An introduction to epigenetics

Master in Omics Data Analysis

Beatrice Borsari
University of Vic, Vic
Computational Biology of RNA Processing, CRG, Barcelona

UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Master in Omics Data Analysis

CRG
Centre for Genomic Regulation

Contact info

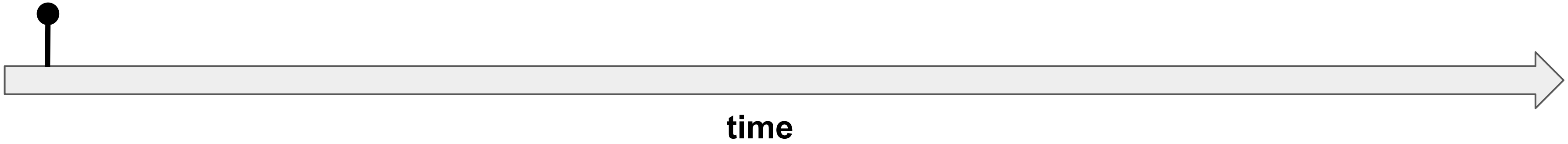
Beatrice Borsari: beatrice.borsari@crg.eu

Diego Garrido Martín: diego.garrido@crg.eu

A brief history of epigenetics

mid XIX century (Darwin, Mendel, Miescher)

- cells derive from other cells
- hereditary information located in the nucleus



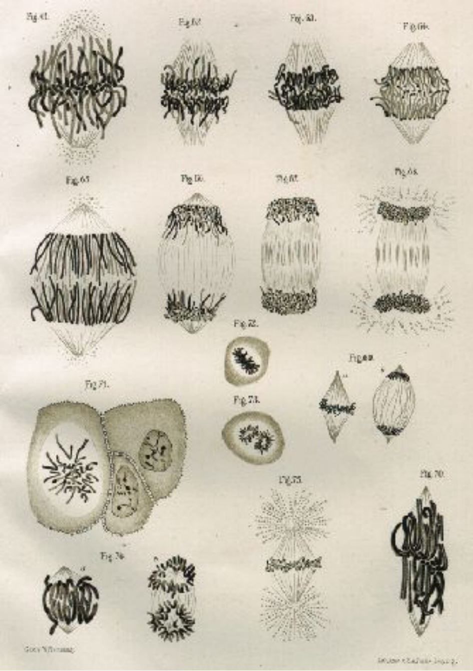
A brief history of epigenetics

chromatin:

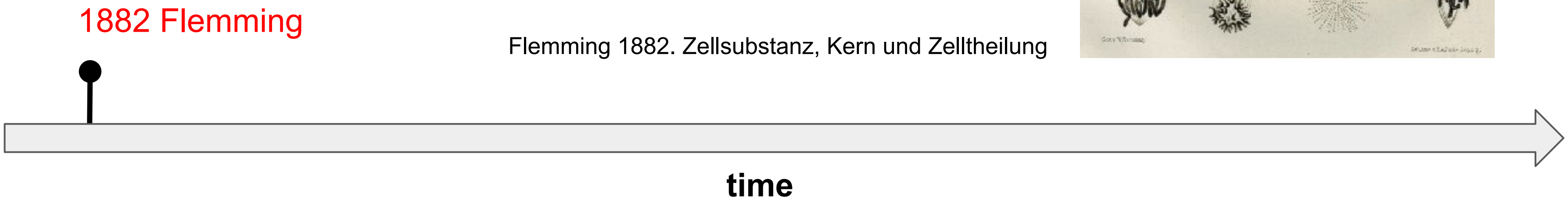
- means "stainable material", coined by Walther Flemming
- is the complex of DNA, histone and non-histone proteins

Flemming

- studies cell division
- discovers chromosomes
- describes chromosome splitting during mitosis



Flemming 1882. Zellsubstanz, Kern und Zelltheilung



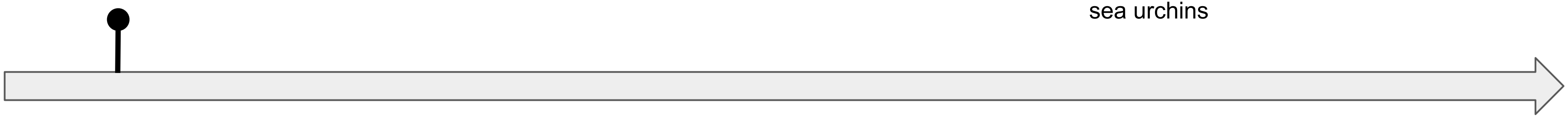
A brief history of epigenetics

- Chromosomes remain organized throughout the process of cell division
- Sperm and egg contribute the same number of chromosome to the zygote

Ascaris megalocephala

sea urchins

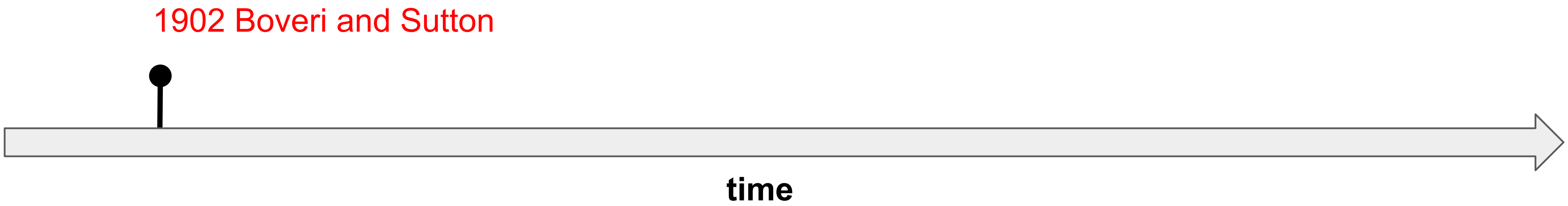
1888 Boveri



time

A brief history of epigenetics

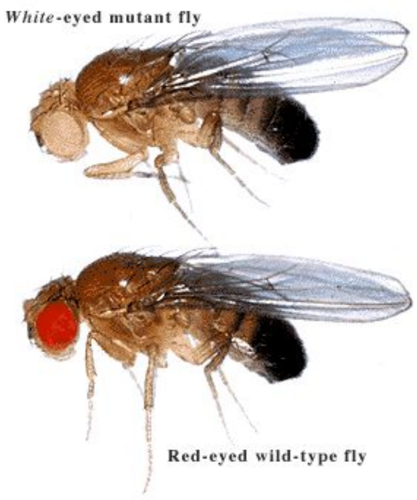
- A specific assortment of chromosomes is responsible for normal development (Boveri)
- During meiosis, the number of chromosomes is reduced by half in gametes, while it is restored in the zygote (Sutton)
- The Boveri-Sutton or chromosome theory: chromosomes are the carriers of genetic material, and thus transmit hereditary characteristics



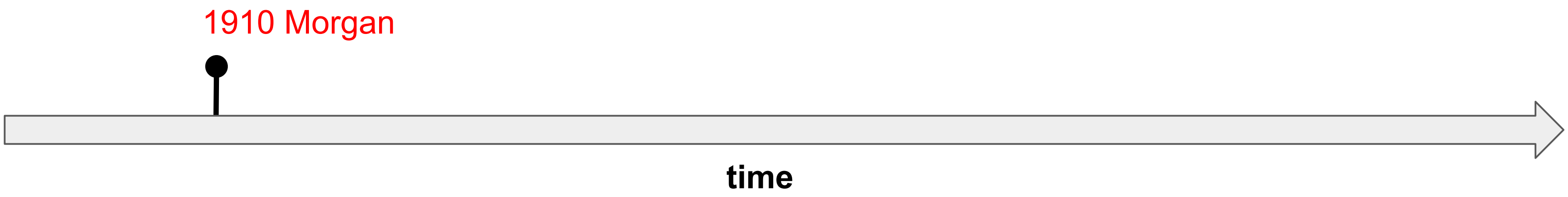
1902 Boveri and Sutton

time

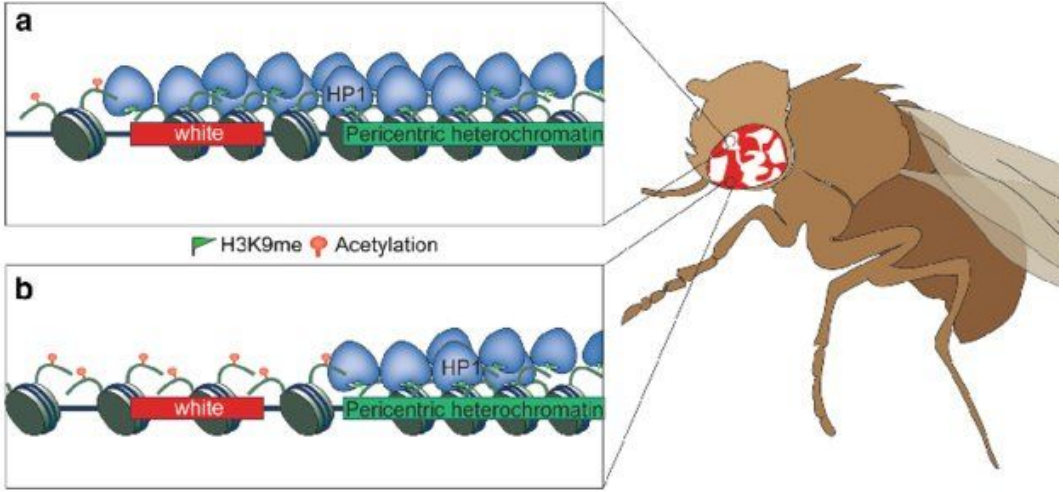
A brief history of epigenetics



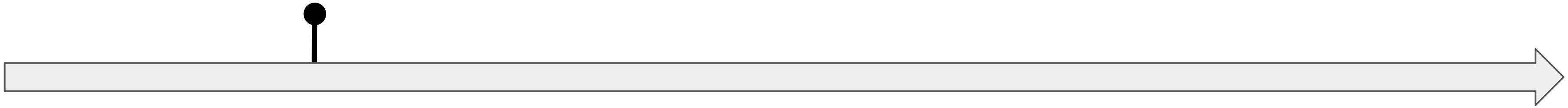
- Morgan discovers an association between eye color and sex in the fruit fly
- He unequivocally demonstrates the chromosome theory, by showing that genes are located on chromosomes
- Chromosomes segregation confirms Mendel's laws of heredity



A brief history of epigenetics



- Muller discovers position-effect variegation (PEV)
- Genes do not function as independent portions of chromosomes
- Context matters! Changes in the location of genes within the genome can have the same impact of changes in the DNA sequence

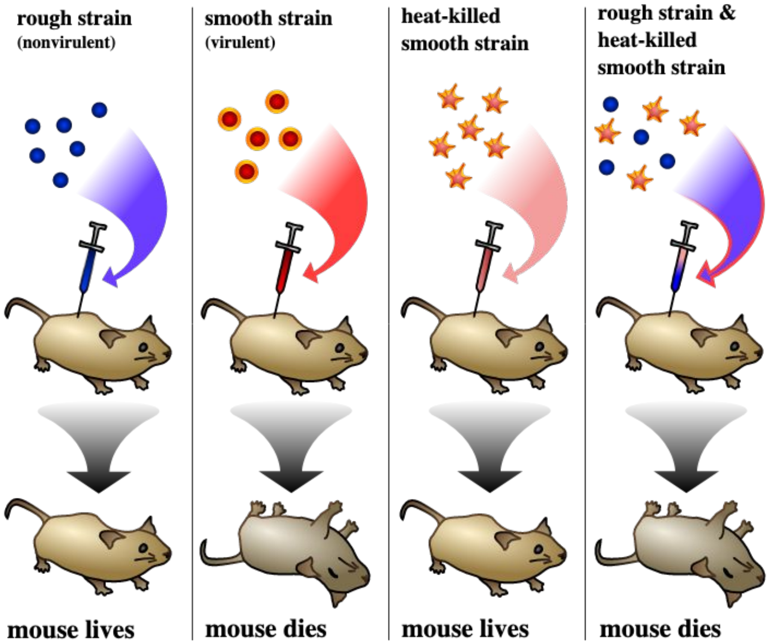


1930 Muller

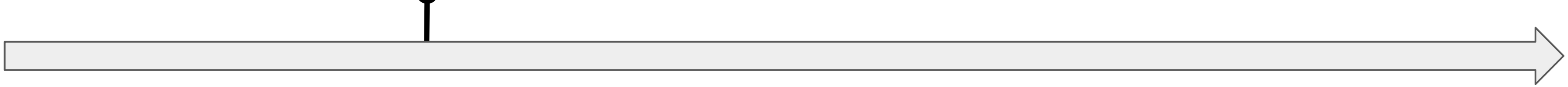
time

A brief history of epigenetics

- They discover that DNA is the transforming principle that accounts for the stable inheritance of specific characteristics



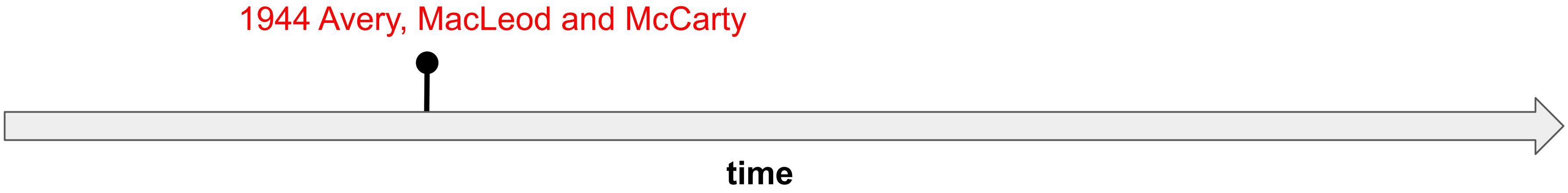
1944 Avery, MacLeod and McCarty



time

A brief history of epigenetics

- They discover that DNA is the transforming principle that accounts for the stable inheritance of specific characteristics
- This discovery relights the debate on developmental processes
- Karyotype studies suggested that all somatic cells present the same set of chromosomes
- ...but is this the same genetic makeup of the zygote?



A brief history of epigenetics

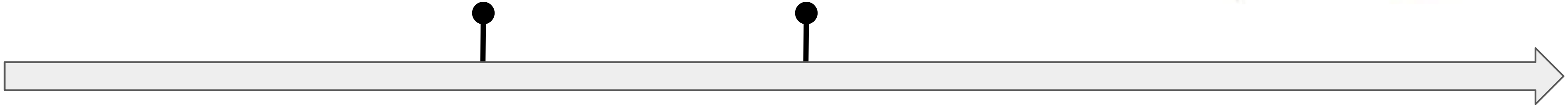
- Briggs and King: transplantation of the nucleus of an embryonic stem cell into an enucleated egg of *R. pipiens* develops into a normal embryo

- Laskey and Gurdon: obtain tadpoles by transplanting nuclei of somatic cells into enucleated cells of *Xenopus*

- The signals that orchestrate development and differentiation do not affect the germline DNA sequence, but lay on top of it (epi-genetics)

1952 Briggs and King

1970 Laskey and Gurdon



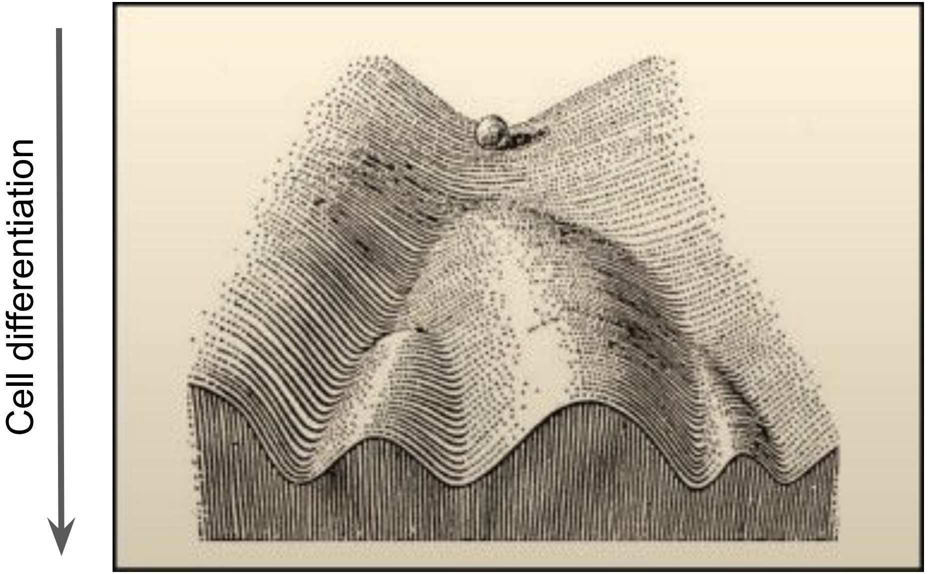
time

A brief history of epigenetics

Conrad Waddington, the father of epigenetics:

- Epigenetics: *the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being* (Waddington, 1942)

- Cellular differentiation is an epigenetic phenomenon, governed mainly by changes in the **epigenetic landscape**, and not in the DNA sequence



Some definitions to start with

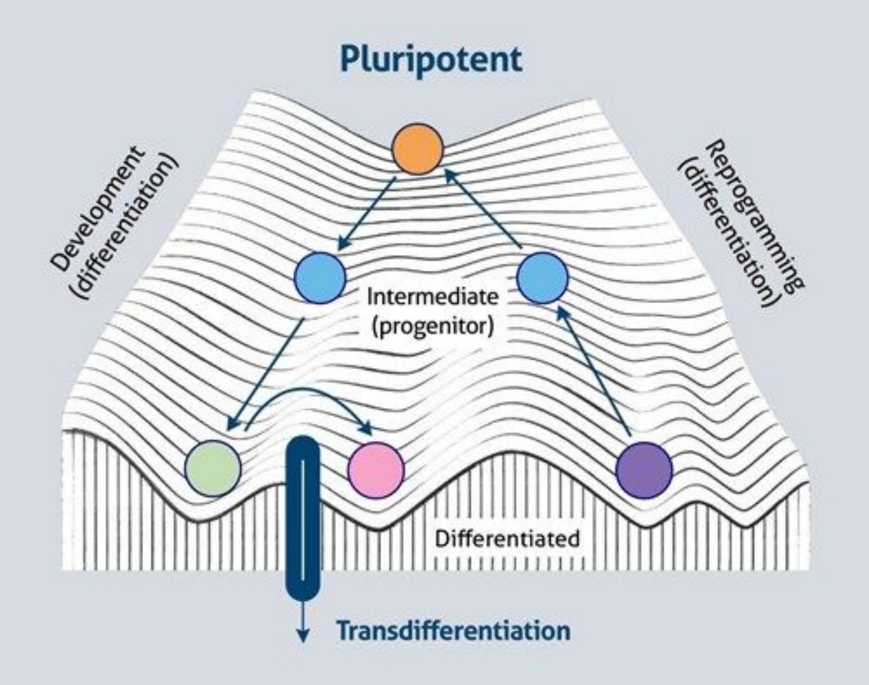
- ***genomics***: the study of the genome, the genetic material of a cell

- ***epigenomics***: the study of the effects of chromatin structure on the genetic material of a cell
    - *epi-* (from ancient Greek επί, “above”): these effects “sit” on top of the DNA, are reversible and do not alter its sequence
    - instead, they *mark* the genome, i.e. provide instructions on when and where to execute specific functions (turning genes on and off, controlling the amount of transcripts produced, etc.)
    - they include
        - DNA methylation
        - packaging of DNA around nucleosomes
        - covalent modifications of histone tails
        - higher order chromatin folding
        - long non-coding RNAs
    - all the cells in our body have the same genetic content; the differences between cells depend on how and when the epigenome turns on and off different sets of genes

Epigenomics Fact Sheet

- **the epigenome can be inherited**
    - when cells divide, often much of the epigenome is passed on to the next generation of cells, helping the cells remain specialized.
    - much of the epigenome is reset when parents pass their genomes to their offspring; still, some of the chemical tags on the DNA and histones of eggs and sperm may be passed on to the next generation.
- **the epigenome can change**
    - most of the epigenetic rearrangements occur during organismal development and cell differentiation.
    - lifestyle and environmental factors (such as smoking, diet and infectious diseases) can lead to changes in the epigenome.
- **epigenomics and disease**
    - in cancer, changes in the epigenome can switch on or off genes involved in cell growth, leading to uncontrolled cell duplication, or in the immune response, causing a failure of the immune system to destroy tumors.

