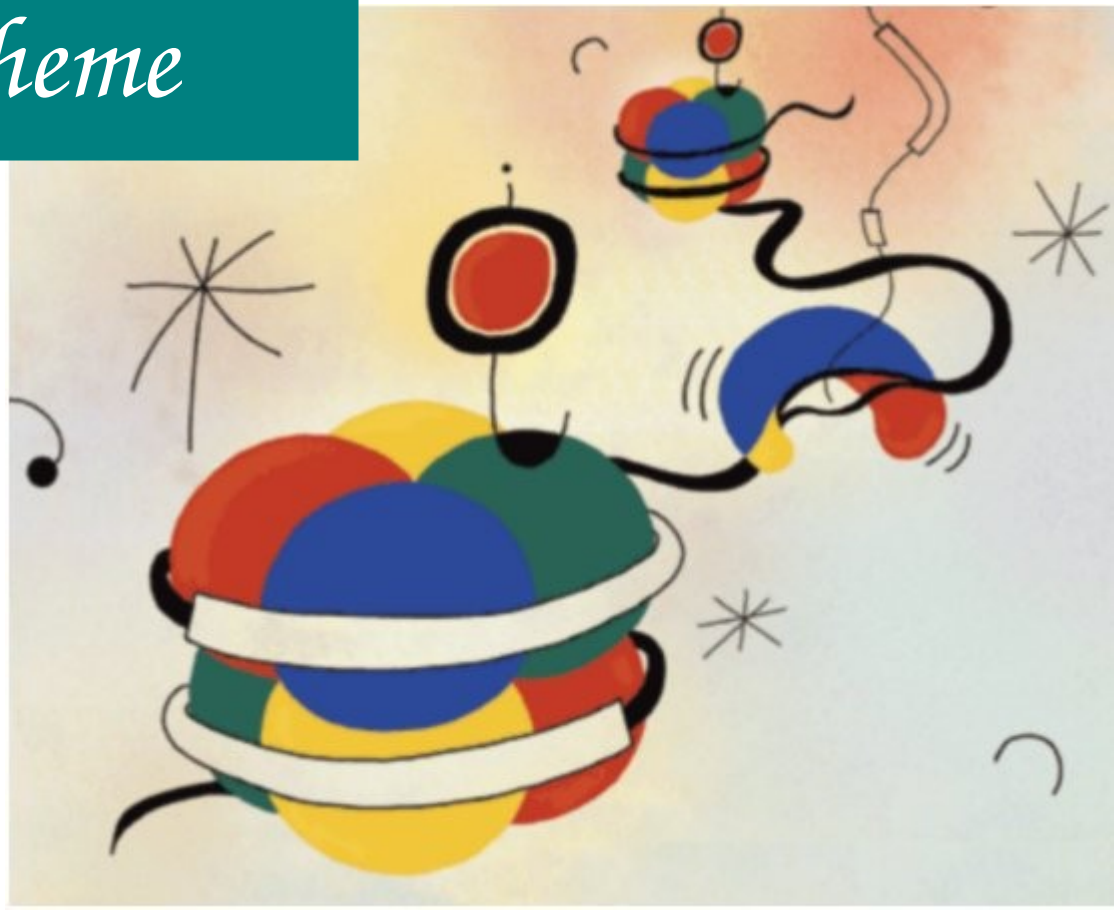
Epigenomics: on the Genome and its Variations on Theme

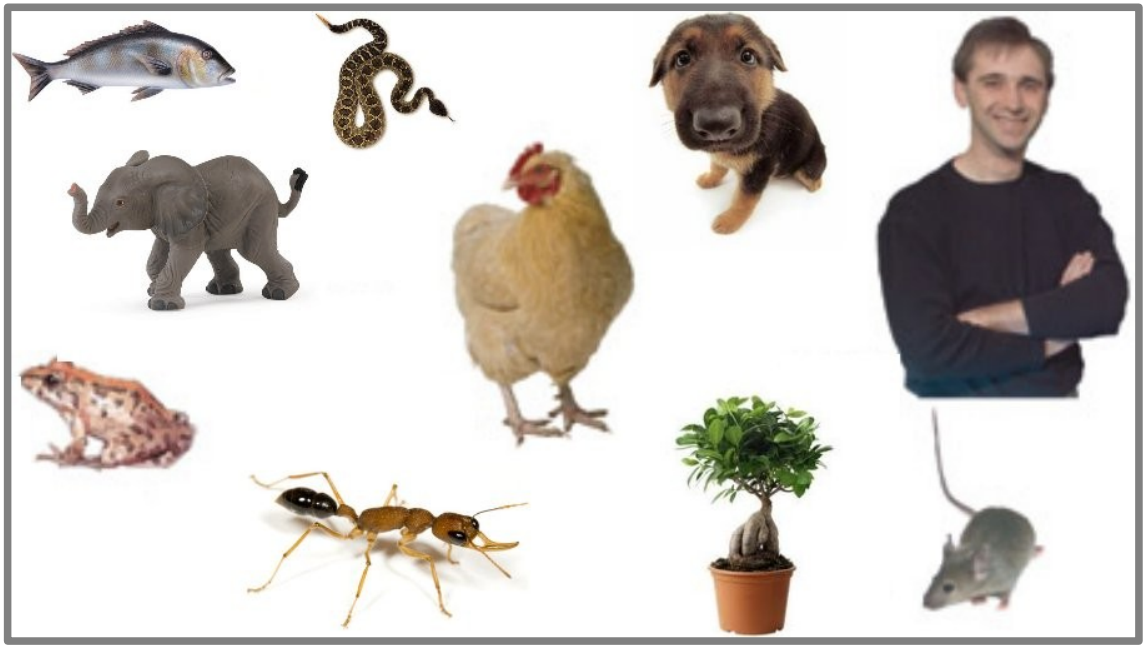


„We certainly need to remember that between genotype and phenotype, and connecting them to each other, there lies a whole complex of developmental processes. It is convenient to have a name for this complex: 'epigenotype' seems suitable"
(Conrad Hal Waddington, 1942)

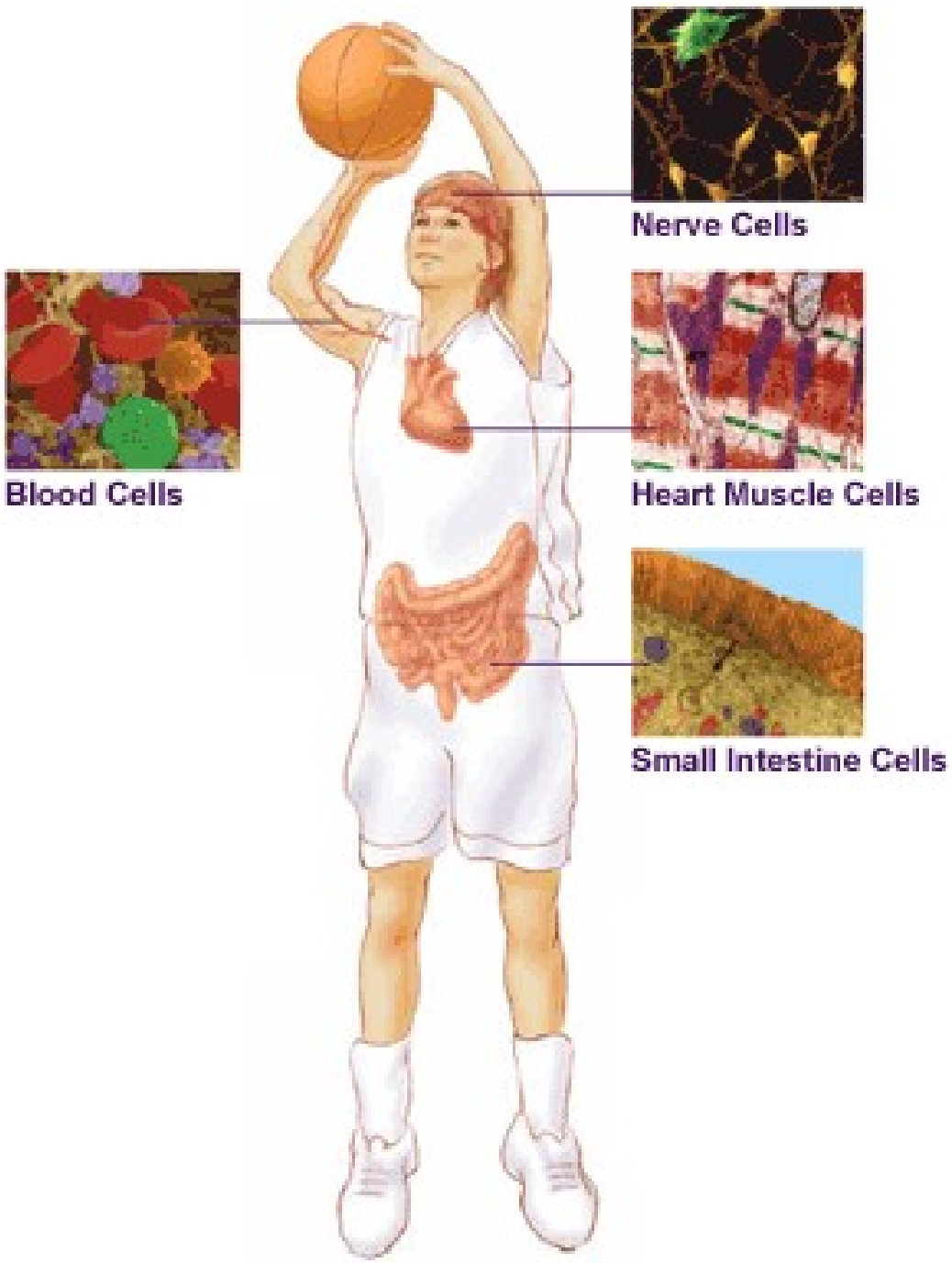
23.10.2024
Ph.D. School in Biology and Molecular Medicine (BEMM)
"BIOINFORMATICS: THEORY AND APPLICATIONS FROM GENOMES TO DRUGS"

Teresa Colombo
teresa.colombo@cnr.it

The key to the difference here lies in the genes



But… Why Do Cells in the Same Body Differ from One Another?



The **genome is the same** in all somatic cells of an organism

Yet, our body is made up of **hundreds different cell types** arranged in individual organs, such as the skin, muscles and nerves, **with their unique cellular phenotype**

The key here lies in **the EPIGENOME**

Different set of genes are active in different cell types

The **epigenome regulates gene expression through chemical modifications** w/o altering the underlying DNA sequence.

Epigenomics, epigenome: a definition

"***Epigenomics*** *is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the* ***epigenome****, which serves as a layer of control beyond the sequence of the DNA itself.*"

National Human Genome Research Institute (NHGRI)

Epigenetic modifications are heritable, reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence.

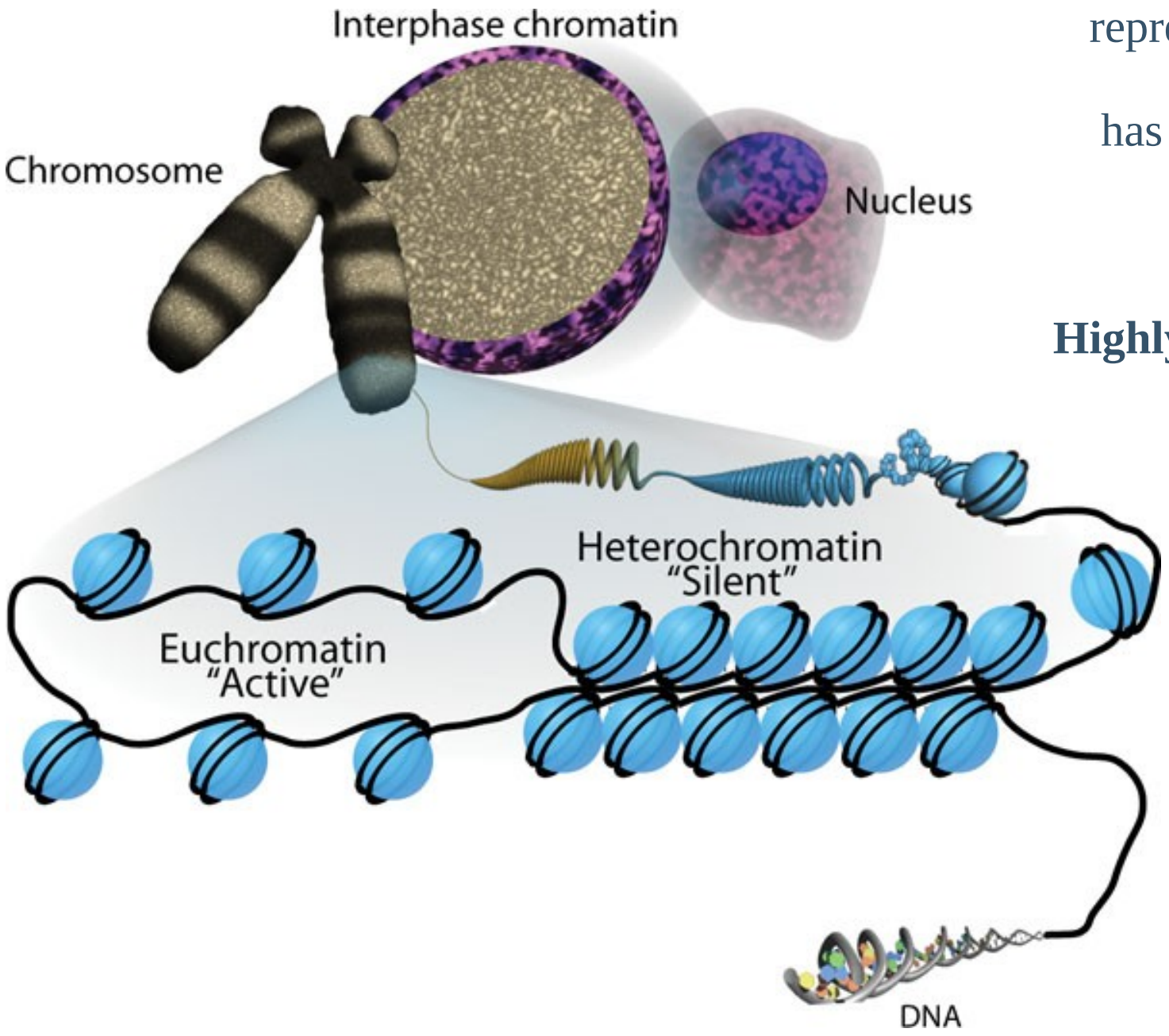
Two Core Information Sets Operate in the Cell

