Master 2 PRIAM 2023: Functional and Spatial annotation of livestocks' genomes

Hervé Acloque

Functional and Spatial annotation of livestocks' genomes

Hervé ACLOQUE
herve.acloque@inrae.fr
Master2 PRIAM
4th of December 2023

Chromosome and DNA: a long history

Chromosome discovery

1842: First observation of chromosomes in plant cells by Karl Wilhelm von Nägeli

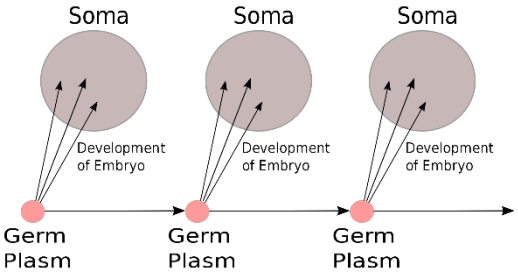
1882: First observation of chromosomes in animal cells by Walther Flemming during mitosis

Flemming, W. Zellsubstanz, Kern und Zelltheilung (F. C. W. Vogel, Leipzig, 1882).

Chromosome theory of inheritance

1866: Mendel's laws on inheritance of traits

1892 : Germ Plasm : a Theory of Heredity (refutation of inheritance of acquired traits, chromosome) by August Weissman & Wilhelm Roux



1902: Chromosome theory of inheritance by Boveri-Sutton. Chromosomes are the carriers of genetic material and worked by pairs in accordance to Mendel's laws

Chromosome and DNA: a long history

DNA is the molecule that carries the genetic information

1869 : First isolation of nuclein by Friedrich Miescher

1919 : Phoebus Levene identified the molecular structure of DNA

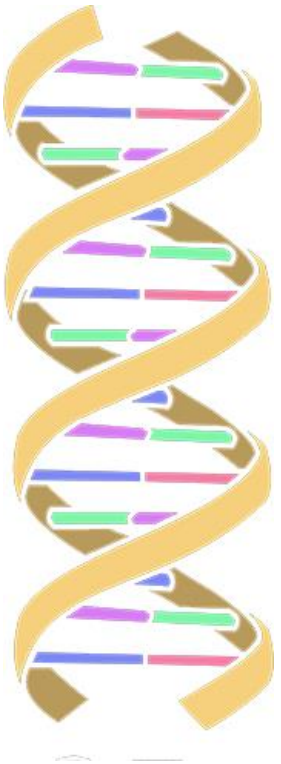
1944: Avery-MacLeod-McArty showed that DNA is the material which genes and chromosomes are made

1953: Double Helix structure of DNA by Watson and Crick

1977: Sanger method for DNA sequencing
1978: DNA sequenced ΦX174 5,386 nt

1995: Genome sequence of Haemophilus influenza: first bacteria to be sequenced (Craig Venter group): 1.8M bases

2001: First version of the human genome : 3.2 billion of bases



Livestock genomes: a younger history

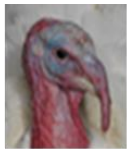
2001
Genomic selection proposed

2002
Mouse draft genome sequence

2003
Human genome sequence "finished"
$3 billion

2004
Chicken (Red Jungle Fowl) genome sequence

2008
Human 1000 Genomes Project launched



2010
Turkey genome sequenced

2009
Cattle genome sequenced

Horse genome Sequenced

Mouse genome "finished"

2012
Pig genome sequenced

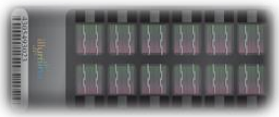
2003
ENCODE (1%) launched

2007
ENCODE genome-wide

2008
Bovine 50K SNP chip

2009
Pig 60K SNP chip

Sheep 60K SNP chip



2010
750K bovine SNP chips

Book of life – How do you read it ?

Book of life – How do you read it ?

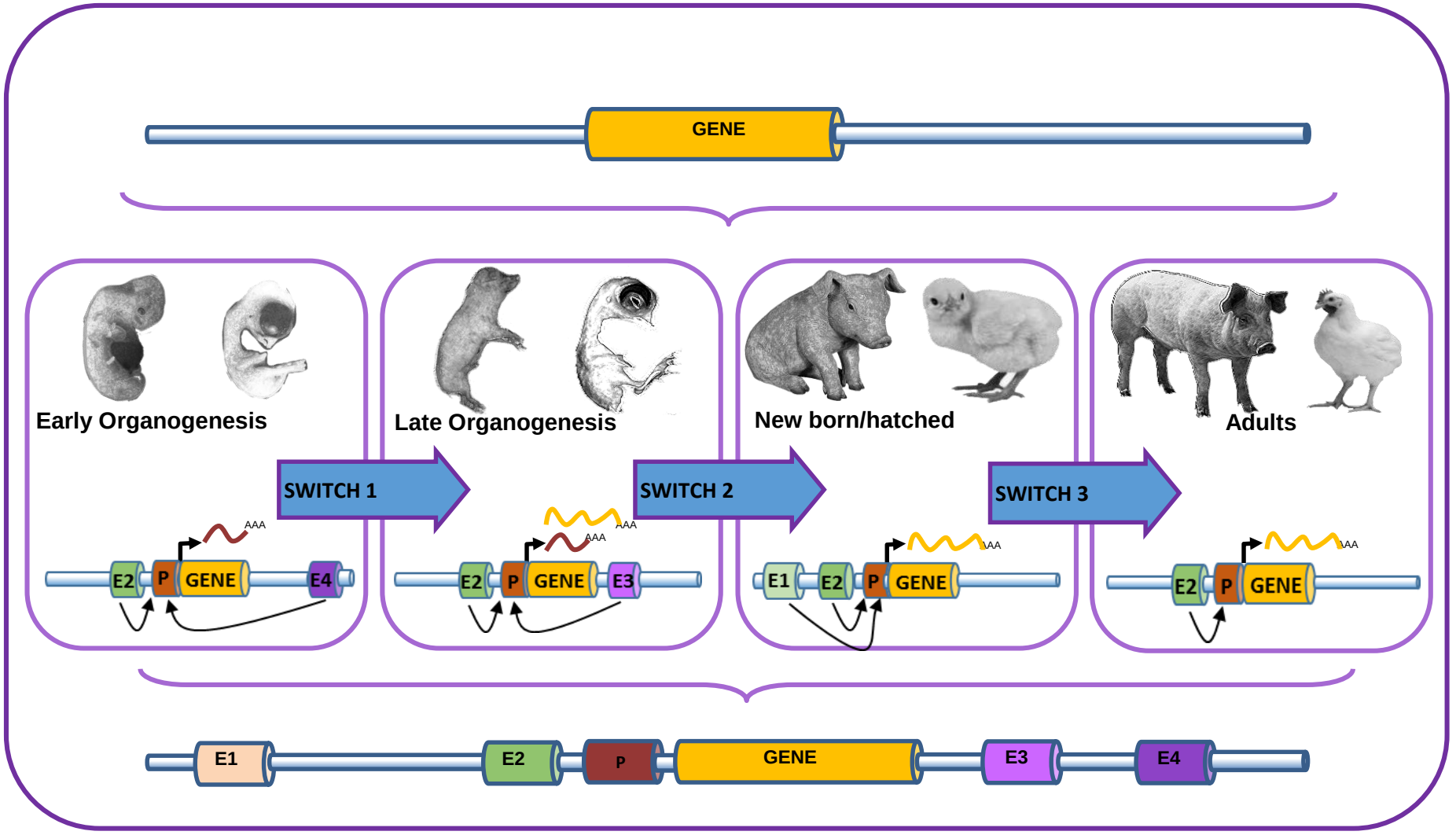
DANSLAPREMIEREPARTIEDECELIVRECOMBRAYLENARRATEUREVOQUELESSEJOURSPASSE
SDANSLAMAISONDETANTELEONIEALORSQUILETAITENFANTILSESOUVIENTAVECNOSTALGI
EDUBAISERDUSOIRDESAMEREBAISERTANTATTENDUMAISPARFOISRETARDEPARUNINVITES
OUVENTMSWANNILNOUSFAITCONNAÎTREDESPERSONNESDESONENTOURAGESATANTELEO
NIEMALADEGARDANTTOUJOURSLACHAMBRESAGRANDMEREUNPEUFANTASQUEQUIAIME
SEPROMENERSOUSLAPLUIEFRANÇOISELAFIDELECUISINIERELESHABITANTSDUVILLAGEI
LEVOQUESESGOÛTSPOURLALECTURELESLONGUESPROMENADESAVECSESPARENTSDUCÔ
TEDECHEZSWANNOUDEGUERMANTESLASECONDEPARTIEUNAMOURDESWANNSEDEROUL
EQUELQUESANNEESAVANTLANAISSANCEDUNARRATEURCHARLESSWANNRICHECOLLECTI
ONNEURDOBJETSDARTVAFINIRPARCEDERAUXAVANCESPASTOUTÀFAITDESINTERESSEESD
ODETTEDECRECYDEMIMONDAINEQUILEFERADAILLEURSBEAUCOUPSOUFFRIRONDECO
UVRELESALONDESVERDURINFREQUENTEPARDENOMBREUXPERSONNAGESQUIFIGURER
ONTTOUTAULONGDELŒUVRECOTTARDSANIETTEBRICHOTFORCHEVILLEETBIENDAUTRE
SLASSEPARLESNOMBREUSESINFIDELITESDODETTESWANNRECOUVREENFINSALIBERTESE
TONNANTDAVOIRETEAMOUREUXDUNEFEMMEQUILNAJAMAISVRAIMENTAIMEE

Book of life – How do you read it ?

Dans la première partie de ce livre, Combray, le narrateur évoque les séjours passés dans la maison de tante Léonie alors qu'il était enfant. Il se souvient avec nostalgie du baiser du soir de sa mère, baiser tant attendu, mais parfois retardé par un invité, souvent M. Swann. Il nous fait connaître des personnes de son entourage, sa tante Léonie, malade gardant toujours la chambre, sa grand-mère un peu fantasque qui aime se promener sous la pluie, Françoise la fidèle cuisinière, les habitants du village. Il évoque ses goûts pour la lecture, les longues promenades avec ses parents, du côté de chez Swann ou de Guermantes.

La seconde partie, Un amour de Swann, se déroule quelques années avant la naissance du narrateur. Charles Swann, riche collectionneur d'objets d'art va finir par céder aux avances pas tout à fait désintéressées d'Odette de Crécy, demi-mondaine, qui le fera d'ailleurs beaucoup souffrir. On découvre le salon des Verdurin, fréquenté par de nombreux personnages qui figureront tout au long de l'œuvre : Cottard, Saniette, Brichot, Forcheville et bien d'autres. Lassé par les nombreuses infidélités d'Odette, Swann, recouvre enfin sa liberté, s'étonnant d'avoir été amoureux d'une femme qu'il n'a jamais vraiment aimée.

Book of life – How do you read it ?



First annotation on the human genome

- **80.4%** participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type
- promoter functionality can explain most of the variation in RNA expression
- SNPs associated with disease by GWAS are enriched within non-coding functional elements

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

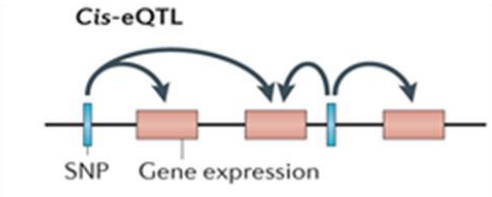
The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.

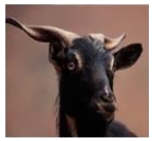


>$250 million

Then annotation on livestocks' genomes

2012
ENCODE

2012
Pig genome sequenced



2013
Goat genome sequenced

2013
Duck genome sequenced

2014
Sheep genome sequenced

2014
Functional Annotation of Animal Genomes (FAANG) launched

2014-17
FR-AgENCODE pilot project

2016
EC Workshop on animal genomics and breeding for sustainable production

2016
FAANG-Europe
COST Action

2012
Chicken 600K SNP chip

2012
AgENCODE (later FAANG) Conceived
EU-US ABWG, ISAG

2013 onwards
Genotype-by-sequence

2014
Salmon SNP chip

2015
Pig 650K SNP chip

2015 onwards
LCseq for genomic selection
SNPs impute to sequence

Fish genomes: Tilapia, Cod, Salmon,……

Then annotation on livestocks' genomes

2016
EC Workshop on animal genomics and breeding for sustainable production

2016-2020
FAANG-Europe
COST Action

2017
EC Horizon 2020
Work Programme 2018-2020
SFS-30-2018-2019-2020: Agri-Aqua Labs

2019
H2020 FAANG/AquaFAANG projects start
AQUA-FAANG
BovReg
GENE-SWitCH

~€18 million
9 species

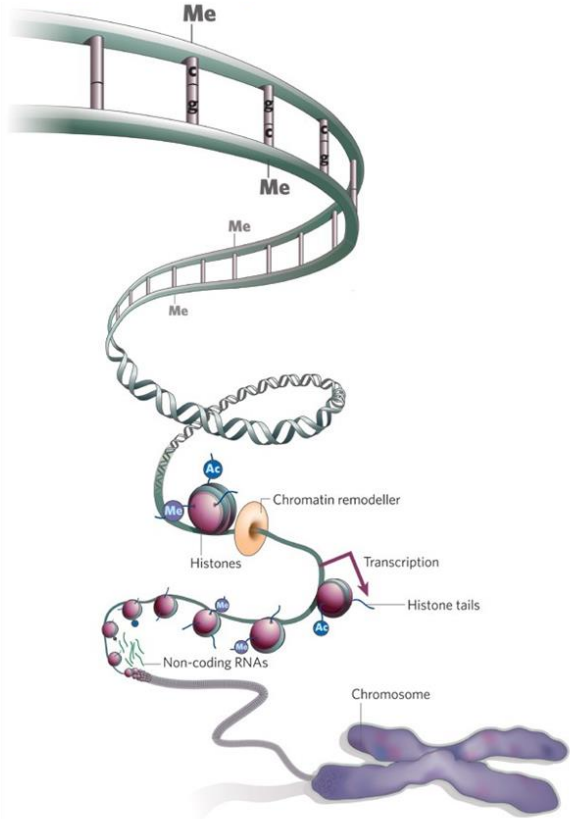
2017
Improved pig genome sequence

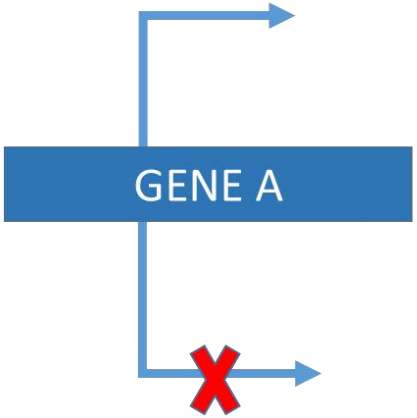
2021
Chicken (broiler and layer) genome sequences

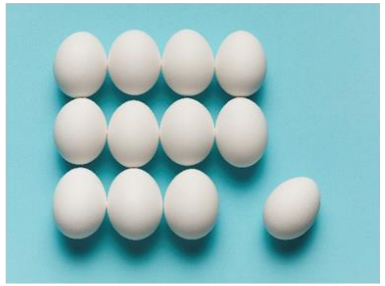
2023 onwards
Telomere-toTelomere (T2T) genome assemblies
Pangenomes

Why studying the functional genome?

Regulatory elements → **Gene expression** → **Traits of interest**

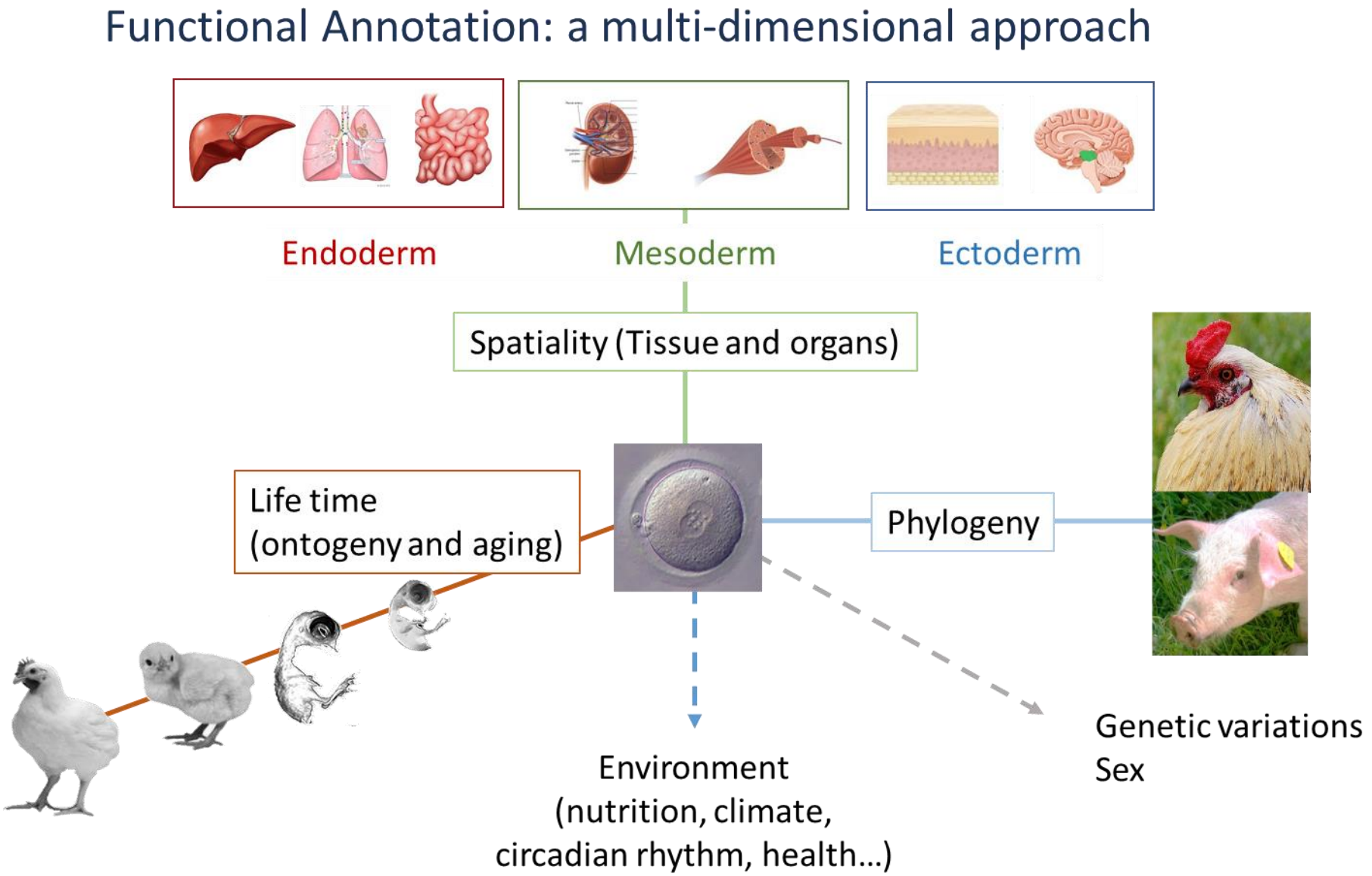






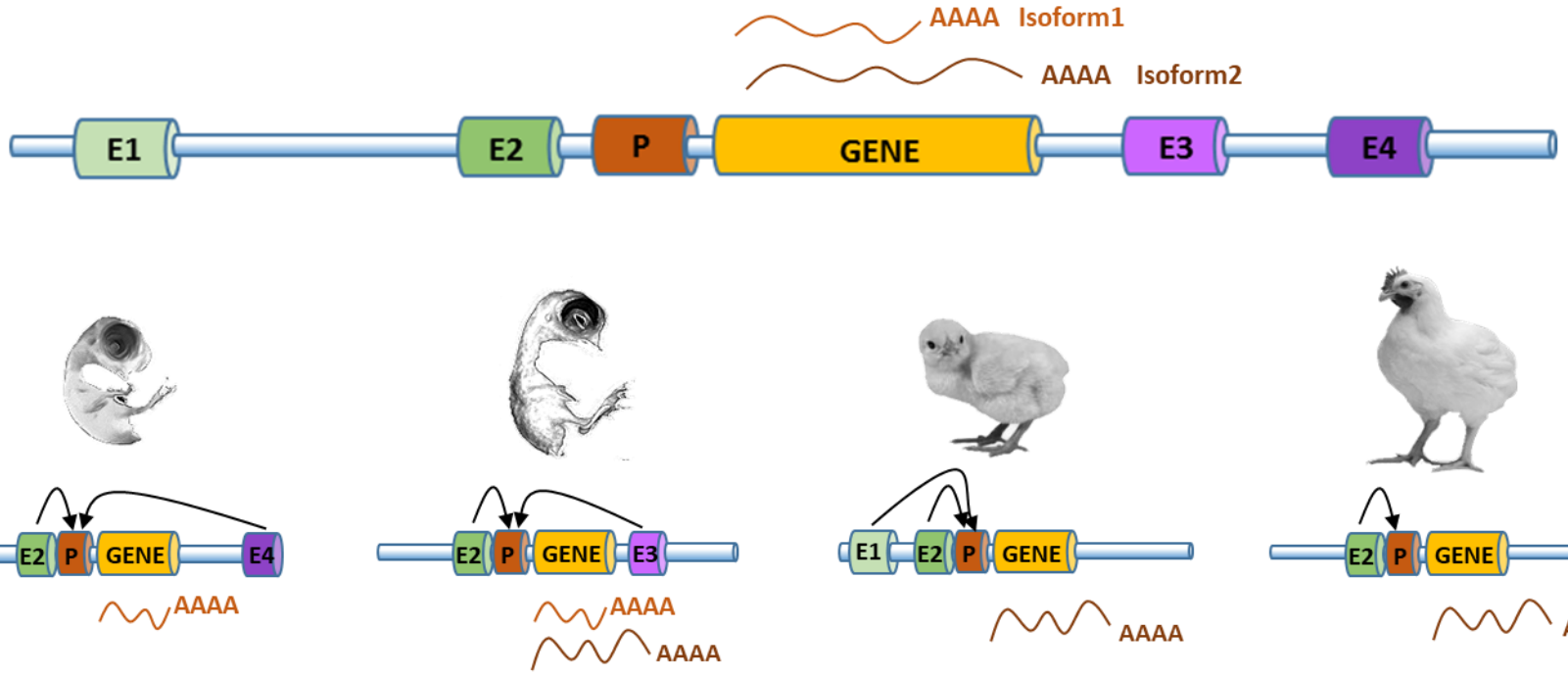
The American Association for Cancer Research Human Epigenome Task Force., European Union, Network of Excellence, Scientific Advisory Board. Moving AHEAD with an international human epigenome project. *Nature* **454,** 711–715 (2008). https://doi.org/10.1038/454711a

Why studying the functional genome?



Why studying the functional genome?

Functional Annotation: how are all these dimensions integrated to regulate genome function ?



Annotation of:
E2, E4, P, I1

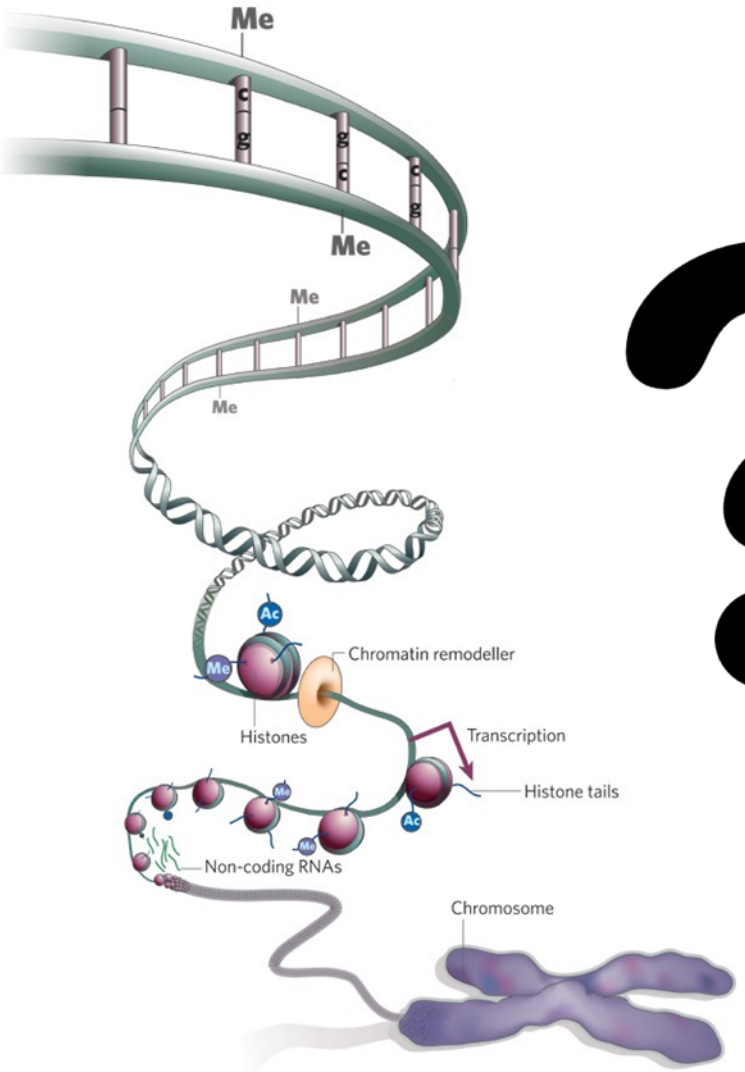
Annotation of:
E2, E3, P, I1, I2

Annotation of:
E1, E2, P, I2

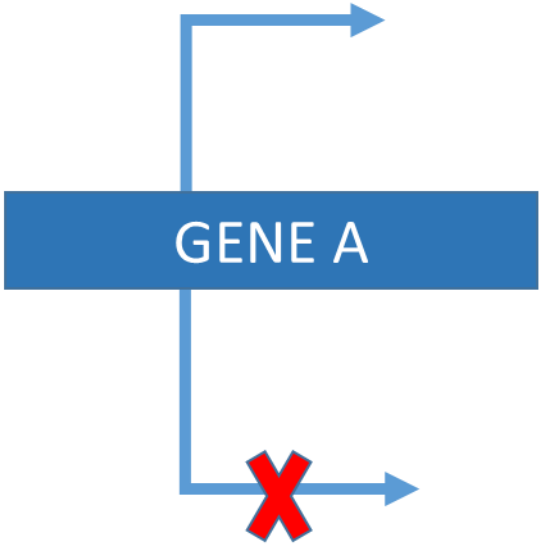
Annotation of:
E2, P, I2

How studying the functional genome?

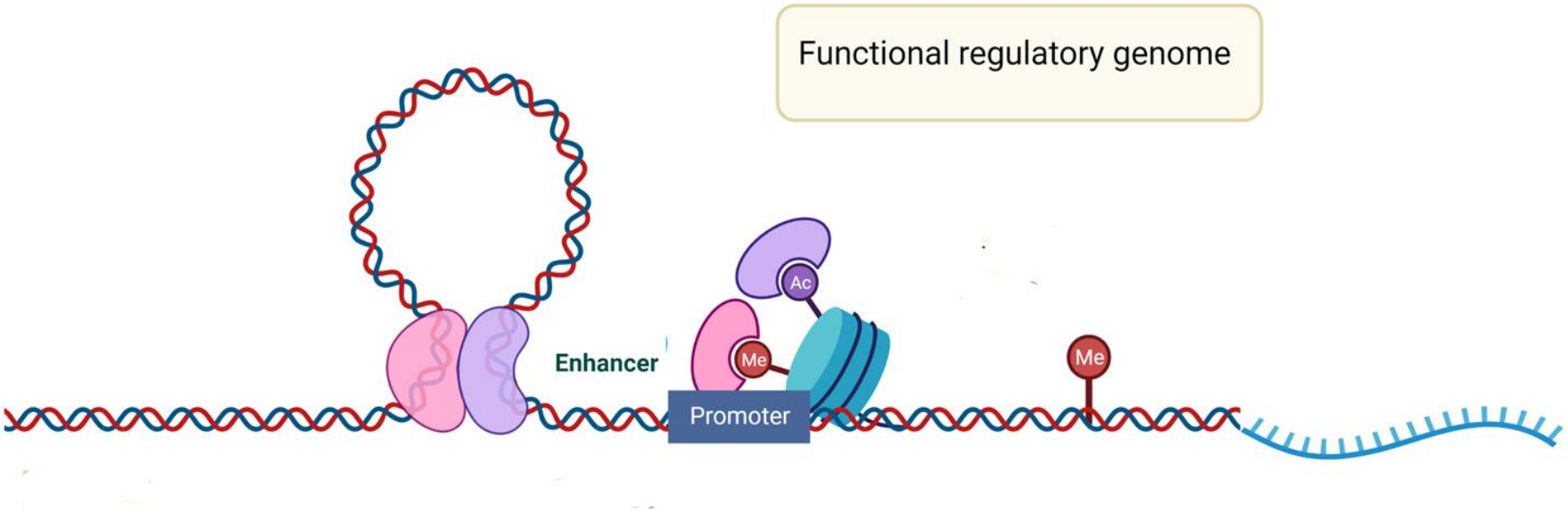
Regulatory elements → **Gene expression**



?

