

# Lecture 13. Methyl-seq, ChipSeq, and HiC

Michael Schatz

March 11, 2019

JHU 601.749: Applied Comparative Genomics



# Assignment 5: Due Monday March 11

## Assignment 5: RNA-seq and differential expression

Assignment Date: Monday, March 4, 2019

Due Date: Monday, March 11, 2019 @ 11:59pm

### Assignment Overview

In this assignment, you will analyze gene expression data and learn how to make several kinds of plots in the environment of your choice. (We suggest Python or R.) Make sure to show your work/code in your writeup! As before, any questions about the assignment should be posted to [Piazza](#).

#### Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here: [ftp://ftp.ensembl.org/pub/release-87/gtf/homo\\_sapiens/Homo\\_sapiens.GRCh38.87.gtf.gz](ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz)

- Question 1a. How many annotated protein coding genes are on each autosome of the human genome? [Hint: Protein coding genes will have "gene" in the 3rd column, and contain the following text: gene\_biotype "protein\_coding"]
- Question 1b. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes? [Hint: use the genes identified in 1b]
- Question 1c. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform for each protein coding gene]

#### Question 2. Time Series [10 pts]

[This file](#) contains pre-normalized expression values for 100 genes over 10 time points. Most genes have a stable background expression level, but some special genes show increased expression over the timecourse and some show decreased expression.

- a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical clustering are common clustering algorithms you could try.]
- b. Calculate the first two principal components of the expression matrix. Show the plot and color the points based on their cluster from part (a). Does the PC1 axis, PC2 axis, neither, or both correspond to the clustering?
- c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.

#### Question 3. Sampling Simulation [10 pts]

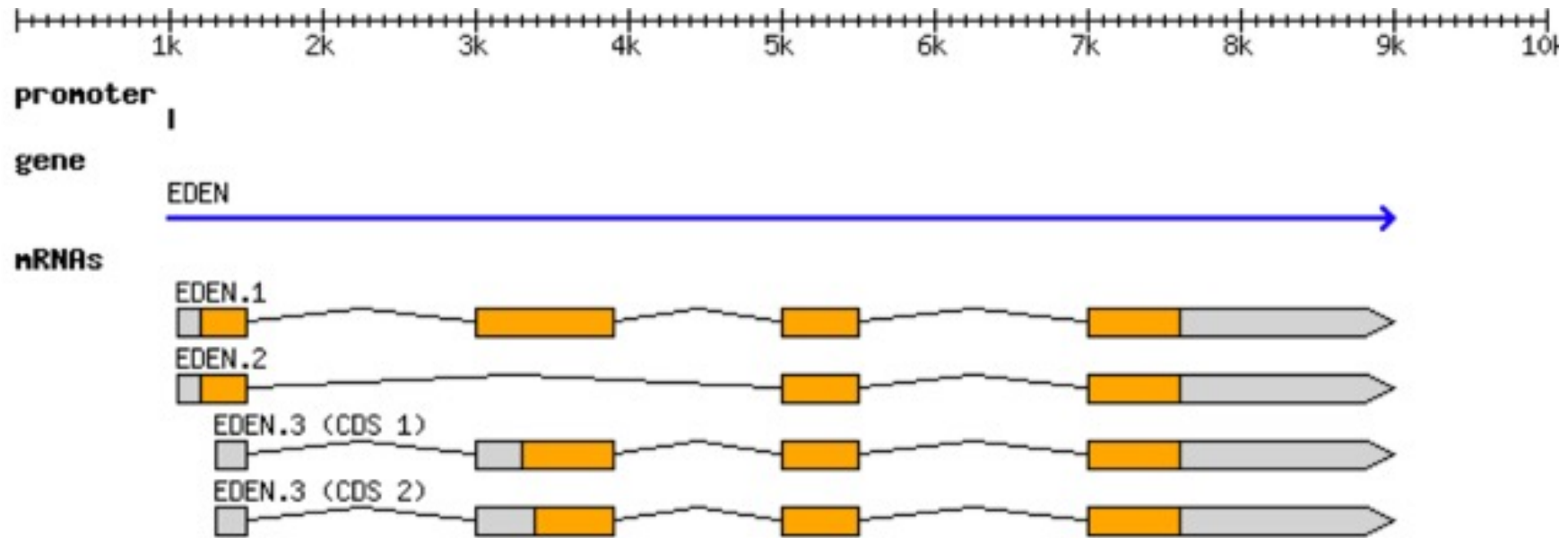
A typical human cell has ~250,000 transcripts, and a typical bulk RNA-seq experiment may involve millions of cells. Consequently in an RNAseq experiment you may start with trillions of RNA molecules, although your sequencer will only give a few million to billions of reads. Therefore your RNAseq experiment will be a small sampling of the full composition. We hope the sequences will be a representative sample of the total population, but if your sample is very unlucky or biased it may not represent the true distribution. We will explore this concept by sampling a small subset of transcripts (1000 to 5000) out of a much larger set (1M) so that you can evaluate this bias.

In [data1.txt](#) with 1,000,000 lines we provide an abstraction of RNA-seq data where normalization has been performed and the number of times a gene name occurs corresponds to the number of transcripts sequenced.

- a. Randomly sample 1000 rows. Do this simulation 10 times and record the relative abundance of each of the 15 genes. Plot the mean vs. variance.
- b. Do the same sampling experiment but sample 5000 rows each time. Again plot the mean vs. variance.
- c. Is the variance greater in (a) or (b)?, and explain why. What is the relationship between abundance and variance?



# Gene Models



- “Generic Feature Format” (GFF) records genomic features
  - Coordinates of each exon
  - Coordinates of UTRs
  - Link together exons into transcripts
  - Link together transcripts into gene models

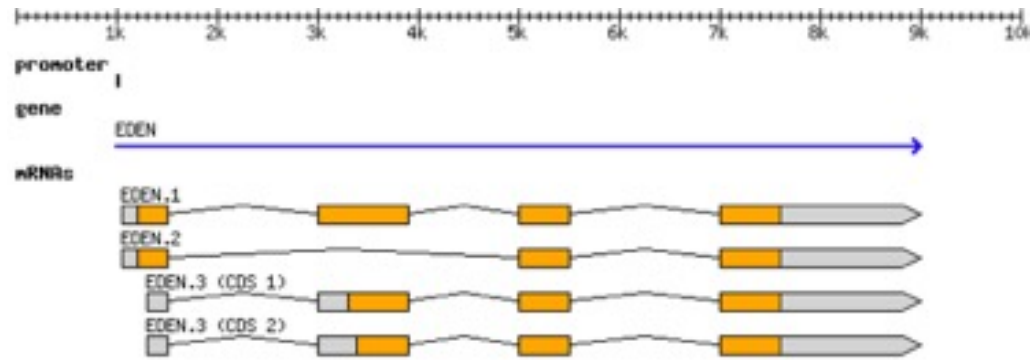
# GFF File format

GFF3 files are nine-column, tab-delimited, plain text files

- 1. *seqid*:** The ID of the sequence
- 2. *source*:** Algorithm or database that generated this feature
- 3. *type*:** *gene/exon/CDS/etc...*
- 4. *start*:** 1-based coordinate
- 5. *end*:** 1-based coordinate
- 6. *score*:** E-values/p-values/index/colors/...
- 7. *strand*:** “+” for positive “-” for minus, “.” not stranded
- 8. *phase*:** For "CDS", where the feature begins with reference to the reading frame (0,1,2)
- 9. *attributes*:** A list of tag=value features
  - Parent: Indicates the parent of the feature (group exons into transcripts, transcripts into genes, ...)