Epigenomics Fact Sheet


Pluripotent
Development (differentiation)
Reprogramming (differentiation)
Intermediate (progenitor)
Differentiated
Transdifferentiation

Both cell differentiation and cell plasticity are thought to be epigenetic phenomena, and can be described as movements / trajectories within Waddington’s epigenetic landscape

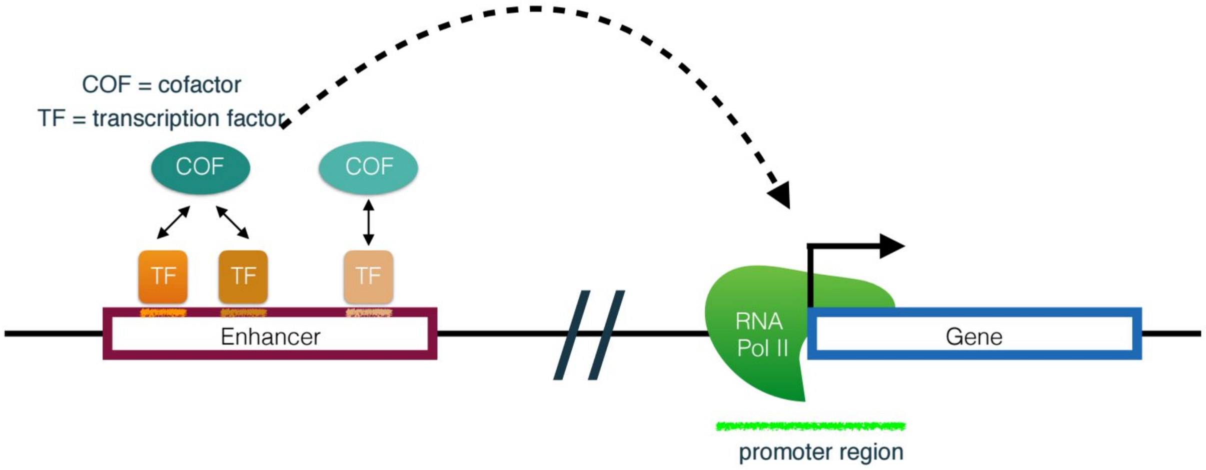
Cell plasticity: the ability of some cells to take on characteristics of other cells in an organism.

Some definitions to start with

- ***genomics***: the study of the genome, the genetic material of a cell

- ***epigenomics***: the study of the effects of chromatin structure on the genetic material of a cell
  - *epi-* (from ancient Greek επί, “above”): these effects “sit” on top of the DNA, are reversible and do not alter its sequence
  - instead, they *mark* the genome, i.e. provide instructions on when and where to execute specific functions (turning genes on and off, controlling the amount of transcripts produced, etc.)
  - they include
    - DNA methylation
    - packaging of DNA around nucleosomes
    - covalent modifications of histone tails
    - higher order chromatin folding
    - long non-coding RNAs
  - all the cells in our body have the same genetic content; the differences between cells depend on how and when the epigenome turns on and off different sets of genes

A schematic view of gene expression

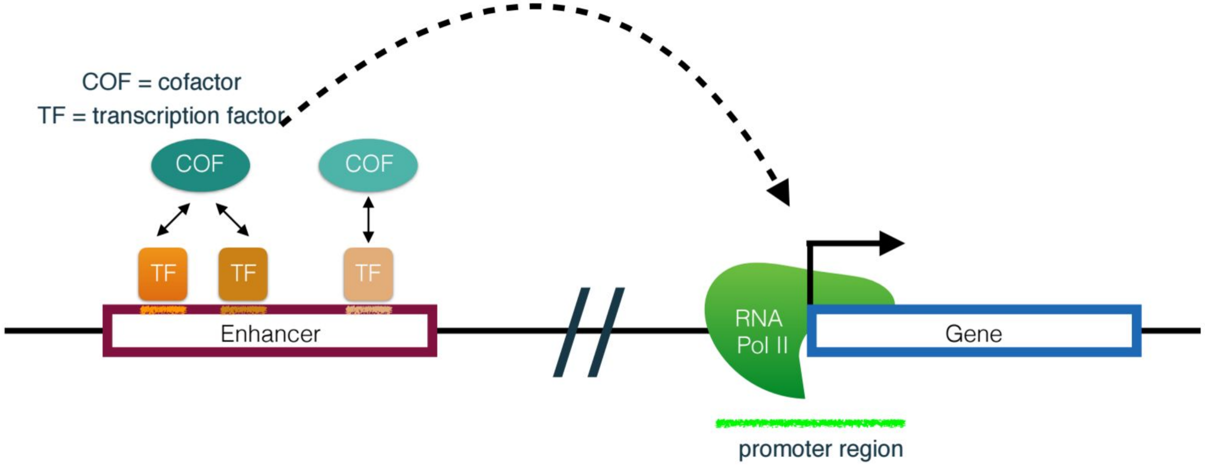


promoter: a DNA sequence typically located directly up-stream or at the 5' end of the transcription start site (TSS)

- bound by RNA polymerase II complex and the TFs required to initiate transcription
- defines the direction of transcription, therefore which DNA strand will be transcribed
- for some eukaryotic genes, promoter regions comprise the TATA box (25-35 bp up-stream of the TSS)

http://workshop.bcdata.ca/2017/project/project-3/

A schematic view of gene expression

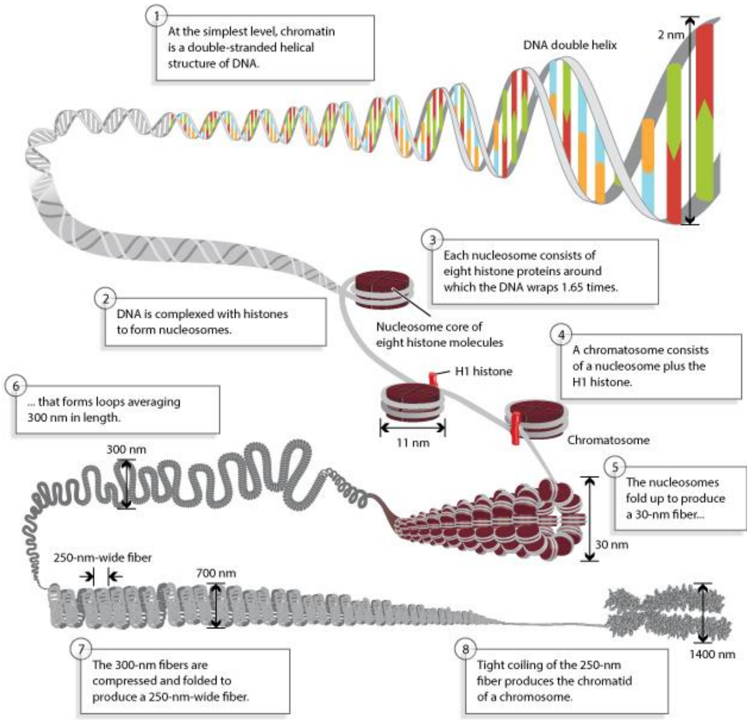


enhancer: a regulatory DNA sequence that, when bound by specific proteins called transcription factors (TFs), increases the likelihood of transcription of an associated gene

- can be proximal or distal to the gene
- one of the main challenges is to associate enhancers to target genes
- enhancer-gene pairs can be context-specific, i.e. can vary across cell lines and tissues

http://workshop.bcdata.ca/2017/project/project-3/

DNA packaging: nucleosomes and chromatin



1 At the simplest level, chromatin is a double-stranded helical structure of DNA.

2 DNA is complexed with histones to form nucleosomes.

3 Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.

4 A chromatosome consists of a nucleosome plus the H1 histone.

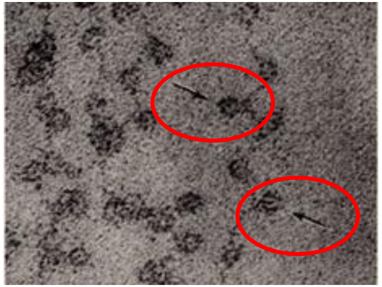
5 The nucleosomes fold up to produce a 30-nm fiber...

6 ... that forms loops averaging 300 nm in length.

7 The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.

8 Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

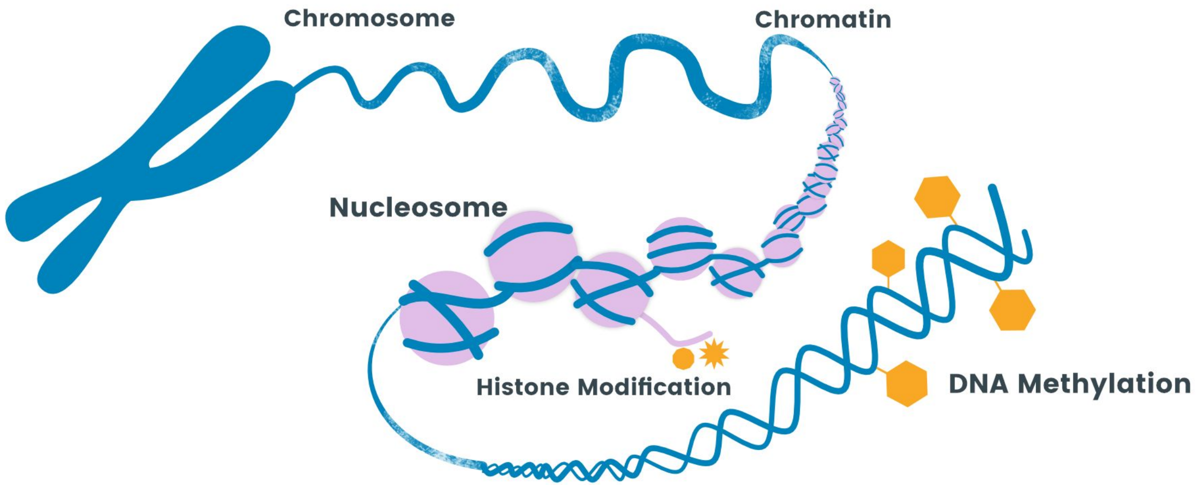
- eukaryotic cells pack their genetic information inside the nucleus
- haploid human genome
  - ~3 billion base pairs of DNA
  - 23 chromosomes
- histones are proteins that provide a scaffold to pack DNA
- chromatin: histone proteins + DNA
- nucleosome: the basic unit of chromatin



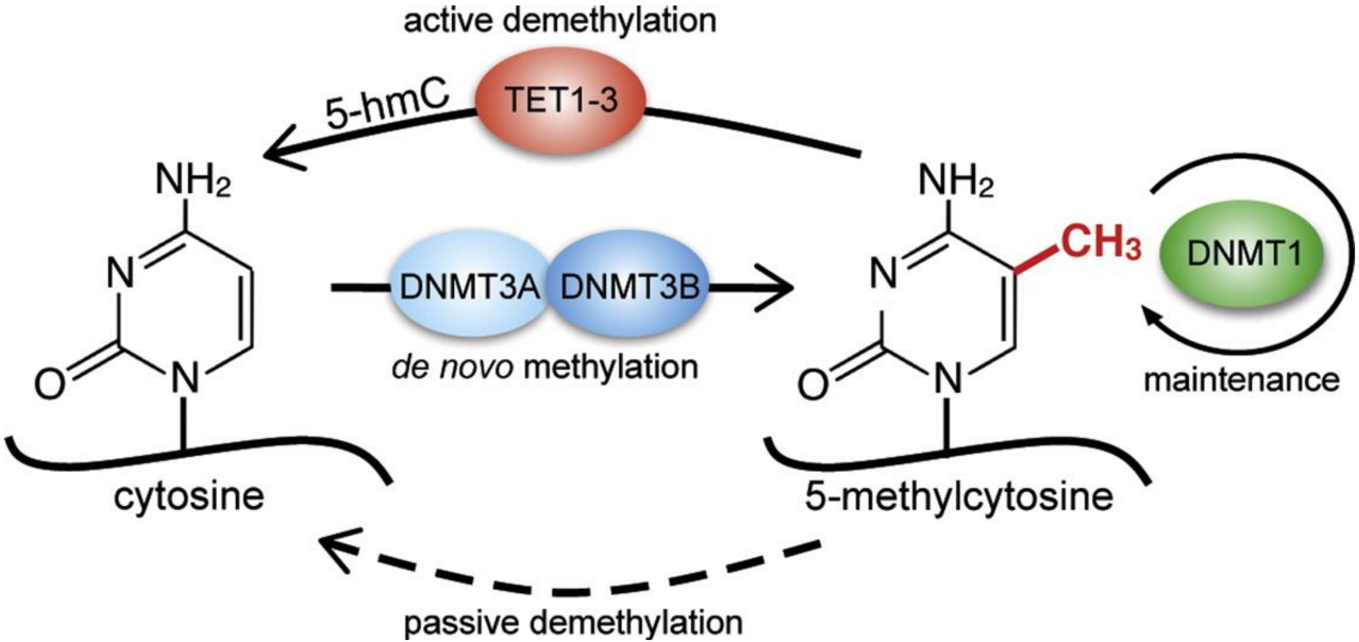
Electron micrograph of chromatin:
the beads on a string

Olins and Olins, 1974
Woodcock et al., 1976

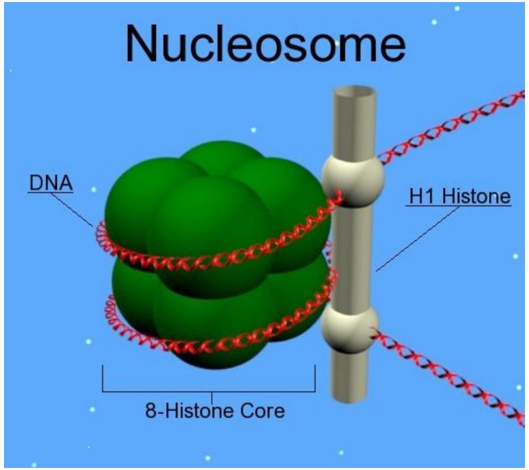
Histone and DNA modifications



DNA methylation

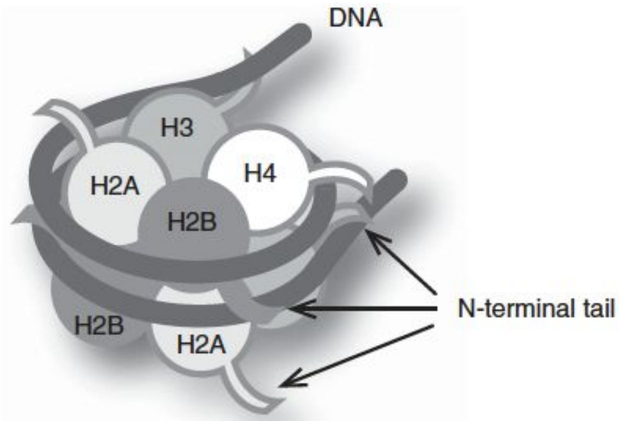


Nucleosome structure



- chromatin is made possible by the electrostatic interactions between histones (positively charged) and DNA (negatively charged)
- nucleosome structure:
  - histone octamer: H2A, H2B, H3 and H4 (x2)
  - ~146 bp of DNA are wrapped around the histone octamer
- *linker DNA*: the fragment of DNA between two consecutive nucleosomes (20-80 bp)
- one copy of histone H1 sits on top of the nucleosome
  - keeps in place the DNA wrapped around the nucleosome
  - binds to the linker DNA
  - helps stabilize the chromatin fiber

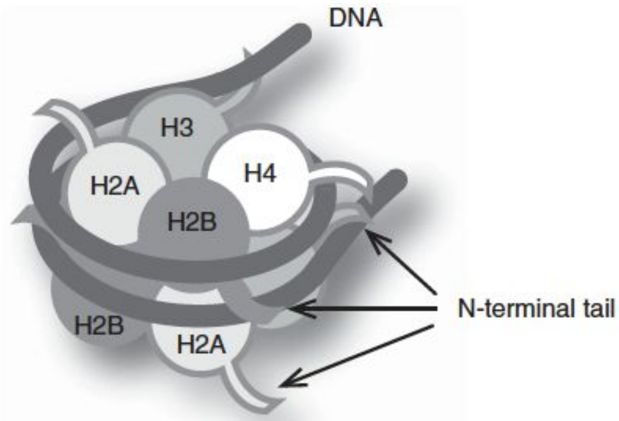
Histone modifications: definition and nomenclature



- the four core histone proteins have N-terminal tails that are extruded from the octamer
- this makes them easily accessible to transcription factors (TFs) and other proteins involved in gene expression regulation
- histone marks are chemical tags deposited on these tails
- different modifications occur on different amino-acid residues:
  - lysine (**K**): acetylation (**ac**), methylation (**me**), ubiquitination
  - arginine (**R**): methylation and citrullination
  - serine (**S**), threonine (**T**), tyrosine (**Y**): phosphorylation
- lysine acetylation and methylation on **histone 3** are the most studied modifications in the field of epigenetic regulation
- methylation can occur at three levels: mono- (**1**), di- (**2**), tri- (**3**)

Kimura 2013

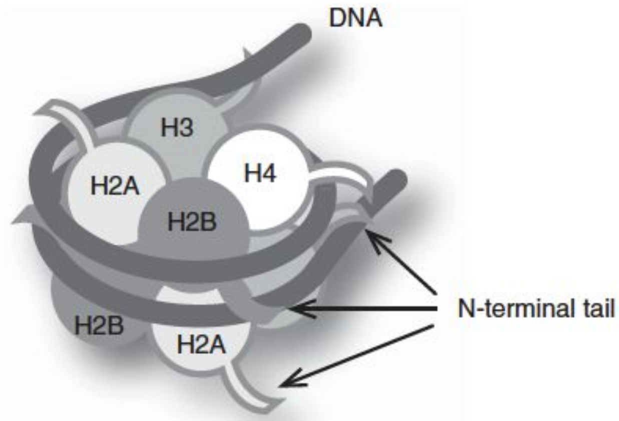
Histone modifications: definition and nomenclature



- acetylation is generally correlated with active gene expression, independently of the aa residue
- in the case of methylation, the correlation with gene expression depends on the degree of methylation (i.e. number of methyl groups) and the aa residue
    - H3K4me3 ⟶ active expression
    - H3K9me3 ⟶ repressed expression

Kimura 2013

Histone modifications: definition and nomenclature

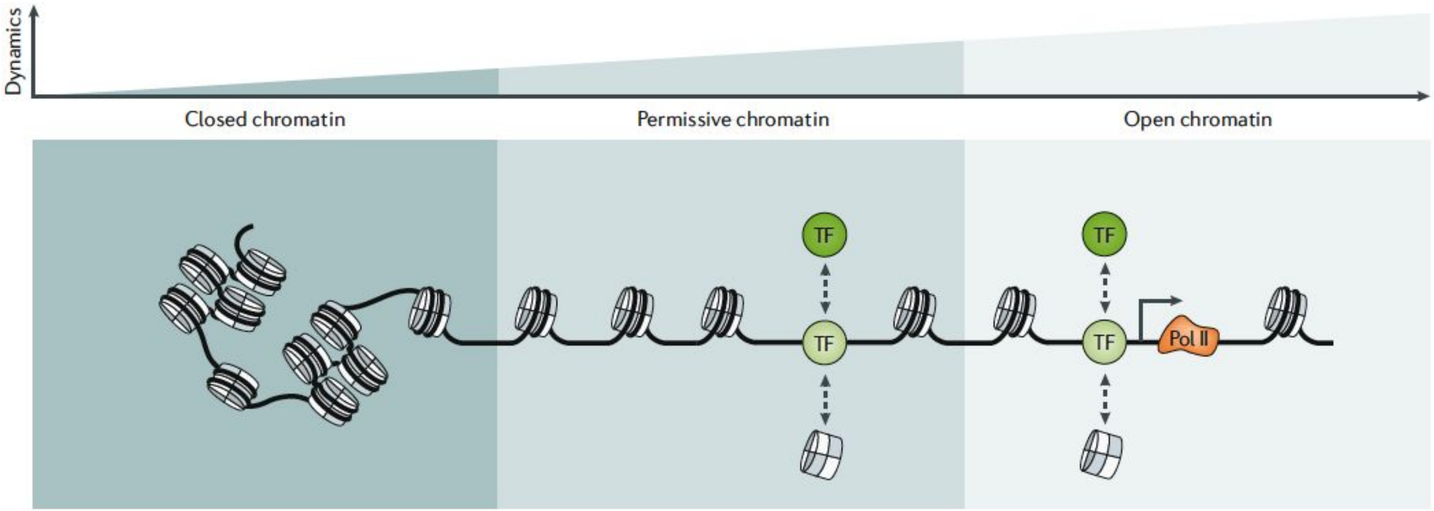


Kimura 2013

How to read the code for a histone mark:
H3K4me3: *tri-methylation of lysine 4 on histone 3*