- The Genome
- The Epigenome

The Genome provides building and functional blocks needed for the cell to survive and operate (such as protein-coding/non-coding genes and regulatory elements)

The Epigenome provides additional instructions on ***how, when and where*** these information should be used

Chromatin Structure influences Gene Expression



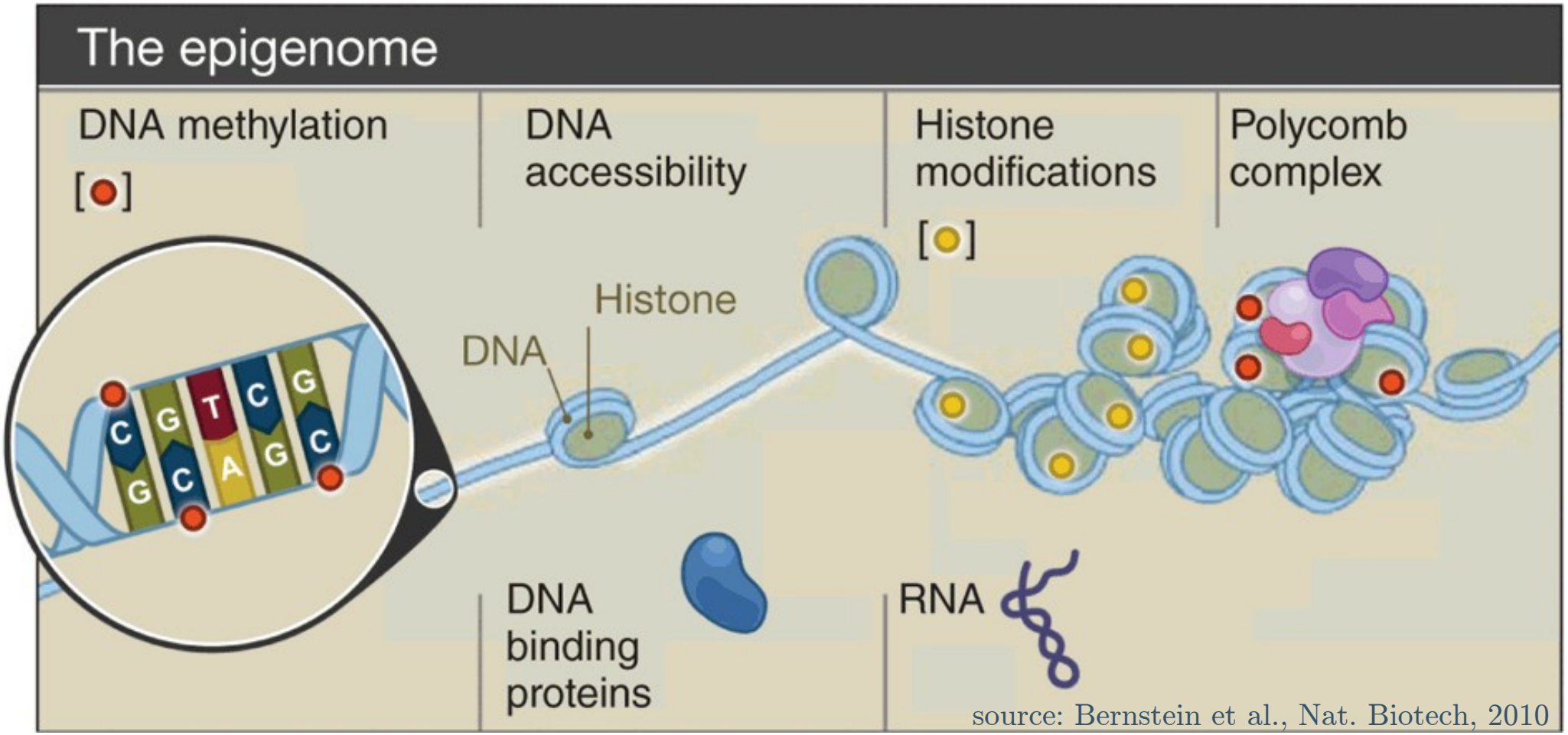
Decondensed chromatin, representing transcriptionally active DNA regions, has been traditionally called ***Euchromatin***

Highly condensed chromatin, with regions of silenced DNA, has been traditionally called ***Heterochromatin***

Heterochromatin can be **constitutive** (irreversibly silenced) or **facultative** (with the ability to become transcriptionally active)

source: http://www.stembook.org/

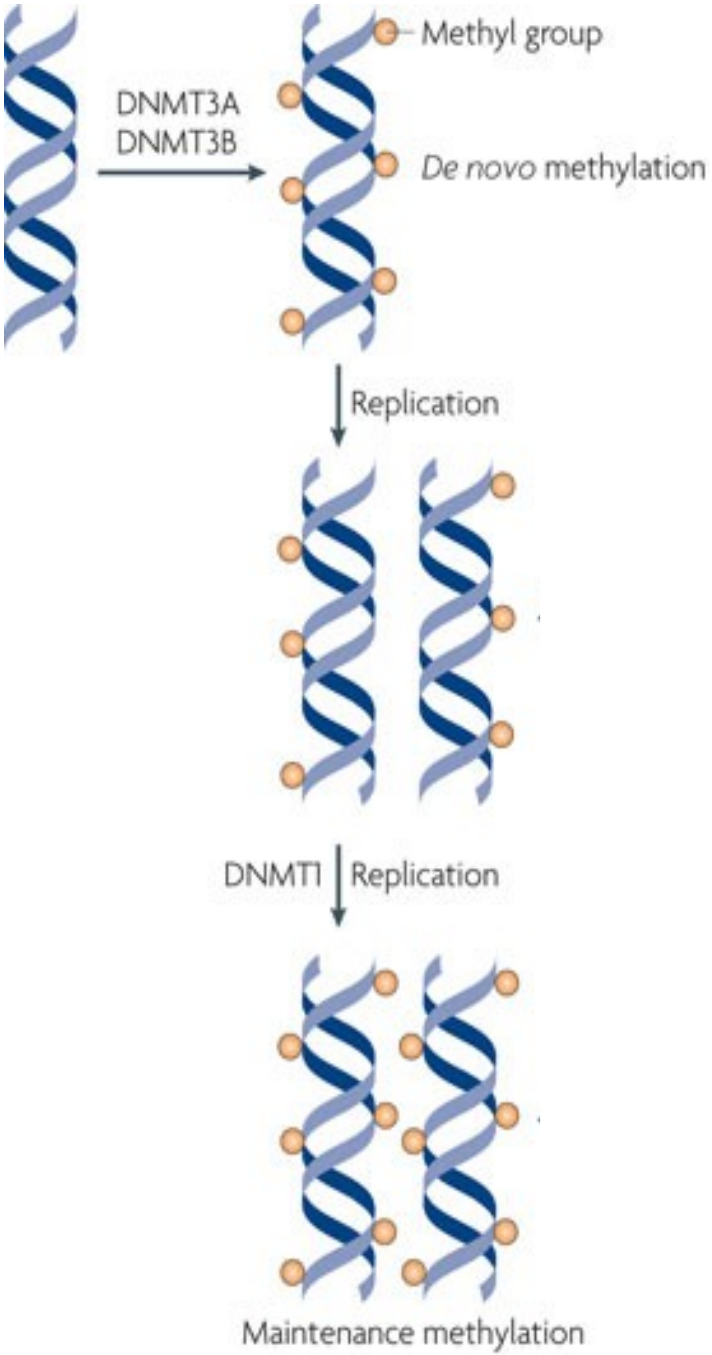
Molecular mechanisms that contribute to the epigenome

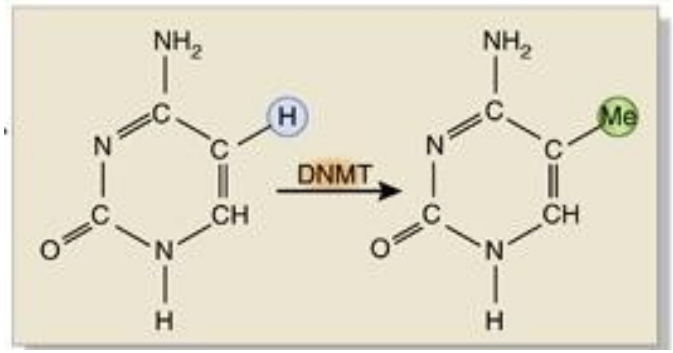


For didactic purposes, epigenetic players can be grouped into few main categories: **DNA methylation**, **histone modifications**, **non-coding RNAs**, **and 3D chromatin structure**.

In reality, a complex interplay among several epigenetic levels of regulation is responsible for an observed phenotype.

DNA methylation



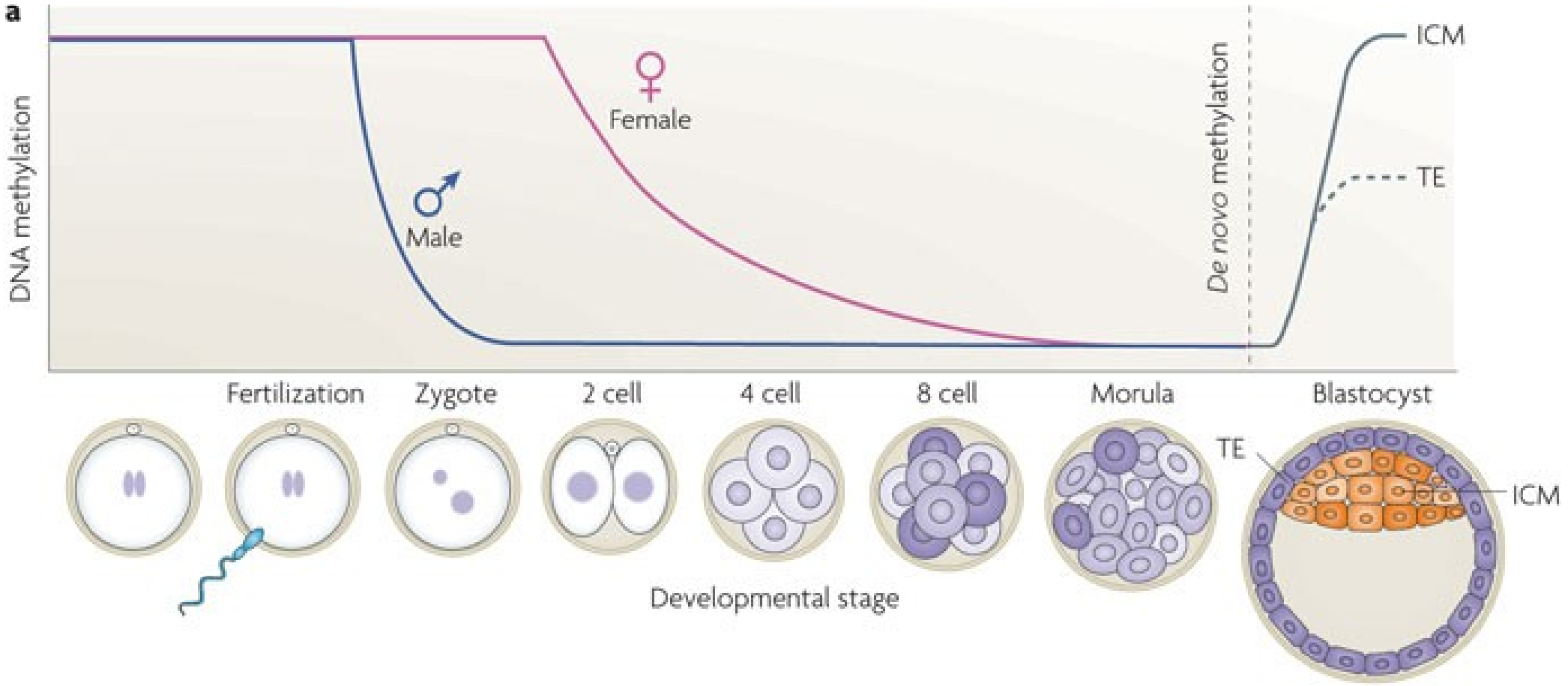


During early development, methylation patterns are initially established by the ***de novo* DNA methyltransferases** (DNMT3A and DNMT3B)