# GFF Example

Gene “EDEN” with 3 alternatively spliced transcripts, isoform 3 has two alternative translation start sites



```
##gff-version 3
##sequence-region    ctg123 1 1497228
ctg123 . gene         1000  9000  .  +  .  ID=gene00001;Name=EDEN

ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

ctg123 . mRNA         1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA         1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA         1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3

ctg123 . exon         1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon         1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon         3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon         5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon         7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003

ctg123 . CDS          1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS          3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS          5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS          7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1

ctg123 . CDS          1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS          5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS          7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2

ctg123 . CDS          3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS          5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS          7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3

ctg123 . CDS          3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS          5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS          7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```



# Proposal: Due Friday March 15

## Project Proposal

---

Assignment Date: Wednesday March 6, 2019

Due Date: Friday, March 15, 2019 @ 11:59pm

Review the [Project Ideas](#) page

Work solo or form a team for your class project (no more than 3 people to a team).

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)

Submit the proposal as a single page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at

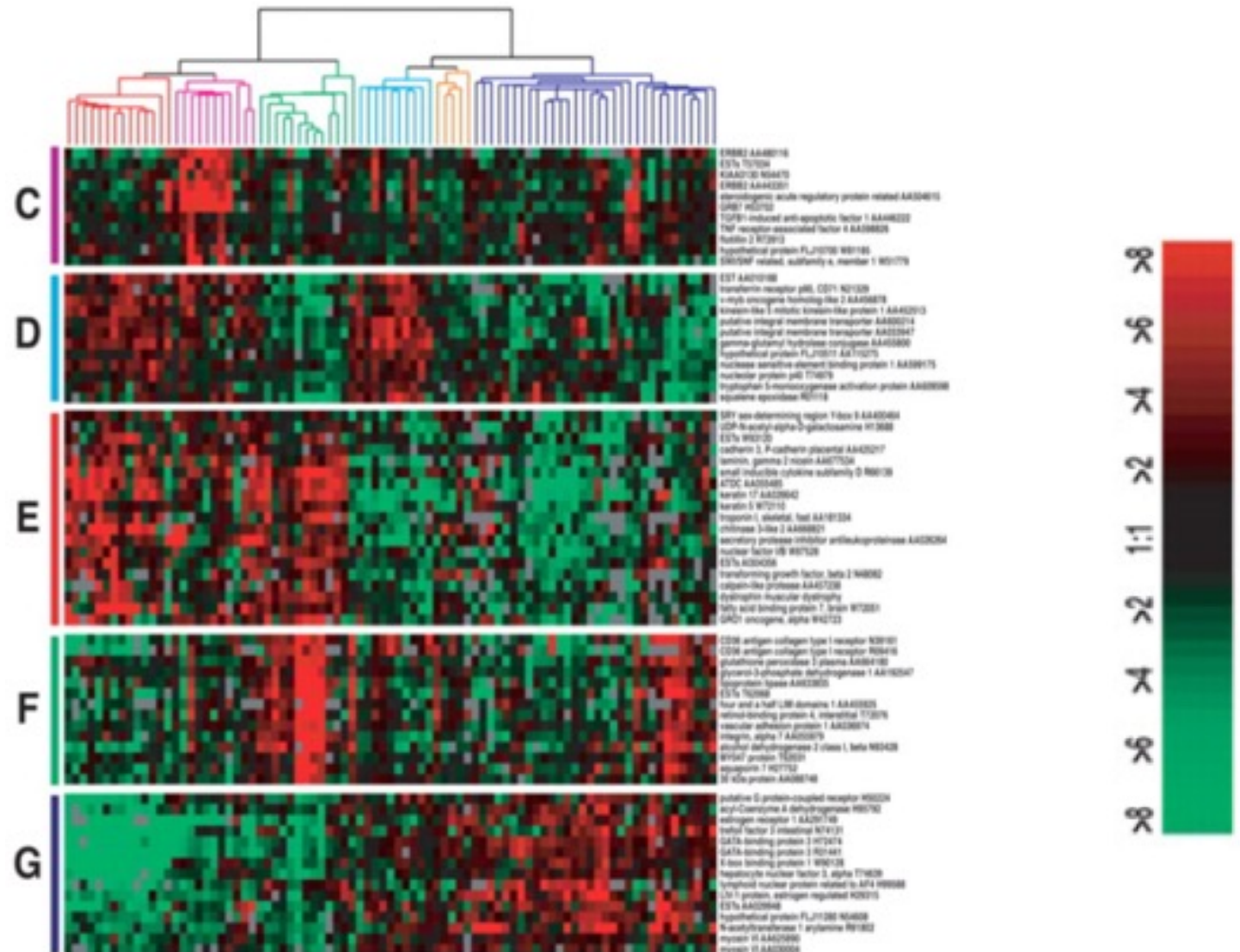
[https://academic.oup.com/bioinformatics/pages/submission\\_online](https://academic.oup.com/bioinformatics/pages/submission_online)

Please use Piazza to coordinate proposal plans!



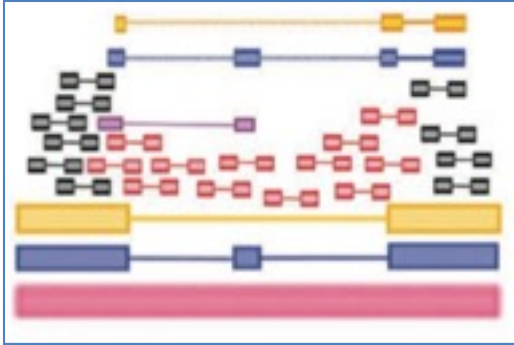


# RNA-seq



**Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.**  
 Sørli et al (2001) *PNAS*. 98(19):10869-74.

# RNA-seq Challenges

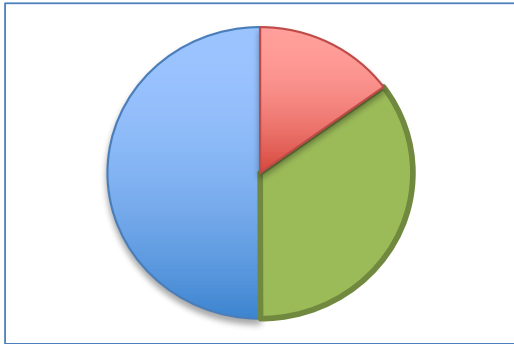


## Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

### TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111

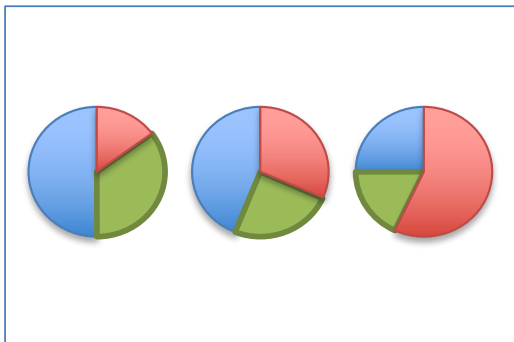


## Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

### Transcript assembly and quantification by RNA-seq

Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



## Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

