Introduction to Epigenetics and Three-Dimensional Genome Organization

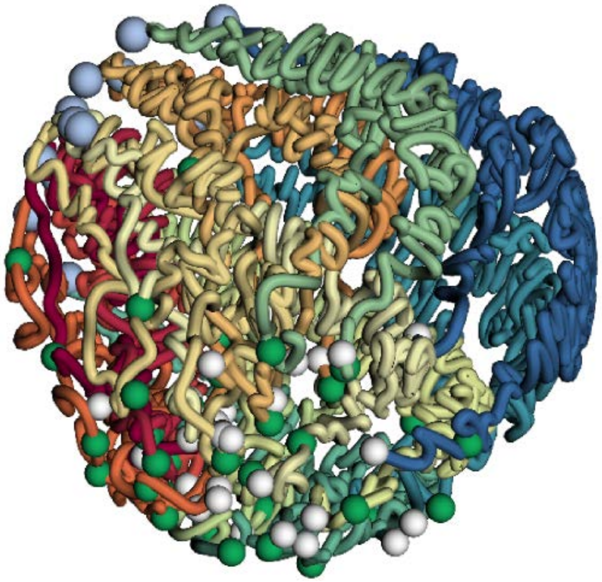
Ferhat Ay

Assistant Professor of Computational Biology

La Jolla Institute for Immunology

Genome Informatics Division, Department of Pediatrics, UCSD

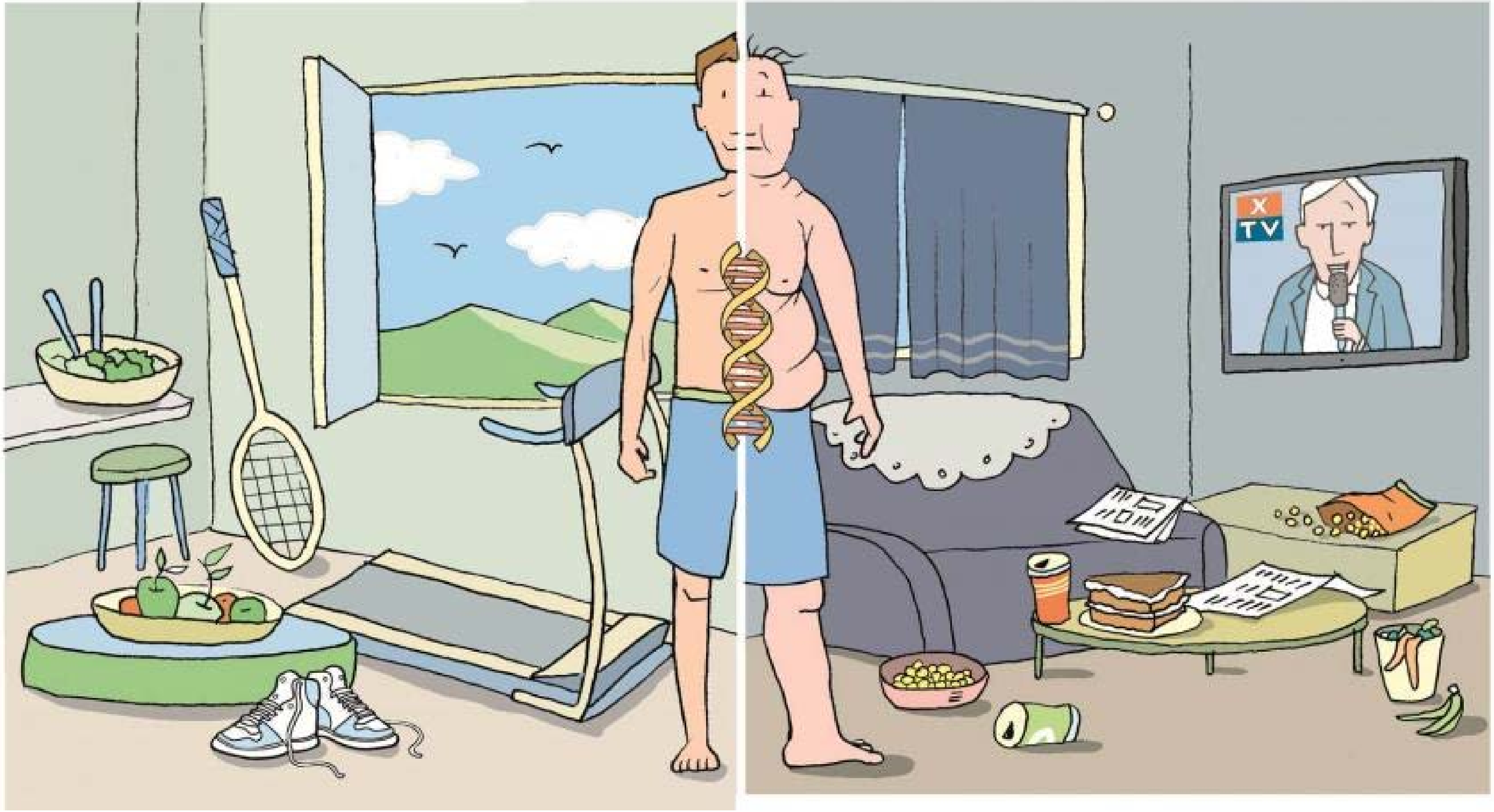
BGGN-213 – Guest Lecture - W2020



What is Epigenetics?

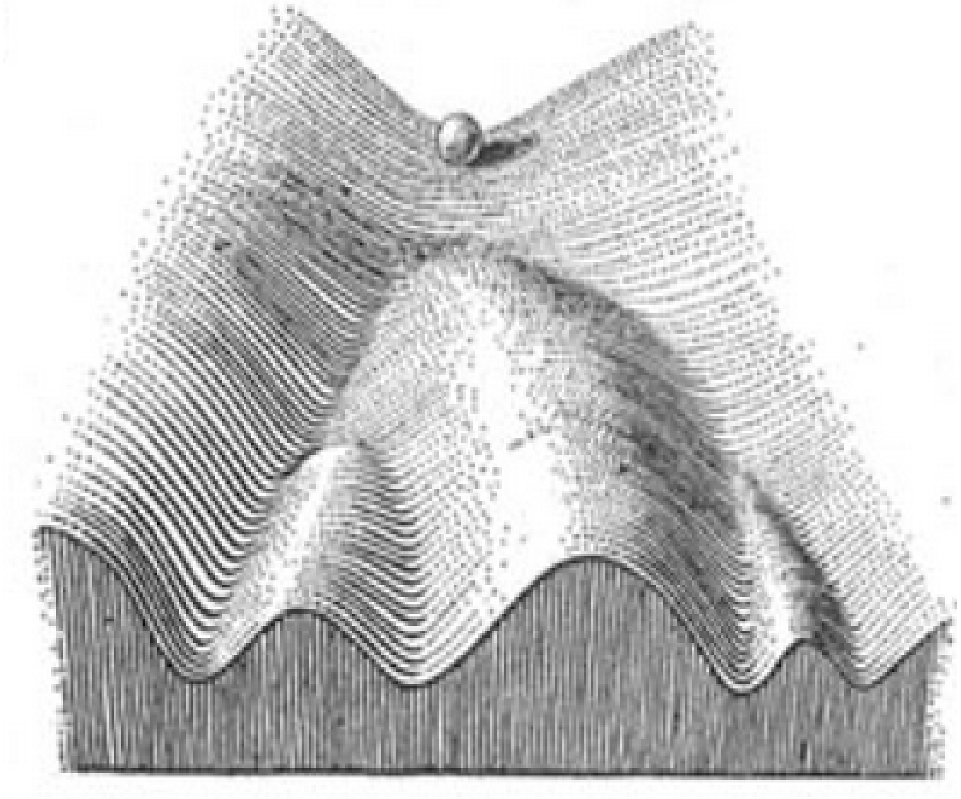
- **Epigenetics** is the study of heritable phenotype changes that do not involve alterations in the DNA sequence. The Greek prefix epi- (above, over, outside of) in epi-genetics implies features that are *on top of* or *in addition to* the traditional genetic basis for inheritance

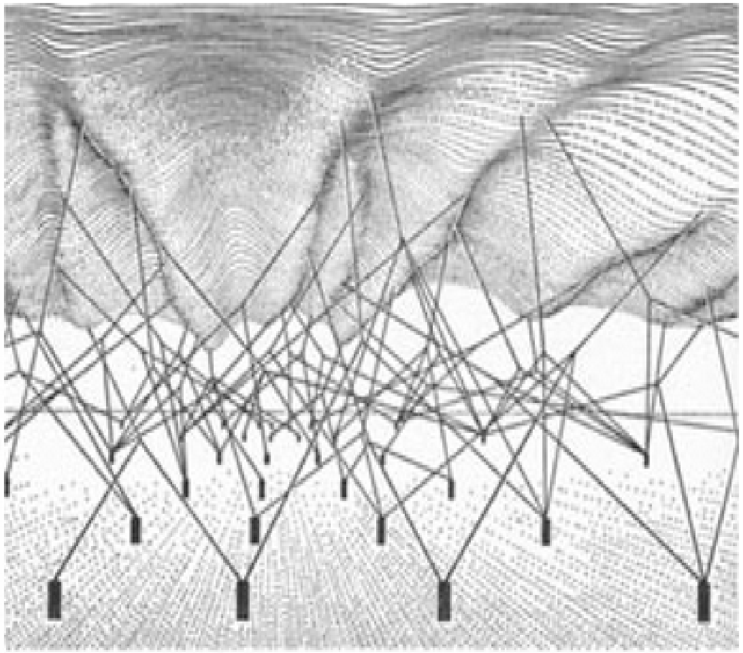
Environmental effects influence how genes are turned on and off



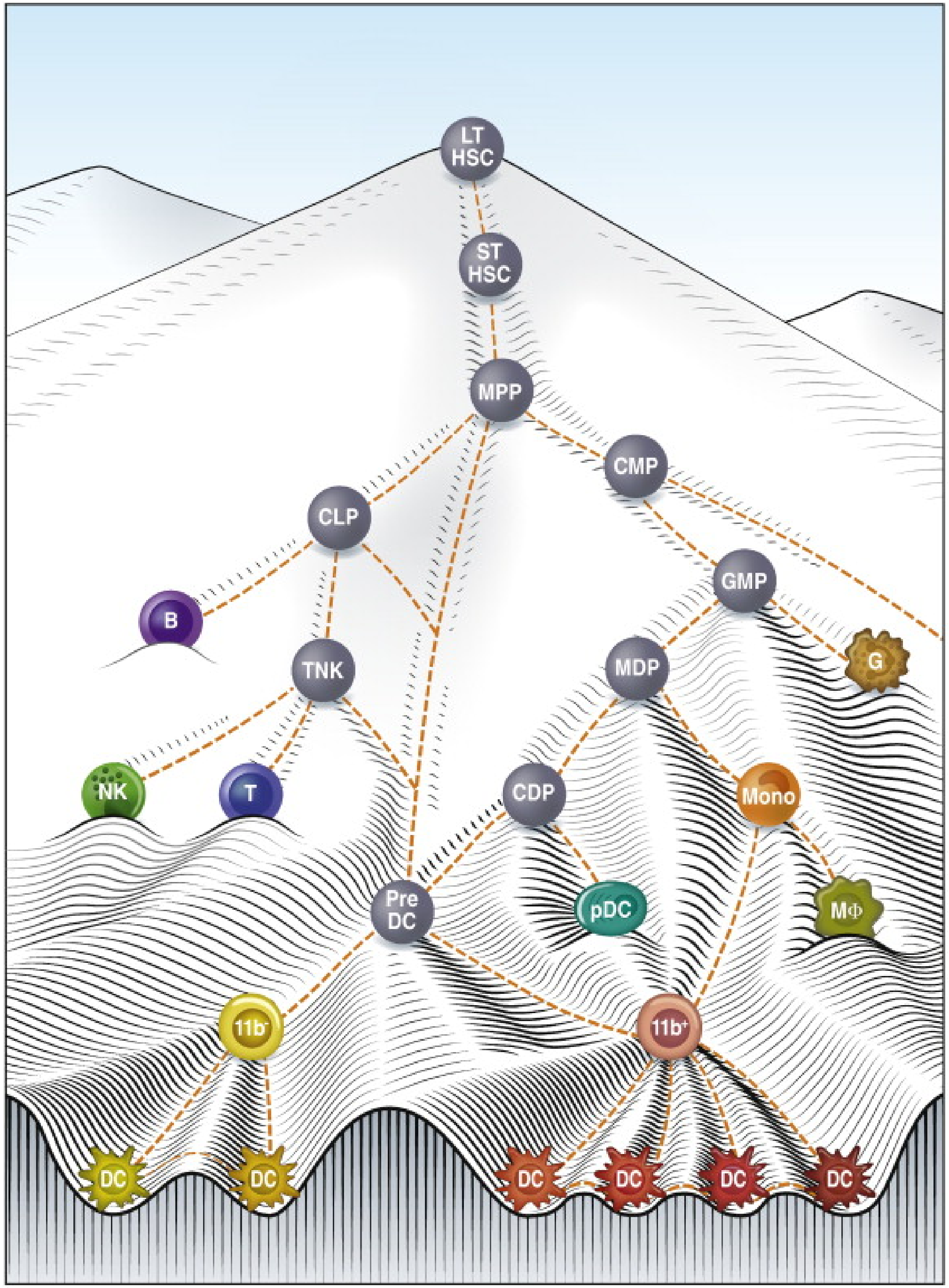
Credit: Weizmann Institute of Science

Waddington’s epigenetic landscape



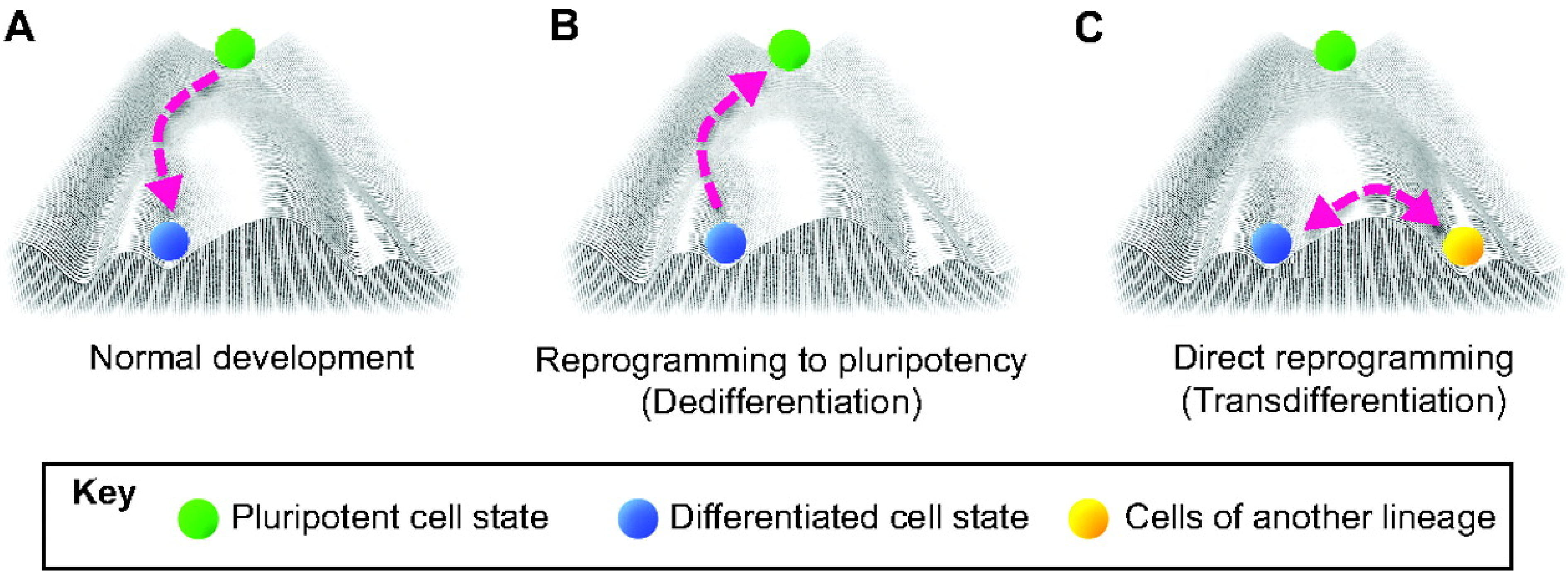


Hematopoietic Cell Lineage Tree



Current Opinion in Immunology

Hematopoietic Cell Lineage Tree?



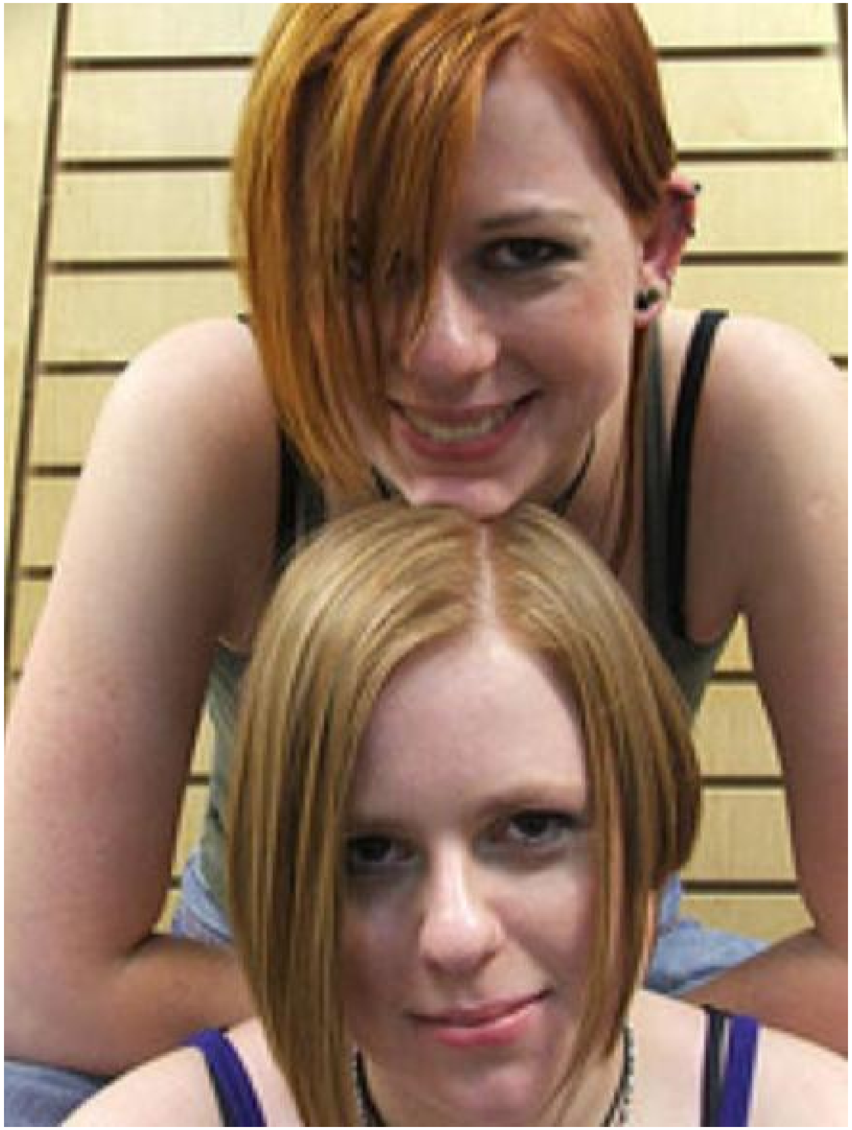
Normal development

Reprogramming to pluripotency
(Dedifferentiation)

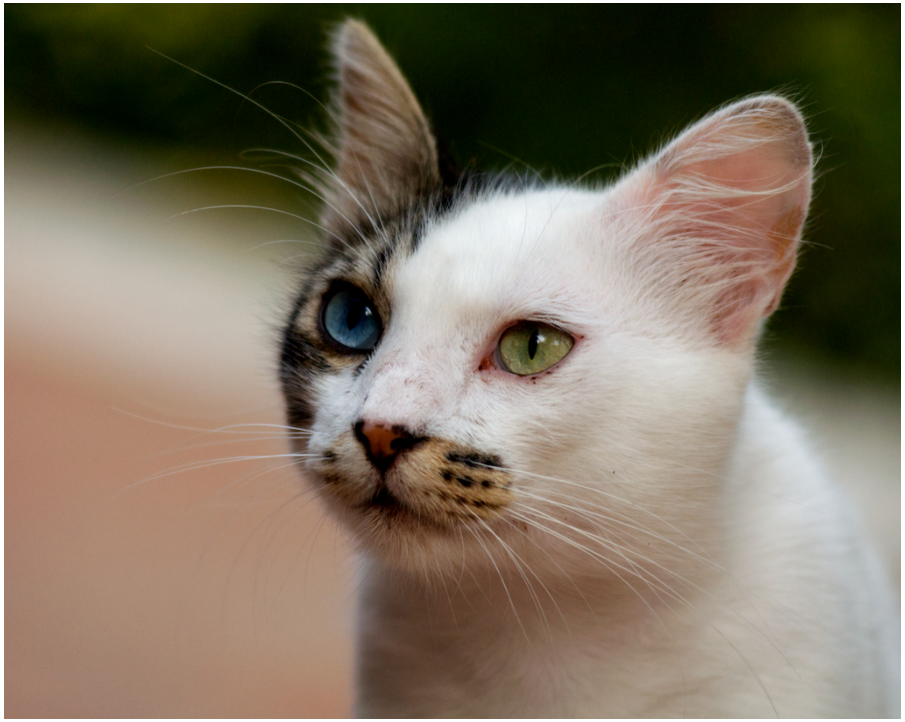
Direct reprogramming
(Transdifferentiation)

Examples of epigenetic inheritance

Identical twins with different hair color



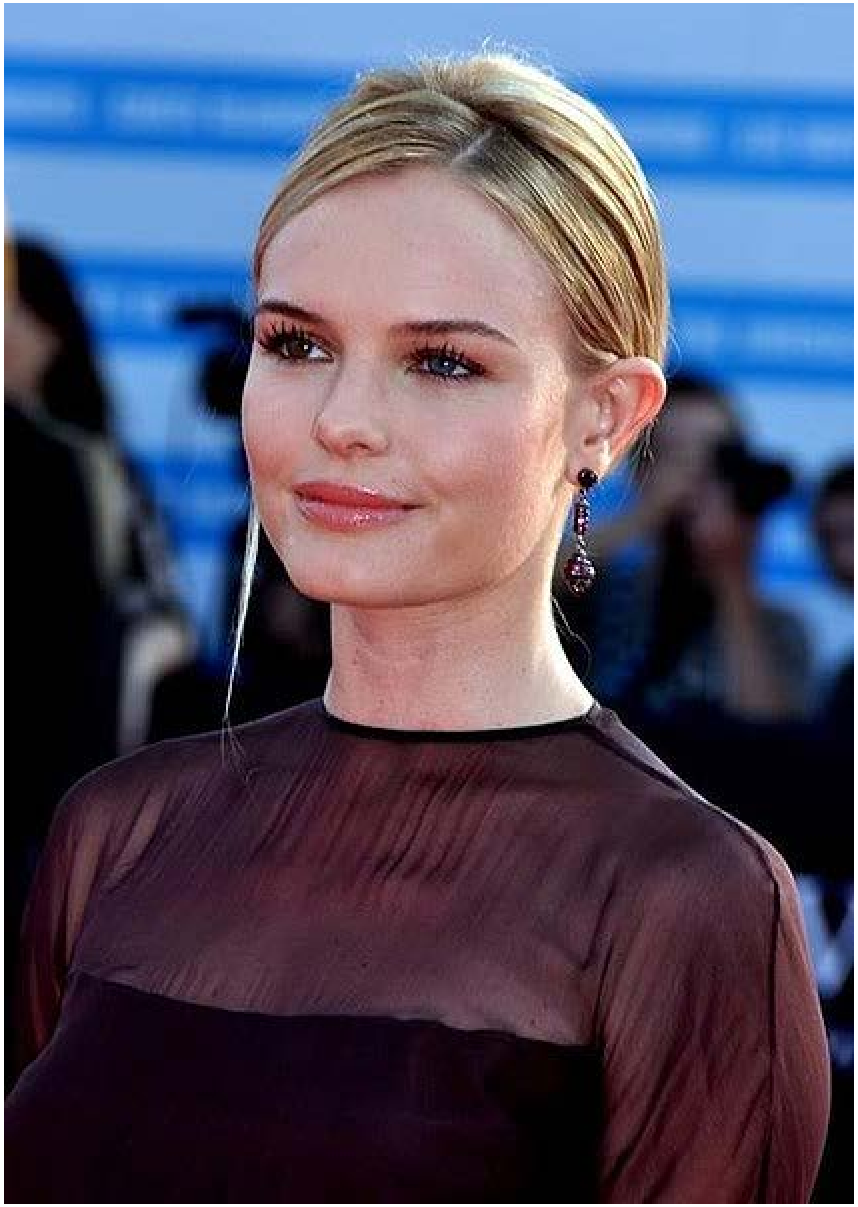
Mosaicism: presence of multiple populations of cells with different genotypes in one individual