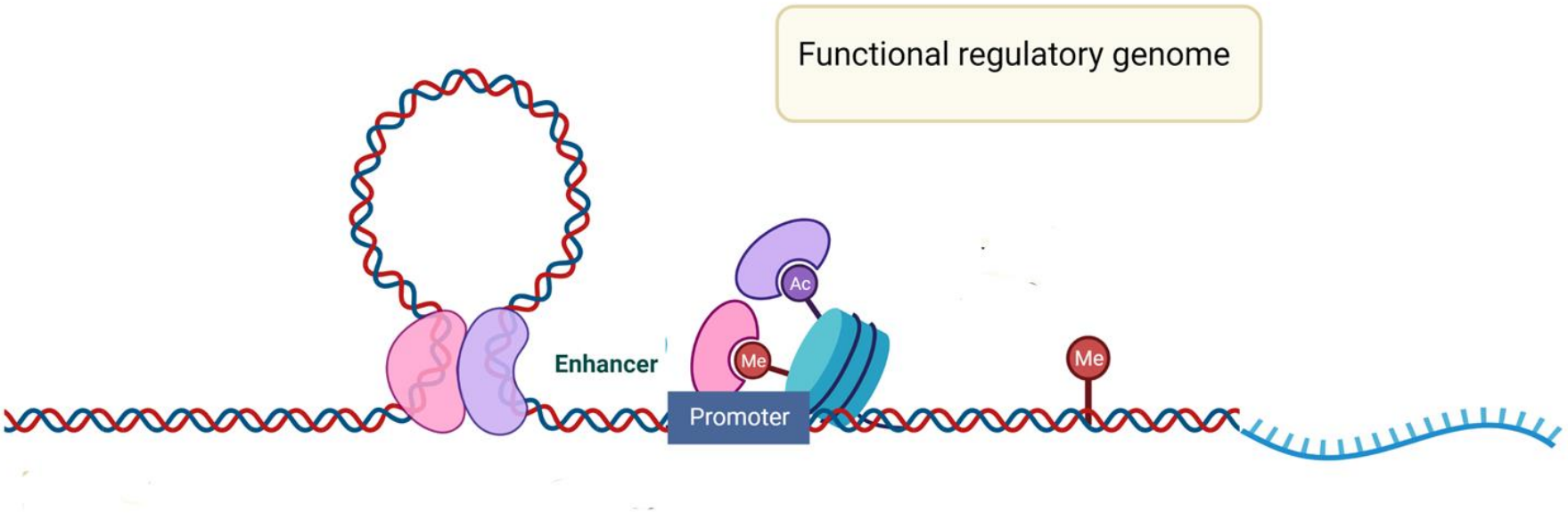


How studying the functional genome?

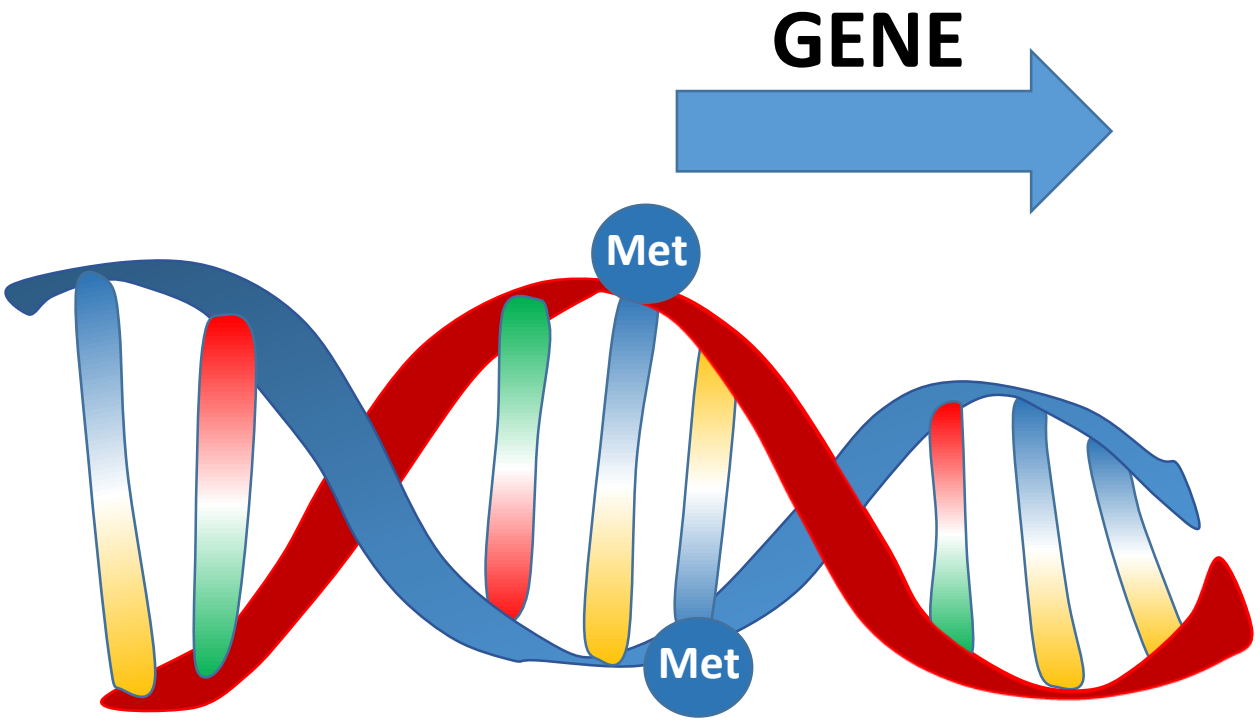


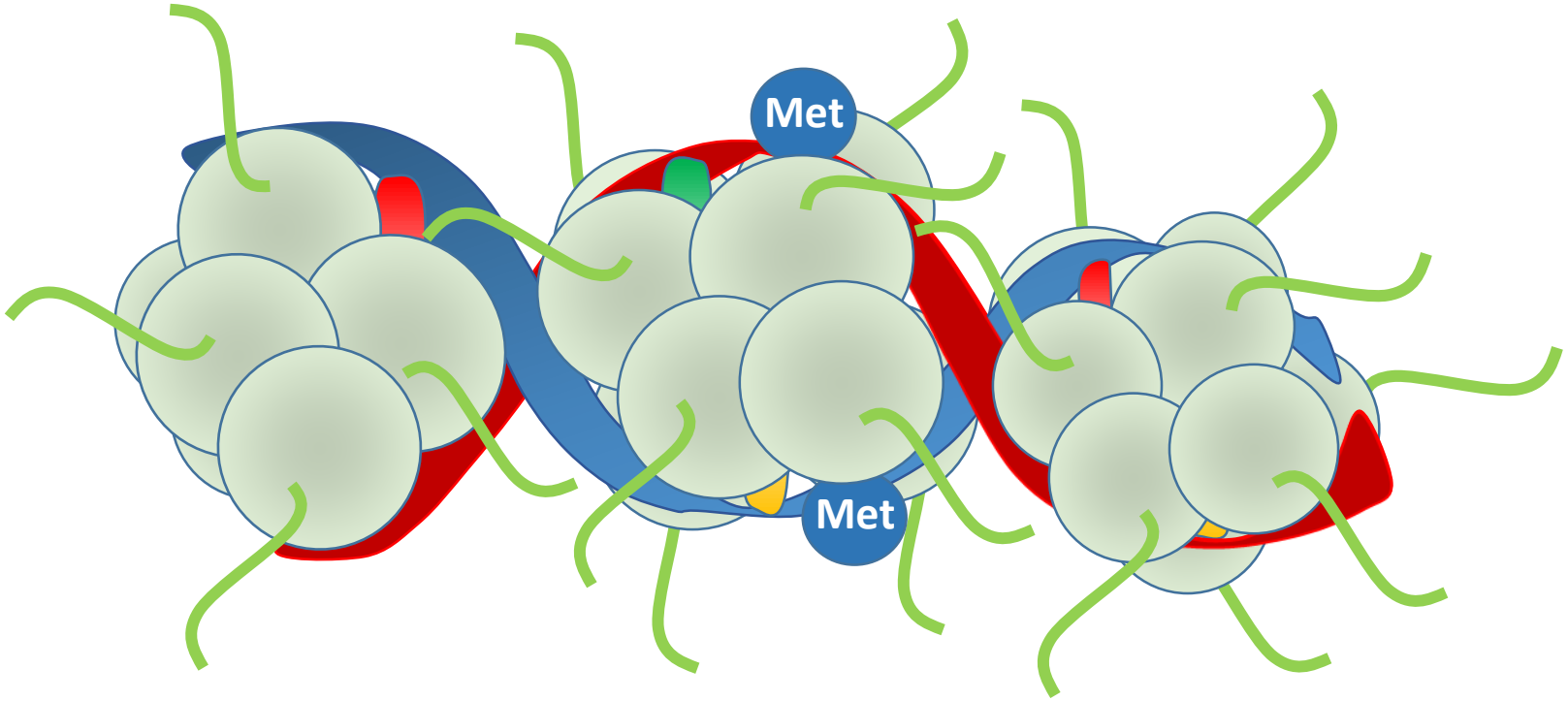
How studying the functional genome?

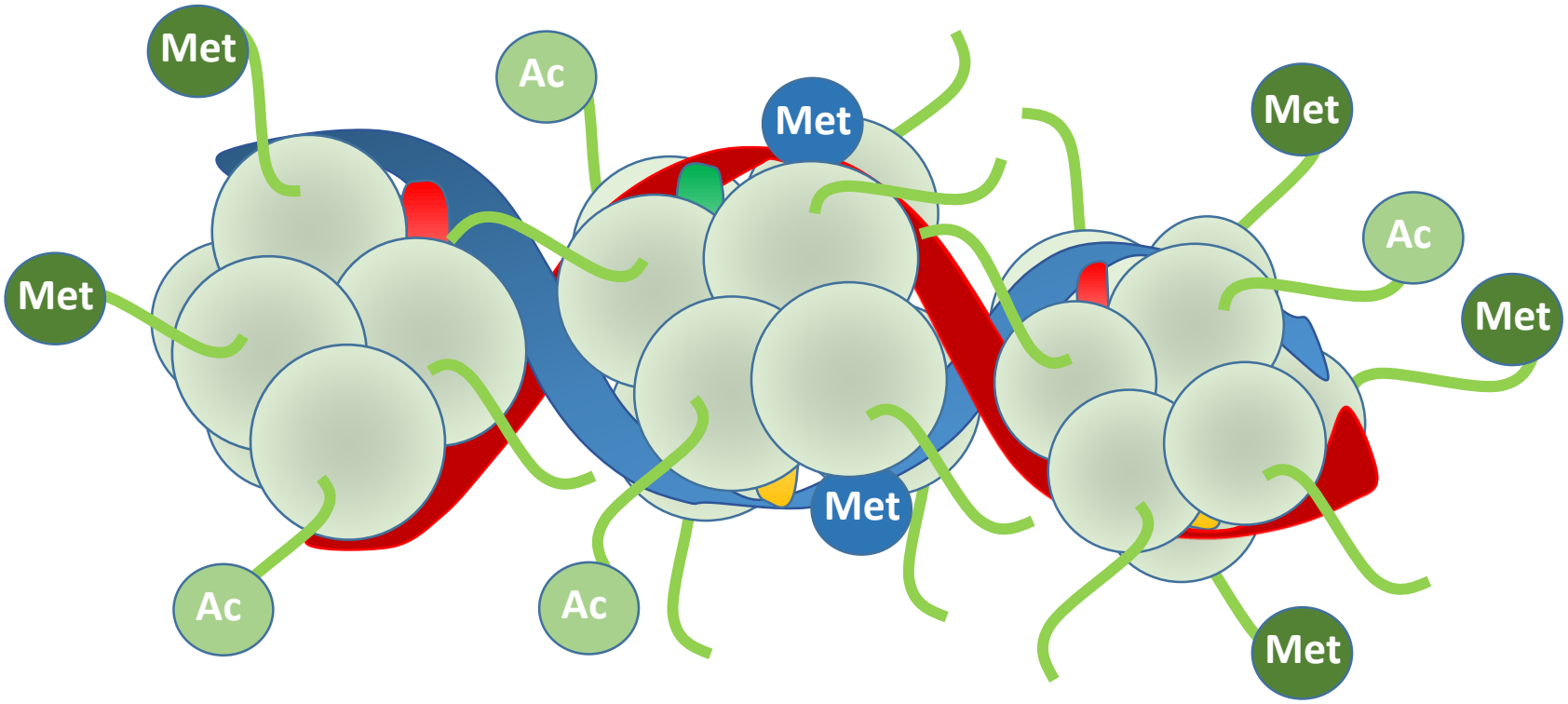
- Gene expression
- Epigenetic information

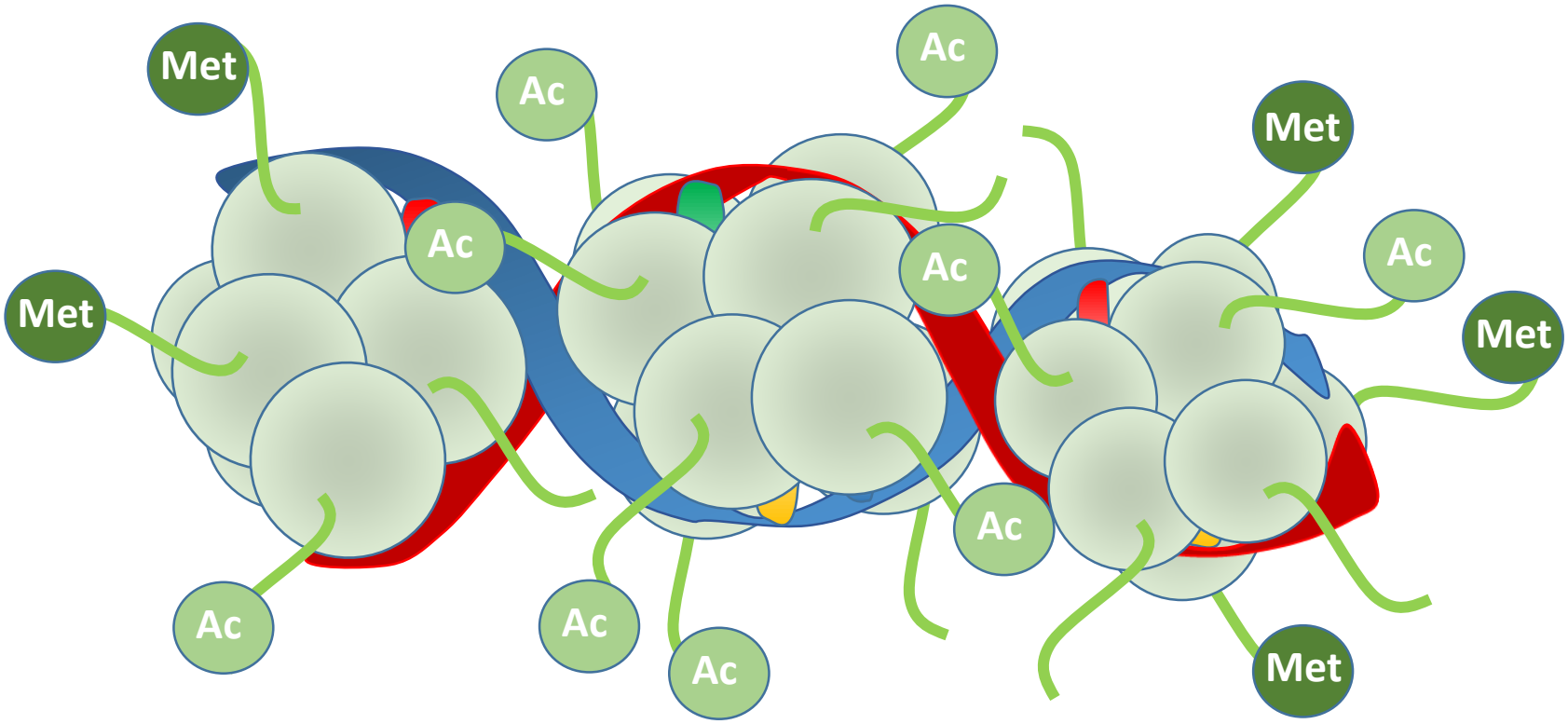


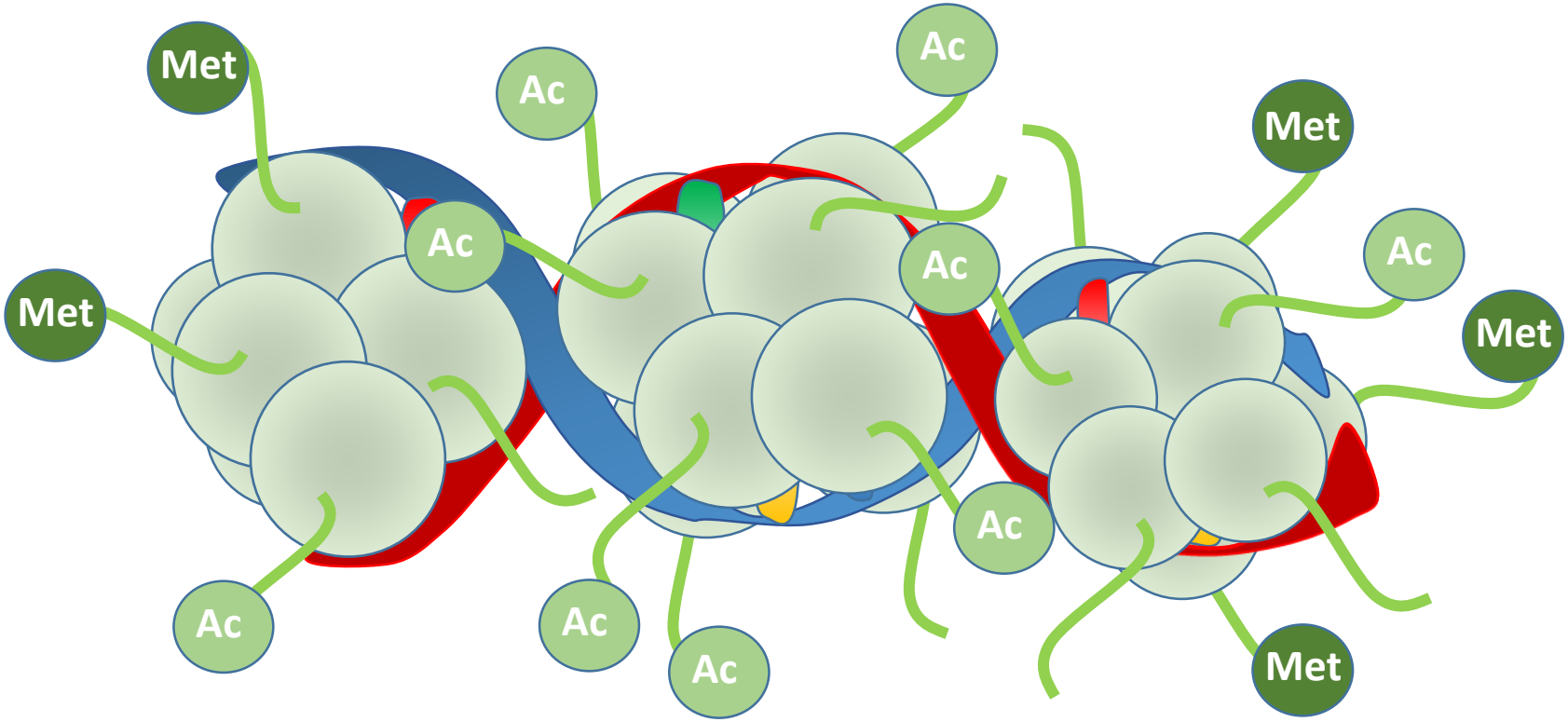
What is epigenetics?