### RNA-seq differential expression studies: more sequence or more replication?

Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

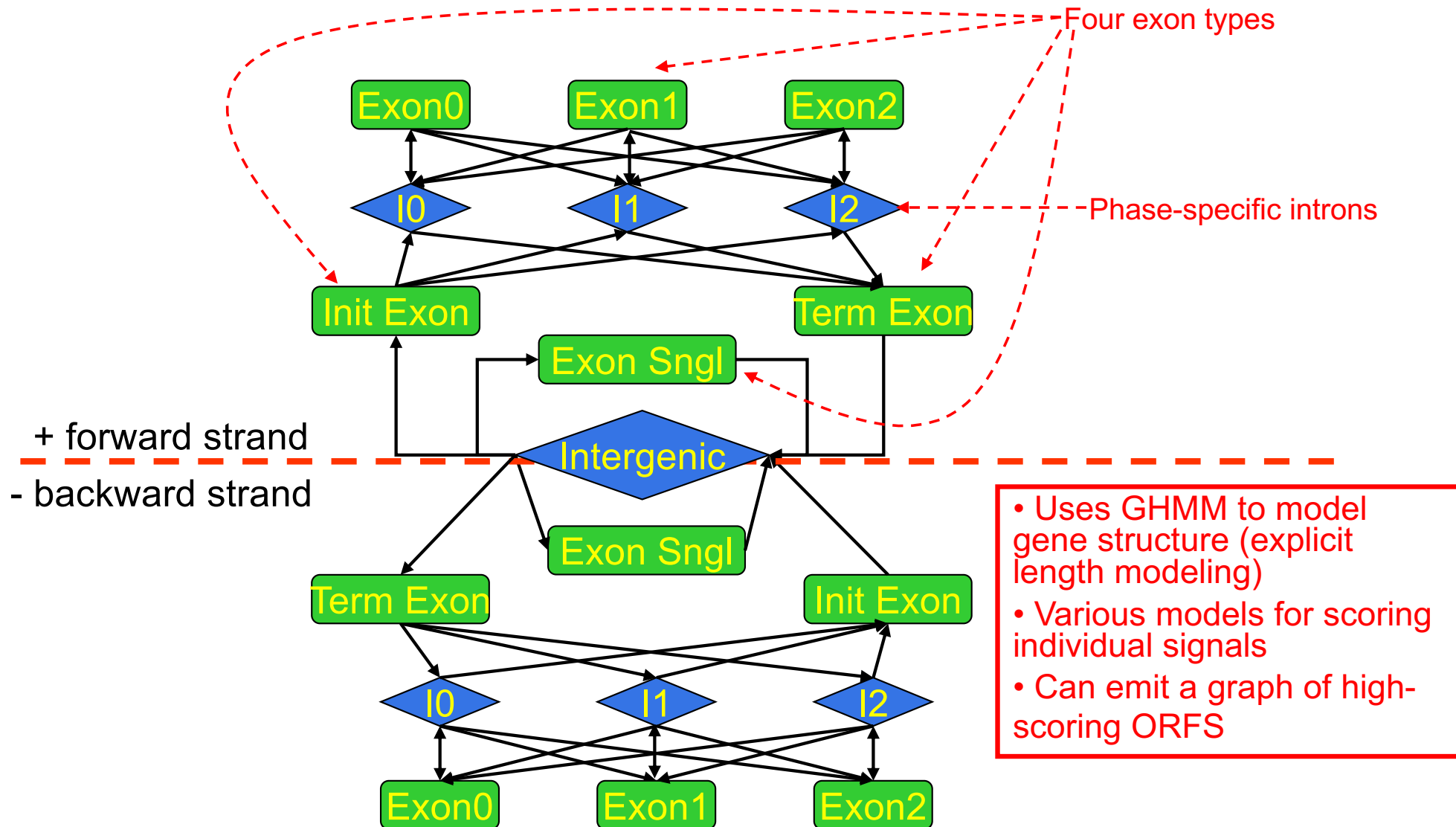




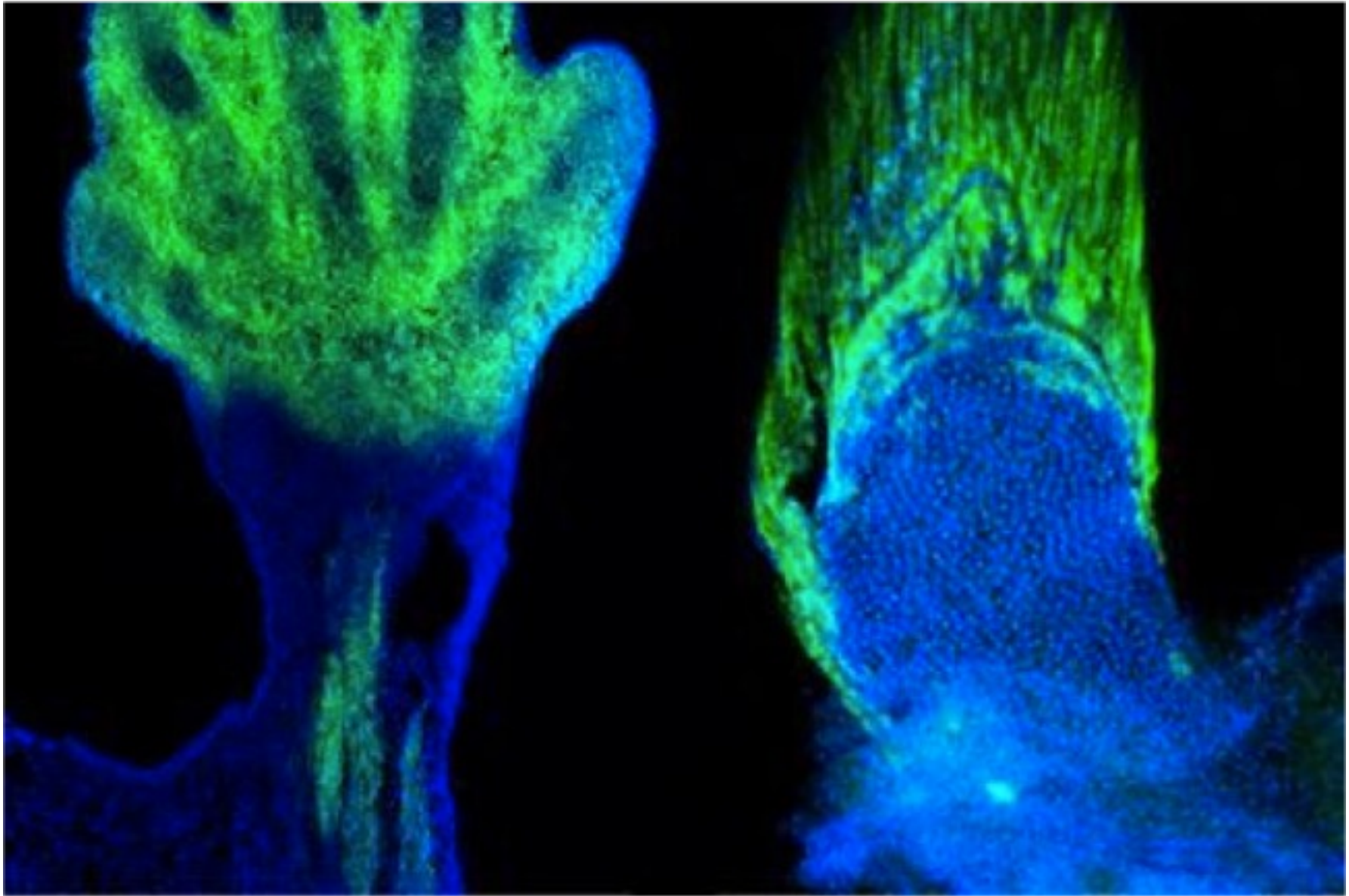
# Gene Identification Approaches

1. Experimental: RNAseq
2. Homology: Alignment to other genomes
3. Prediction: “Gene Finding”

# GlimmerHMM architecture



# Human Evolution

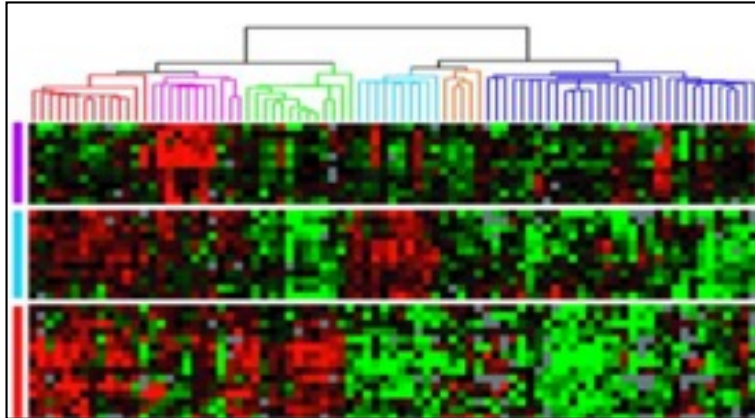


***Digits and fin rays share common developmental histories***

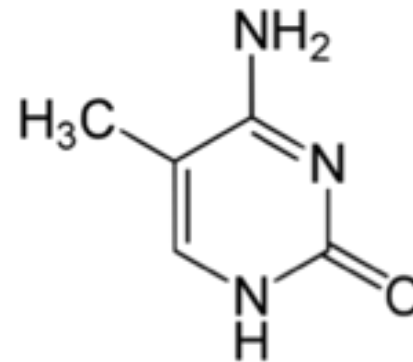
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

# \*-seq in 4 short vignettes

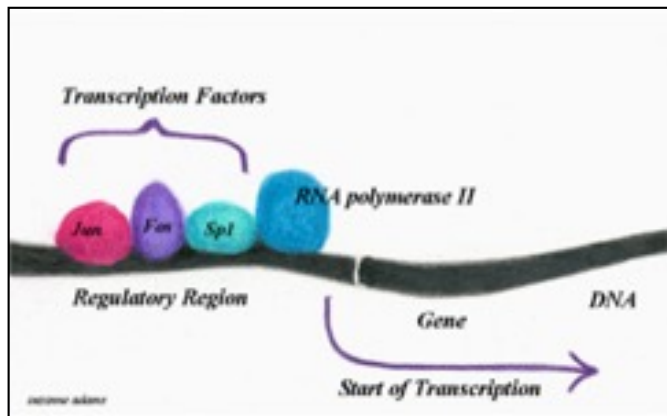
## RNA-seq



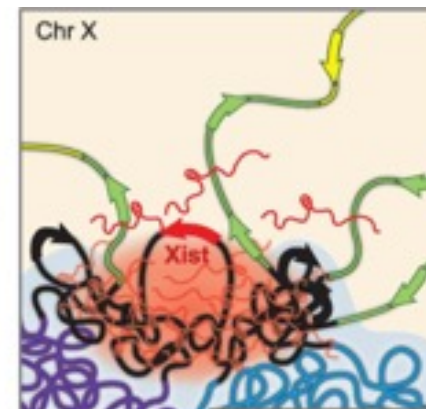
## Methyl-seq



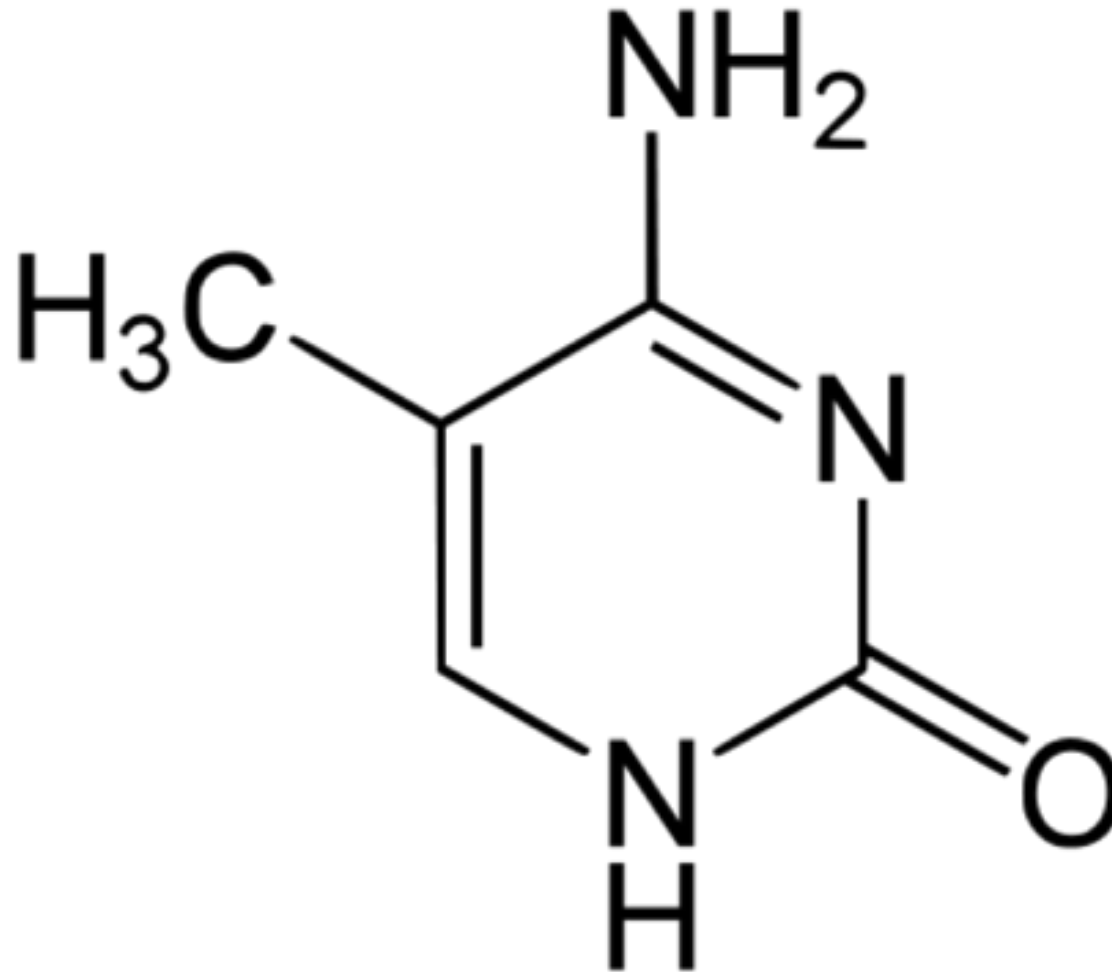
## ChIP-seq



## Hi-C



# Methyl-seq

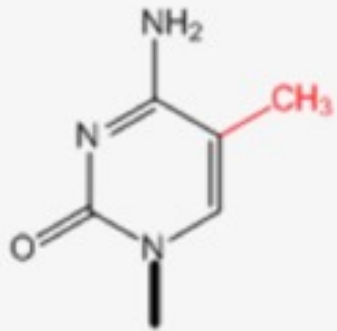


**Finding the fifth base: Genome-wide sequencing of cytosine methylation**

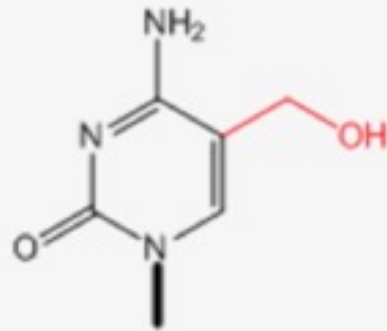
Lister and Ecker (2009) *Genome Research*. 19: 959-966



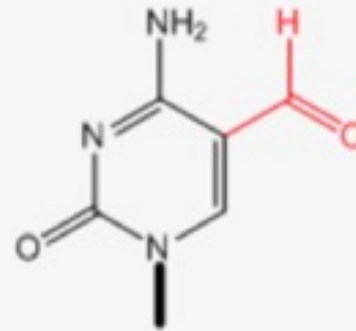
# Epigenetic Modifications to DNA



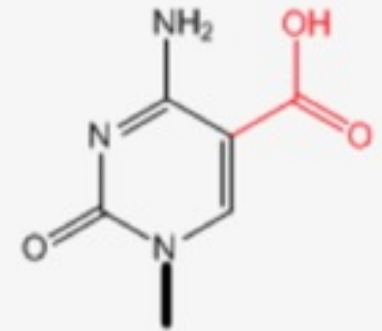
5-mC



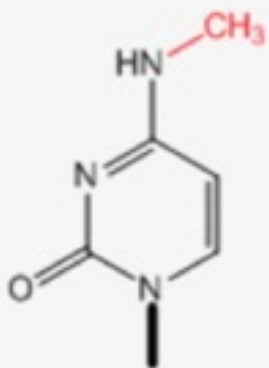
5-hmC



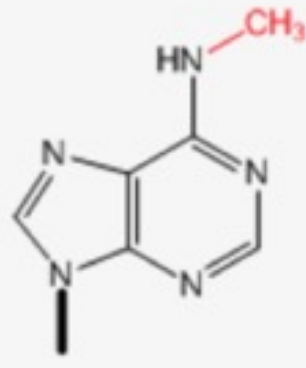
5-fC



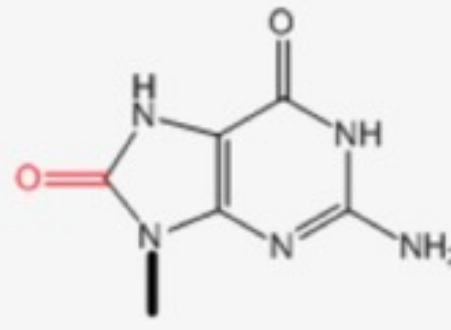
5-caC



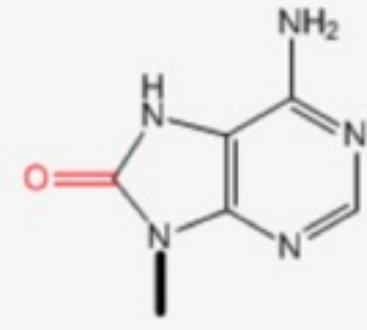
4-mC



6-mA



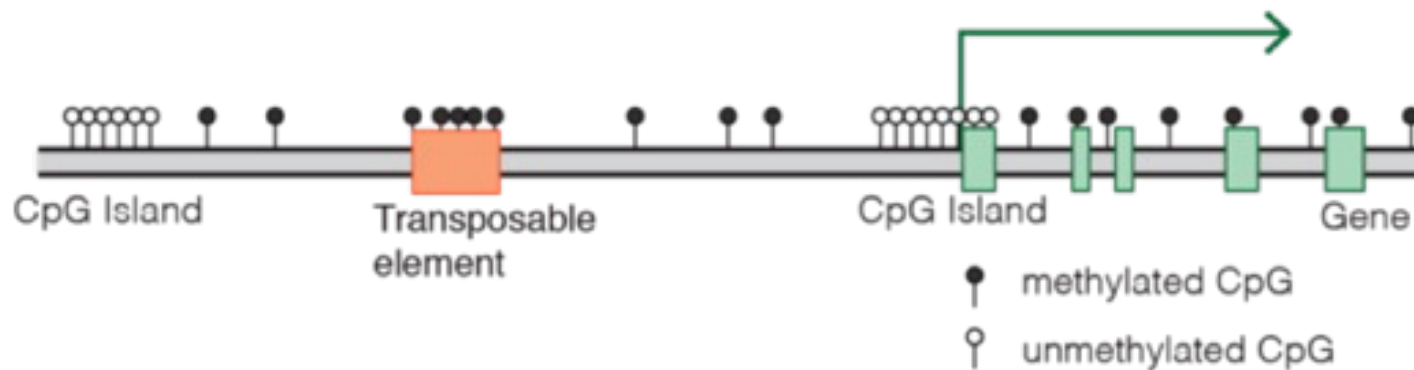
8-oxoG



8-oxoA

# Methylation of CpG Islands

Typical mammalian DNA methylation landscape



***CpG islands are (usually) defined as regions with***

- 1) a length greater than 200bp,
- 2) a G+C content greater than 50%,
- 3) a ratio of observed to expected CpG greater than 0.6

***Methylation in promoter regions correlates negatively with gene expression.***

- CpG-dense promoters of actively transcribed genes are never methylated
- In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene
- Methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene
- Methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) that can modify histones, thereby forming compact, inactive chromatin, termed heterochromatin.

# The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko<sup>1§</sup>, Sylvain Foret<sup>2§</sup>, Robert Kucharski<sup>3</sup>, Stephan Wolf<sup>4</sup>, Cassandra Falckenhayn<sup>1</sup>, Ryszard Maleszka<sup>3\*</sup>

**1** Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, **2** ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, **3** Research School of Biology, the Australian National University, Canberra, Australia, **4** Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany



*“The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently ensconced within the colony and uses a powerful chemical influence to exert control. Remarkably, these two female castes are generated from identical genomes. The key to each female's developmental destiny is her diet as a larva: future queens are raised on royal jelly. This specialized diet is thought to affect a particular chemical modification, methylation, of the bee's DNA, causing the same genome to be deployed differently.”*





**Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm**

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365





**Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm**

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365





Somaclonal variation arises in plants and animals when differentiated somatic cells are induced into a pluripotent state, but the resulting clones differ from each other and from their parents. In agriculture, somaclonal variation has hindered the micropropagation of elite hybrids and genetically modified crops, but the mechanism responsible remains unknown. The oil palm fruit 'mantled' abnormality is a somaclonal variant arising from tissue culture that drastically reduces yield, and has largely halted efforts to clone elite hybrids for oil production. Widely regarded as an epigenetic phenomenon, 'mantling' has defied explanation, but here we identify the MANTLED locus using epigenome-wide association studies of the African oil palm *Elaeis guineensis*. DNA hypomethylation of a LINE retrotransposon related to rice Karma, in the intron of the homeotic gene *DEFICIENS*, is common to all mantled clones and is associated with alternative splicing and premature termination. **Dense methylation near the Karma splice site (termed the Good Karma epiallele) predicts normal fruit set, whereas hypomethylation (the Bad Karma epiallele) predicts homeotic transformation, parthenocarpy and marked loss of yield.** Loss of Karma methylation and of small RNA in tissue culture contributes to the origin of mantled, while restoration in spontaneous revertants accounts for non-Mendelian inheritance. The ability to predict and cull mantling at the plantlet stage will facilitate the introduction of higher performing clones and optimize environmentally sensitive land resources.

**Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm**

Ong-Abdullah, et al (2015) *Nature*. doi:10.1038/nature15365



# Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein

Cell Structure and Function Laboratory, The Oncology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations<sup>1,2</sup>, rearrangements<sup>3-5</sup> and changes in methylation<sup>6-8</sup> have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9-11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture<sup>8,12-30</sup>. The results of these studies have varied; depending on the techniques and systems used, an increase<sup>12-19</sup>, decrease<sup>20-24</sup>, or no change<sup>25-29</sup> in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After *Hpa*II cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

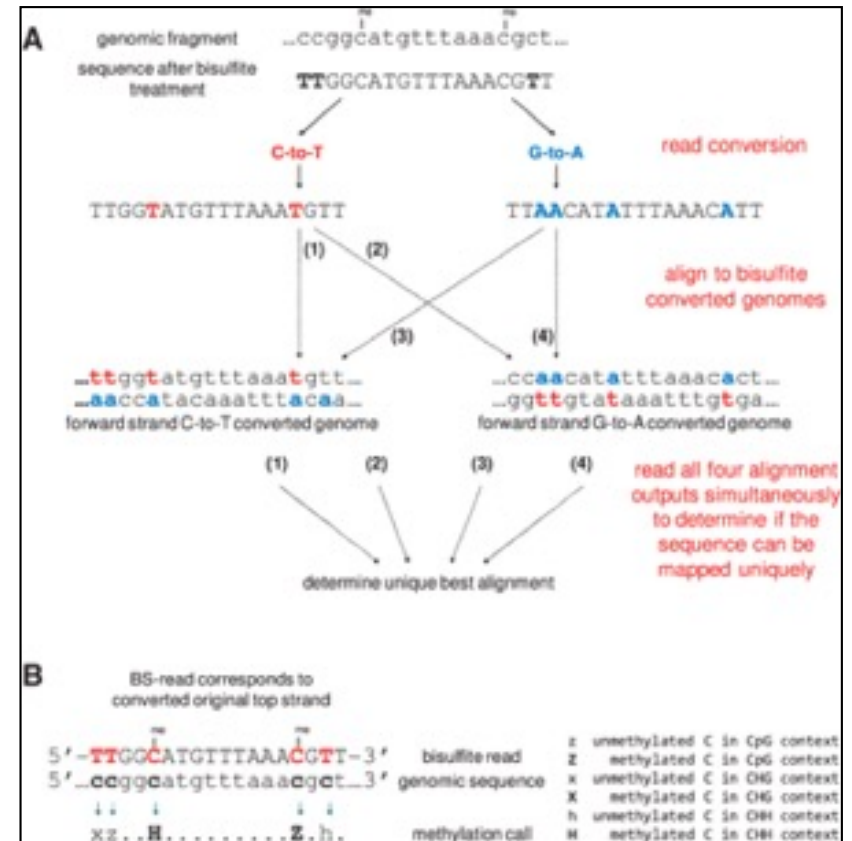
**Table 1** Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	{ <i>Hpa</i> II	<10	35	—
			{ <i>Hha</i> I	<10	39	—
		$\gamma$ -Globin	{ <i>Hpa</i> II	<10	52	—
			{ <i>Hha</i> I	<10	39	—
		$\alpha$ -Globin	{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
2	Colon	HGH	{ <i>Hpa</i> II	<10	76	—
			{ <i>Hha</i> I	<10	85	—
		$\gamma$ -Globin	{ <i>Hpa</i> II	<10	58	—