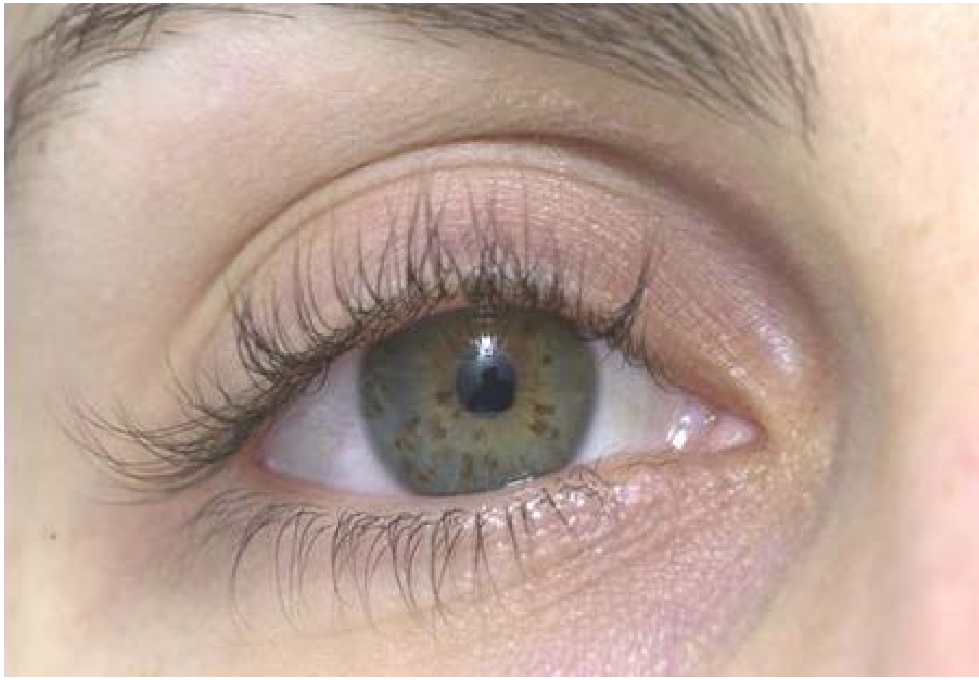
heterochromia

~~Persian cat~~

Van kedisi

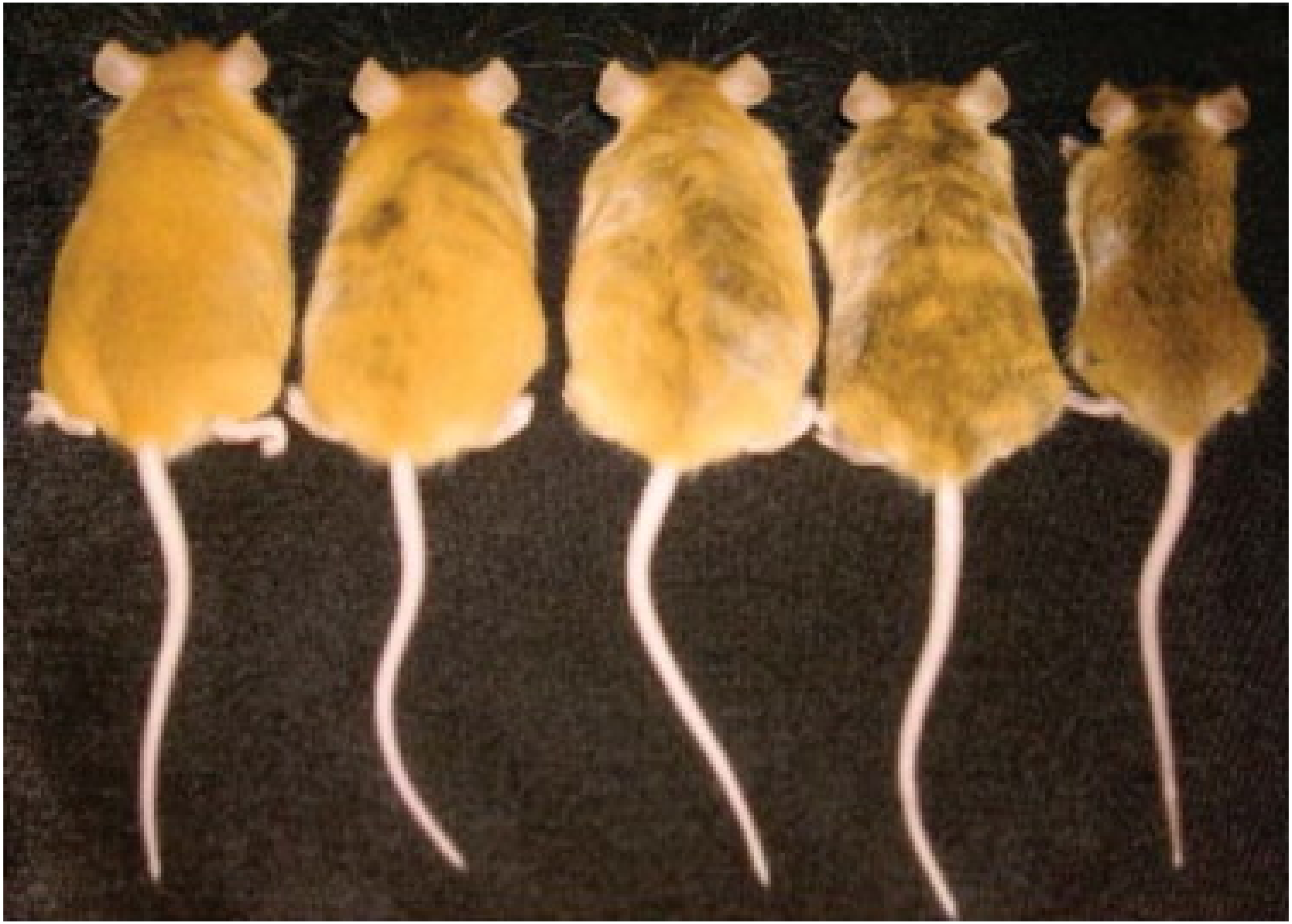


Complete heterochromia

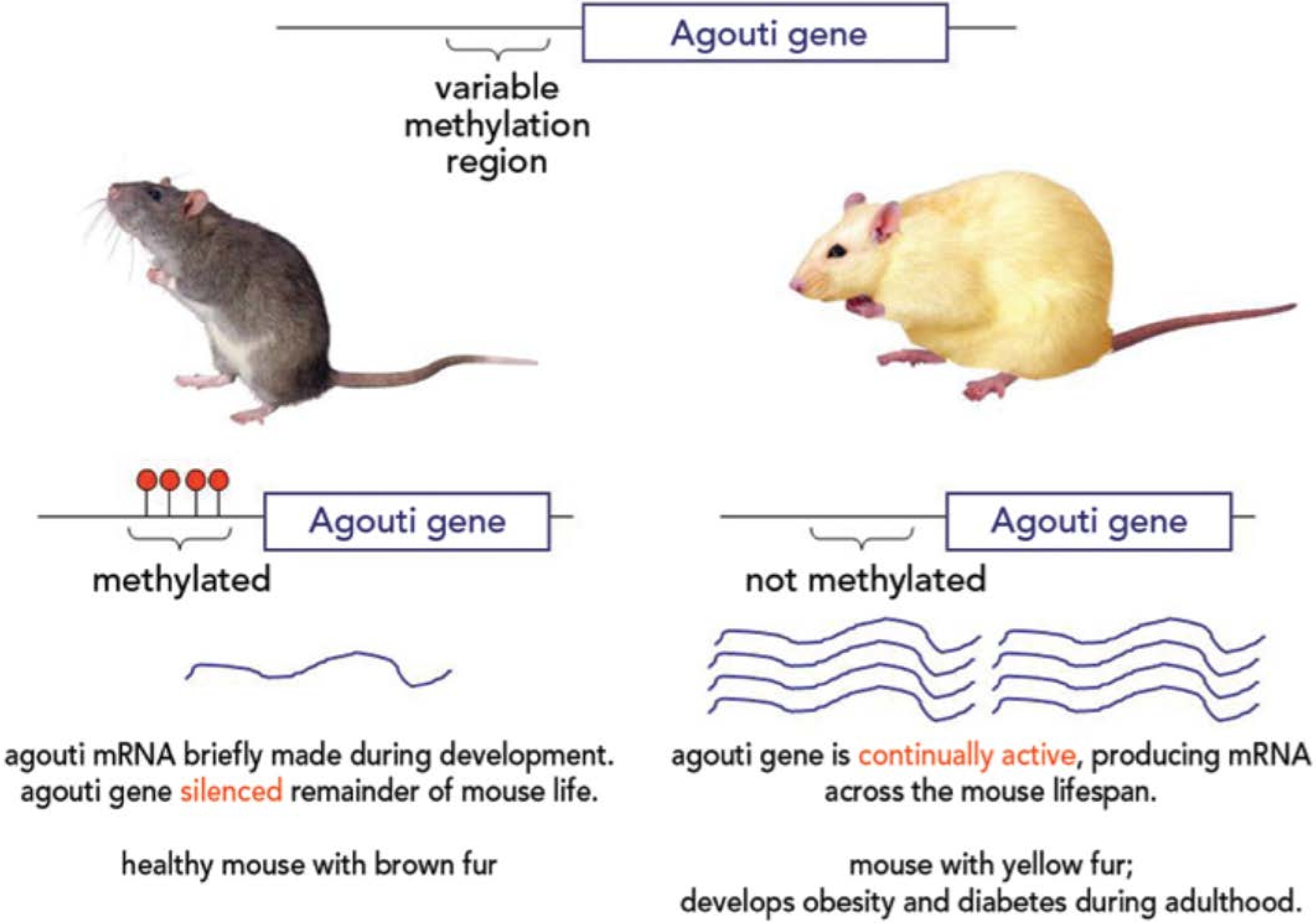


Sectoral heterochromia

Genetically Identical Agouti Mice Littermates



Genetically Identical Agouti Mice Littermates



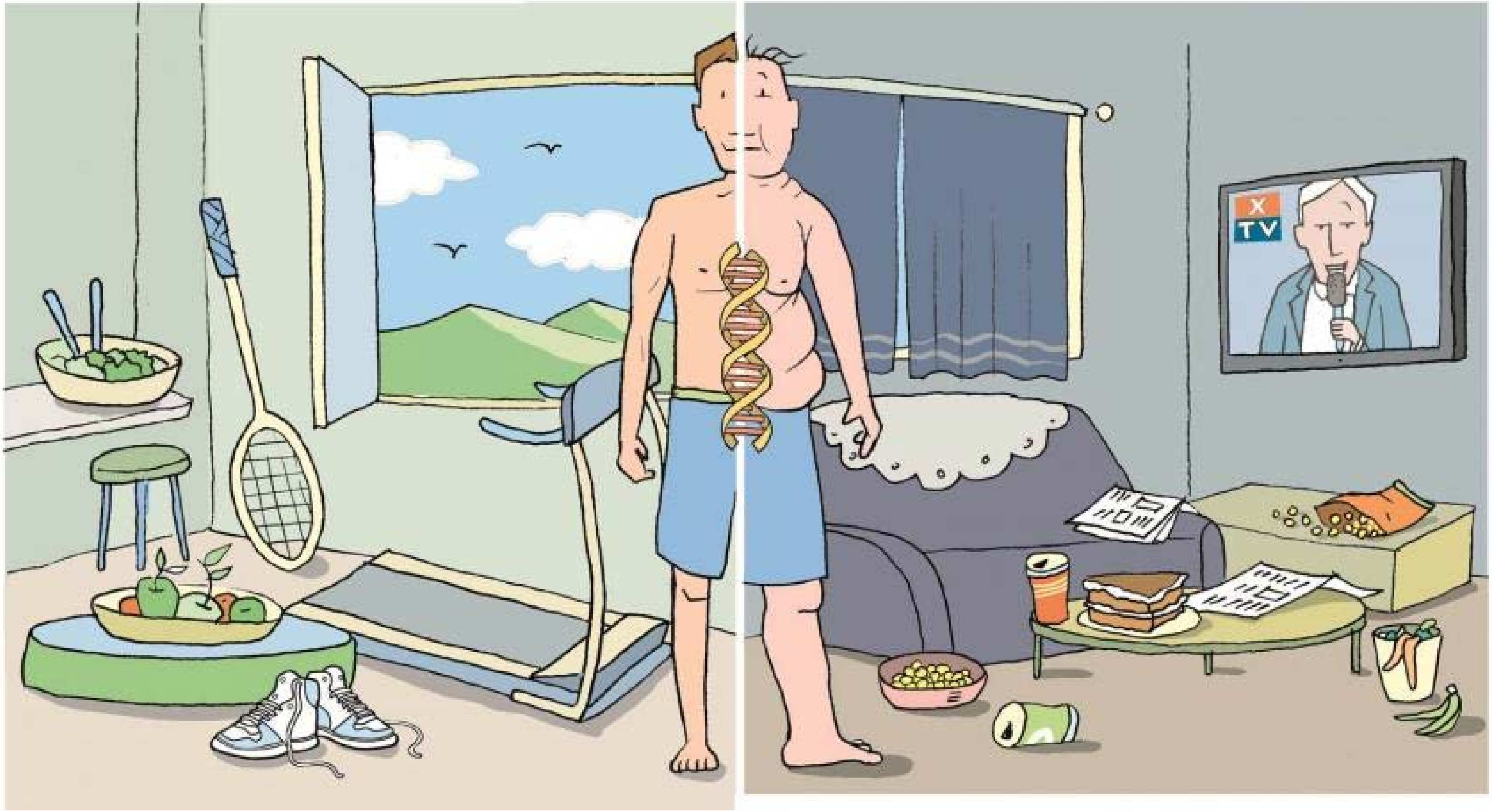
agouti mRNA briefly made during development. agouti gene silenced remainder of mouse life.

healthy mouse with brown fur

agouti gene is continually active, producing mRNA across the mouse lifespan.

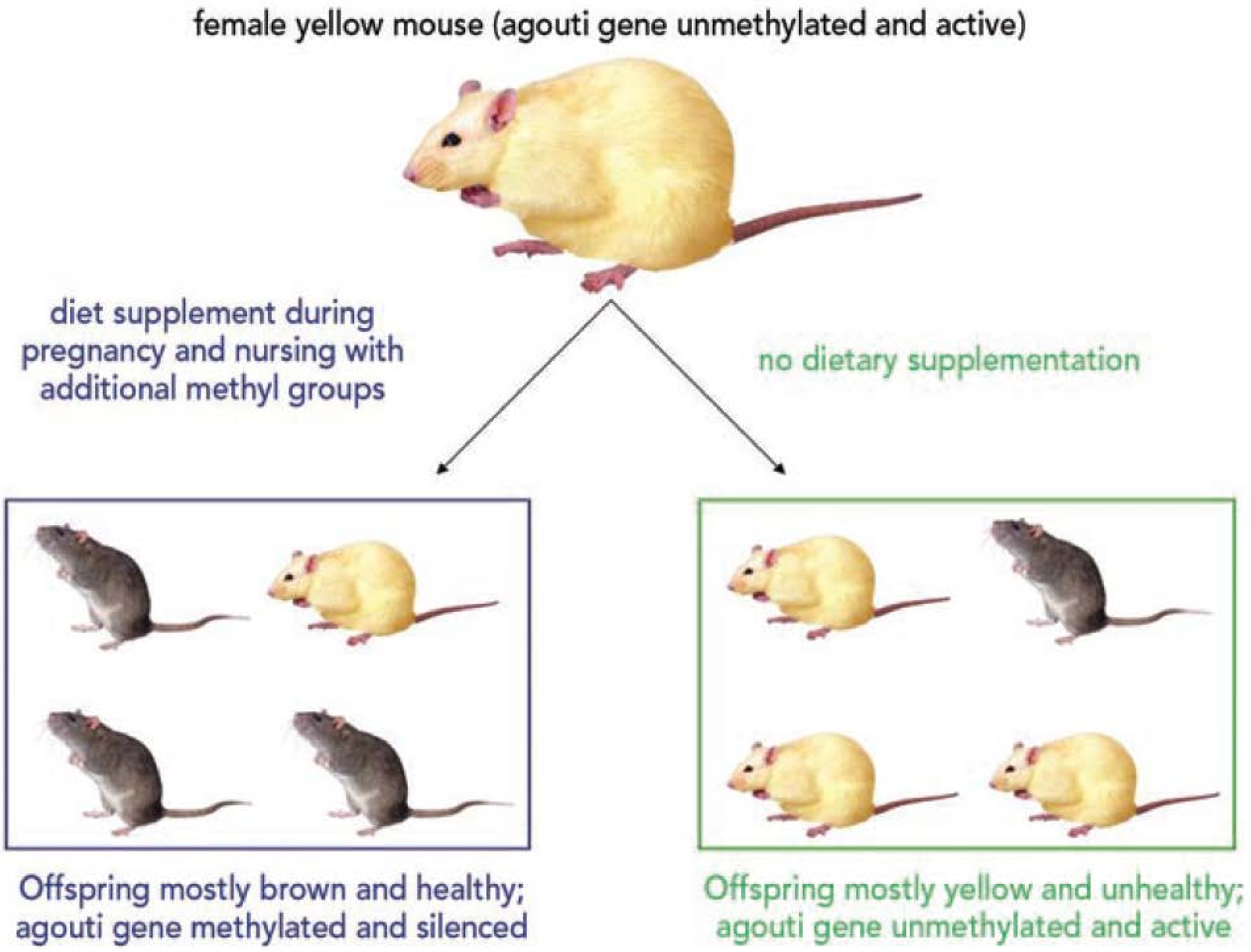
mouse with yellow fur;
develops obesity and diabetes during adulthood.

Environmental effects influence how genes are turned on and off



Credit: Weizmann Institute of Science

Role of Diet in Agouti Mice



female yellow mouse (agouti gene unmethylated and active)

diet supplement during pregnancy and nursing with additional methyl groups

no dietary supplementation

Offspring mostly brown and healthy; agouti gene methylated and silenced

Offspring mostly yellow and unhealthy; agouti gene unmethylated and active

The Dutch Famine (Hongerwinter)

- German's blocked food to the Dutch in the winter of 1944.
- Calorie consumption dropped from 2,000 to 500 per day for 4.5 million.
- Children born or raised in this time were small, short in stature and had many diseases including, edema, anemia, diabetes and depression.
- The Dutch Famine Birth Cohort study showed that women living during this time had children 20-30 years later with the same problems despite being conceived and born during a normal dietary state.
- Also when these children grew up and had children those children were thought to also be smaller than average

Slide adapted from Doug Brutlag - Stanford:
http://biochem158.stanford.edu/Epigenetics.html

Recap

- Changes in the epigenome do not change a gene’s sequence (DNA sequence in general), but rather its activity status.

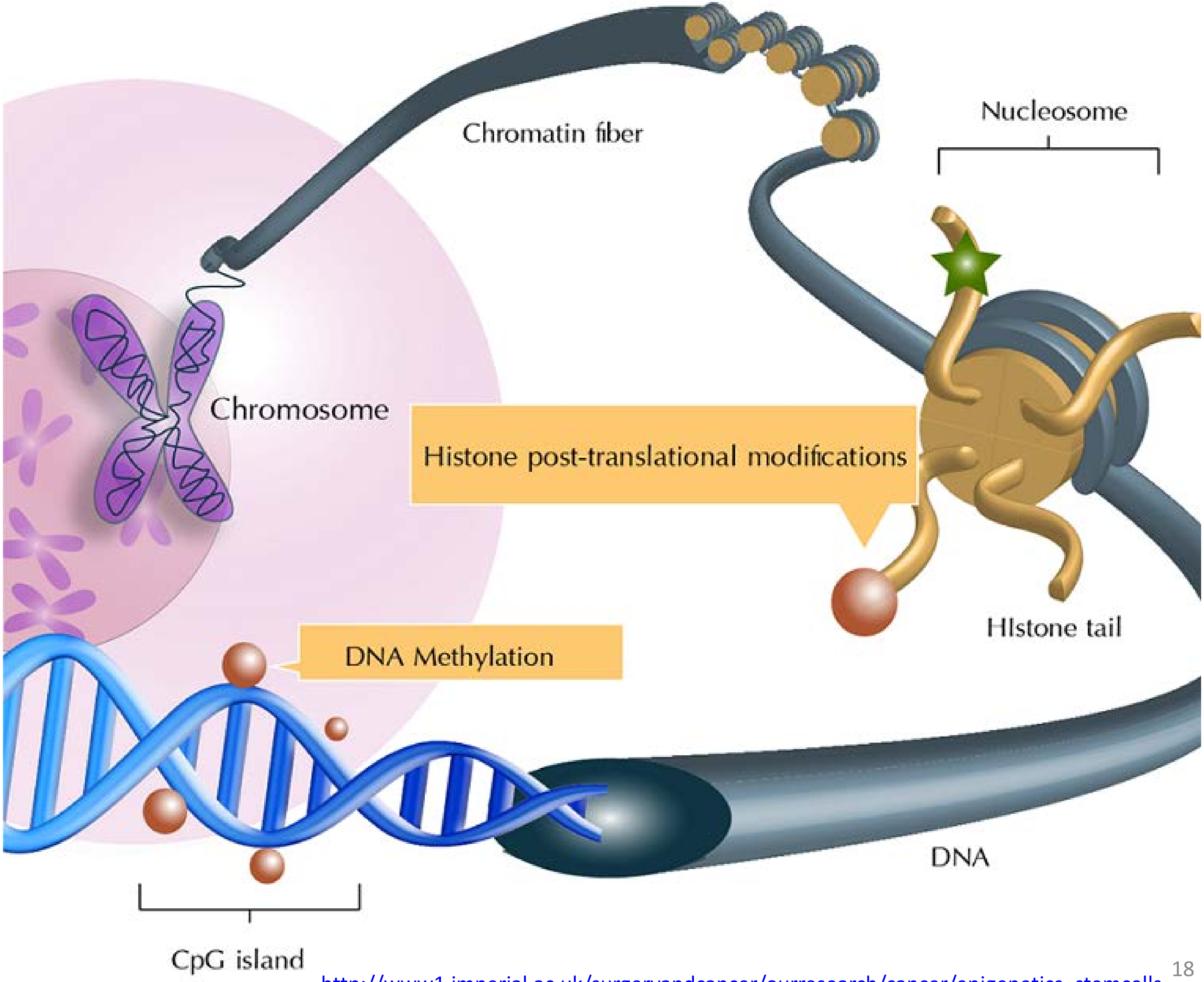
- Genes can switch between active (directing protein production) or silent (no protein produced) phases.

- Patterns of activation and silencing, known as the epigenome, exist across all the genes in a cell.

- The environment can alter the epigenome, changing the activity level of genes.

- Some environmental factors, such as diet, not only change an individual’s epigenome, but appear to influence the epigenome of future generations.

Nucleus of a cell



http://www1.imperial.ac.uk/surgeryandcancer/ourresearch/cancer/epigenetics_stemcells

epigeneticmodificationscanbeconsideredasthepunctuationmarksinthe genomealackofpriorknowledgemakesthechallengegreater

Epigenetic modifications can be considered as the punctuation marks in the genome. A lack of prior knowledge makes the challenge greater.

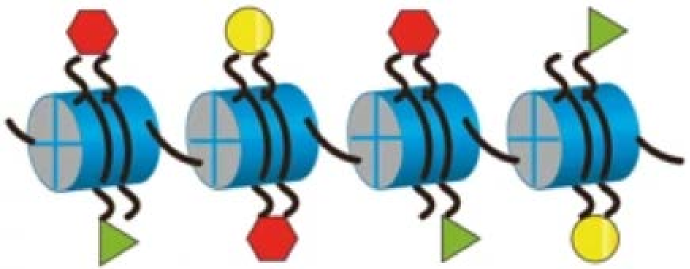
Epigenetic marks

- Demarcate the start and end of genes, like the start and end of sentences and words in the sentence
- Provide structure to the chromosome, like paragraph breaks or chapter breaks
- Alter how we read each and every gene, like the punctuation marks in each sentence
- Lead to genes being expressed (active) or not expressed (silent), or more subtle changes (fine tuning)

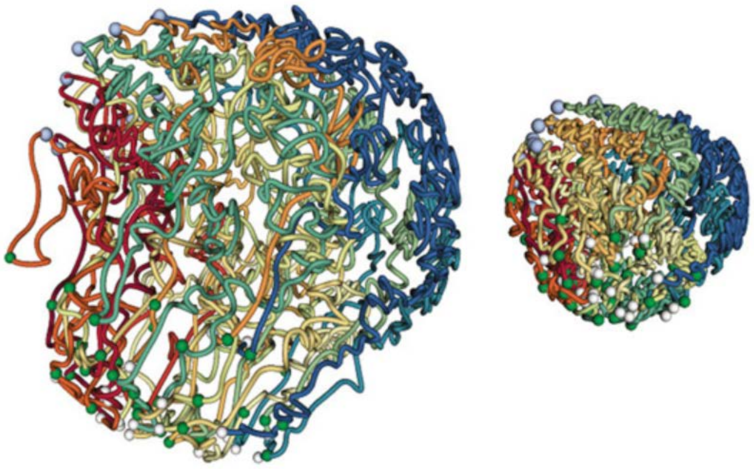
Part 1: DNA Methylation

Me
Me
C
G
A
T
C
A
C
G
T
C

Part 2: Nucleosome Positioning and Histone Modifications



Part 3: Three-dimensional Structure and Folding of the Genome

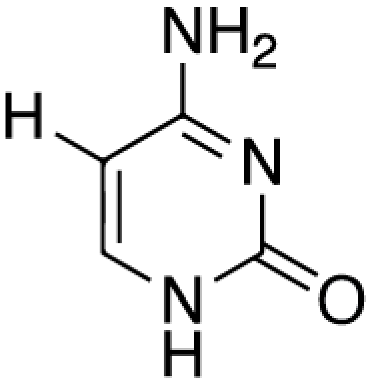


Part 1: DNA Methylation

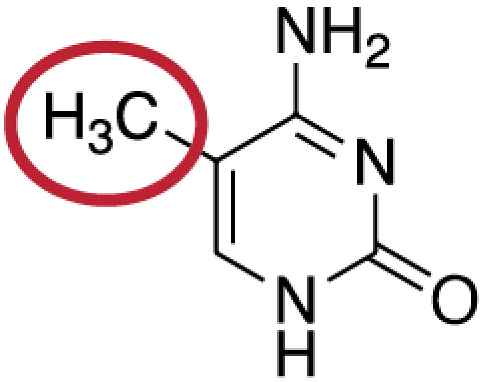
Me
Me
C
G
A
T
C
A
C
G
T
C

- Establishment and maintenance of DNA methylation
- Inheritance of DNA methylation
- DNA demethylation
- Bisulfite conversion for detecting DNA methylation
- Exercise: Simulation and alignment of WGBS reads

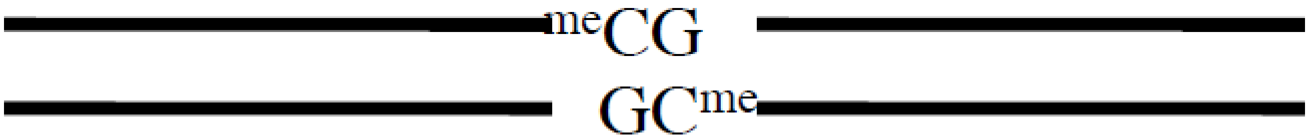
Addition of a methyl group to DNA



Cytosine

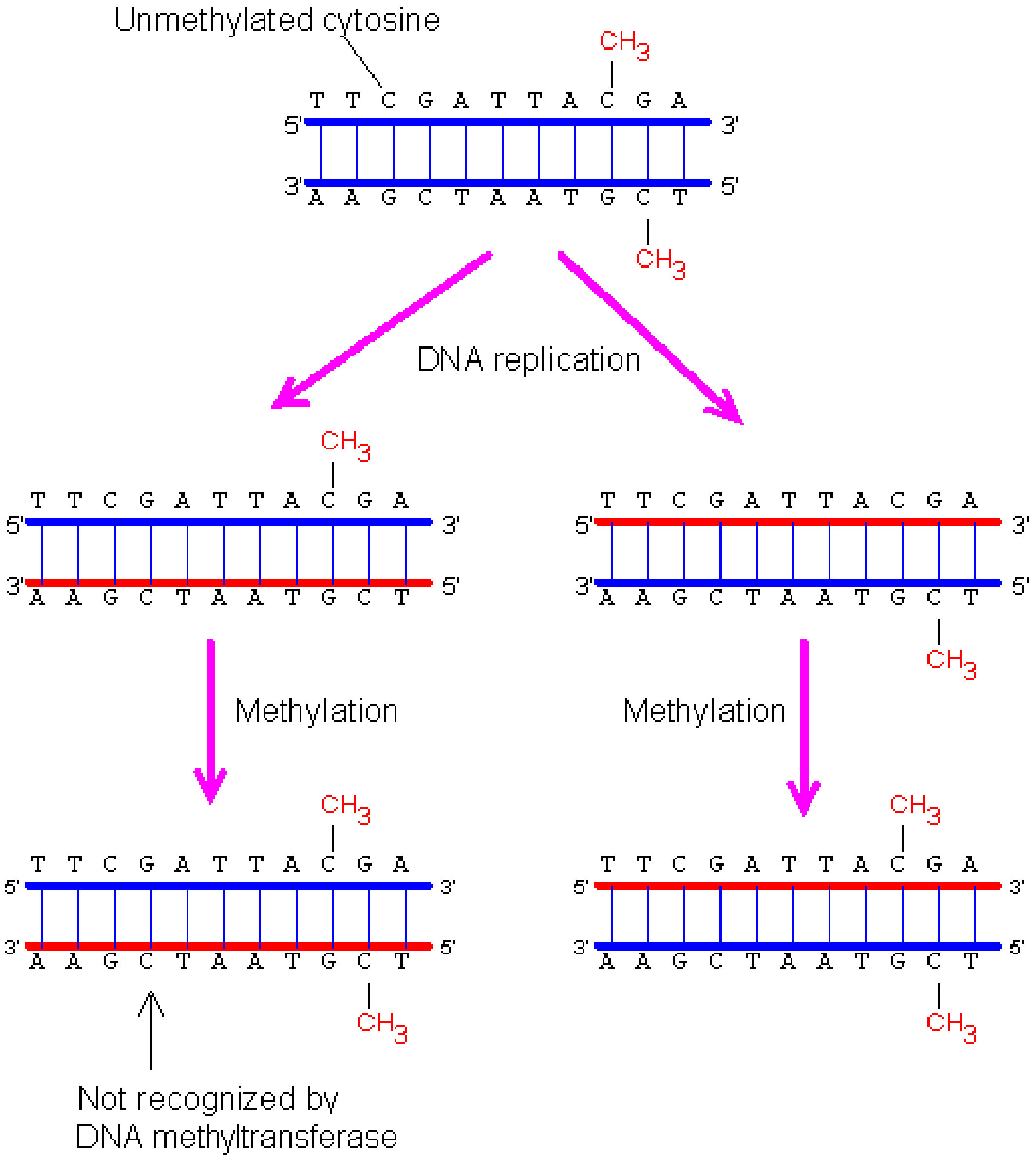


methylated Cytosine



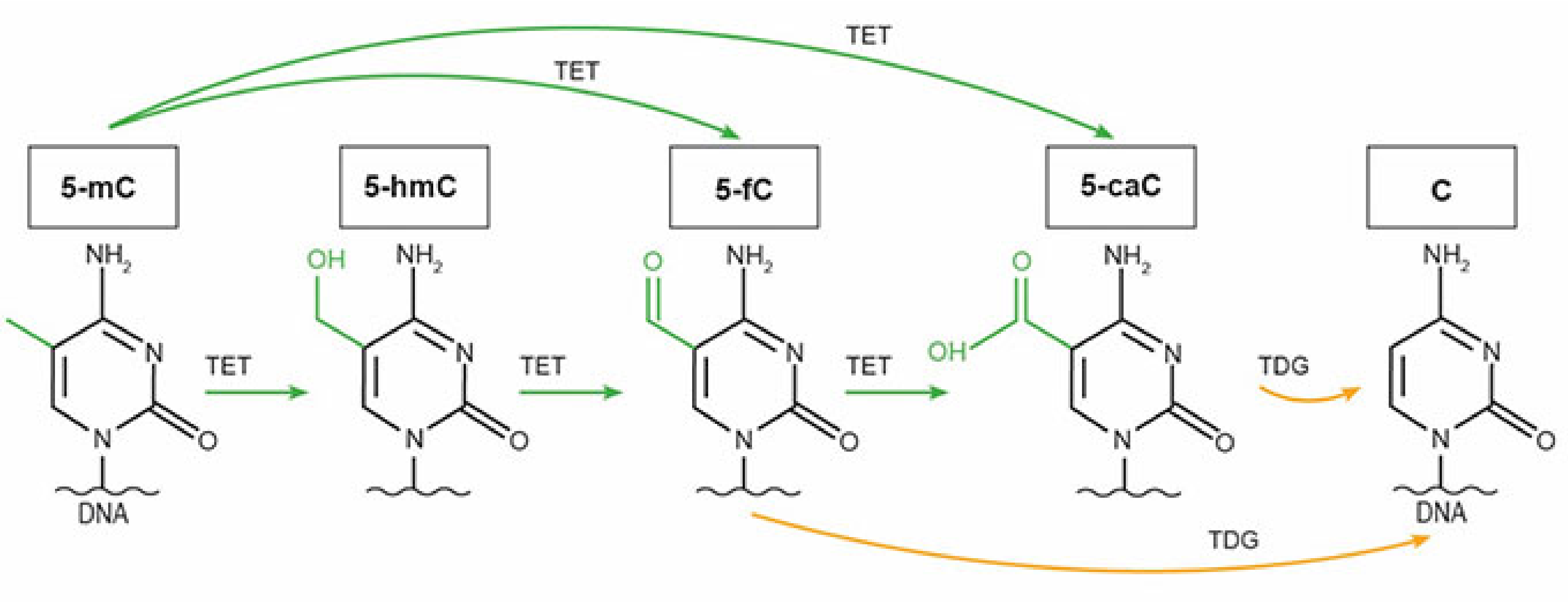
Symmetric DNA methylation at CpG dinucleotides established *de novo* by enzymes **DNMT3a** and **DNMT3b** in mammals