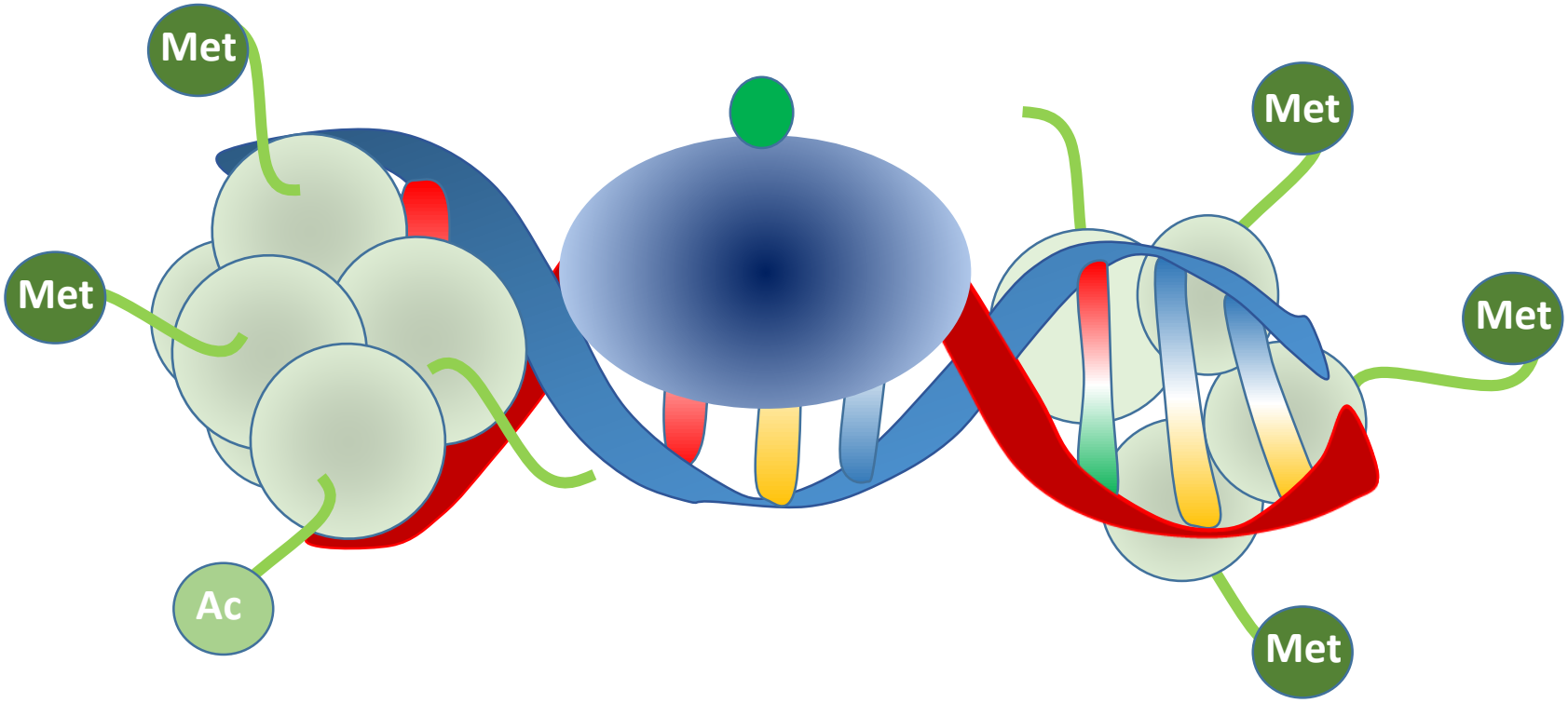


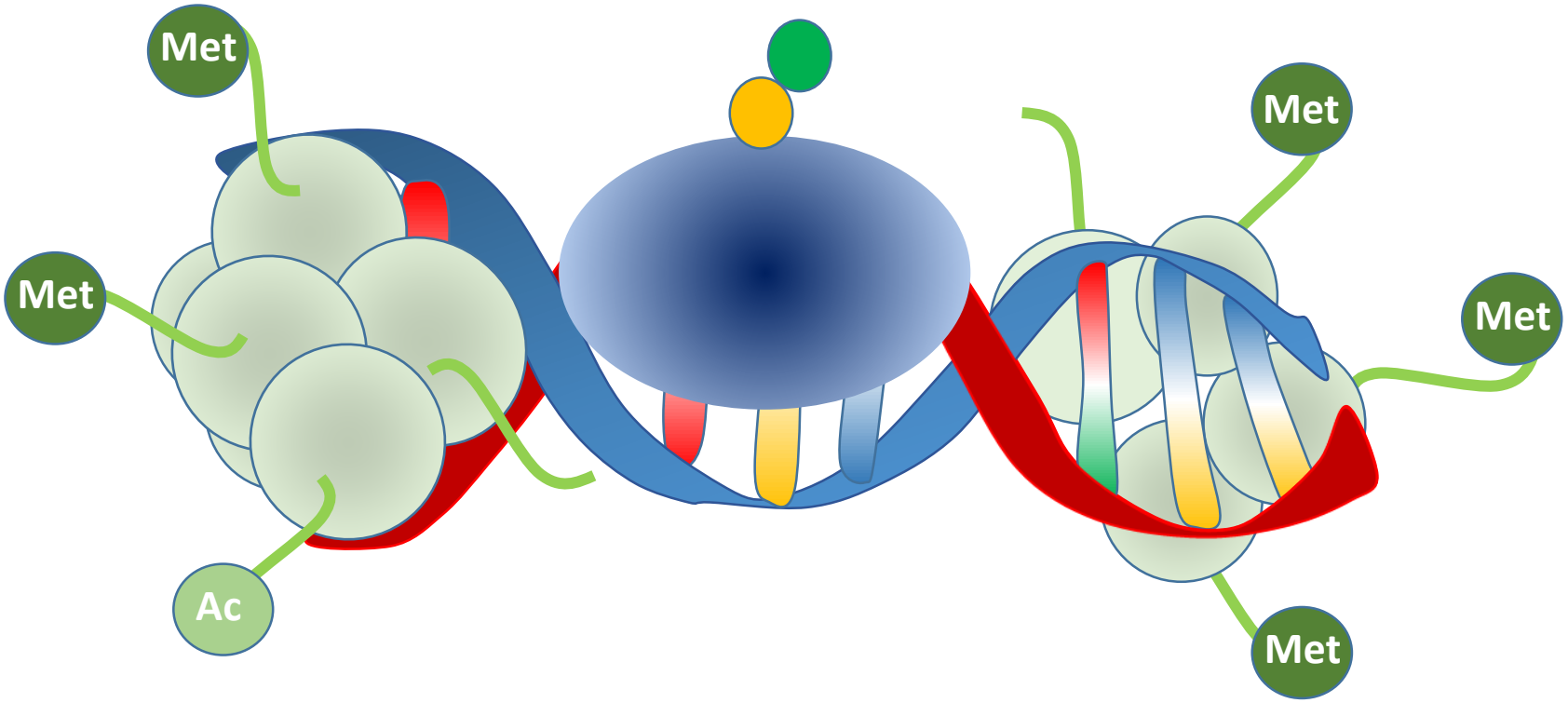

Met
Met

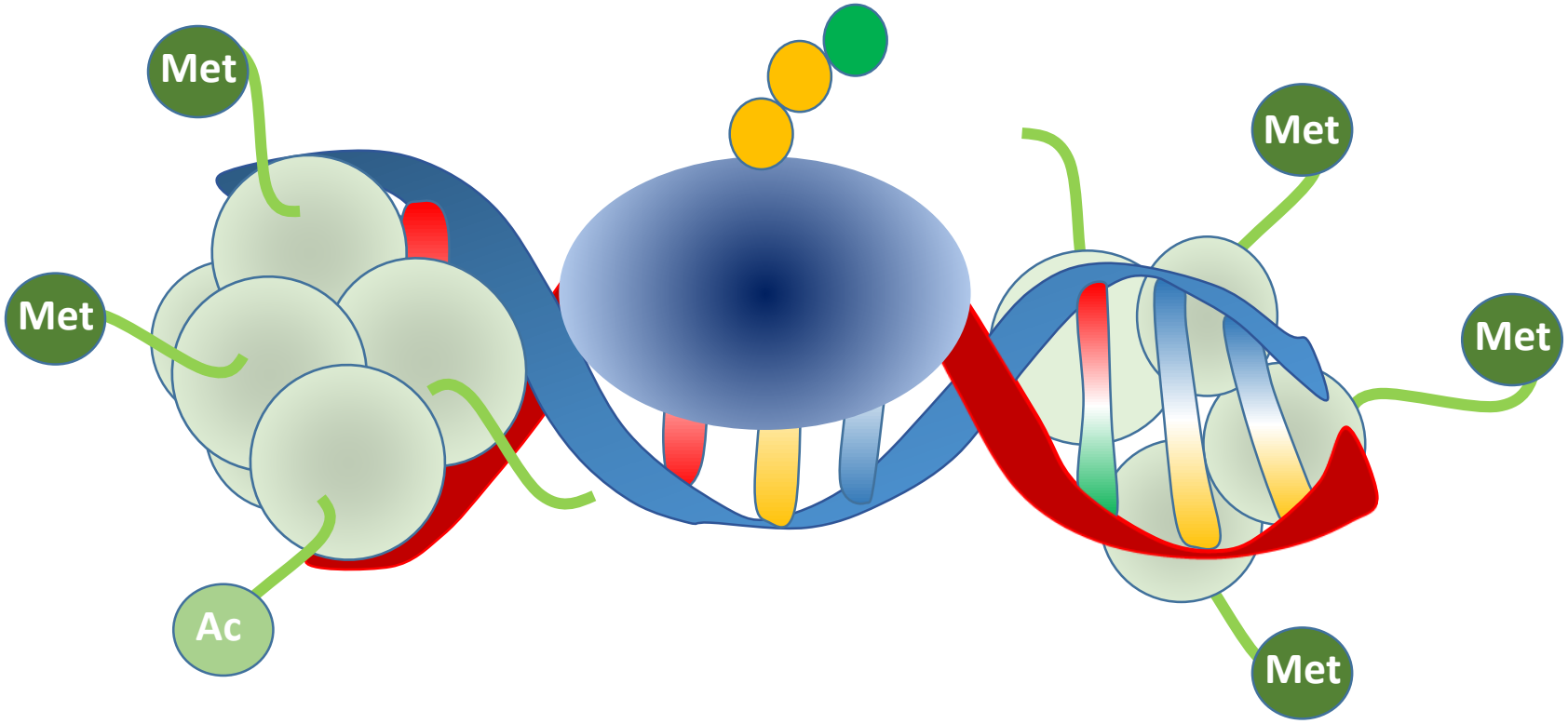

Met
Ac
Met
Met
Met
Ac
Met
Met
Ac
Ac
Met

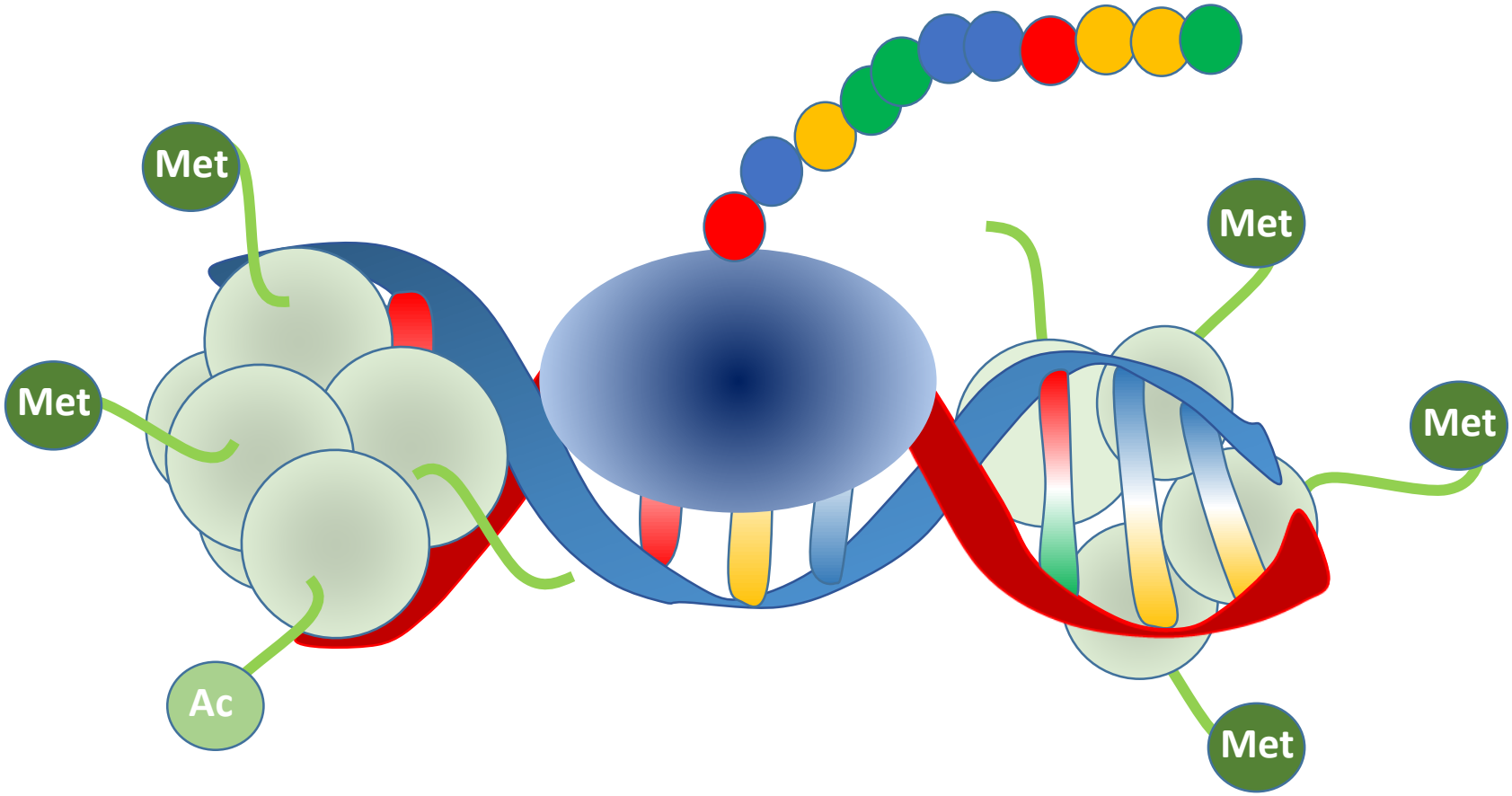

Met
Ac
Ac
Ac
Met
Met
Ac
Ac
Met
Ac
Ac
Ac
Ac
Met

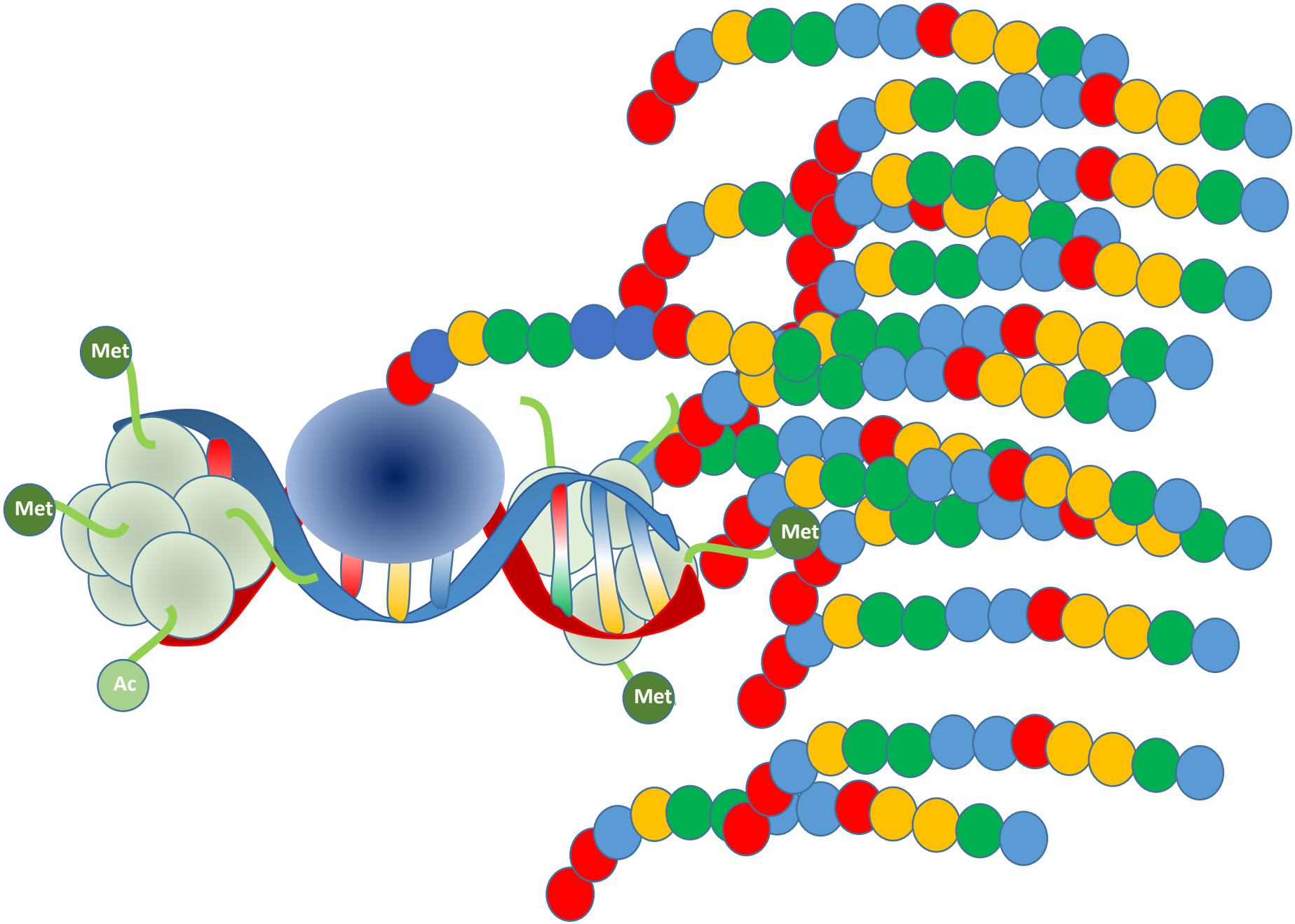

Met
Ac
Ac
Ac
Met
Met
Ac
Ac
Met
Ac
Ac
Ac
Ac
Met


Met
Met
Met
Met
Met
Ac

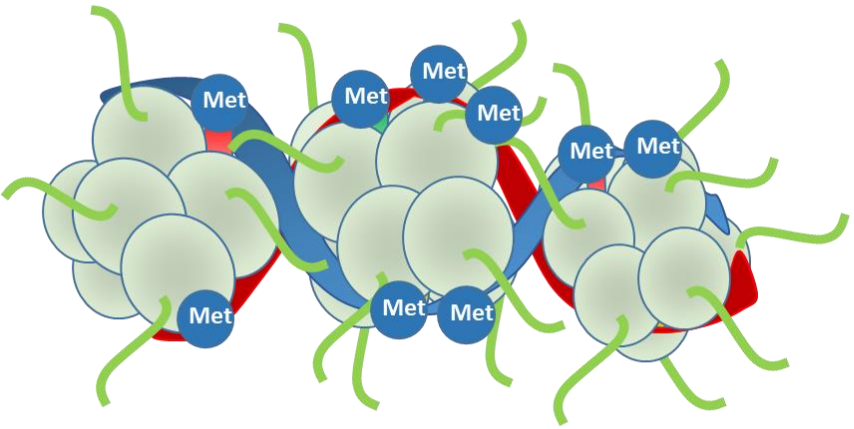

Met
Met
Ac
Met
Met
Met


Met
Met
Met
Met
Met
Ac


Met
Met
Ac
Met
Met
Met

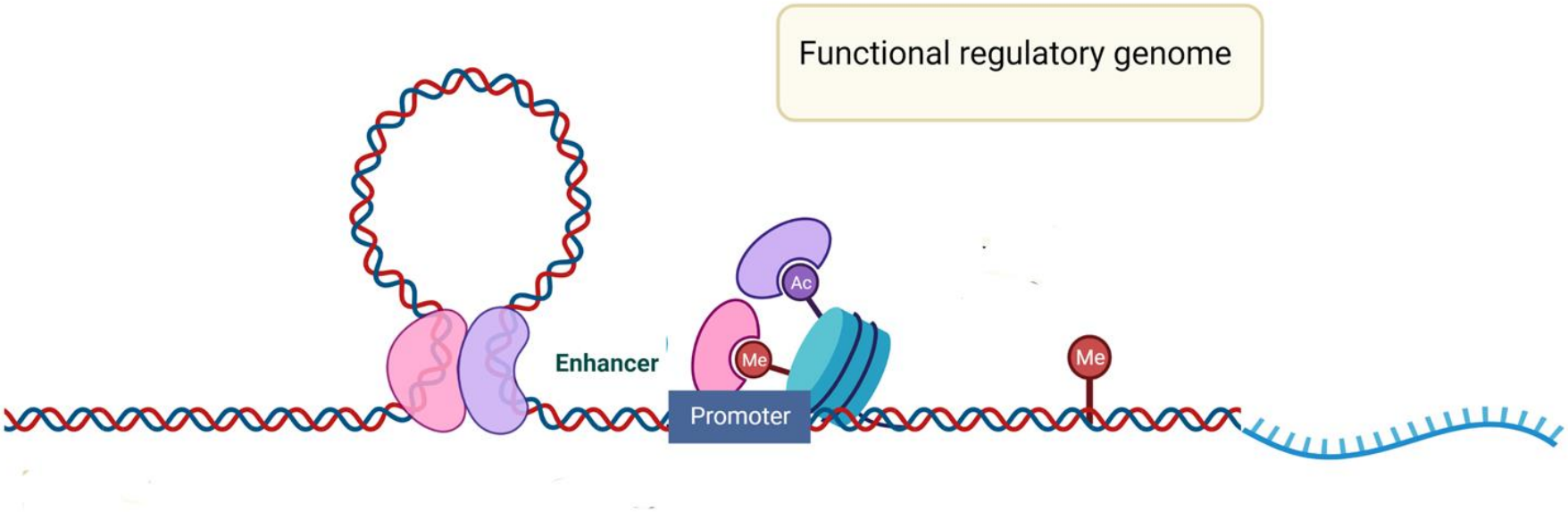

Met
Met
Ac
Met
Met

Epigenetics modulates gene activity and phenotypes



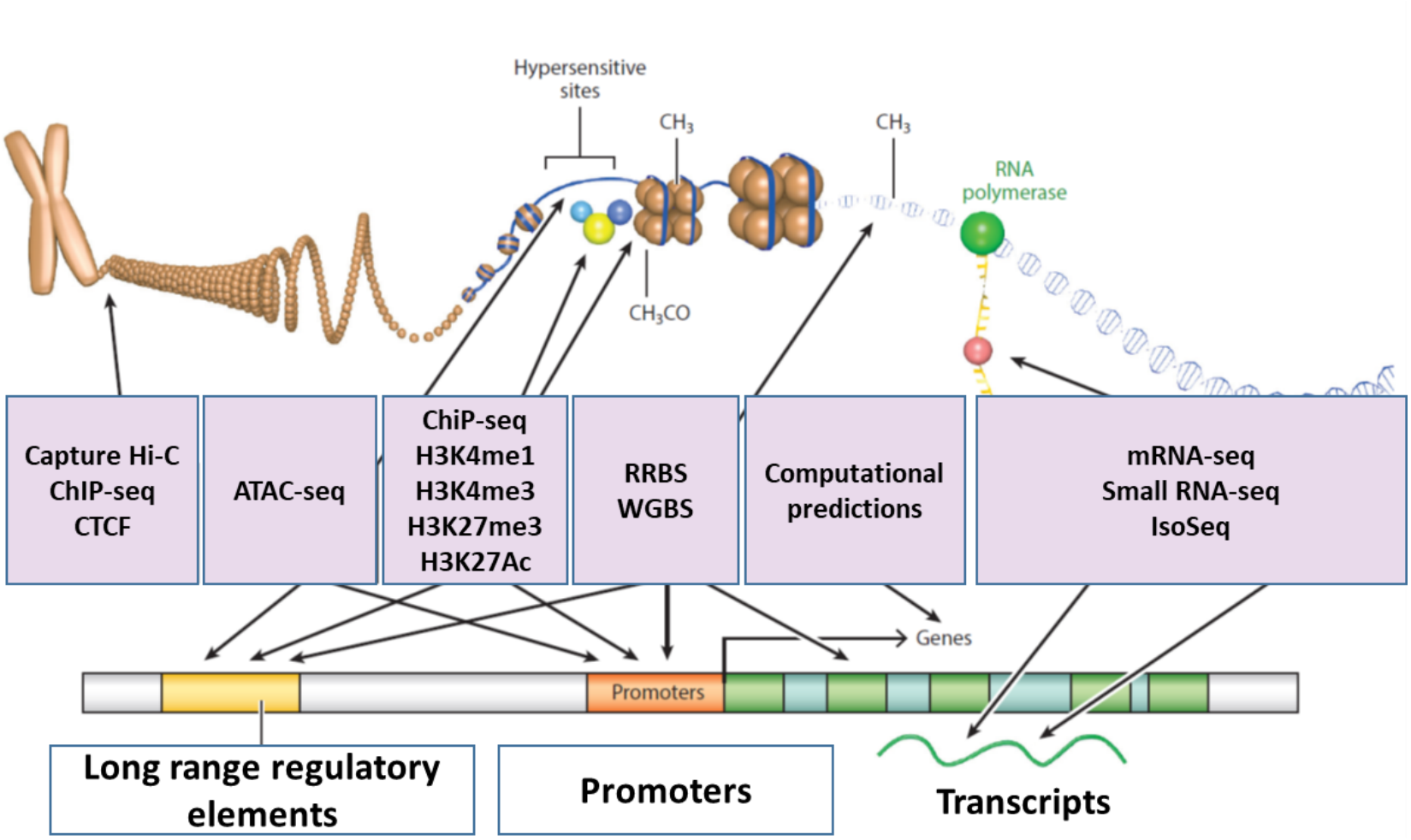
Epigenetic regulation of the *agouti* gene modulates coat color in mice

How studying the functional genome?

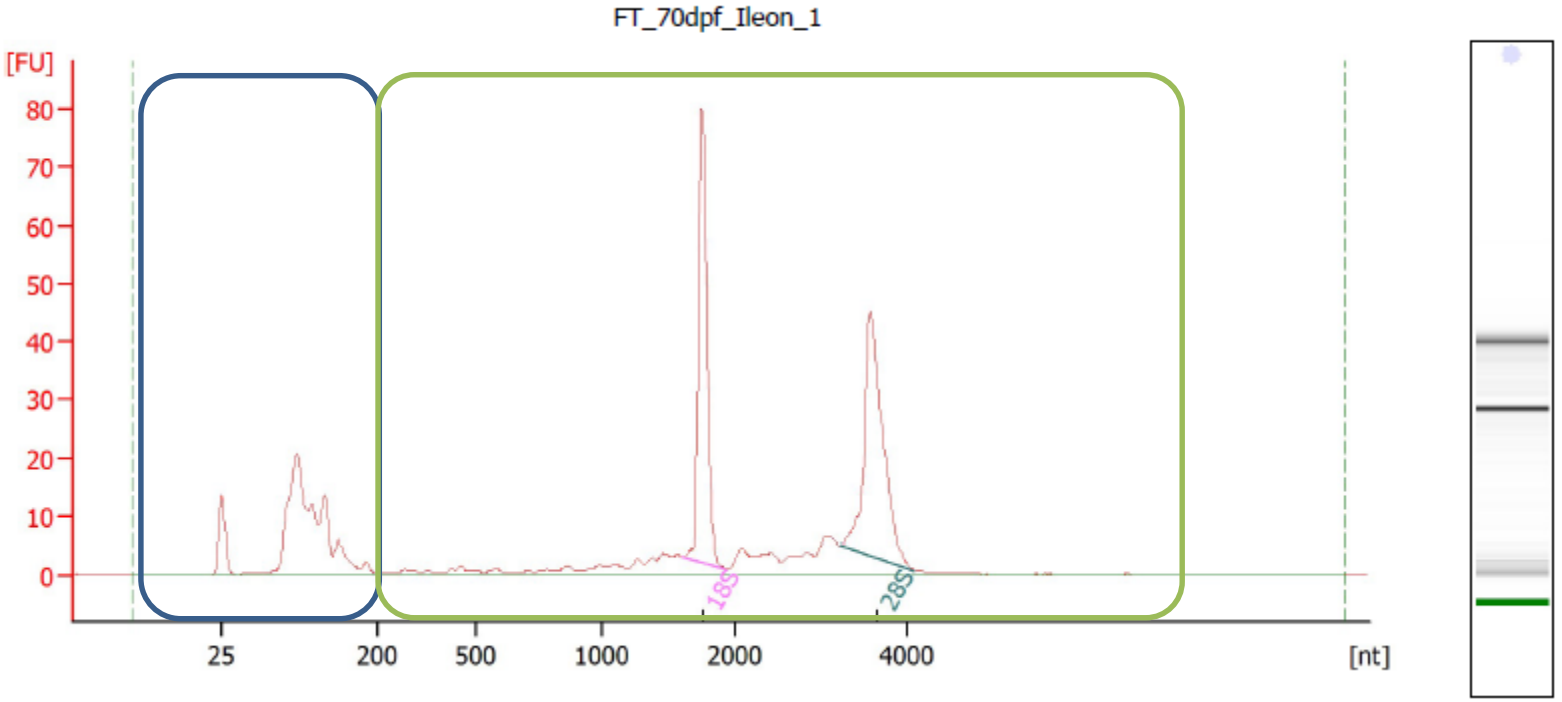


How studying the functional genome?