When DNA replication and cell division occur, these marks are maintained in the daughter cells by the **maintenance methyltransferase** (DNMT1)

modified from: Wu and Zhang, Nat. Rev. Mol. Cell Biology, 2010

Dynamics of DNA methylation during early embryo development

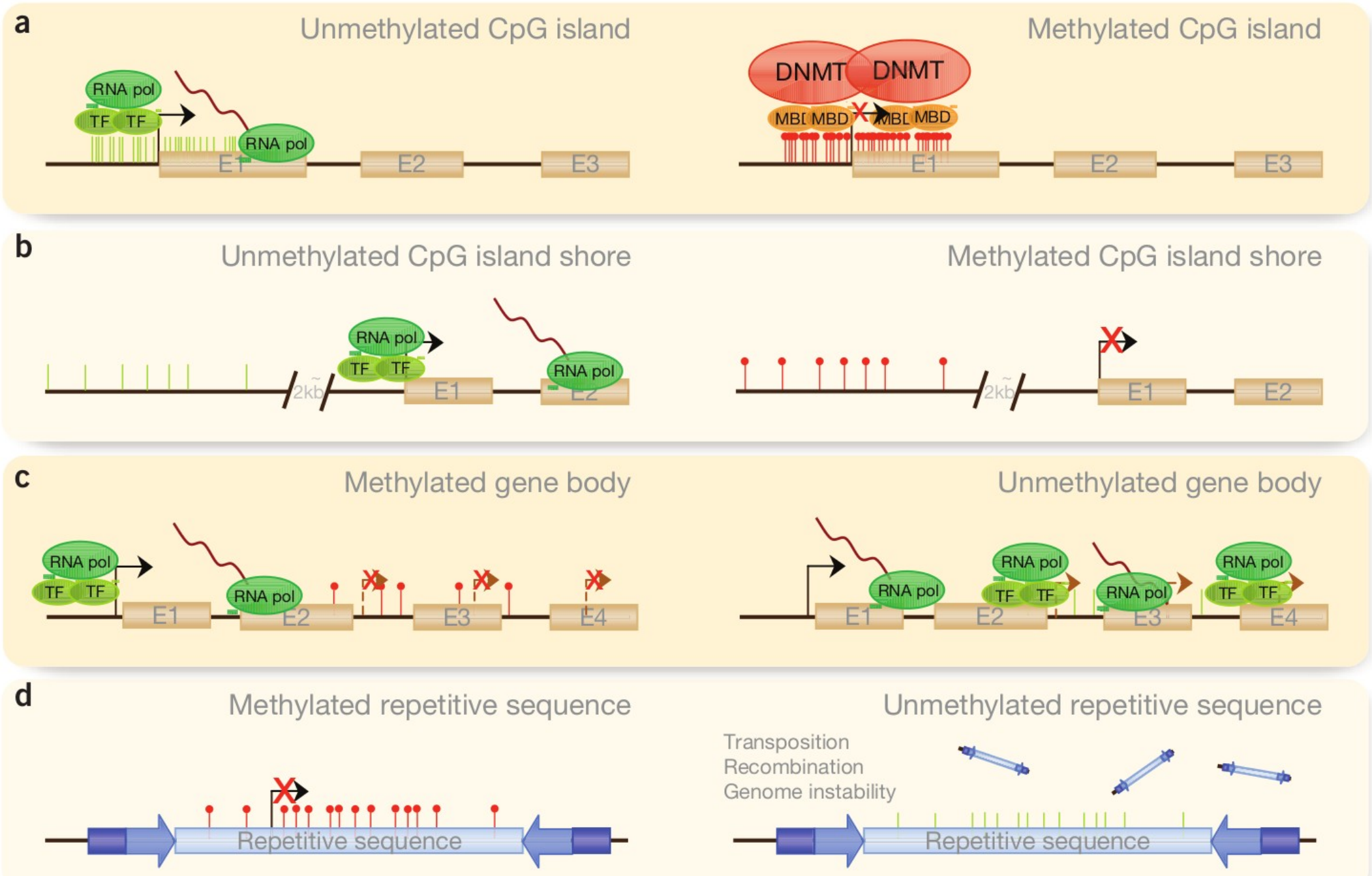


modified from: Wu and Zhang, Nat. Rev. Mol. Cell Biology, 2010

Both the **establishment and maintenance of DNA methylation patterns are crucial for development** as mice deficient in DNMT3B or DNMT1 are embryonic lethal and DNMT3A-null mice die by 4 weeks of age.

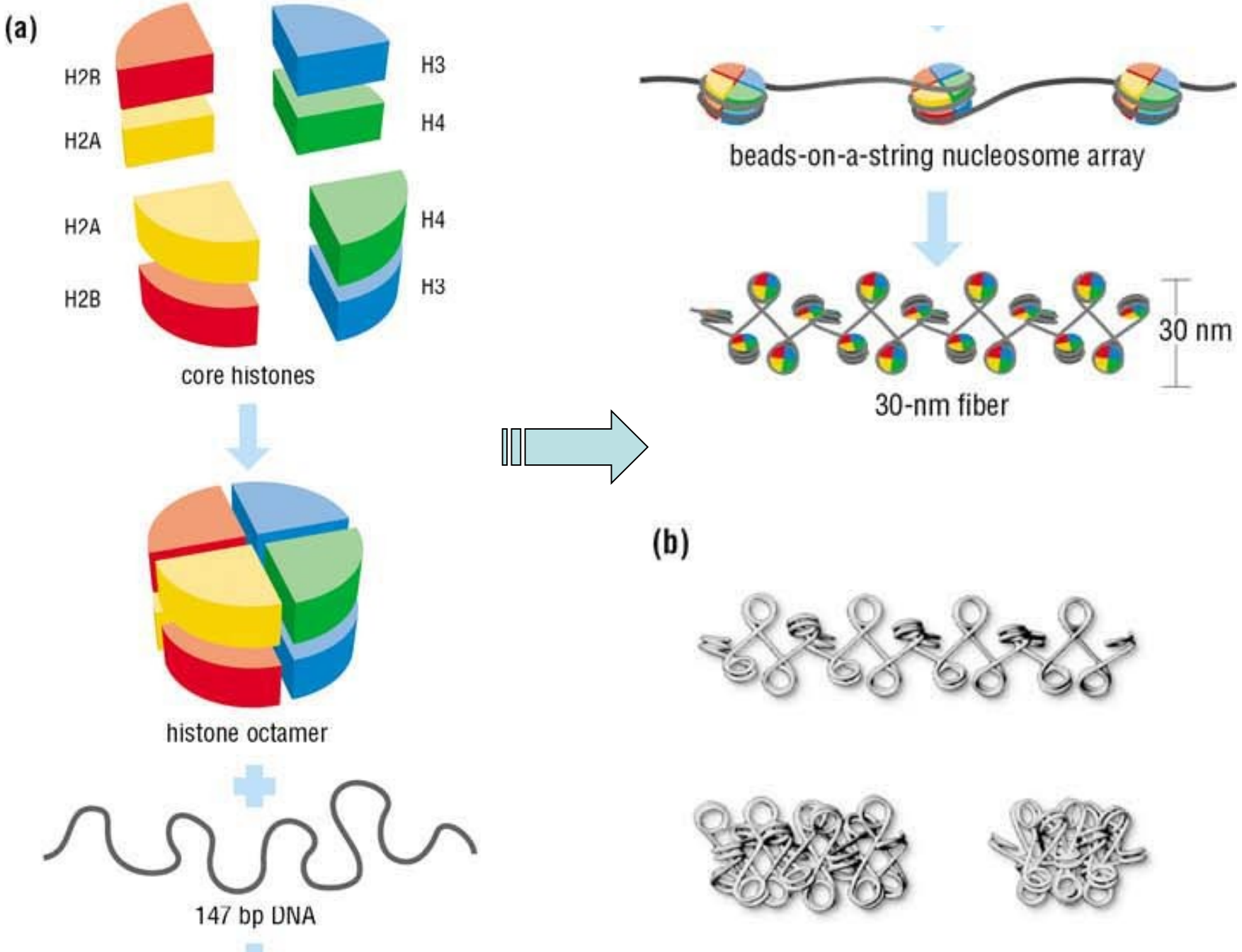
Similarly, long-term silencing provided by DNA methylation is crucial for an organism and **aberrant DNA methylation has been associated with cancer and imprinting-related diseases**.

DNA methylation: normal and aberrant patterns



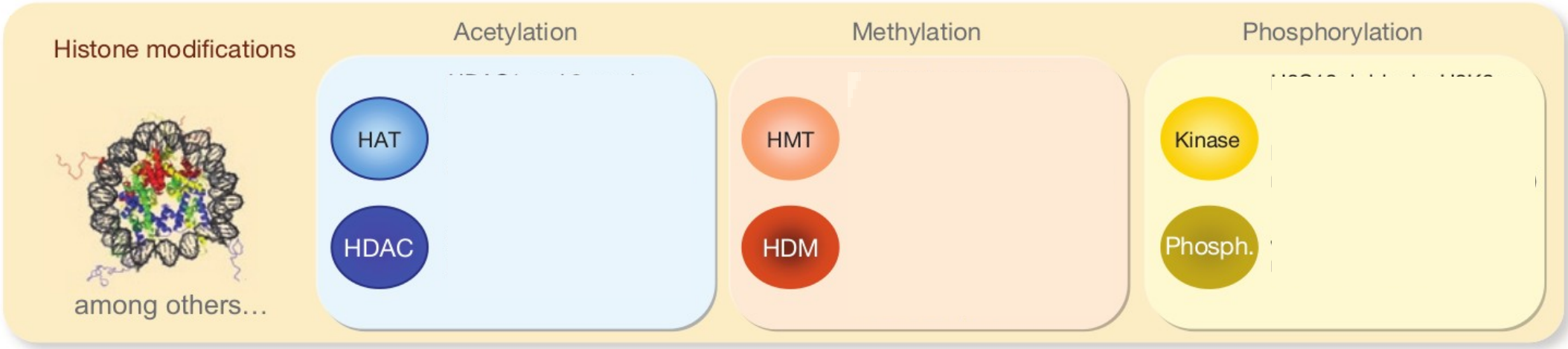
source: Portela and Esteller, Nat. Biotech, 2010

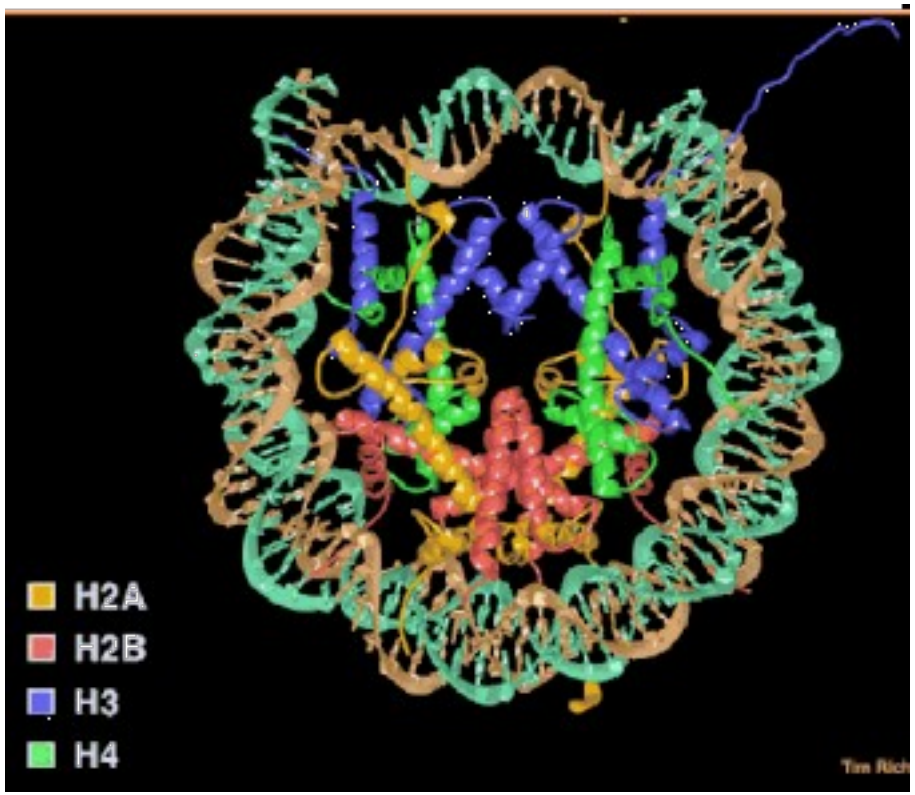
Nucleosomes



The core histones group into two H2A-H2B dimers and one H3-H4 tetramer to form the nucleosome