Inheritance of DNA methylation

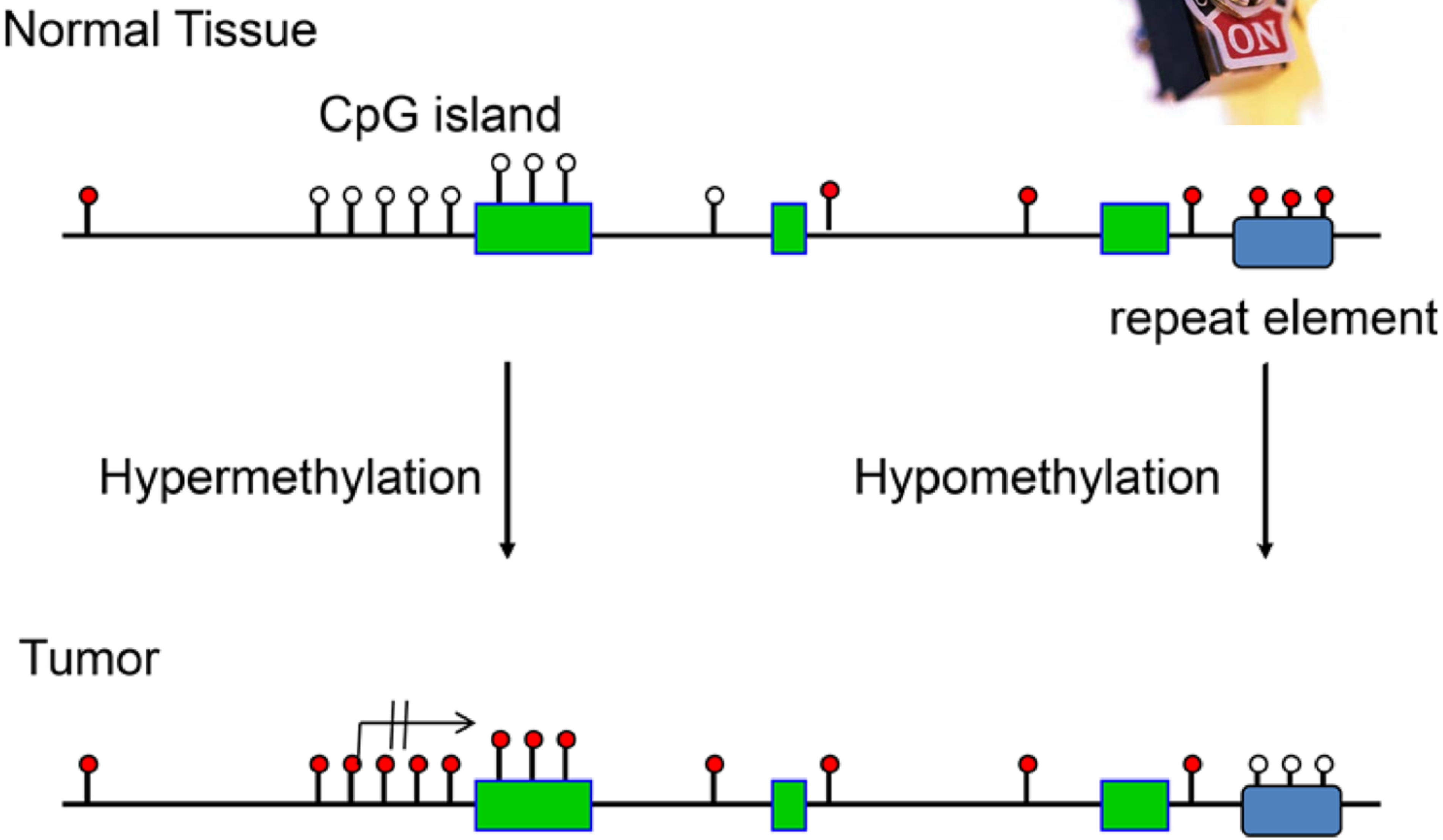


Hemi-methylated DNA is recognized by DNMT1 (maintenance)

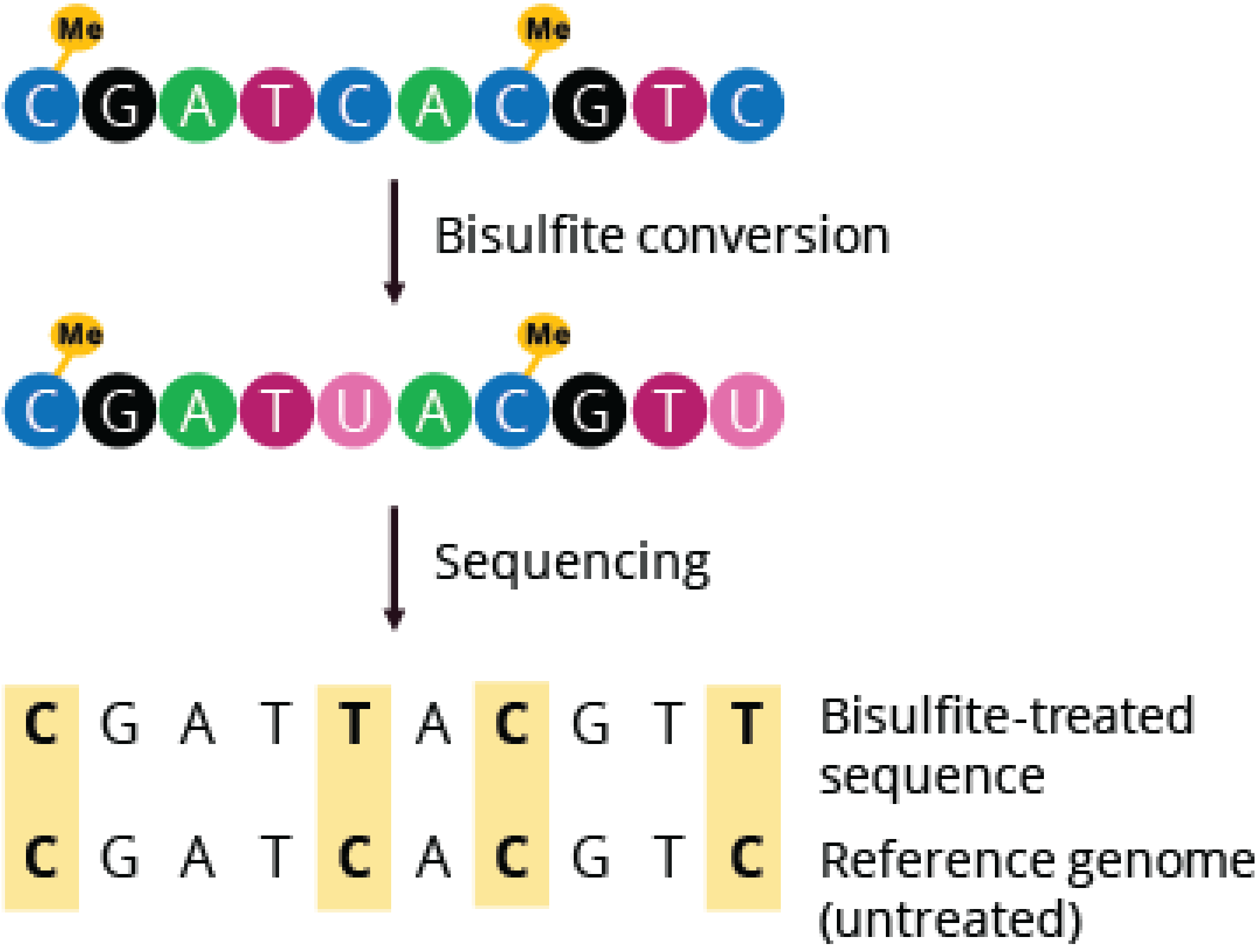
Active DNA demethylation



Why does it matter?

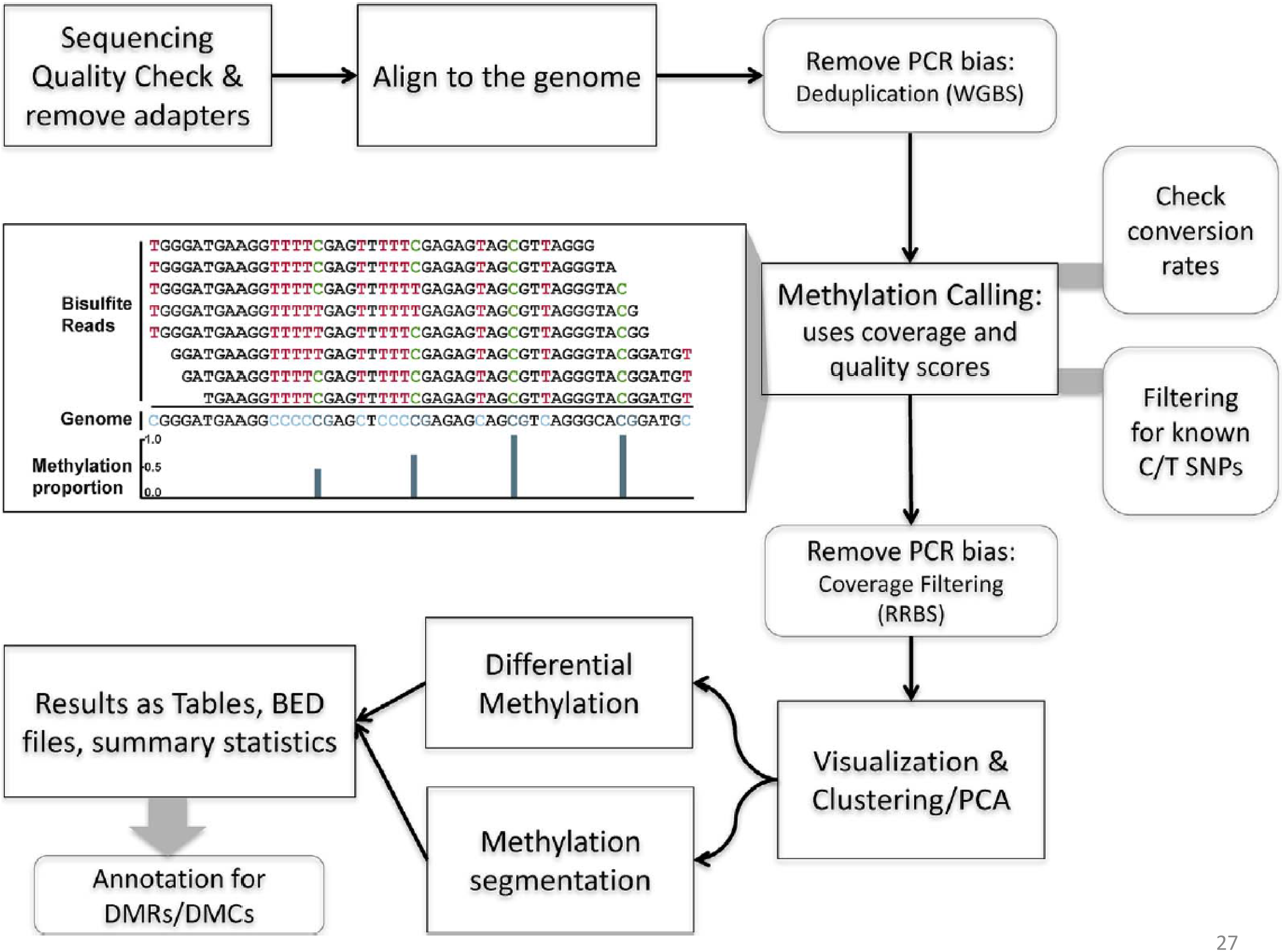


How do we detect methylated vs unmethylated DNA?



C G A T T A C G T T Bisulfite-treated sequence

C G A T C A C G T C Reference genome (untreated)



Sequencing Quality Check & remove adapters

Align to the genome

Remove PCR bias: Deduplication (WGBS)

Check conversion rates

Methylation Calling: uses coverage and quality scores

Filtering for known C/T SNPs

Remove PCR bias: Coverage Filtering (RRBS)

Visualization & Clustering/PCA

Differential Methylation

Methylation segmentation

Results as Tables, BED files, summary statistics

Annotation for DMRs/DMCs

Bisulfite Reads
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGG
TGGGATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTA
TGGGATGAAGGTTTTCGAGTTTTTTGAGAGTAGCGTTAGGGTAC
TGGGATGAAGGTTTTTGAGTTTTTTGAGAGTAGCGTTAGGGTACG
TGGGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGG
GGATGAAGGTTTTTGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
GATGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
TGAAGGTTTTCGAGTTTTTCGAGAGTAGCGTTAGGGTACGGATGT
Genome
CGGGATGAAGGCCCCCGAGCTCCCCGAGAGCAGCGTCAGGGCACGGATGC
Methylation proportion
1.0
0.5
0.0

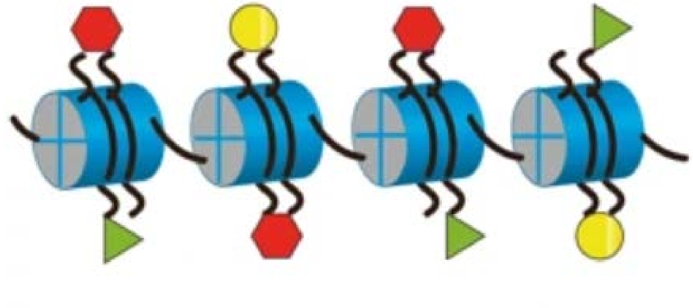
Exercise: Quantification of DNA methylation levels from WGBS

Reference genome:
CGGGATGAAGGCCCCCGAGCTCCCCGAGAGCAGCGTCAGGGCACGGATGC

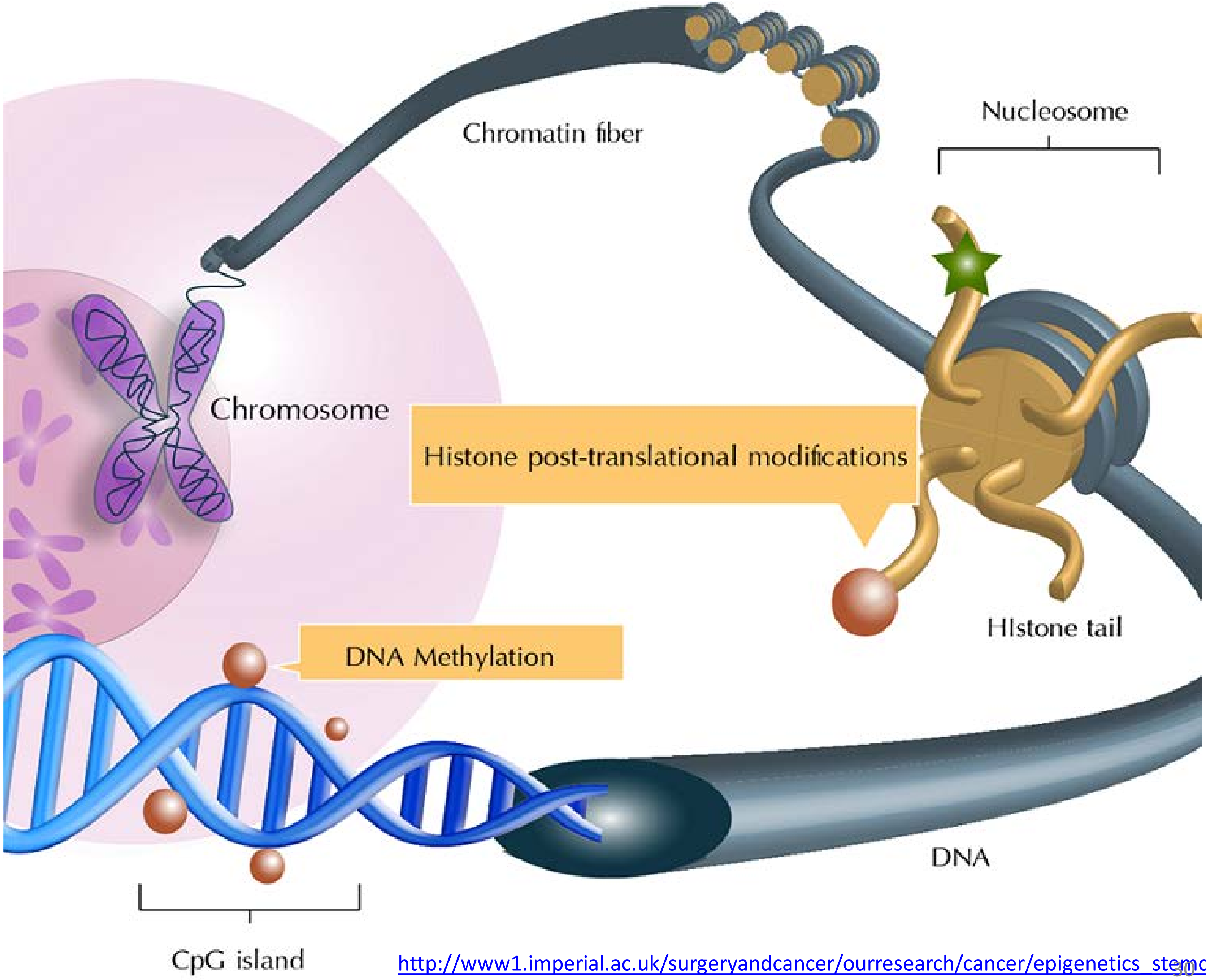
1. Take this reference genome and pick randomly n=100 substrings (i.e., simulated short read), each of length say k=8 bp
2. For each such read check to see if it has a CpG dinucleotide in it
3. For each CG in the substring, flip a biased coin (p=0.6) and if tails/fail change the CpG to TpG (unmethylated CpG)
4. Align the new k bp reads (what would come out of the sequencer for a WGBS experiment) back to reference genome allowing 1 mismatch
5. Count the number of reads that overlap each CpG with an exact match (ref CG – read CG) or a 1-bp mismatch (ref CG – read TG)
6. Report the ratio of C/(C+T) as the methylation level of each CpG

Big thanks to Abhijit Chakraborty who wrote the initial version of the R code

Part 2: Nucleosome Positioning and Histone Modifications

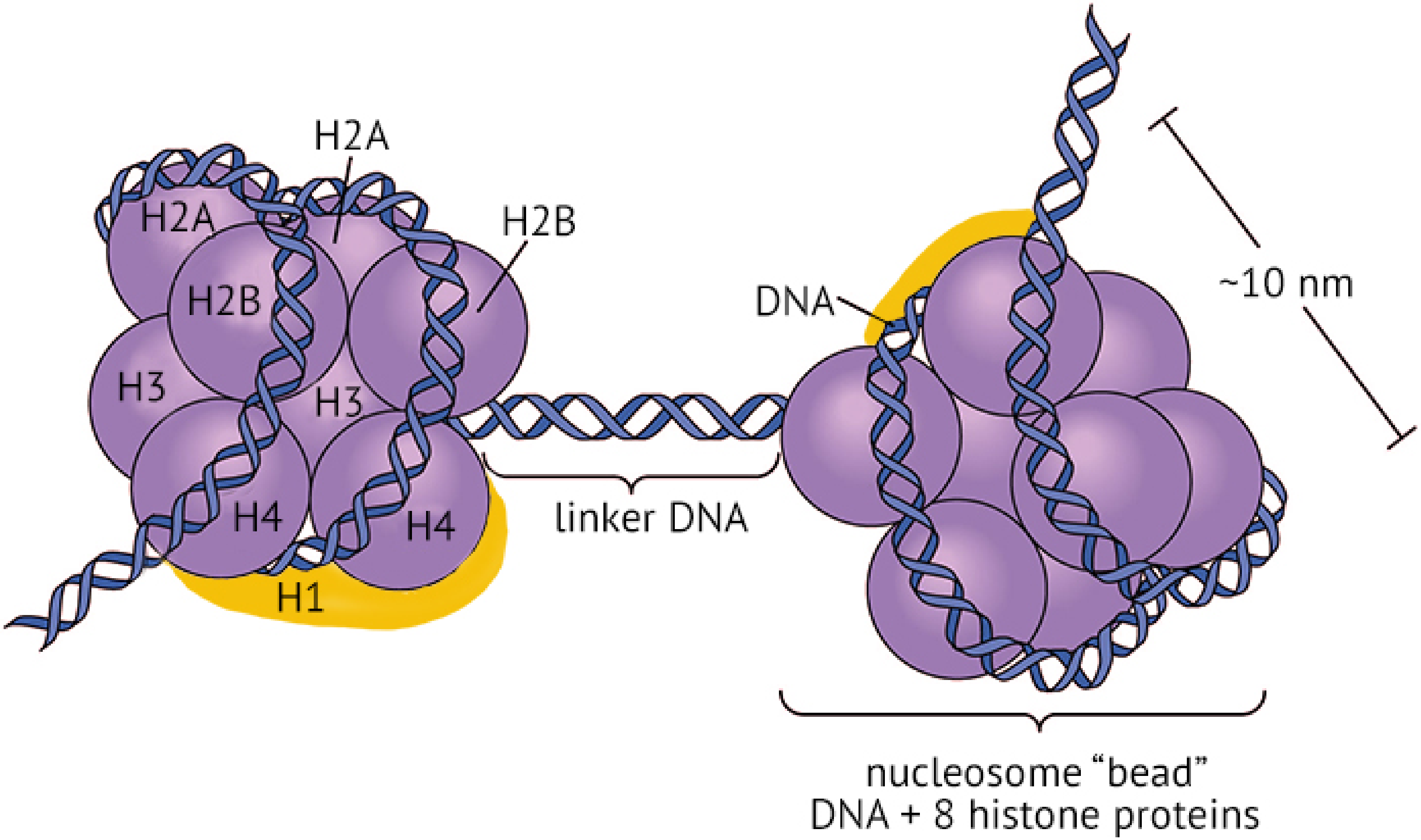


- Nucleosomes
- Histone code
- Different types of histone modifications
- The concept of euchromatin vs heterochromatin
- ChIP-seq for histone modifications
- Exercise: Genome Browser visualization of ChIP-seq data

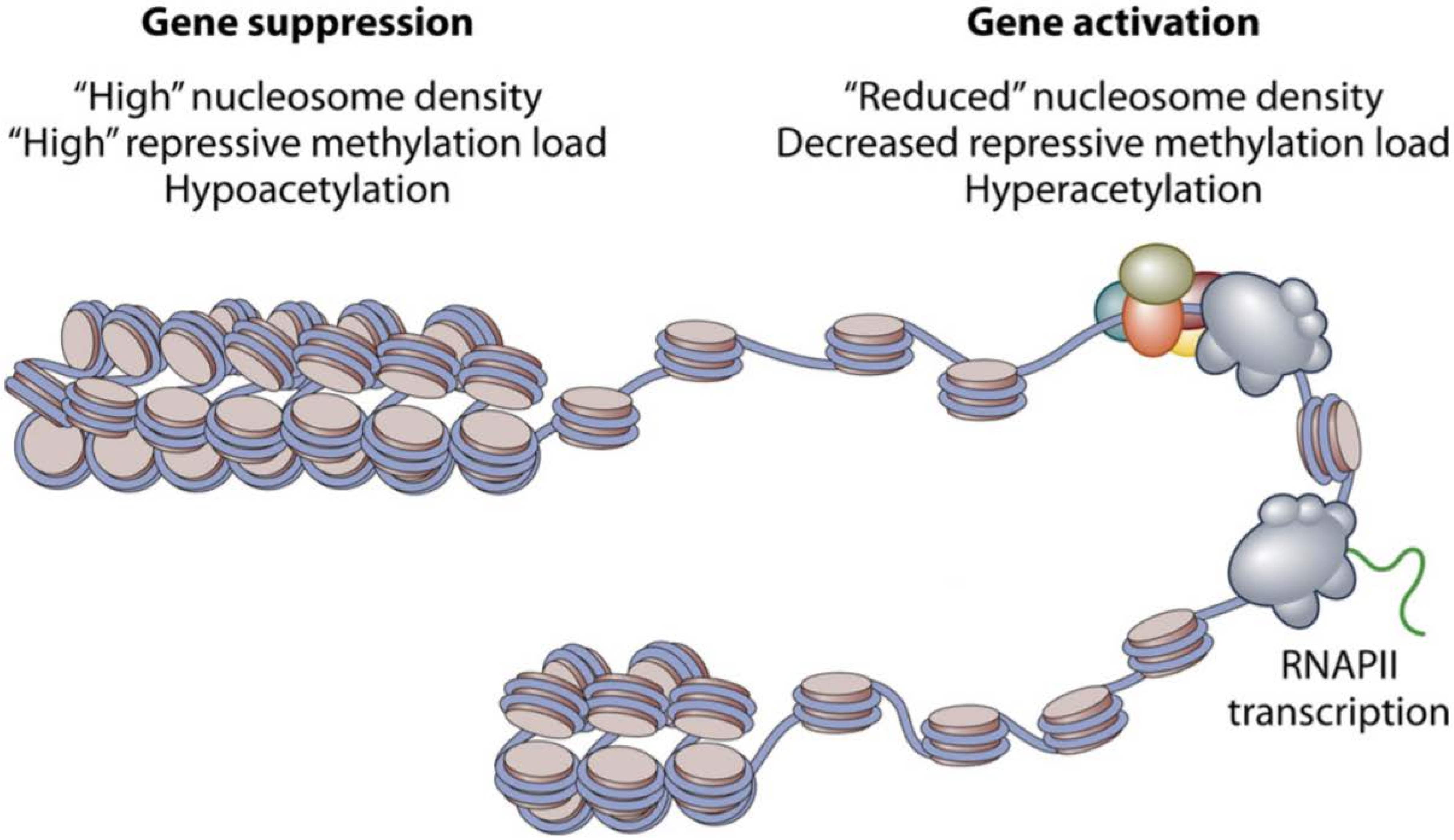


http://www1.imperial.ac.uk/surgeryandcancer/ourresearch/cancer/epigenetics_stemcells

