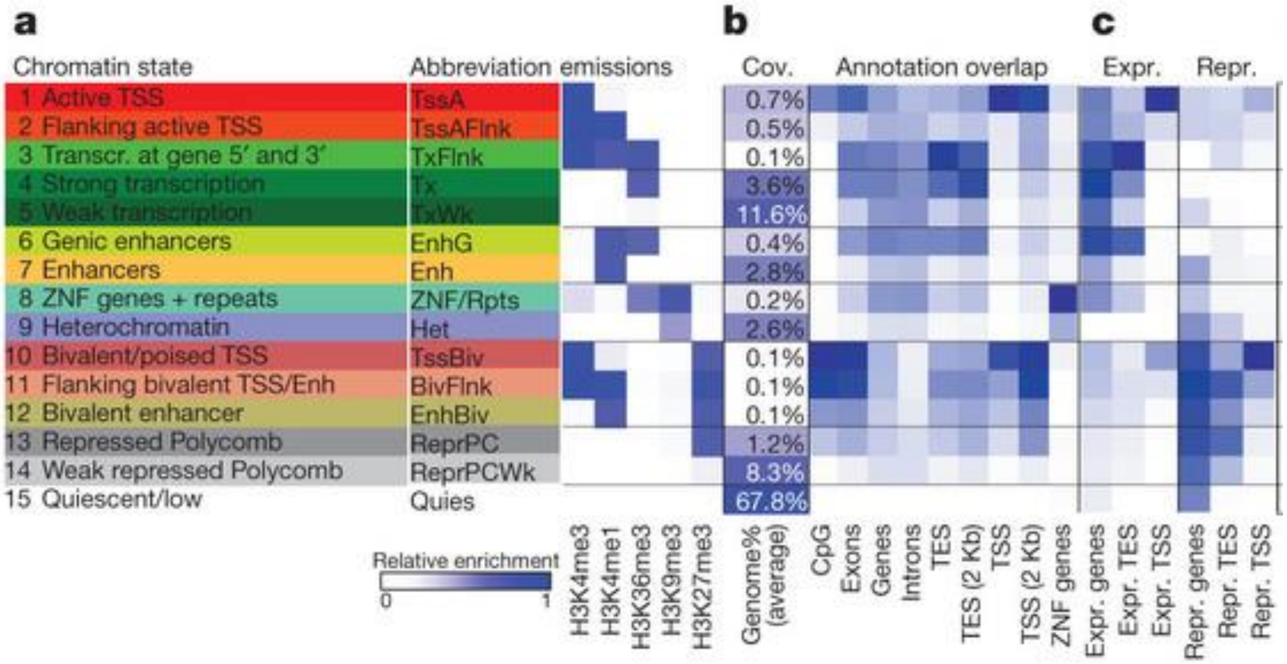
Lessons from epigenomics assays

- 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.
  - 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 *without* H3K4me3 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 but, not all H3K4me1 marks are in enhancers
- Other marks: H3K27Ac or H3K27me3
  - Mutually exclusive marks for open (Ac) versus closed (Me3) chromatin regions
  - H3K27Ac is perhaps the most general mark of open chromatin: promoters and enhancers
  - H3K27Ac often found in combination with H3K4 me1/me3

Application 1: Chromatin "states": an unbiased, systematic characterization

- ChromHMM tool combines information from 38 different histone marks, Pol2 and CTCF profiles to identify different 'states'