H3	histone carrying the modification
K	the name of the aa residue
4	the number of the aa residue
me	the type of modification
3	in the case of methylation, the number of methyl groups

Heterochromatin and euchromatin



heterochromatin

1. tightly packed, high nucleosome occupancy
2. low nucleosome turnover
3. little or no transcriptional activity
4. constitutive vs. facultative

euchromatin

1. loosely packed, low nucleosome occupancy
2. high nucleosome turnover
3. can be bound by chromatin-binding factors
4. can be transcriptionally active

Klemm et al., 2019

Constitutive vs. facultative heterochromatin

- A major function of heterochromatin is to protect DNA from being used as a substrate for transcription or for other DNA-based transactions, such as repair.
- Facultative heterochromatin contains genes that must be kept silent under specific time and space conditions
  - varies between cell types
  - can be present in different chromosomal regions
- Constitutive heterochromatin typically occurs at the same genomic regions in every cell type ⟶ more static
  - mainly found at pericentromeric, telomeric, and ribosomal regions
  - pericentromeric and telomeric regions are typically gene poor and contain many repetitive regions

Saksouk et al., 2015

International Human Epigenome Consortium

IHEC

International Human Epigenome Consortium

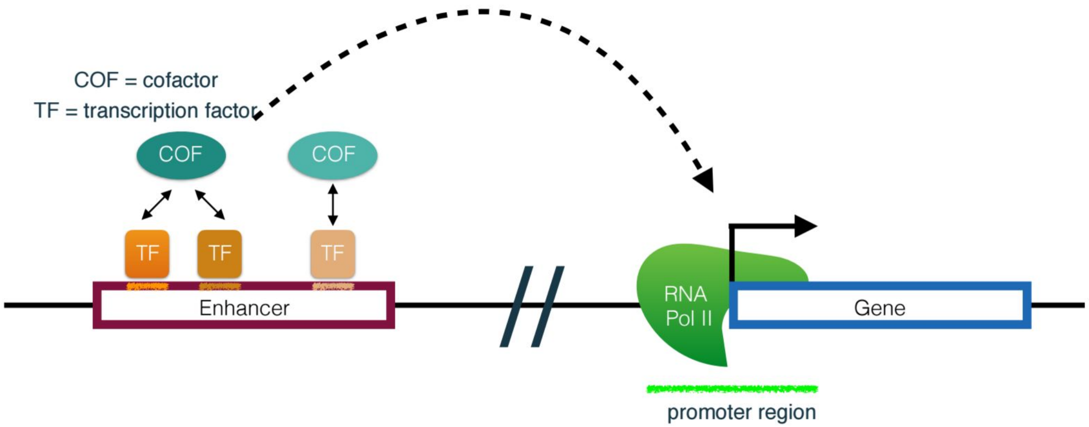
http://ihec-epigenomes.org/

- aims to produce high-resolution reference human epigenome maps in health and diseases
- helps disseminate knowledge and standards regarding new technologies and software in epigenomics
- recommends the profiling of six histone modifications:
  - H3K4me1
  - H3K4me3
  - H3K36me3
  - H3K27me3
  - H3K9me3
  - H3K27ac

lysine methylation

lysine acetylation

A schematic view of gene expression

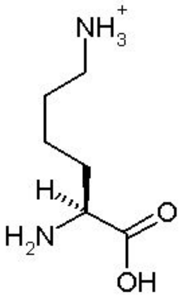


Different steps in the process of gene expression are under epigenetic regulation:

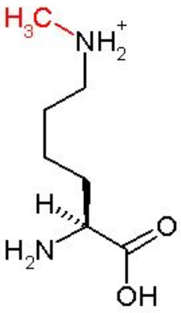
- binding of TFs and cofactors to enhancer regions
- interaction between enhancer and promoter regions
- binding of RNA Pol II complex to promoters
- transcription elongation, termination and polyadenylation
- splicing

http://workshop.bcdata.ca/2017/project/project-3/

Lysine methylation

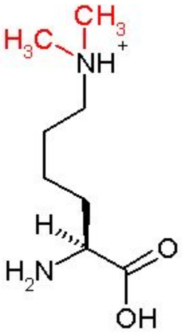


Lysine



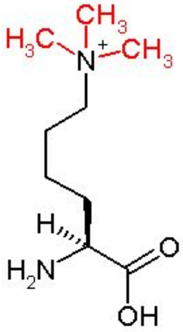
Mono-methyl lysine

H3K4me1



Dimethyl lysine

H3K4me2



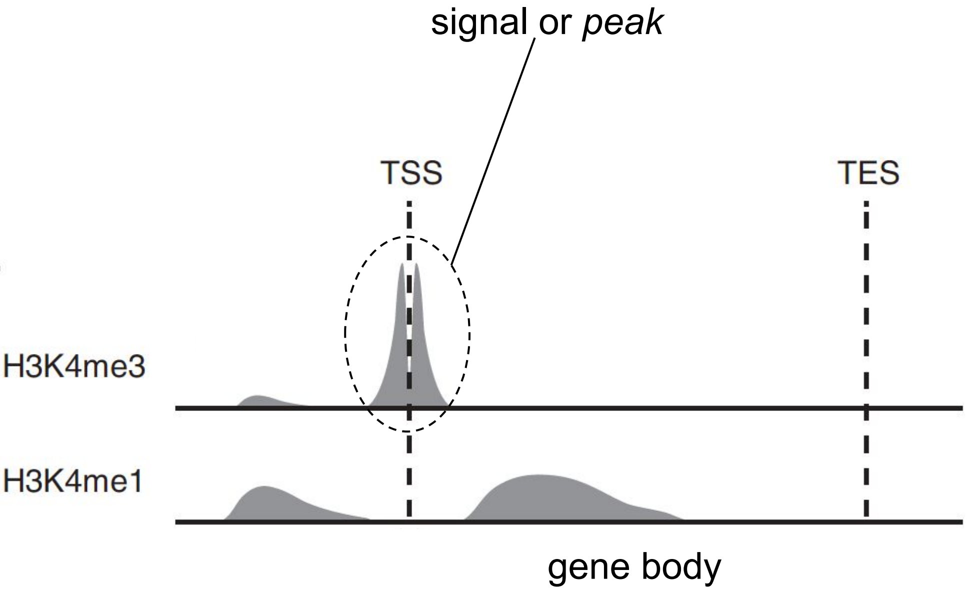
Trimethyl lysine

H3K4me3

Kimura 2013

- methyl-groups are deposited by enzymes called *methyltransferases* that contain the SET protein domain
  - in the case of H3K4, these enzymes belong to the Set1/COMPASS complex
    - includes MLL1-4 (mixed-lineage leukemia 1-4) genes
    - rearrangements of the MLL1 gene by translocations are associated with aggressive acute leukemias
- methyl-groups are removed by enzymes called *demethylases*
  - in the case of H3K4, these enzymes belong to KDM and Jumonji/JARID1 families
    - dysregulation of KDM5 genes are observed in cancer

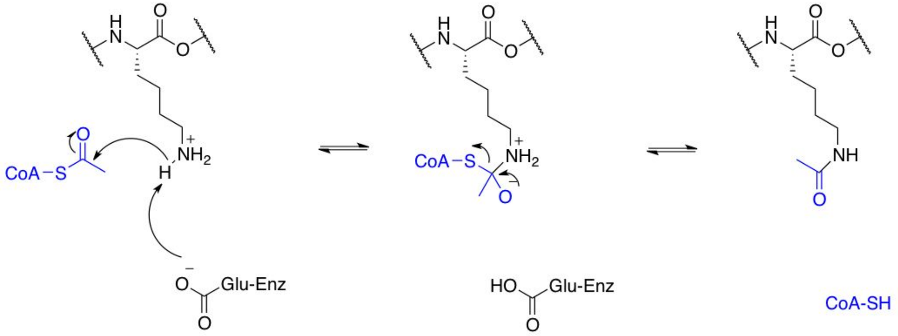
H3K4me1 & H3K4me3



Kimura 2013

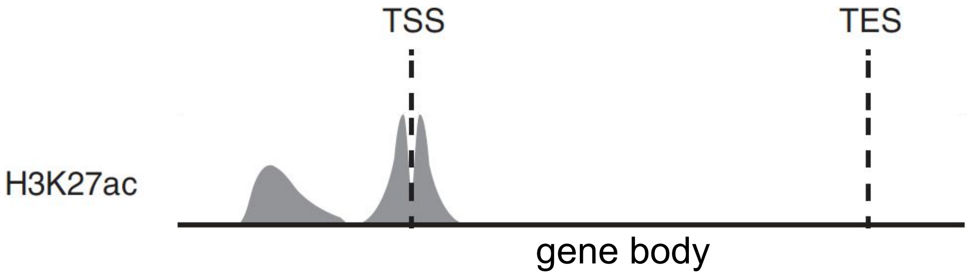
- H3K4me3 is enriched around the transcription start site (TSS) and is typically characterized by a narrow peak
- H3K4me1 is enriched both up-stream and down-stream of the TSS, and is typically characterized by a broader peak
- unmethylated H3K4 and DNA methylation are associated to form *silent* chromatin
    - unmethylated H3K4 is recognized by DNA methyltransferase 3L
- methylated H3K4 is associated with *active transcription*
    - H3K4me3 interacts with TAF3 (component of TATA-binding protein factor)
    - interaction enhanced by K9/14ac (other active marks)

Lysine acetylation



- acetyl-groups are deposited by enzymes called *histone acetyl-transferases* (HATs)
  - type *A* HATs are nuclear and comprise 3 main families (GNAT, MYST and CBP/p300)
  - type *B* HATs are cytoplasmic and target newly synthetized histones H4 and H3
- acetyl-groups are removed by enzymes called *histone deacetylases* (HDACs)
  - 4 sub-families
  - lack of specificity for target substrate: they can target multiple aa sites

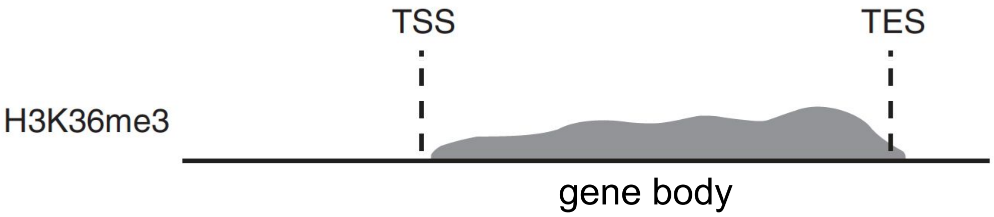
H3K27ac



- H3K27ac is typically enriched around the TSS, with a peak very similar to H3K4me3
- it can be found also up-stream of the TSS, forming a broader peak
- co-occupancy of H3K4me1 and H3K27ac is considered a good indicator of active enhancers

Kimura 2013

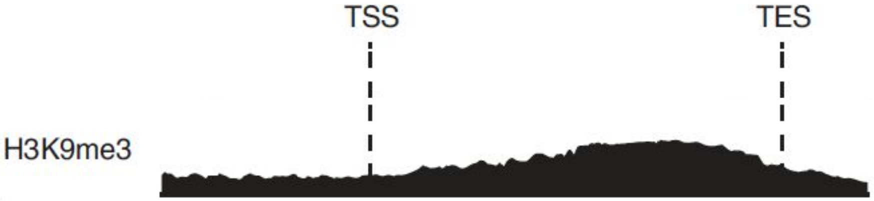
H3K36me3



Kimura 2013

- deposited by methyltransferase SETD2 and removed by demethylase JHDM3
- prevents abortive initiation of transcription within the gene body and controls transcription elongation
    - as Pol II moves over the gene body, newly incorporated nucleosomes acquire H3K36me3 and H3K9me3, and lose H3K4me
- can occasionally mark promoters: prevents the accumulation of another mark (H3K27me3)
- involved in DNA damage repair, especially via homologous recombination
- has been associated with splicing events (prevents exon inclusion)

H3K9me3

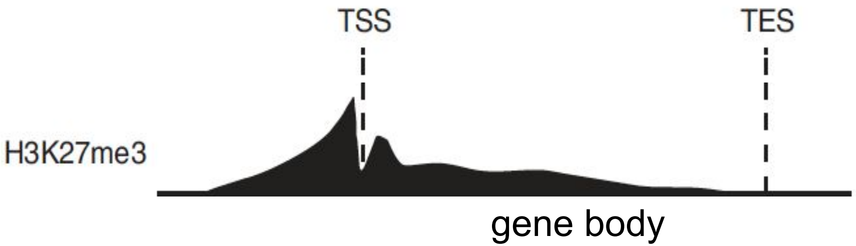


gene body

- associated with higher nucleosome occupancy
- indicator of silenced transcription (broad peak towards the 3’ end of the gene)
- indicator of constitutive heterochromatin
  - enriched in pericentromeric regions, major satellite repeats, centromeric alpha-satellite repeats
  - maintains transposon repression in pericentromeric heterochromatin
- deposited by methyltransferases SUV39H1 and SUV39H2; removed by JHDM3 demethylases
- cross-talk between H3K9me and DNA methylation during DNA replication
- occasionally associated with active gene expression and splicing (enhances exon inclusion)

Kimura 2013

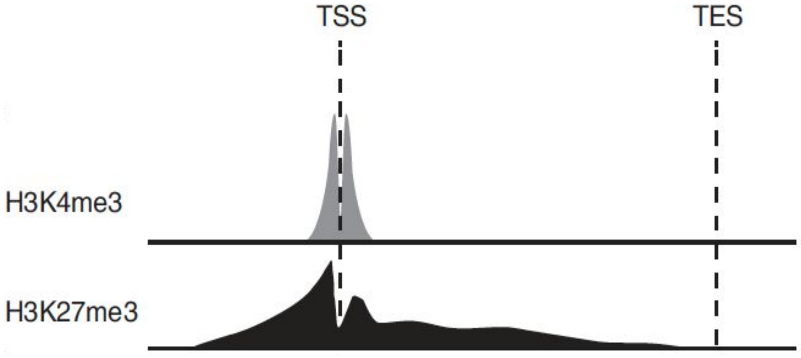
H3K27me3



Kimura 2013

- a hallmark of transcriptional repression (facultative heterochromatin with low DNA methylation)
  - e.g. the inactivated X chromosome in mammalian cells is enriched in H3K27me3
- deposited by polycomb repressive complex 2 (PRC2) mainly at promoter regions
  - binding of PRC2 to H3K27me3 spreads H3K27me to neighbouring nucleosomes → positive feedback mechanism in maintaining a repressed gene expression status
- removed by UTX/KDM6A, UTY/KDM6C and JMJD3/KDM6B demethylases
- H3K9me3 and H3K27me3 are mutually exclusive and do not exist in the same loci

Bivalent promoters: H3K4me3 & H3K27me3

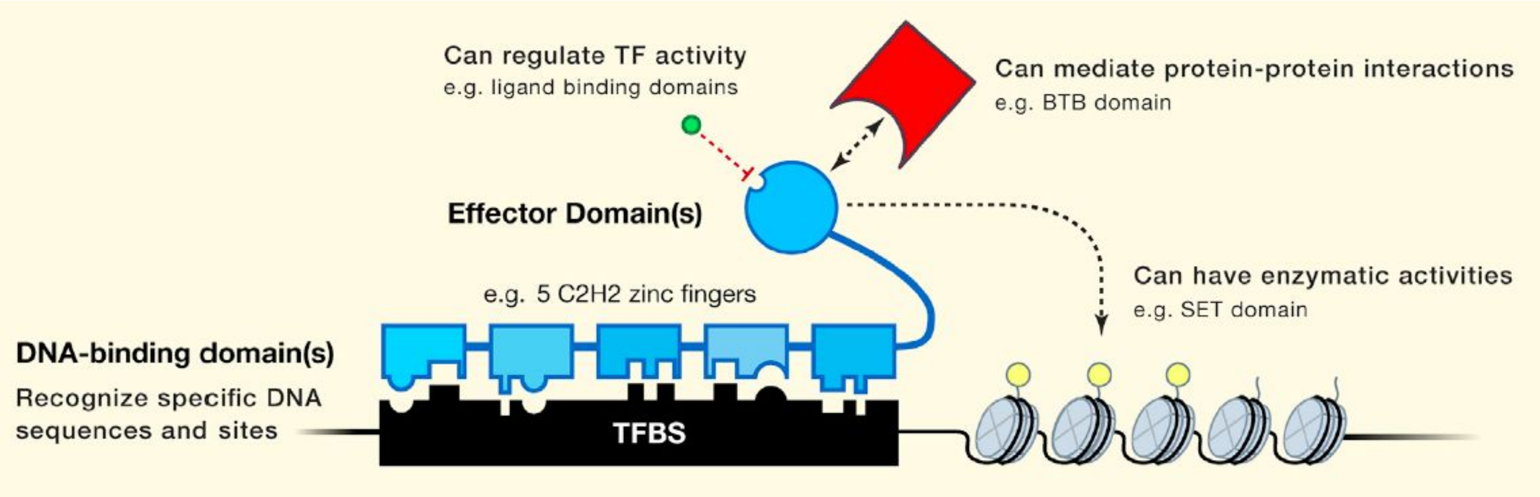


- bivalent or poised chromatin is characterized by the simultaneous presence of histone modifications associated with both gene activation and repression
- in embryonic stem cells (ESCs), promoters of lineage-specific regulatory genes present peaks of both H3K4me3 and H3K27me3
- during differentiation, poised chromatin is progressively lost
  - either H3K4me3 or H3K27me3 is removed to establish repressed or activated genes

Kimura 2013

Transcription Factors (TFs)

- Any protein involved in transcription and/or capable of altering gene expression levels
- More specifically, a protein capable of
  - binding DNA in a sequence-specific manner
  - regulating transcription
- Since they can occlude the binding of other TFs, often the first condition implies the second


Can regulate TF activity
e.g. ligand binding domains
Can mediate protein-protein interactions
e.g. BTB domain
Effector Domain(s)
Can have enzymatic activities
e.g. SET domain
e.g. 5 C2H2 zinc fingers
DNA-binding domain(s)
Recognize specific DNA sequences and sites
TFBS

Lambert et al., 2018

Transcription Factors (TFs)

- The main TF families in eukaryotes include: i) C2H2-zinc finger (ZF), ii) Homeodomain, iii) basic helix-loop-helix (bHLH), iv) basic leucine zipper (bZIP), v) nuclear hormone receptor (NHR)
- There are currently ~1600 known human TFs
- Many function as *master regulators* and *selector genes*. They can control:
    - specific cellular pathways, such as immune response
    - cell differentiation (pluripotent → differentiated cell), de-differentiation (differentiated cell → iPSC), trans-differentiation (cell type A → cell type B)
    - mutations in TFs and TF-binding sites underlie many human diseases
- *Cooperativity* and *sinergy* define the TF activity: most TFs form complexes called **regulatory networks**
- The same TF can regulate different genes in different cell types with opposite effects on gene expression (either positive or negative)
    - e.g. MAX:
        - is a transcriptional activator when binding DNA as a heterodimer with MYC
        - acts as an inhibitor when binding as a heterodimer with MNT or MXD1

Lambert et al., 2018

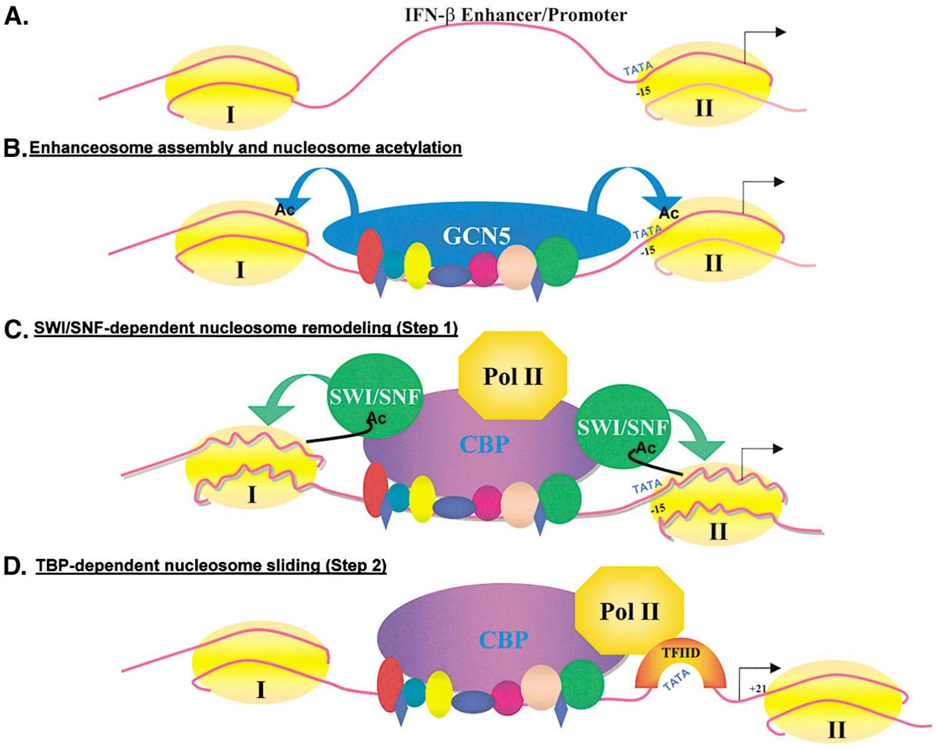
Transcription Factors (TFs)

TFs can exert their functions:

- by directly recruiting the RNA Polymerase Machinery, like the TATA-box binding protein (TBP)
- most commonly by recruiting accessory factors (*cofactors*) that promote specific phases of transcription.
  - cofactors are often multi-subunit protein complexes or multi-domain proteins performing different functions (chromatin binding, nucleosome remodelling, histone marking, etc.)
- an example of coactivator recruitment is the cytokine IFNβ enhanceosome activation (Panne 2008)

Lambert et al., 2018

Transcription Factors (TFs)


A.
IFN-β Enhancer/Promoter
TATA
-15
I
II
B. Enhanceosome assembly and nucleosome acetylation
Ac
GCN5
Ac
TATA
-15
I
II
C. SWI/SNF-dependent nucleosome remodeling (Step 1)
Pol II
SWI/SNF
Ac
CBP
SWI/SNF
Ac
TATA
-15
I
II
D. TBP-dependent nucleosome sliding (Step 2)
Pol II
CBP
TFIID
TATA
+21
I
II

The IFN-β enhancer/promoter is flanked by two nucleosomes (I and II). Nucleosome II masks the TATA box and the start site of transcription.