By using a set of complementary molecular assays

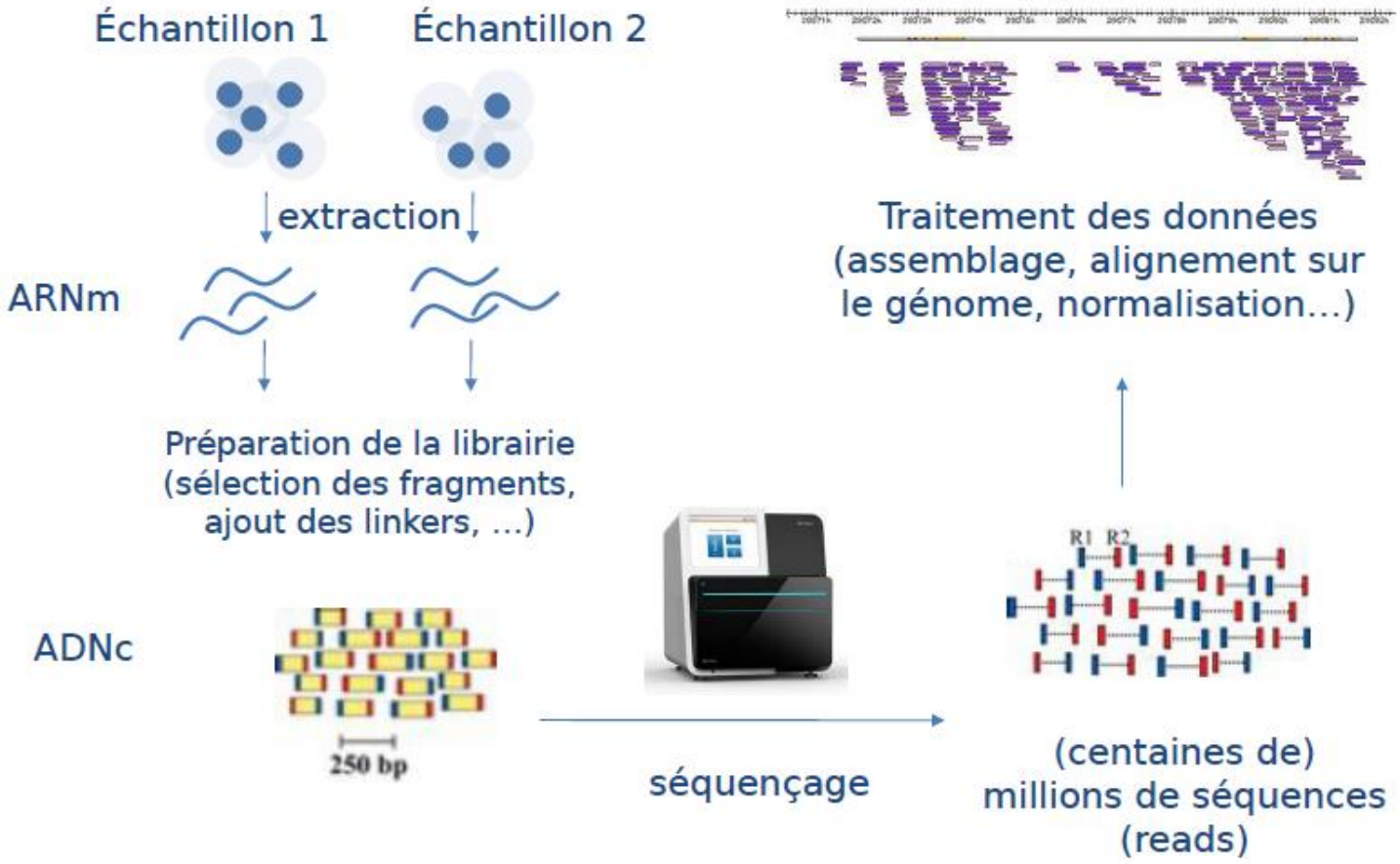


The different assays used for annotation: RNA-seq

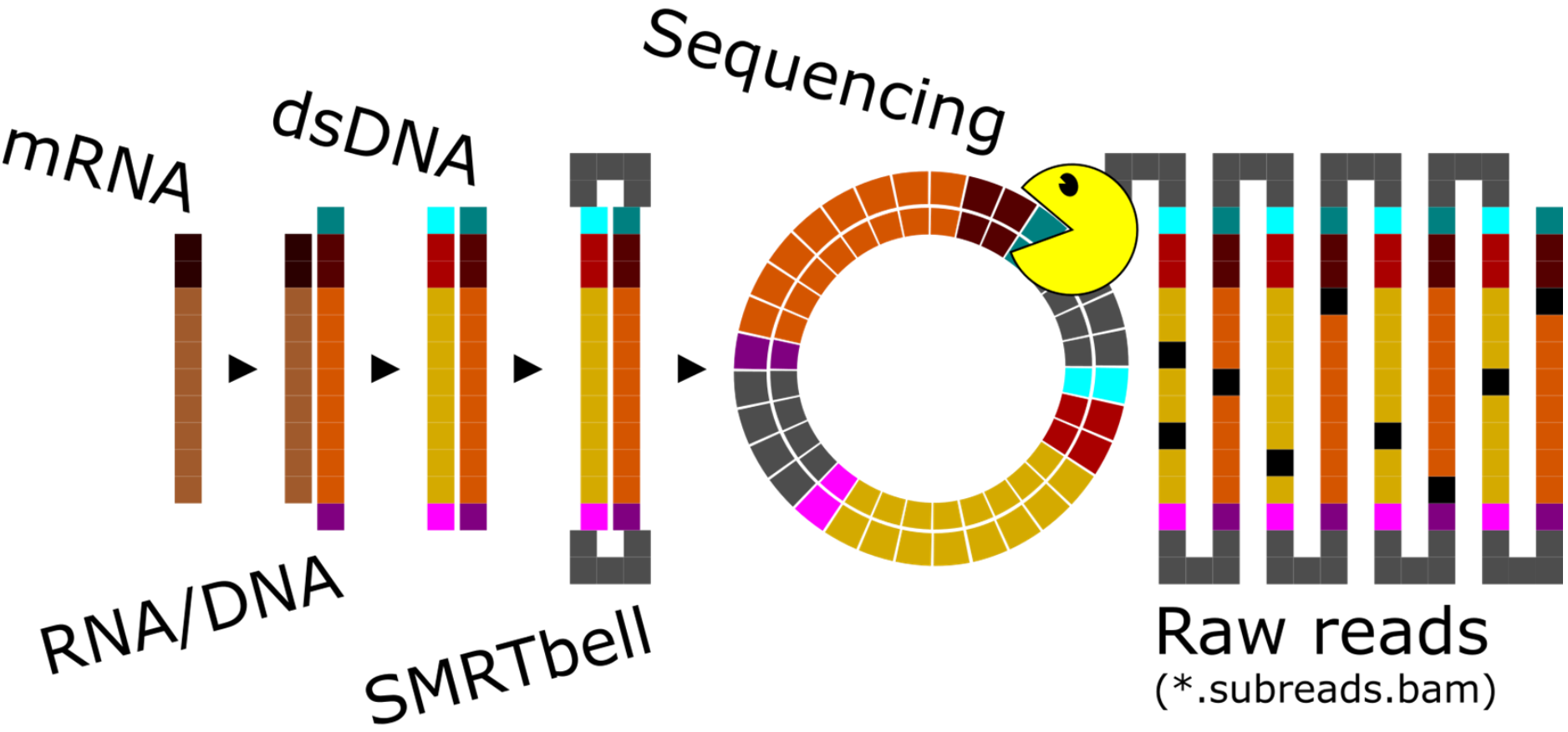


smallRNAseq Long-RNAseq

The different assays used for annotation: mRNA-seq

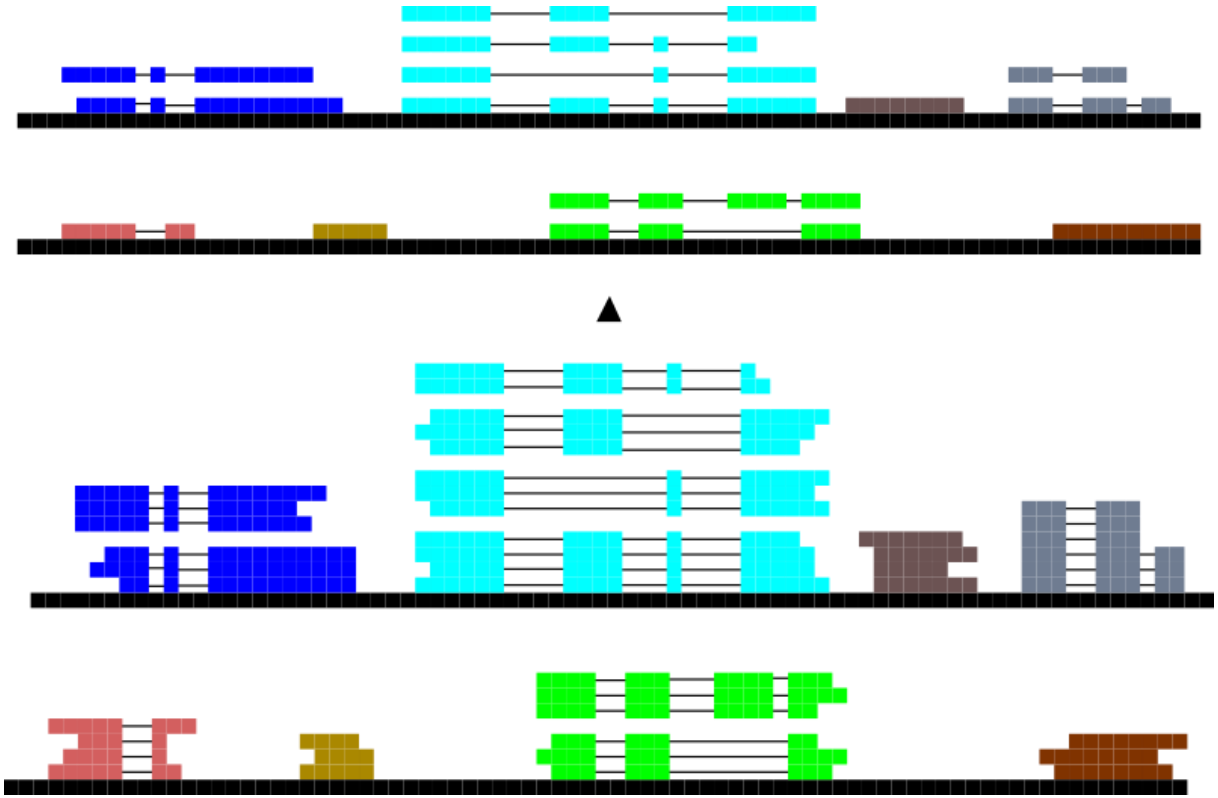


The different assays used for annotation: iso-seq



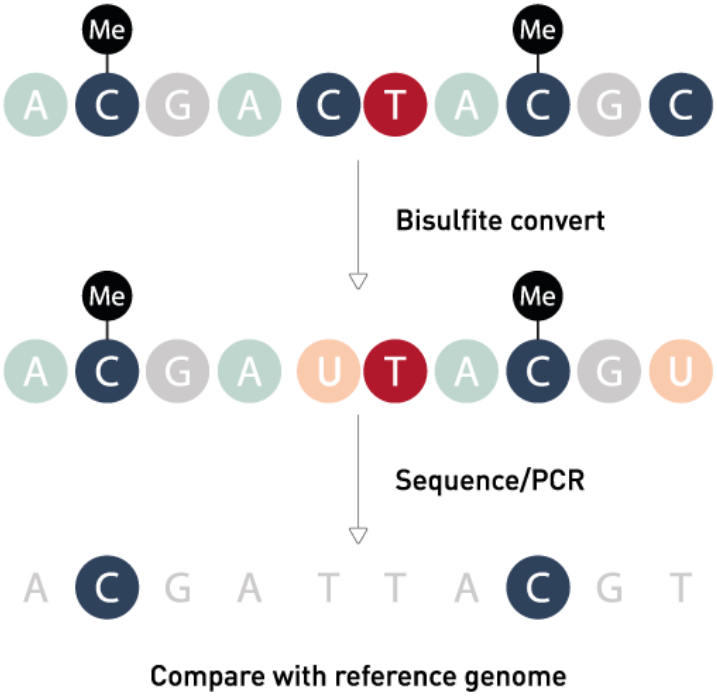
The different assays used for annotation: RNA-seq

Repertoire of transcripts



The different assays used for annotation: RRBS/WGBS

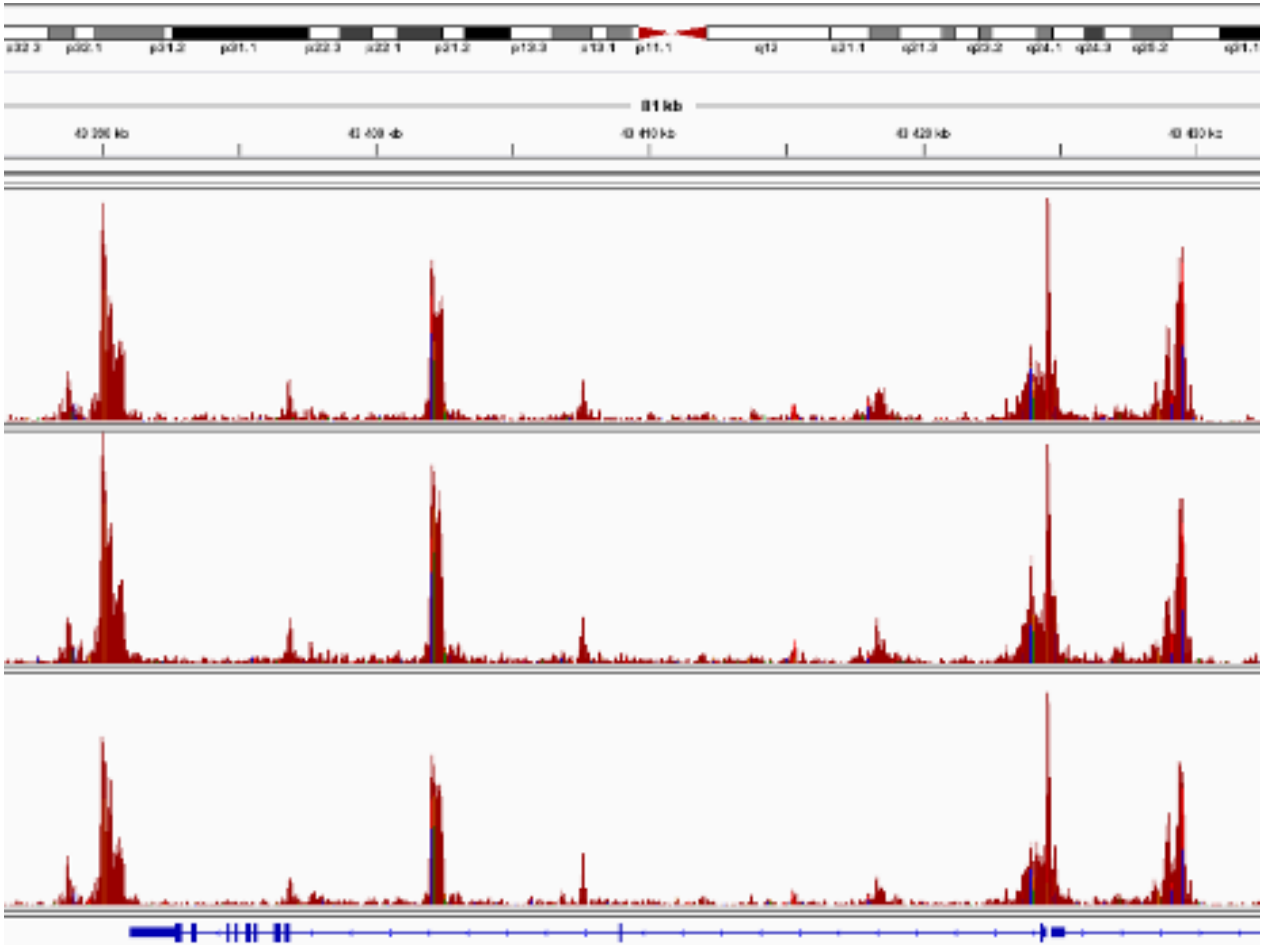
Chemical conversion of unmethylated cytosine by bisulfite treatment



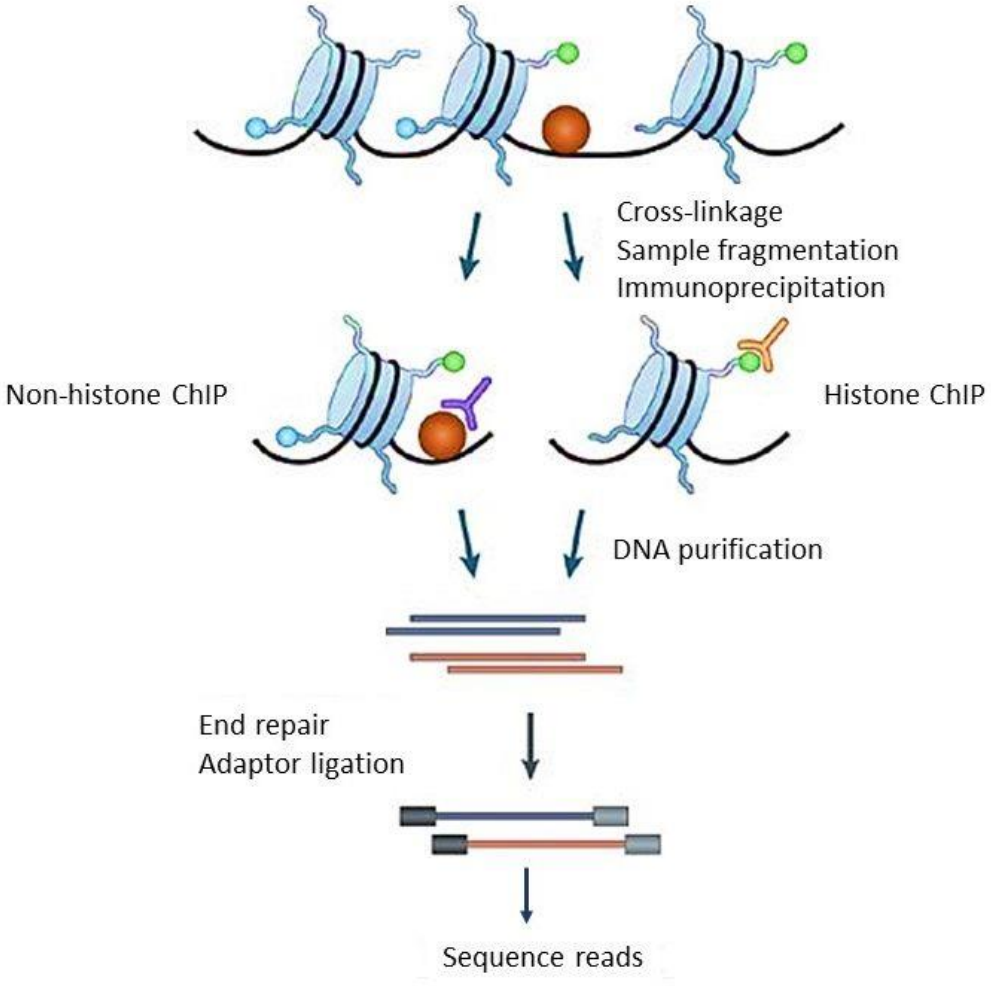
CpG 1: 5%
CpG 2: 10%
…
CpG n: 98%

The different assays used for annotation: ATAC-seq

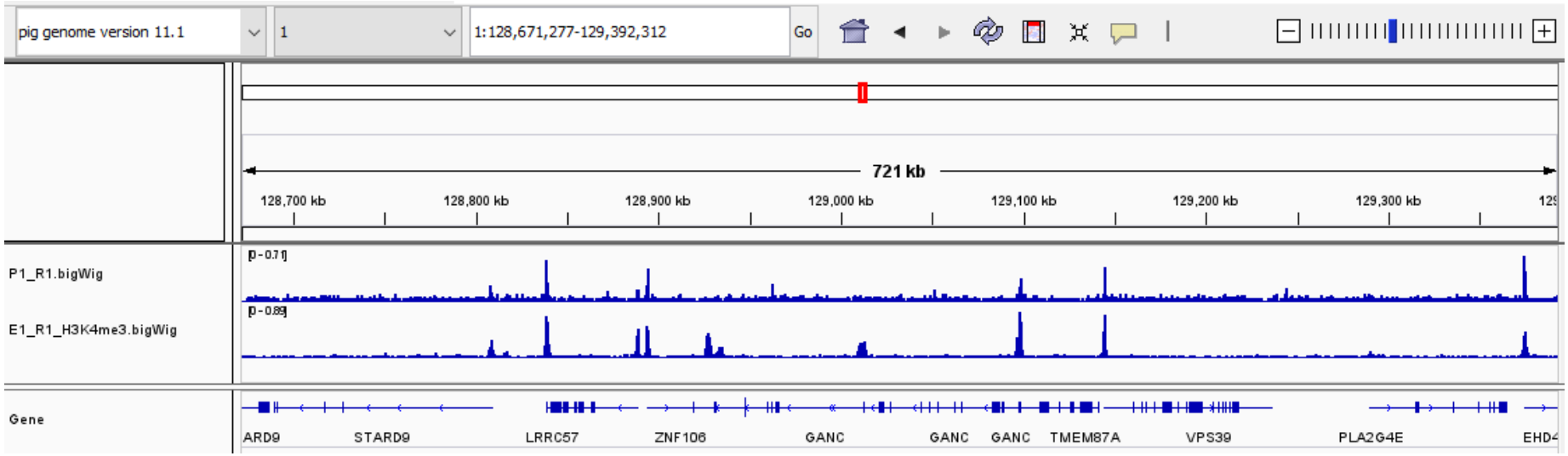
- Measure of chromatin accessibility



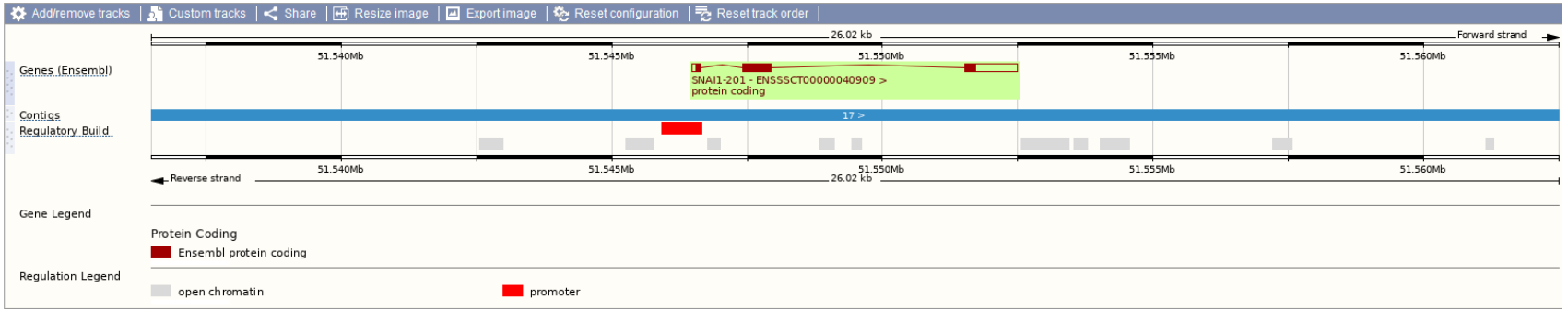
The different assays used for annotation: ChIP-seq

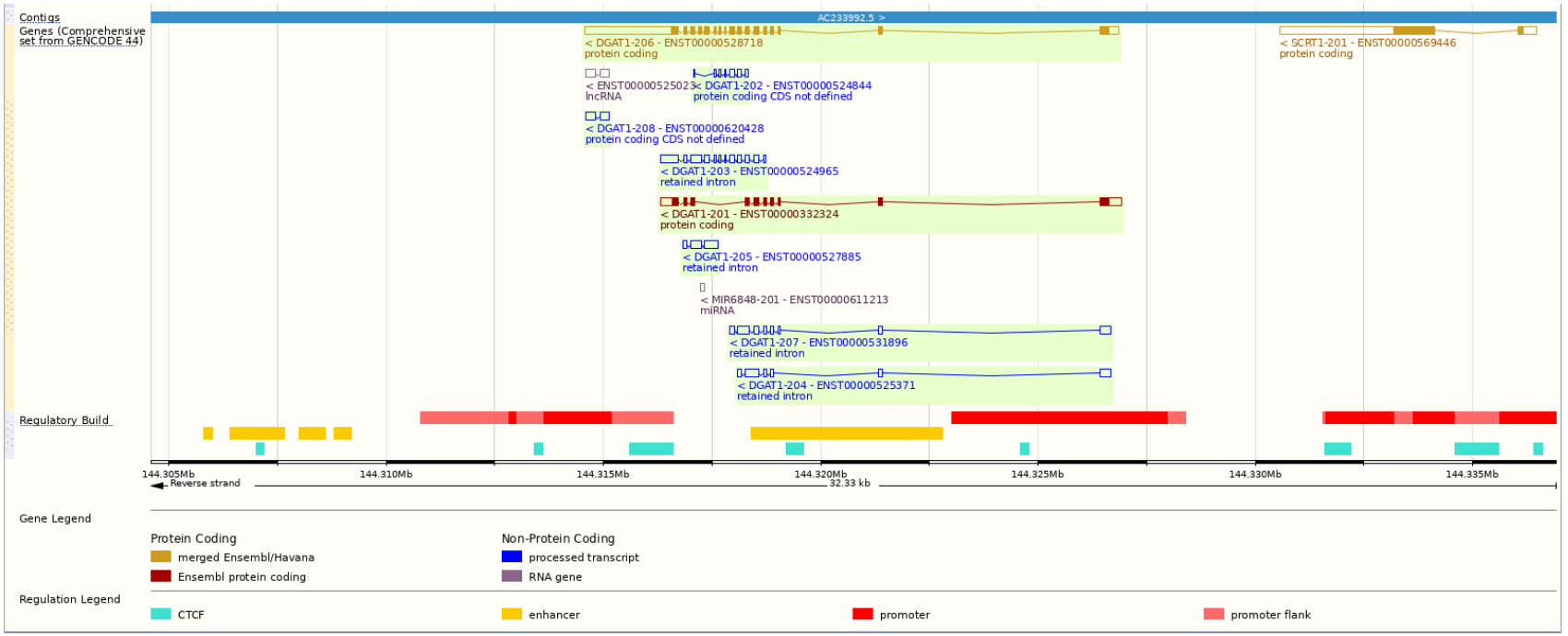


The different assays used for annotation: ChIP-seq



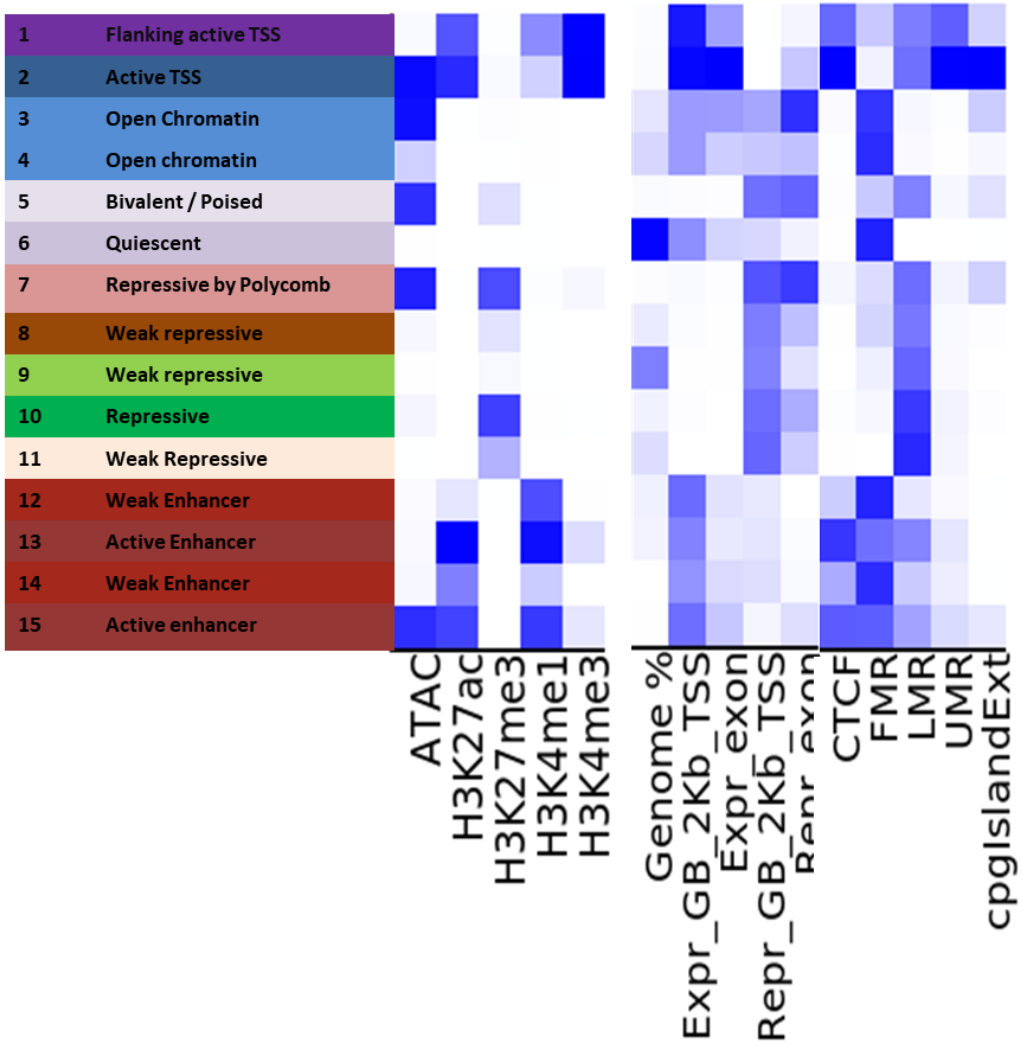
The different assays used for annotation: outputs





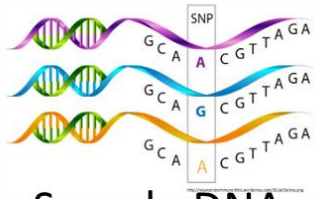
The different assays used for annotation: outputs

Epigenetic states



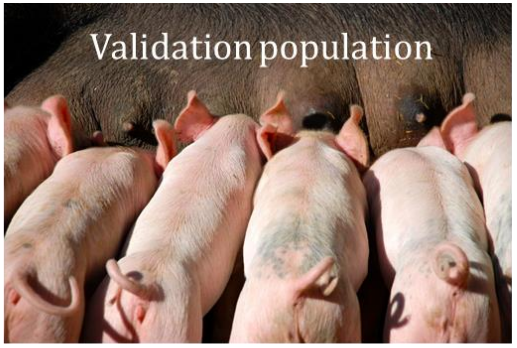
Functional annotations: its use for breeding animals

Functional annotations: its use for breeding animals



Sample DNA

Trait phenotypes

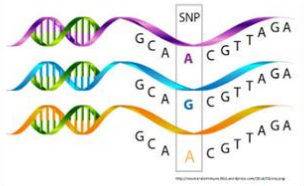


Genomic Prediction Model

$$Y = \mathbf{Z}\theta + \mathbf{X}\beta + \varepsilon$$

+0.2 +0.7 +0.4 +0.5

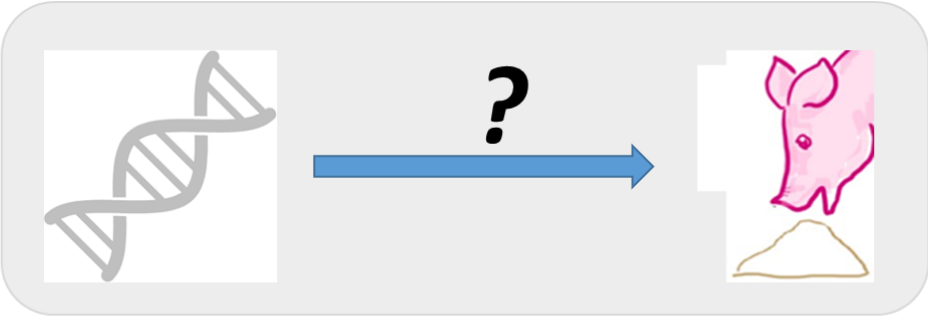
Genomic Breeding Values (kg/day)