Nucleosome structure

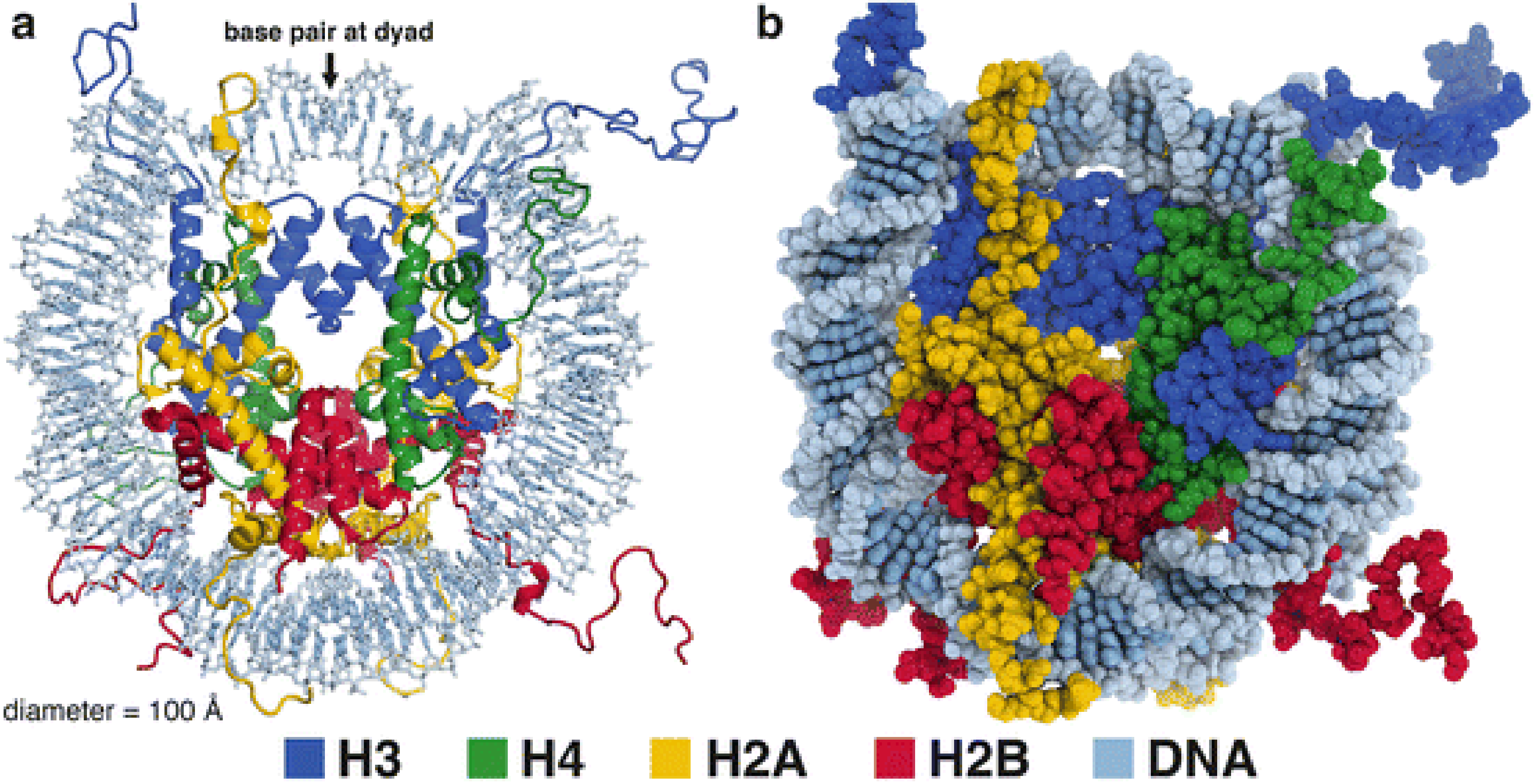


Nucleosome density and positioning

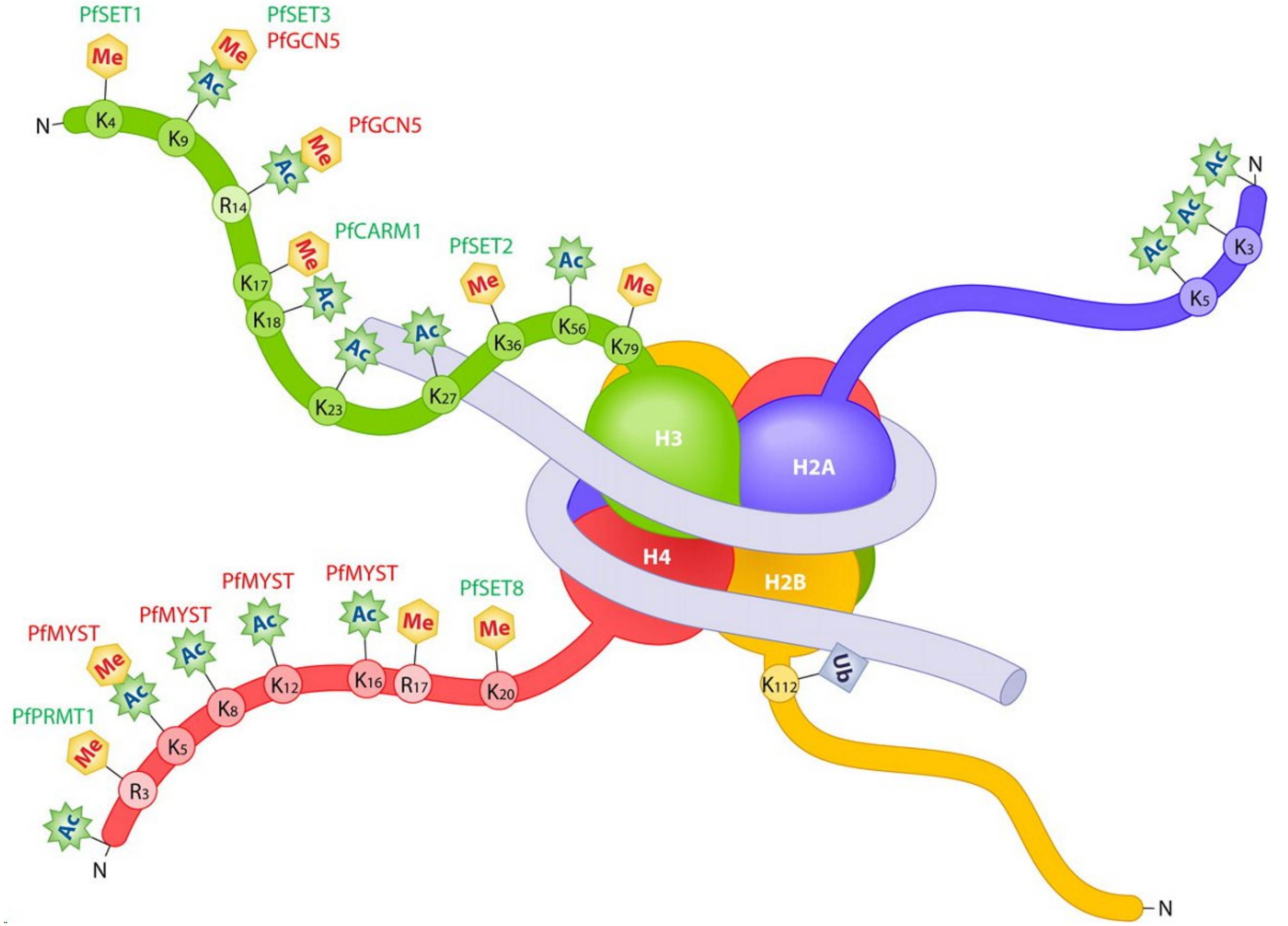

Gene suppression
"High" nucleosome density
"High" repressive methylation load
Hypoacetylation
Gene activation
"Reduced" nucleosome density
Decreased repressive methylation load
Hyperacetylation
RNAPII
transcription

Kristie, mBio , 2016

Histone proteins



Histone code



- Predominantly on the tails of H3 and H4 and on Lysine (K)
- Over 50 sites/residues can be modified
- Some sites can be both Acetylated (K) and Methylated (R,K)

Histone acetylation

- Acetyl groups are laid on the histones by **histone acetyltransferases (HATs)**, and are removed by **histone deacetylases (HDACs)**

- Histone acetylation is positively correlated with gene activity

- Acetylation reduces positive charge of histones, neutralizes positive lysine residues and decreases attraction between +ve charged histones and –ve charged DNA

- Acetylated histones act as docking sites for other proteins, which further open the chromatin or recruit other proteins that do so

- Very dynamically established and removed

- No clear mechanism for inheritance on its own (unlike DNA methylation)

Histone methylation

- Methyl groups are laid on the histones by **lysine methyltransferases (HMT/KMT)** and are removed by **lysine demethylases (HDM/KDM)** which are specific to a particular residue (H3K4, H3K9, H3K27)

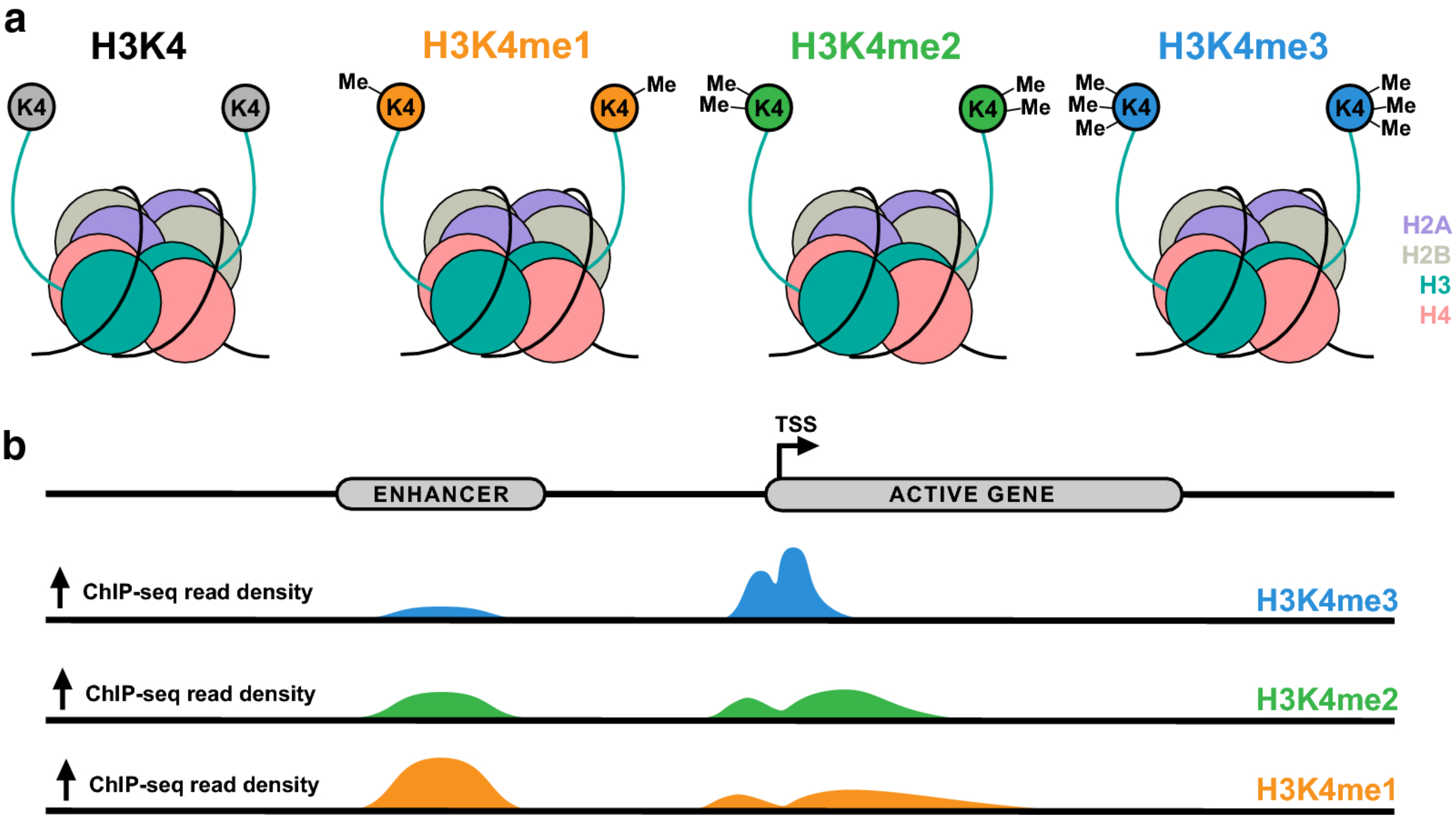
- Methylation can happen in mono, di or tri form (me1/2/3)

- Methylation does not change the electrical charge of histones

- Histone methylation can be positively (H3K4me1/2/3) or negatively correlated with gene activity (H3K9me3, H3K27me3)

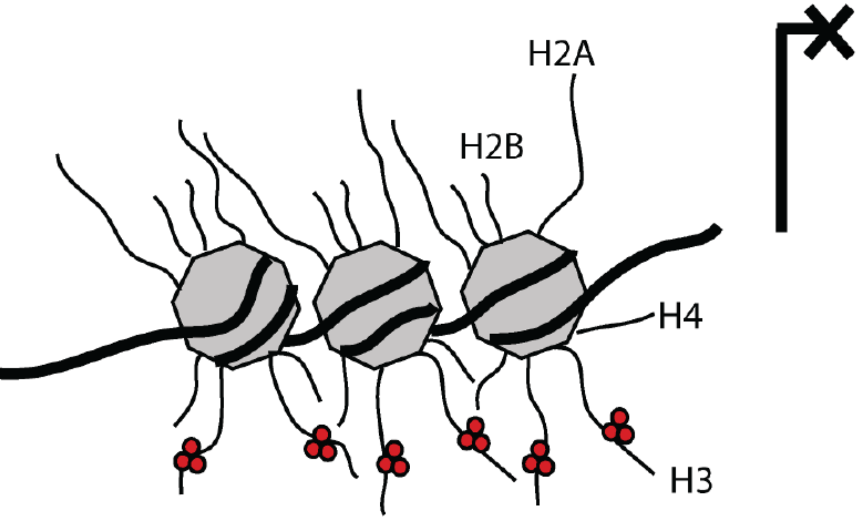
- Repressive histone methylation act as docking site for other proteins (chromodomain) that stabilize the closed/repressive chromatin state

Histone methylation: H3K4 vs H3K9 vs H3K27

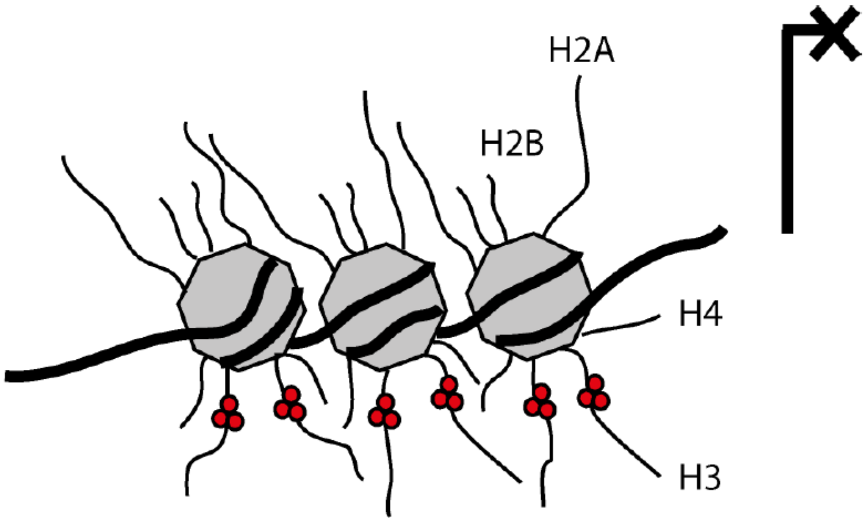


Collins et al. 2019

Histone methylation: H3K4 vs H3K9 vs H3K27

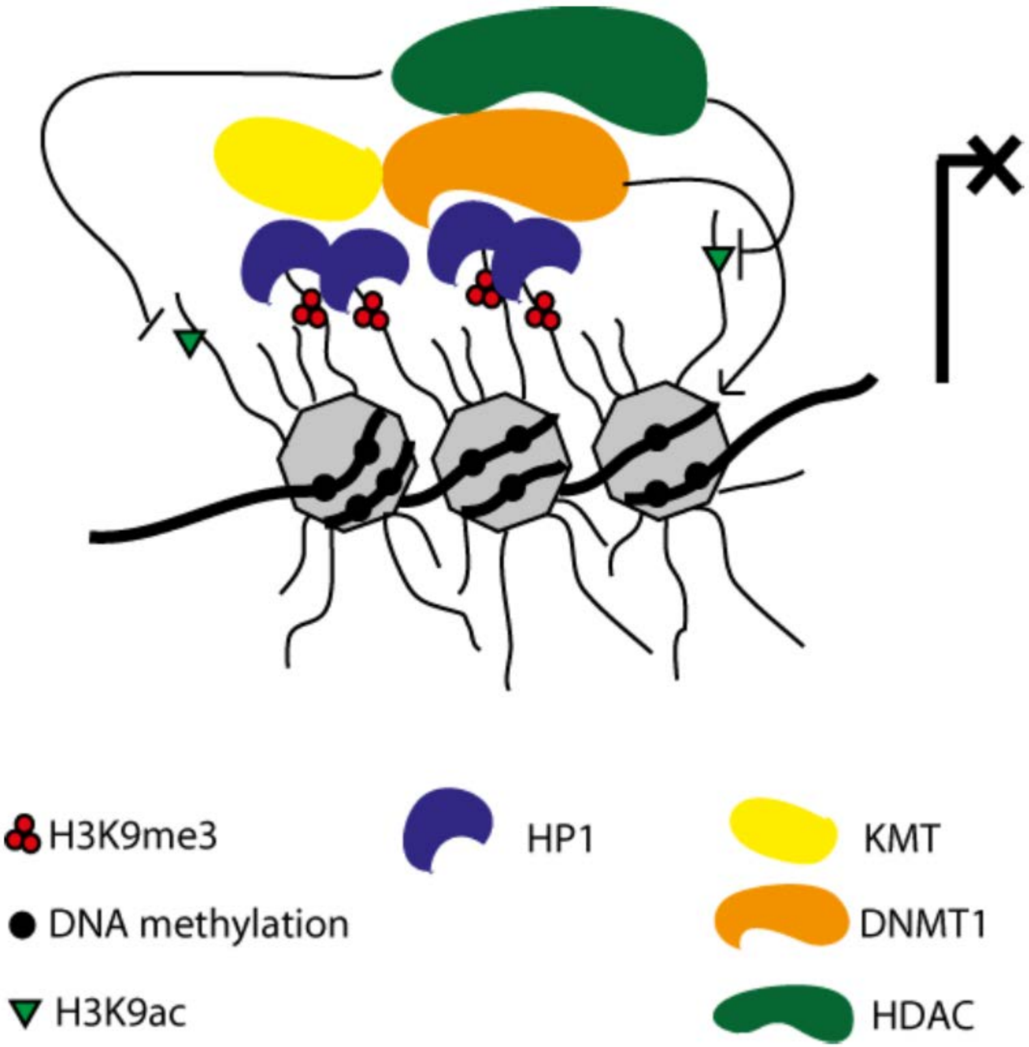


H3K9me - Inactive locus
Spread over the gene
Constitutive heterochromatin

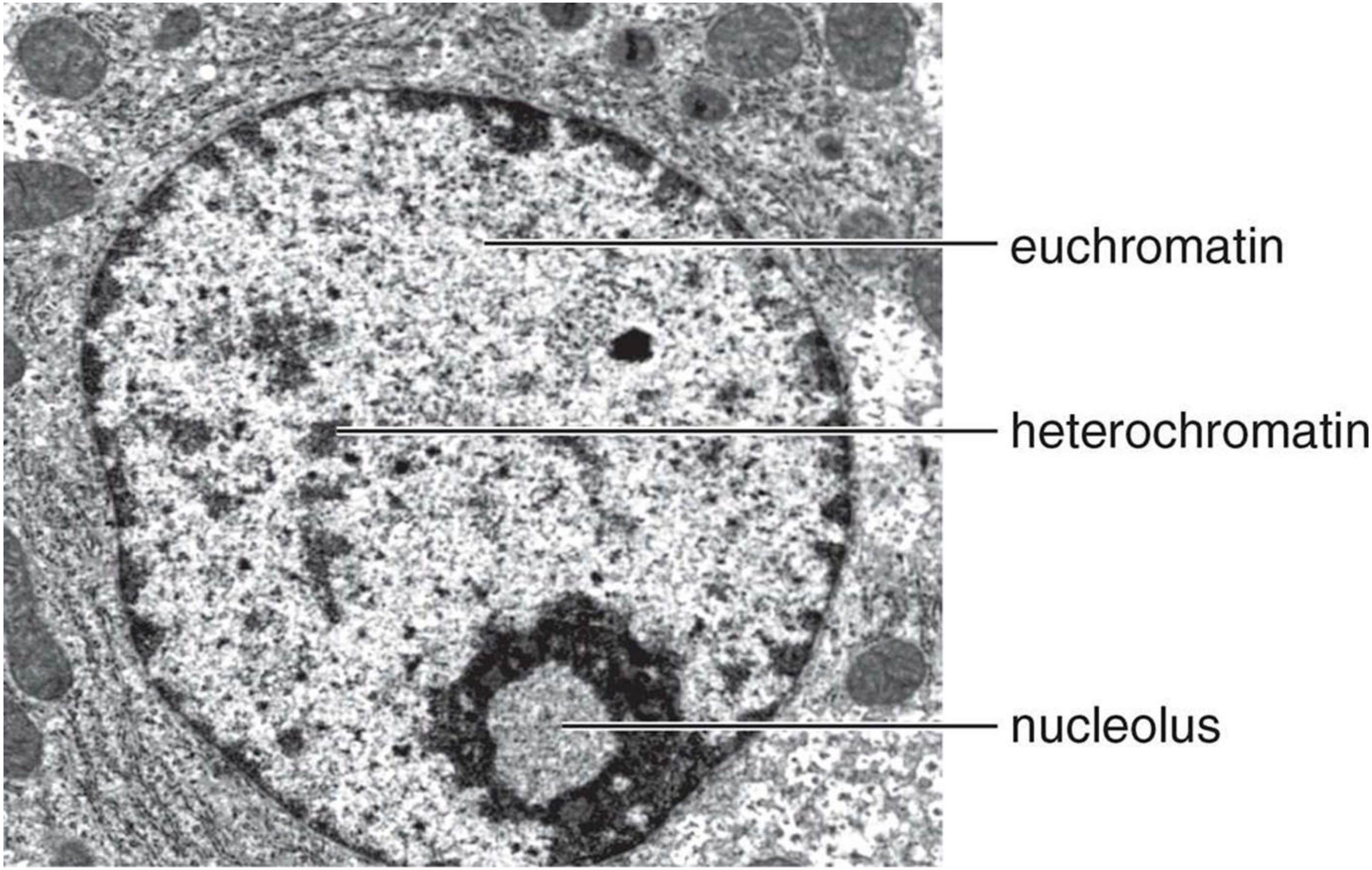


H3K27me - Inactive locus
Spread over the gene
Facultative heterochromatin

Histone methylation: H3K4 vs H3K9 vs H3K27

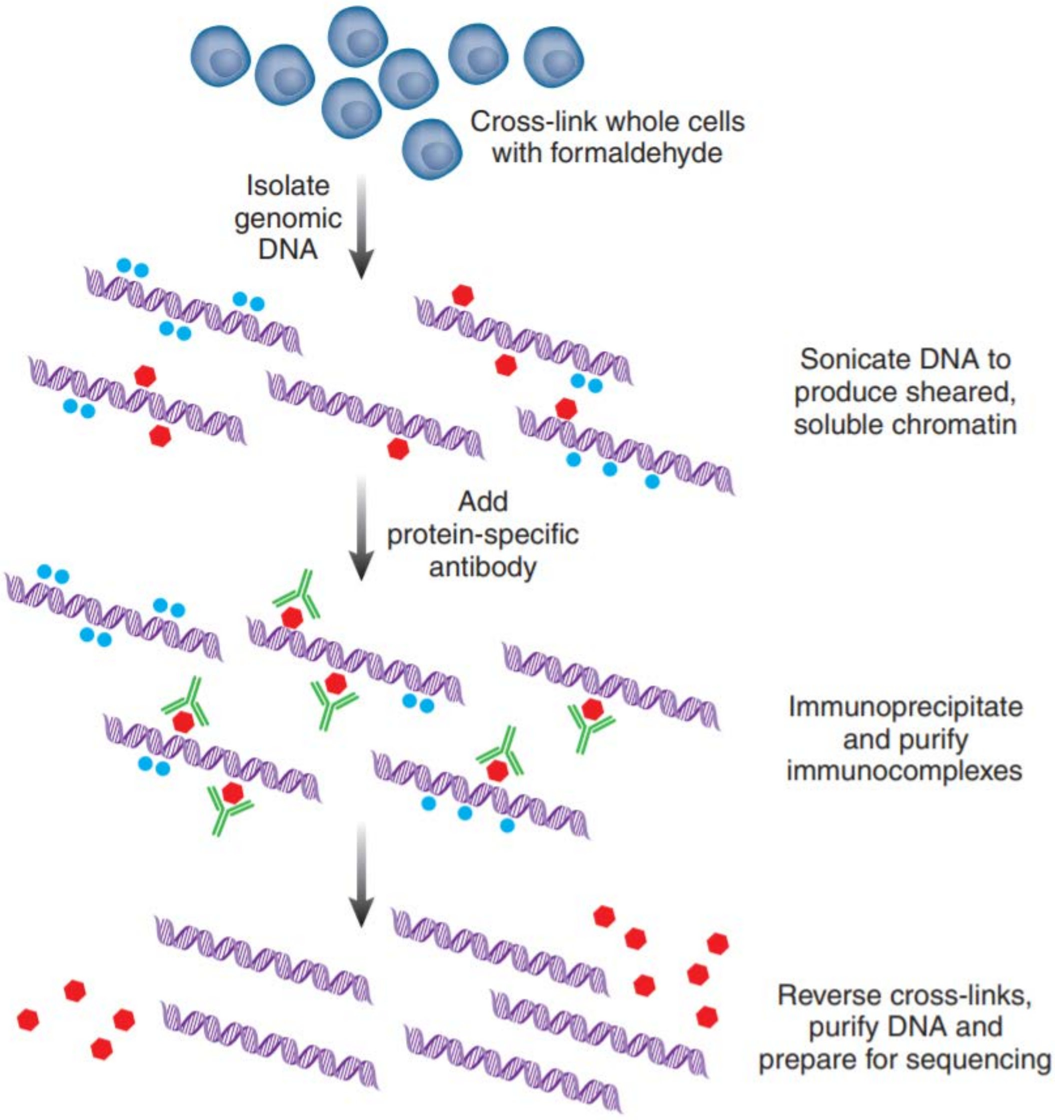


Euchromatin vs heterochromatin

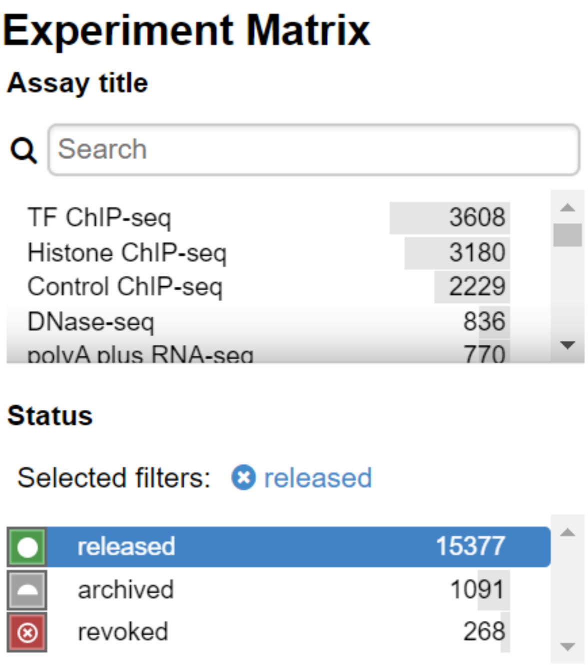


light microscopy

How do we measure histone modifications genome-wide?


Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing

ChIP-seq: Chromatin immunoprecipitation coupled with high-throughput sequencing - Wold lab (2007)



Experiment Matrix

Assay title

Search

TF ChIP-seq	3608
Histone ChIP-seq	3180
Control ChIP-seq	2229
DNase-seq	836
polyA plus RNA-seq	770

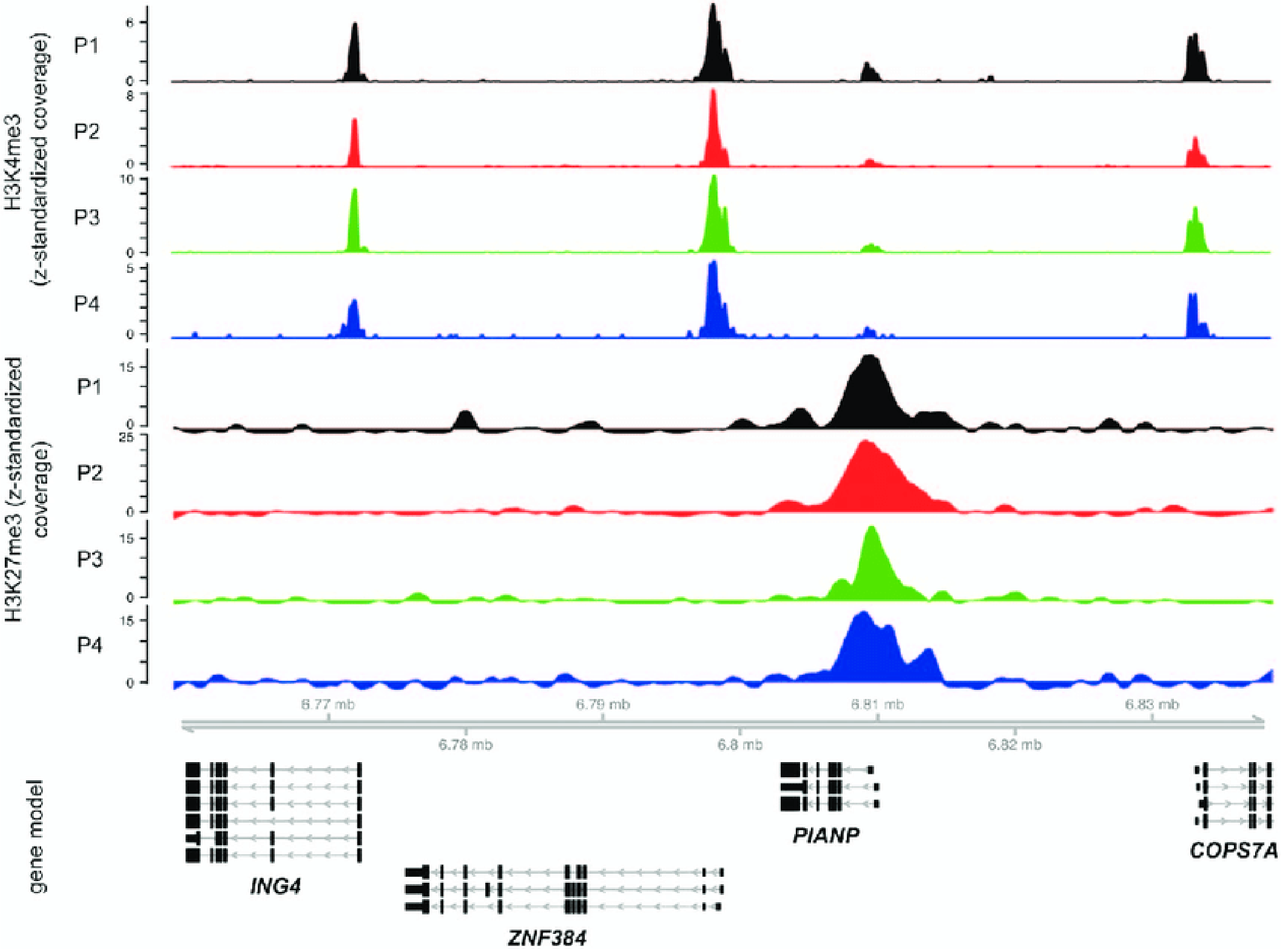
Status

Selected filters: ☒ released

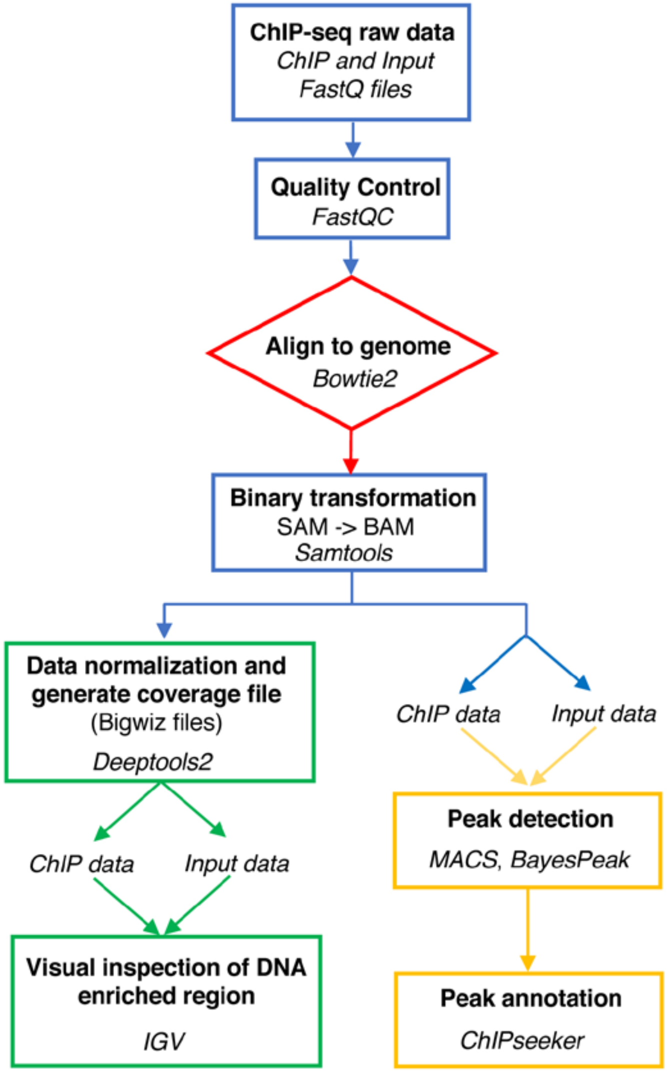
●	released	15377
◓	archived	1091
⊗	revoked	268