Virus infection induces enhanceosome assembly and recruitment of the GCN5 complex that acetylates both nucleosomes.

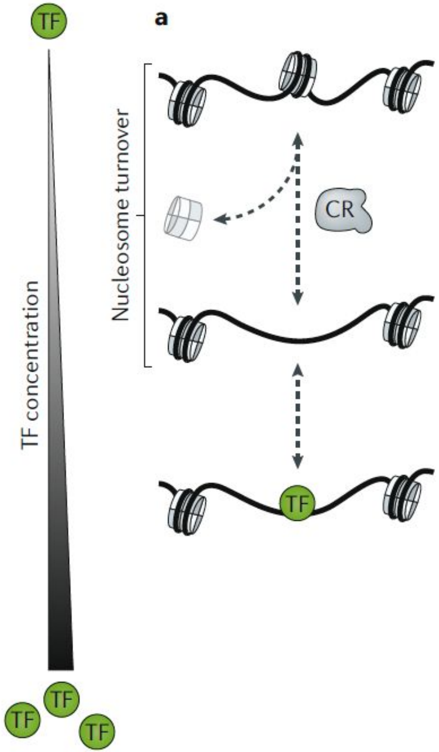
The enhanceosome recruits the CBP/PolII holoenzyme complex and SWI/SNF whose recruitment is stabilized by the acetylated nucleosome. SWI/SNF acts on the nucleosome by modifying the histone-DNA contacts (DNA shown as ruffled lines).

Nucleosome modification by SWI/SNF allows recruitment of TFIID to the promoter by the enhanceosome. The radical DNA bend induced by TBP binding promotes nucleosome sliding, thus fully exposing the core promoter and allowing initiation of transcription.

Lomvardas and Thanos, 2001

Chromatin accessibility

- is the extent to which macromolecules inside the nucleus can physically contact chromatinized DNA
- is non-uniform across the genome
- the accessible genome comprises
  - ~2-3% of total DNA sequence
  - ~90% of regions bound by TFs
- depends on the distribution of nucleosomes and other chromatin-binding factors
- can alter transcription factor (TF) binding
- can change in response to external stimuli and developmental cues
- ATP-dependent chromatin remodelling complexes can move, eject or restructure nucleosomes

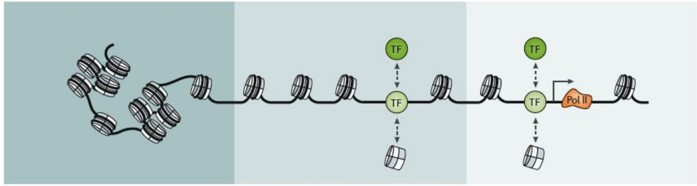


Klemm et al., 2019

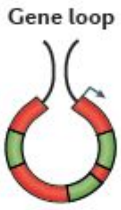
Measuring chromatin accessibility with massive parallel DNA sequencing

- DNase-seq (Crawford et al., 2006; Sabo et al., 2006)
- ATAC-seq (Buenrostro et al., 2013; Corces et al., 2017)
- MNase-seq (Mieczkowski et al., 2016; Mueller et al., 2017)
- NOMe-se (Krebs et al., 2017; Kelly et al., 2012)

Hierarchical folding of chromatin

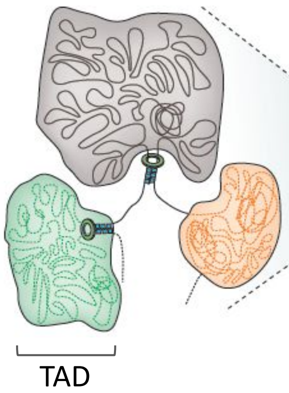


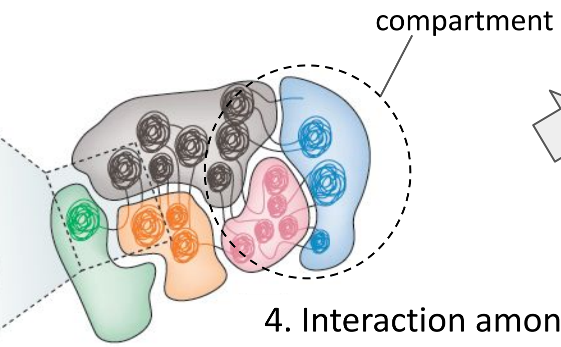
1. For a long time, nucleosomes were thought to form arrays with either solenoid or zig-zag shapes



2. However, they do not form a rigid structure, but are rather involved in long-distance loops

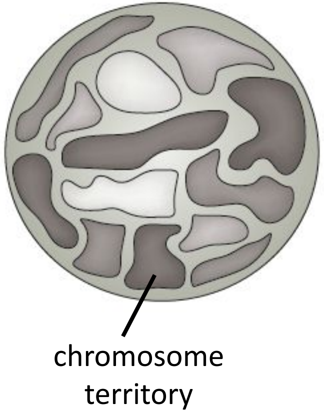
3. Some of these loops give rise to sub-megabase domains, called “topologically associated domains” (TADs)





4. Interaction among TADs of the same epigenomics type forms compartments

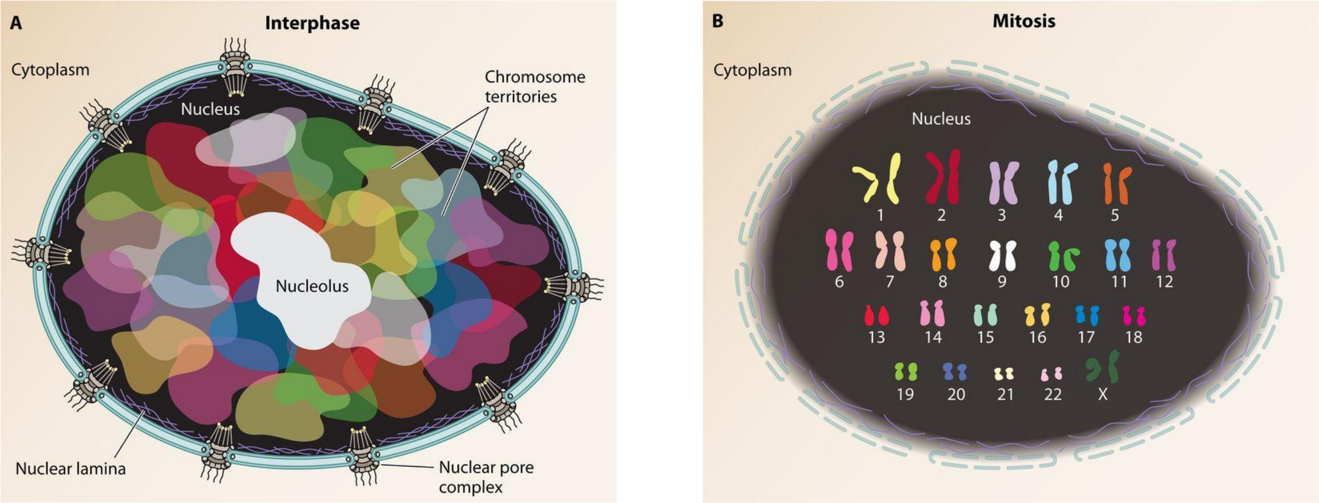
5. Compartments of the same chromosome form a chromosome territory



Bonev and Cavalli, 2016

The genetic material inside a cell: interphase vs. mitosis chromatin

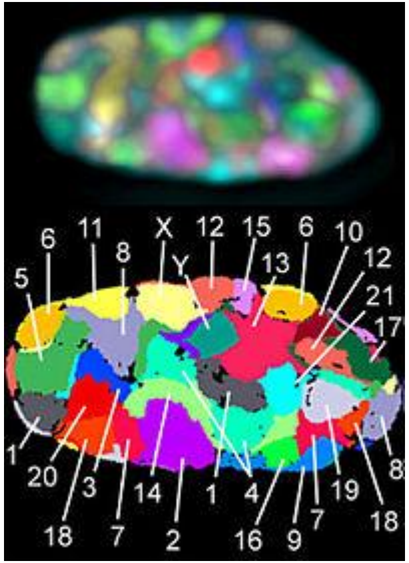
- During interphase, individual chromosomes are not visible, and the chromatin appears diffuse and unorganized



Fraser et al., 2015

Studying 3D genome organization

chromosome territories during prometaphase (human fibroblasts)

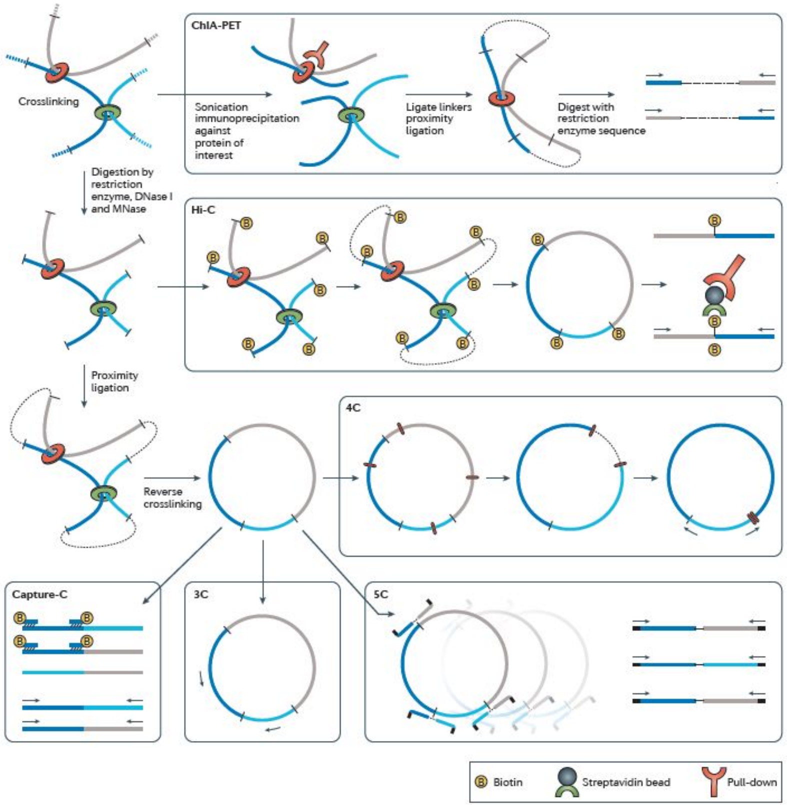


Bolzer et al., 2005

Microscopy-based techniques to visualize the genome in 3D

- include FISH, STORM and PALM techniques
- can provide information about the dynamics of single chromosome domains or generic chromatin
- still limited to a small number of genetic loci
- do not allow the analysis of the whole genome architecture

Studying 3D genome organization



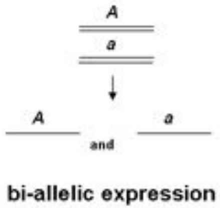
ChIA-PET
Crosslinking
Sonication immunoprecipitation against protein of interest
Ligate linkers proximity ligation
Digest with restriction enzyme sequence
Digestion by restriction enzyme, DNase I and MNase
Hi-C
Proximity ligation
4C
Reverse crosslinking
Capture-C
3C
5C
Biotin
Streptavidin bead
Pull-down

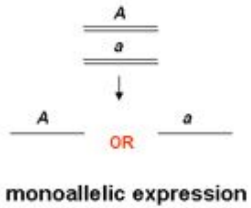
Chromosome **C**onformation **C**apture (**3C**)-based approaches

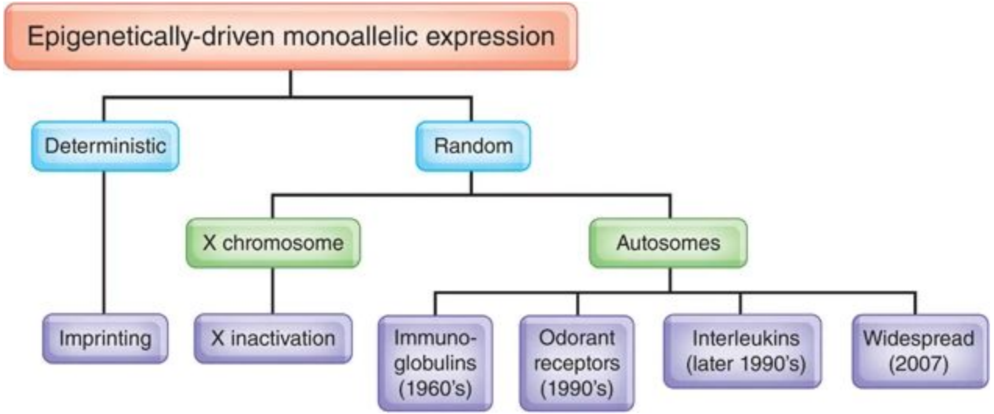
- include ChIA-PET, Hi-C, Capture-C, 3C, 4C, 5C techniques
- are genome-wide approaches
- detect DNA fragments that preferentially interact together on the basis of their proximity in the 3D
- are bulk methods: the results may be the superimposition of individual genome conformations in a population of cells

Bonev and Cavalli, 2016

Monoallelic expression

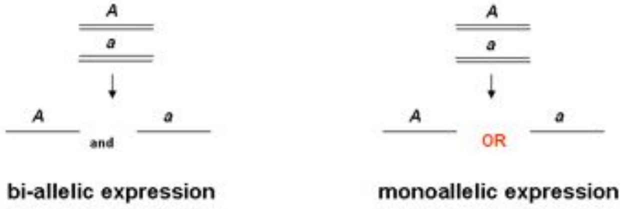


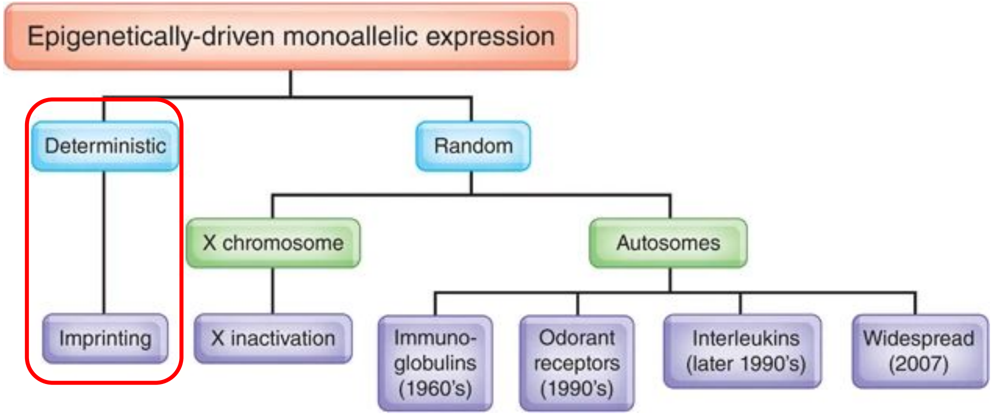



Epigenetically-driven monoallelic expression
Deterministic
Random
X chromosome
Autosomes
Imprinting
X inactivation
Immuno-globulins (1960's)
Odorant receptors (1990's)
Interleukins (later 1990's)
Widespread (2007)

Singer-Sam, J. (2010)

Genomic imprinting in mammals




Epigenetically-driven monoallelic expression
Deterministic
Random
X chromosome
Autosomes
Imprinting
X inactivation
Immuno-globulins (1960's)
Odorant receptors (1990's)
Interleukins (later 1990's)
Widespread (2007)

- Discovered in the early 1980s by embryological studies in mice

- Mouse *Igfr2*: 1st imprinted gene to be discovered

- Genomic imprinting is required for normal development, fetal growth, metabolism, behavior etc.