- Other tools exist, e.g., ChromaSig, Segway

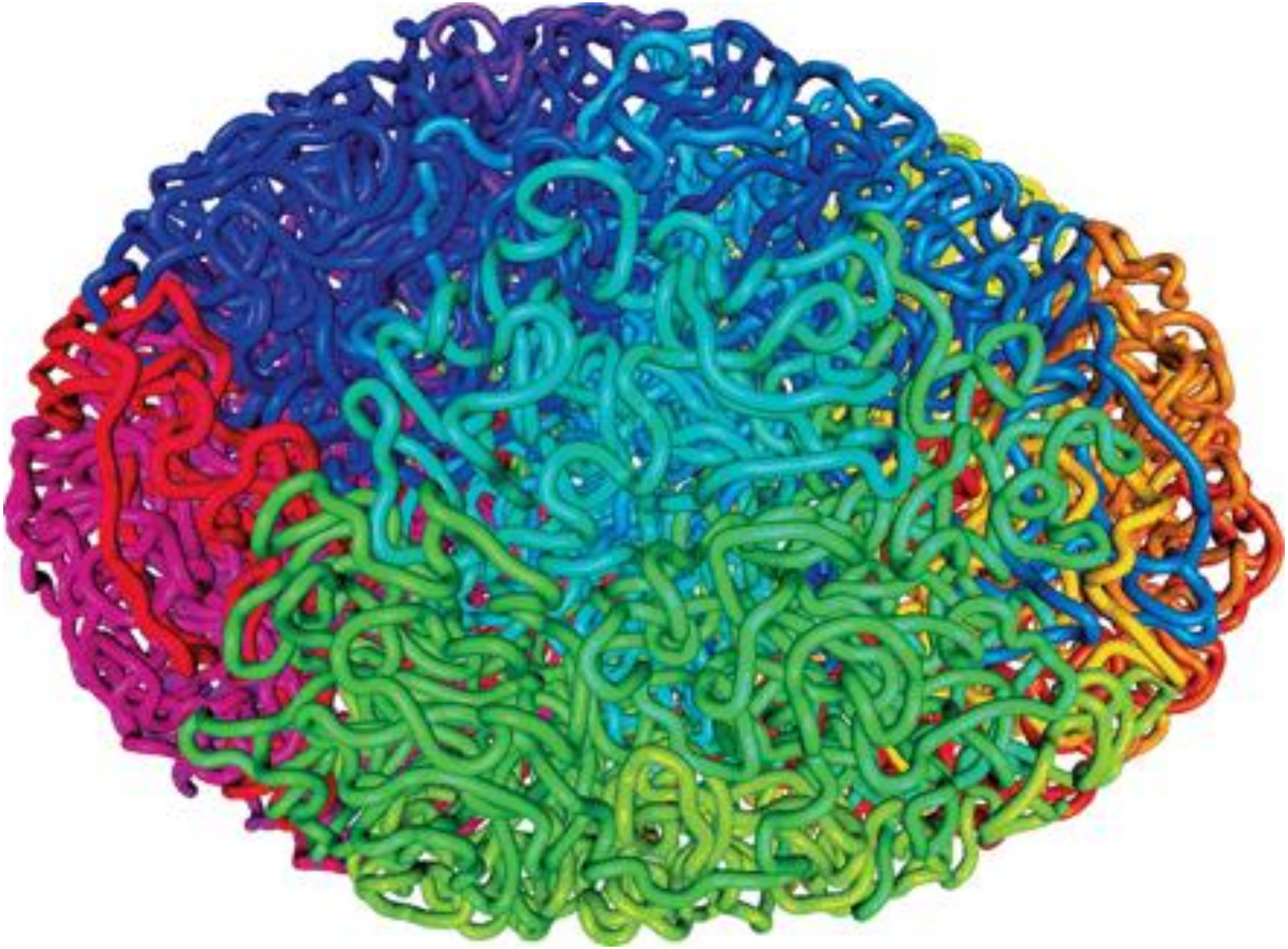


ChromHMM: automating chromatin-state discovery and characterization. Jason Ernst & Manolis Kellis. Nature Methods 9, 215–216 (2012) http://www.ncbi.nlm.nih.gov/pubmed/22373907

Application 2: DNA Methylation profiles in cancer and aging

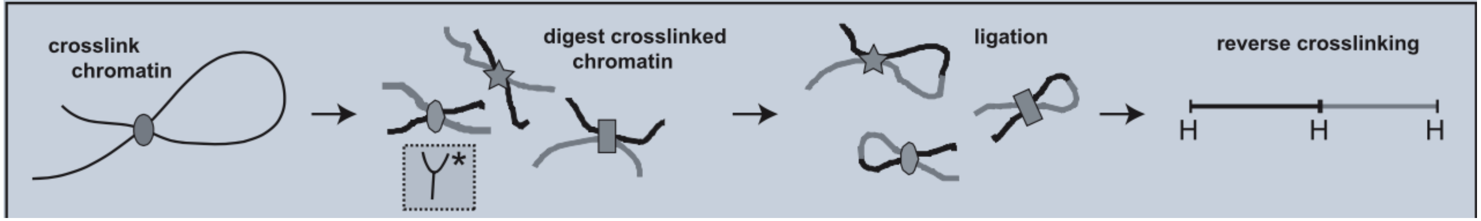
- DNA Methylation levels can be condition-dependent
  - Aberrant methylation patterns in cancer (e.g., hypermethylation of tumor suppressors and hypomethylation of oncogenes)
  - Progressive increase in global methylation levels with age. Also aging-correlated hypomethlation at some genes.