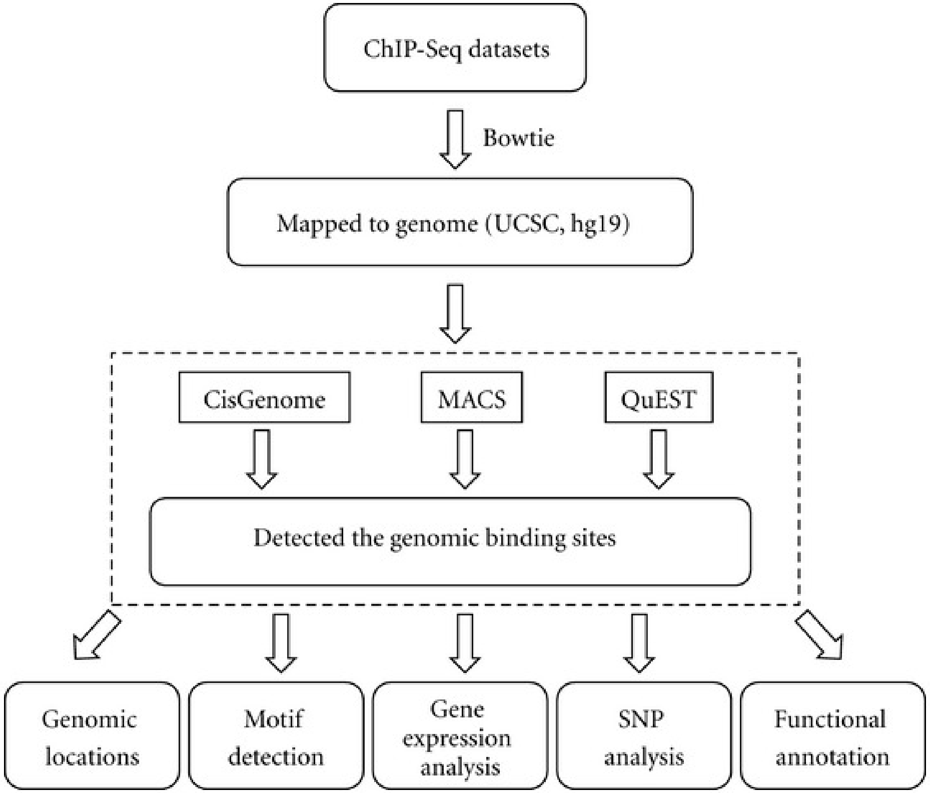
https://www.encodeproject.org/

Analysis of ChIP-seq data

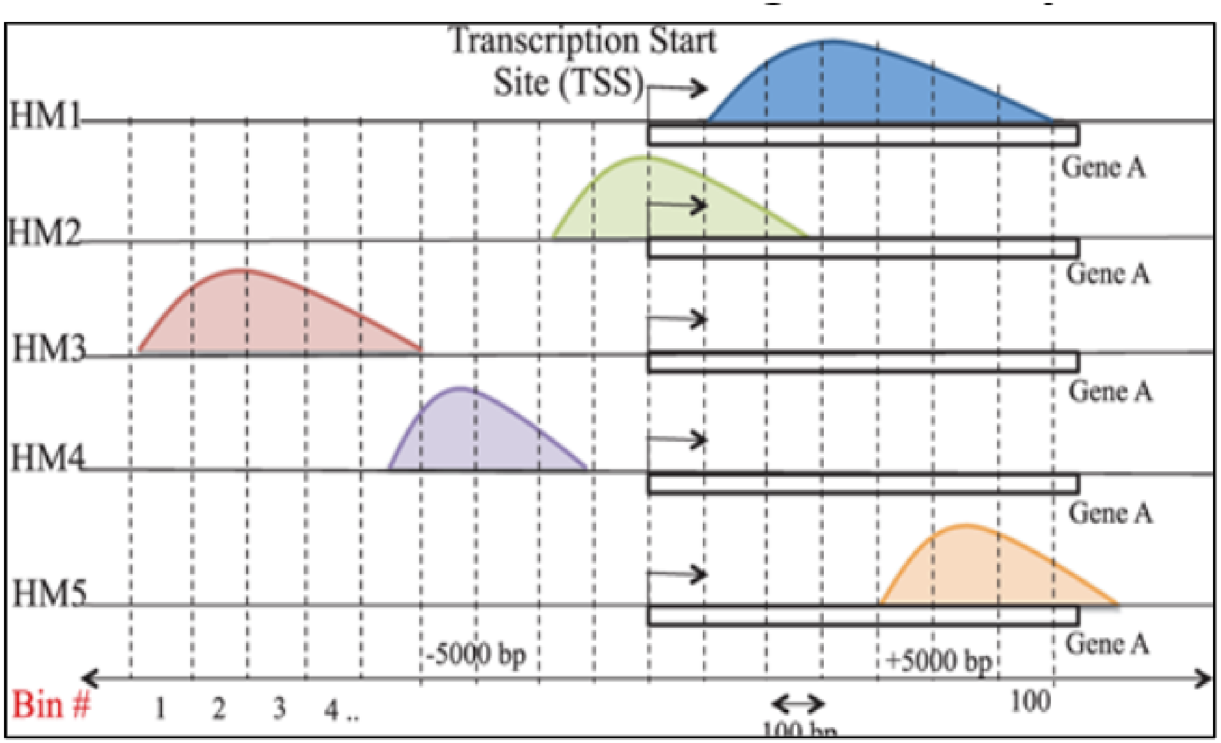


Analysis of ChIP-seq data





Combinatorial patterns of histone modifications



H3K36me3
Transcription

H3K4me3
Promoters

H3K9me3
Heterochromatin

H3K4me1
Enhancers

H3K27me3
Polycomb

Computational venues opened-up by ChIP-seq

- Prediction of gene expression from histone modifications
- Semi-supervised annotation of chromatin states (clustering of patterns)
- Motif discovery
- Prediction of enhancers and their target genes

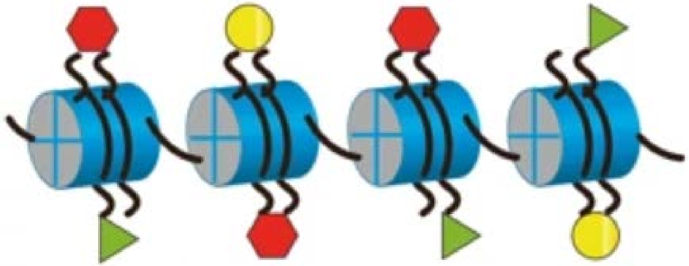
Exercise: Visualization of ChIP-seq data

1. Go to: http://epigenomegateway.wustl.edu/browser/
2. Select Human -> hg19 -> Go
3. Select Tracks -> Custom Tracks -> Add custom data hub
4. Choose datahub file -> Load “ImmuneCell-ChIPseq-PCHiC.json”
5. Wait a bit then Click red X on top-right
6. Navigate using zoom in/out and other controls
7. To jump to another region/gene click the gray coordinate (top left) and enter the name of your favorite gene
8. Select the top entry and see the H3K27ac pattern in cell for that gene
9. Some good examples are: *PAX5, LYZ, CD4, CD8A, YWHAZ*

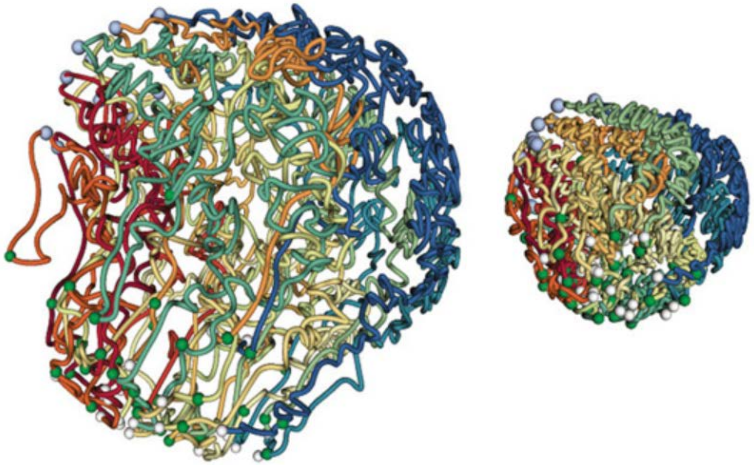
Part 1: DNA Methylation

Me
Me
C
G
A
T
C
A
C
G
T
C

Part 2: Nucleosome Positioning and Histone Modifications



Part 3: Three-dimensional Structure and Folding of the Genome



Finishing the Job:

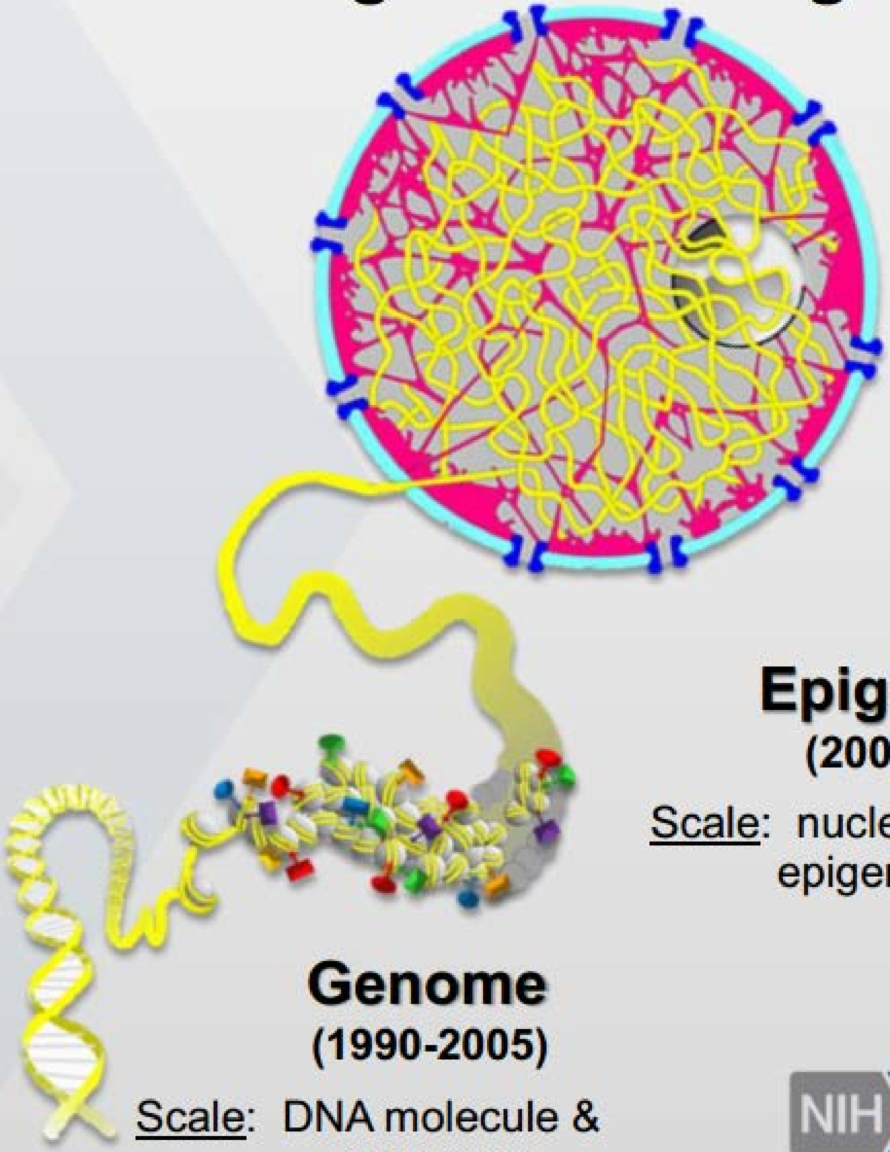
Understanding Genome Organization

3D Nucleome

ENCODE

ROADMAP epigenomics PROJECT

Human Genome Project



3D Nucleome
(2015-2022?)
Scale: cell nucleus & chromosome domains

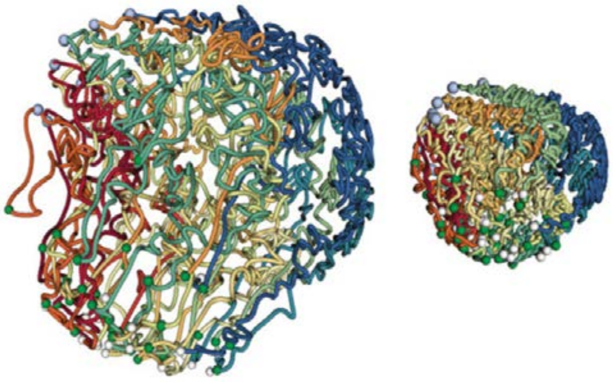
Epigenome
(2005-2015)
Scale: nucleosome & epigenetic marks

Genome
(1990-2005)
Scale: DNA molecule & sequence

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

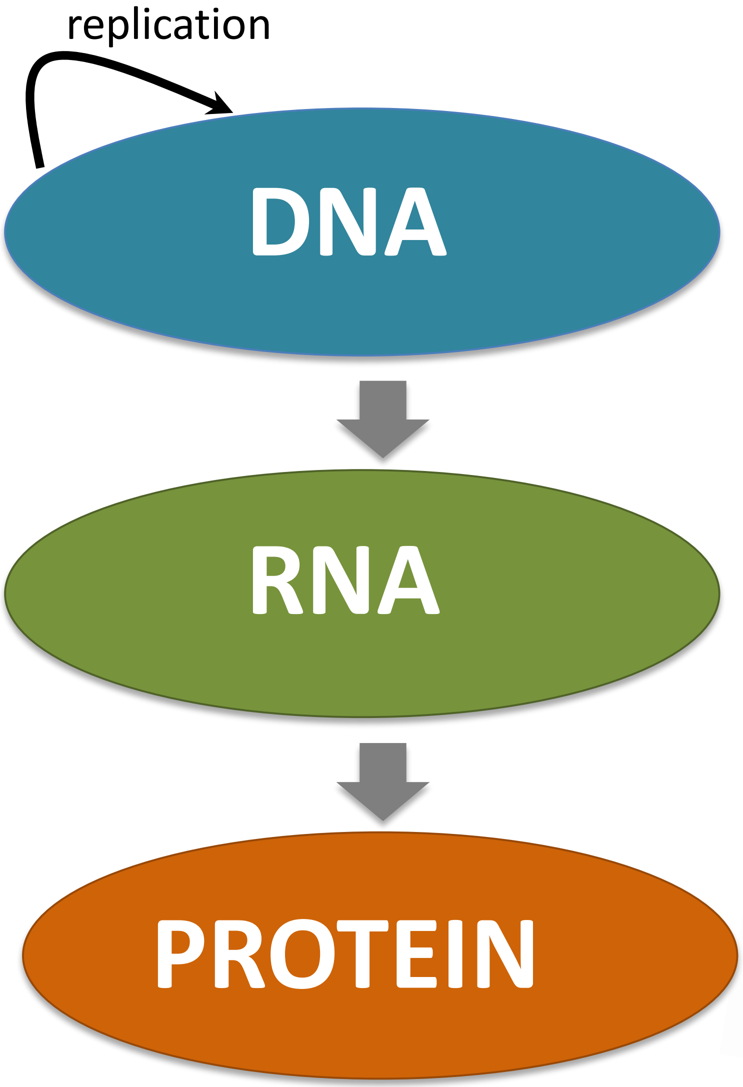
http://dpcpsi.nih.gov/sites/default/files/CoC_3D_Nucleome_28_final_29.pdf

Part 3: Three-dimensional Structure and Folding of the Genome



- Why ALL/MOST of the genome matters?
- Distal gene regulation
- Introduction to conformation capture methods
- Uses of Hi-C and similar experiments
- Examples from Ay lab research interest in 3D genome
- Exercise: Visualize Hi-C data

Central Dogma ("The BIG Idea") of Biology



replication

DNA

RNA

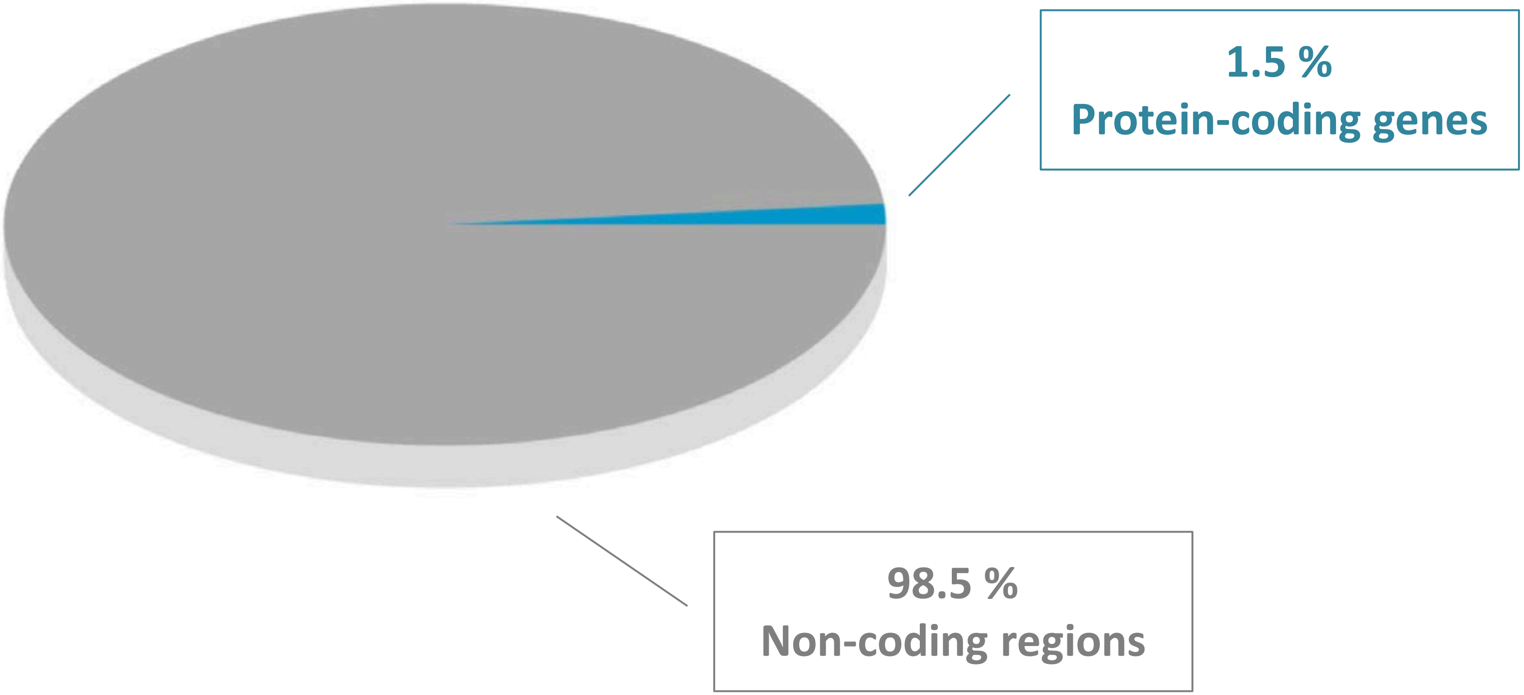
PROTEIN

DNA stores information to run cell

RNA's function is to make proteins

Proteins actually do the work inside the cell

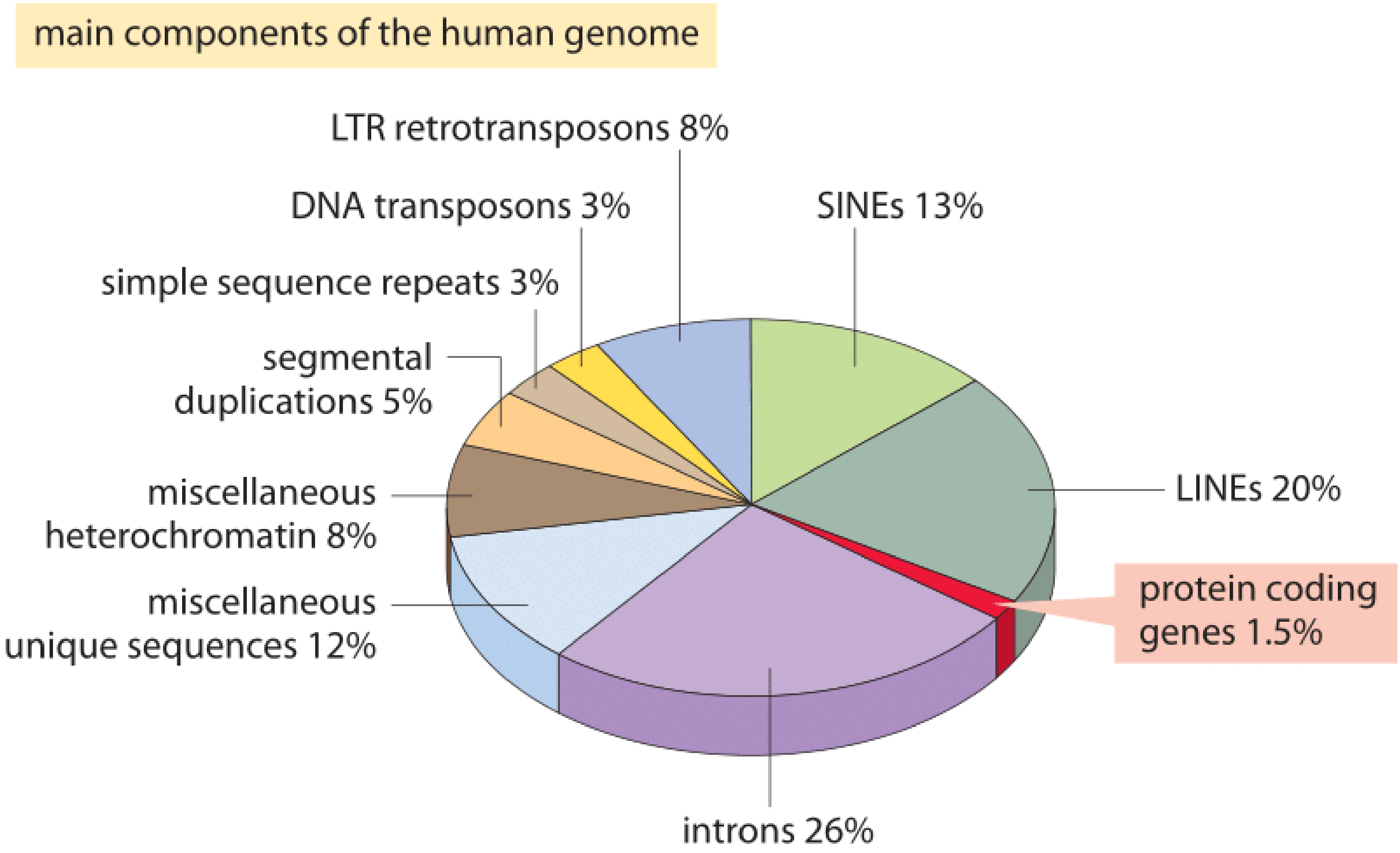
Only a small fraction of our genome encodes genes



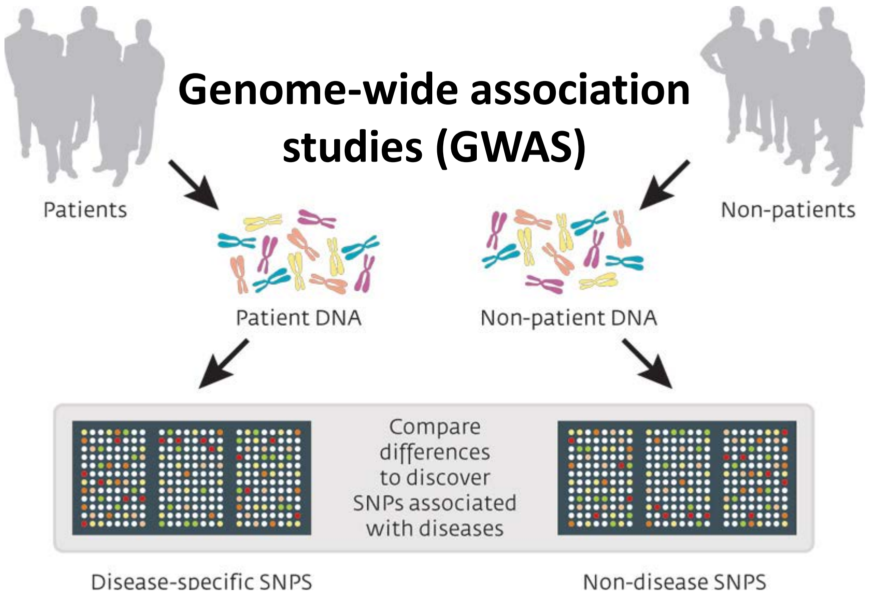
1.5 %
Protein-coding genes

98.5 %
Non-coding regions

Only a small fraction of our genome encodes genes



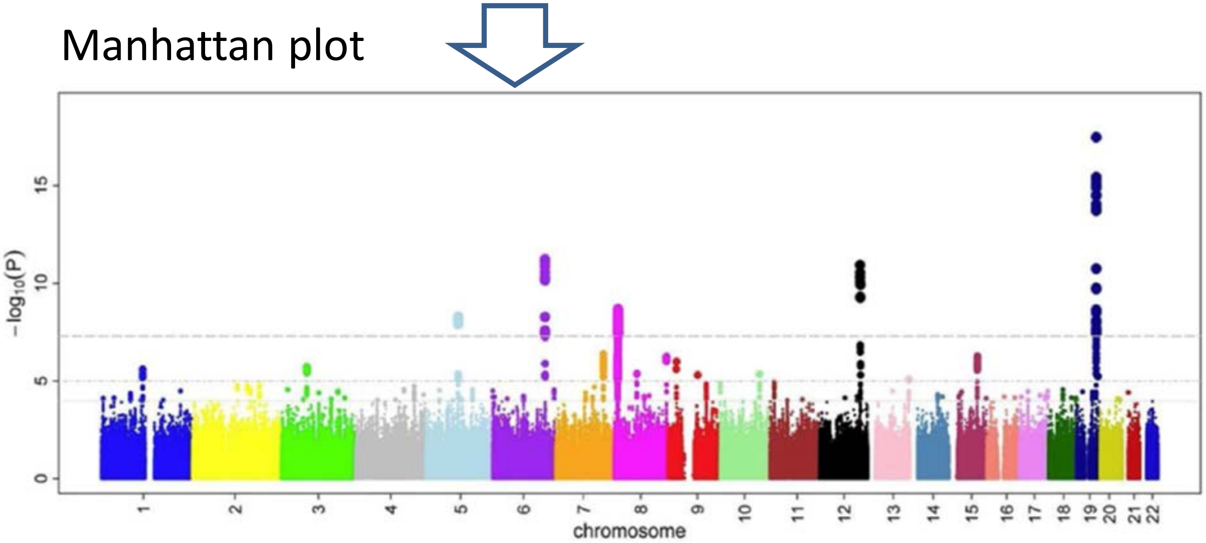
Variation in the noncoding genome plays a huge role in disease association



Genome-wide association studies (GWAS)