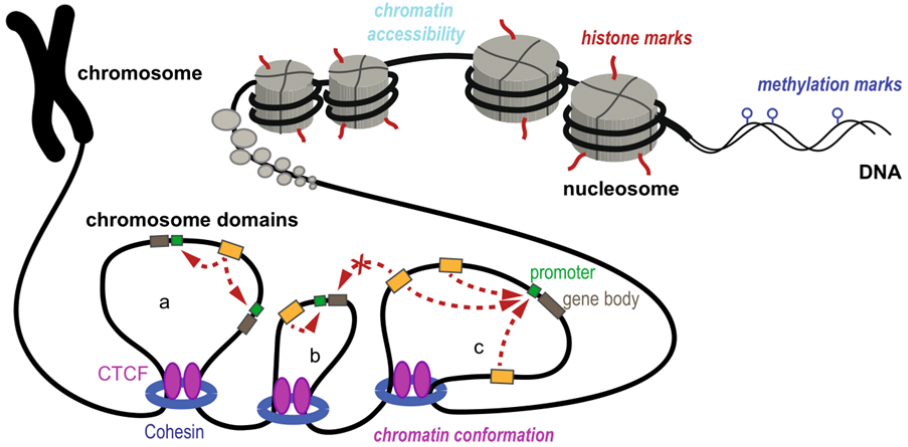


Sample DNA

Selected? X ✓ X ✓

Functional annotations: its use for breeding animals



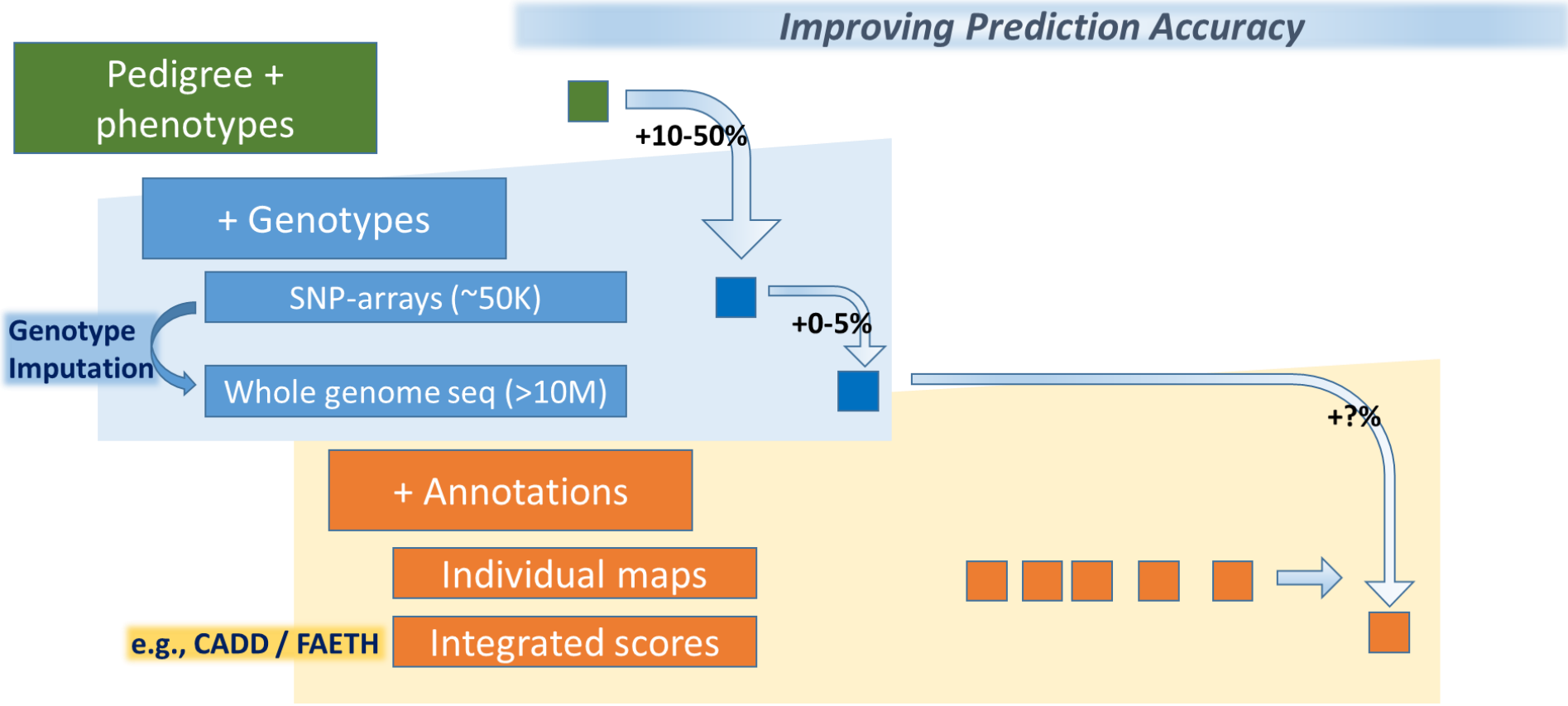


Adapted from: Ohnmacht et al, 2020. J Neural Transmission, 127; 729-748

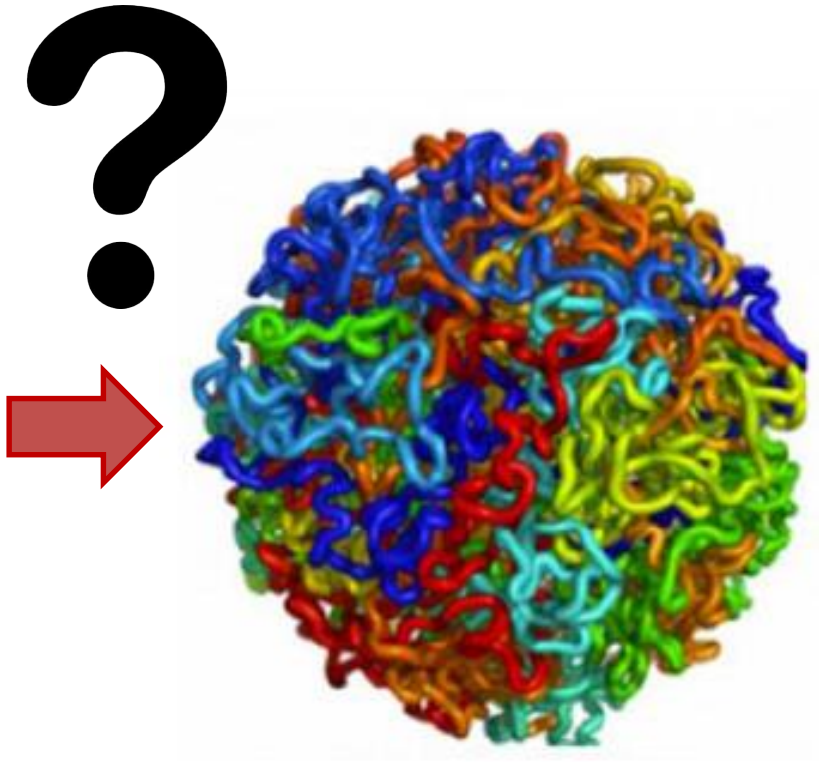
Two paths to capture intermediate stages:

1. Use animals' omics data in genomic prediction
   "individual level"
   Generate individual level methylation & gene expression data
2. Use functional annotation in genomic prediction
   "population level"
   Generate omics data with many different assays

Functional annotations: its use for breeding animals

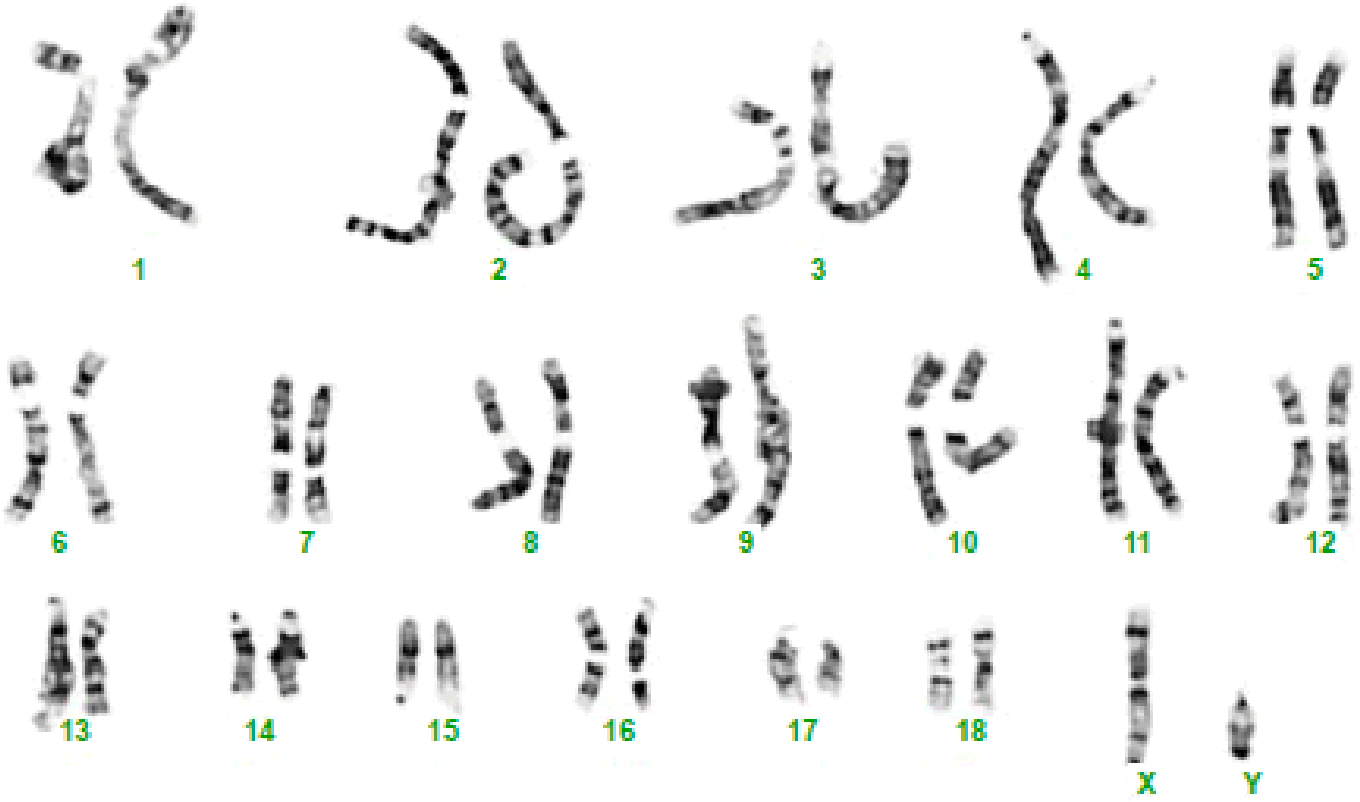


The complex way to perform spatial annotations



Chromosome and DNA: a long history

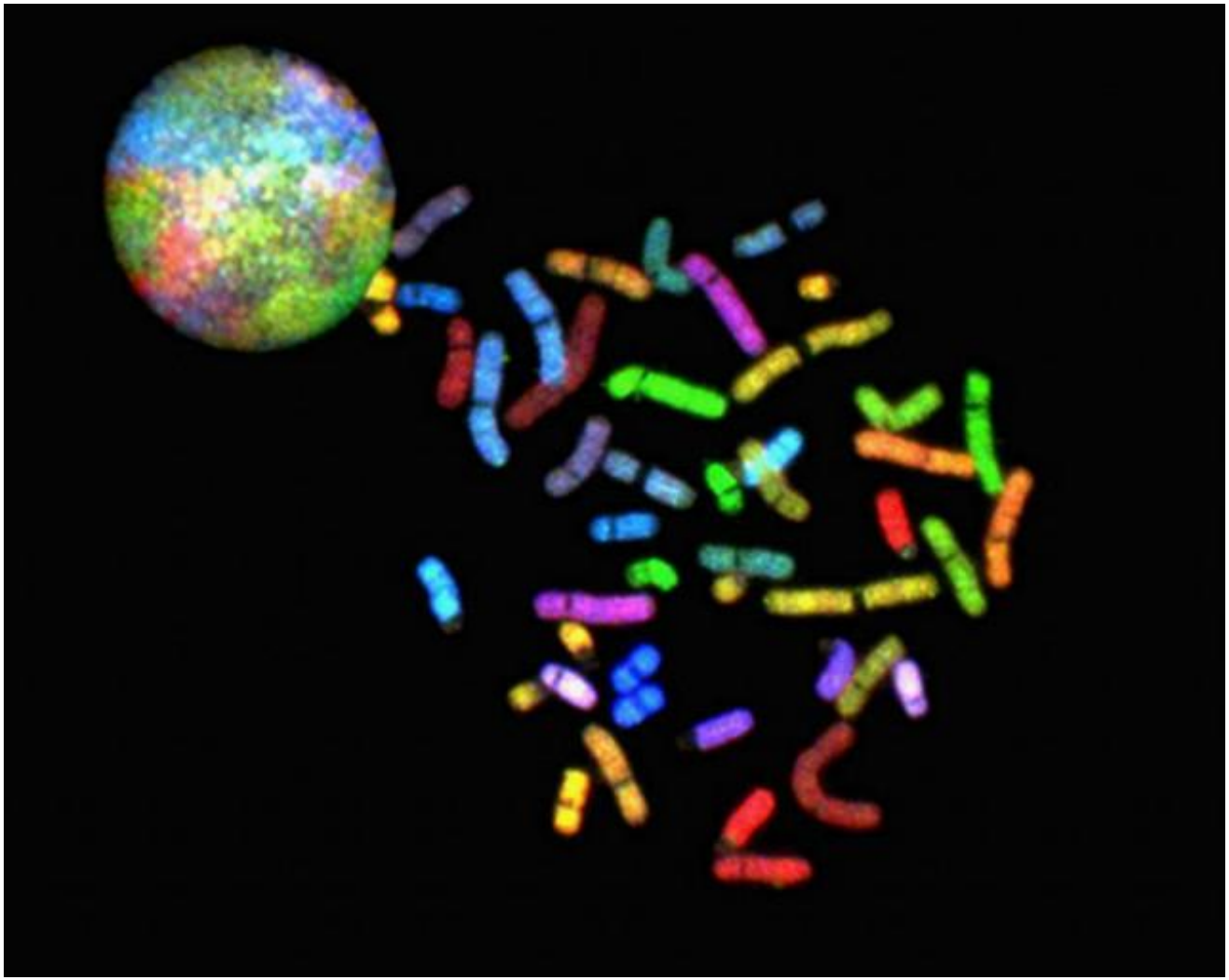
DNA is organized around chromosomes



For each diploid cell, you have 2n molecules of DNA where n is the number of chromosmes

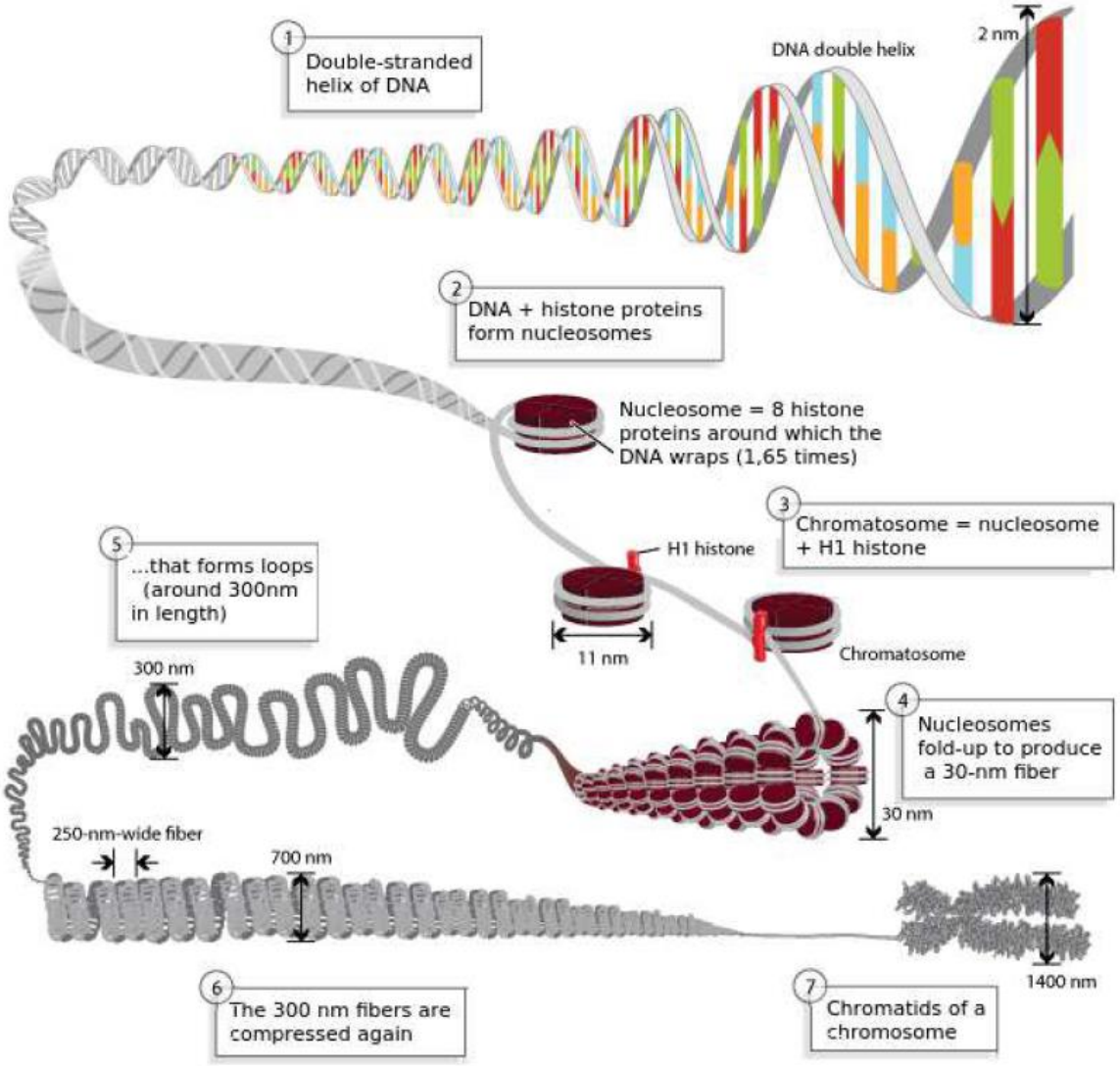
Chromosome and DNA: a long history

DNA is organized around chromosomes



Chromosome and DNA: a long history

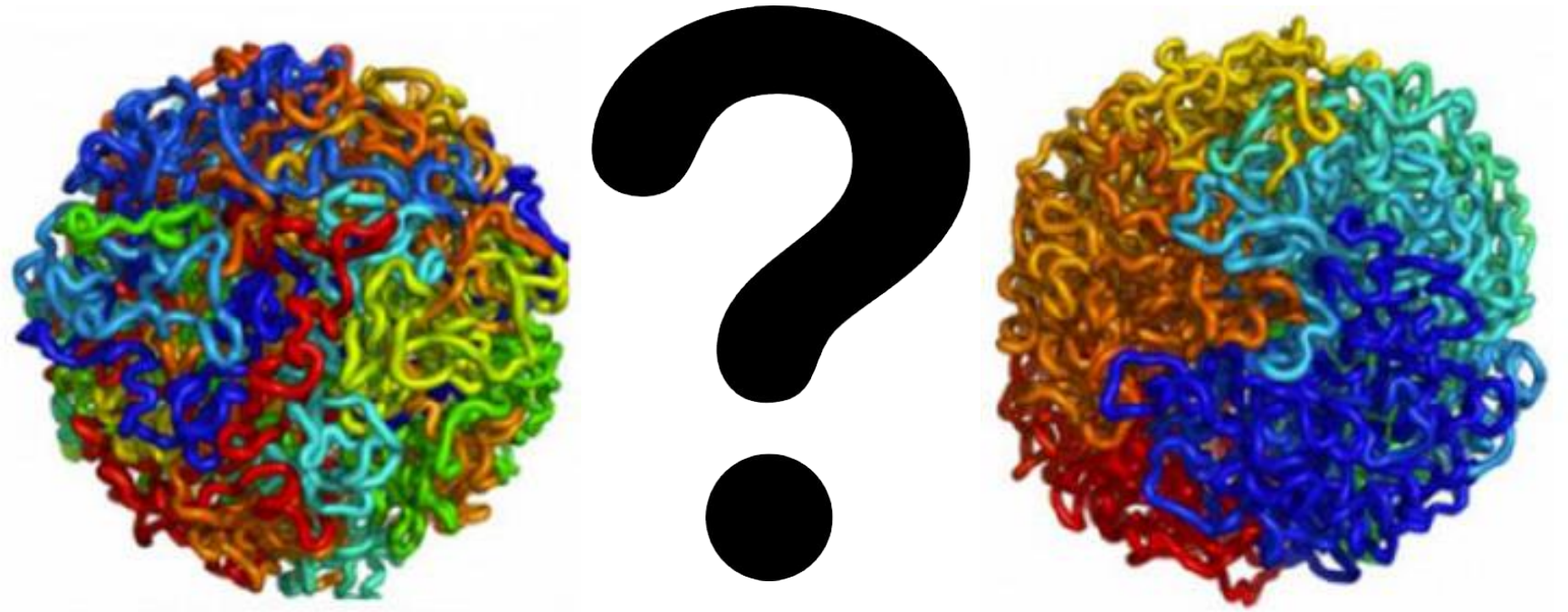
DNA is organized around chromosomes



Pierce, Benjamin. Genetics : A Conceptual Approach

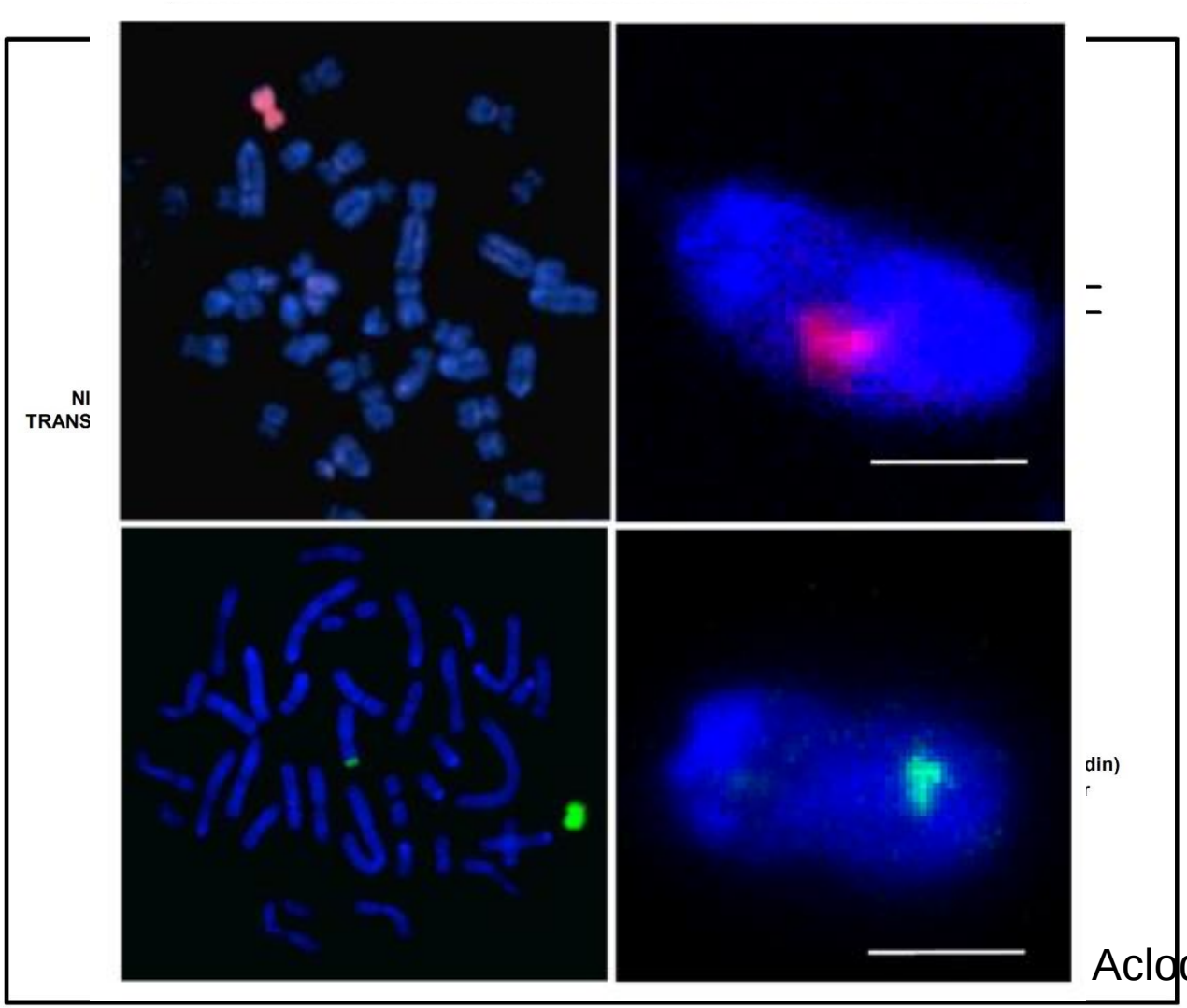
Chromosome and DNA: a long history

DNA is organized around chromosomes



Looking at chromosome organisation

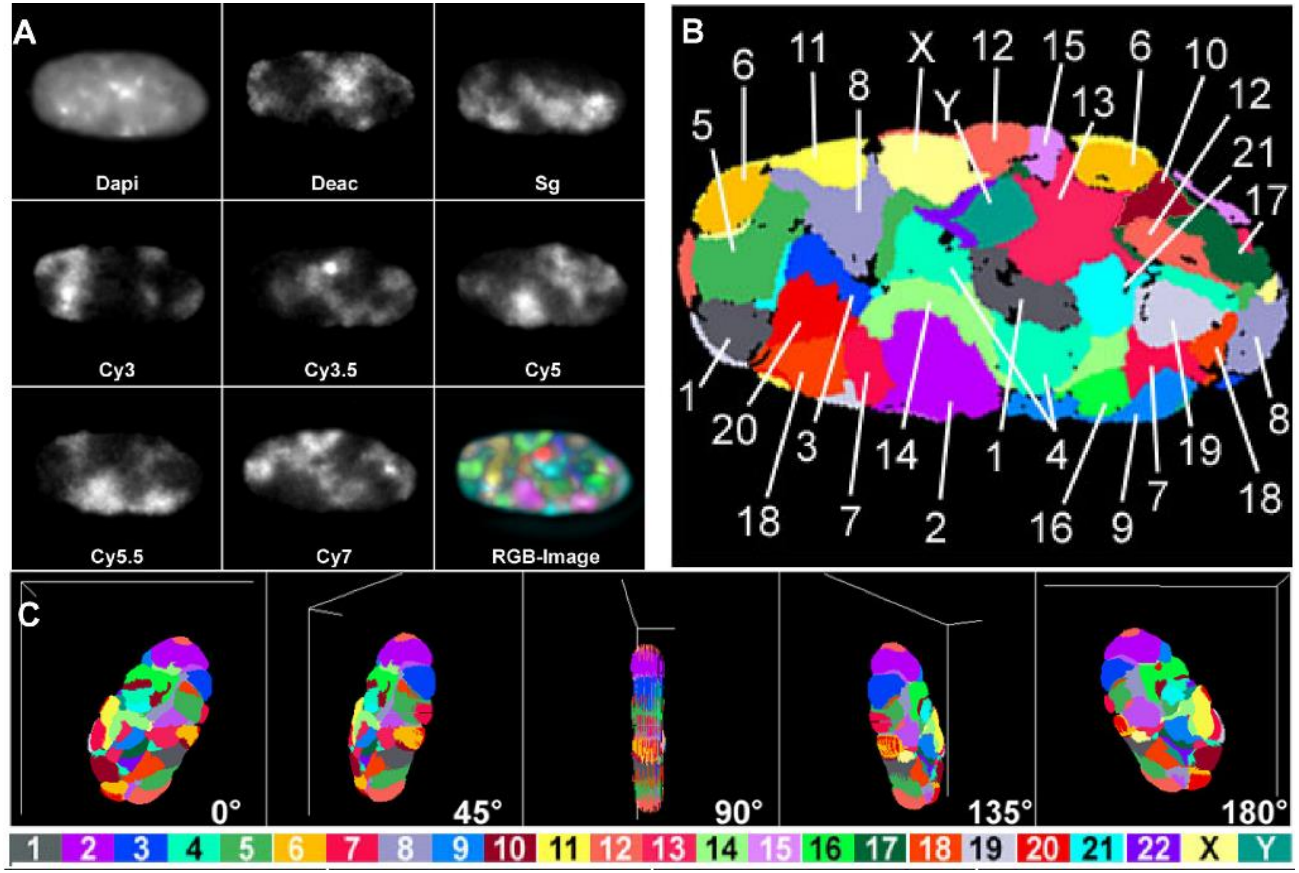
Fluorescent in situ hybridization allows to visualize interphase chromosome into the nucleus