			{ <i>Hha</i> I	<10	23	—
		$\alpha$ -Globin	{ <i>Hpa</i> II	<10	<10	—
			{ <i>Hha</i> I	<10	<10	—
3	Colon	HGH	{ <i>Hpa</i> II	<10	41	—
			{ <i>Hha</i> I	<10	38	—
		$\gamma$ -Globin	{ <i>Hpa</i> II	<10	50	—
			{ <i>Hha</i> I	<10	22	—

# Bisulfite Conversion

**Treating DNA with sodium bisulfite will convert unmethyated C to T**

- 5-MethylC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a “reduced alphabet” where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



# Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications

Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.



# Bisulfite Conversion

T  
W

- 
- 
- 



ersion

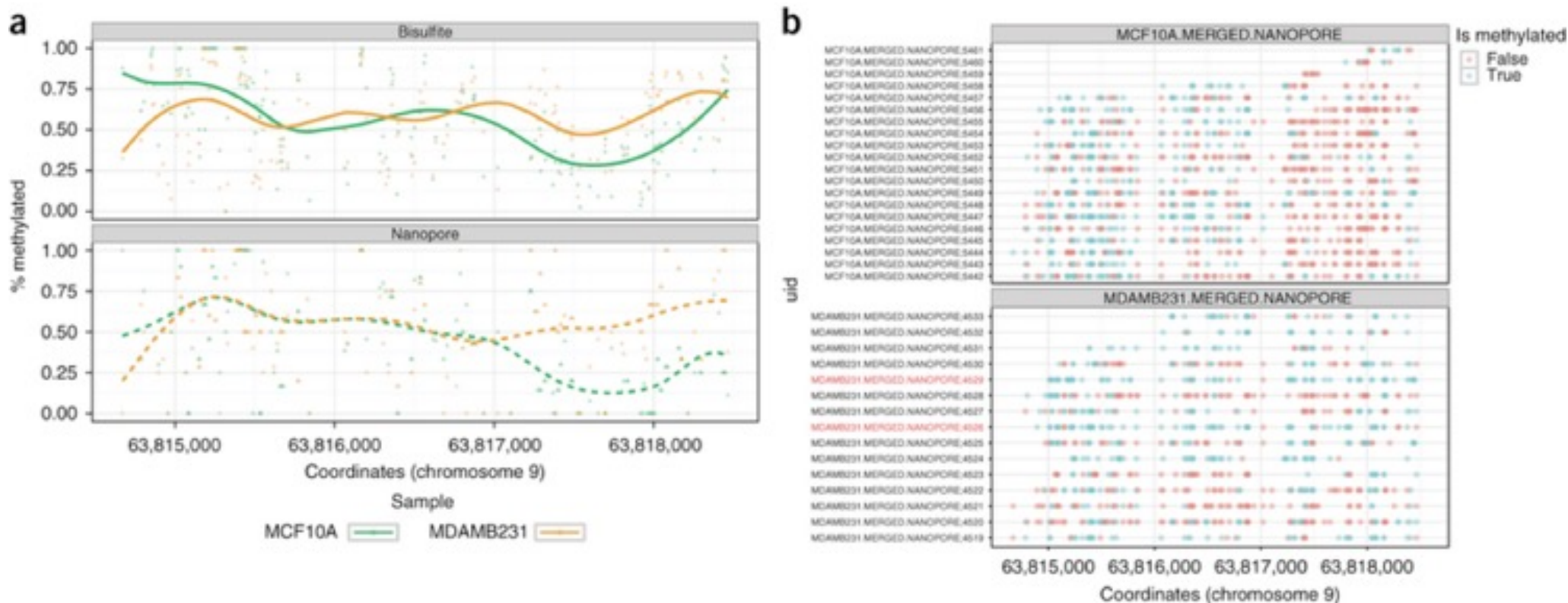
sulfite  
enomes

alignment  
taneously  
e if the  
can be  
niquely

pg context  
pg context  
pg context  
pg context  
pg context

**Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**  
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

# Methylation changes in cancer detected by Nanopore Sequencing



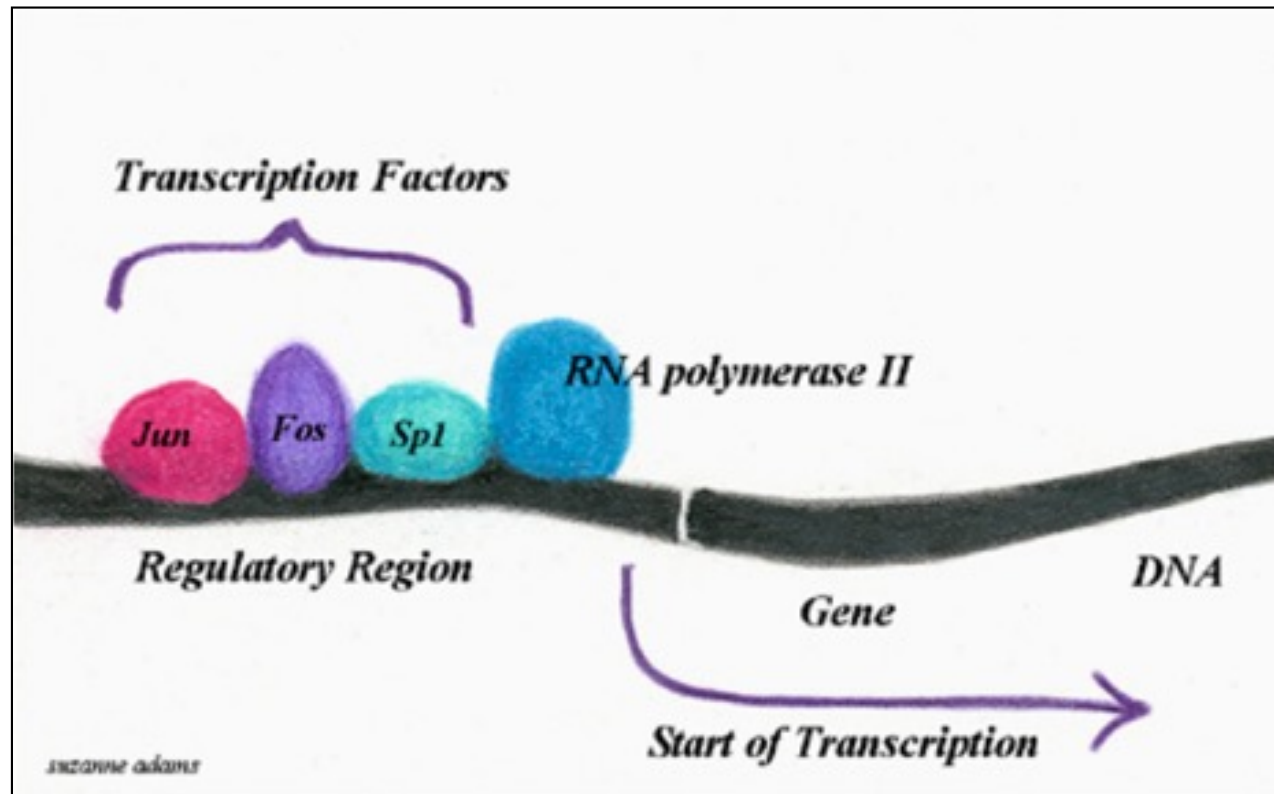
Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in a genomic region from the human mammary epithelial cell line MCF10A (green) and metastatic mammary epithelial cell line MDA-MB-231 (orange). (b) Same region as in a but with individual nanopore reads plotted separately. Each CpG that can be called is a point. Blue indicates methylated; red indicates unmethylated.

## Detecting DNA cytosine methylation using nanopore sequencing

Simpson, Workman, Zuzarte, David, Dursi, Timp (2017) Nature Methods. doi:10.1038/nmeth.4184



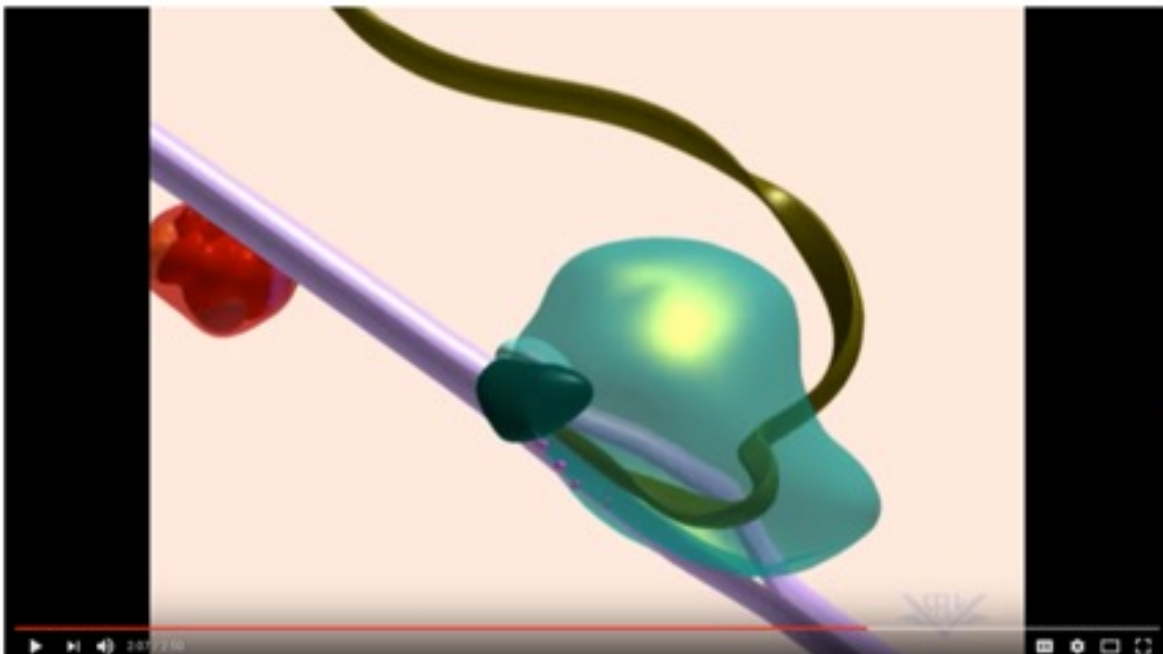
# ChIP-seq



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

Johnson et al (2007) *Science*. 316(5830):1497-502

# Transcription



Transcription

2,018,430 views

ndsvirtualcell  
uploaded on Jan 30, 2008

NDsu Virtual Cell Animations Project animation "Transcription". For more information please see <http://vcell.ndsu.edu/animations>

Up next

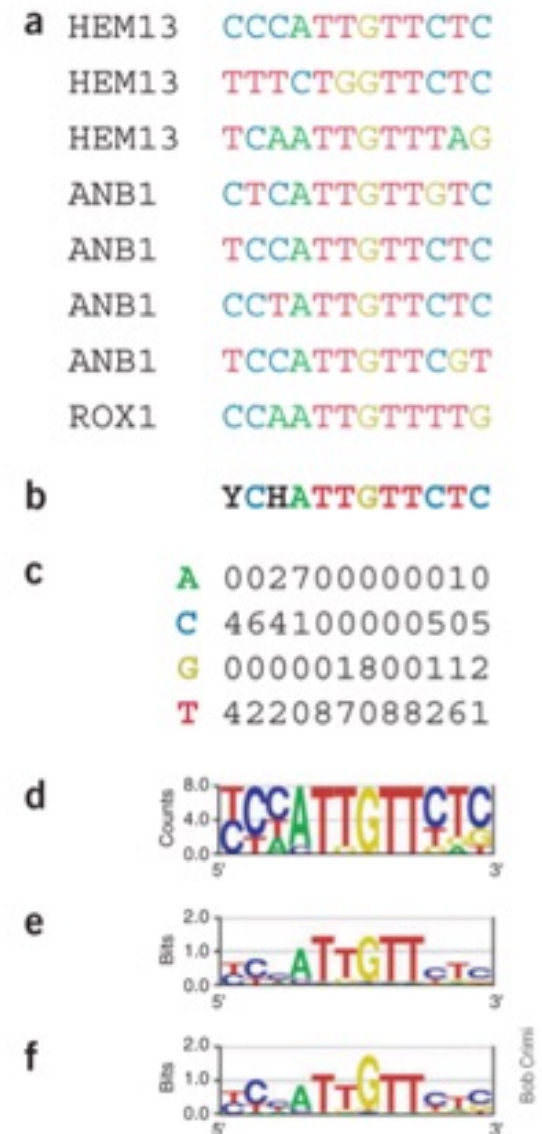
- Transcription and Translation: From DNA to Protein  
Professor Dave Explains  
151K views
- DNA - transcription and translation  
Wiam Kabala  
40K views
- Transcription and mRNA processing | Biomolecules | Khan Academy  
106K views
- DNA transcription and translation Animation  
mader abel  
45K views
- Translation  
ndsvirtualcell  
2.1M views
- Transcription and Translation Overview  
Armando Hasudungan  
611K views
- DNA, Hot Pockets, & The Longest Word Ever: Crash  
CrashCourse  
2.2M views
- TRANSCRIPTION 1  
KhanAcademyMedicine  
262K views
- TRANSCRIPTION  
congharhng  
795K views
- Moana - Best Scenes (FHD)

<https://www.youtube.com/watch?v=WsofH466lqk>

# Transcription Factors

***A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.***

- Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.
- A defining feature of transcription factors is that they contain at least one DNA-binding domain (DBD)
- Figure (a) Eight known genomic binding sites in three *S. cerevisiae* genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo (f) Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*.