Compare differences to discover SNPs associated with diseases

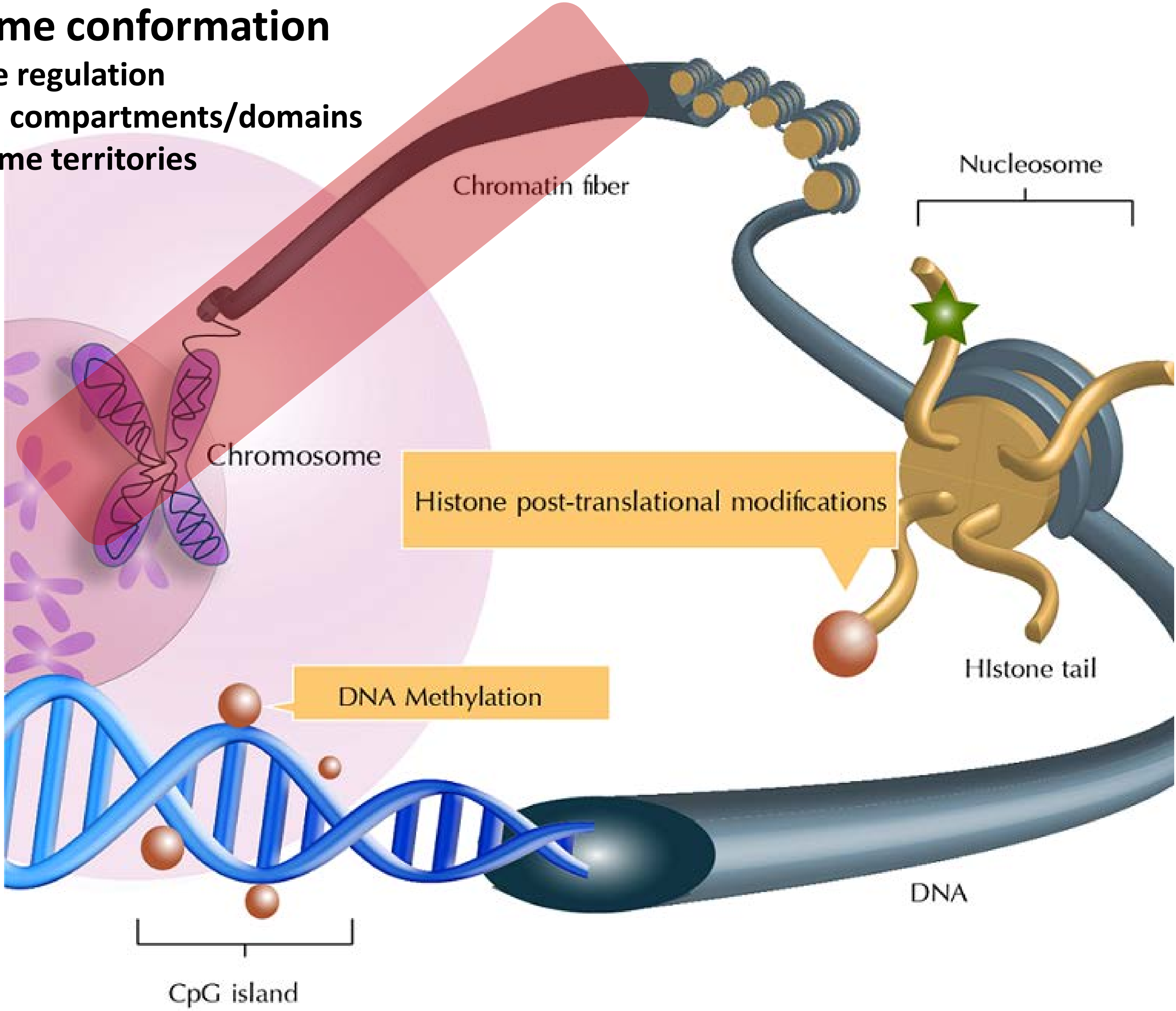
Manhattan plot



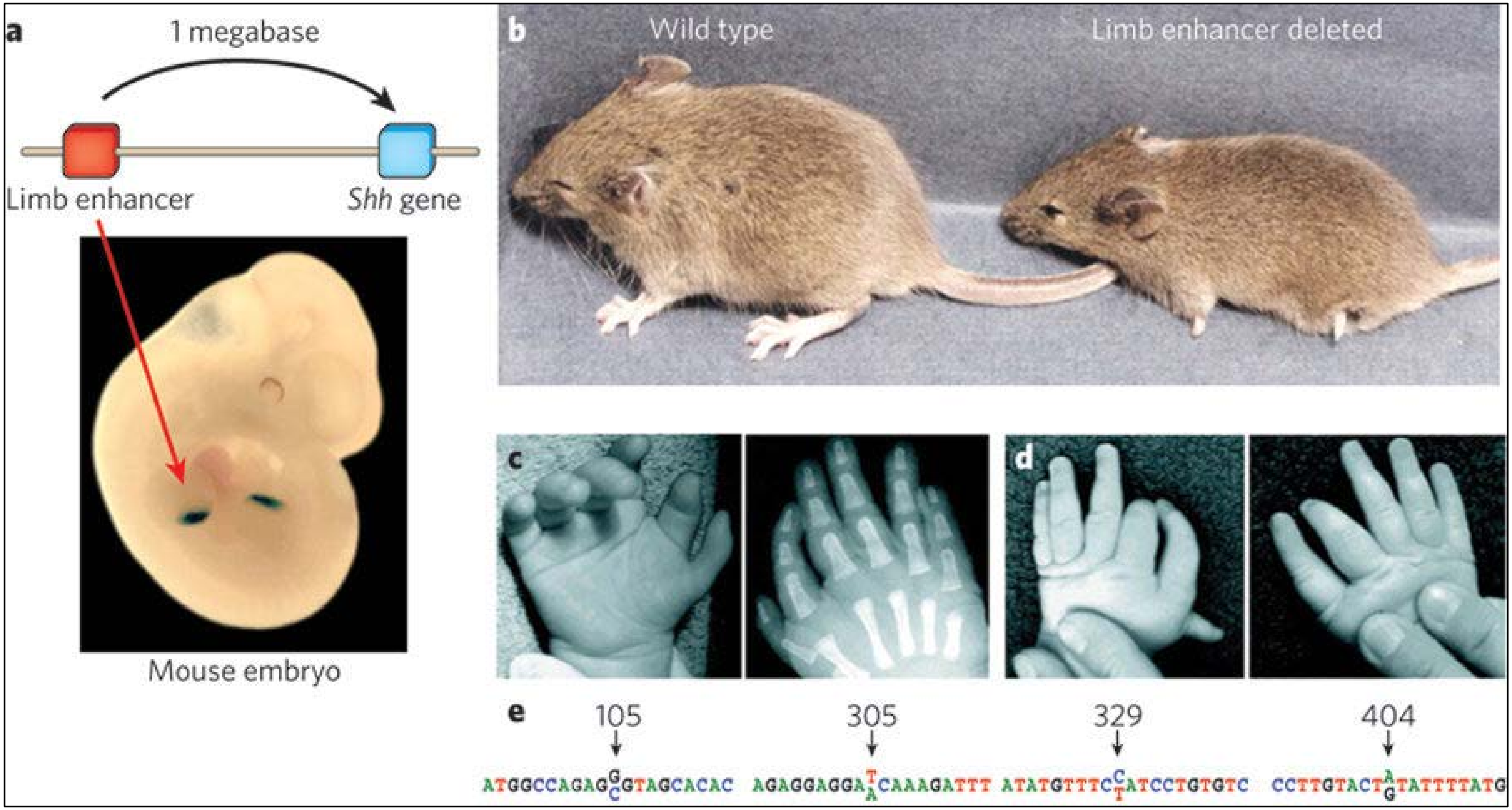
More than 90% of disease-associated genetic variants reside in noncoding regions with unknown gene targets.

Chromosome conformation

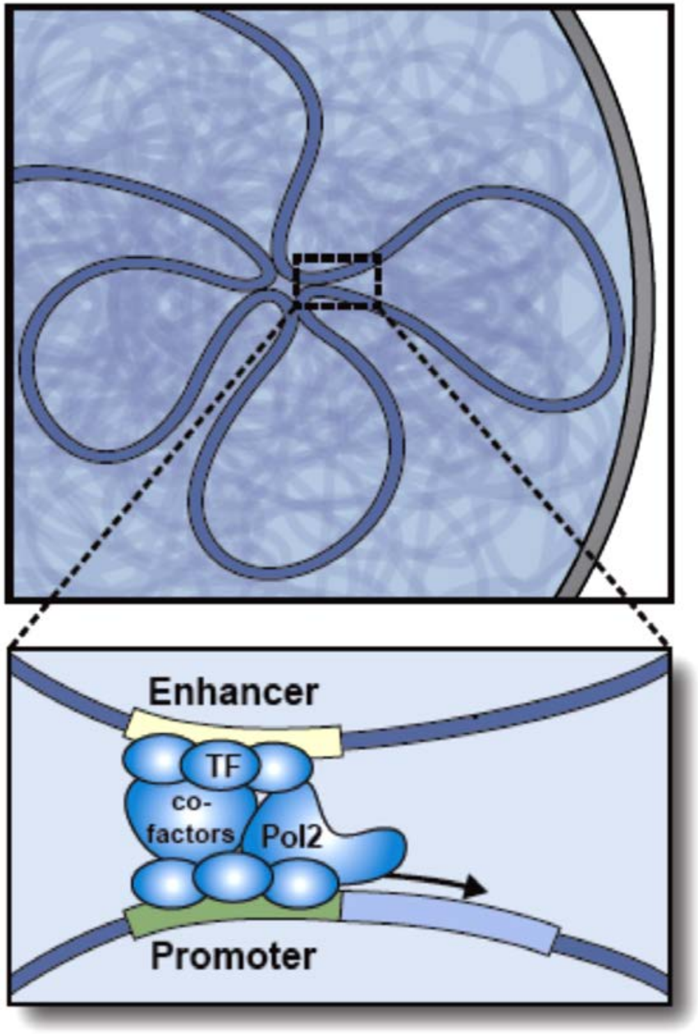
- Distal gene regulation
- Chromatin compartments/domains
- Chromosome territories

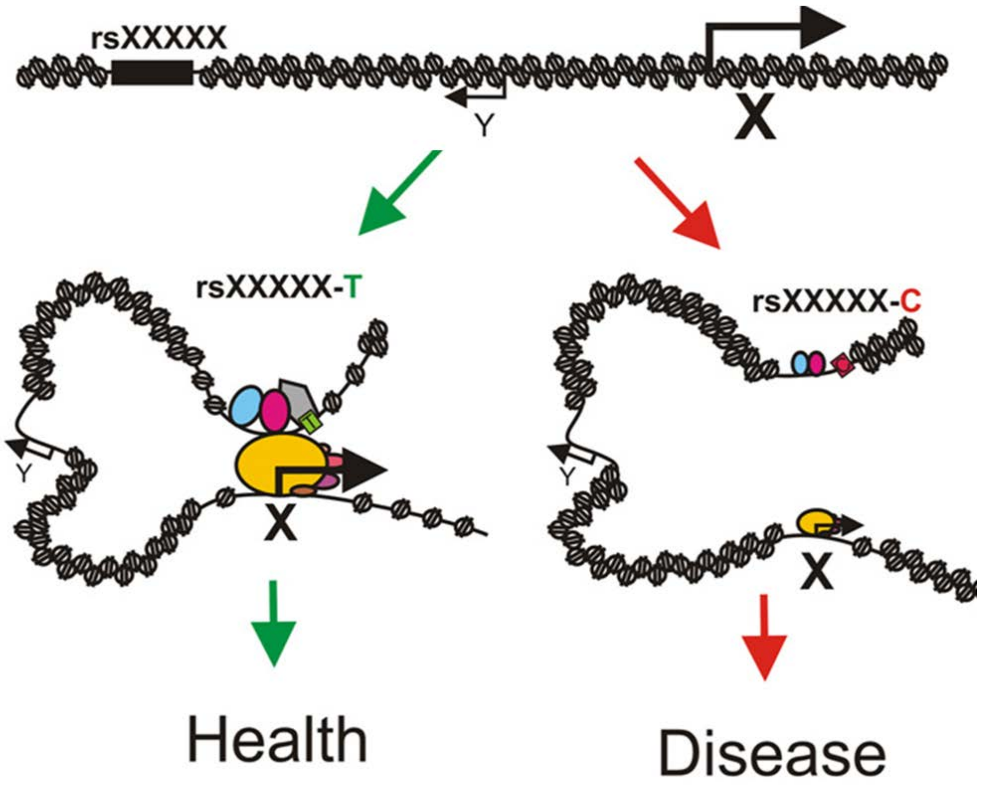


Genetic changes in enhancer regions may regulate distal genes

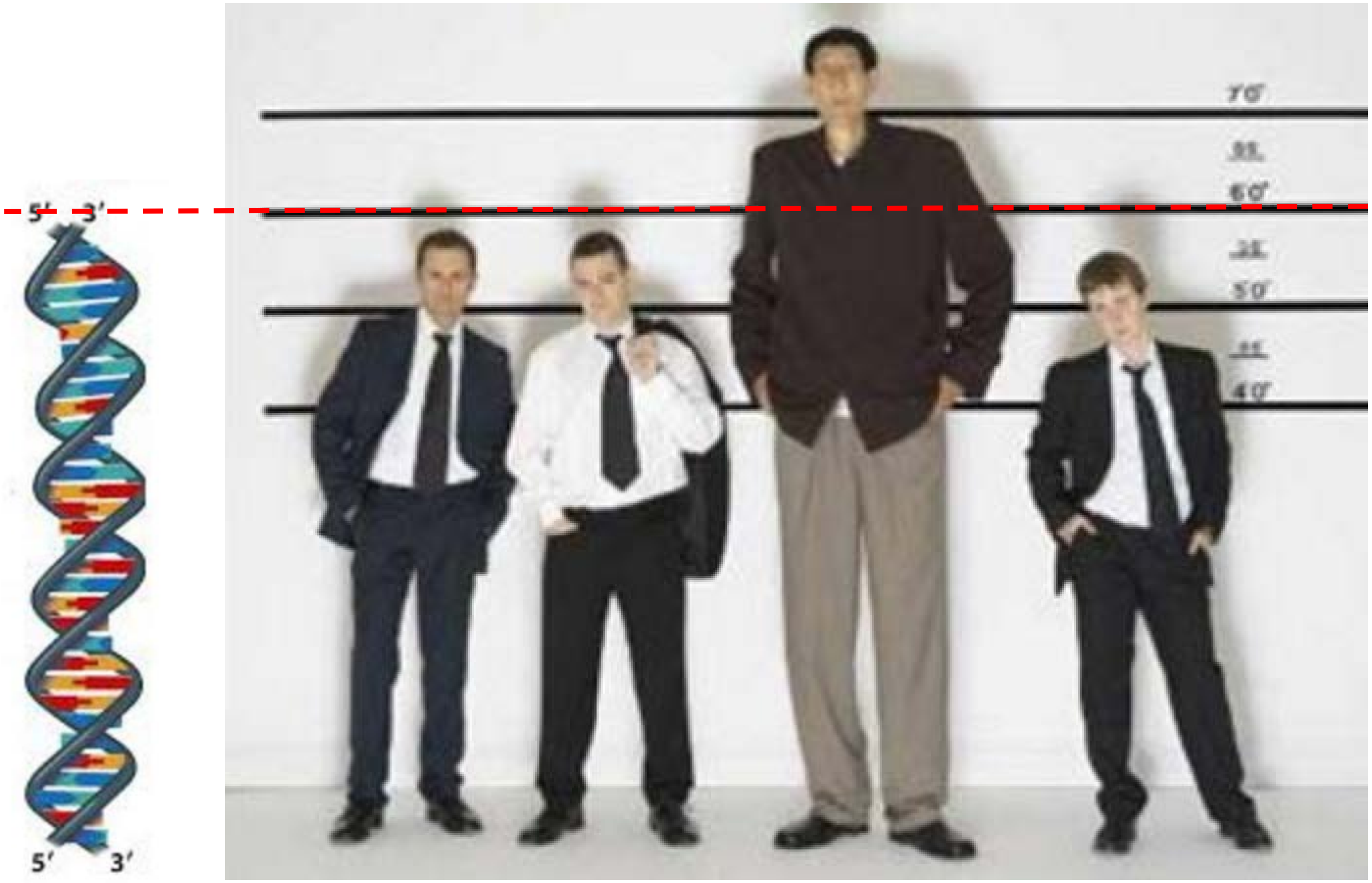


Genetic changes in enhancer regions may regulate distal genes

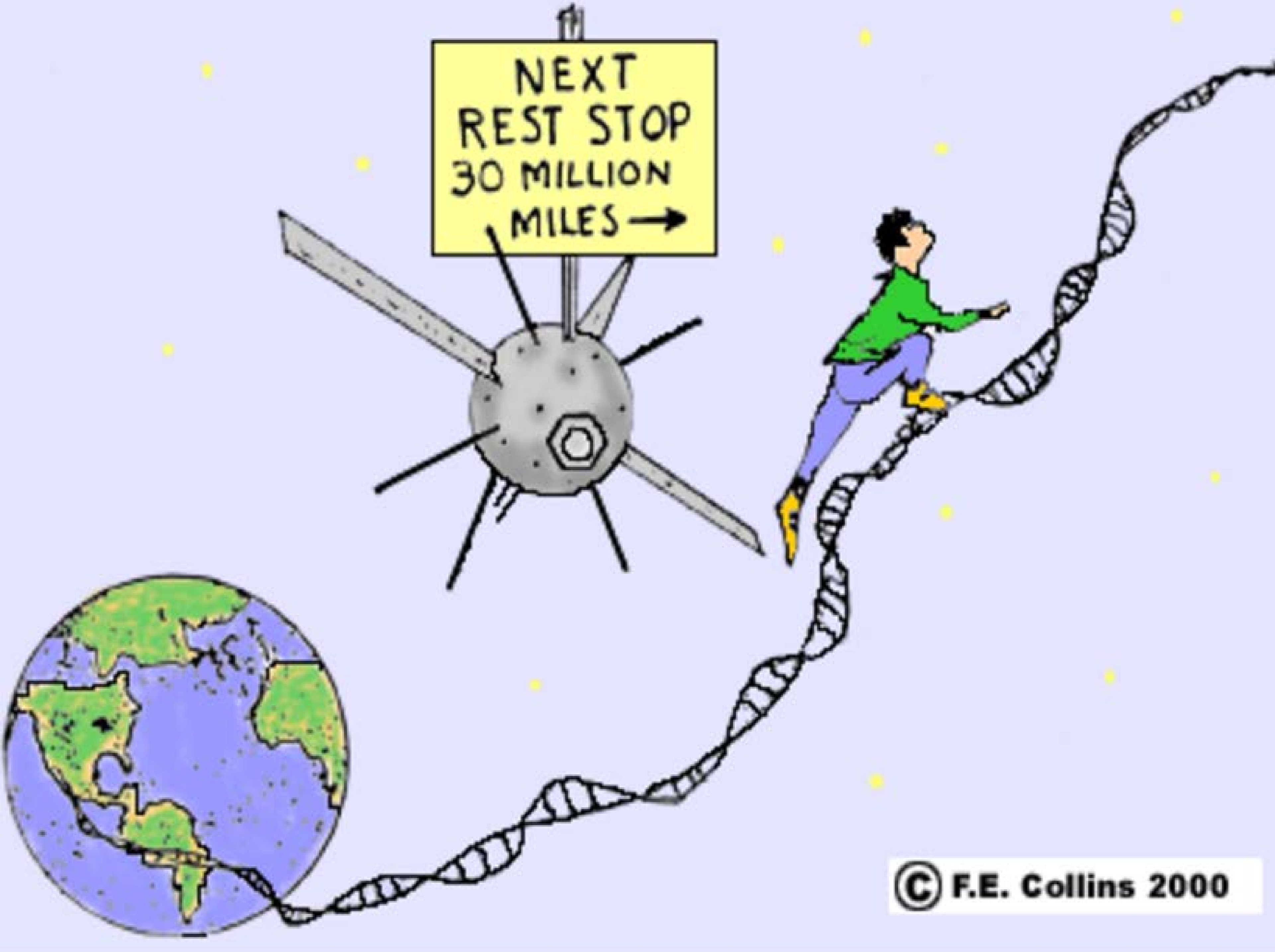




The DNA from a single one of our cells is taller than ...

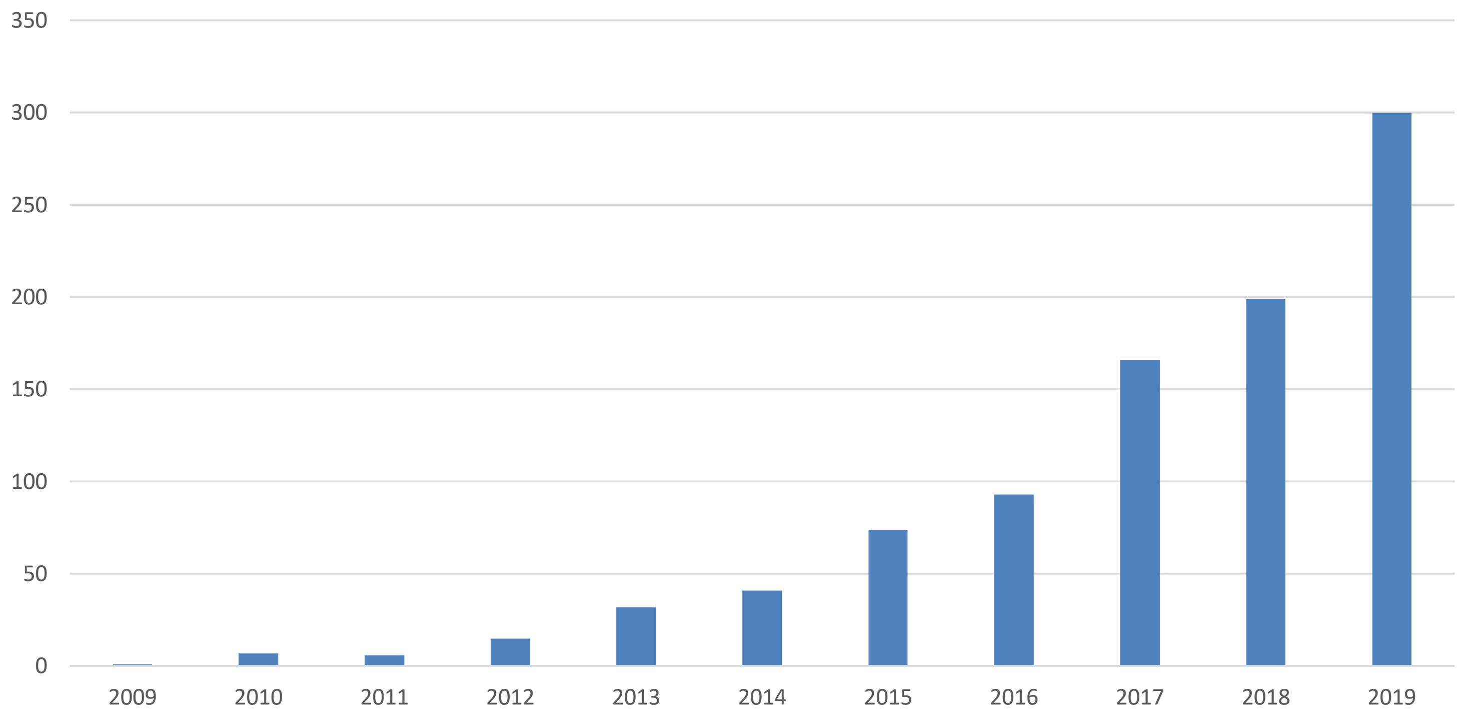


most of us


NEXT
REST STOP
30 MILLION
MILES →

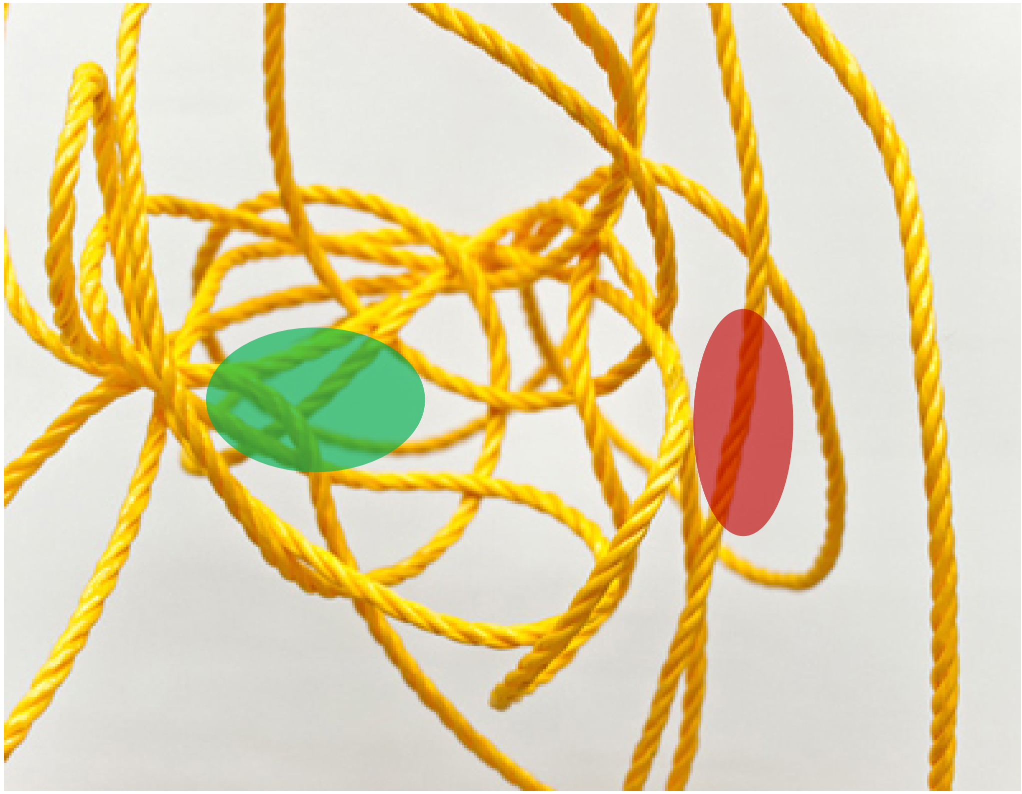
Another good motivation

Number of publications per year involving keyword **“Hi-C”**



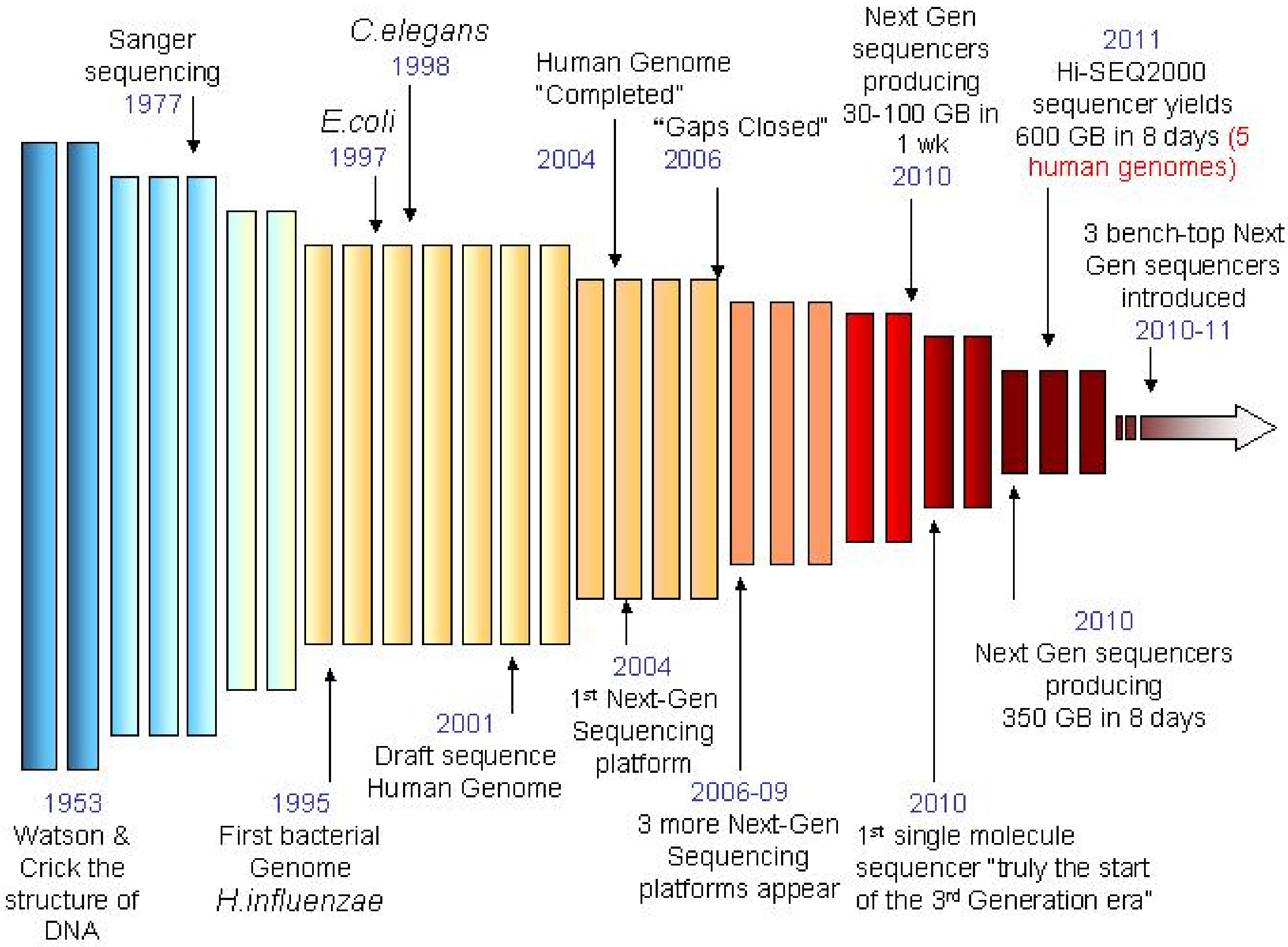
Source: Pubmed

That's all great but…
How can we measure and model how DNA folds?