Histone modifications

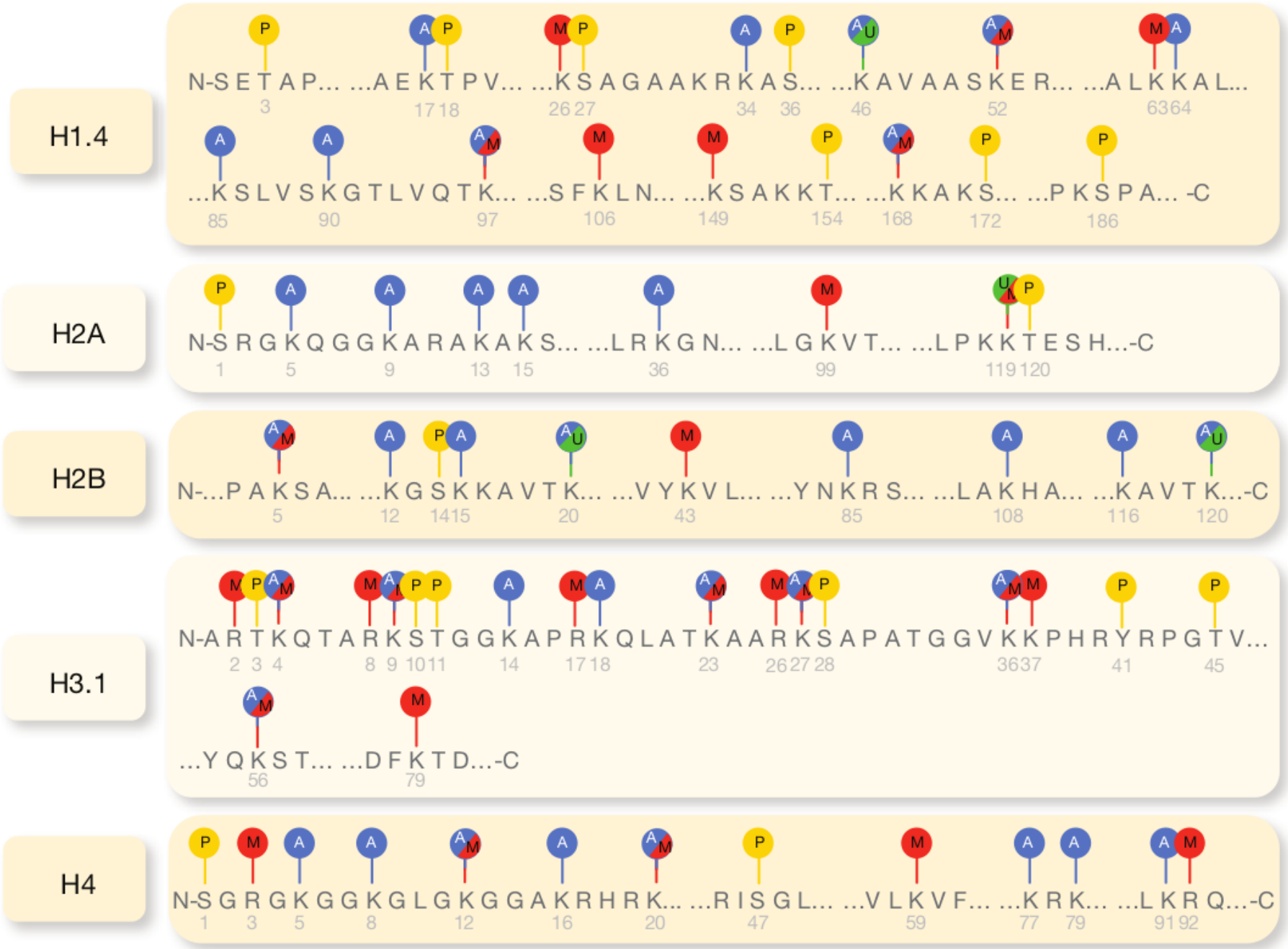




The core histones are predominantly globular except for their **N-terminal tails**, which **are unstructured**.

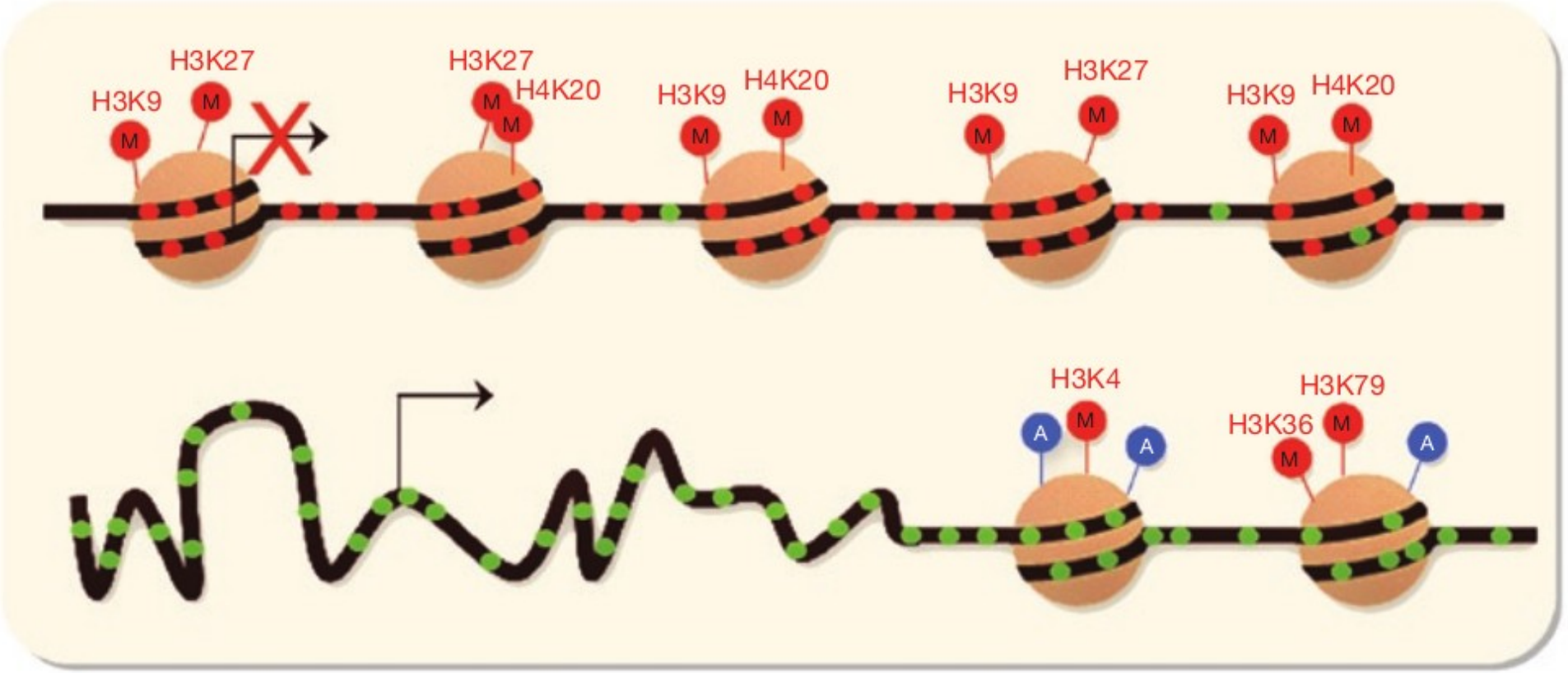
All histones are subject to **post-translational modifications** which mainly occur **in their N-terminal tails**: such as Acetylation, Methylation, Phosphorylation, Ubiquitination ...

Different histone modifications are possible



source: Portela and Esteller, Nat. Biotech, 2010

The Histone Code

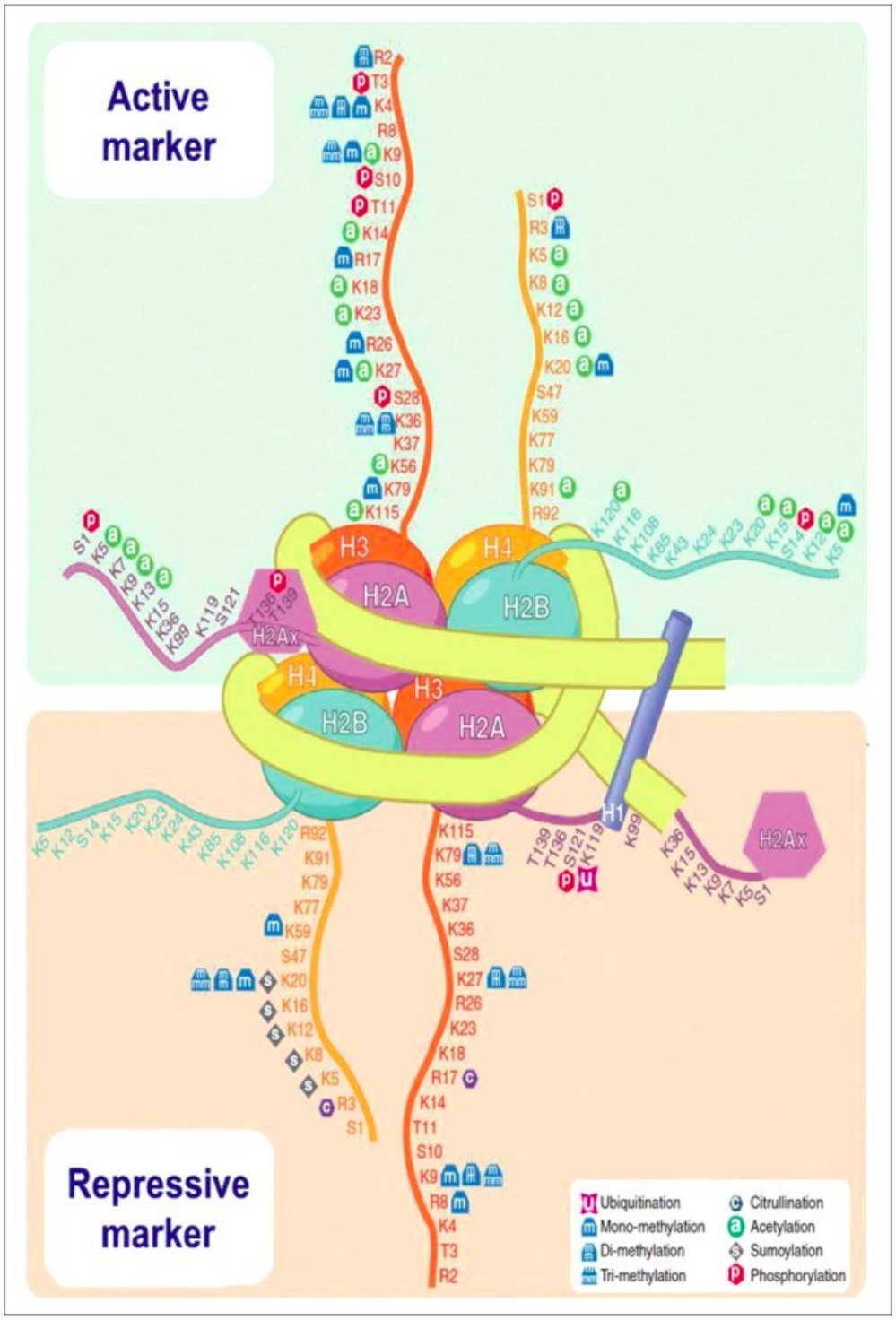


Histones can be modified at multiple sites simultaneously.

Combinations of "marks" in a nucleosome (and of a group of nucleosomes in a genomic region) **specifiy the final outcome** (eg. Active / Repressed / Poised chromatin domain).

Specific combinations of histone modifications are called **'chromatin states**'
and their catalogue (still incomplete) is referred to as '**the histone code'**.

The Histone Code

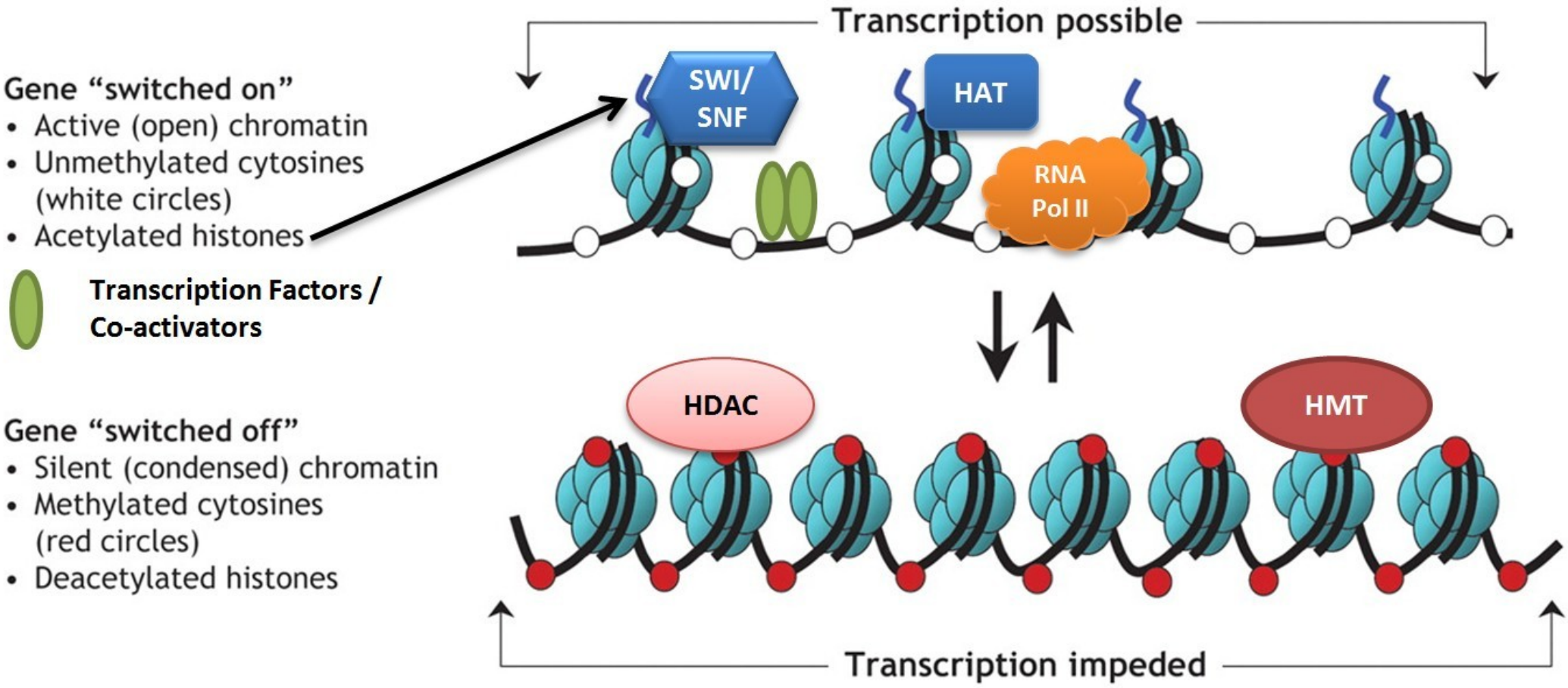


source: Brain Tumor Res Treat. 2014 Apr;2(1):7-21

What is Epigenome?