Genomic imprinting in mammals

- Disturbance in expression of imprinted genes can lead to diseases

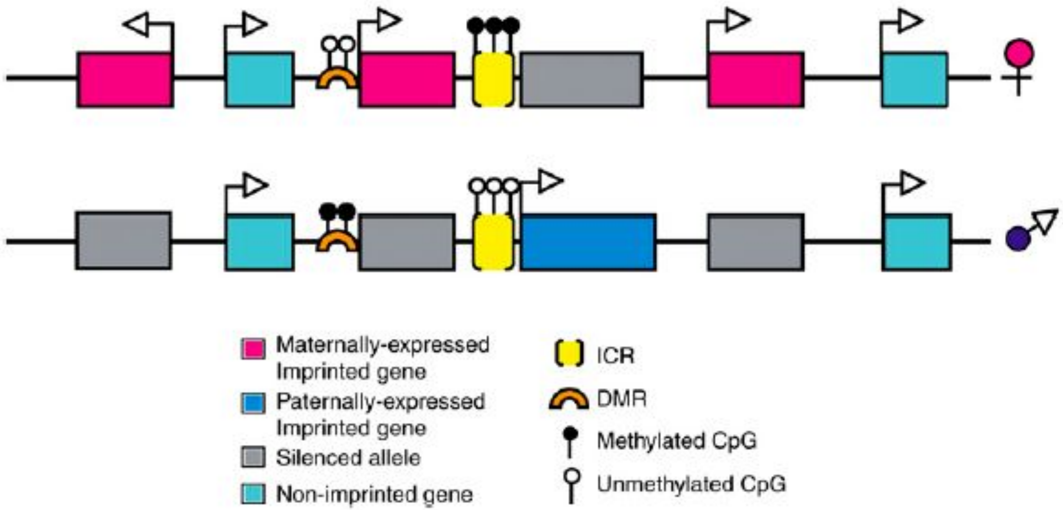
- The best known examples are the Prader-Willi and Angelman “sister” syndromes

- Both caused by deletions or non correct inheritance of the region chr15q11-q13, which contains several imprinted genes

    - Prader-Willi: loss of the paternal allele
        - SNRPN, NDN, MAGEL2 and snRNA cluster are not expressed
        - Associated with obesity and hypogonadism

    - Angelman: loss of the maternal allele
        - UBE3A not expressed
        - Associated with epilepsy and tremors

Genomic imprinting in mammals

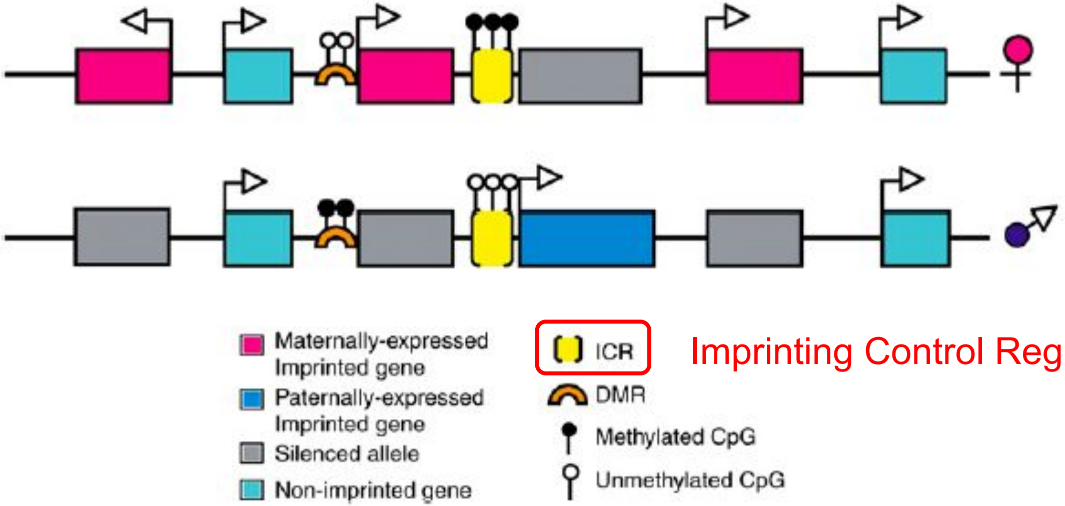

Maternally-expressed Imprinted gene
Paternally-expressed Imprinted gene
Silenced allele
Non-imprinted gene
ICR
DMR
Methylated CpG
Unmethylated CpG

Clusters of imprinted genes include:

- Maternally expressed genes
- Paternally expressed genes
- Non-imprinted genes
- Insulators and non-coding RNAs
- Differentially DNA Methylated Regions (DMRs)

Bartolomei 2009

Genomic imprinting in mammals


Maternally-expressed Imprinted gene
Paternally-expressed Imprinted gene
Silenced allele
Non-imprinted gene
ICR
DMR
Methylated CpG
Unmethylated CpG

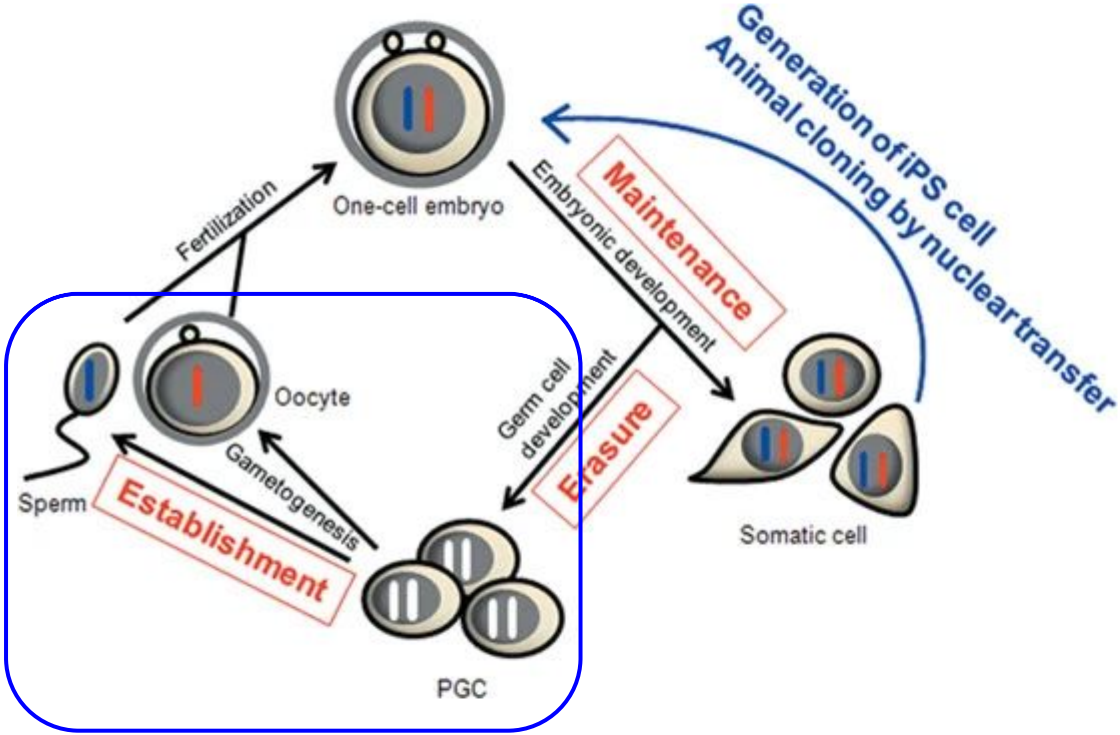
Imprinting Control Region

ICR:

- a gametic DMR (shows parental allele-specific DNA methylation and histone modifications)
- a few kilobase long
- directly associated with the mono-allelic expression of the locus
    - when deleted, loss of imprinted gene expression in the locus

Bartolomei 2009

Establishment, maintenance and erasure of genomic imprinting



Occurs during gametogenesis

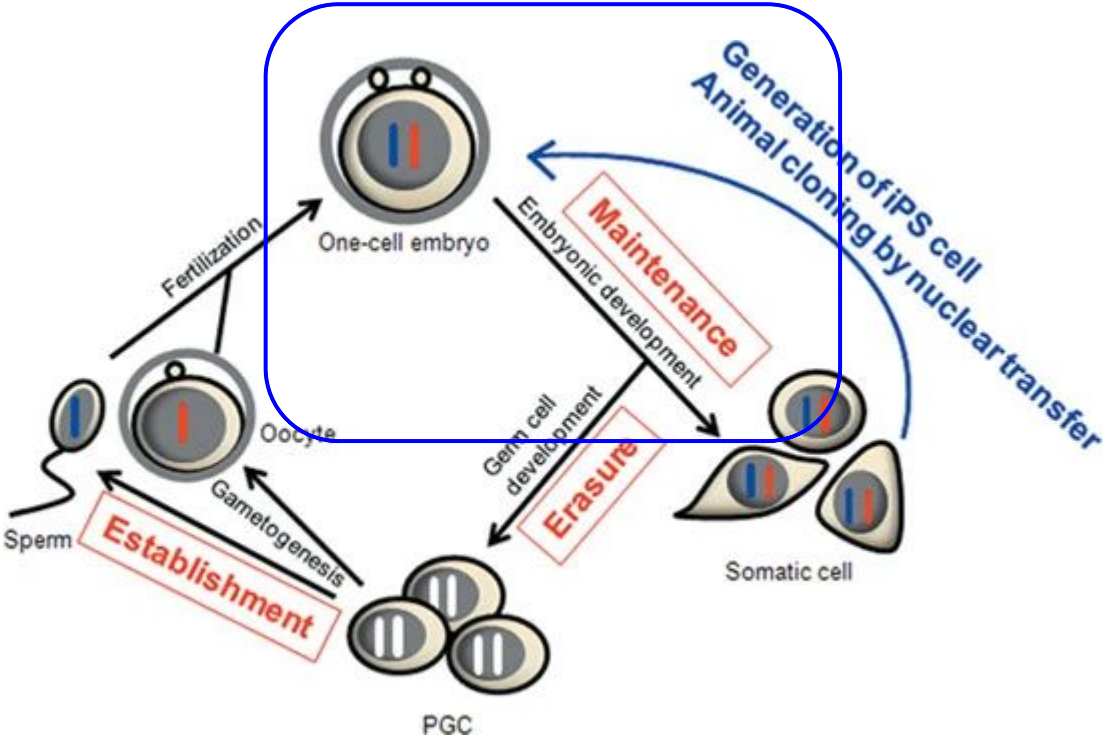
- males: starts at E14.5 at 4 gametic DMRs
- females: starts during oocyte growth asynchronously at different DMRs

de novo methyltransferases involved: DNMT3A, DNMT3B, DNMT3L

Targets:

- 8-10 bp CpG islands
- unmethylated H3K4
- binding sites of ZFP57
- transcribed DMRs

Establishment, maintenance and erasure of genomic imprinting



Genomic imprinting needs to survive two global waves of epigenome reprogramming

1. Pre-implantation: global demethylation
2. Post-implantation: methylation of germ cell-specific and pluripotency genes

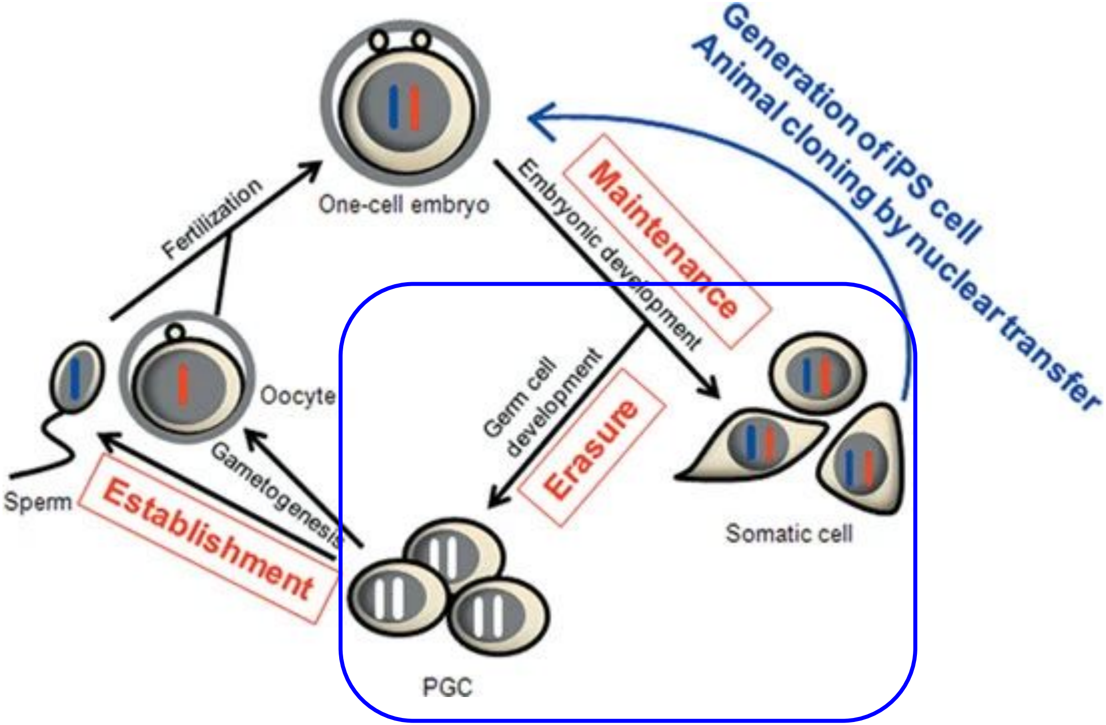
Factors involved:

- DNMT1, ZPF57, PGC7, MBD3

Histone modifications:

- H3K4me, acetylation (unmethylated allele)
- H3K9me2/3, H3K27me3, H4K20me3 (methylated allele)

Establishment, maintenance and erasure of genomic imprinting

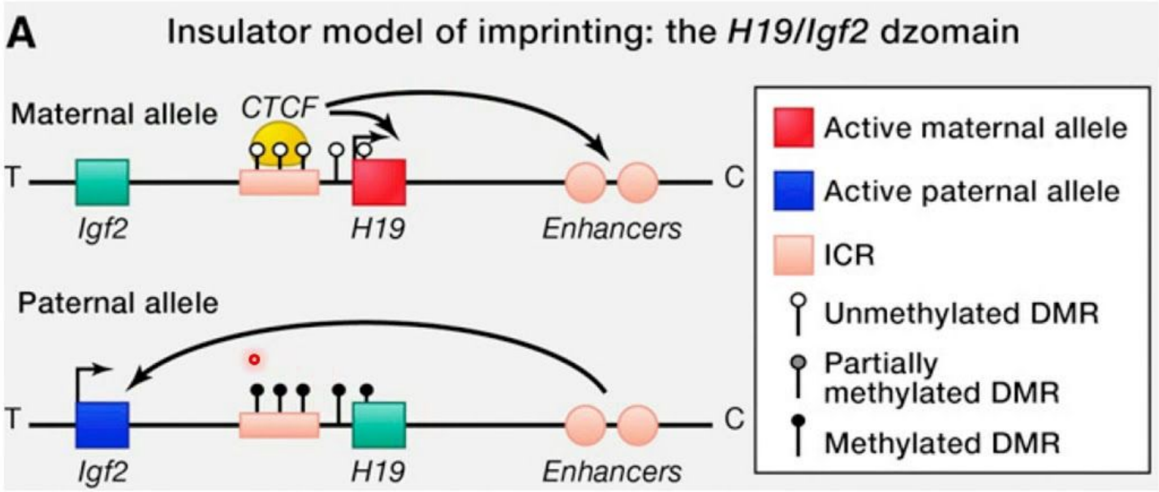


- Occurs in primordial germ cells (PGCs)
- Parental origin-specific DNA methylation is erased
- Imprinted genes go back to biallelic expression / silencing

DNA demethylation carried out by TET1, TET2 and TET3

CTCF mediates monoallelic expression in H19/IGF2 cluster

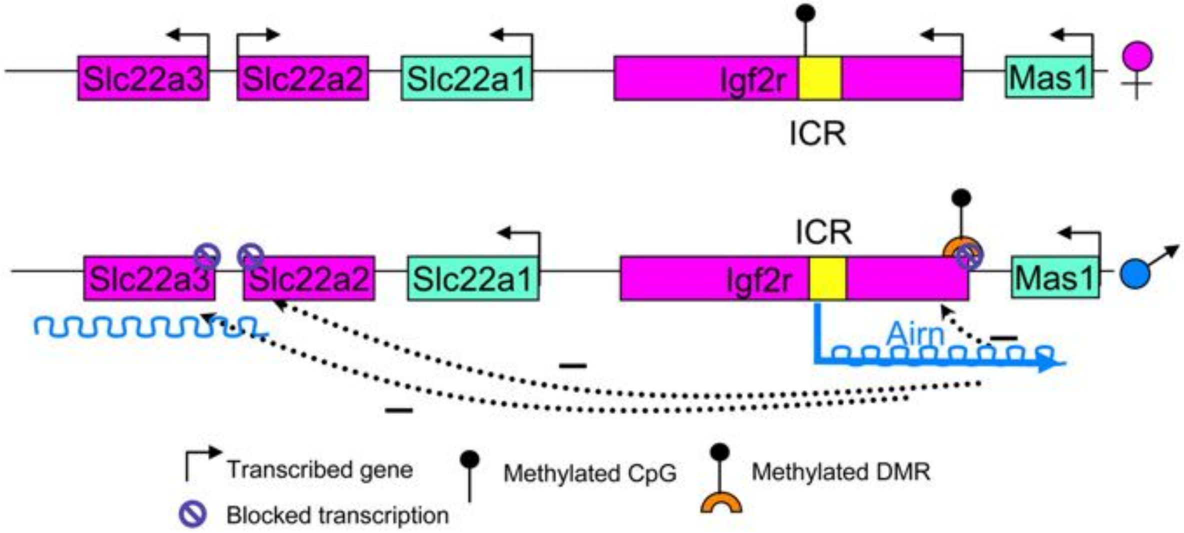
ICR (Imprinting Control Region) is within a **DMR**



The region upstream of *H19* is one of the four gametic DMRs methylated in the paternal germline

- *H19* expressed from the maternal allele
- *Igf2* expressed from the paternal allele

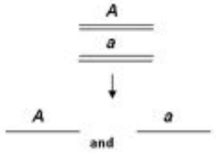
The long non-coding RNA Airn is involved in the imprinted expression of Igf2r



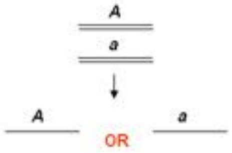
- ICR methylated on the maternal allele
- *Igf2r, Slc22a2/3* expressed from the maternal allele
- *Airn* expressed from the paternal allele

Bartolomei 2009

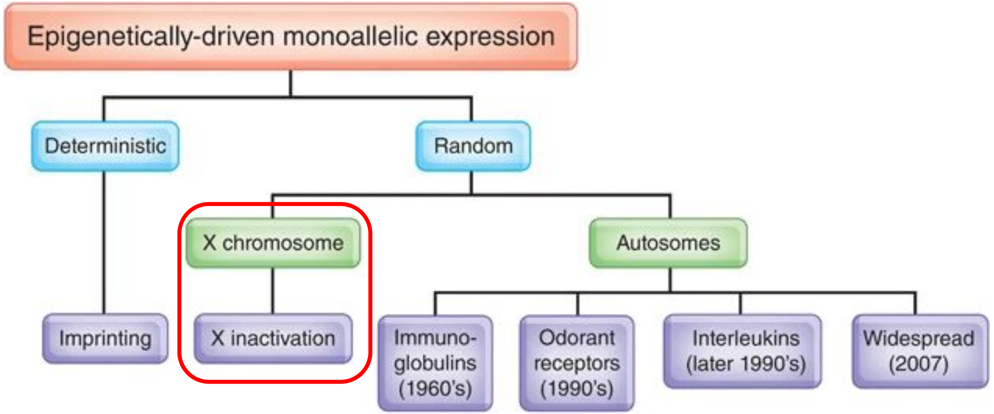
X chromosome inactivation (XCI)



bi-allelic expression

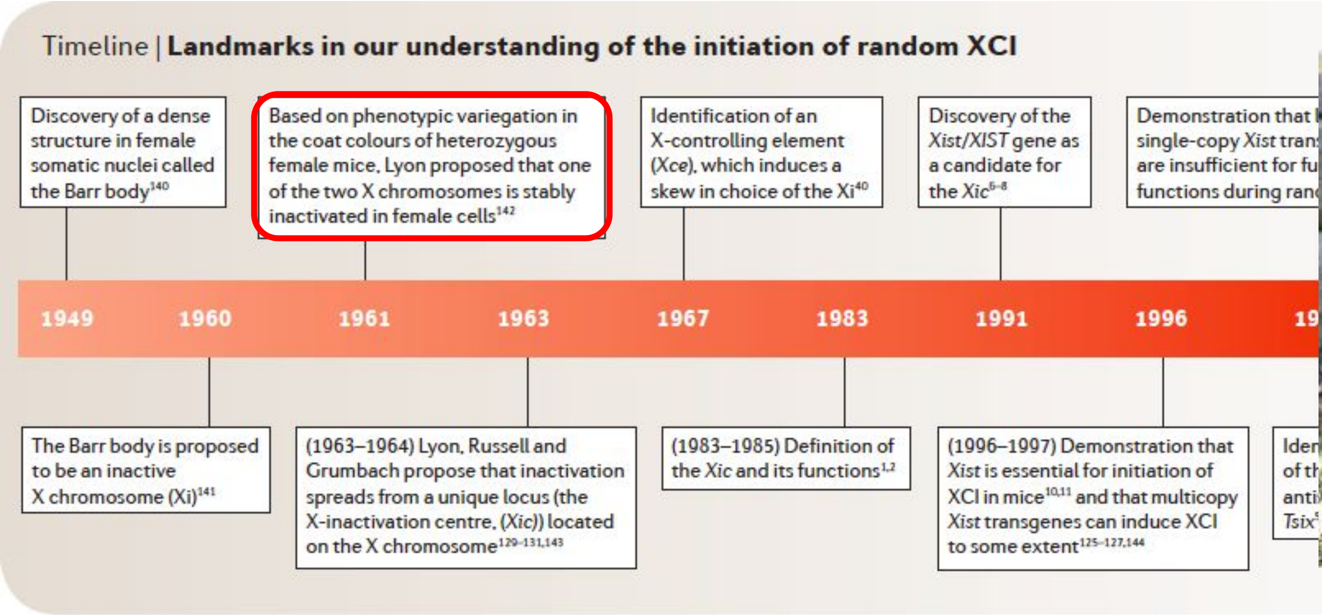


monoallelic expression


Epigenetically-driven monoallelic expression
Deterministic
Random
X chromosome
Autosomes
Imprinting
X inactivation
Immuno-globulins (1960's)
Odorant receptors (1990's)
Interleukins (later 1990's)
Widespread (2007)