## What are DNA sequence motifs?

D'haeseleer (2006) Nature Biotechnology 24, 423 – 425 doi:10.1038/nbt0406-423

# Transcription Factors Database

The screenshot displays the JASPAR database interface. At the top, there's a search bar with filters for Name, Species, and Class. Below this is a table titled "JASPAR matrix models:" with columns for ID, name, species, class, family, and Sequence logo. The table lists several transcription factors, including Ahr, Ahr-Ahr, Oat3-Catpa, NFIL3, Meom, Fork head / winged helix factors (FOX2, FOXD1), and GRI. Each entry includes a sequence logo. To the right of the table is a panel titled "ANALYZE selected matrix models:" which contains buttons for CLUSTER, RANDOMIZE, and PERMUTE, along with a section for scanning a sequence with selected matrix models.

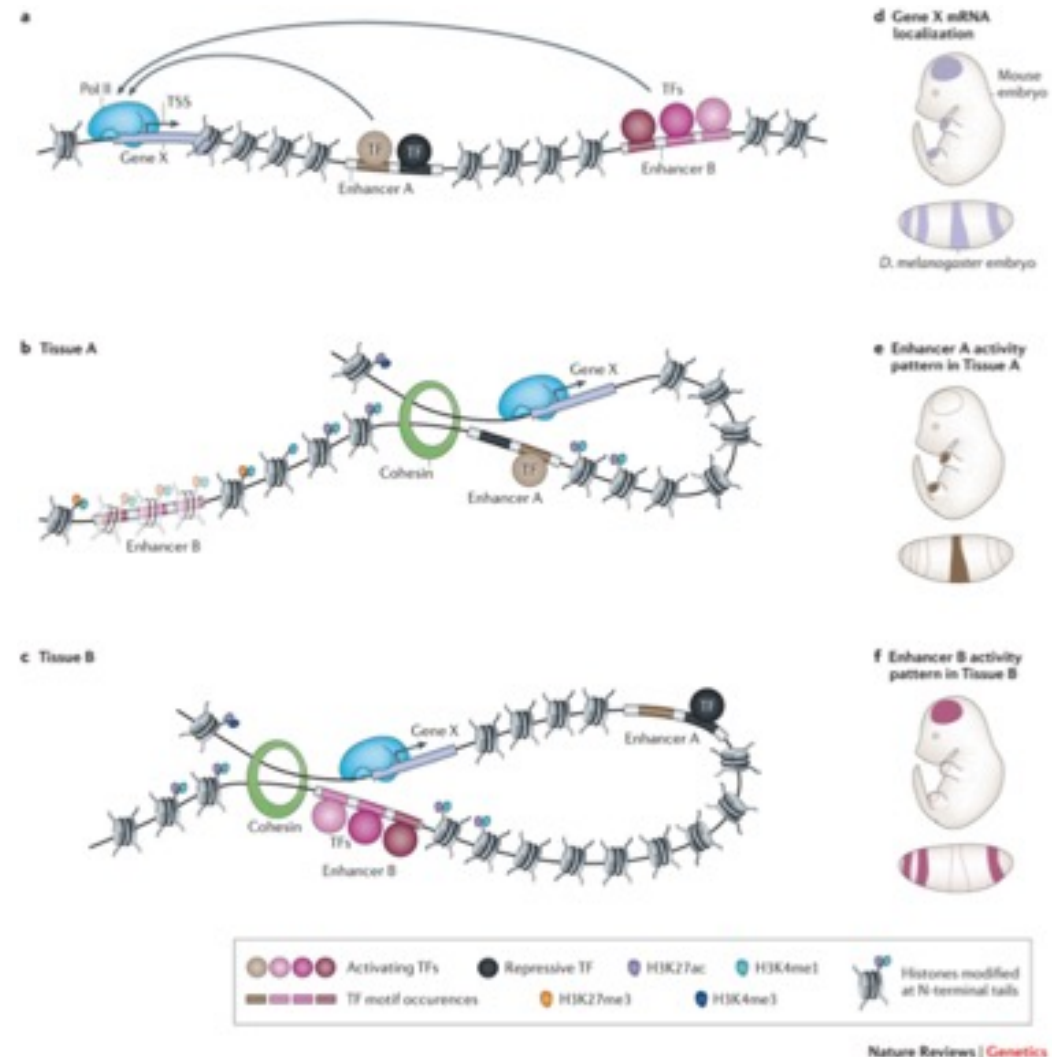
ID	name	species	class	family	Sequence logo
MA004.1	Ahr	mus musculus	Basic helix-loop-helix factors (bHLH)	PAS domain factors	
MA006.1	Ahr-Ahr	mus musculus	Basic helix-loop-helix factors (bHLH); Basic helix-loop-helix factors (bHLH)	PAS domain factors; PAS domain factors	
MA019.1	Oat3-Catpa	Rattus norvegicus	Basic leucine zipper factors (bZIP); Basic leucine zipper factors (bZIP)	C/EBP-related; C/EBP-related	
MA025.1	NFIL3	Homo sapiens	Basic leucine zipper factors (bZIP)	C/EBP-related	
MA029.1	Meom	mus musculus	C2H2 zinc finger factors	Factors with multiple dispersed zinc fingers	
MA030.1	FOX2	Homo sapiens	Fork head / winged helix factors	Forkhead box (FOX) factors	
MA031.1	FOXD1	Homo sapiens	Fork head / winged helix factors	Forkhead box (FOX) factors	
MA038.1	GRI	Rattus norvegicus	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	
MA040.1	Foxq1	Rattus norvegicus	Fork head / winged helix factors	Forkhead box (FOX) factors	

**JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles**  
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: <https://doi.org/10.1093/nar/gkt997>

# Enhancers

**Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.**

- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory elements are marked by various biochemical features
- Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities

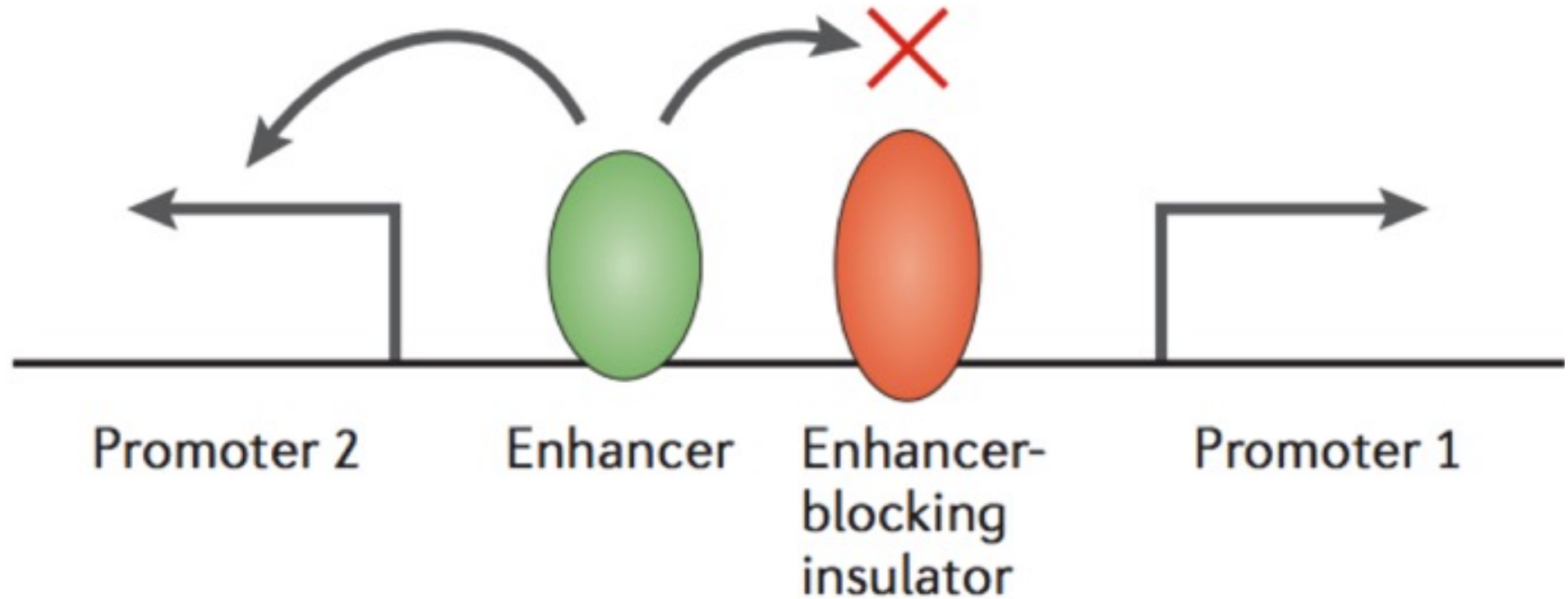


**Transcriptional enhancers: from properties to genome-wide predictions**

Shlyueva et al (2014) *Nature Reviews Genetics* 15, 272–286



# Insulators



***Insulators are DNA sequence elements that prevent “inappropriate interactions” between adjacent chromatin domains.***

- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

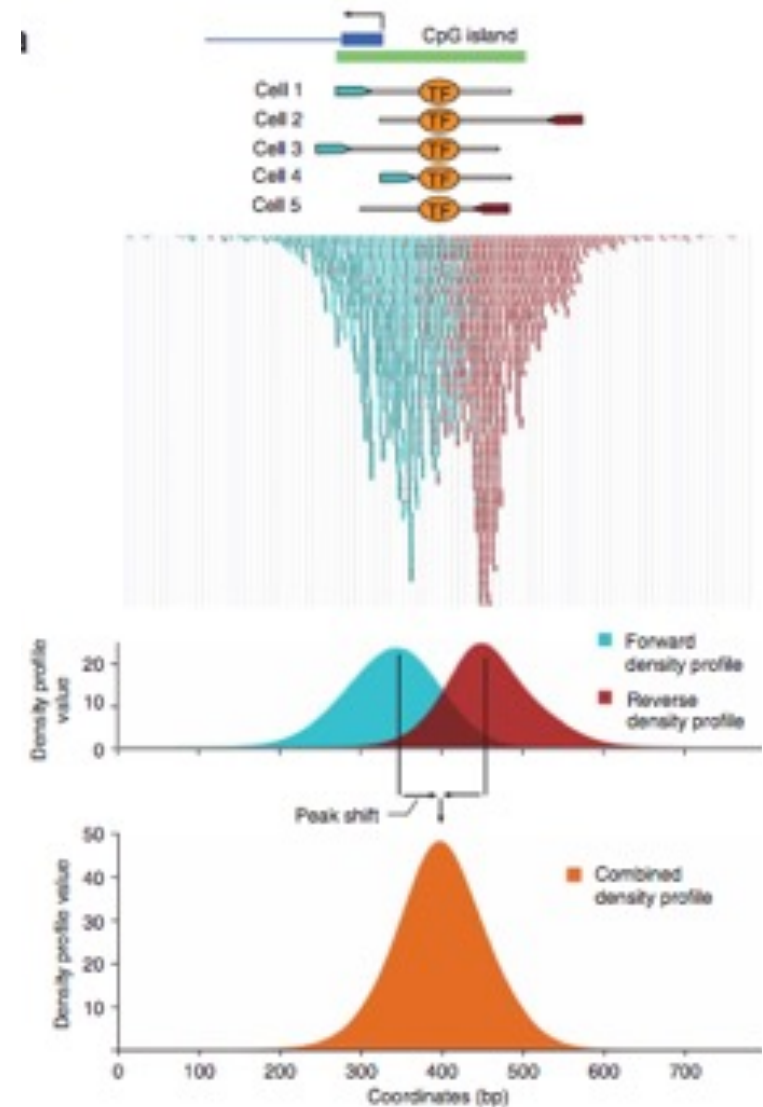
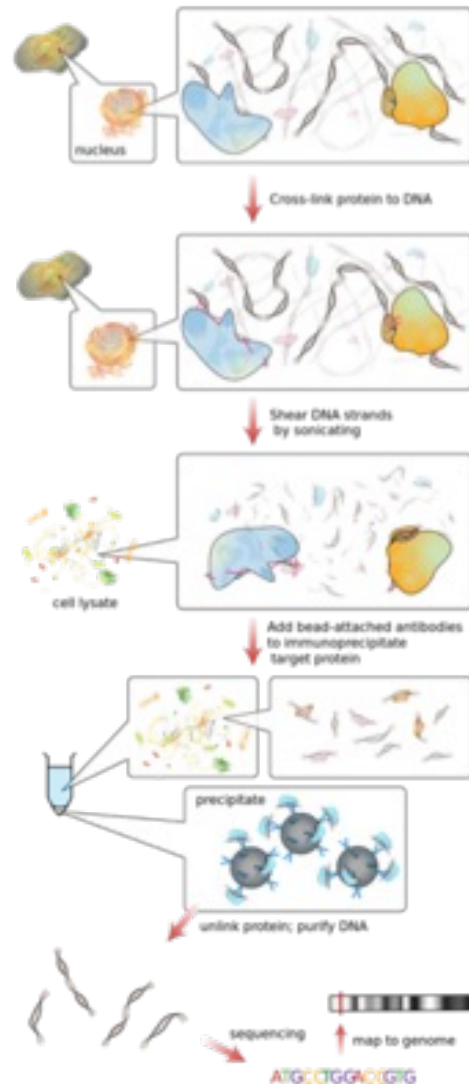
**Insulators: exploiting transcriptional and epigenetic mechanisms**

Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

# ChIP-seq:TF Binding

## Goals:

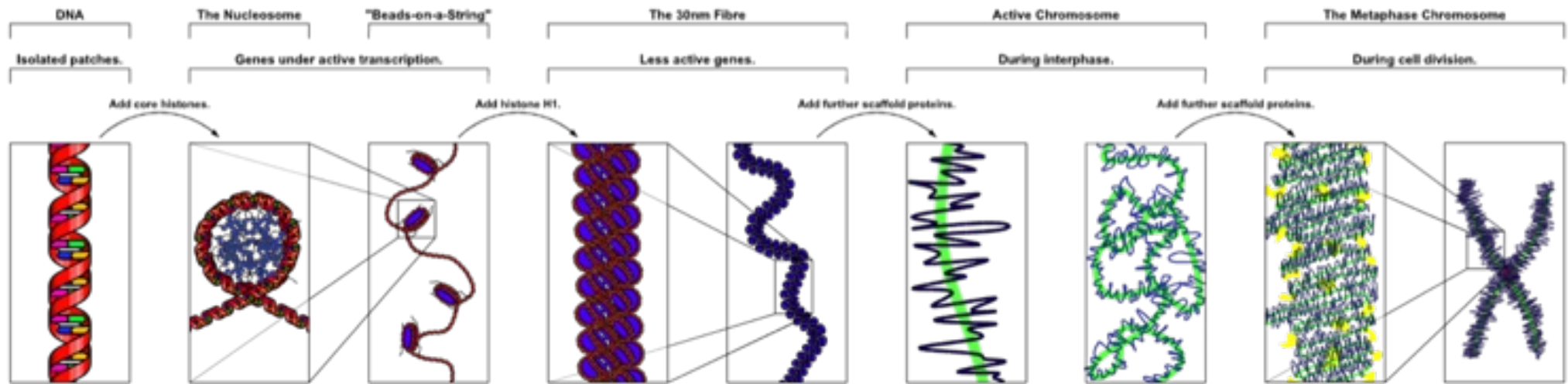
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



**Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data**

Valouev et al (2008) *Nature Methods*. 5, 829 - 834