https://learn.genetics.utah.edu/content/epigenetics/intro/

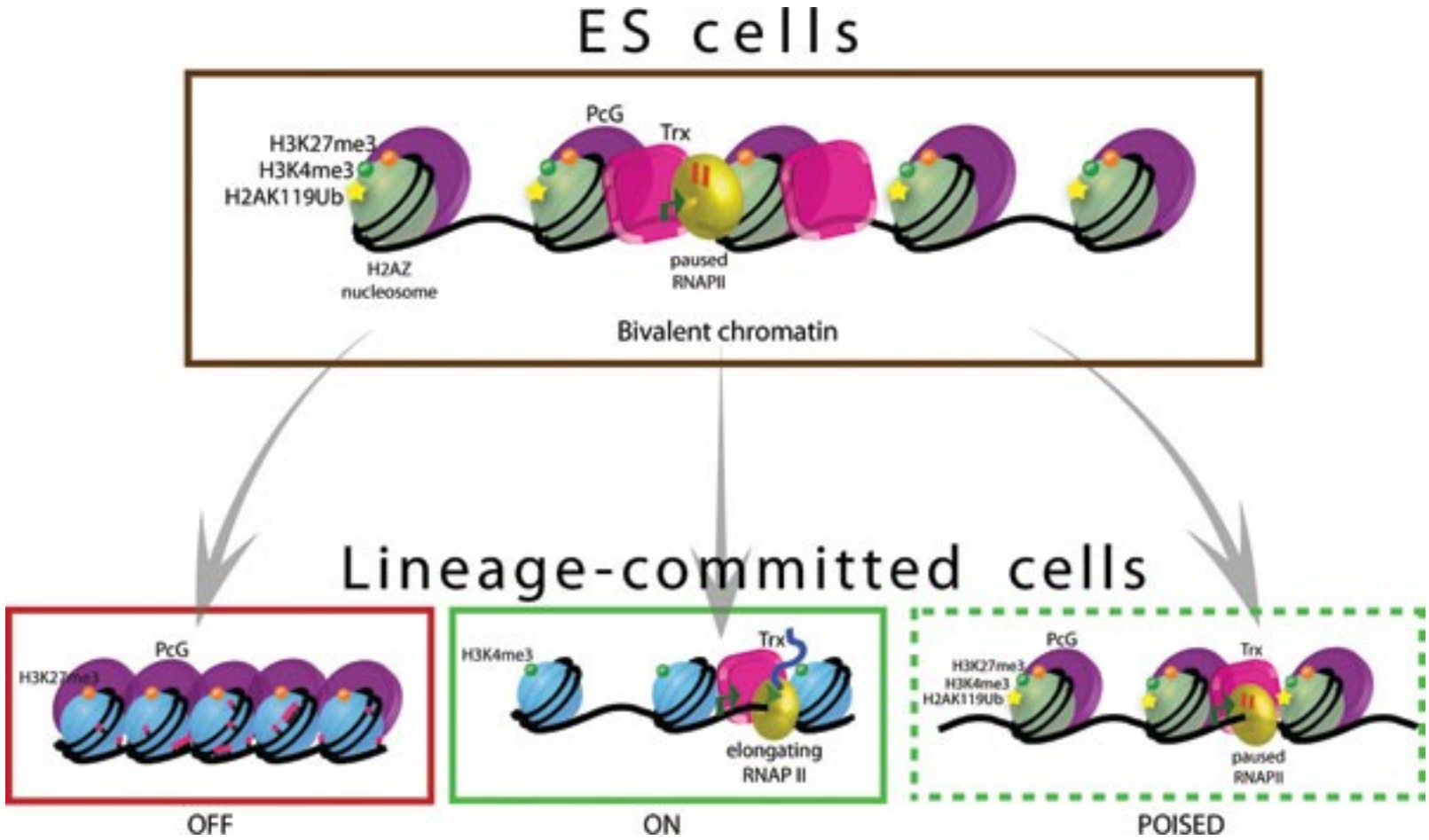
Chromatin remodeling



Source: The role of epigenomics in personalized medicine Expert Review of Precision Medicine and Drug Development. 2. 1-13 (2017)

Dynamic **modifications of chromatin architecture** can **facilitate or hinder the access** of genomic DNA to the regulatory transcription machinery proteins

The chromatin signature of pluripotent cells

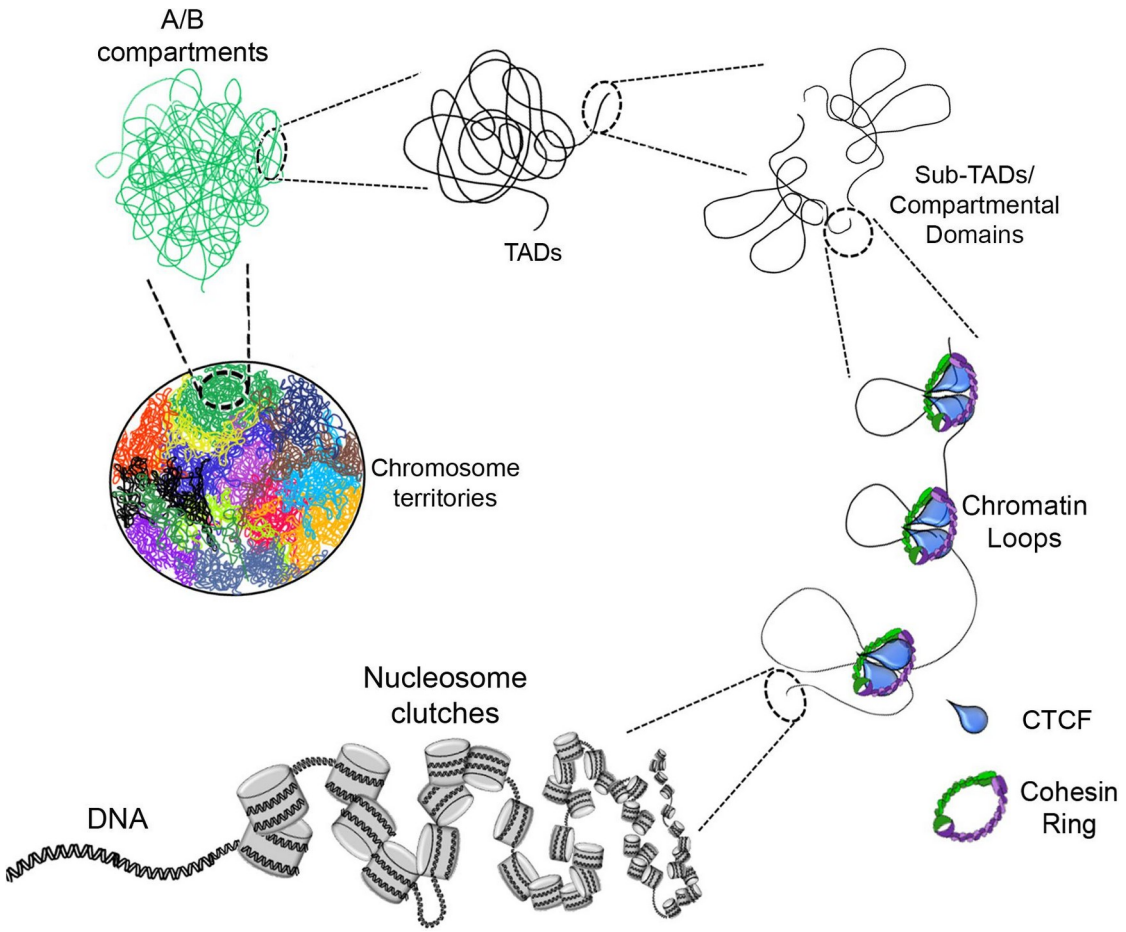


source: http://www.stembook.org/

An interesting case of co-existing histone modifications is found in embryonic stem (ES) cells within the '**bivalent domains**', where the H3k4me3 active mark is found together with the H3K27me3 repressive mark **at promoters of developmentally important genes**. As such, bivalent genes are said to be silent, yet poised for activation.

Pluripotency, a property ascribed to the cells that constitute the early embryo as well as to ES cells, is progressively lost during differentiation with the activation of lineage-specific programs.

Eukaryotic Chromatin Structure is Well Organized and Highly Dynamic



The DNA interacts with histone octamers and aggregates forming **nucleosome** clutches. In the next level of compaction are the **chromatin loops** (in the kb scale) which are formed by loop extrusion and in a greater extent **stabilized by CTCF and the cohesin ring.**

Chromatin loops are the base of **compartmental domains, sub-TADs** (~200 kb) **and TADs** (tens of kb to Mb) with delimited boundaries and **high-rate interactions inside** of these domains.

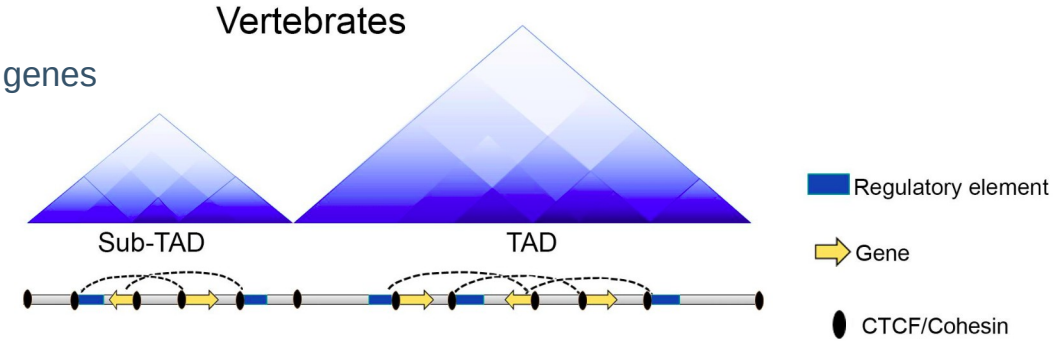
A/B compartments is the next level. Can be determined by gene content, epigenetic marks, DNase hypersensitivity and nuclear localization.

Finally, **Chromosome territories** emerge from non-random spatial arrangement of chromocsomes inside the nucleus

TADs: Topologically Associating Domains

A compatments: high content of transcriptionally active genes

B compatments: high content of silenced genes



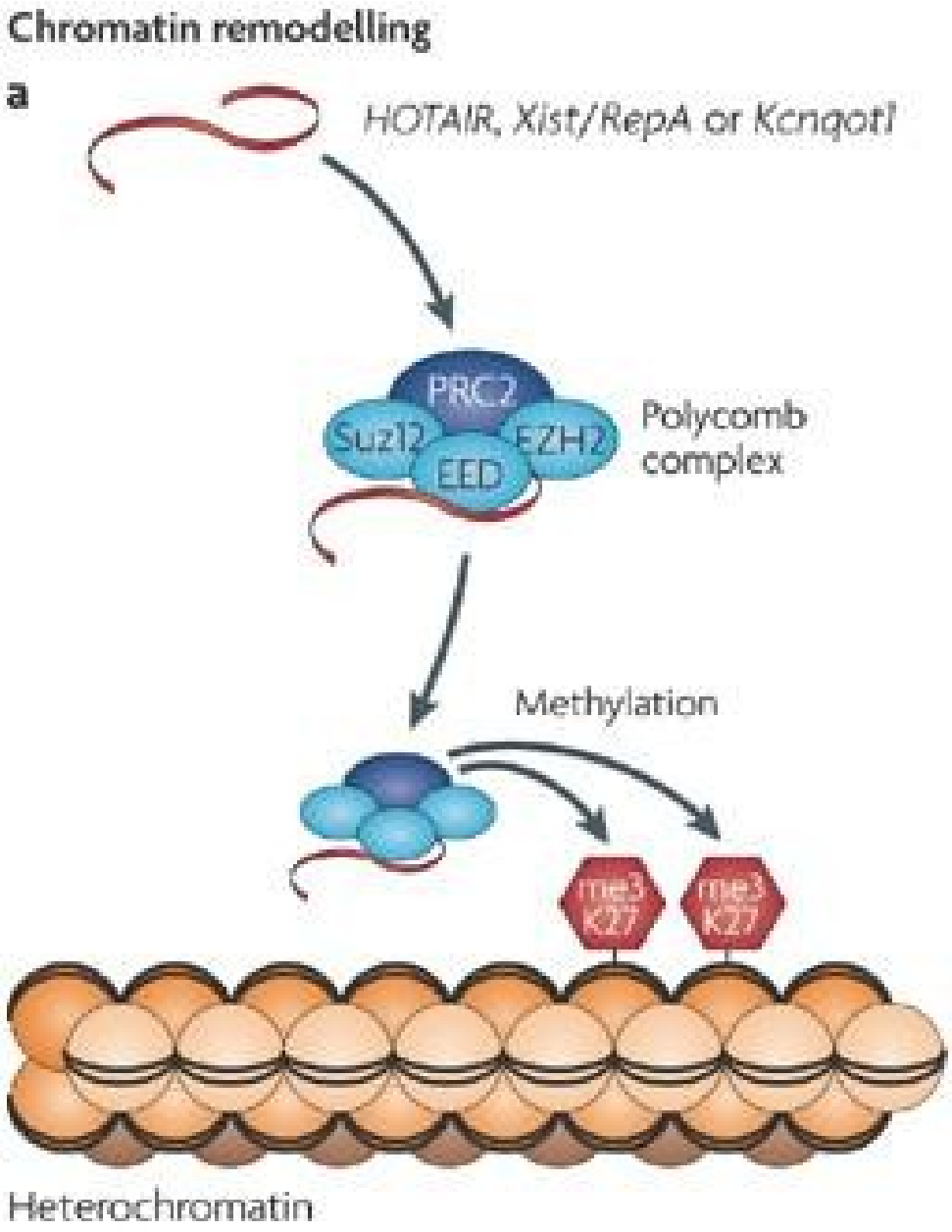
Source:
Front. Genet., 03 November 2020 Sec. Epigenomics and Epigenetics Volume 11 - 2020

Non-coding RNA contribution to the epigenome

More recently, it has become evident that RNA, particularly non-coding RNAs, has a hand in controlling multiple epigenetic phenomena.