Acloque et al. 2013

Looking at chromosome organisation

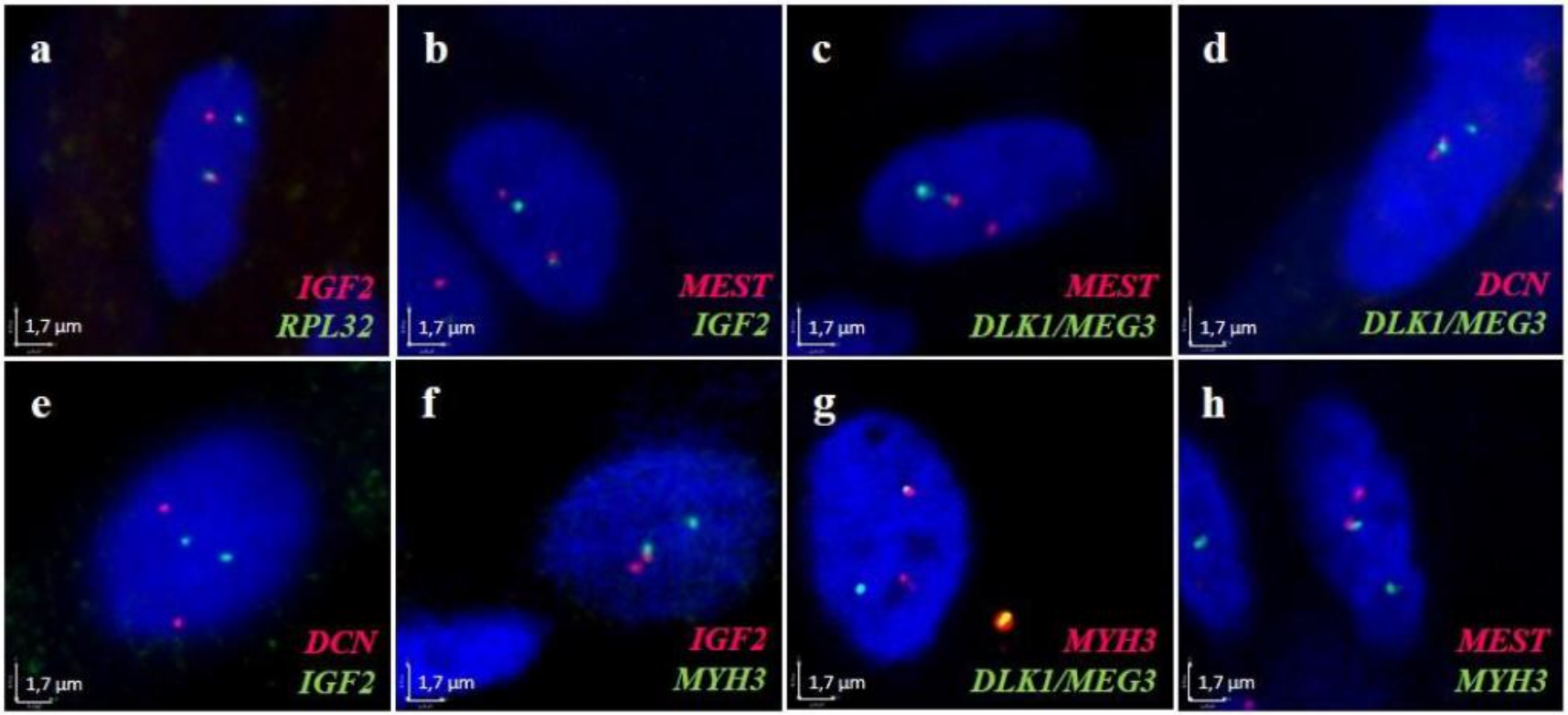
Within the interphase nucleus chromosomes are organized in chromosome territories



Bolzer et al. 2005

Looking deeper: gene-gene spatial interactions

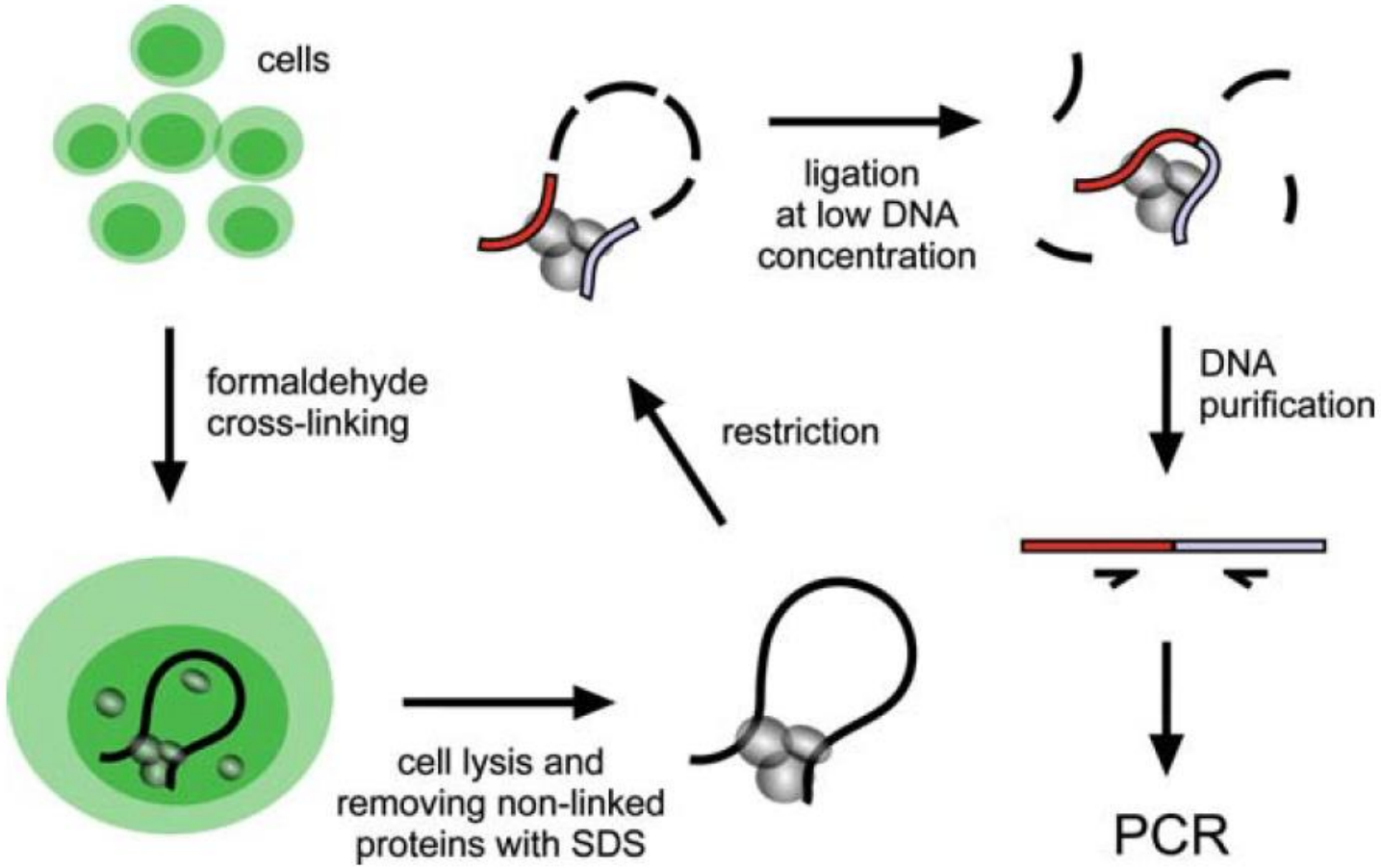
by FISH: low throughput



Marti-Marimon et al. 2018

Looking at gene-gene spatial interactions

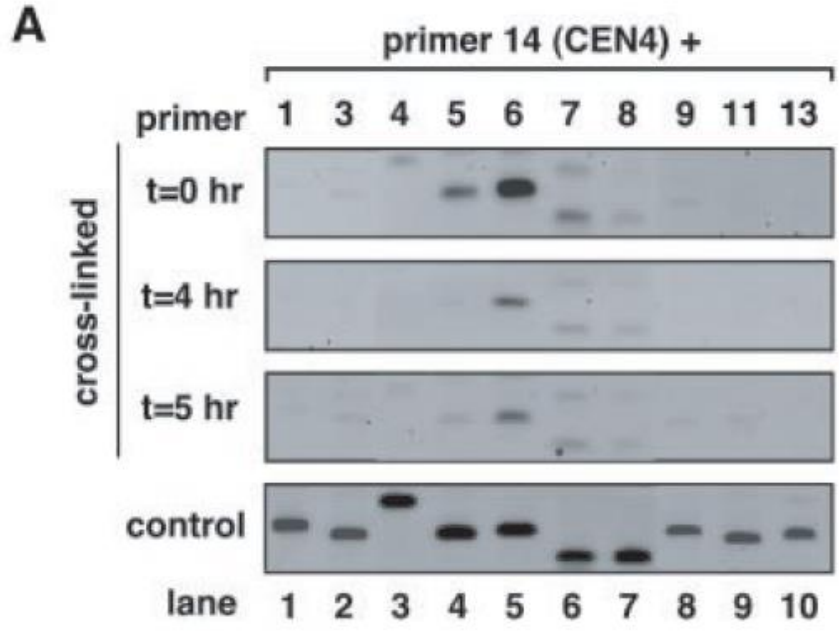
by Capture of Chromosome Conformation (3C)

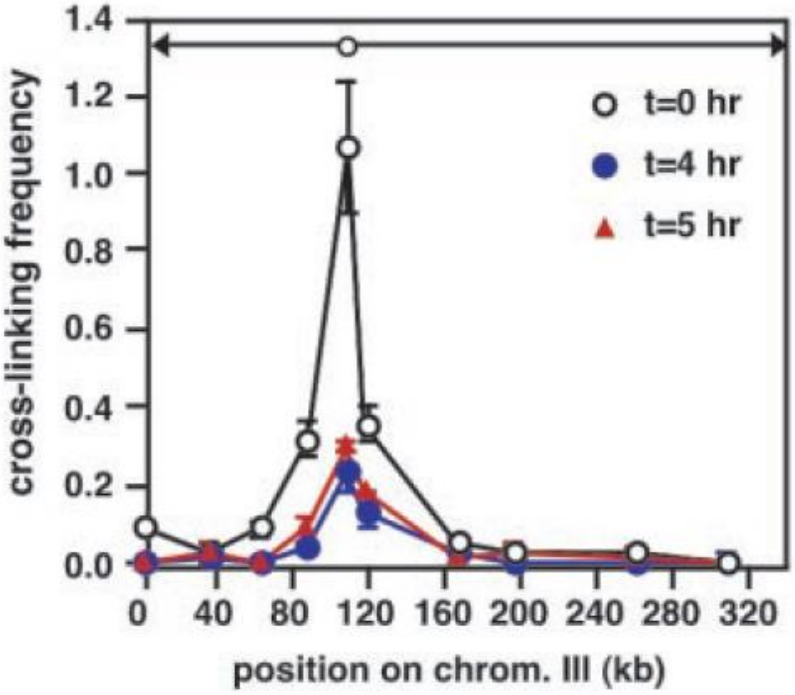


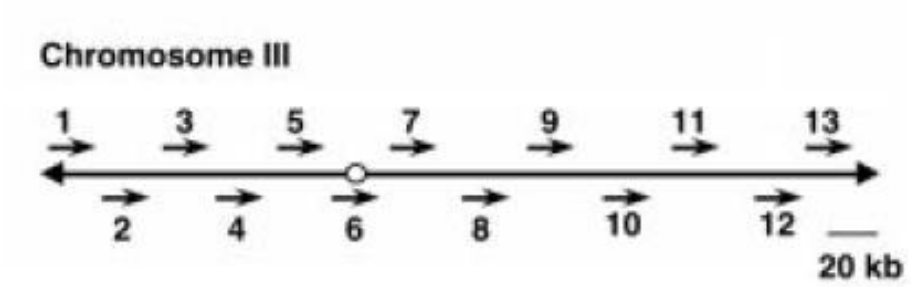
Gavrilov et al. 2009 (from Dekker et al. 2002)

Looking at gene-gene spatial interactions

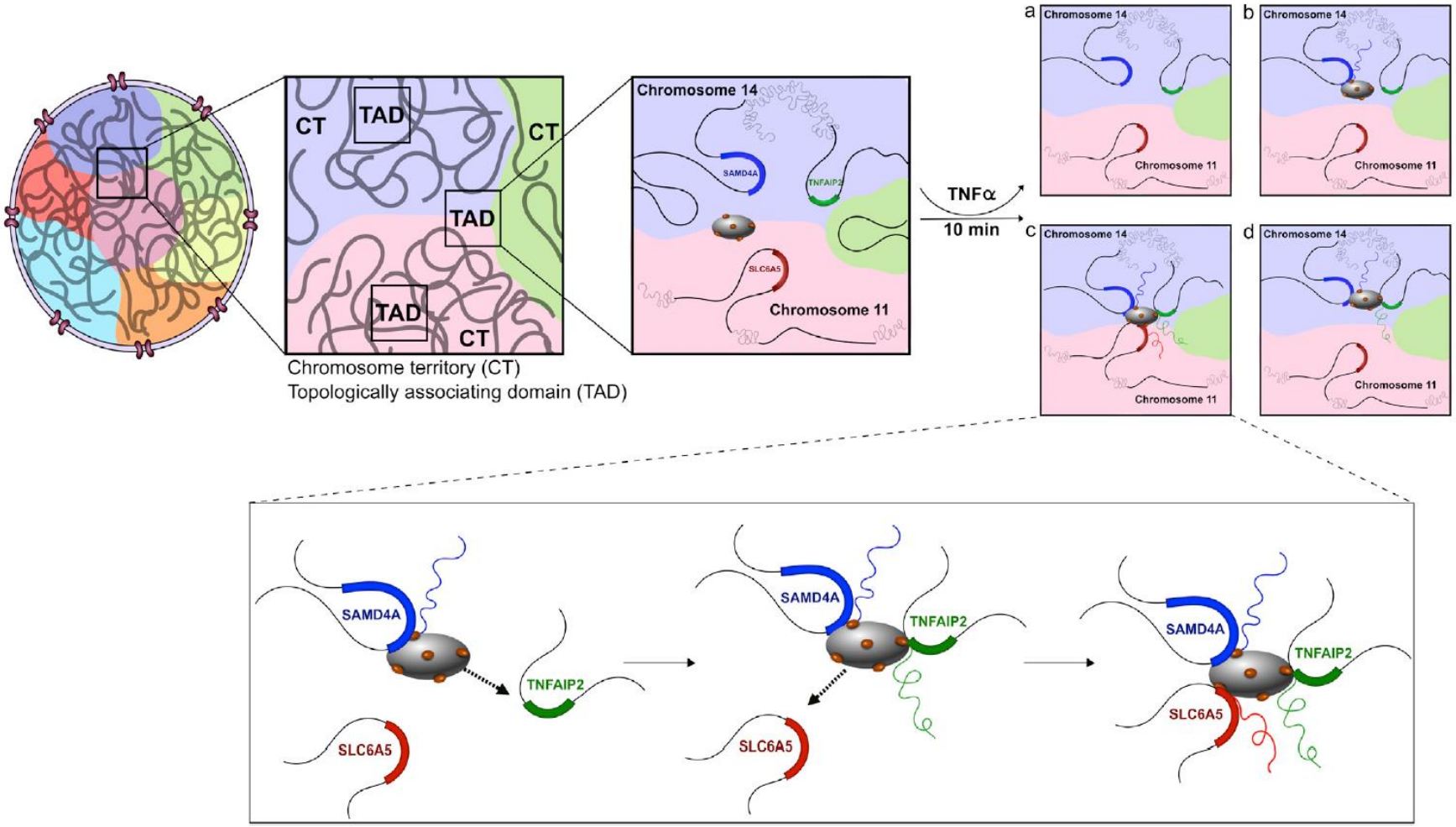
by Capture of Chromosome Conformation (3C)







3D organisation is important for gene regulation and nucleus function



Fanucchi et al. 2013

Toward a 3D genome map ?

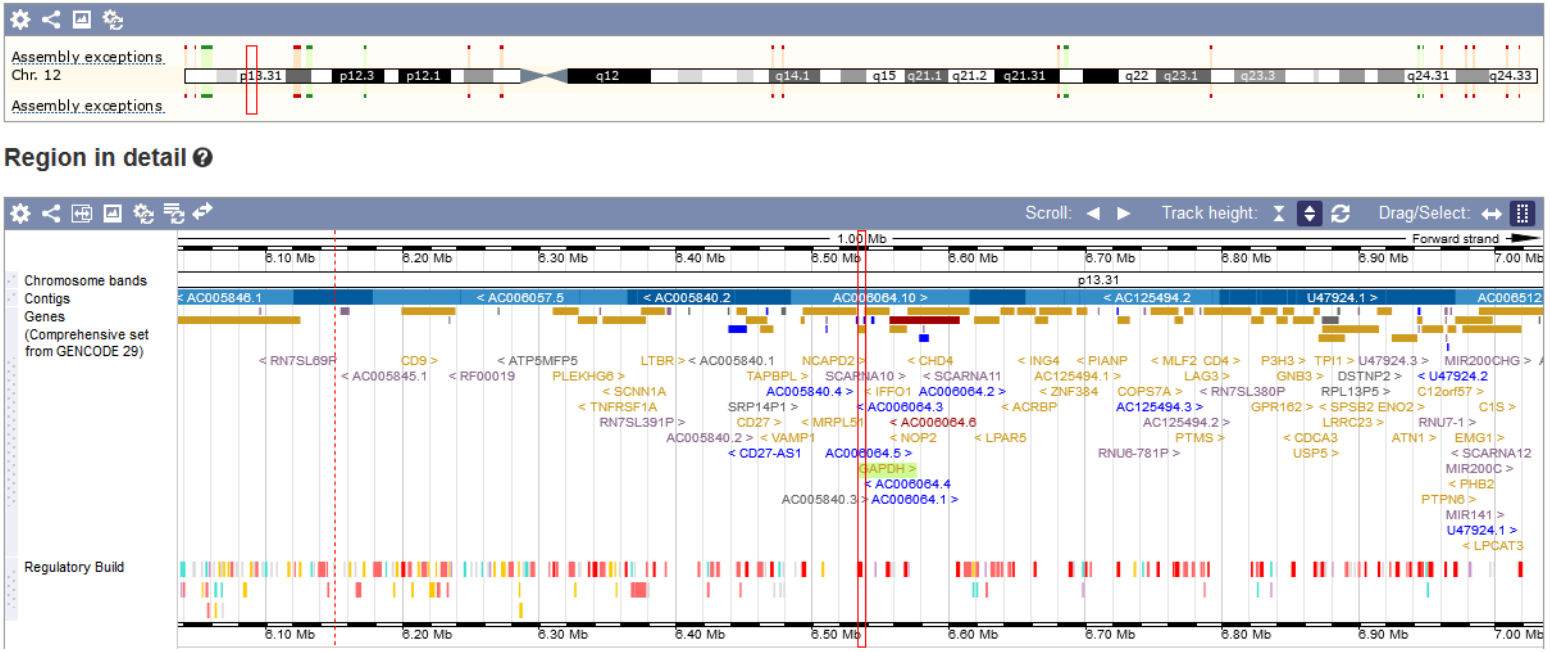
- Low input approaches are working well but:
  - Time and money-consuming
  - Only few genomic regions can be studied
  - Difficult to draw general rules on genome function and activities
  - You have a supervised analysis (identified candidates regions)

From a single dimension genome…

- First reference genomes between 2004-2012 for livestock species

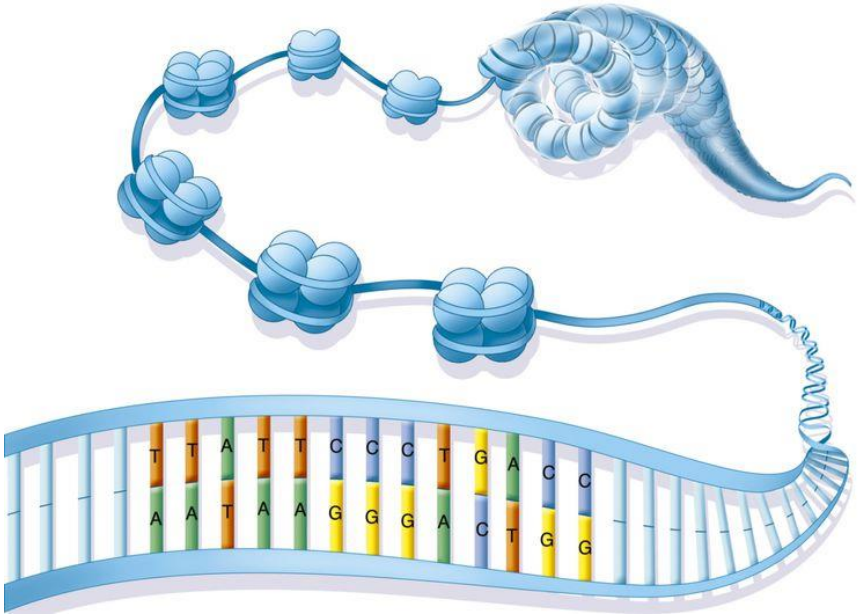
This is the ultimate physical map of the genome!

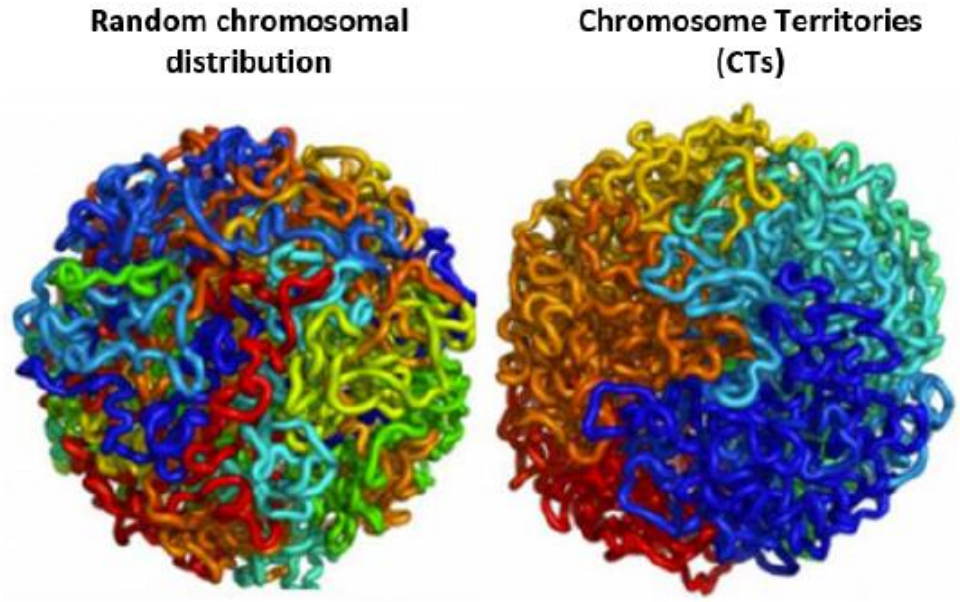
But this is still a linear information:



Toward a 3D genome map

- The genome is condensed into a small volume, the nucleus, thanks to chromatin compaction



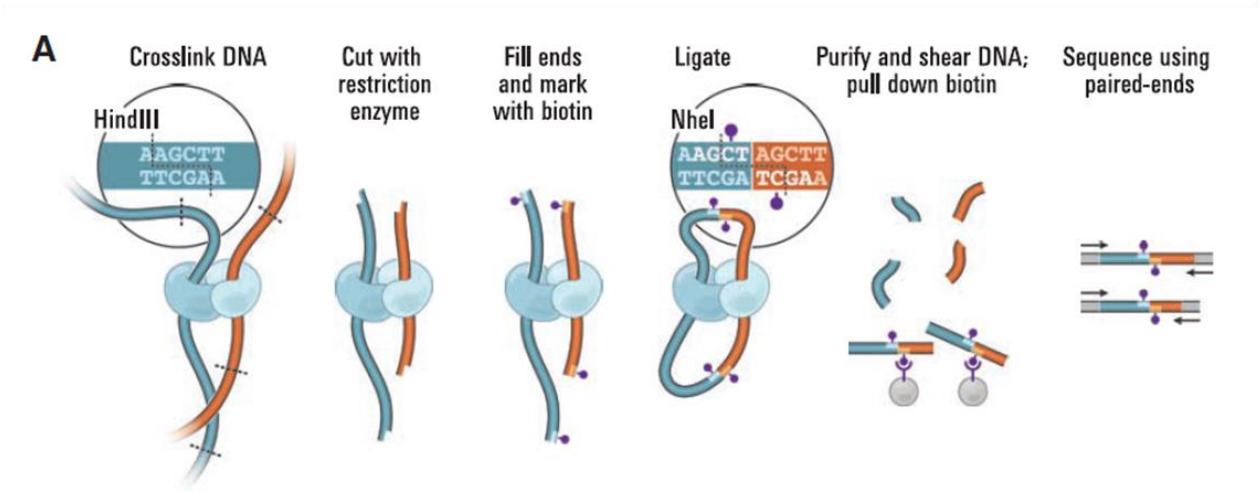


Toward a 3D genome map: the Hi-C revolution

By combining 3C and next generation sequencing, Erez Lieberman-Aiden from Lander's and Dekker's lab filled the gap by inventing Hi-C and allowed important breakthroughs in the knowledge of genome organization and function.