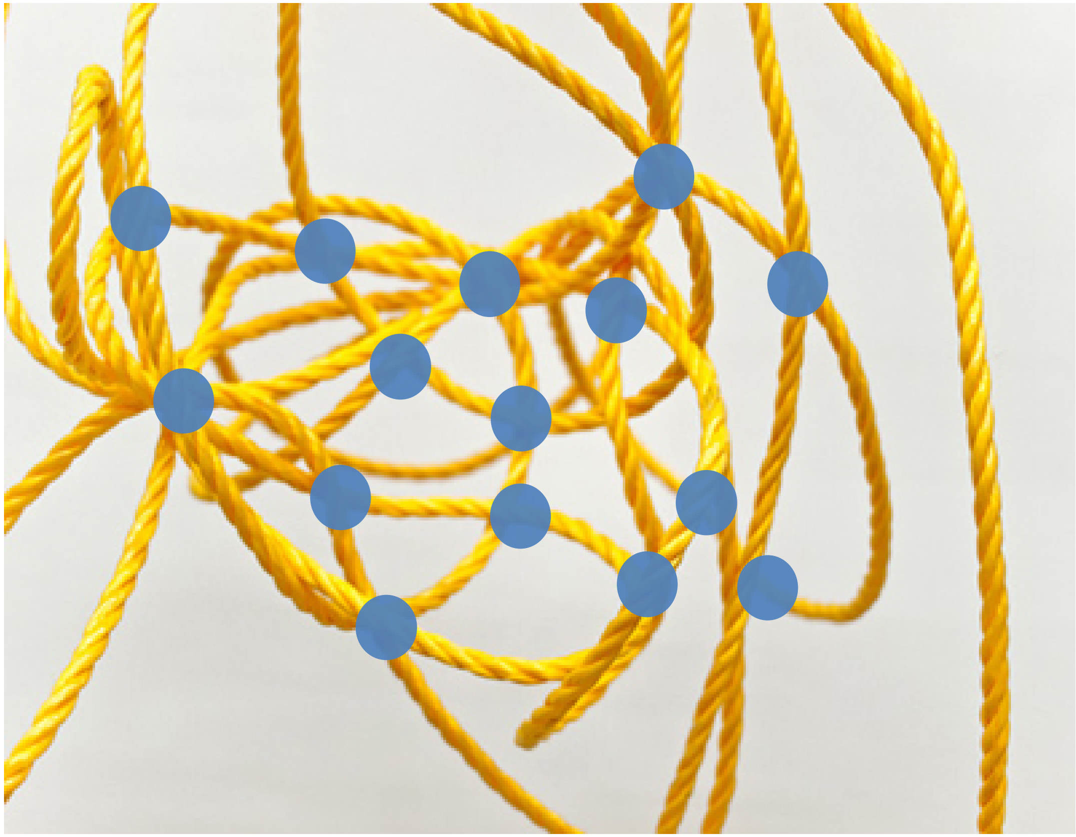
- Has been the only way up until last decade
- Low resolution: only large chunks of DNA can be visualized/colored
- Low throughput: only a few points can be visualized at once
- Not feasible to generate 3D models from it but good for validation once you have them

The revolution of next generation sequencing

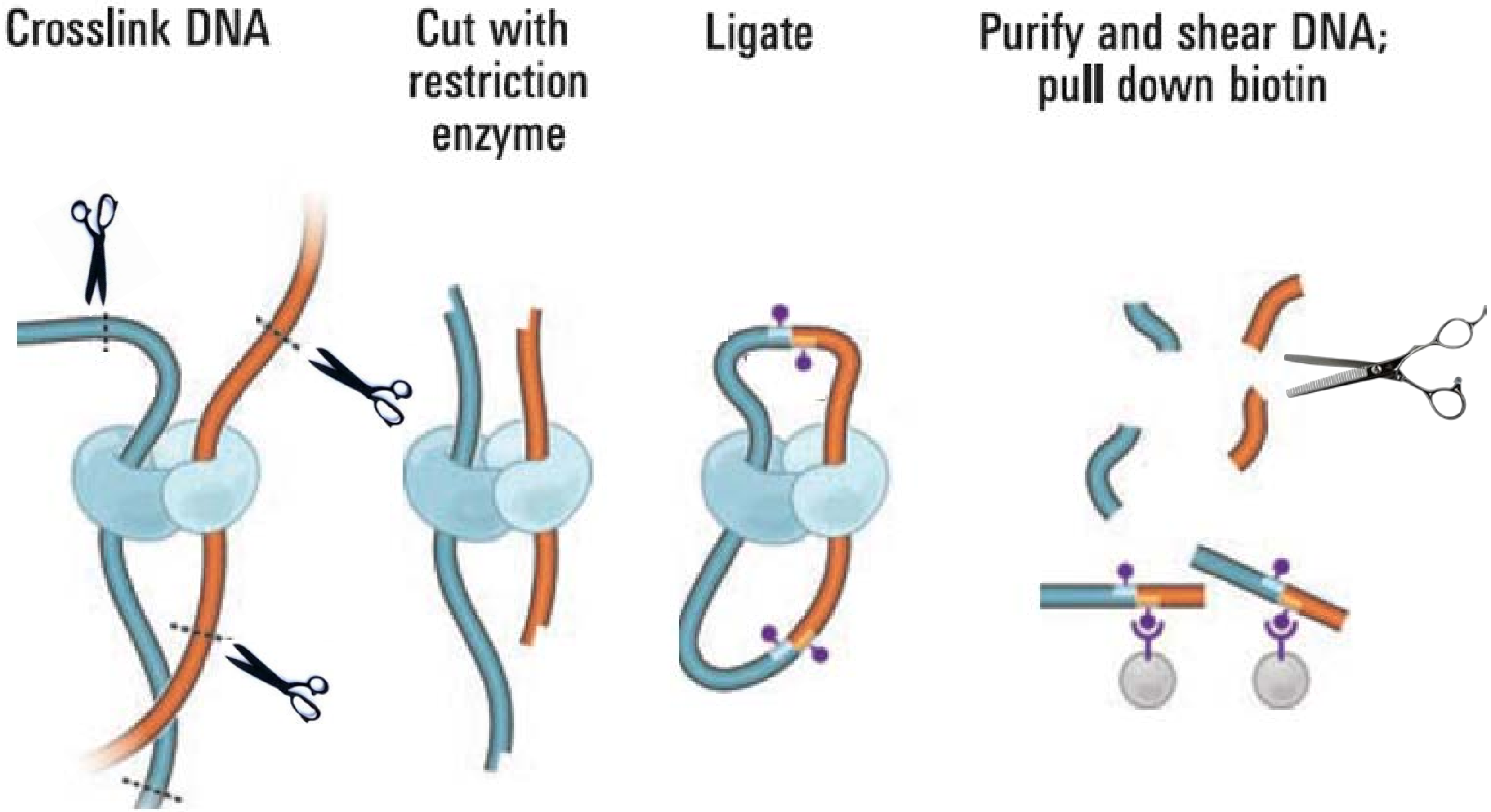

Sanger sequencing 1977
C.elegans 1998
E.coli 1997
Human Genome "Completed" 2004
"Gaps Closed" 2006
Next Gen sequencers producing 30-100 GB in 1 wk 2010
2011 Hi-SEQ2000 sequencer yields 600 GB in 8 days (5 human genomes)
3 bench-top Next Gen sequencers introduced 2010-11
1953 Watson & Crick the structure of DNA
1995 First bacterial Genome *H.influenzae*
2001 Draft sequence Human Genome
2004 1st Next-Gen Sequencing platform
2006-09 3 more Next-Gen Sequencing platforms appear
2010 1st single molecule sequencer "truly the start of the 3rd Generation era"
2010 Next Gen sequencers producing 350 GB in 8 days

http://www.ipc.nxgenomics.org/newsletter/no11.htm

Next generation sequencing-based assays to measure 3D structure genome-wide

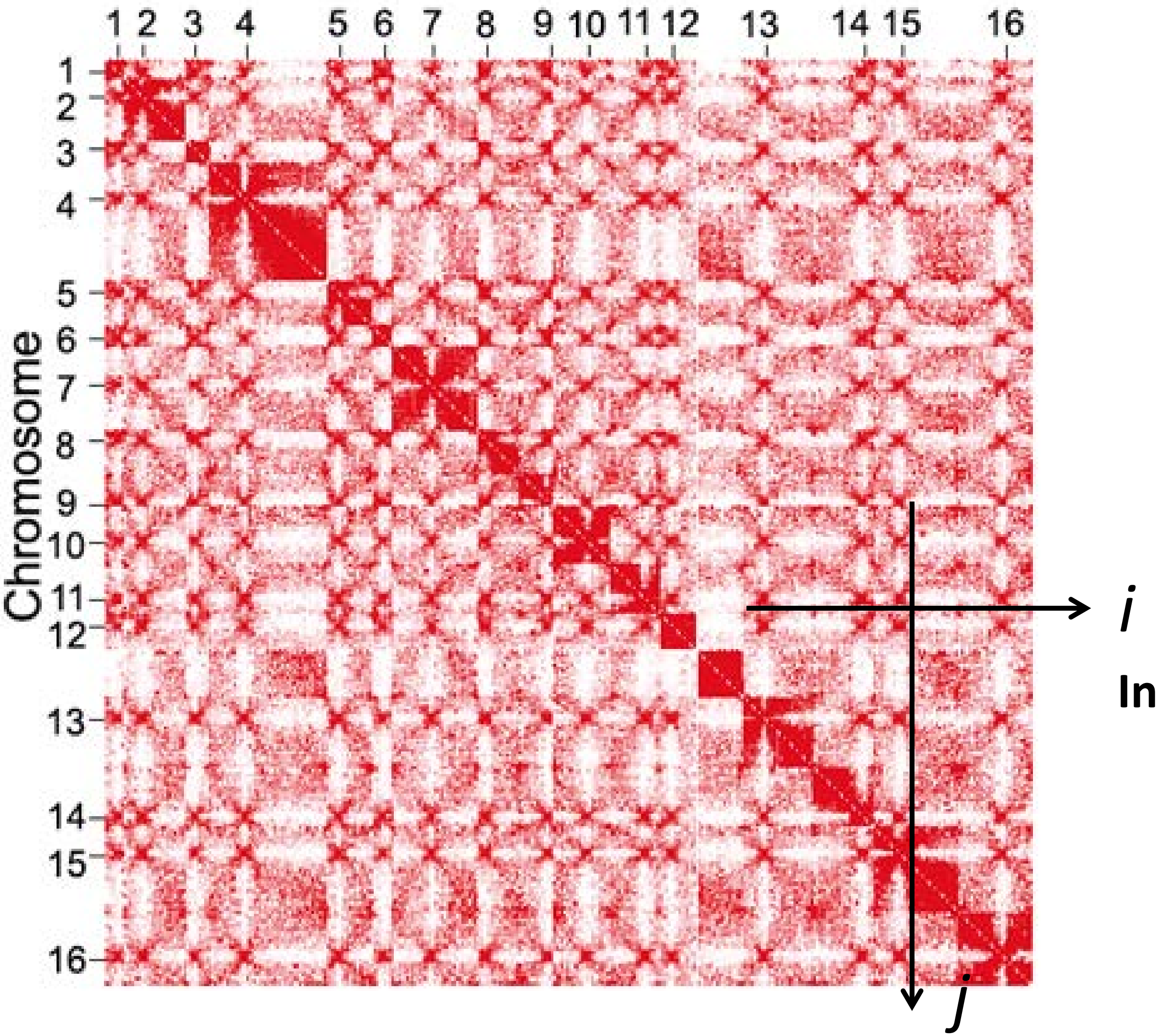


The revolution of next generation sequencing technology in measuring the 3D structure

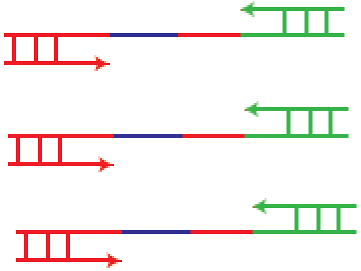

Crosslink DNA
Cut with restriction enzyme
Ligate
Purify and shear DNA; pull down biotin

Hi-C: L.-Aiden et al. *Science* 2009

The readout from Hi-C is a contact matrix



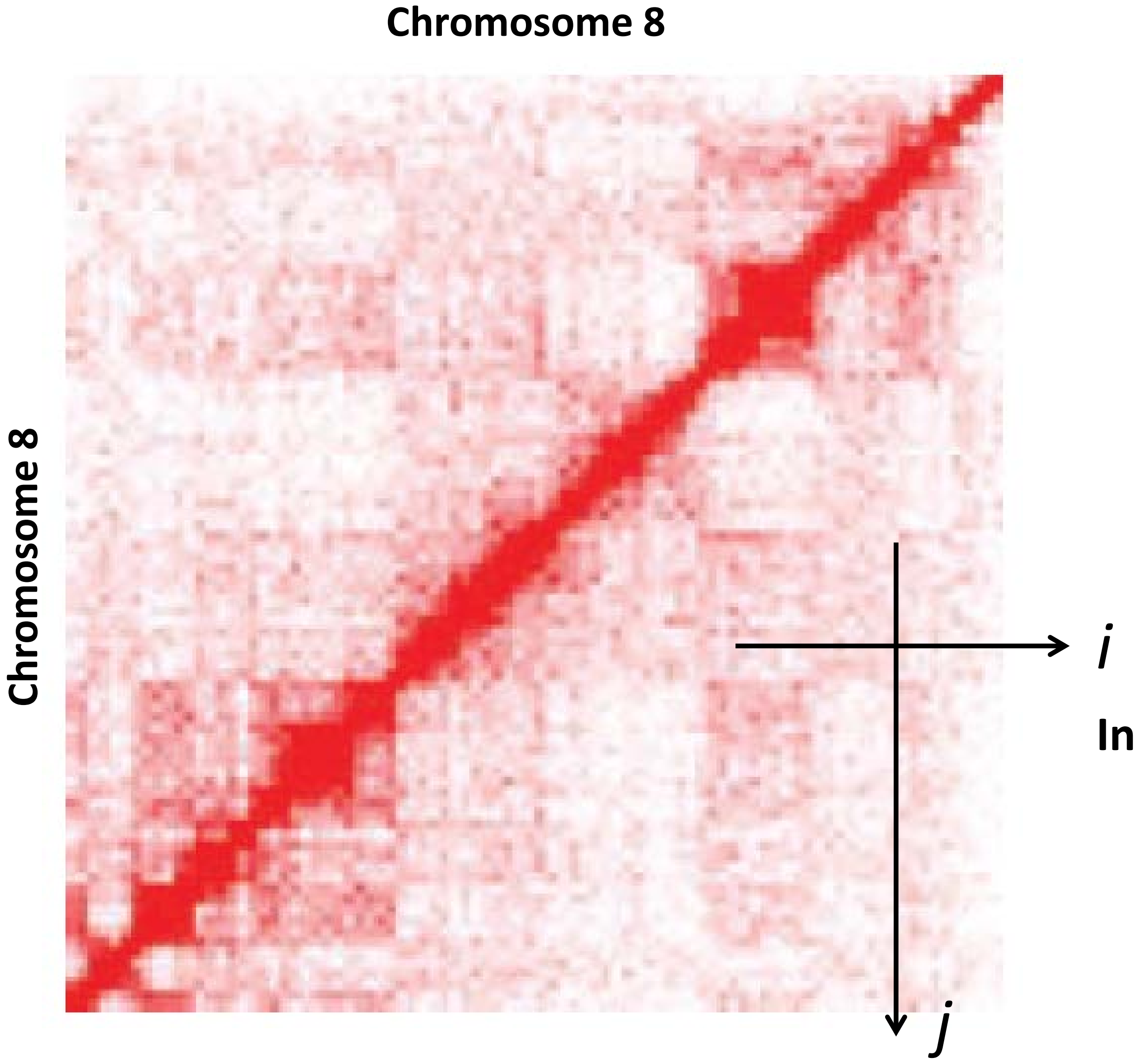
paired-end reads



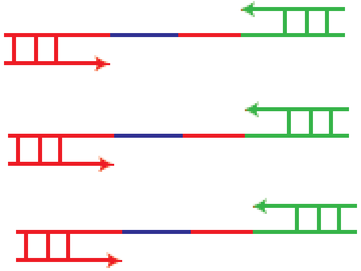
C(i,j) = How many times locus ***i*** is linked to locus ***j*** by a paired-end read?

Inter-chromosomal contact

The readout from Hi-C is a contact matrix



paired-end reads

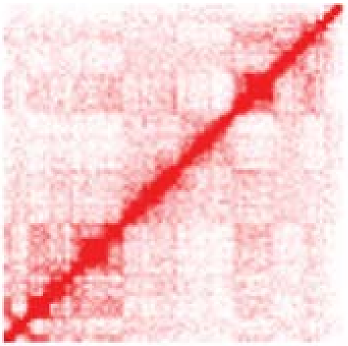


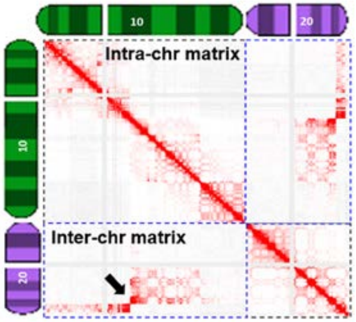
C(i,j) = How many times locus ***i*** is linked to locus ***j*** by a paired-end read?

Intra-chromosomal contact

What can we see with Hi-C?

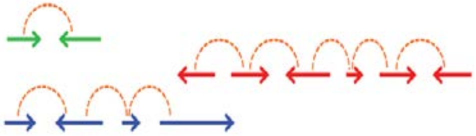
Hi-C contact map





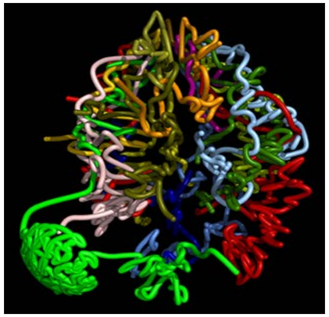
Identifying genomic rearrangements

Chakraborty & Ay. Bioinformatics, 2017.
Dixon *et al.* Nature Genetics, 2018.



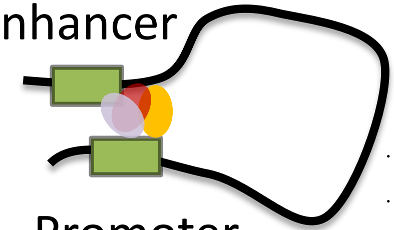
Genome assembly and phasing

Nature Biotech, Dec 2013.



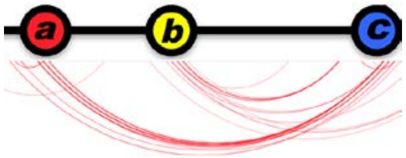
3D modeling of genomes

- Duan *et al*. Nature, 2010 *(S. cerevisae)*,
- Ay *et al.* Genome Res., 2014a *(P. fal)*,
- Varoquaux, Ay, *et al*. ISMB, 2014.



Long-range chromatin contacts

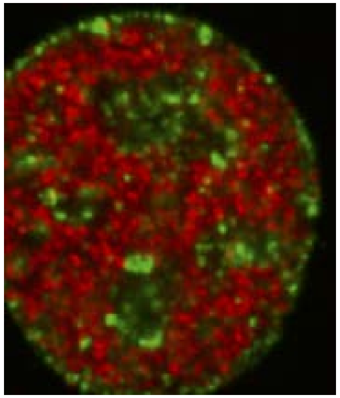
- Ay *et al*. Genome Res., 2014b
- Ma, Ay, *et al*. Nature Methods, 2015.

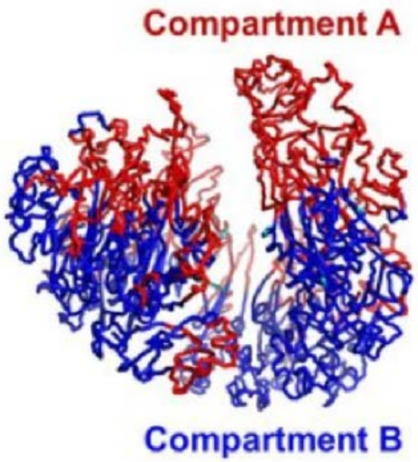


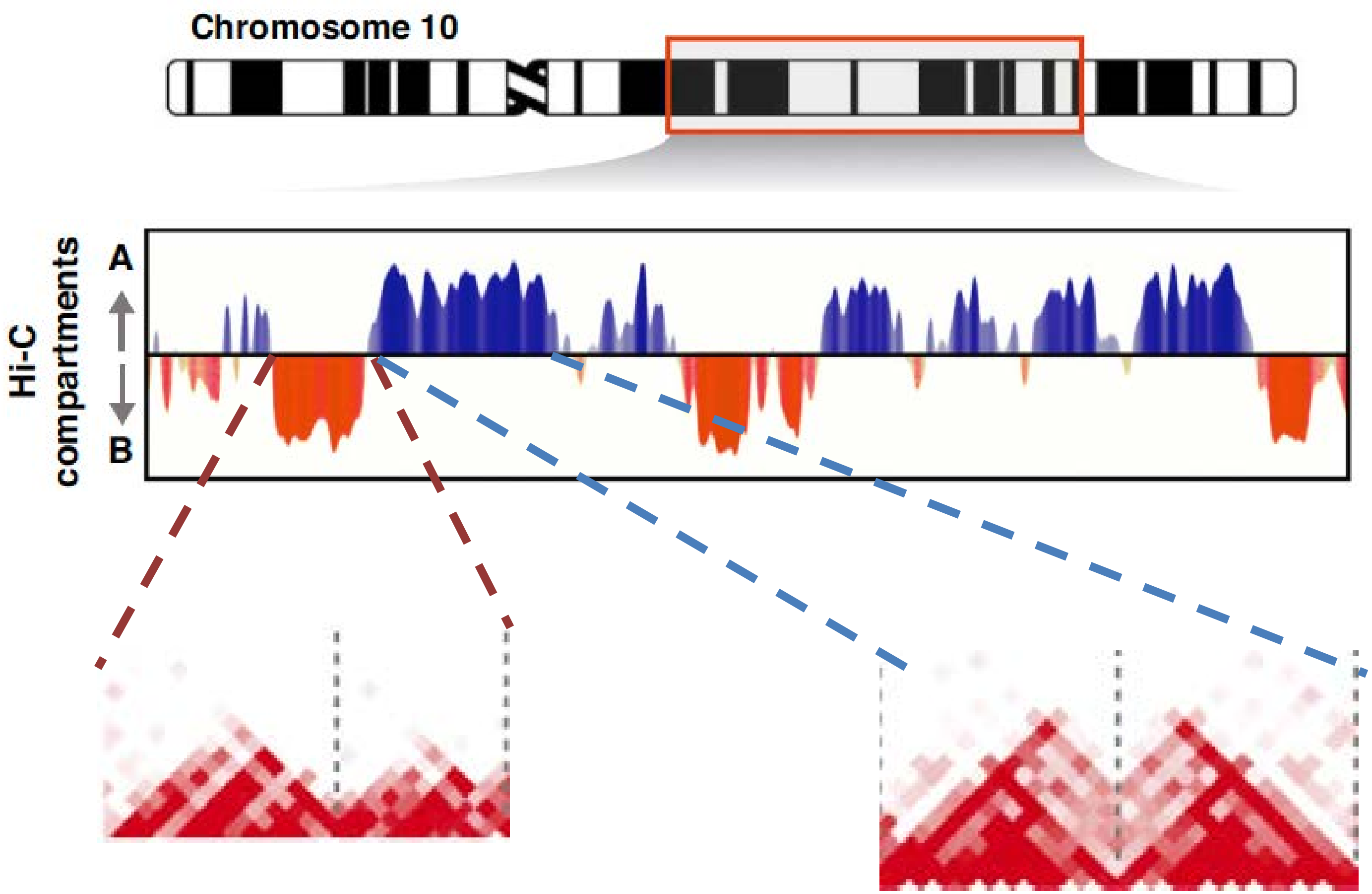
Discovery of non-linear effects on function

Sima, Chakraborty *et al*. *Cell*, 2019.

What can we see with Hi-C?







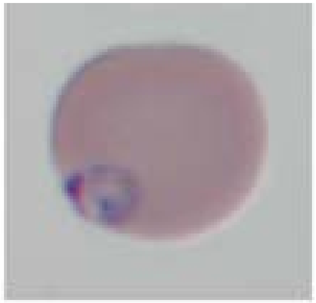
TADs

Rivera-Mulia & Gilbert, 2016

Importance of 3D genome organization: examples from our own work

Malaria

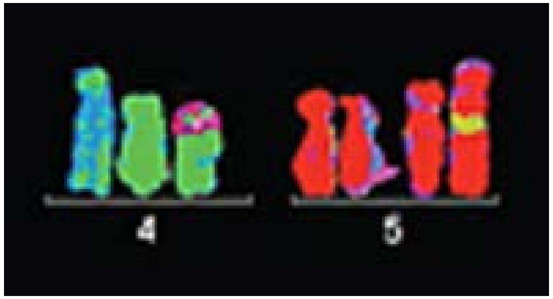
Vector



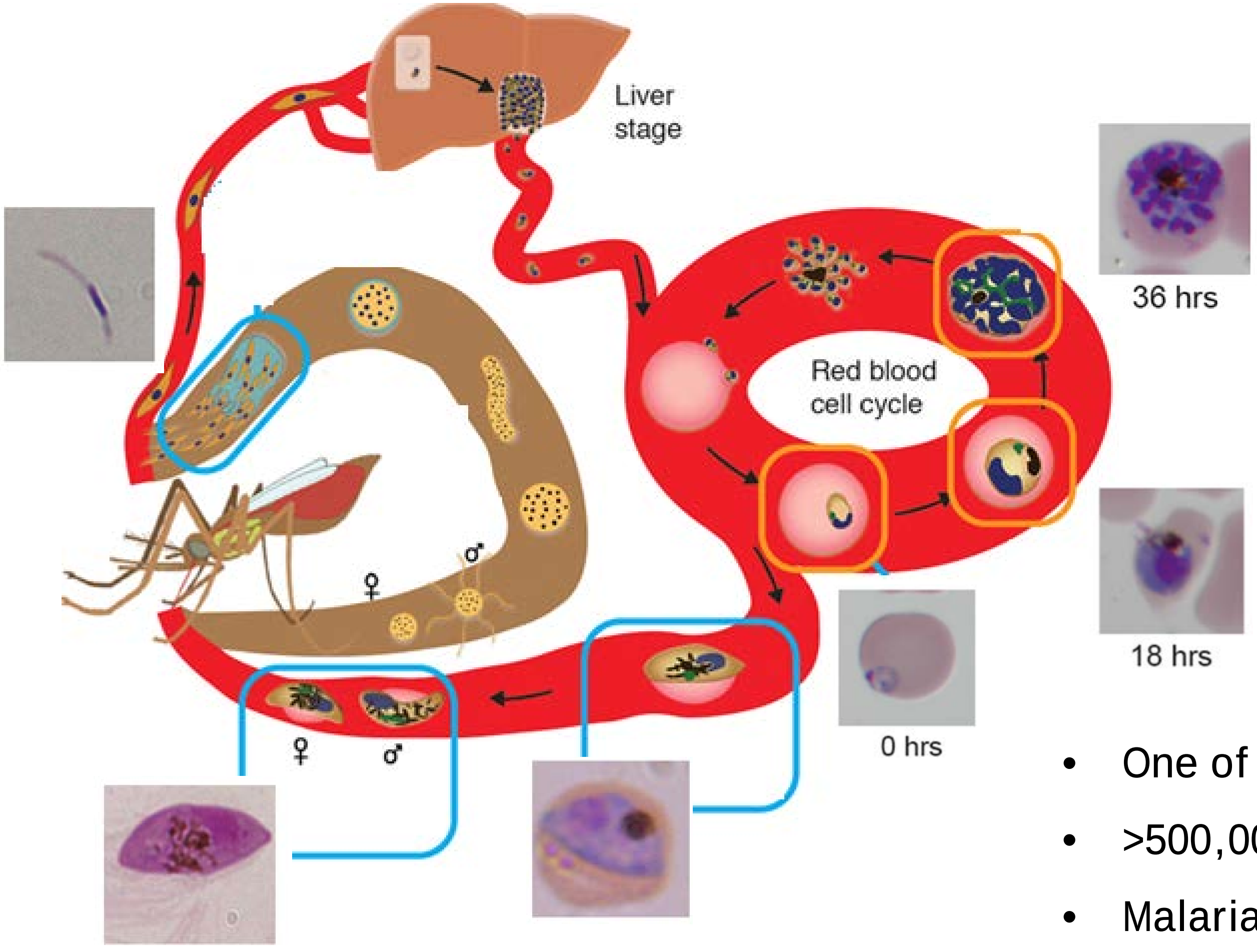
Plasmodium falciparum

Asthma

Cancer



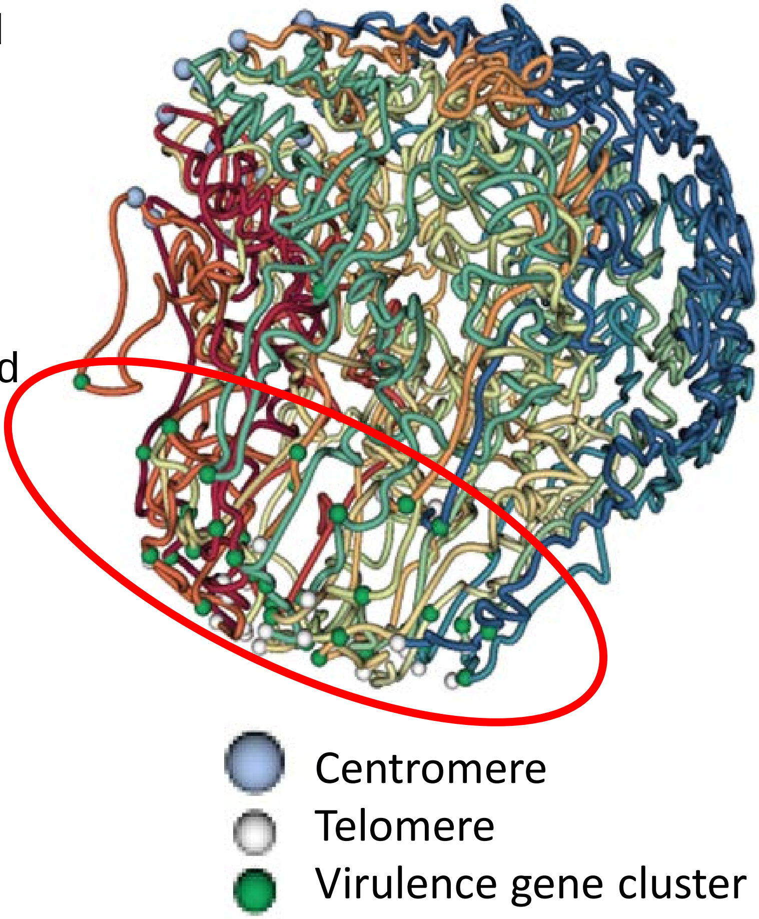
P. falciparum: The deadliest human malarial parasite