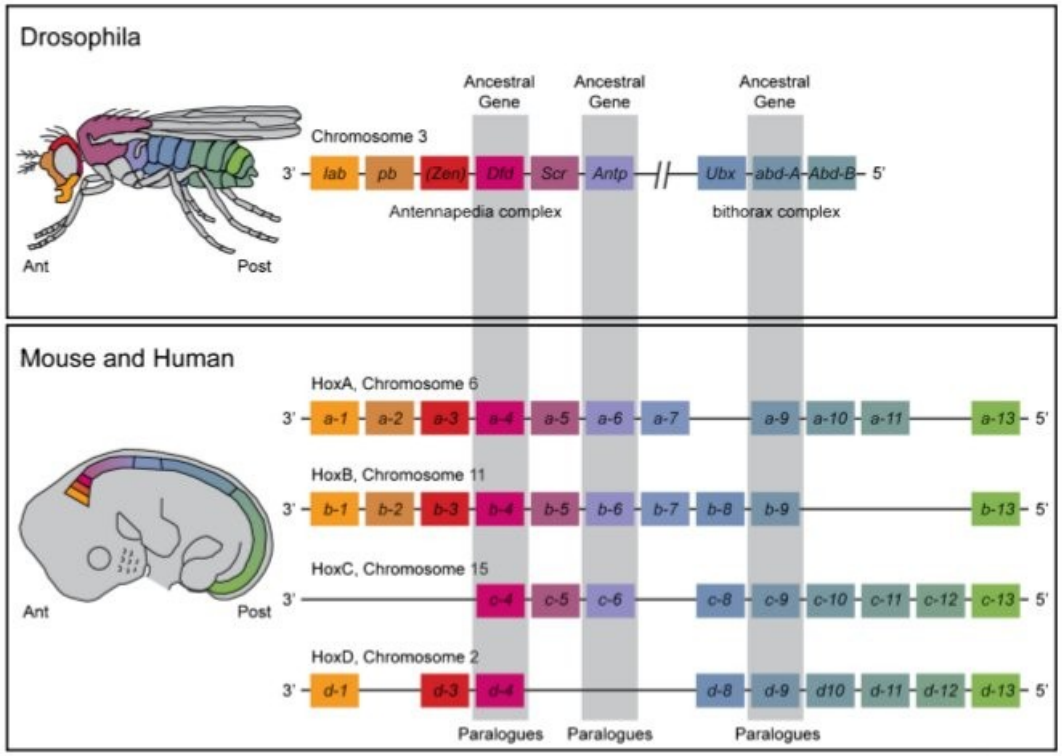
A non-coding RNA is a functional RNA molecule that is not translated into a protein.

These RNAs often act in concert with various component of the cell's chromatin and DNA methylation machinery to achieve stable silencing.

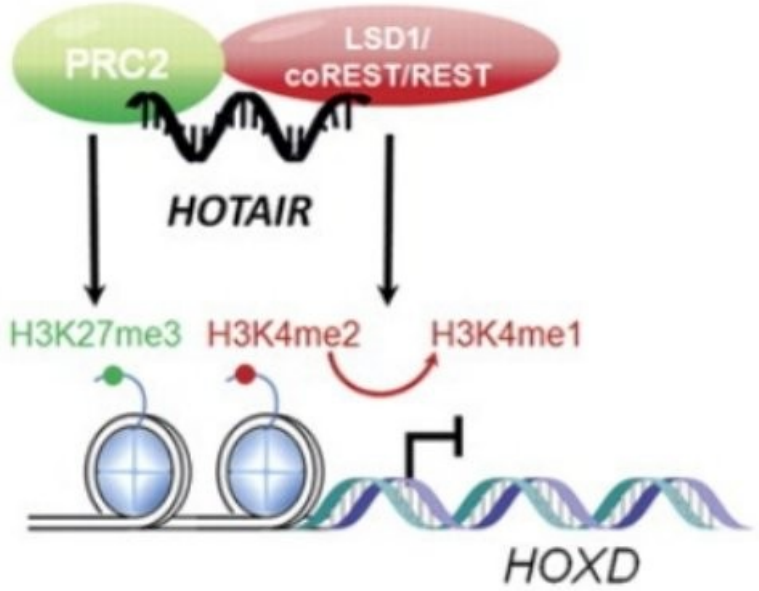


Non-coding RNA contribution to the epigenome

HOTAIR and the HoxC cluster:
LncRNA as Scaffolds for Histone Modifiers



HOTAIR is expressed from an intergenic region of the HoxC cluster and is necesasry for PRC2 occupancy, H3K27me3 and silencing of the HOXD locus (which is located on a completely different chromosome)



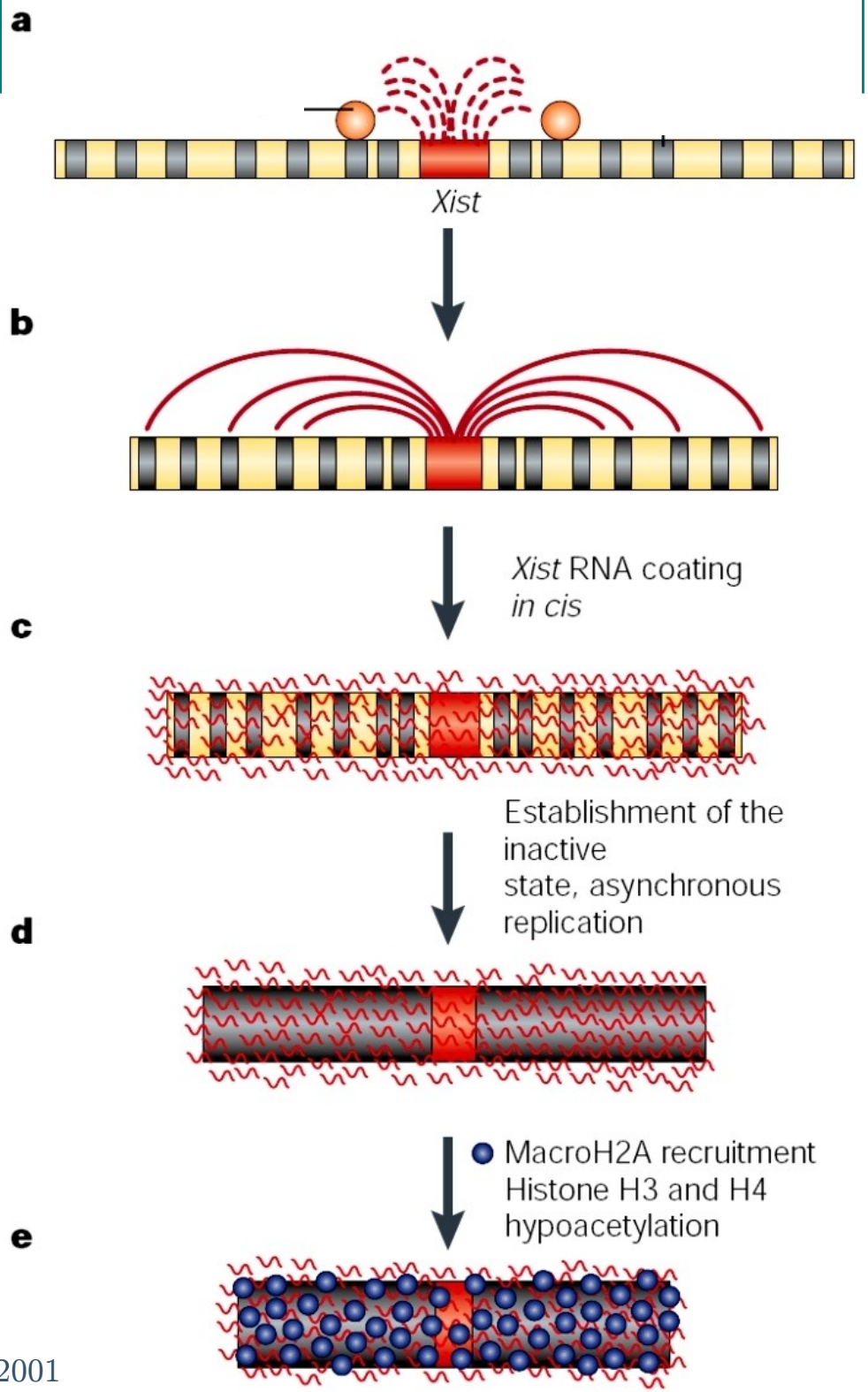
Modified from: Kaikkonen M U et al. Cardiovasc Res 2011;cvr.cvr097

Non-coding RNA contribution to the epigenome

In mammal females one X chromosome must be silenced to compensate gene dosage of X-linked genes between females (2X) and males (1X).

X chromosome inactivation results in random transcriptional inactivation of one of the two X chromosomes present in normal, female mammalian cells.

The **XIST non-coding RNA** is expressed exclusively from the inactive X chromosome and **is required for silencing** by packaging most of genes in the inactive X into transcriptionally inactive heterochromatin.



source: Avner et al., Nature Reviews Genetics, 2001

The epigenome: dynamic modifications at the interface between genome and environment

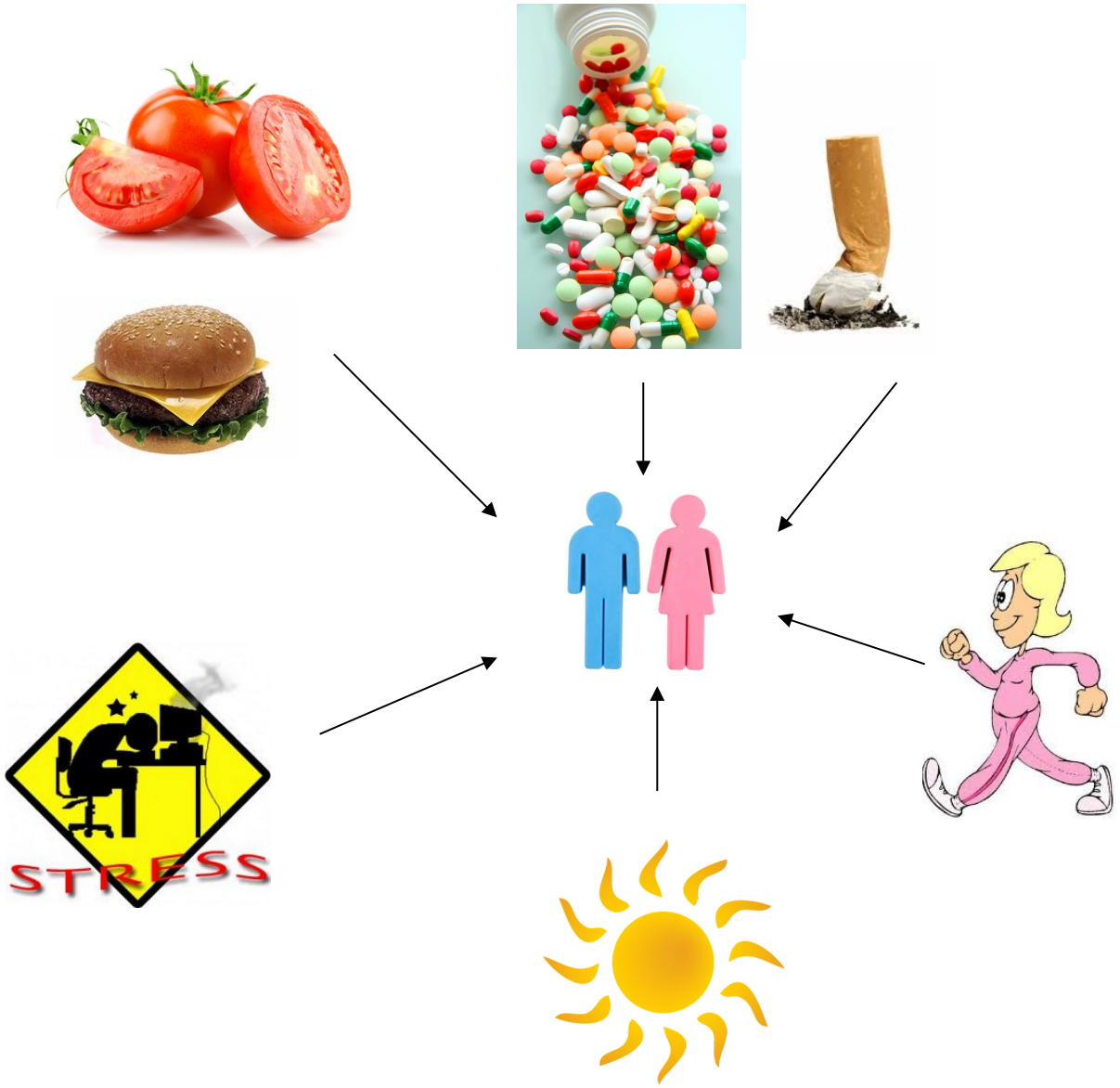
Cell's epigenetic **marks developmentally established are dynamic**

They **can be modified by the external environment** (diet, toxins, physical activity...) **or the internal environment** (hormones)