- An example of mono-allelic expression at the chromosomal level

X chromosome inactivation (XCI)



Timeline | **Landmarks in our understanding of the initiation of random XCI**

1949 — Discovery of a dense structure in female somatic nuclei called the Barr body[140]

1960 — The Barr body is proposed to be an inactive X chromosome (Xi)[141]

1961 — Based on phenotypic variegation in the coat colours of heterozygous female mice, Lyon proposed that one of the two X chromosomes is stably inactivated in female cells[142]

1963 — (1963–1964) Lyon, Russell and Grumbach propose that inactivation spreads from a unique locus (the X-inactivation centre, (*Xic*)) located on the X chromosome[129–131,143]

1967 — Identification of an X-controlling element (*Xce*), which induces a skew in choice of the Xi[40]

1983 — (1983–1985) Definition of the *Xic* and its functions[1,2]

1991 — Discovery of the *Xist*/*XIST* gene as a candidate for the *Xic*[6–8]

1996 — (1996–1997) Demonstration that *Xist* is essential for initiation of XCI in mice[10,11] and that multicopy *Xist* transgenes can induce XCI to some extent[125–127,144]

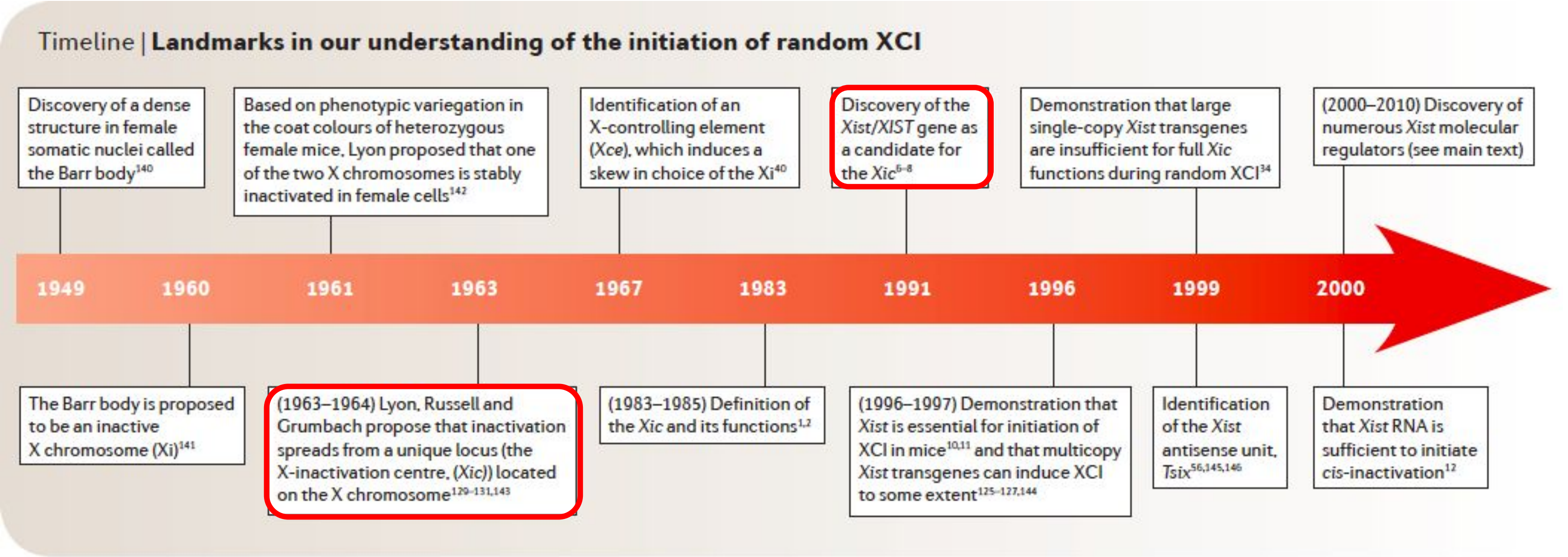
19

Demonstration that single-copy *Xist* tran are insufficient for fu functions during ran

Iden of th anti *Tsix*[5

- Dosage compensation mechanism to avoid overexpression of genes on the X chromosome
    - XCI not observed in normal males XY or abnormal females X0
    - XXX or XXXX individuals have only one active X chromosome

X chromosome inactivation (XCI)



Timeline | **Landmarks in our understanding of the initiation of random XCI**

Discovery of a dense structure in female somatic nuclei called the Barr body[140]

The Barr body is proposed to be an inactive X chromosome (Xi)[141]

Based on phenotypic variegation in the coat colours of heterozygous female mice, Lyon proposed that one of the two X chromosomes is stably inactivated in female cells[142]

(1963–1964) Lyon, Russell and Grumbach propose that inactivation spreads from a unique locus (the X-inactivation centre, (*Xic*)) located on the X chromosome[129–131,143]

Identification of an X-controlling element (*Xce*), which induces a skew in choice of the Xi[40]

(1983–1985) Definition of the *Xic* and its functions[1,2]

Discovery of the *Xist*/*XIST* gene as a candidate for the *Xic*[5–8]

(1996–1997) Demonstration that *Xist* is essential for initiation of XCI in mice[10,11] and that multicopy *Xist* transgenes can induce XCI to some extent[125–127,144]

Demonstration that large single-copy *Xist* transgenes are insufficient for full *Xic* functions during random XCI[34]

Identification of the *Xist* antisense unit, *Tsix*[56,145,146]

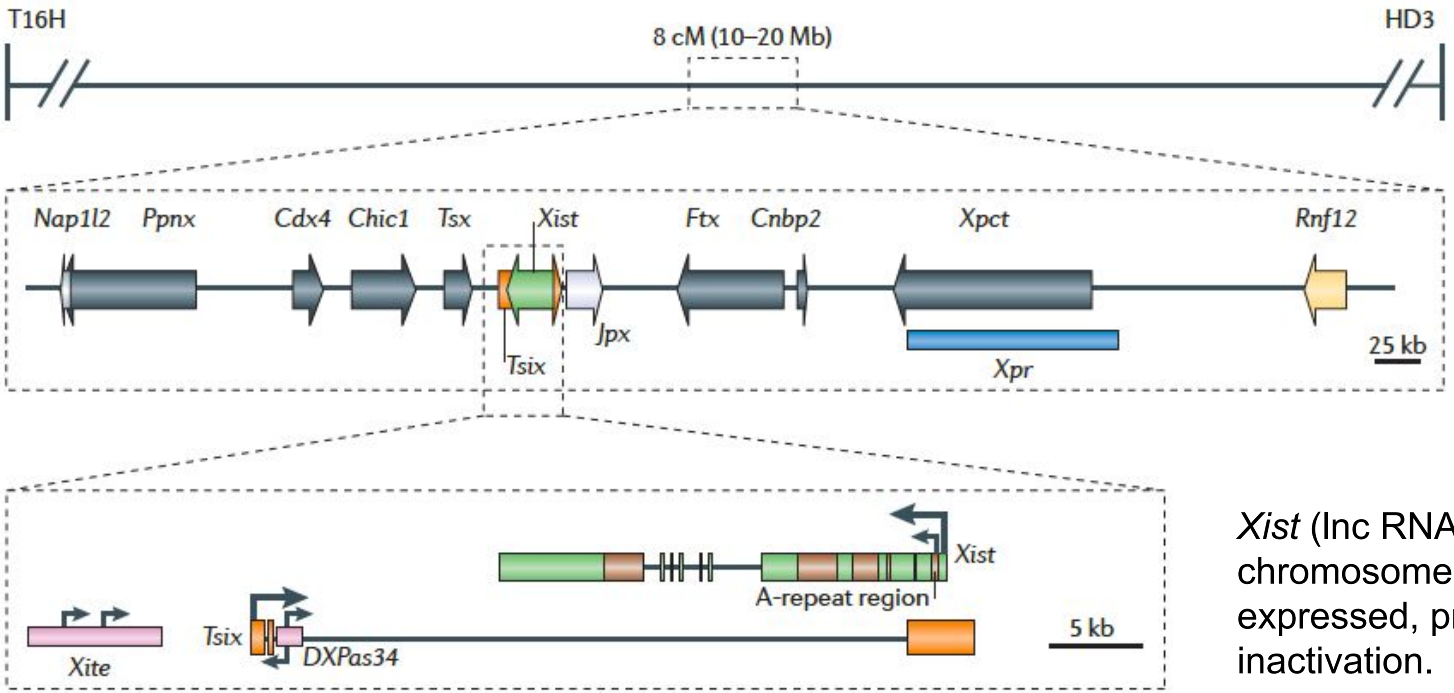
(2000–2010) Discovery of numerous *Xist* molecular regulators (see main text)

Demonstration that *Xist* RNA is sufficient to initiate cis-inactivation[12]

- Dosage compensation mechanism to avoid overexpression of genes on the X chromosome
  - XCI not observed in normal males XY or abnormal females X0
  - XXX or XXXX individuals have only one active X chromosome

X chromosome inactivation (XCI)

The minimum candidate region for the X-inactivation Center (xic) on the X chromosome



Xist (lnc RNA) coats the X chromosome from which it is expressed, promoting its inactivation.

X chromosome inactivation (XCI)

Waves of X chromosome inactivation and reactivation in the mouse embryo

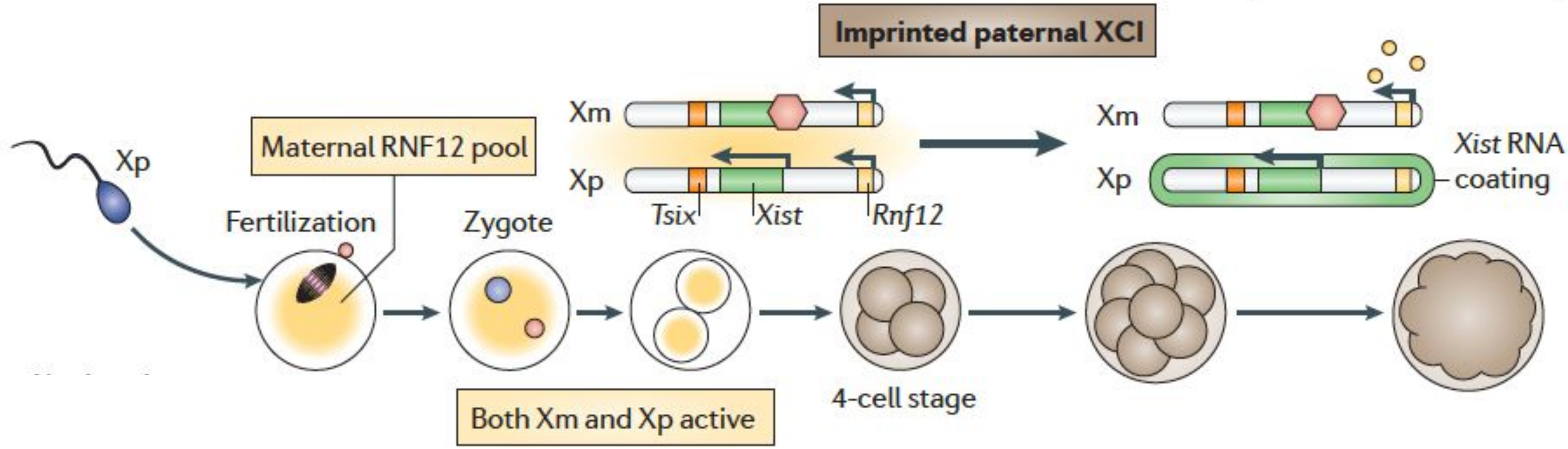
1. pre-implantation embryo: **imprinted** inactivation of paternal X chromosome (Xp)

A gametic maternal imprint (not related to DNA methylation) prevents expression of Xist from the Xm

⬢ Maternal imprint represses *Xist*

● Pluripotency factors repress *Xist*

● High dose of RNF12 activates *Xist*

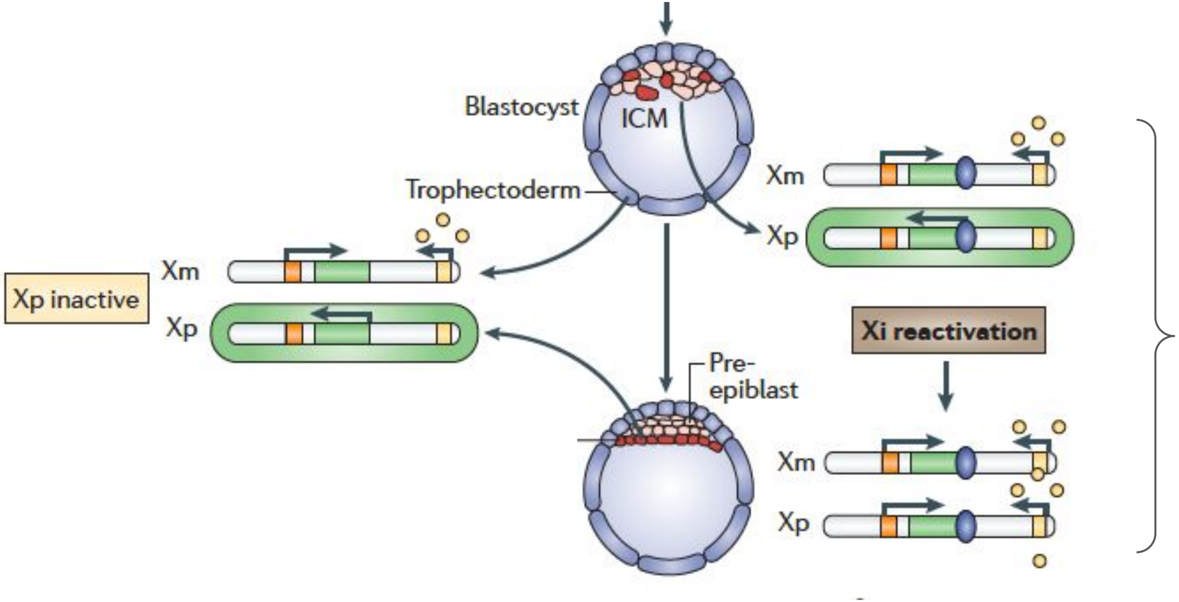


X chromosome inactivation (XCI)

Waves of X chromosome inactivation and reactivation in the mouse embryo

2. inner cell mass: Xp is reactivated → both X chromosomes are active
Xp remains inactive in extra-embryonic tissues (trophectoderm and placenta)

- ⬢ Maternal imprint represses *Xist*
- ⬮ Pluripotency factors repress *Xist*
- ● High dose of RNF12 activates *Xist*



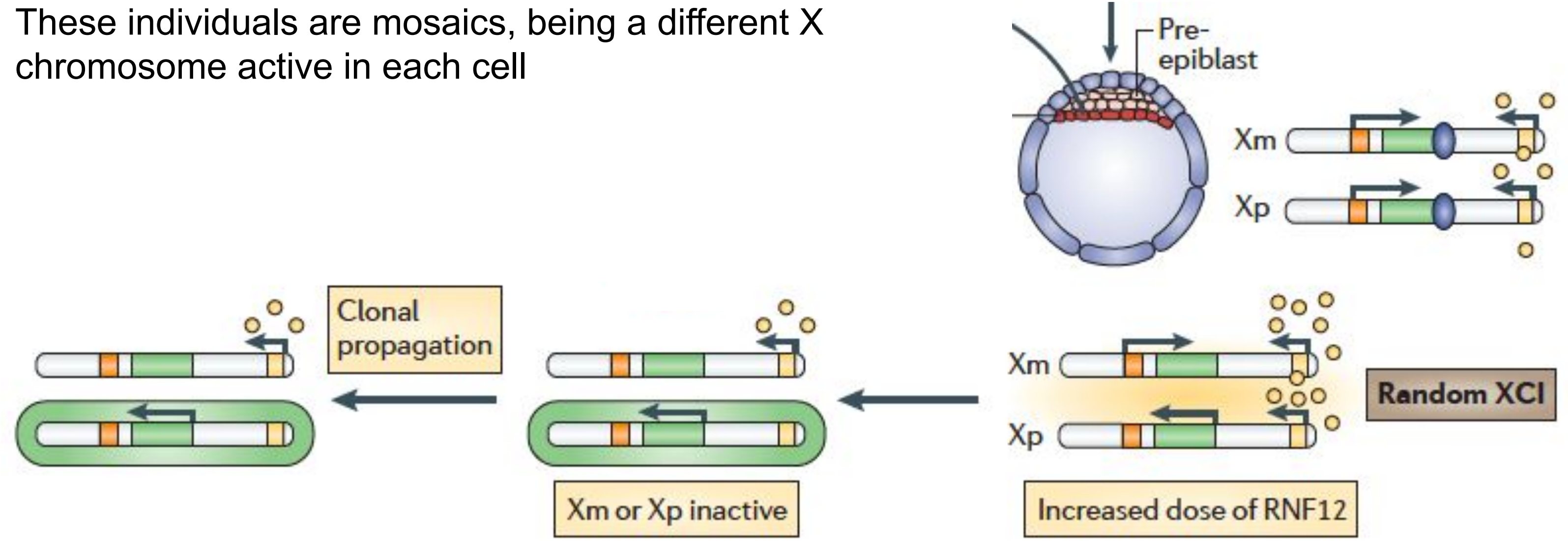
Pluripotency factors bind to *Xist* and silence its expression

X chromosome inactivation (XCI)

Waves of X chromosome inactivation and reactivation in the mouse embryo

3. inner cell mass: **random** inactivation of Xp or Xm and faithful maintenance in daughter cells

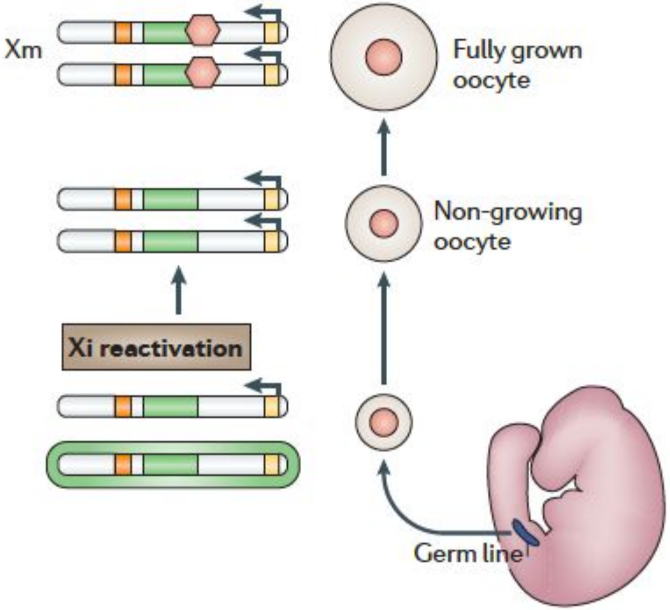
- It occurs only in individuals with ≥ 2 X chromosomes
- These individuals are mosaics, being a different X chromosome active in each cell



X chromosome inactivation (XCI)

Waves of X chromosome inactivation and reactivation in the mouse embryo

4. Gametogenesis in individuals with ≥ 2 X chromosomes: re-activation of the inactivated X chromosome → each oocyte carries an active X chromosome



X chromosome inactivation (XCI)

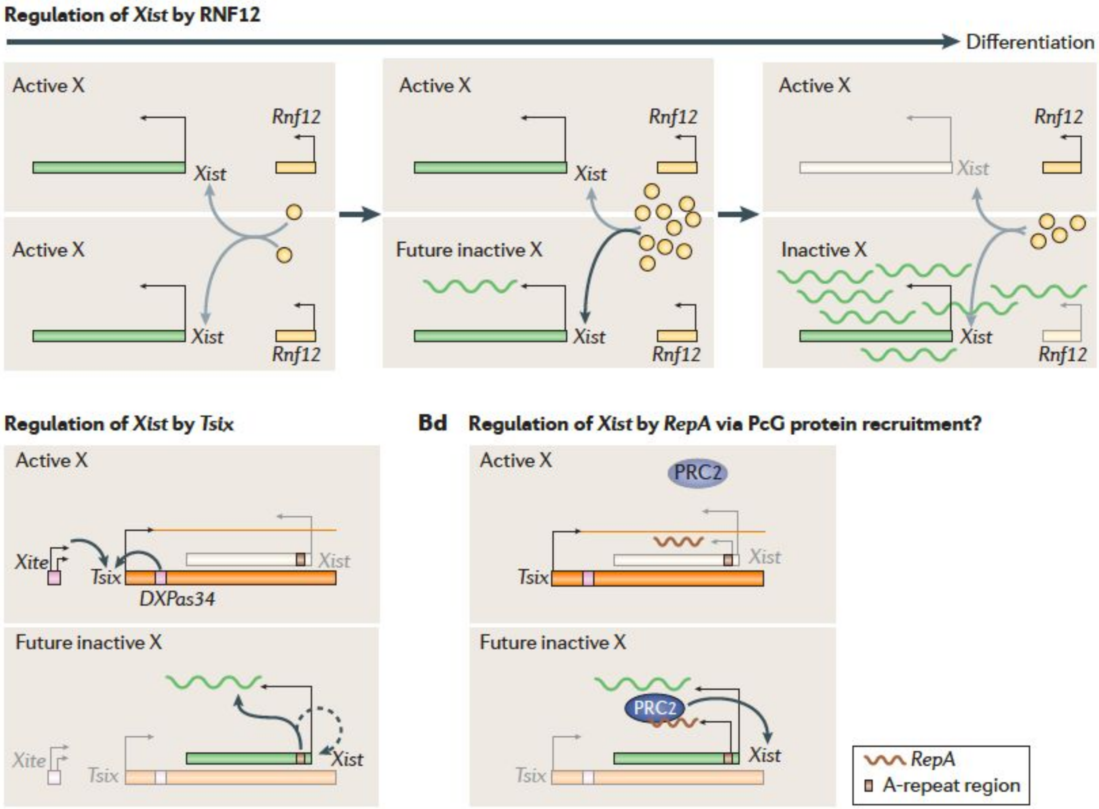
Random inactivation of a X chromosome requires:

- Competence / sensing: XCI happens only in cells with ≥ 2 Xic bearing chromosomes

- Counting: the process by which the cell decides the number of X chromosomes to be inactivated, based on the ratio X/autosomal chromosomes. There will always be N - 1 inactive X chromosomes

- Choice: of which X chromosome will be inactivated

X chromosome inactivation (XCI)



Regulation of *Xist* by RNF12

Regulation of *Xist* by *Tsix*

Bd Regulation of *Xist* by *RepA* via PcG protein recruitment?