3D genome



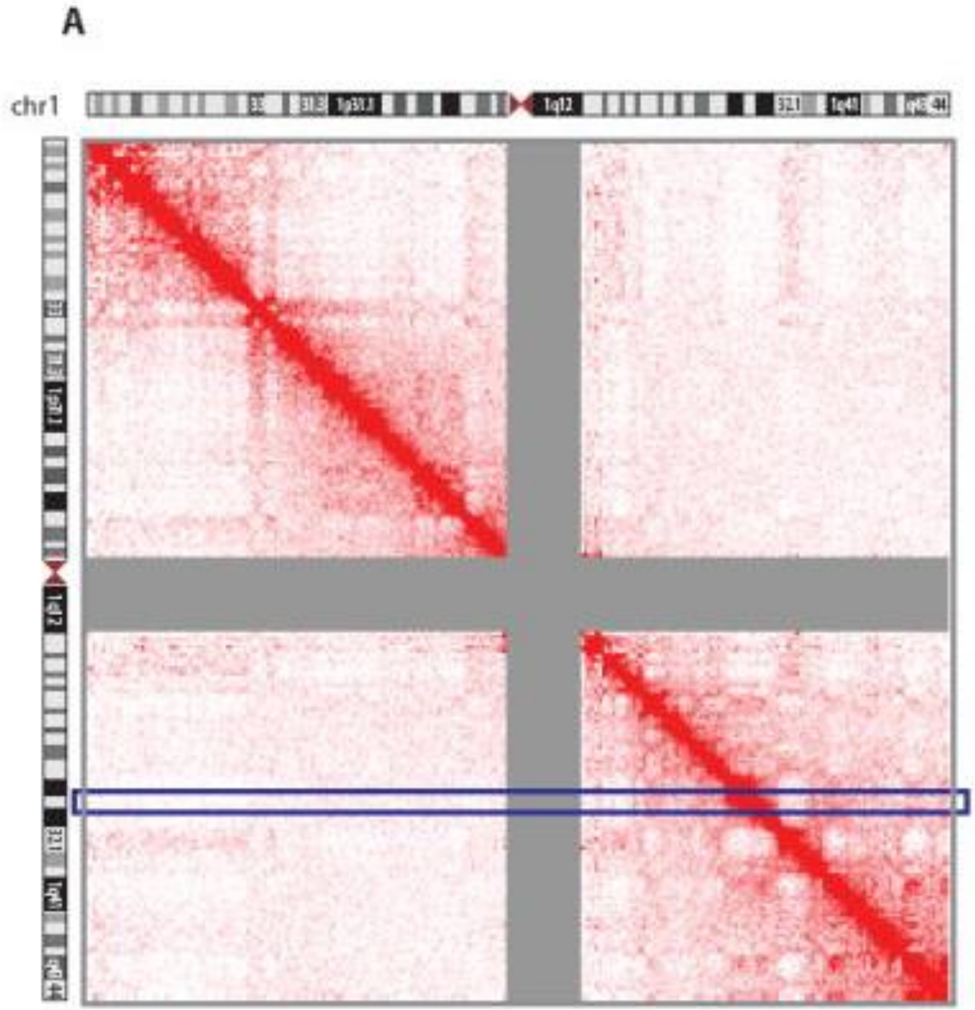
Source: DOI: 10.1126/science.347.6217.10

Probing 3-dimensional chromatin structure with conformation capture

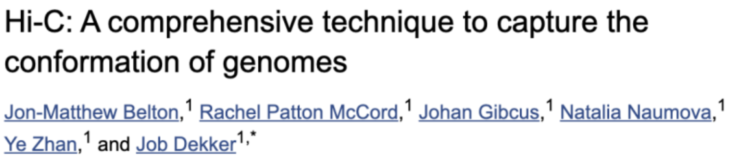


from Wit and de Laat, 2012

Hi-C “output”



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.


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

Jon-Matthew Belton,[1] Rachel Patton McCord,[1] Johan Gibcus,[1] Natalia Naumova,[1] Ye Zhan,[1] and Job Dekker[1,*]

Requires analysis methods that are different from ChIP

- **Provides the essential “big picture” view, since it is otherwise impossible to predict long-range enhancer-enhancer or enhancer-promoter interactions**

- Sequenced fragments contain a bit of DNA from two distant regions
  - Data need to be trimmed and mapped to allow non-contiguous sequences

- Long-distant contacts are numerous, and each contact point is relatively rare: peaks are small and require deep sequencing

- Hi C kits are now readily available and quite reliable, giving a whole-genome view of interactions
  - Lots of interactions and lots of noise! Computational issues are tricky
  - All 3D methods require deep sequencing and paired-end reads

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 ?