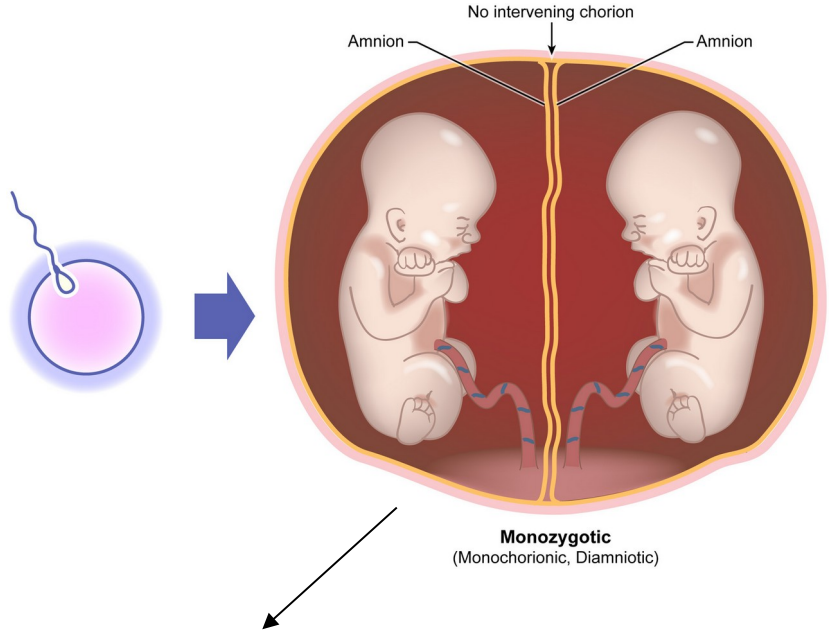
This ability of epigenetics to sense and react to the environment nourishes **increased interest** in the field and has important implications **in human health**

On the one hand many complex diseases (diabetes, rheumatoid arthritis, cancer) have been causally related to aberrant epigenetic patterns

On the other hand reversibility of these marks offers the potential for targeted therapeutics



Identical Twins: an elective case study



Identical twins results when a single embryo splits in two

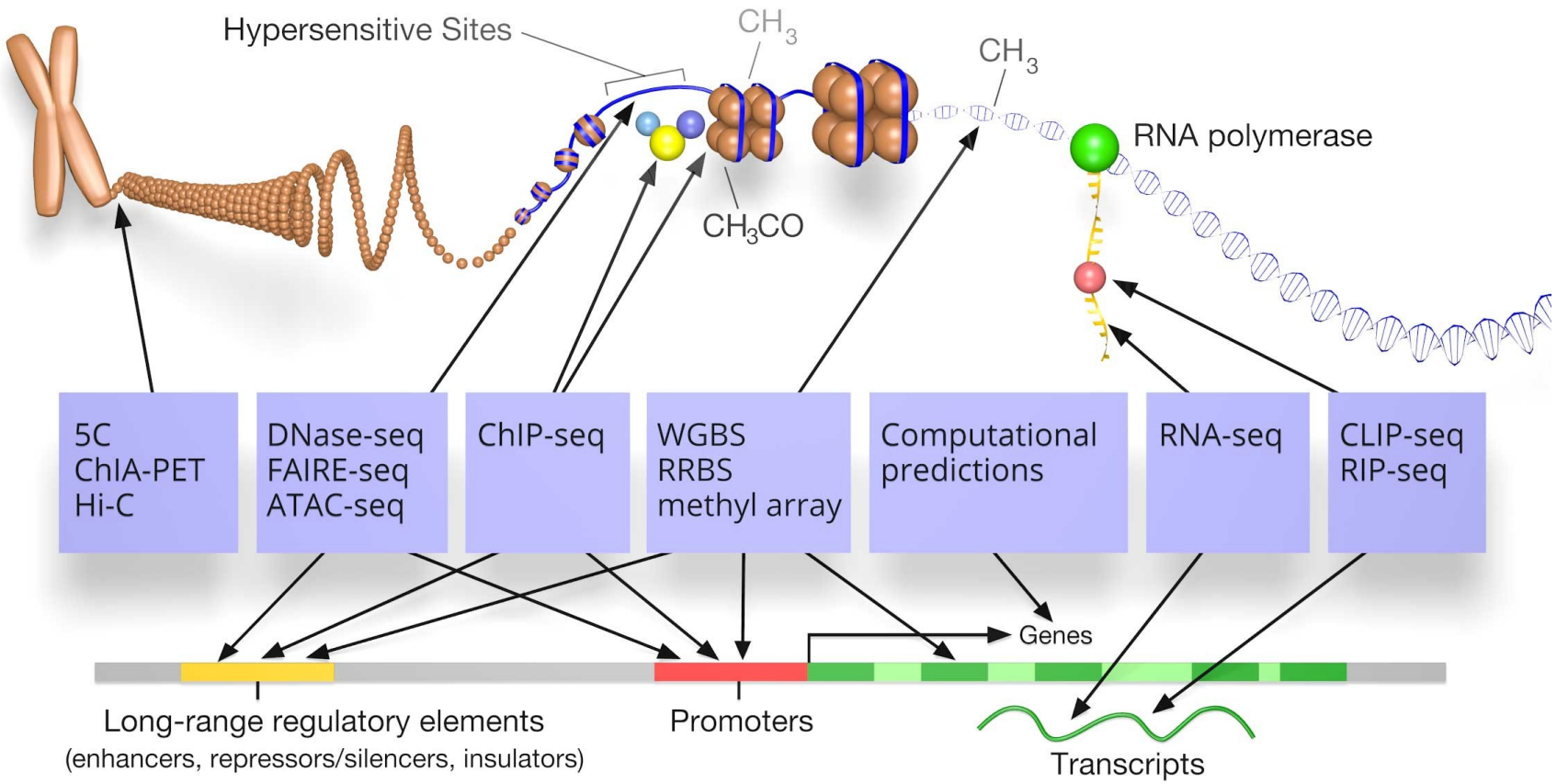
Each embryo has the same genome and the same epigenome

With their identical genetic make-up, monozygotic twins have been electively studied for sorting out genetic effects from those of the environment

As the twins age, their environment begin to differ, ultimately resulting in two different epigenomes

Widespread epigenetic differences between twins that **accumulate over the years** indicate high impact of age and environment on cell state and ultimately human health

Overview of High-Throughput Sequencing Technologies for Genome-Wide Epigenomic Data Collection



ENCODE

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

DNA methylation profiling

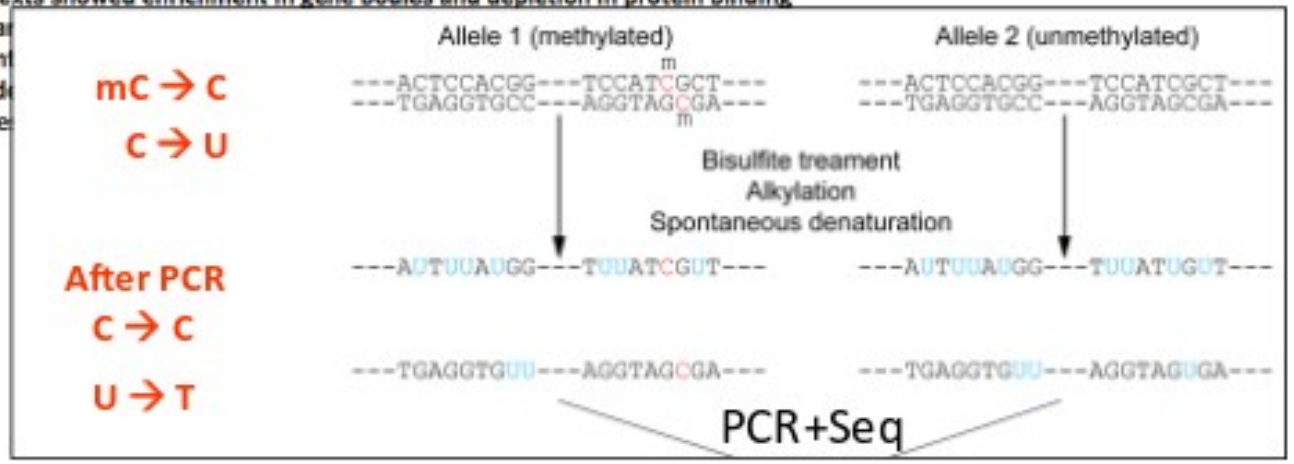
Vol 462 | 19 November 2009 | doi:10.1038/nature08514 nature

ARTICLES

Human DNA methylomes at base resolution show widespread epigenomic differences

Ryan Lister[1]*, Mattia Pelizzola[1]*, Robert H. Dowen[1], R. David Hawkins[2], Gary Hon[2], Julian Tonti-Filippini[4], Joseph R. Nery[1], Leonard Lee[2], Zhen Ye[2], Que-Minh Ngo[2], Lee Edsall[2], Jessica Antosiewicz-Bourget[5,6], Ron Stewart[5,6], Victor Ruotti[5,6], A. Harvey Millar[4], James A. Thomson[5,6,7,8], Bing Ren[2,3] & Joseph R. Ecker[1]

DNA cytosine methylation is a central epigenetic modification that has essential roles in cellular processes including genome regulation, development and disease. Here we present the first genome-wide, single-base-resolution maps of methylated cytosines in a mammalian genome, from both human embryonic stem cells and fetal fibroblasts, along with comparative analysis of messenger RNA and small RNA components of the transcriptome, several histone modifications, and sites of DNA–protein interaction for several key regulatory factors. Widespread differences were identified in the composition and patterning of cytosine methylation between the two genomes. Nearly one-quarter of all methylation identified in embryonic stem cells was in a non-CG context, suggesting that embryonic stem cells may use different methylation mechanisms to affect gene regulation. Methylation in non-CG contexts showed enrichment in gene bodies and depletion in protein binding sites and enhancers. Non-CG methylation disappear restored in induced pluripotent stem cells. We ident involved in pluripotency and differentiation, and wid transcriptional activity. These reference epigenome modification in human disease and development.



ChIP sequencing

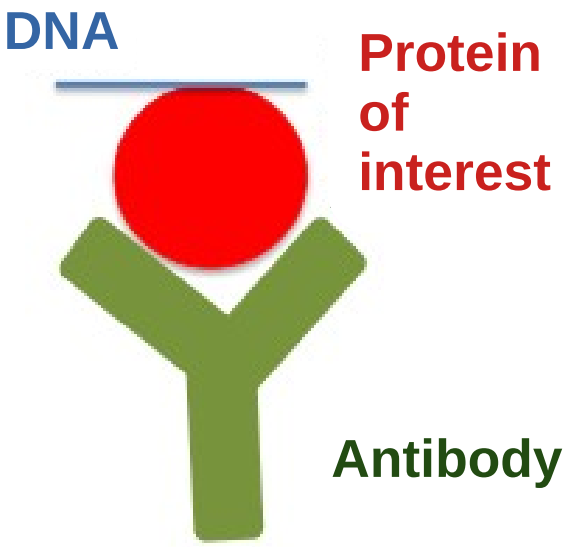
Genome-Wide Mapping of in Vivo Protein-DNA Interactions

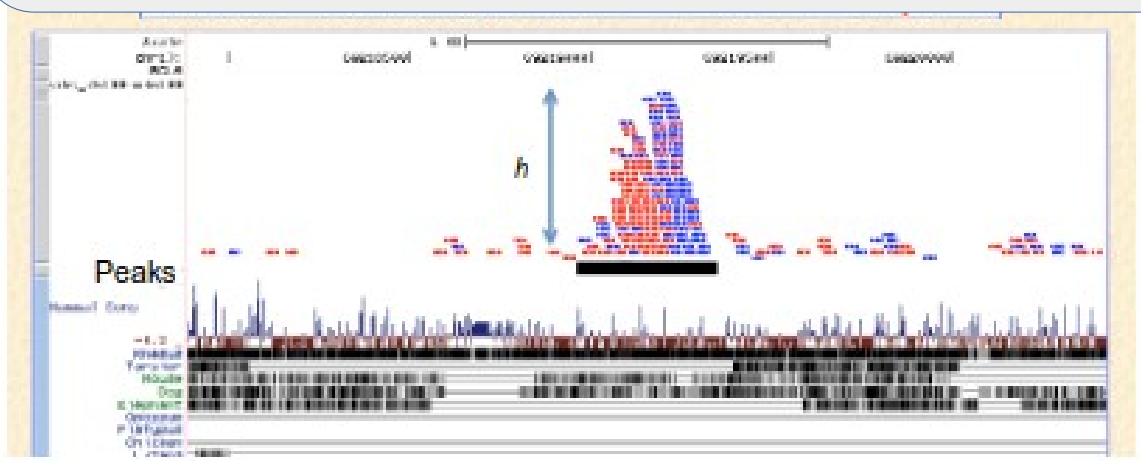
David S. Johnson,[1]* Ali Mortazavi,[1]* Richard M. Myers,[1]† Barbara Wold[2,3]†

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position (±50 base pairs (bp)), which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity (ROC (receiver operator characteristic) area ≥ 0.96) and statistical confidence ($P < 10^{-6}$), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

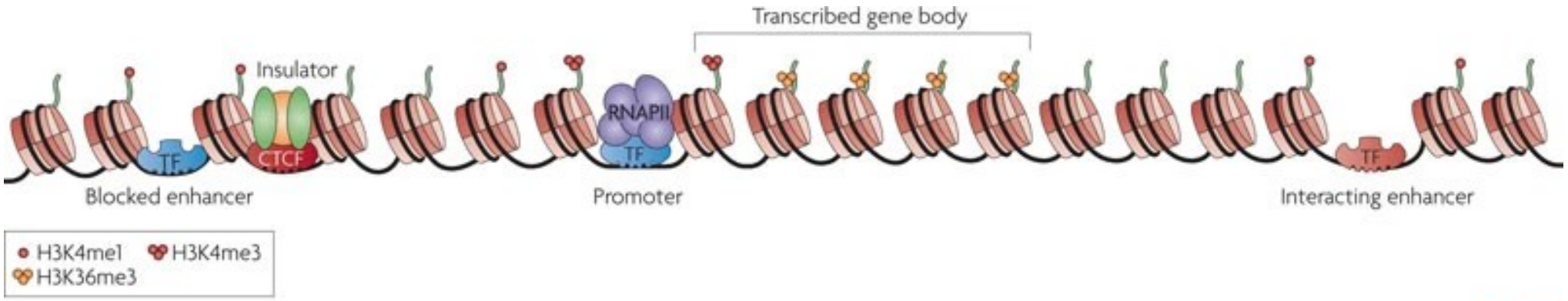
Although much is known about transcription factor binding and action at specific genes, far less is known about the composition and function of entire factor-DNA chromosome can be detected by chromatin immunoprecipitation (ChIP) (*1*). In ChIP experiments, an immune reagent specific for a DNA binding factor is used to enrich target DNA

1. **Chromatin Crosslinking:** Treat cells with formaldehyde to crosslink proteins to DNA.
2. **Chromatin Shearing:** Fragment the chromatin into smaller pieces using sonication or enzymatic digestion.
3. **Immunoprecipitation:** Use antibodies specific to the target protein or histone modification to pull down DNA-protein complexes.
4. **Reverse Crosslinking:** Reverse the crosslinks to free the DNA from the protein.
5. **DNA Purification:** Purify the DNA fragments.
6. **Library Preparation:** Add adapters to the purified DNA fragments to prepare for sequencing.
7. **High-Throughput Sequencing:** Sequence the DNA using a platform like Illumina.