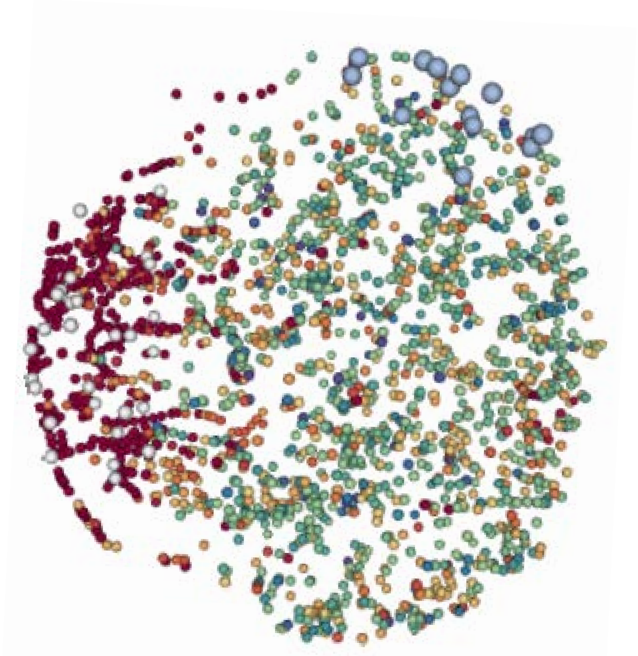


- One of the deadliest infectious diseases
- >500,000 deaths per year
- Malarial death → *P. falciparum*
- No effective vaccine
- Spreading resistance to drugs

Repression of virulence genes by 3D clustering

- Virulence genes encode proteins that are inserted into the infected red blood cell surface
- *P. falciparum* encodes ~60 virulence genes
- Exactly one virulence gene is expressed per cell
- This antigenic variation allows immune evasion and avoidance of antibody-mediated clearance

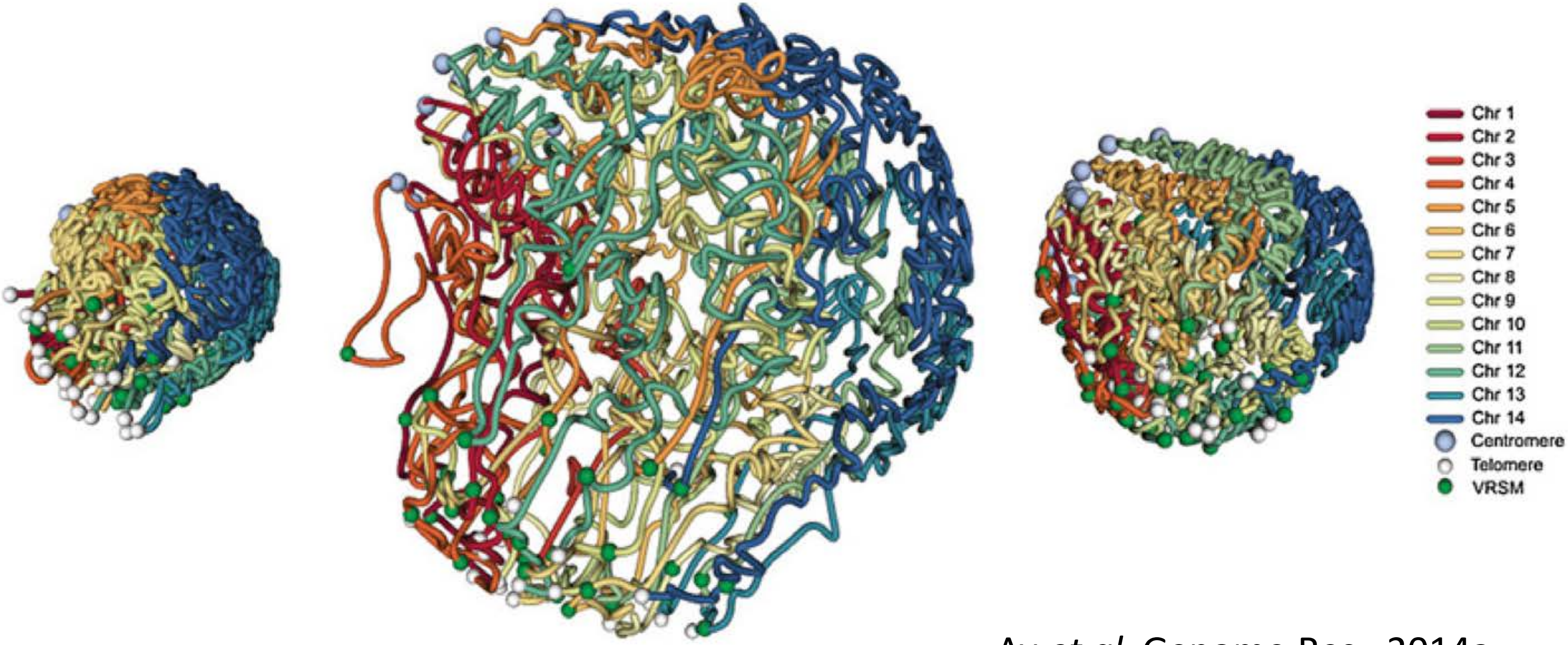




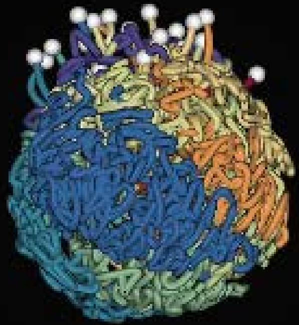
Gene expression

Ay et al. *Genome Research* 2014a

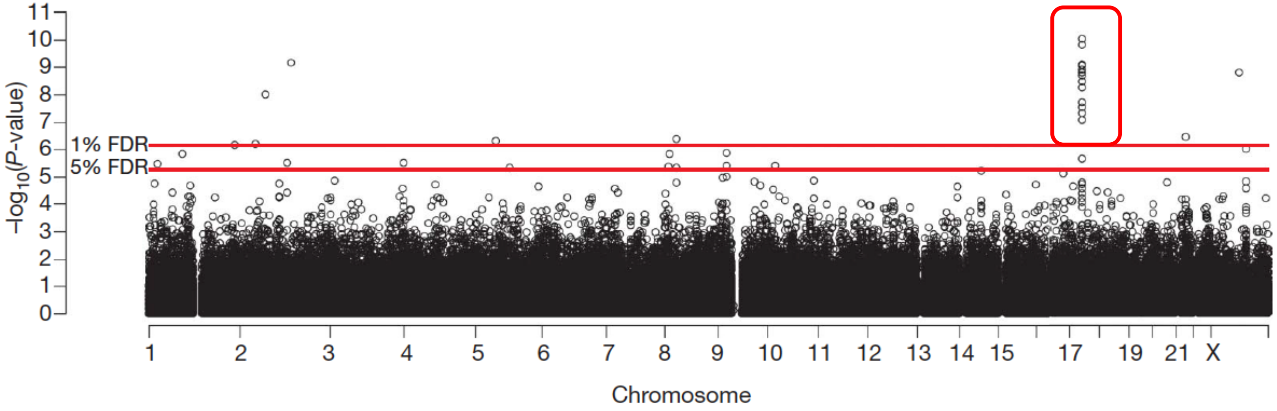
3D genome structure of the deadliest malaria parasite (*P. falciparum*)



- Ay *et al.* Genome Res., 2014a



Asthma-risk locus on chromosome 17 identified by genome-wide association studies (GWAS)

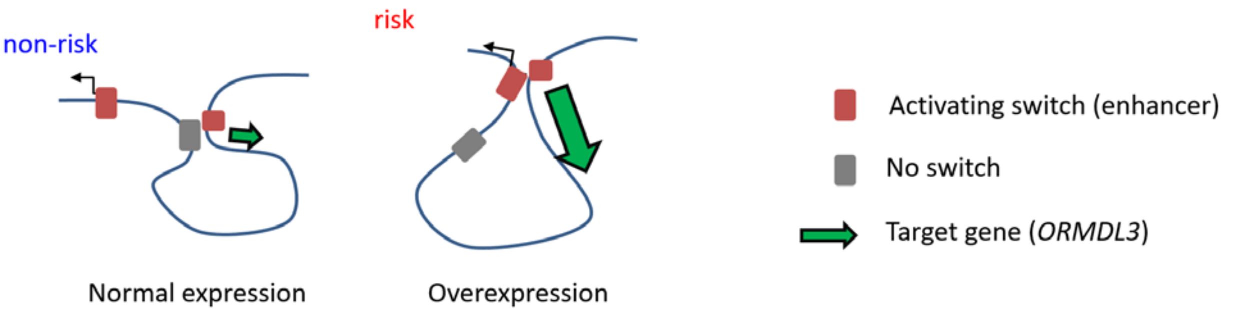


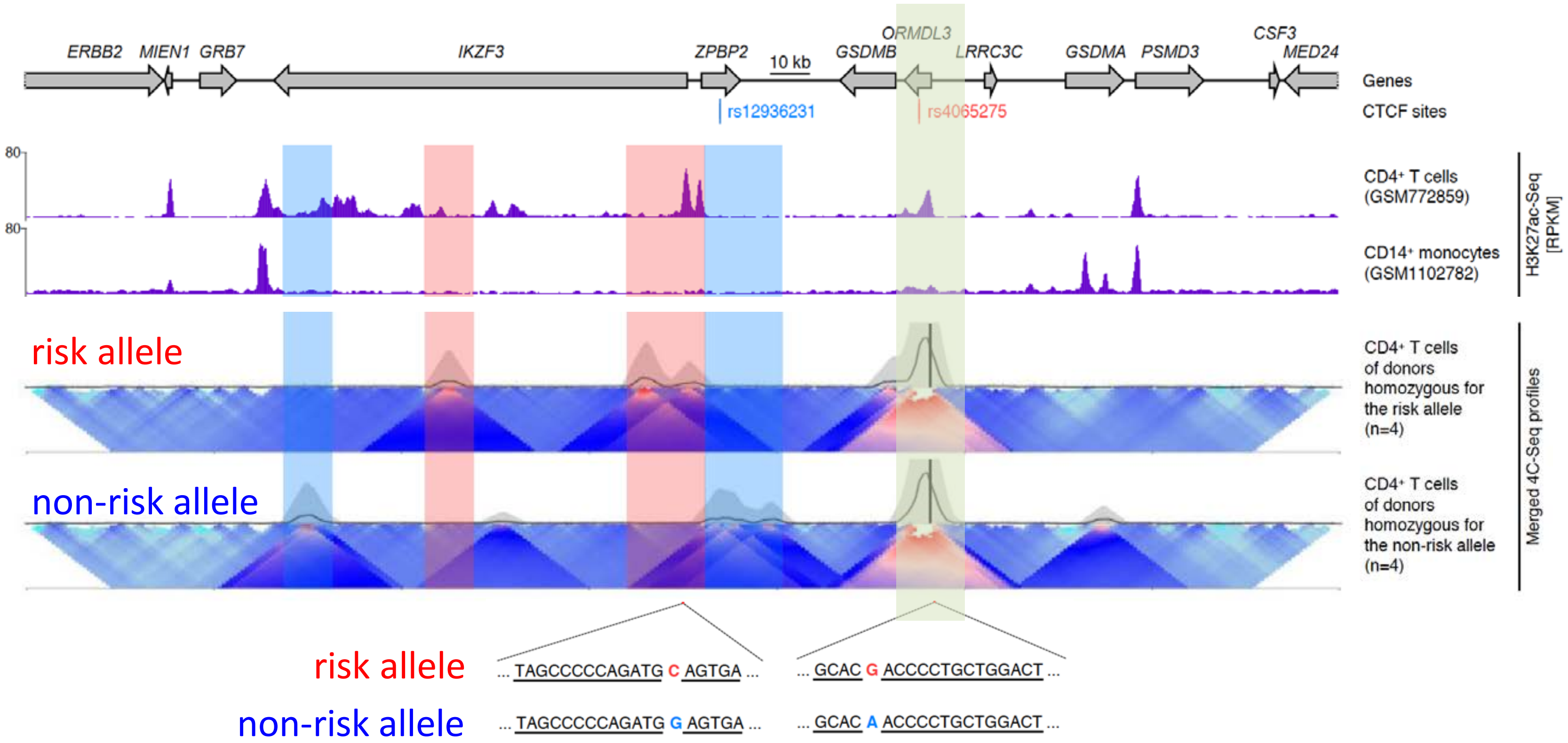
Moffatt *et al. Nature* 2007

17q21 locus is associated with several immune-mediated disorders:

- **Asthma** (Moffatt *et al. Nature* 2007)
- **Type 1 diabetes** (Barrett *et al. Nat Genet* 2009)
- **Rheumatoid arthritis** (Stahl *et al. Nat Genet* 2010)
- **Primary biliary cirrhosis** (Liu *et al. Nat Genet* 2010)
- **Crohn’s disease** (Franke *et al. Nat Genet* 2010)
- **Ulcerative colitis** (McGovern *et al. Nat Genet* 2010; Anderson *et al. Nat Genet* 2011)

Changes in the looping of an asthma-risk related gene

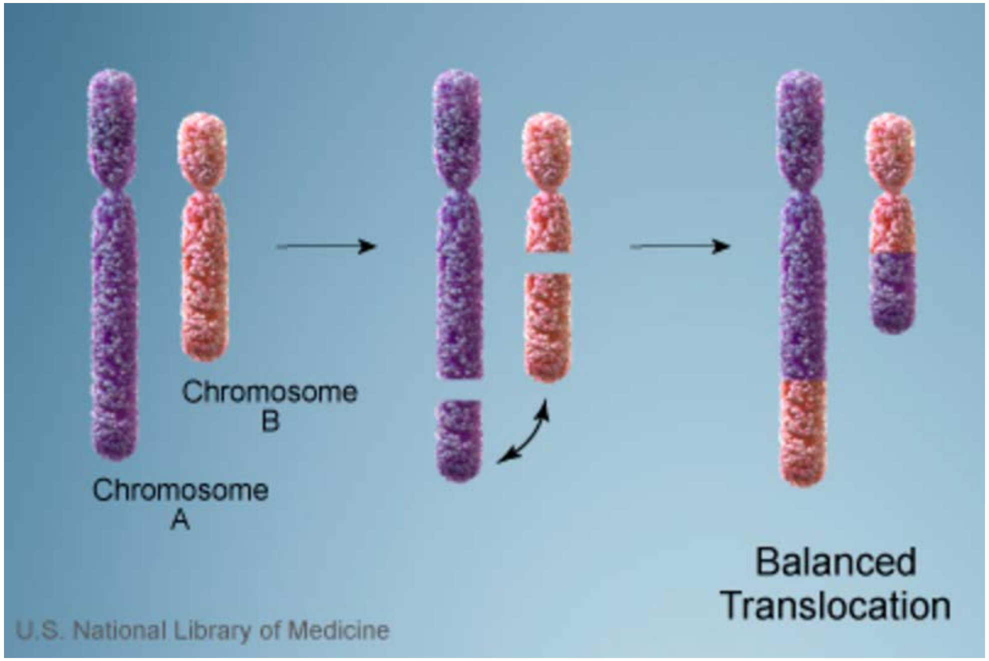


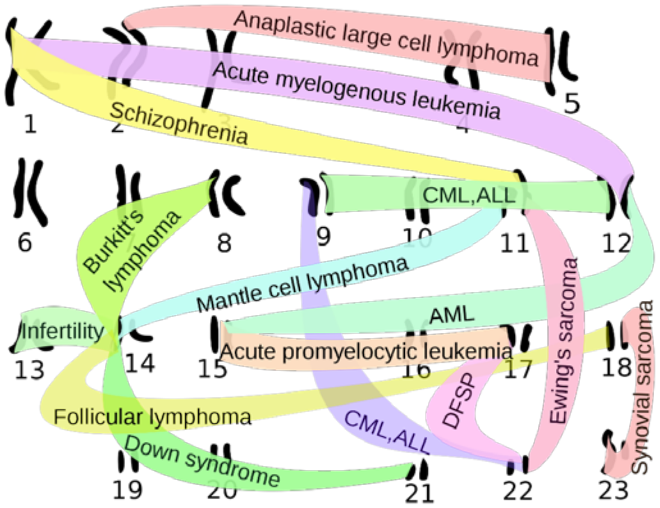


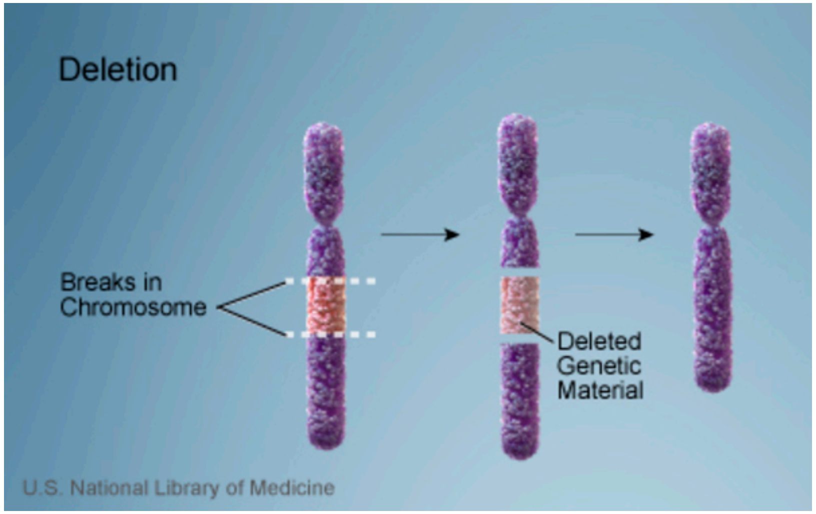
Collaboration with Vijay Lab @ LJI

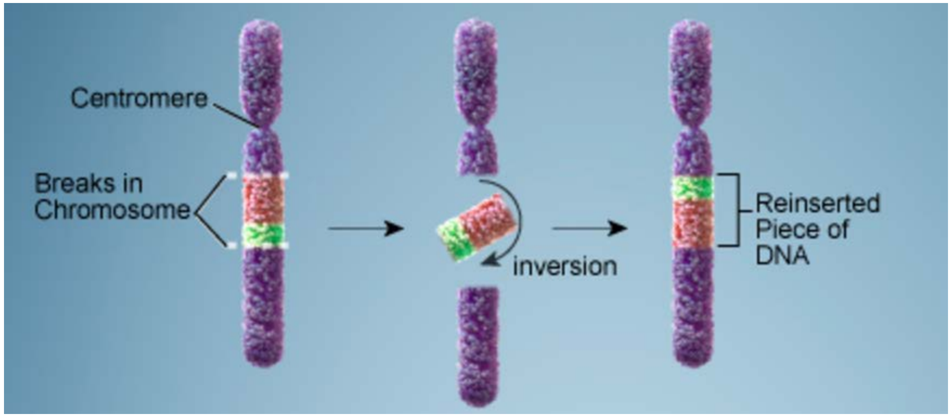
Schmiedel *et al. Nature Communications* 2016

Chromosomal rearrangements are common in cancer





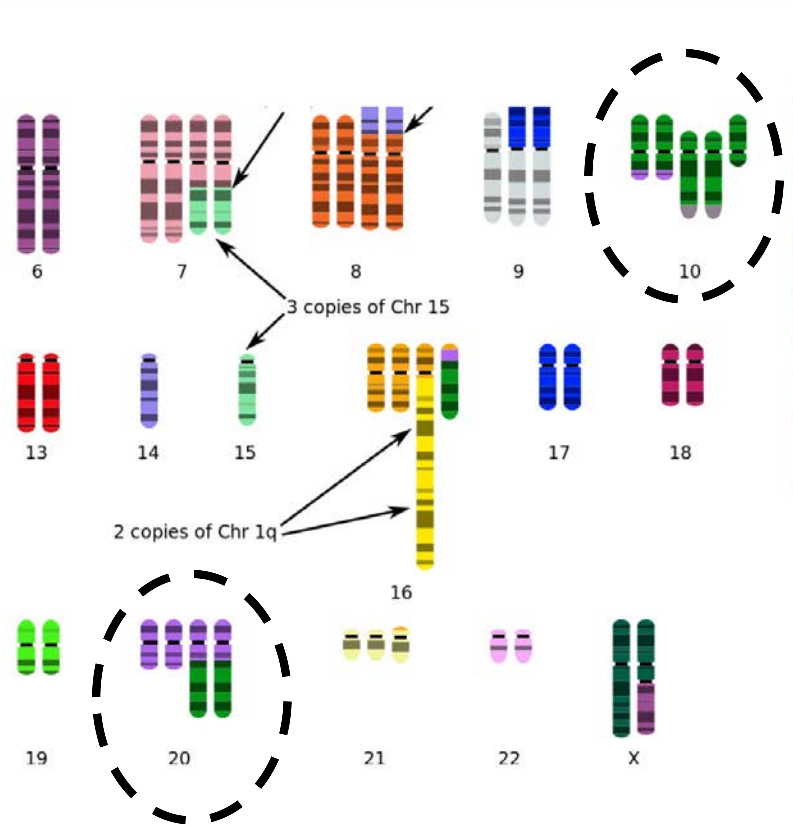


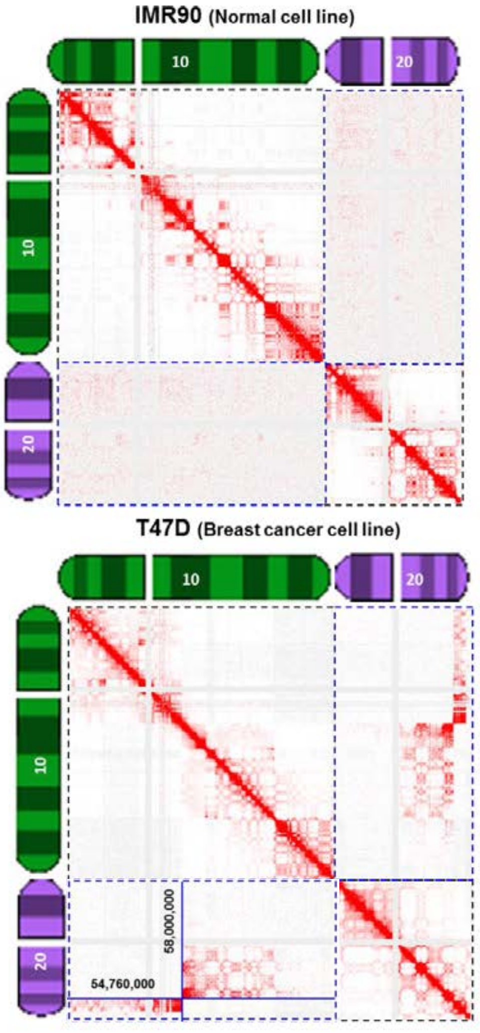


Identification of copy number variations and translocations in cancer cells from Hi-C data

Abhijit Chakraborty, Ferhat Ay ✉ **Published:** 18 October 2017

Bioinformatics, btx664, https://doi.org/10.1093/bioinformatics/btx664

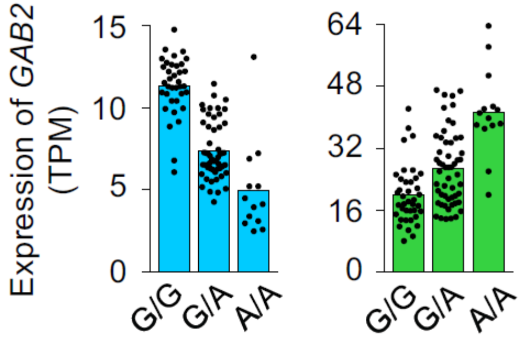




Karyotypically normal cells (fibroblasts)

Breast cancer cells with a translocation

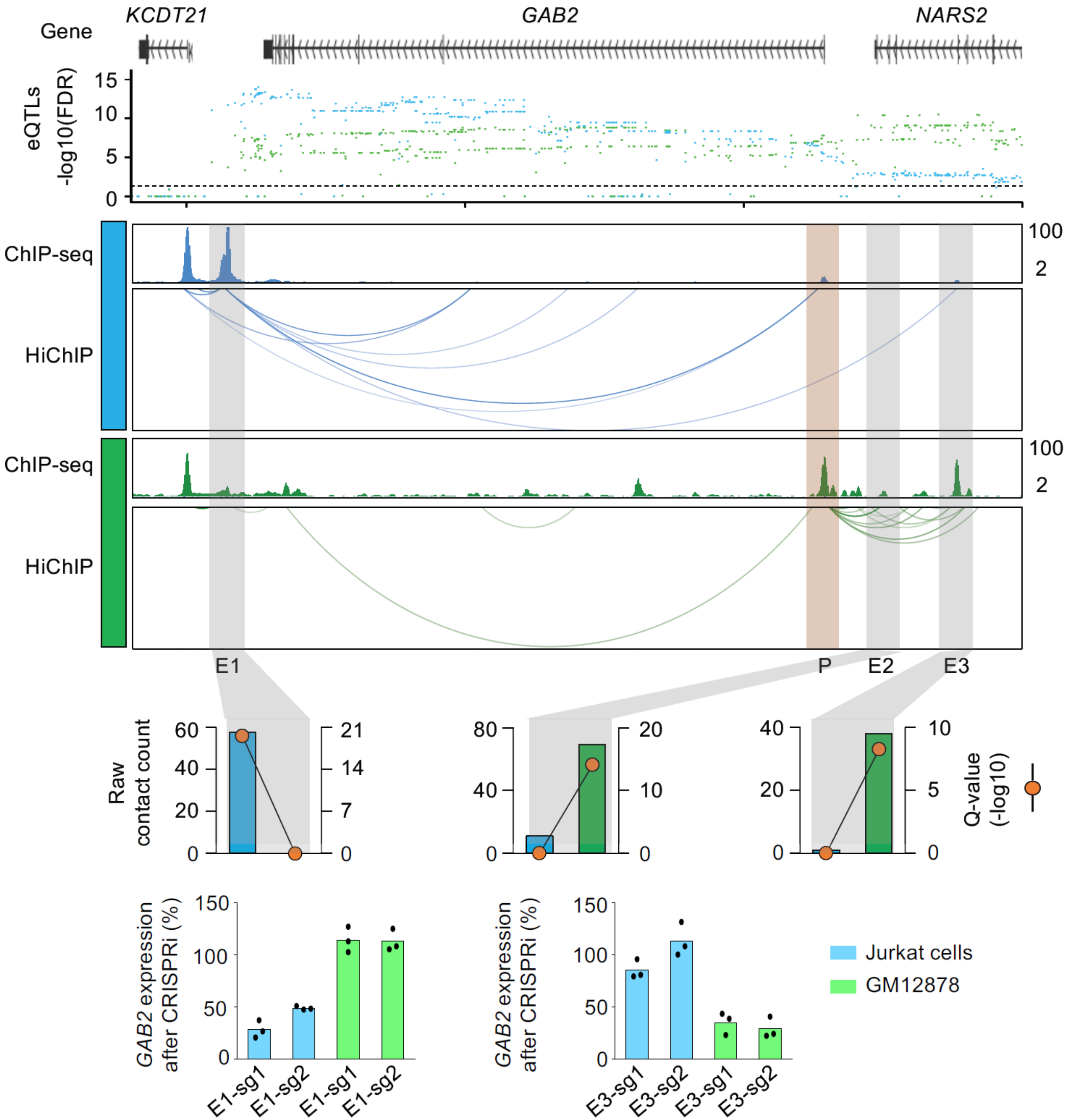
Cell-specific Enhancer function



rs2512539

🟦 Naïve CD4⁺ T cells

🟩 Naïve B cells



Exercise: Visualization of Hi-C data

1. Go to: http://higlass.io
2. Pick a chromosome of your choice
3. Zoom in enough to see A/B compartment patterns corresponding to euchromatin/heterochromatin – Can you guess which one is which?
4. Zoom more to see topological domains (TADs) which are strong square patterns on the diagonal.
5. Find a TAD with a strong corner dot that likely corresponds to a loop between two convergent CTCF binding sites.

References & Course Material

- DNA & Epigenetics: https://ie.unc.edu/dna-epigenetics
- PBS: https://www.pbs.org/wgbh/nova/genes
- Hudson Alpha: https://hudsonalpha.org/wp-content/uploads/2014/04/epigenetics.pdf
- Wikipedia: https://en.wikipedia.org
- Doug Brutlag of Stanford: http://biochem158.stanford.edu/Epigenetics.html
- Epigenetics Game: http://www.letsgethealthy.org/students/games/epigenetics-game
- Coursera – Epigenetic Control of Gene Expression by University of Melbourne