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

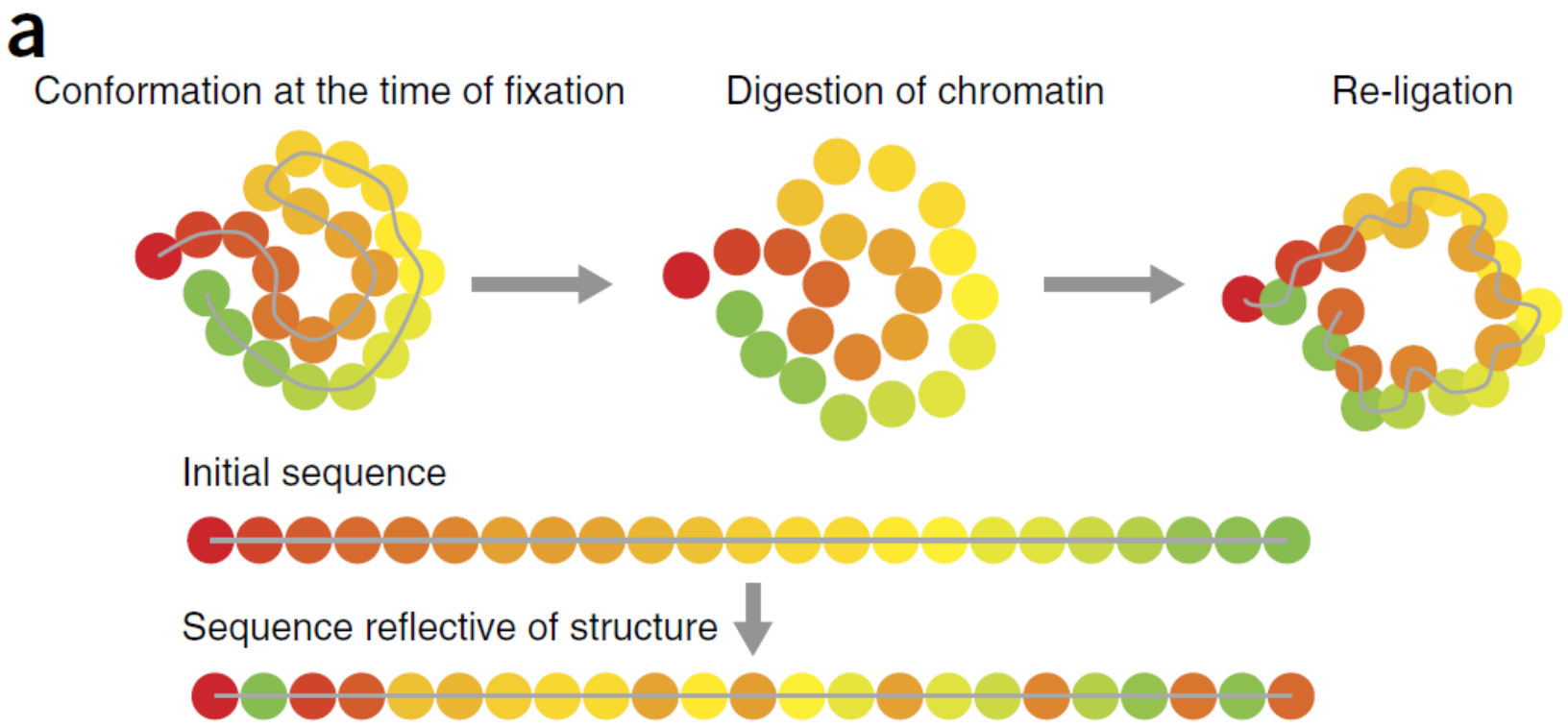
Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†



Hi-C: how it works

Chromatin conformation using Hi-C

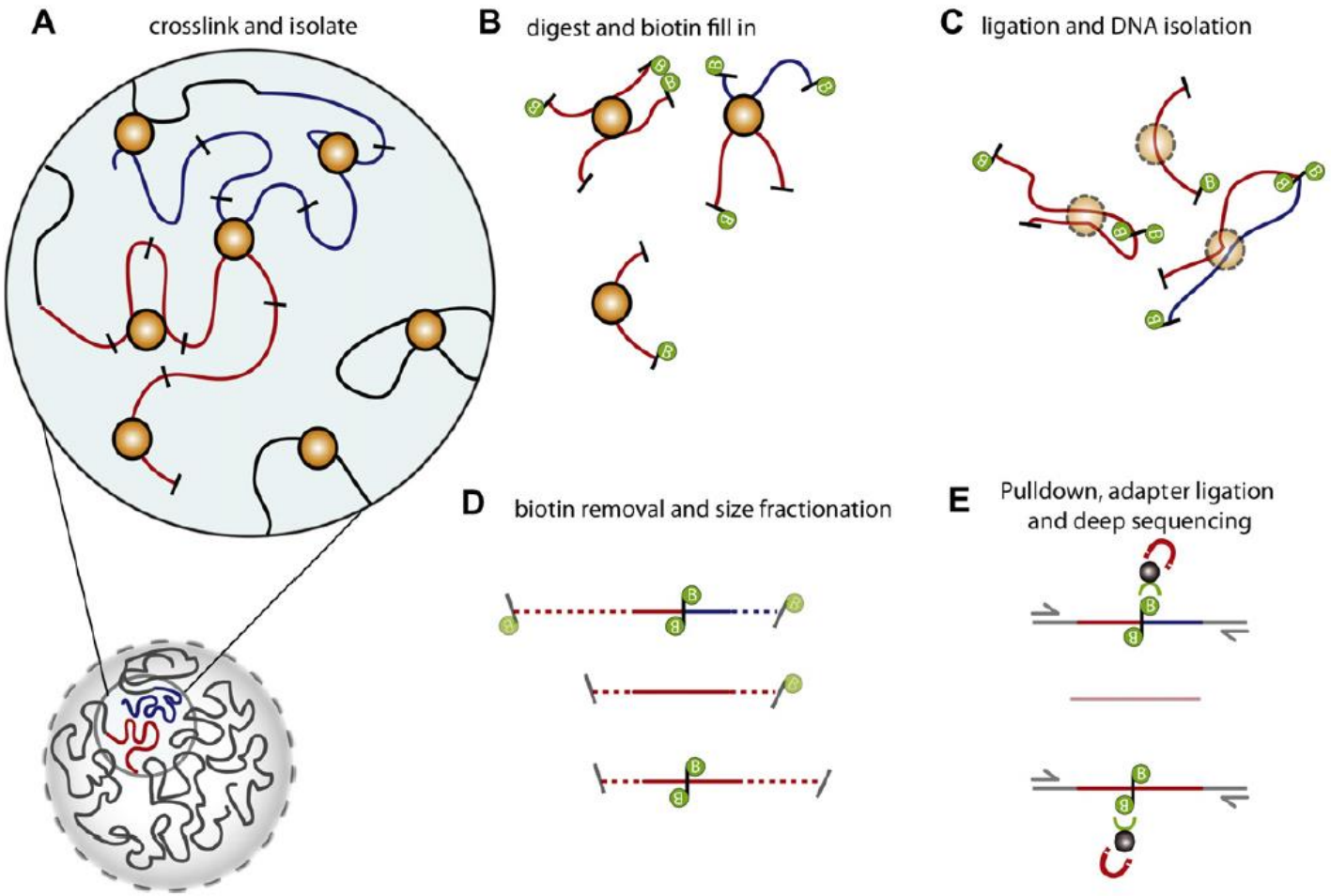
Principle of Chromosome Conformation Capture



from Davies et al. 2017

Hi-C: how it works

Molecular strategy



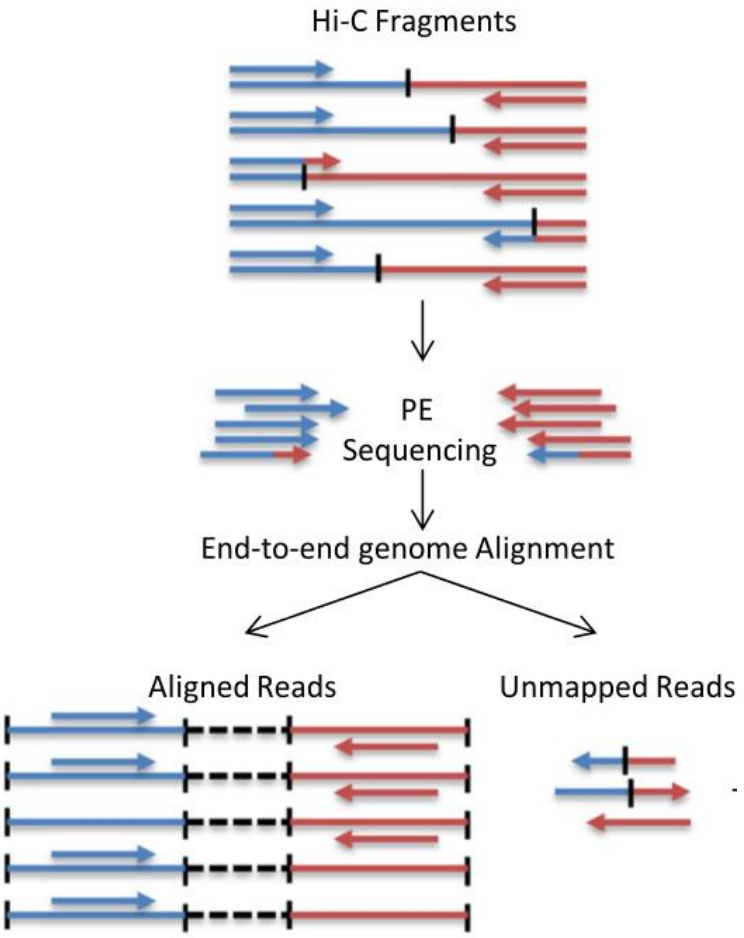
from Belton et al. 2012, Rao et al. 2014

Hi-C: how it works

Analysis of raw sequences data:

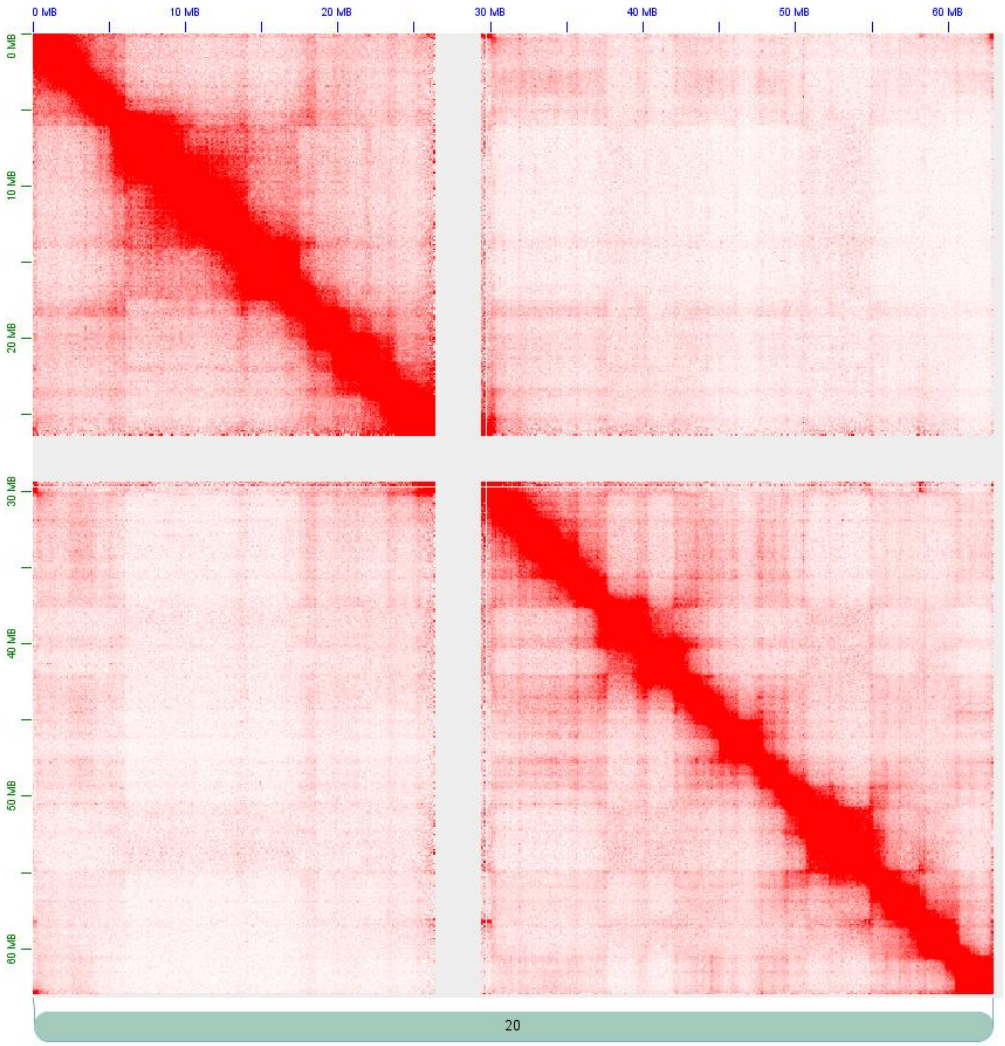
Pipeline
- Trim reads (religation site)
- Map on reference genome
- Discard inconsistent pairs
- Build contact matrix
- Normalize contact matrix
- Generate html report
- Find **TADs**
- Find **A and B compartments**

Software
- HiC-Pro pipeline (Servant et al 2015)
- Bowtie2 mapping (Langmead et al, 2009)
- ICE normalization (Imakaev et al, 2012)
- HiTC display & A/B comp. (Servant et al, 2012)
- HiFive pipeline (Sauria et al, 2015)
- Armatus TAD finding (Filippova et al, 2014)
- Juicebox browser (Durand et al, 2016)



Hi-C: how it works

Main output data: Matrix of proximity / Interaction maps



Observed interactions for human chromosome 20 in human ES cells

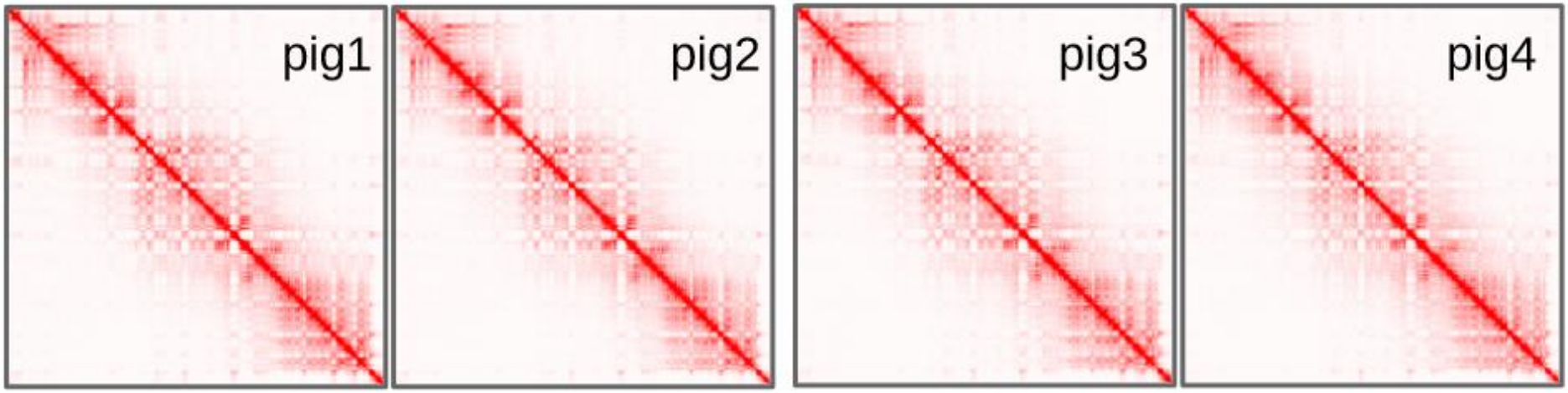
Data from Dekker's lab on Juicer

Resolution : 100kb

The chromosome is divided in bin of 100kb. For each bin, we counted paired reads with one read inside the bin, and the other on another bin (cis or trans),

Hi-C: how it works

Main output data: Matrix of proximity / Interaction maps



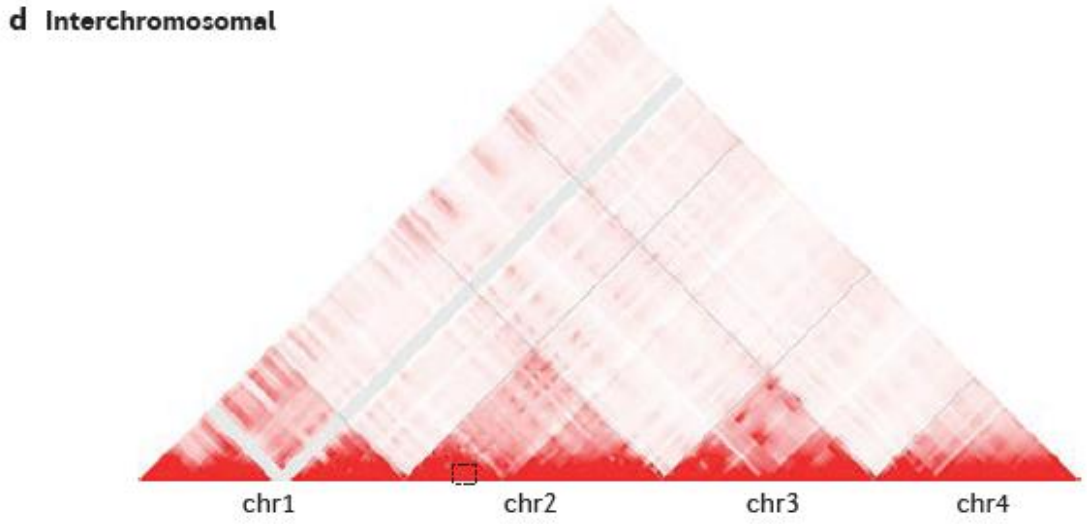
The Hi-C contact matrices from the 4 replicates (*Sus scrofa*, chromosome 1)

From Foissac et al. 2019

Hi-C: what can we learn on genome organisation and function from these data ?

1. Compartments

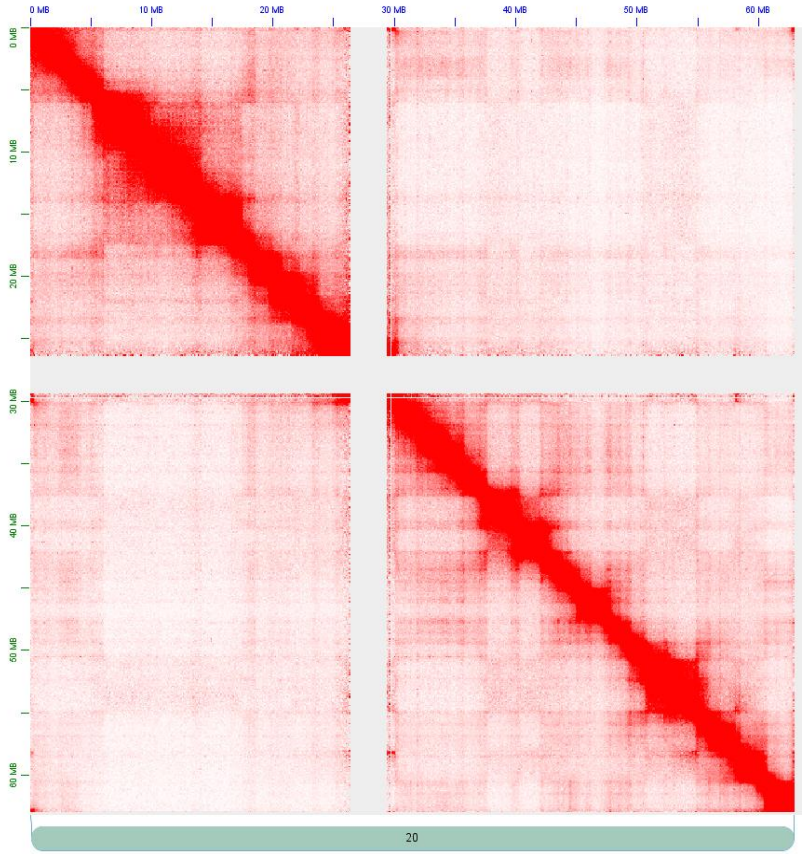
Low resolution Hi-C (500kb- 1Mb) provides cues on global genome organization in the nucleus



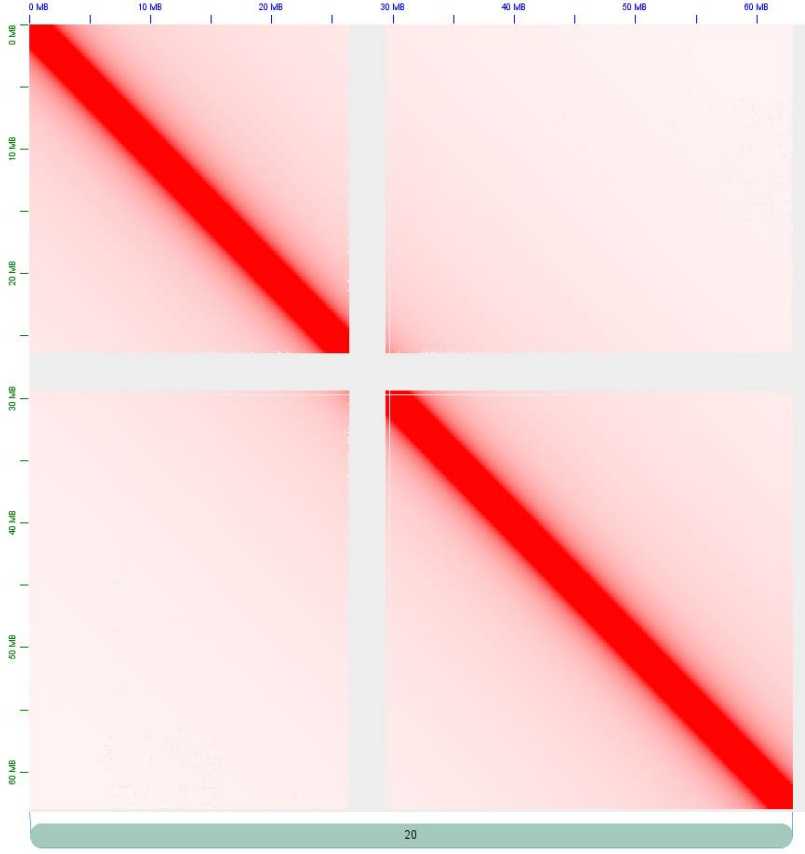
Bonev and Cavalli 2016

1. Discovery of compartments

Low resolution Hi-C (500kb- 1Mb):
Normalization for reducing “proximity noise”



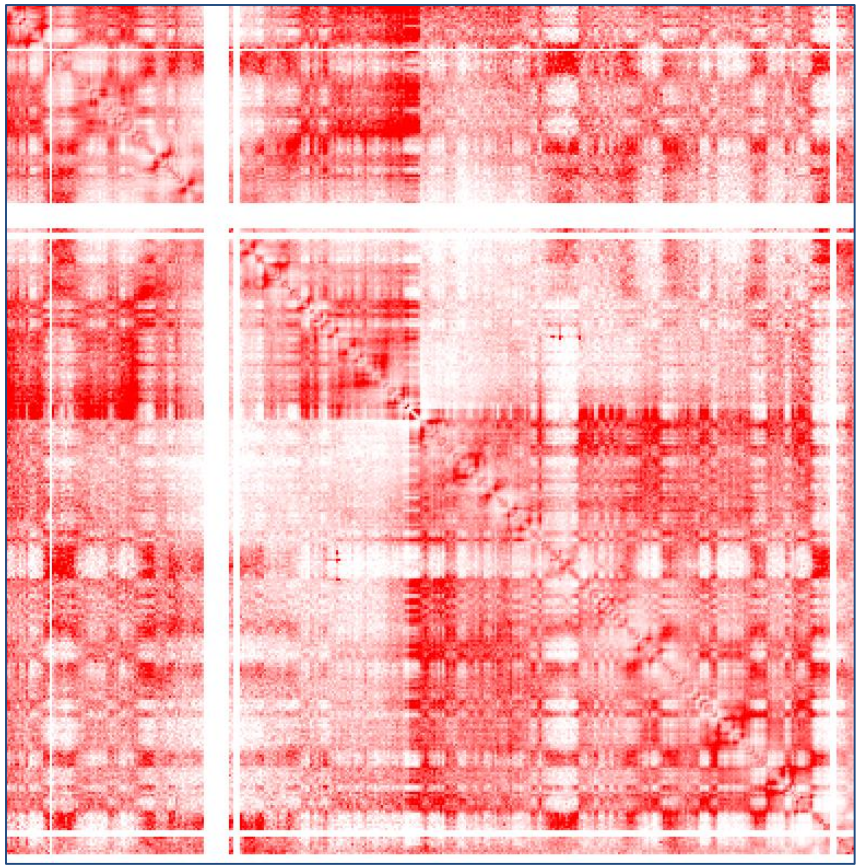
Observed



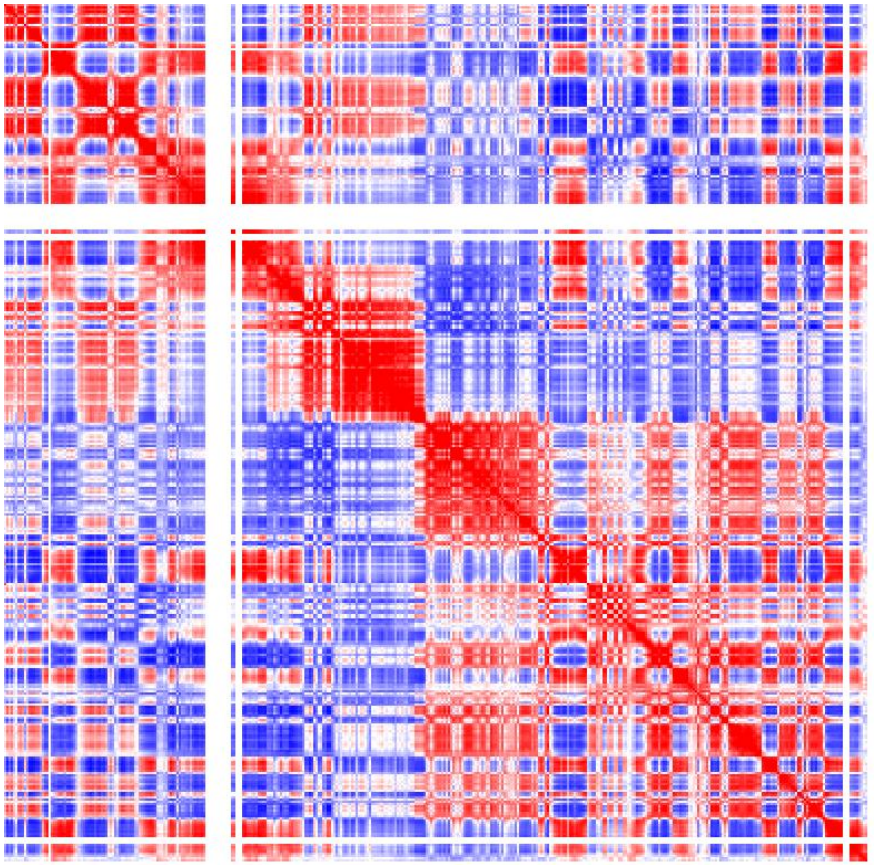
Expected

1. Discovery of compartments

Low resolution Hi-C (500kb- 1Mb):
Normalization for reducing “proximity noise”



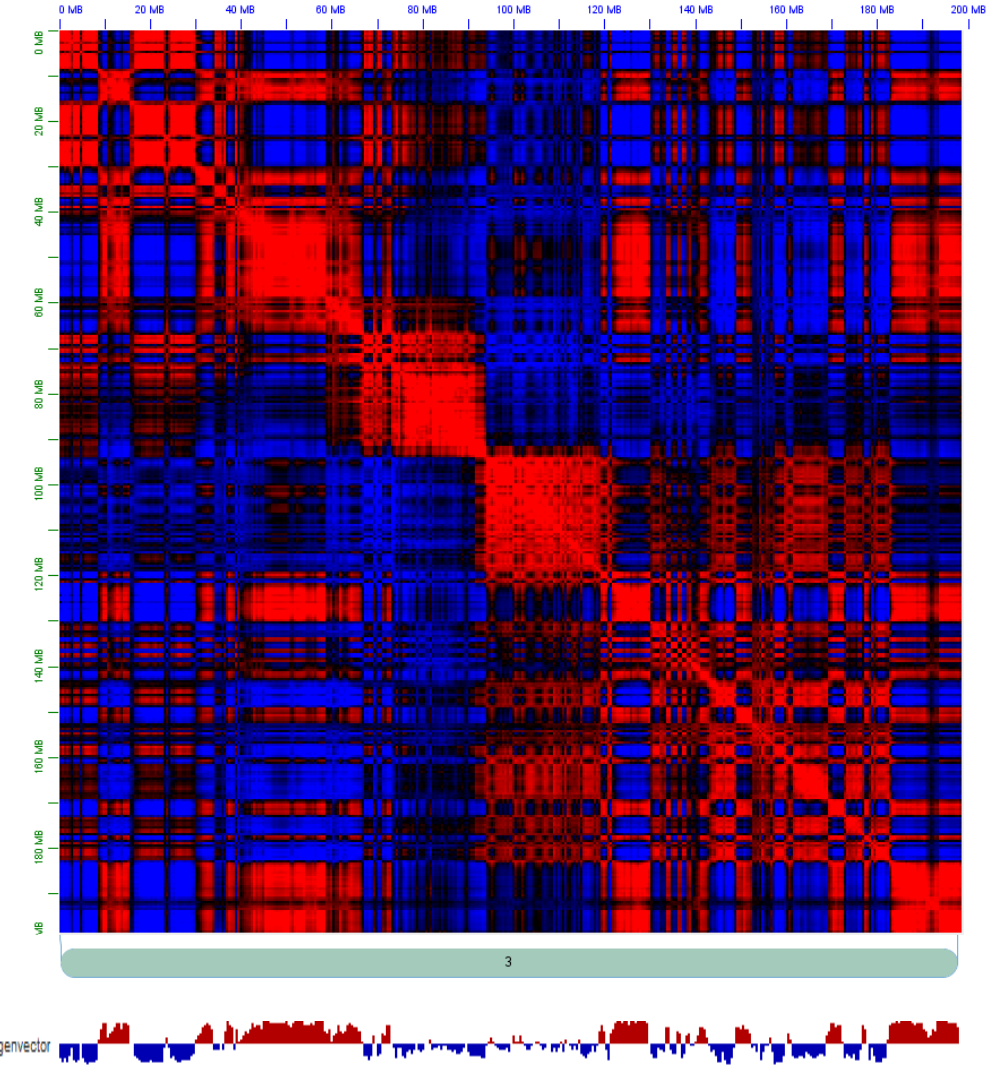
Observed/Expected



Pearson correlation matrix

1. Discovery of compartments

Low resolution Hi-C (500kb- 1Mb): A and B compartments



Pearson correlation matrix

PCA: eigen values of PC1 or PC2

Succession of positive and negative values reflecting a global nuclear organization