# Chromatin compaction model



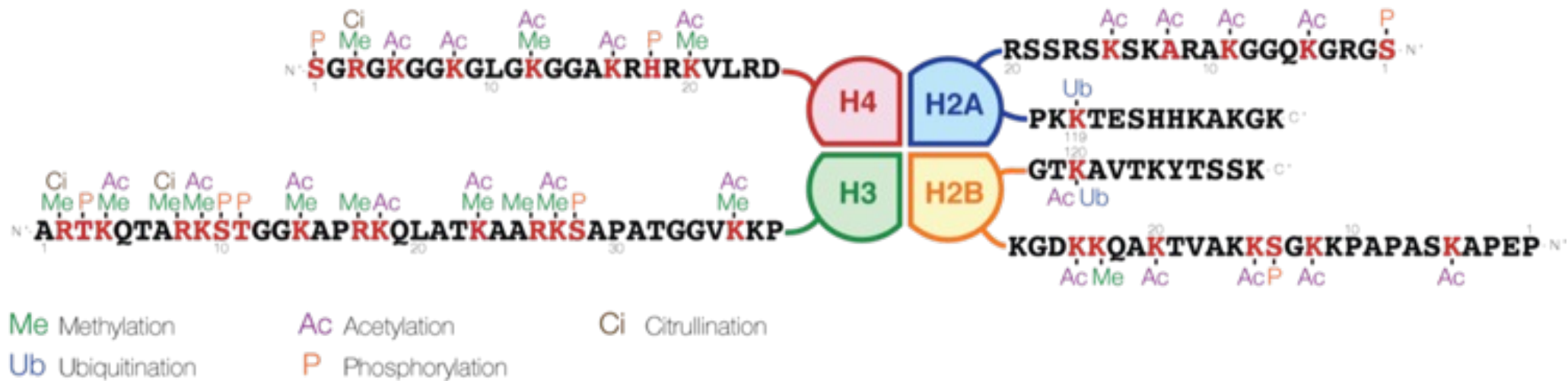
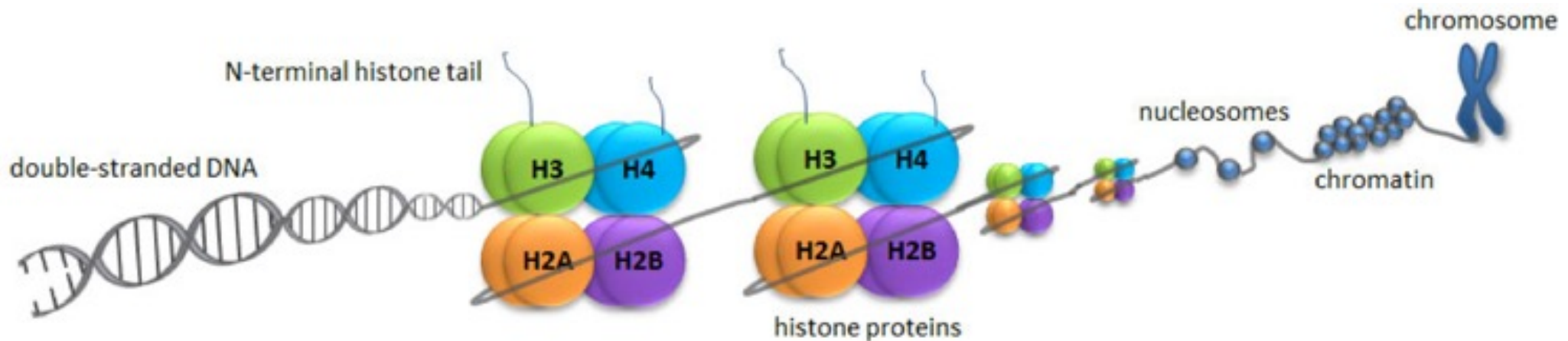
## ***Nucleosome is a basic unit of DNA packaging in eukaryotes***

- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

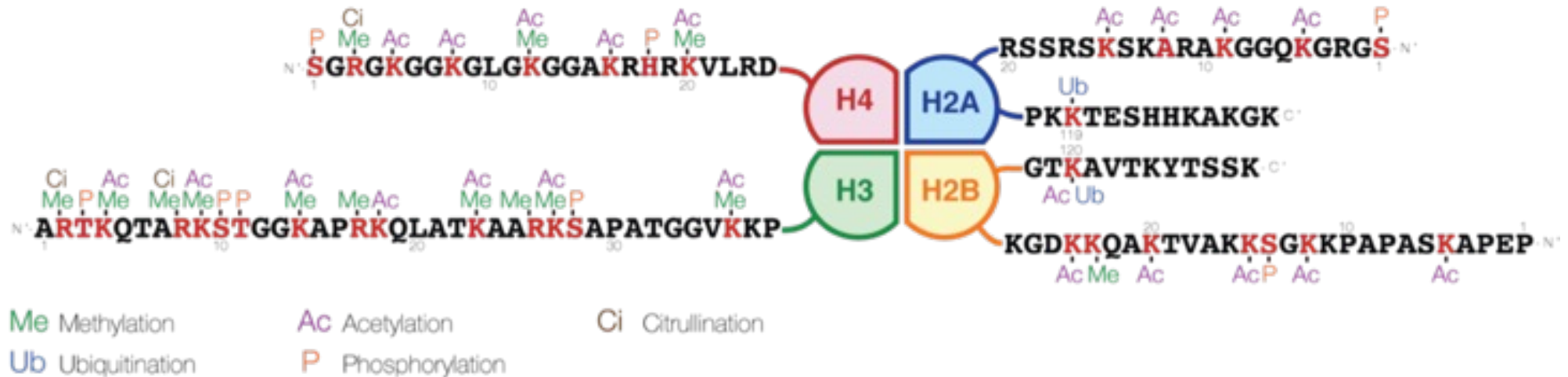
## ***Nucleosomes form the fundamental repeating units of eukaryotic chromatin***

- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10  $\mu\text{m}$  diameter).

# ChIP-seq: Histone Modifications



# ChIP-seq: Histone Modifications



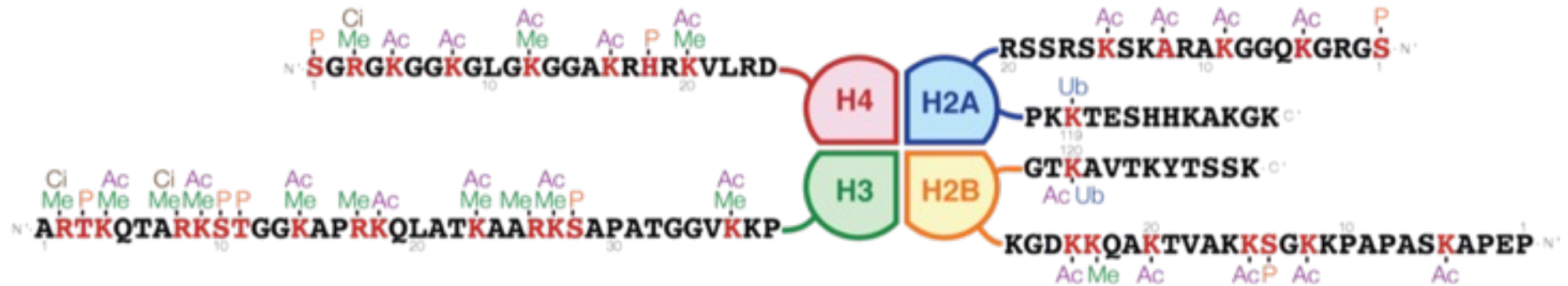
***The common nomenclature of histone modifications is:***

- The name of the histone (e.g., H3)
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification (Me: methyl, P: phosphate, Ac: acetyl, Ub: ubiquitin)
- The number of modifications (only Me is known to occur in more than one copy per residue. 1, 2 or 3 is mono-, di- or tri-methylation)

***So H3K4me1 denotes the monomethylation of the 4th residue (a lysine) from the start (i.e., the N-terminal) of the H3 protein.***



# ChIP-seq: Histone Modifications

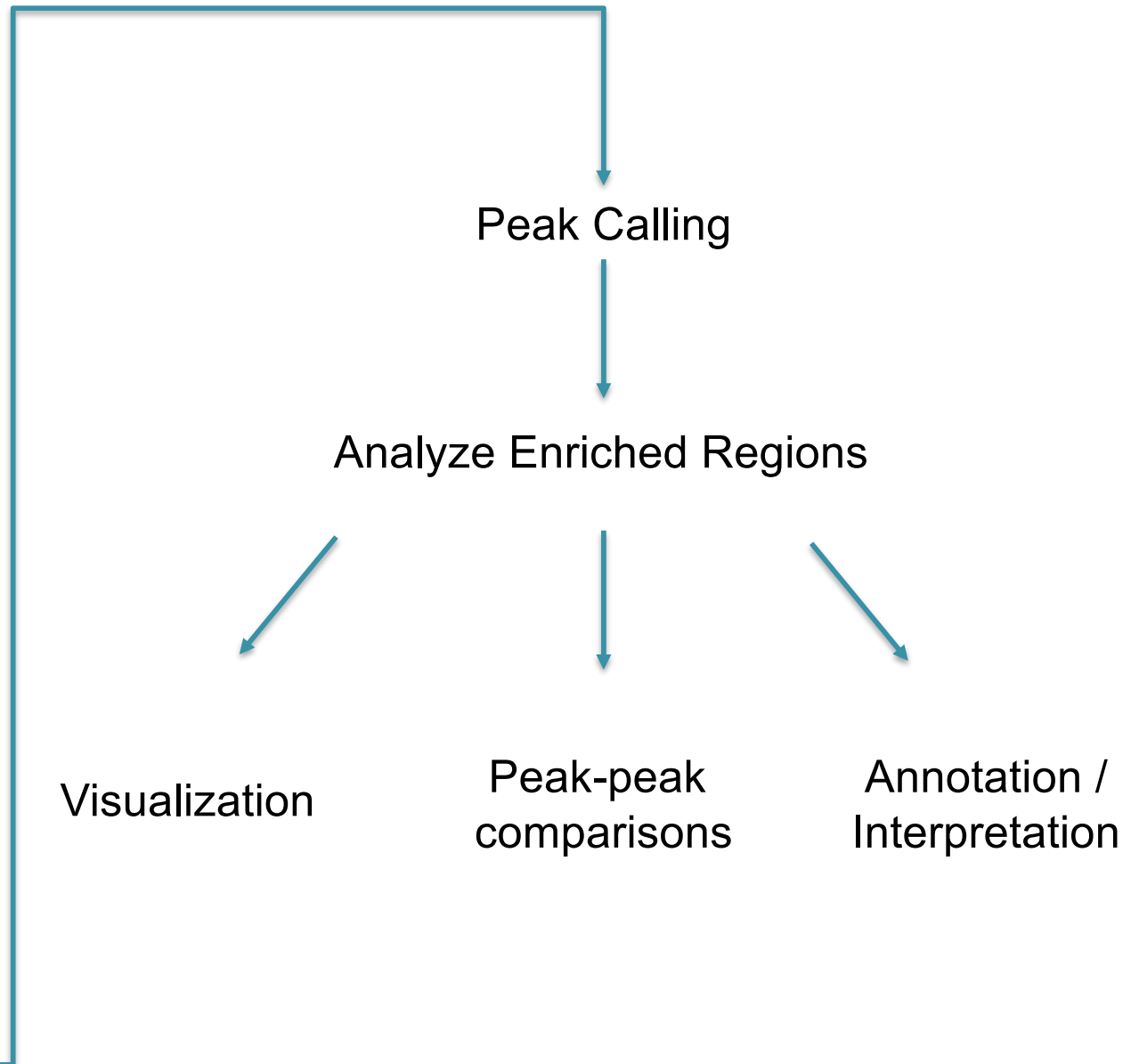
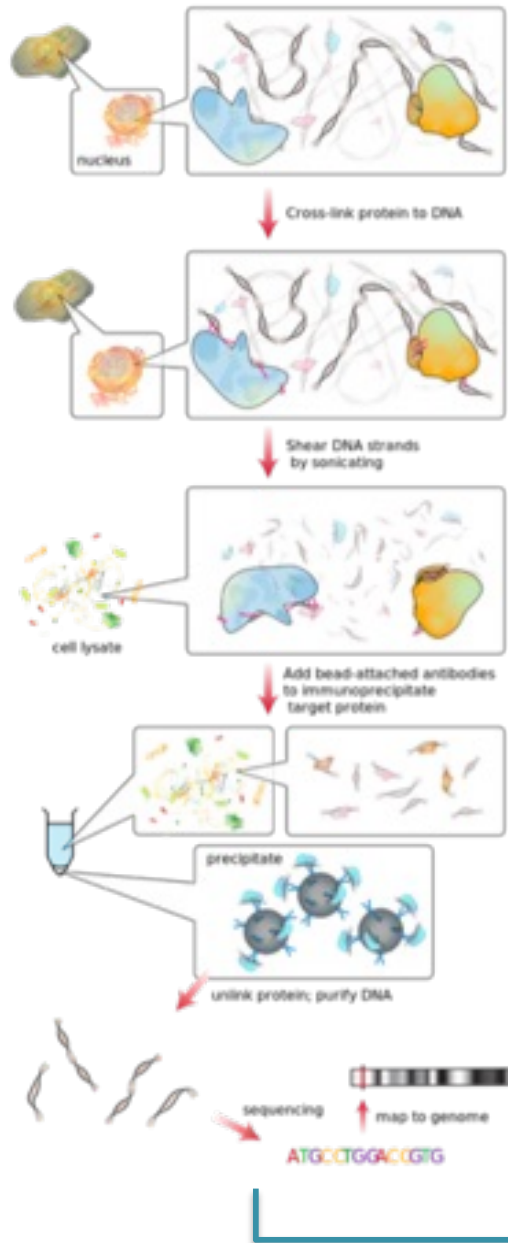


Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation <sup>[6]</sup>	activation <sup>[7]</sup>		activation <sup>[7]</sup>	activation <sup>[7][8]</sup>		activation <sup>[7]</sup>	activation <sup>[7]</sup>
di-methylation	activation	repression <sup>[3]</sup>		repression <sup>[3]</sup>	activation <sup>[8]</sup>			
tri-methylation	activation <sup>[9]</sup>	repression <sup>[7]</sup>		repression <sup>[7]</sup>	activation, <sup>[8]</sup> repression <sup>[7]</sup>			repression <sup>[3]</sup>
acetylation		activation <sup>[9]</sup>	activation <sup>[9]</sup>	activation <sup>[10]</sup>		activation <sup>[11]</sup>		

- H3K4me3 is enriched in transcriptionally active promoters.<sup>[12]</sup>
- H3K9me3 is found in constitutively repressed genes.
- H3K27me3 is found in facultatively repressed genes.<sup>[7]</sup>
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.
- H3K27ac distinguishes active enhancers from poised enhancers.
- H3K122ac is enriched in poised promoters and also found in a different type of putative enhancer that lacks H3K27ac.

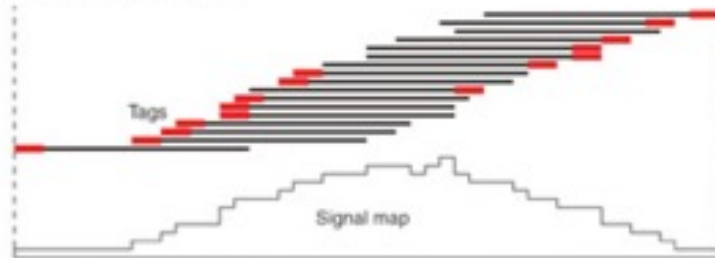


# General Flow of ChIP-seq Analysis



# PeakSeq

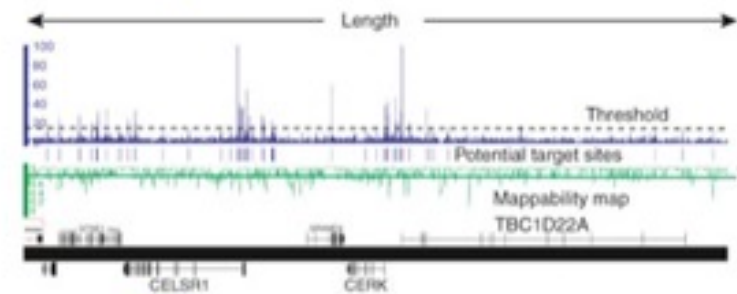
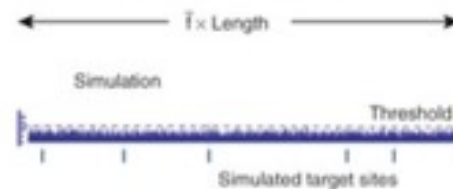
## 1. Constructing signal maps



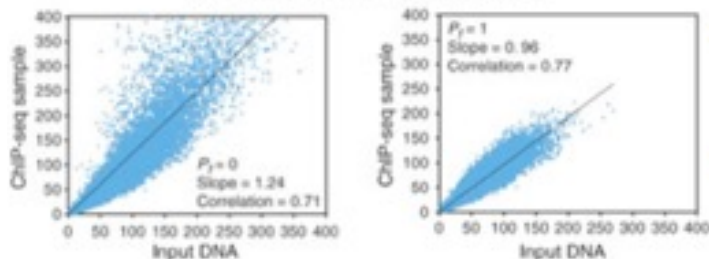
- Extend mapped tags to DNA fragment
- Map of number of DNA fragments at each nucleotide position

## 2. First pass: determining potential binding regions by comparison to simulation

- Simulate each segment
- Determine a threshold satisfying the desired initial false discovery rate
- Use the threshold to identify potential target sites



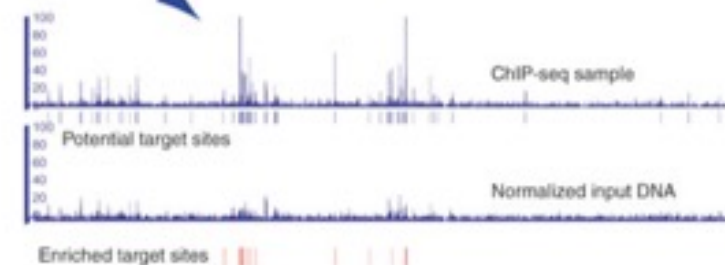
## 3. Normalizing control to ChIP-seq sample



- Select fraction of potential peaks to exclude (parameter  $P_2$ )
- Count tags in bins along chromosome for ChIP-seq sample and control