ChIP-seq to study epigenomics



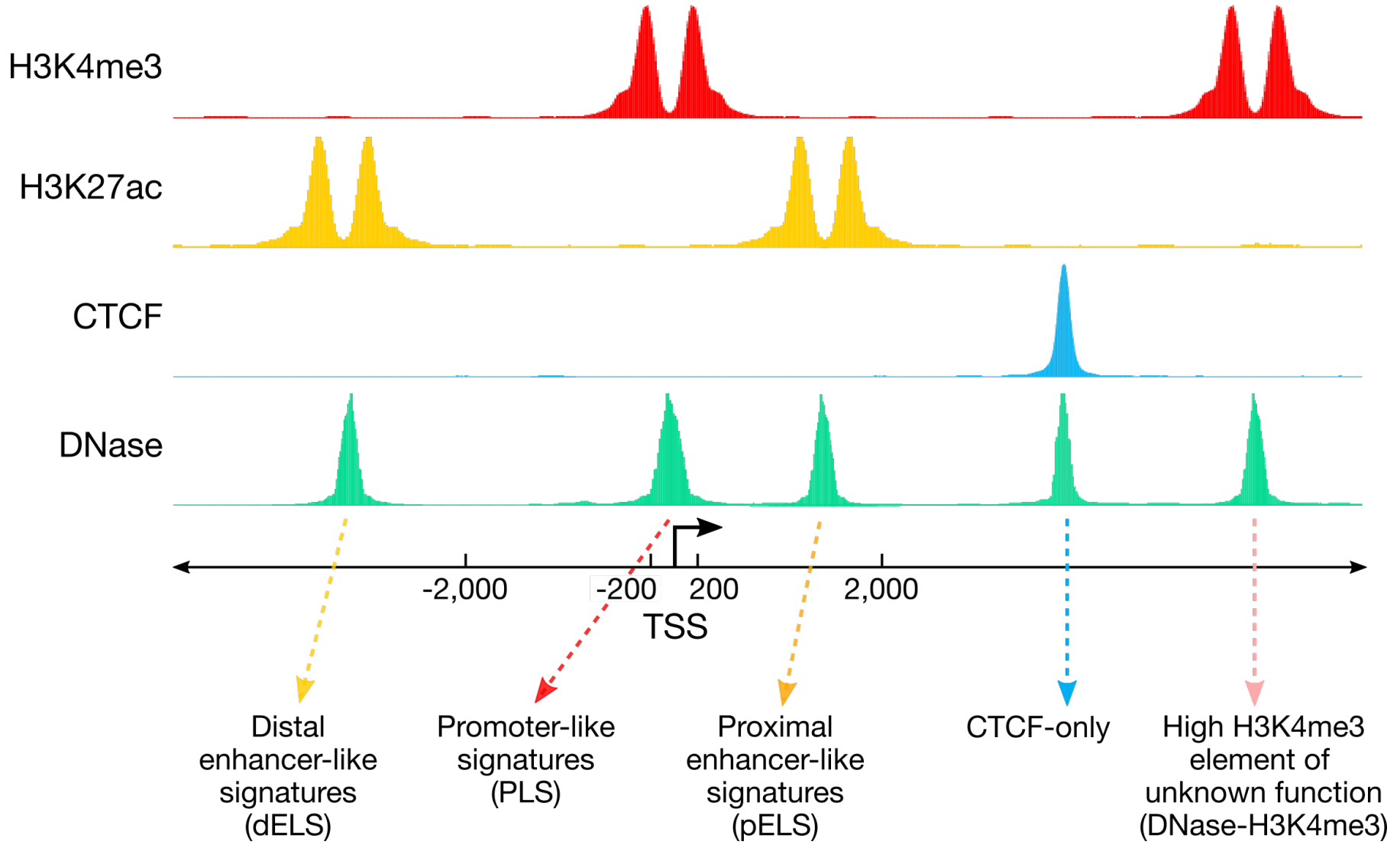
Nature Reviews | Genetics

Source: Nat Rev Genet 11, 476–486 (2010).

Antibodies targeting selected histon marks or proteins
can be used to localize genomic sites enriched in that modification.

For instance:
- Promoters can be mapped by the localization of histone 3 lysine 4 trimethylation (**H3K4me3**).
- Bodies of transcribed genes and non-coding RNAs are marked by **H3K36me3**.
- Enhancers are marked by **H3K4me1**.
- Insulators can be mapped **by the localization of CCCTC-binding factor (CTCF) binding sites**.

ENCyclopedia of DNA Elements



https://www.encodeproject.org/

High-throughput mapping of chromatin interactions (HiC)

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

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[1]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[4,5] Agnes Telling,[4,5] Ido Amit,[1] Bryan R. Lajoie,[1] Peter J. Sabo,[6] Michael O. Dorschner,[6] Richard Sandstrom,[6] Bradley Bernstein,[3,7] M. A. Bender,[10] Mark Groudine,[4,5] Andreas Gnirke,[1] John Stamatoyannopoulos,[6] Leonid A. Mirny,[2,3,11] Eric S. Lander,[1,12,13]† Job Dekker[9]†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase. These maps confirm the presence of chromosome territories and the spatial proximity of small, gene-rich chromosomes. We identified an additional level of genome organization that is characterized by the spatial segregation of open and closed chromatin to form two genome-wide compartments. At the megabase scale, the chromatin conformation is consistent with a fractal globule, a knot-free, polymer conformation that enables maximally dense packing while preserving the ability to easily fold and unfold any genomic locus. The fractal globule is distinct from the more commonly used globular equilibrium model. Our results demonstrate the power of Hi-C to map the dynamic conformations of whole genomes.

The three-dimensional (3D) conformation of

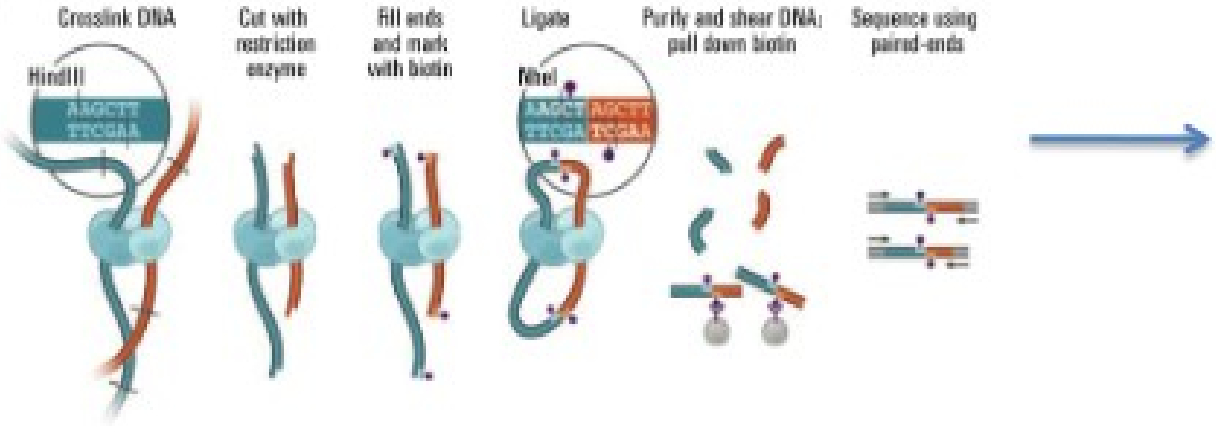
Long-range interactions between specific pairs

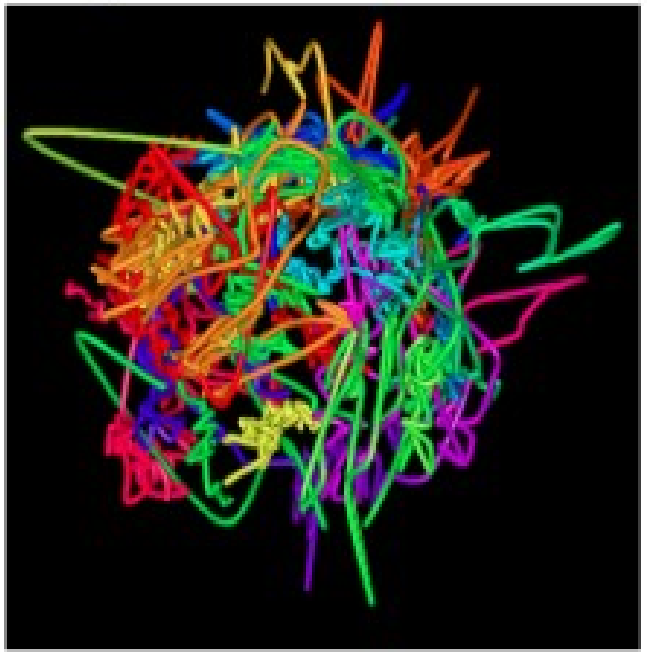
We created a Hi-C library from a karyotypically normal human lymphoblastoid cell line (GM06990) and sequenced it on two lanes of an Illumina Genome Analyzer (Illumina, San Diego, CA), generating 8.4 million read pairs that could be uniquely aligned to the human genome reference sequence; of these, 6.7 million corresponded to long-range contacts between segments >20 kb apart.

We constructed a genome-wide contact matrix M by dividing the genome into 1-Mb regions ("loci") and defining the matrix entry m_{ij} to be the number of ligation products between locus i and locus j (10). This matrix reflects an ensemble average of the interactions present in the original sample of cells; it can be visually represented as a heatmap, with intensity indicating contact frequency (Fig. 1H).

We tested whether Hi-C results were reproducible by repeating the experiment with the same restriction enzyme (HindIII) and with a different one (NcoI). We observed that contact matrices for these new libraries (Fig. 1, C and D) were extremely similar to the original contact matrix [Pearson's $r = 0.990$ (HindIII) and $r = 0.814$

www.sciencemag.org on October 16, 2009





An open source, web-based platform
for data intensive biomedical research

First, let's familiarize ourselves with the Galaxy platform by performing the tutorial:

A short introduction to Galaxy

Next, we will explore (*) the Galaxy tutorial:

An introduction to ChIP-seq analysis with Galaxy

(*) better, we will begin to explore it

References

- [historical value] Waddington CH. **The epigenotype. 1942**. Int J Epidemiol. 2012 Feb;41(1):10-3. doi: 10.1093/ije/dyr184. Epub 2011 Dec 20. PMID: 22186258.

- [review on epigenetics and linked diseases] Portela, A., Esteller, M. **Epigenetic modifications and human disease**. Nat Biotechnol 28, 1057–1068 (2010). https://doi.org/10.1038/nbt.1685

- [recent review on chromatin remodeling] Magaña-Acosta M, Valadez-Graham V. **Chromatin Remodelers in the 3D Nuclear Compartment**. Front Genet. 2020 Nov 3;11:600615. doi: 10.3389/fgene.2020.600615. PMID: 33329746; PMCID: PMC7673392.

- [about data and technologies] Wang KC, Chang HY. **Epigenomics: Technologies and Applications.** Circ Res. 2018 Apr 27;122(9):1191-1199. doi: 10.1161/CIRCRESAHA.118.310998. PMID: 29700067.

- [about Galaxy] Galaxy Community. **The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update.** Nucleic Acids Res. 2024 Jul 5;52(W1):W83-W94. PMID: 38769056.

- [a key project for epigenomics] **ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome.** Nature. 2012 Sep 6;489(7414):57-74.PMID: 22955616

Useful links

- [Galaxy Training material] https://training.galaxyproject.org/training-material/
- [Candidate cis-Regulatory Elements by ENCODE] https://screen.encodeproject.org/
- [Details on the FASTQ format] https://compgenomr.github.io/book/fasta-and-fastq-formats.html
- [How to interpret FastQ Quality scores] https://en.wikipedia.org/wiki/Phred_quality_score
- [Tutorial – DNA methylation data analysis with Galaxy] https://galaxyproject.github.io/training-material/topics/epigenetics/tutorials/methylation-seq/tutorial.html#dna-methylation-data-analysis