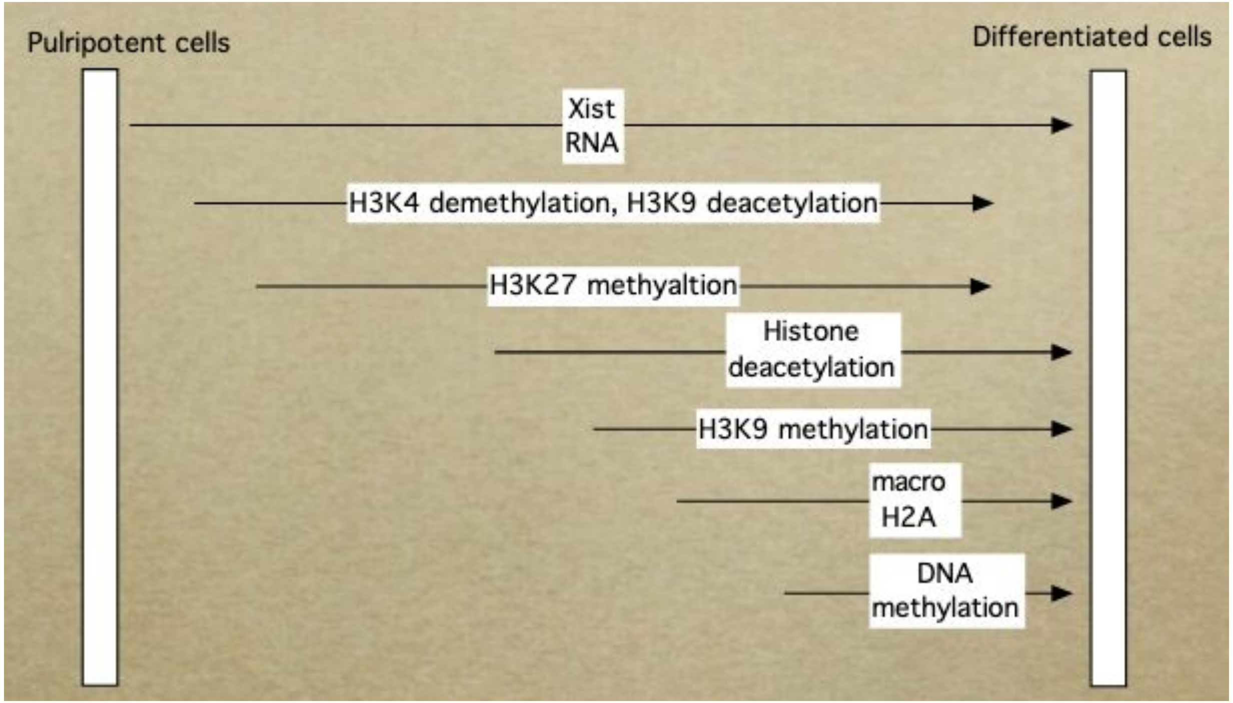
Different factors regulate *Xist* expression in a complementary way during random XCI:

- Dosage dependent expression of ***Rnf12***: it promotes overexpression of *Xist* and consequential negative feedback loop on the expression of *Rnf12*

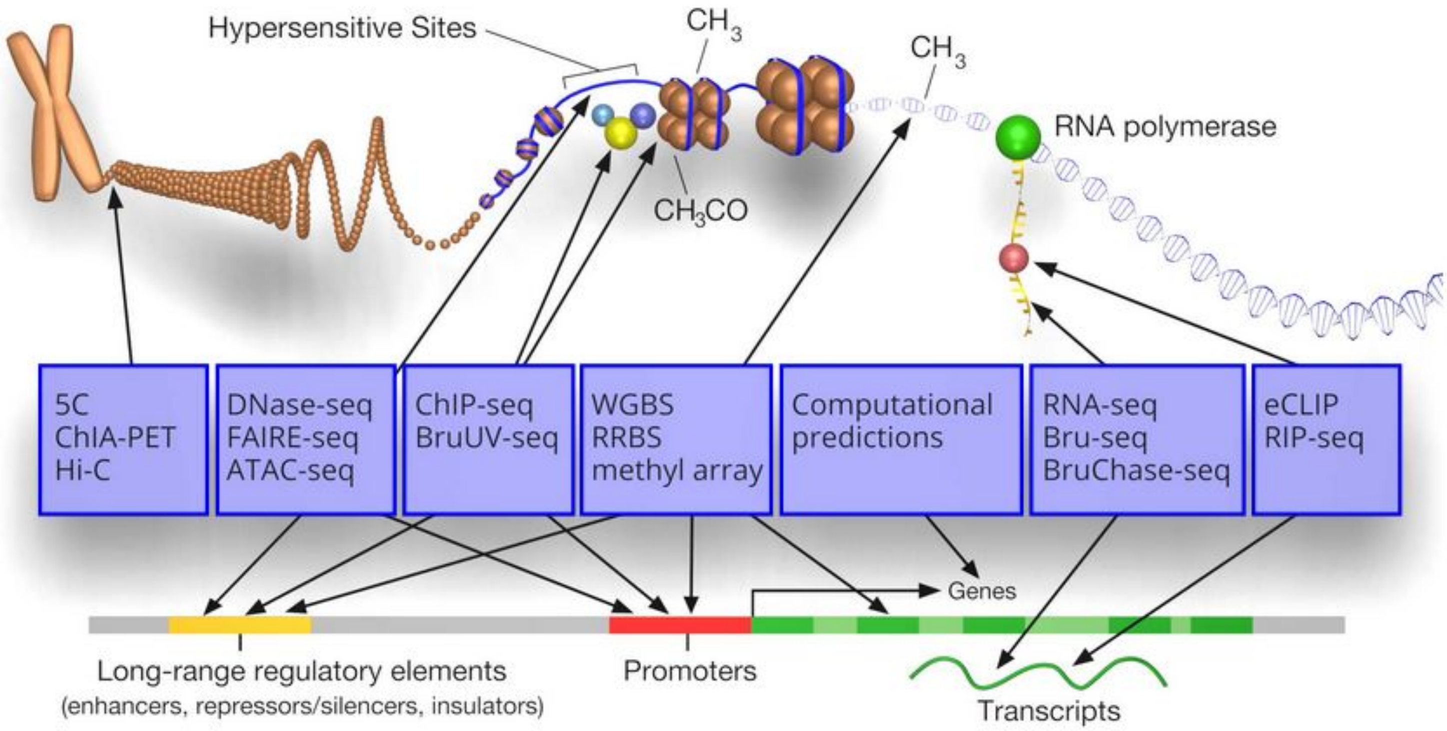
- Antisense transcription of ***Tsix***, counteracts expression of *Xist*

- ***RepA***, a short RNA, recruits PRC2 to the Xist promoter and promotes Xist activation

X chromosome inactivation (XCI)



The Encyclopedia Of DNA Elements (ENCODE)



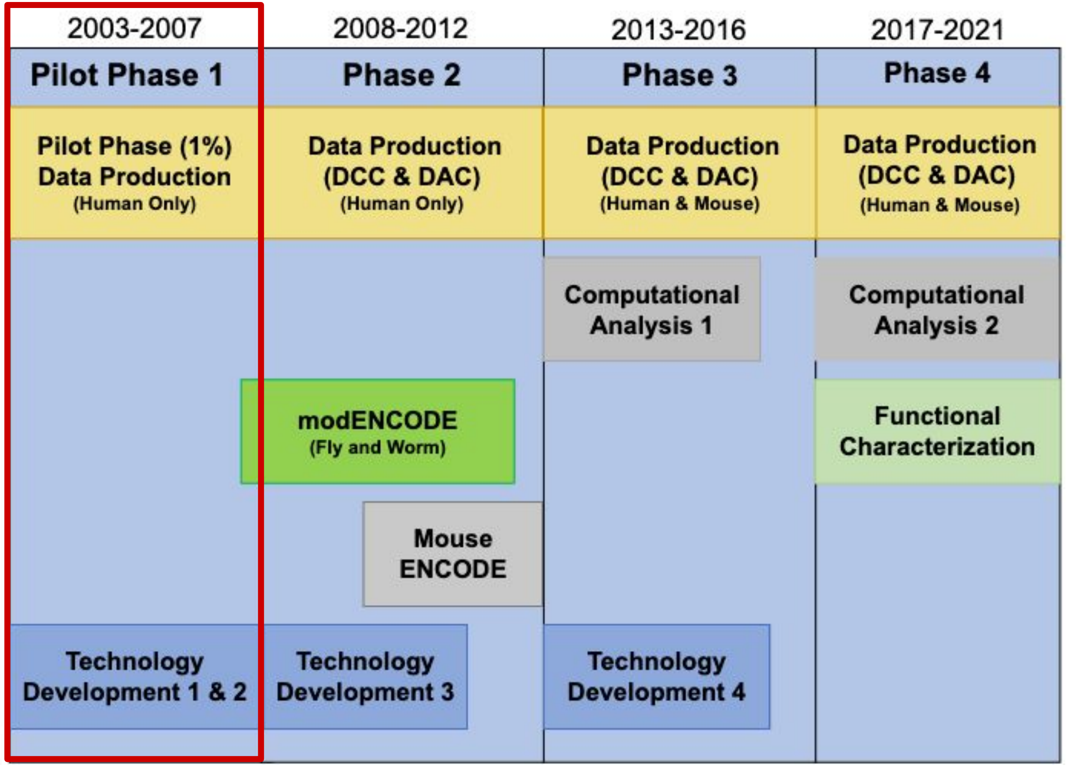
https://www.encodeproject.org/

The Encyclopedia Of DNA Elements (ENCODE)

In 2003, the National Human Genome Research Institute (NHGRI) launches *ENCODE* as a follow-up project to the *Human Genome Project*.

The goal is to enable the scientific and medical communities to interpret the human genome sequence and apply it to understand human biology and improve health

- this is possible through the **comprehensive annotation of all functional elements of the genome**
- functional element: a discrete region that encodes a defined product (e.g. protein) or a reproducible biochemical signature (transcript or chromatin structure)



2003-2007
Pilot Phase 1
Pilot Phase (1%) Data Production (Human Only)
Technology Development 1 & 2

2008-2012
Phase 2
Data Production (DCC & DAC) (Human Only)
modENCODE (Fly and Worm)
Mouse ENCODE
Technology Development 3

2013-2016
Phase 3
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 1
Technology Development 4

2017-2021
Phase 4
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 2
Functional Characterization

The ENCODE Project Consortium, 2011

The Encyclopedia Of DNA Elements (ENCODE)

- During Pilot Phase 1, ENCODE focused on ~1% of the human genome (~30 Mb) and tested and compared different existing methods and protocols in order to select the best approach to scale up for the analysis of the whole genome.
- In parallel, a *Technology Development* step was carried out to
  - develop and advance new technologies towards an accurate, complete and cost-effective throughput
  - establish a paradigm for sharing functional genomics data
- Two main cell lines were used: HeLa S3 (cervical adenocarcinoma) and GM06990 (Epstein-Barr virus-transformed B-lymphocyte)

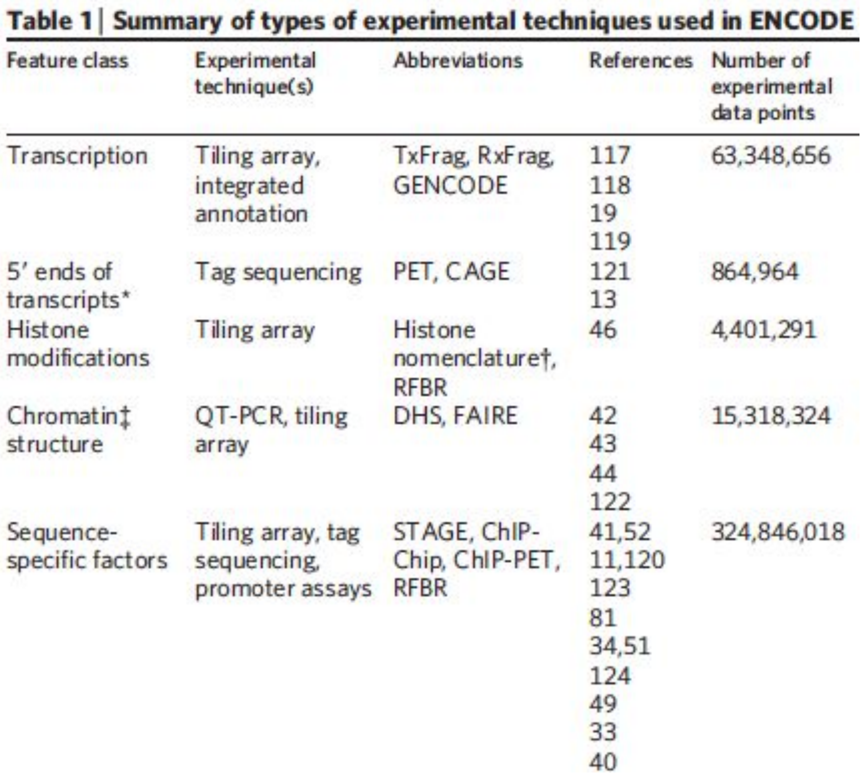


Table 1 | Summary of types of experimental techniques used in ENCODE

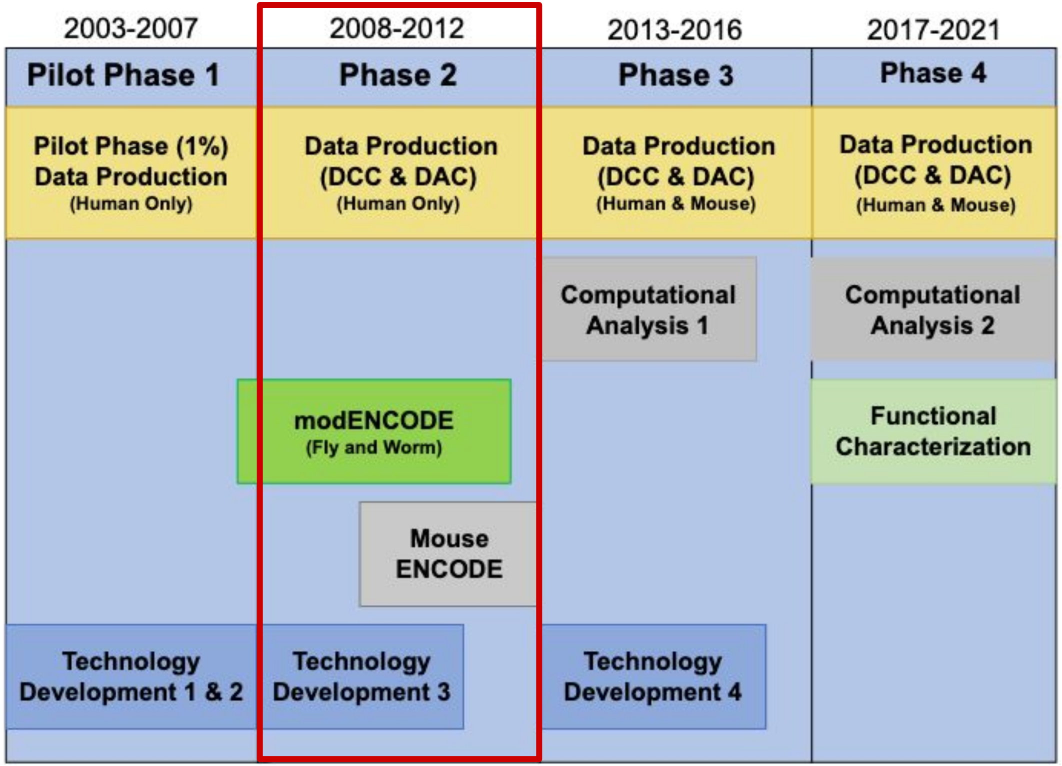
Feature class	Experimental technique(s)	Abbreviations	References	Number of experimental data points
Transcription	Tiling array, integrated annotation	TxFrag, RxFrag, GENCODE	117 118 19 119	63,348,656
5′ ends of transcripts*	Tag sequencing	PET, CAGE	121 13	864,964
Histone modifications	Tiling array	Histone nomenclature†, RFBR	46	4,401,291
Chromatin‡ structure	QT-PCR, tiling array	DHS, FAIRE	42 43 44 122	15,318,324
Sequence-specific factors	Tiling array, tag sequencing, promoter assays	STAGE, ChIP-Chip, ChIP-PET, RFBR	41,52 11,120 123 81 34,51 124 49 33 40	324,846,018

The ENCODE Project Consortium, Birney et al., 2007

The Encyclopedia Of DNA Elements (ENCODE)

ENCODE Phase 2 was mainly characterized by a **large-scale data production** effort which led to the release of the first functional genomics datasets publicly available

- supported by the Data Coordination Center (DCC) and the Data Analysis Center (DAC)
- 3 sets (tiers) of human cell lines were assayed
    - tier 1 (highest priority): K562, GM12878, H1 (ESC)
    - tier 2 (intermediate priority): HeLa-S3, HepG2, UVEC
    - tier 3 (lowest priority): > 100 cell lines


2003-2007
2008-2012
2013-2016
2017-2021
Pilot Phase 1
Phase 2
Phase 3
Phase 4
Pilot Phase (1%) Data Production (Human Only)
Data Production (DCC & DAC) (Human Only)
Data Production (DCC & DAC) (Human & Mouse)
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 1
Computational Analysis 2
modENCODE (Fly and Worm)
Functional Characterization
Mouse ENCODE
Technology Development 1 & 2
Technology Development 3
Technology Development 4

The ENCODE Project Consortium, 2011

The Encyclopedia Of DNA Elements (ENCODE)

- One of the two main goals of Phase 2 is **Gene and Transcripts Annotation**
    - the result of this efforts leads to GENCODE (now an independent consortium dedicated to human and mouse genome annotation)
    - generates gene and transcripts models implementing manual curation supported by automated algorithms
    - defines transcript types, polyA status, localization of transcripts (nuclear vs. cytosolic)
- besides tiling arrays, massively parallel DNA sequencing is introduced (absent in the Pilot Phase)

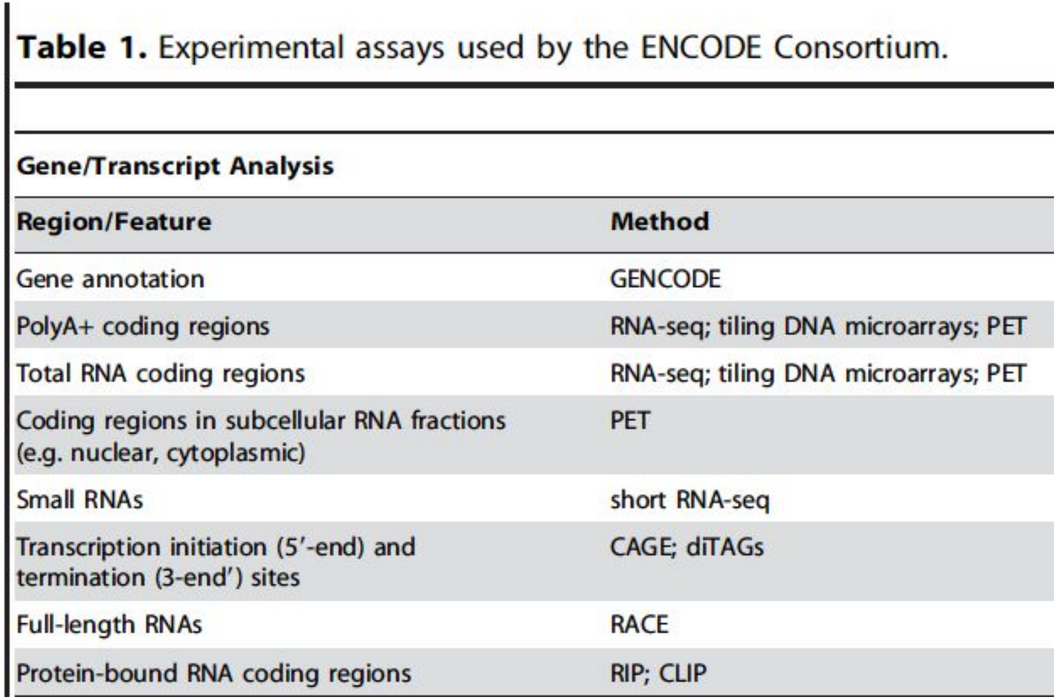


Table 1. Experimental assays used by the ENCODE Consortium.