- Determine slope of least squares linear regression

## 4. Second pass: scoring enriched target regions relative to control

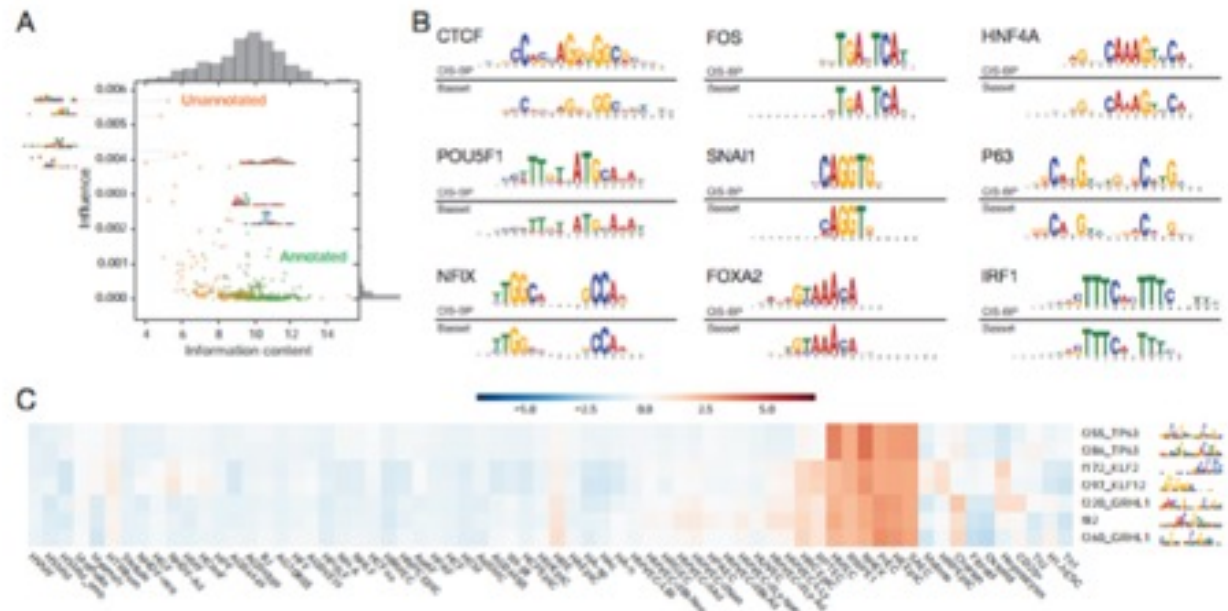
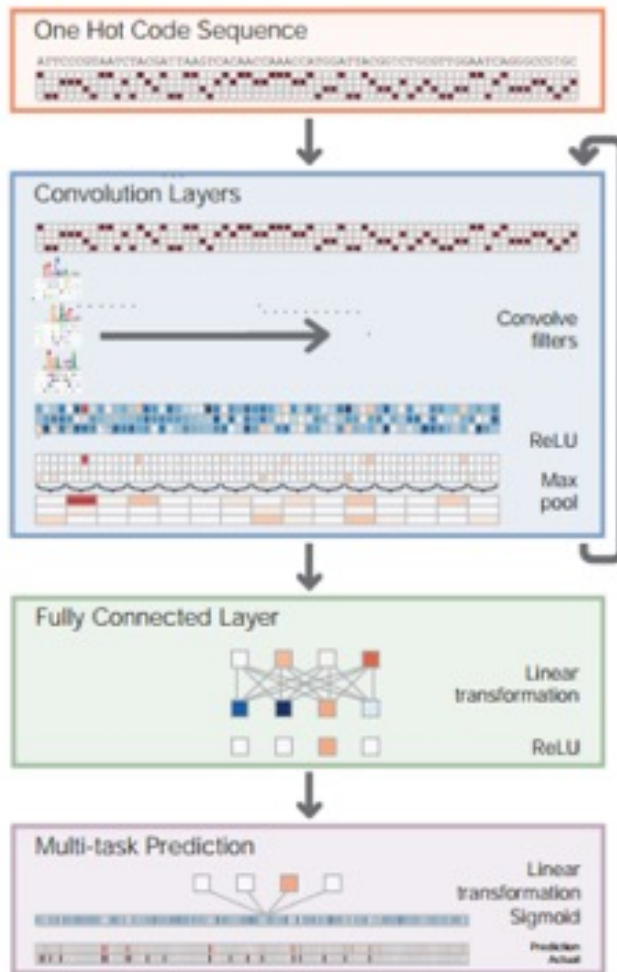
- For potential binding sites calculate the fold enrichment
- Compute a  $P$ -value from the binomial distribution
- Correct for multiple hypothesis testing and determine enriched target sites



**PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls**

Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

# Basset

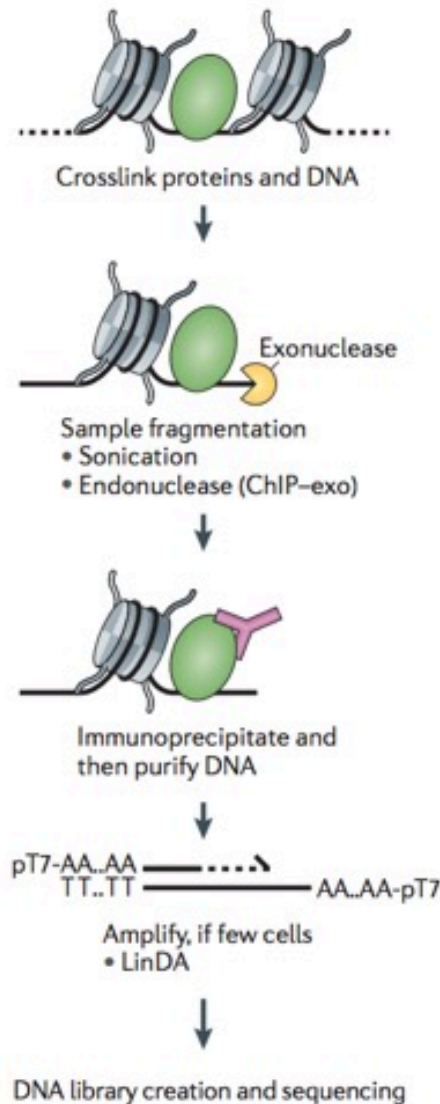


**Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks**

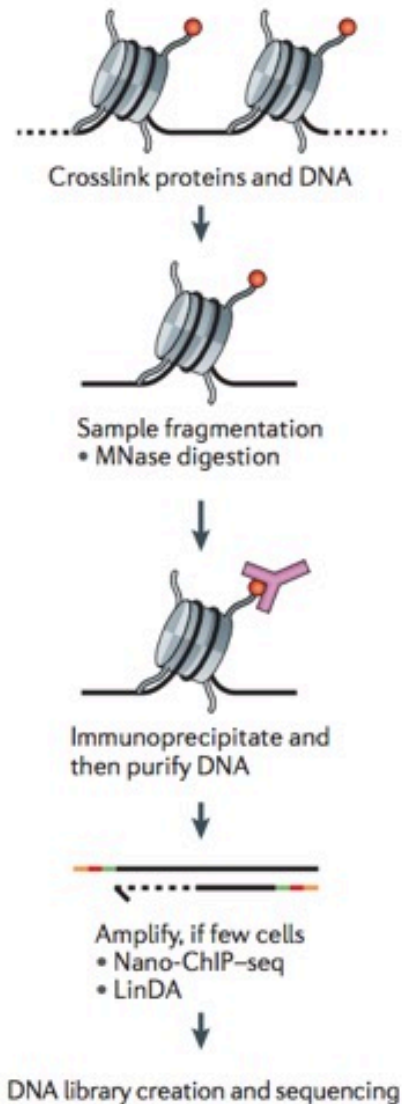
Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115

# Related Assays

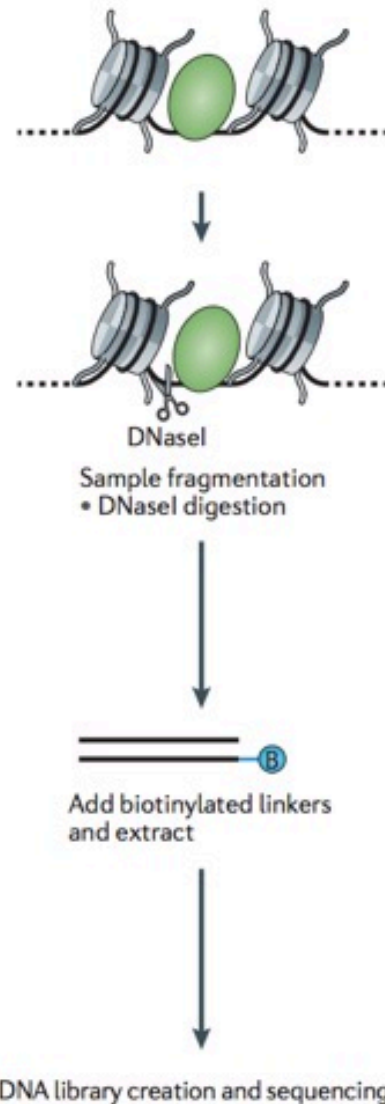
**a DNA-binding protein ChIP-seq**



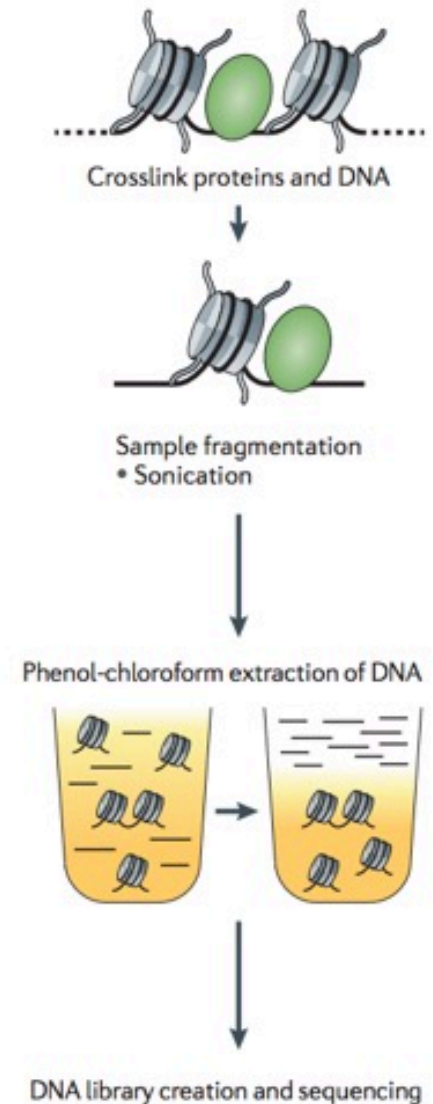
**b Histone modification ChIP-seq**



**c DNase-seq**



**d FAIRE-seq**

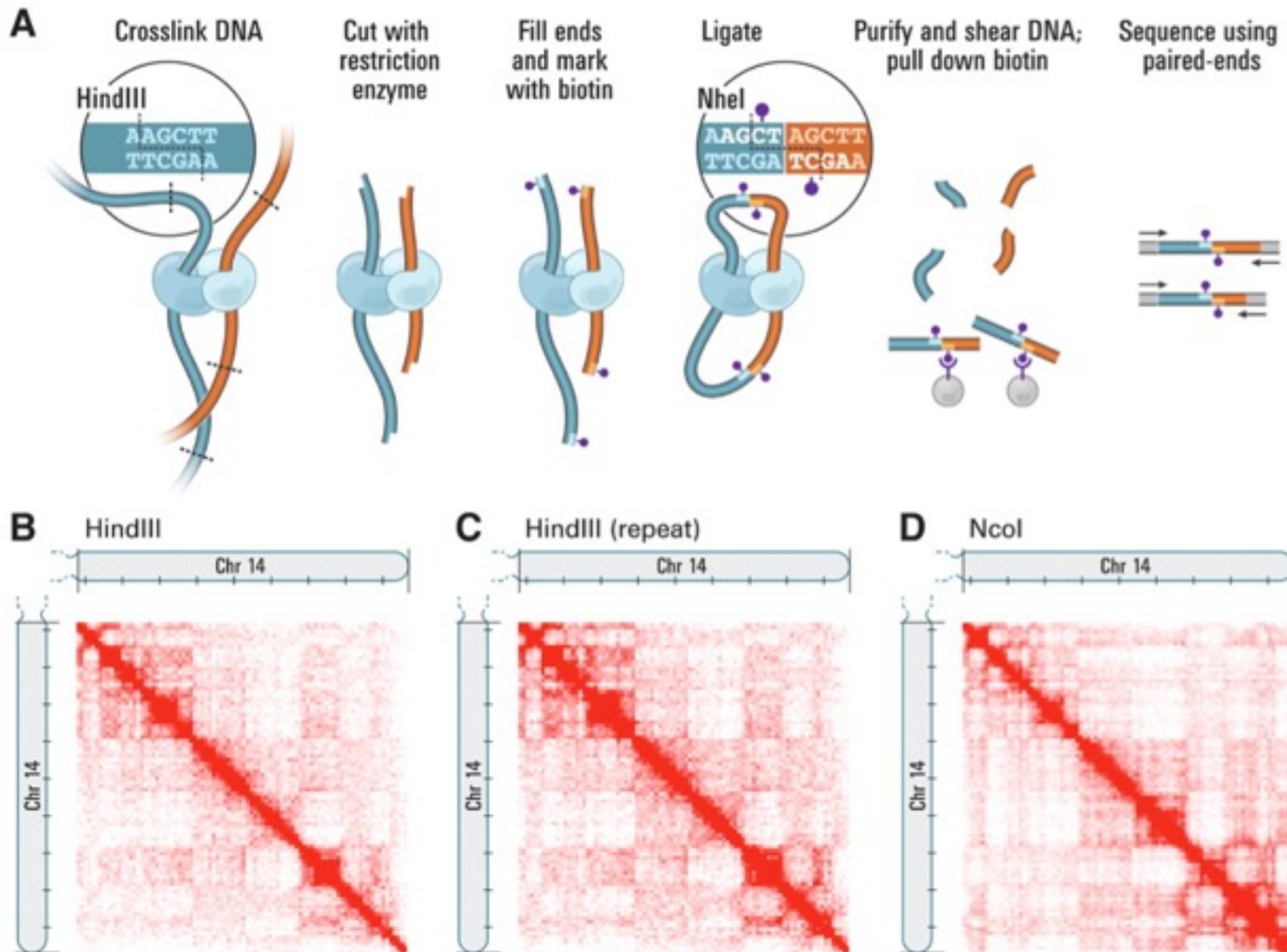


**ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions**

Furey (2012) *Nature Reviews Genetics*. 13, 840-852



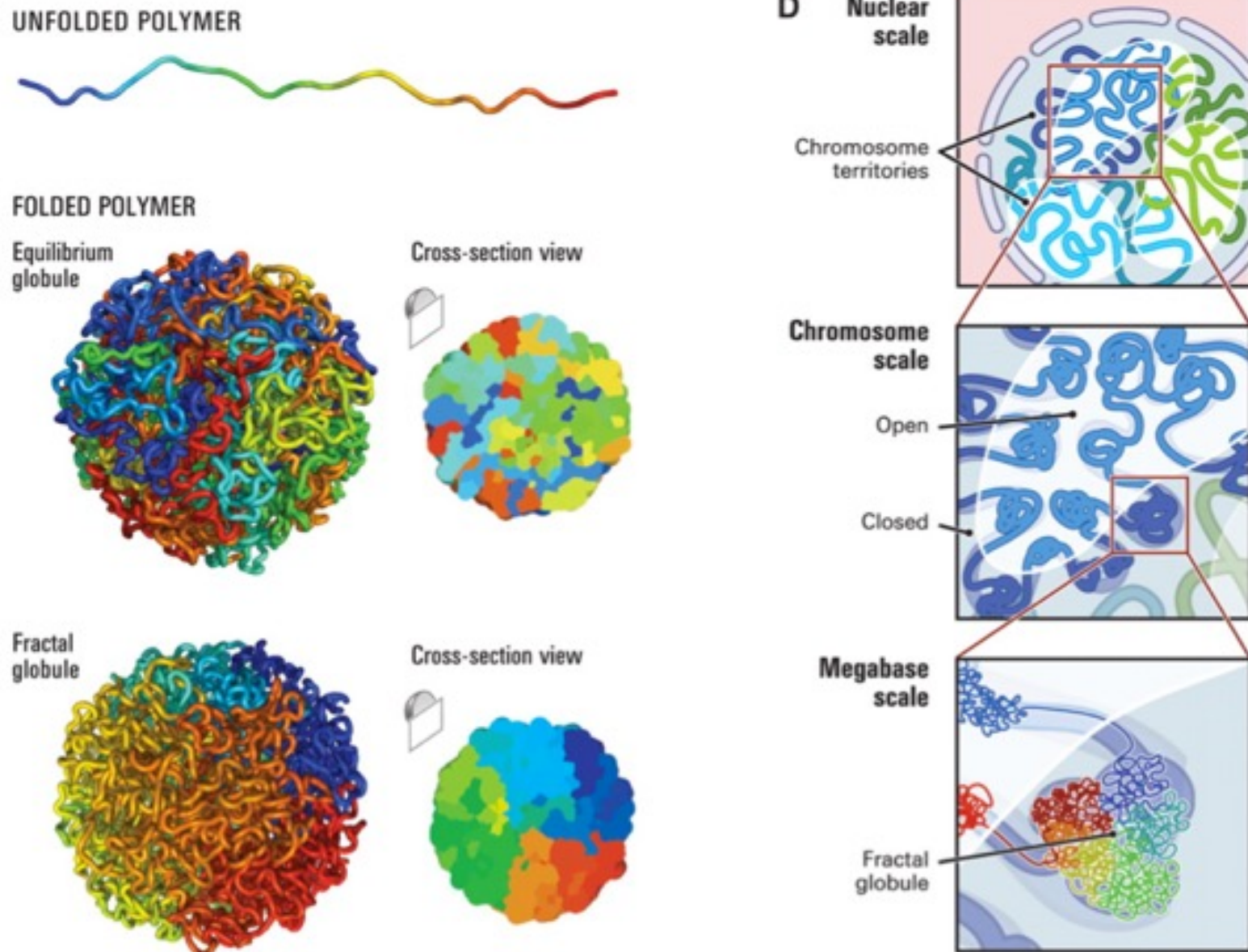
# Hi-C: Mapping the folding of DNA



**Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**

Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

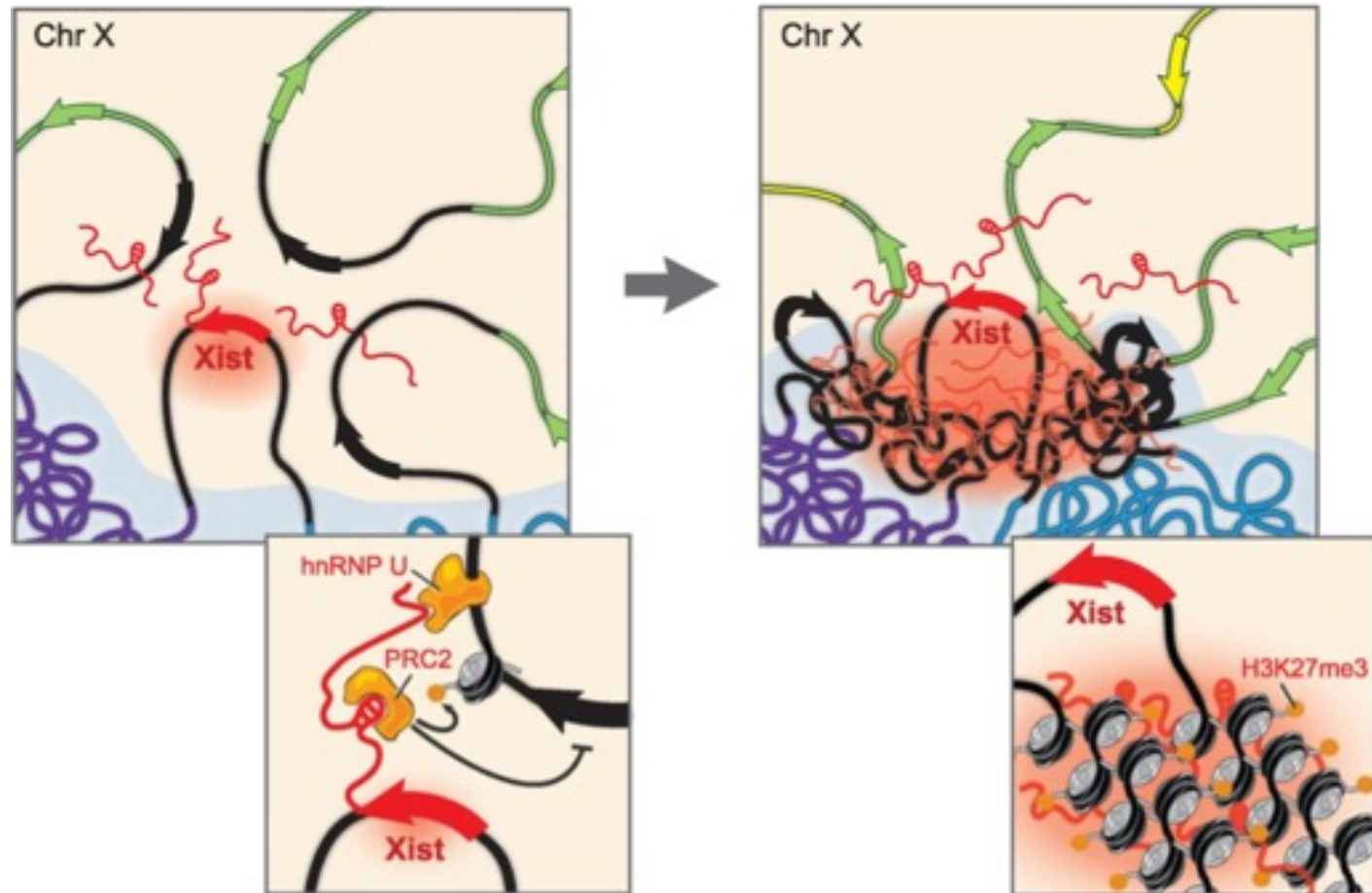
# Hi-C: Mapping the folding of DNA



**Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**

Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

# Gene Regulation in 3-dimensions

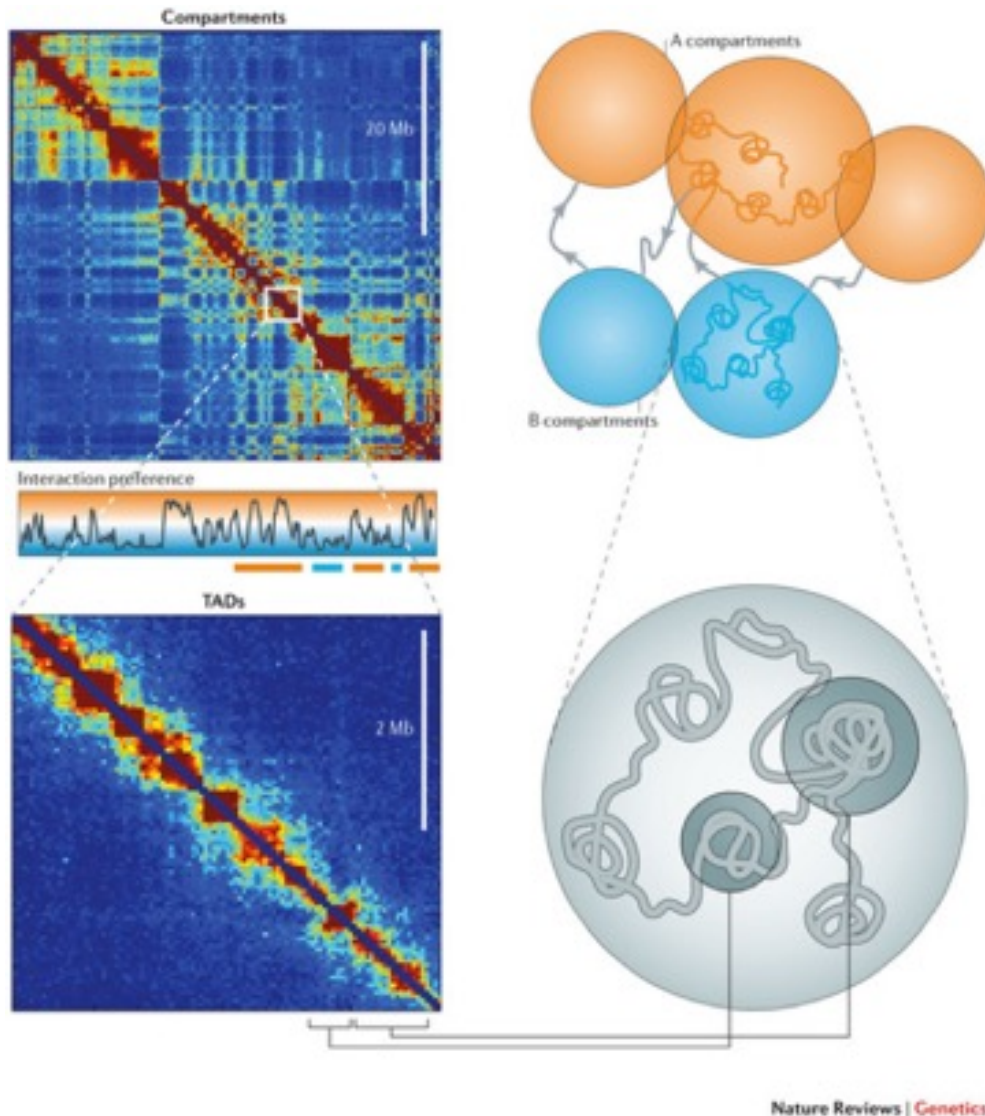


**Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.**

The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome  
Engreitz et al. (2013) *Science*. 341 (6147)



# Genome compartments & TADs



***Mammalian genomes have a pattern of interactions that can be approximated by two compartments called A and B***

- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

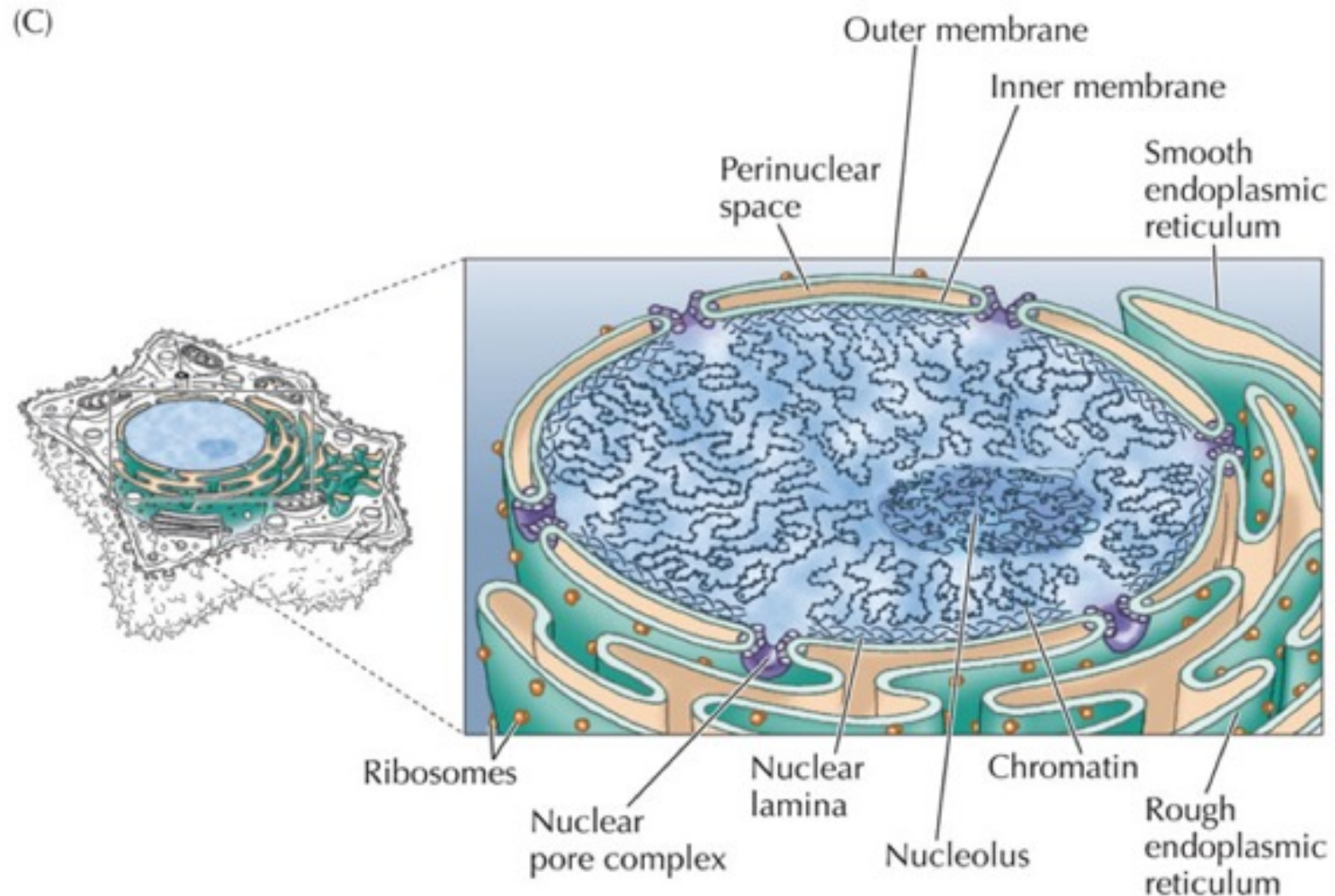
***Topologically associating domains (TADs)***

- TADs are smaller (~400–500 kb)
- Can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status.
- TADs are hard-wired features of chromosomes, and groups of adjacent TADs can organize in A and B compartments

**Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data**

Dekker et al. (2013) *Nature Reviews Genetics* 14, 390–403

# “Lamina-Associated Domains are the B compartment”

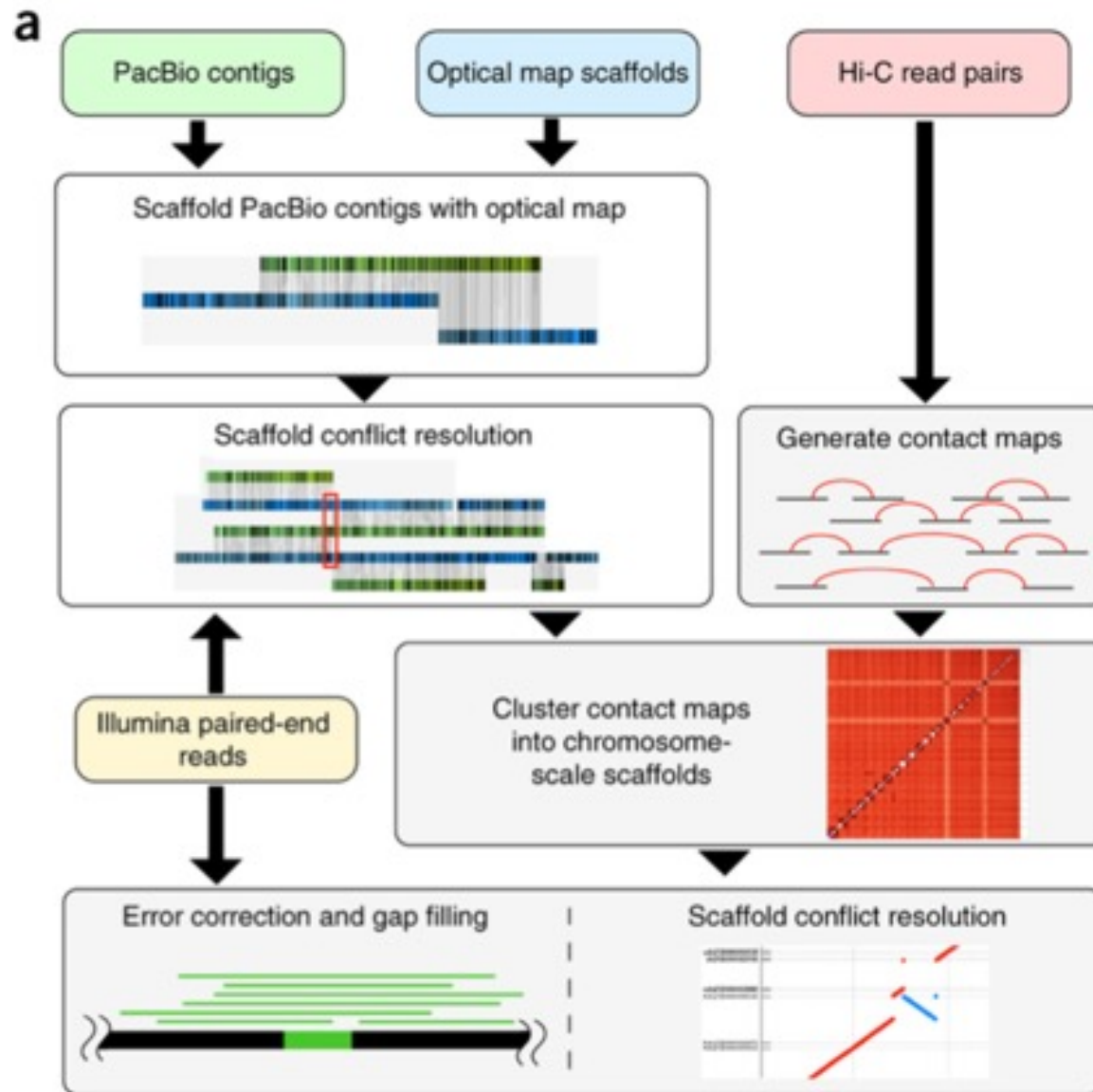


THE CELL, Fourth Edition, Figure 9.1 (Part 3) © 2006 ASM Press and Sinauer Associates, Inc.

**Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation**

Luperchio et al. (2017) bioRxiv. doi: <https://doi.org/10.1101/122226>

# Scaffolding with Hi-C

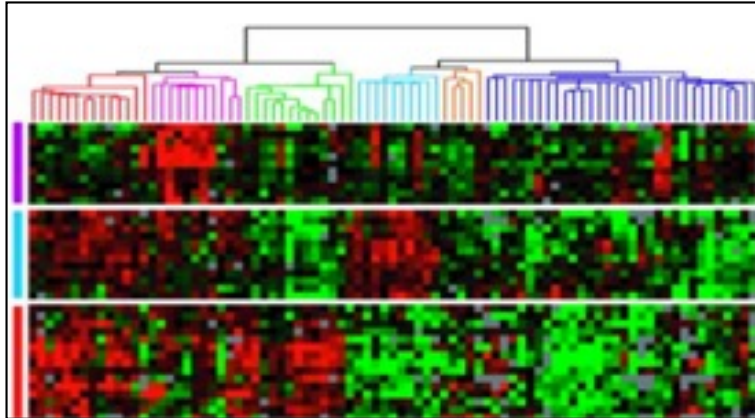


**Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome**  
Bickhart et al (2017) Nature Genetics (2017) doi:10.1038/ng.3802

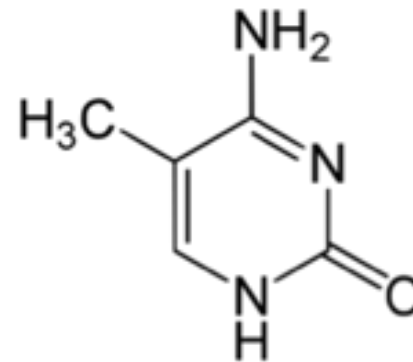


# Putting it all together!

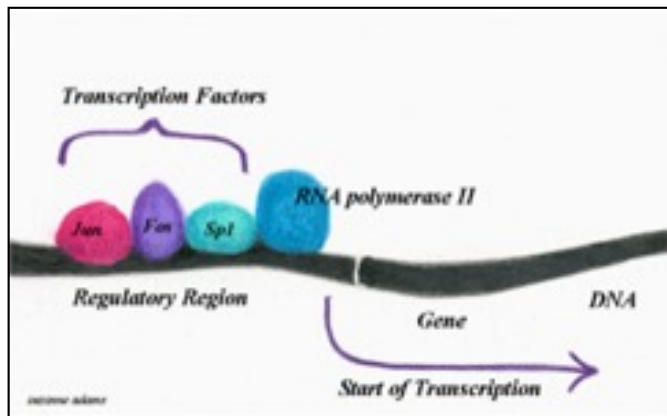
RNA-seq



Methyl-seq



ChIP-seq



Hi-C