Gene/Transcript Analysis

Region/Feature	**Method**
Gene annotation	GENCODE
PolyA+ coding regions	RNA-seq; tiling DNA microarrays; PET
Total RNA coding regions	RNA-seq; tiling DNA microarrays; PET
Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)	PET
Small RNAs	short RNA-seq
Transcription initiation (5′-end) and termination (3-end′) sites	CAGE; diTAGs
Full-length RNAs	RACE
Protein-bound RNA coding regions	RIP; CLIP

The ENCODE Project Consortium, 2011

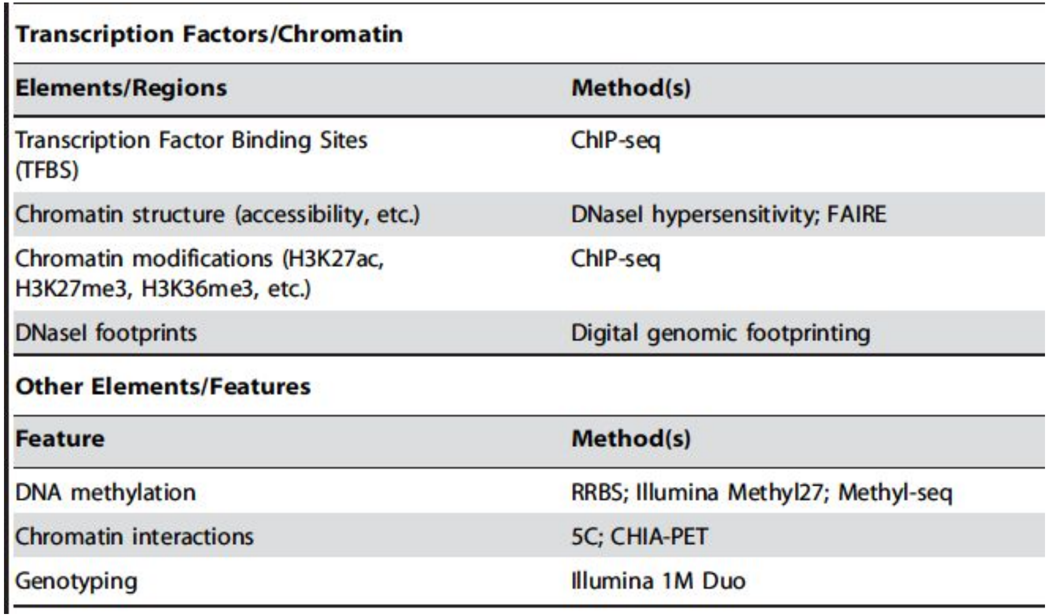
The Encyclopedia Of DNA Elements (ENCODE)

The other main goal is the annotation of cis-regulatory elements that affect the magnitude, timing and cell-specificity of gene expression. The main functional assays target:

- chromatin accessibility and histone modifications distribution
- TFs and RNA polymerase occupancy

Additional assays targeted:

- DNA methylation
- DNaseI footprints
- sequence structural variations
- long-range chromatin interactions
- protein-RNA interactions
- proteomics

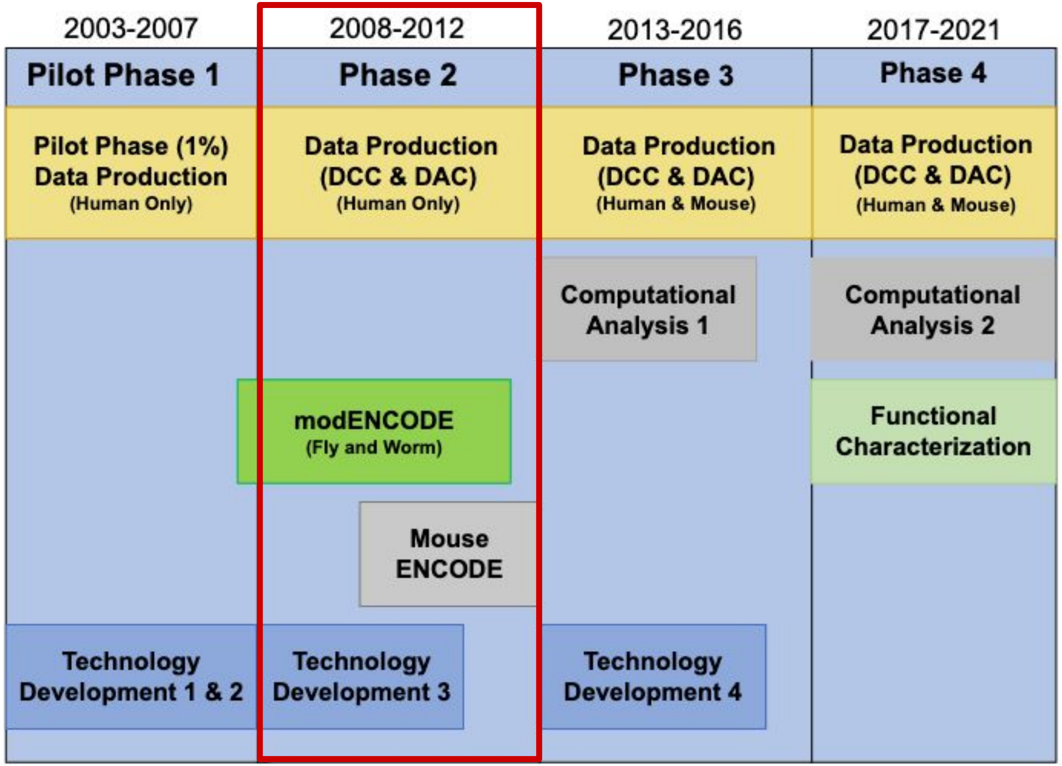


Transcription Factors/Chromatin	
Elements/Regions	**Method(s)**
Transcription Factor Binding Sites (TFBS)	ChIP-seq
Chromatin structure (accessibility, etc.)	DNaseI hypersensitivity; FAIRE
Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)	ChIP-seq
DNaseI footprints	Digital genomic footprinting
Other Elements/Features	
Feature	**Method(s)**
DNA methylation	RRBS; Illumina Methyl27; Methyl-seq
Chromatin interactions	5C; CHIA-PET
Genotyping	Illumina 1M Duo

The ENCODE Project Consortium, 2011

The Encyclopedia Of DNA Elements (ENCODE)

- As concerns the Technology Development step, ENCODE Phase 2 has defined standards for collecting and processing each data type
  - experimental components, including cell growth conditions, antibody characterization,
  - requirements for controls and biological replicates
  - assessment of reproducibility
  - formats for data submission (data parameters, experimental details etc.)
- Mouse ENCODE: annotate functional elements in the mouse genome
- modENCODE: annotate functional elements in *D. melanogaster* and *C. elegans*



2003-2007
Pilot Phase 1
Pilot Phase (1%) Data Production (Human Only)
Technology Development 1 & 2

2008-2012
Phase 2
Data Production (DCC & DAC) (Human Only)
modENCODE (Fly and Worm)
Mouse ENCODE
Technology Development 3

2013-2016
Phase 3
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 1
Technology Development 4

2017-2021
Phase 4
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 2
Functional Characterization

The ENCODE Project Consortium, 2011

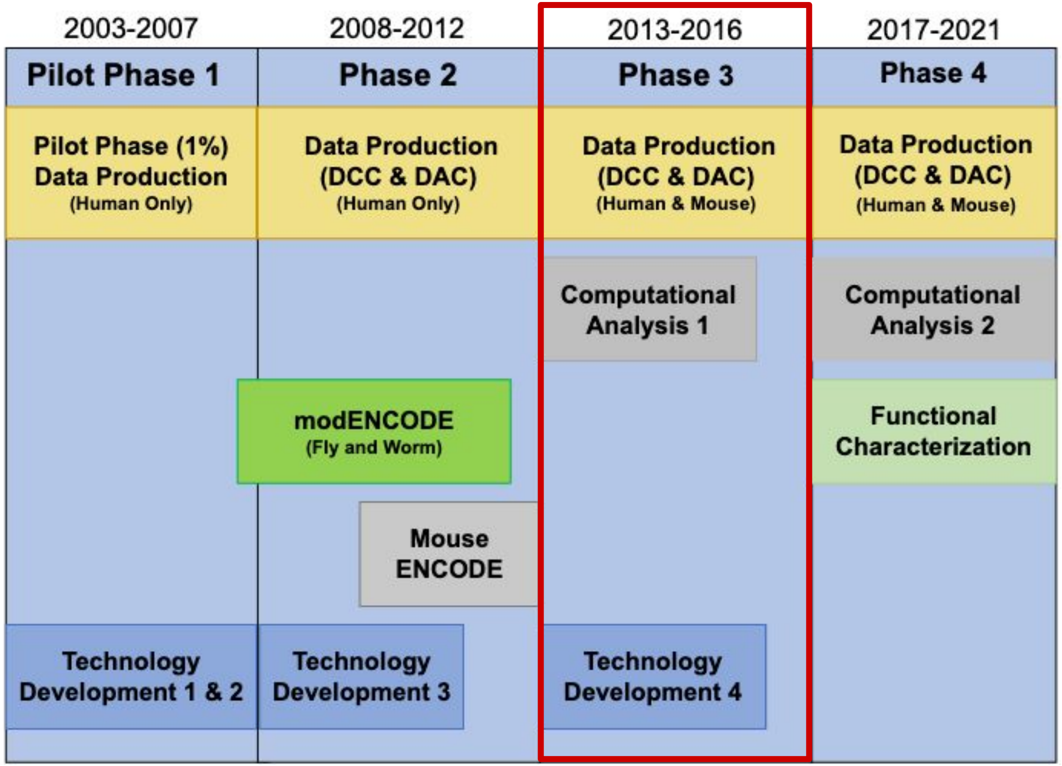
The Encyclopedia Of DNA Elements (ENCODE)

During Phase 3, a new data production phase went on to populate as much as possible the different tiers of human and mouse cell lines with all functional assays.

Computational analysis groups joined ENCODE to help with the multi-omics data integration.

Technology Development 4 is mainly focused on:

- CLIP technologies to study RNA-protein interactions
- Hi-C technology to study long-range interactions
- construction of personalized genomes (see later: EN-TEx)
- Functional characterization
- BruChase-seq and BrUV-seq assays to study RNA metabolism



2003-2007
Pilot Phase 1
Pilot Phase (1%) Data Production (Human Only)
Technology Development 1 & 2

2008-2012
Phase 2
Data Production (DCC & DAC) (Human Only)
modENCODE (Fly and Worm)
Mouse ENCODE
Technology Development 3

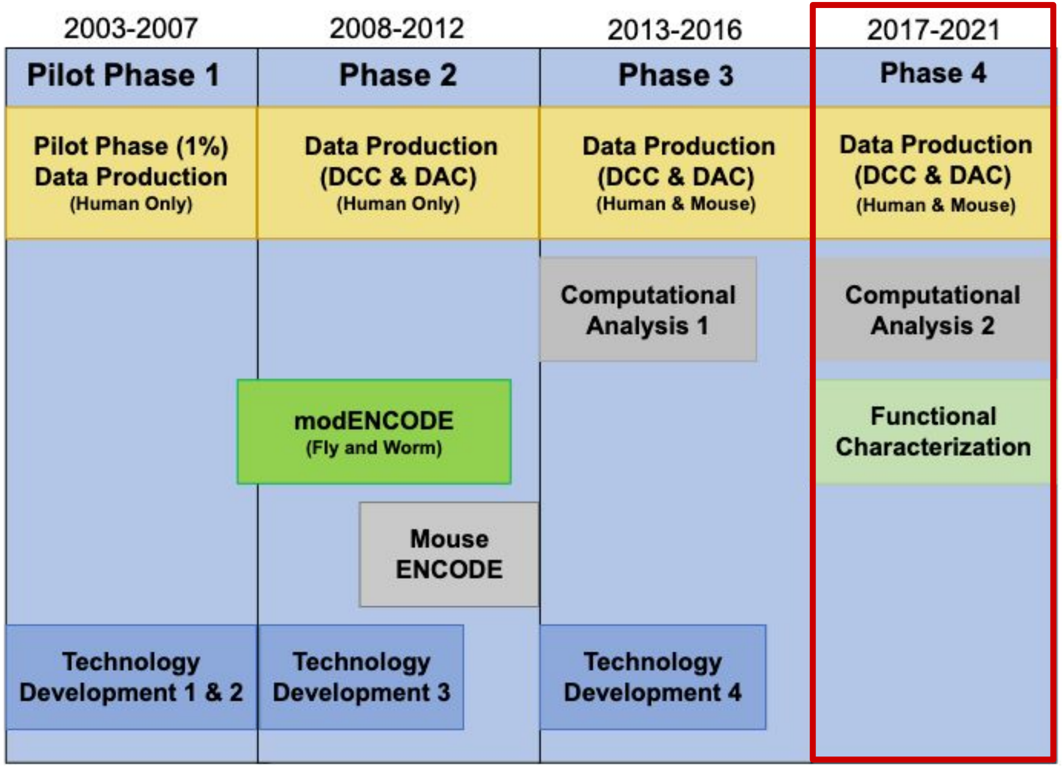
2013-2016
Phase 3
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 1
Technology Development 4

2017-2021
Phase 4
Data Production (DCC & DAC) (Human & Mouse)
Computational Analysis 2
Functional Characterization

The Encyclopedia Of DNA Elements (ENCODE)

Currently (phase 4), ENCODE has three main goals:

- provide for cell lines K562 and HepG2 the most comprehensive list of functional assays
  - this mainly includes the generation of ChIP-seq maps for a large fraction of TFs
- functional characterization of the identified regulatory elements with CRISPR/RNAi technologies
- Integrative analysis of the data generated in previous phases



2003-2007

Pilot Phase 1

Pilot Phase (1%) Data Production (Human Only)

Technology Development 1 & 2

2008-2012

Phase 2

Data Production (DCC & DAC) (Human Only)

modENCODE (Fly and Worm)

Mouse ENCODE

Technology Development 3

2013-2016

Phase 3

Data Production (DCC & DAC) (Human & Mouse)

Computational Analysis 1

Technology Development 4

2017-2021

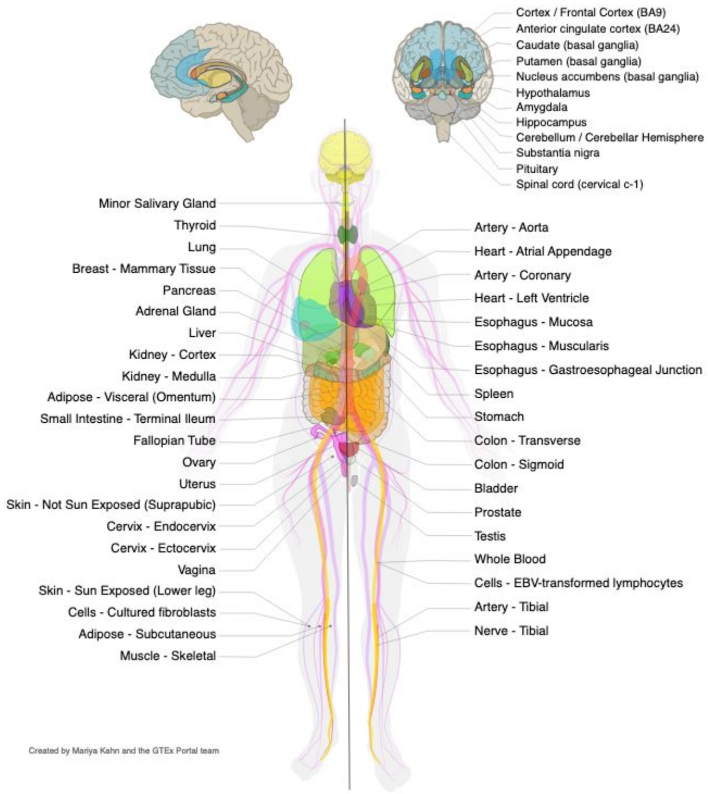
Phase 4

Data Production (DCC & DAC) (Human & Mouse)

Computational Analysis 2

Functional Characterization

The Genotype-Tissue Expression project (GTEx)



- It's an effort to build a comprehensive public resource to study tissue-specific gene expression and regulation
- It collected samples from 54 non-diseased tissue sites across nearly 1000 individuals
- The molecular assays performed are: whole-genome sequencing (WGS), whole-exome sequencing (WES) and RNA-seq
  - WGS and WES performed in whole blood only
  - RNA-seq performed systematically across tissues and organs
- It provides open access to gene expression matrices, QTLs and histology images

GTEx Portal

https://gtexportal.org/home/

EN-TEx = ENCODE + GTEx

ENCODE

+

GTEx

- The goal of the EN-TEx project is the expression and regulation analysis of personalized genomes
  - reference genome: the result of a coordinated effort to assemble the human or another species' genome by assembling DNA fragments from multiple donors. Does not represent the genome sequence of a specific individual
  - personalized genome: the result of *de novo* assembling the genome of a specific individual. It's less accurate but allows to detect structural variations (insertions, deletions, inversions) specific of that person.
- EN-TEx focuses on 4 GTEx donors (2 males & 2 females), for whom has built personalized genomes
  - the construction of personalized genomes is made possible by a combination of different DNA sequencing technologies (long- and short-read). In this case DNA sequencing was performed from *transverse colon* (in the case of GTEx it was done with *whole blood*).