Median size: 5Mb

1. Compartments

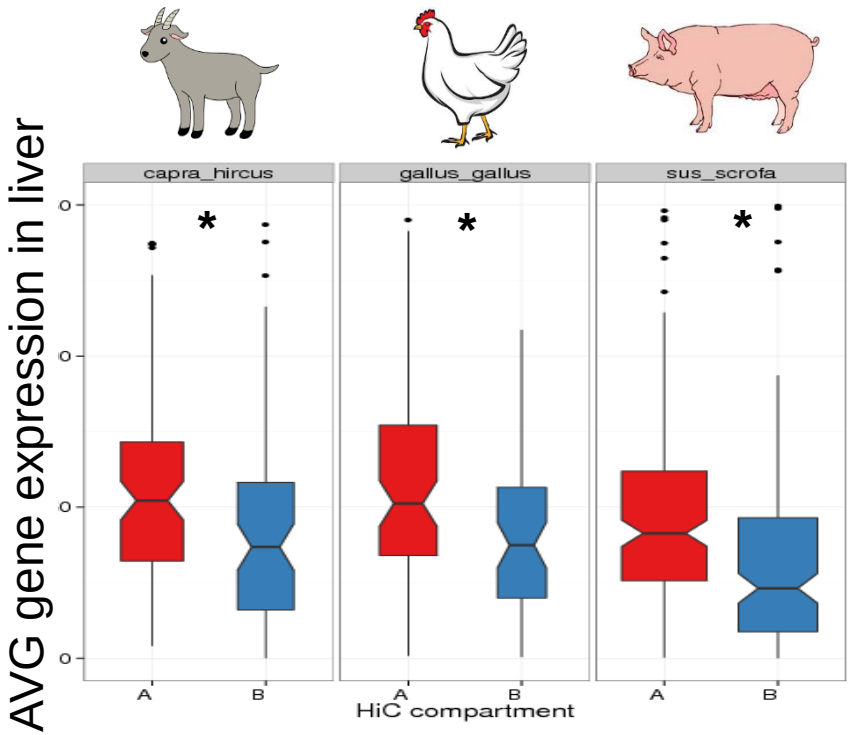
Biological properties of A and B compartments

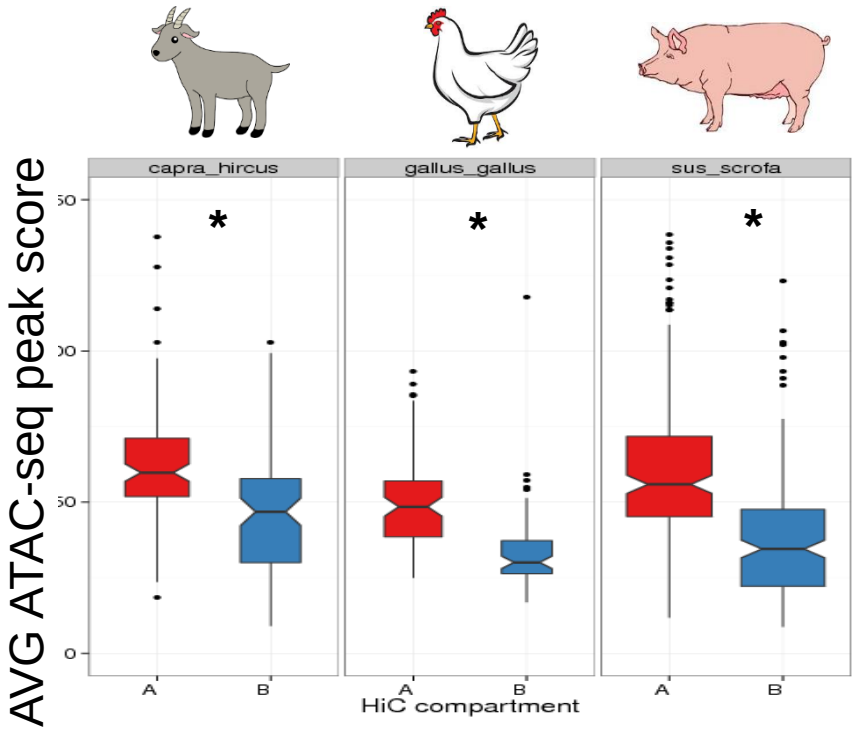
A compartments represent:
Gene-rich domains with an high GC content,
They are enriched for histone marks of active transcription
They tend to be located in the interior of the nucleus
They are enriched ofr early replication origins

B compartments:
Gene-poor domains and closed chromatin
They are enriched for histone marks associated with gene repression
They tend to be located on the nuclear periphery (LADs)
They are enriched for late replication origins

1. Compartments : biological properties

Confirmation by experimental data: the Fr-AgENCODE project (http://www.fragencode.org/)



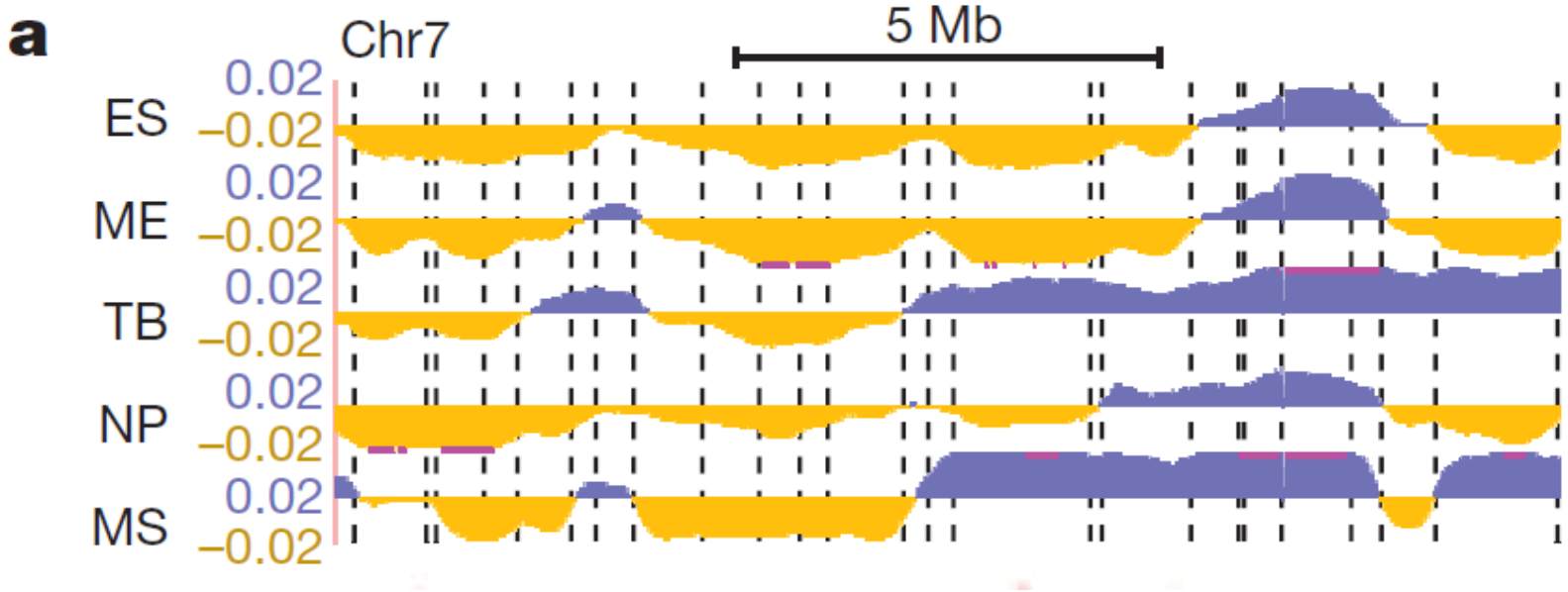


Genes are more transcribed in A vs B compartments

Chromatin is more accessible in A versus B compartments

1. Compartments : biological properties

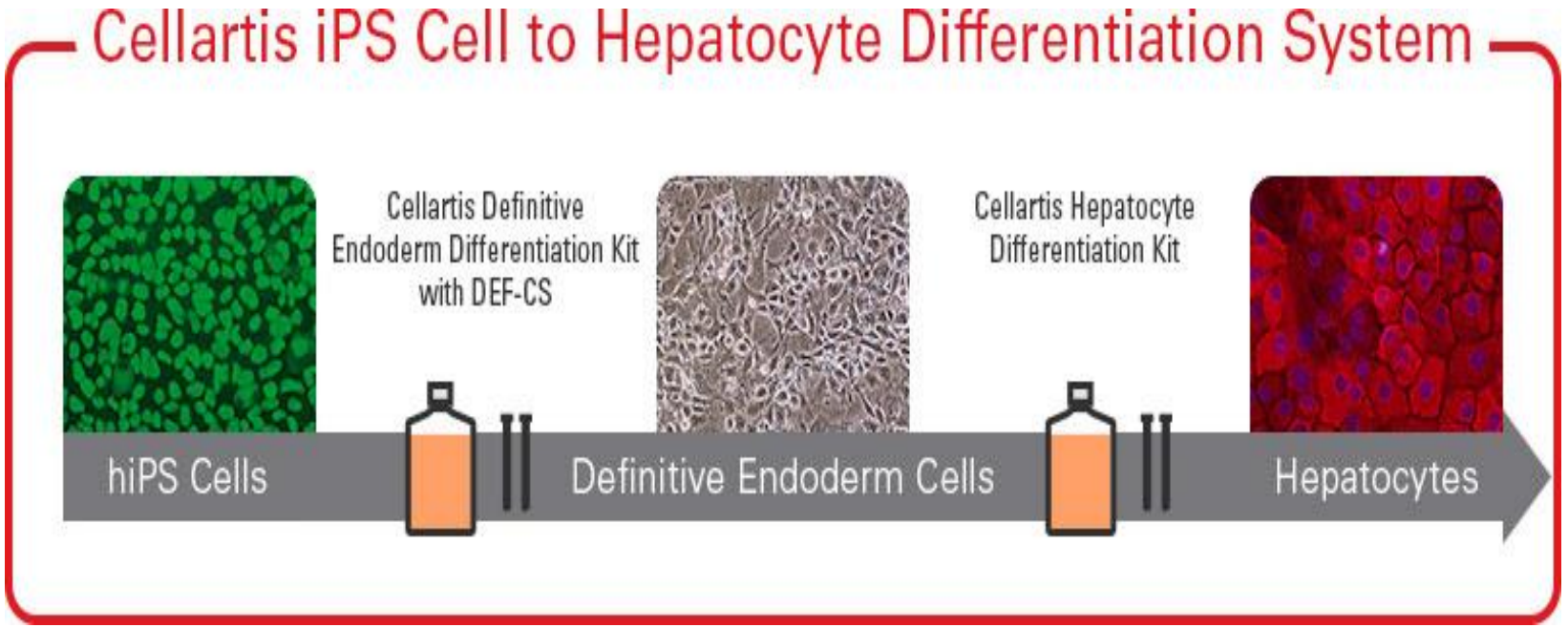
A and B compartments are mostly conserved but exhibit cell specificities.

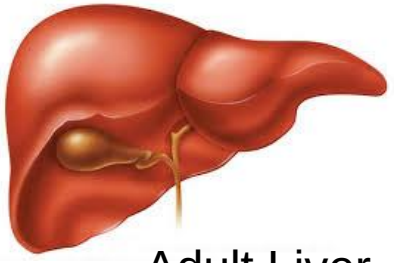


Dixon et al. 2015

1. Compartments : biological properties

A and B compartments are mostly conserved but exhibit cell specificities.

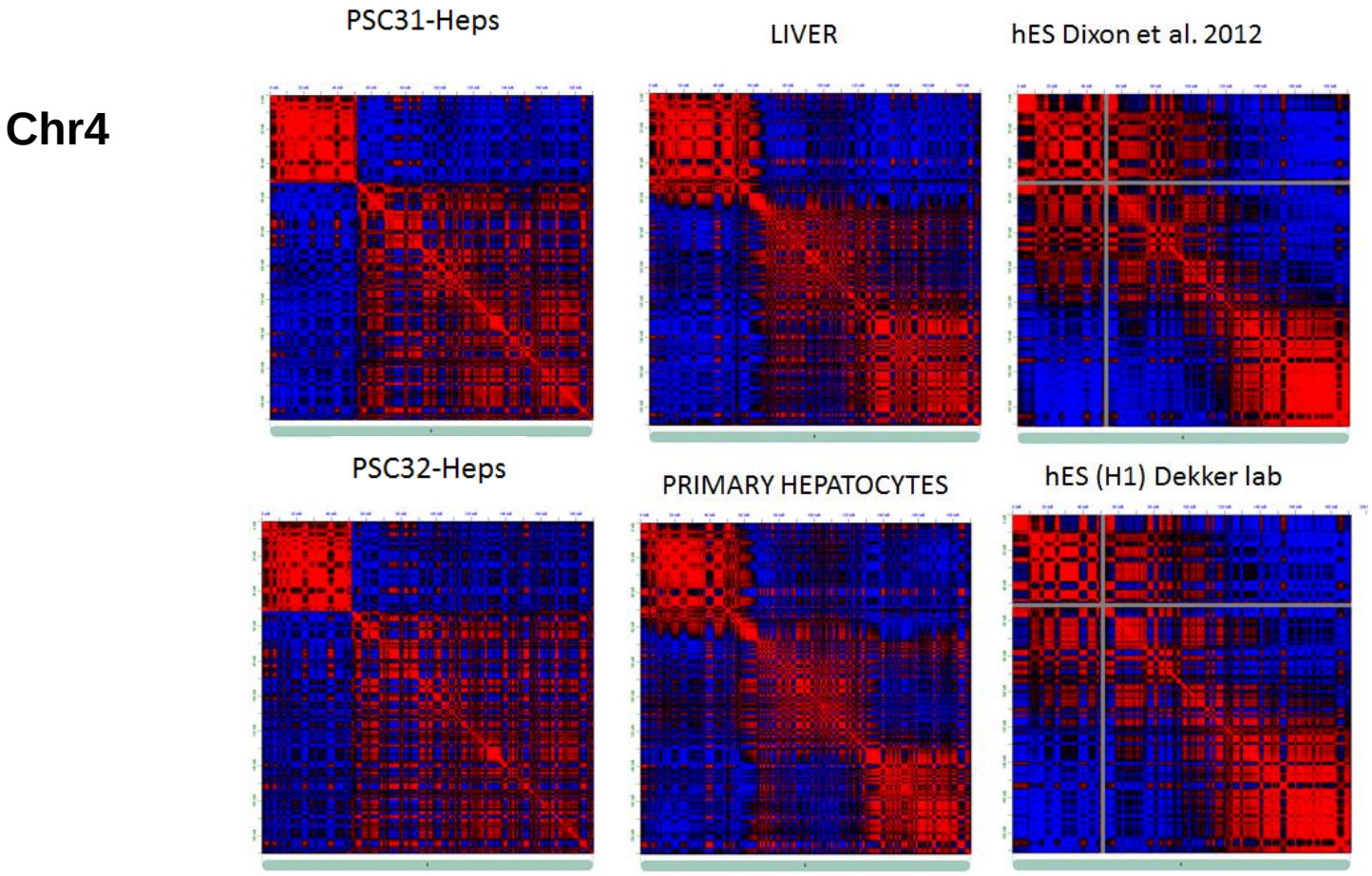




Adult Liver

1. Compartments : biological properties

A and B compartments are mostly conserved but exhibit cell specificities.



1. Compartments : biological properties

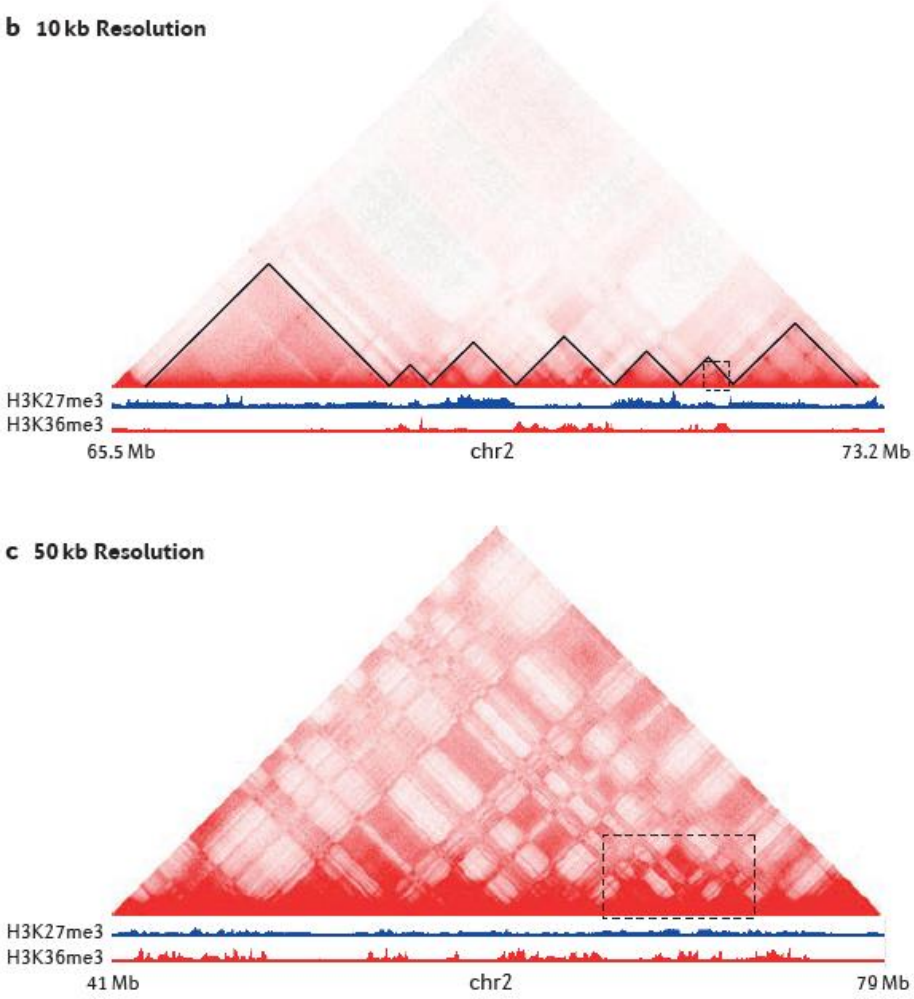
A and B compartments are mostly conserved but exhibit cell specificities.

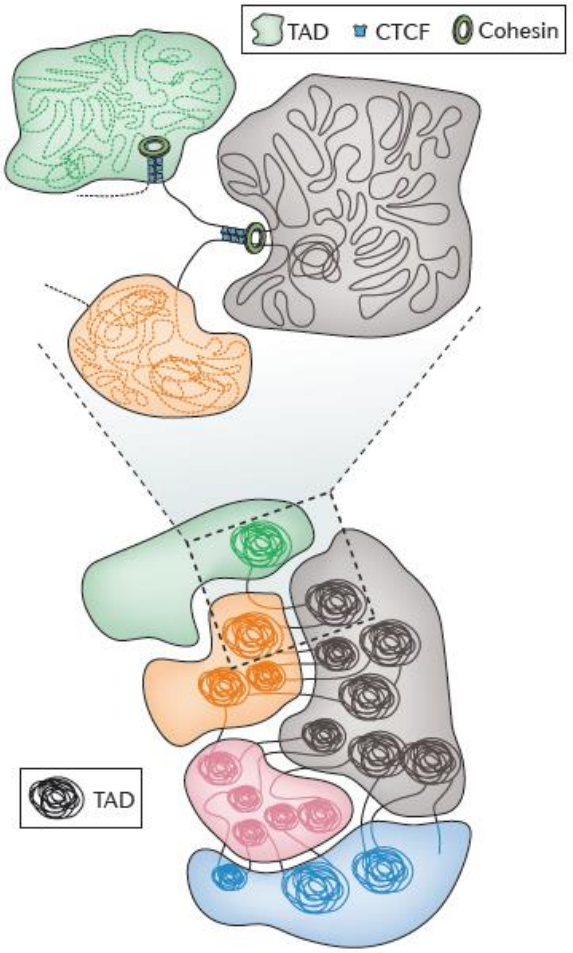
A/B COMPARTMENT SWITCHING

CONSERVATION: about 80%

SWITCHES: about 20%

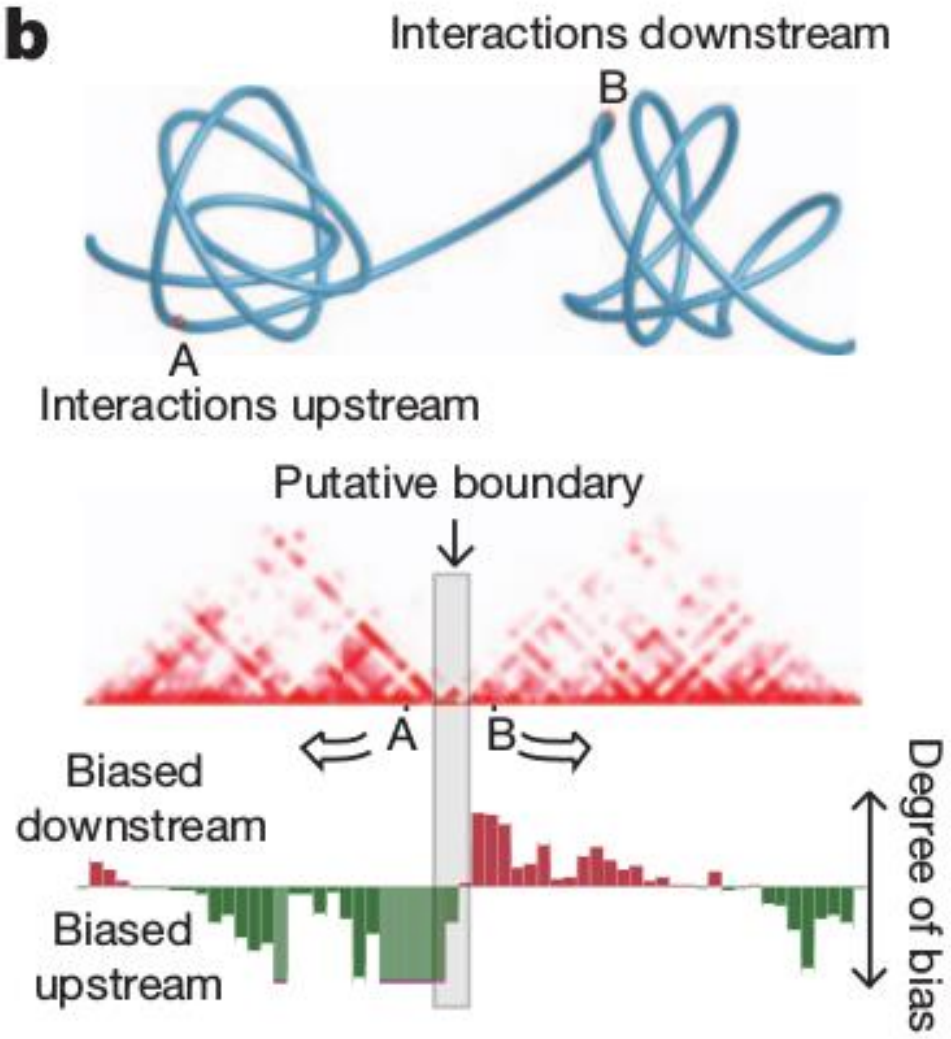
2. Topological Associated Domains (TADs)





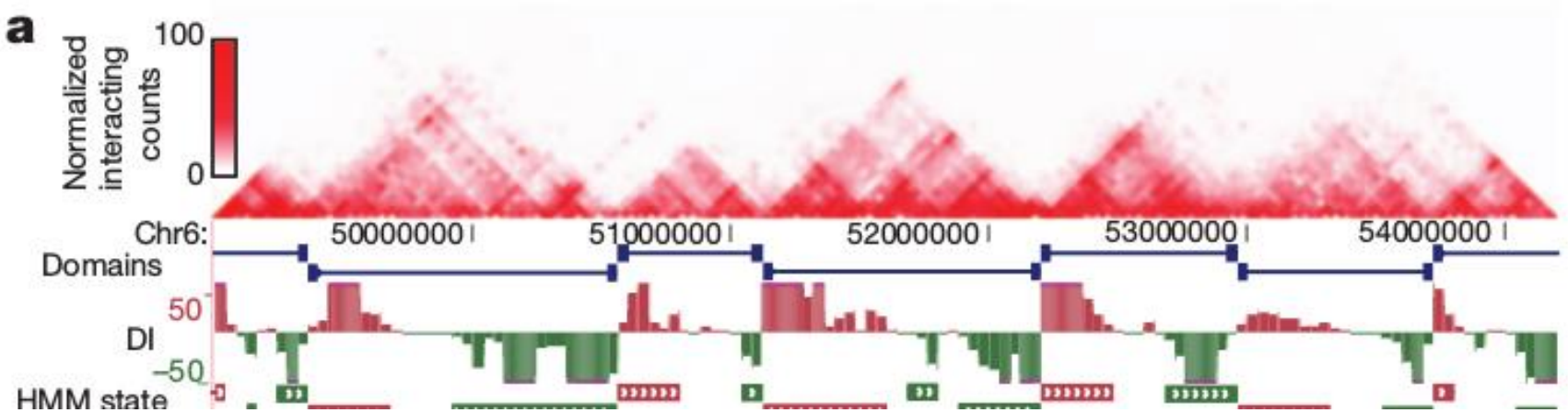
Bonev and Cavalli 2016

2. Identification of TADs



TADs are well detected using resolution (bins) smaller than 100kb

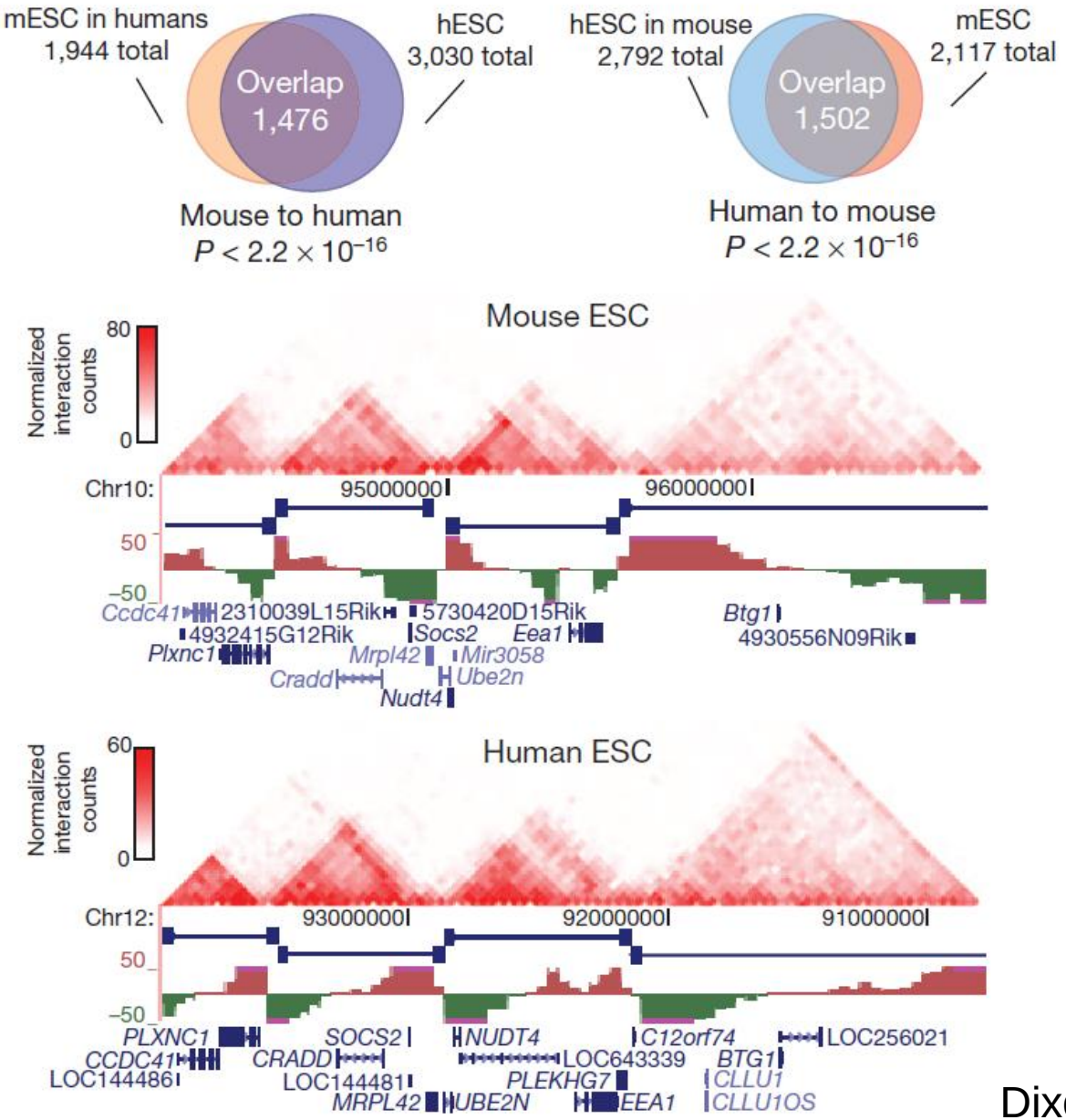
2. Identification of TADs: desequilibrium of interactions



from Dixon et al. 2012

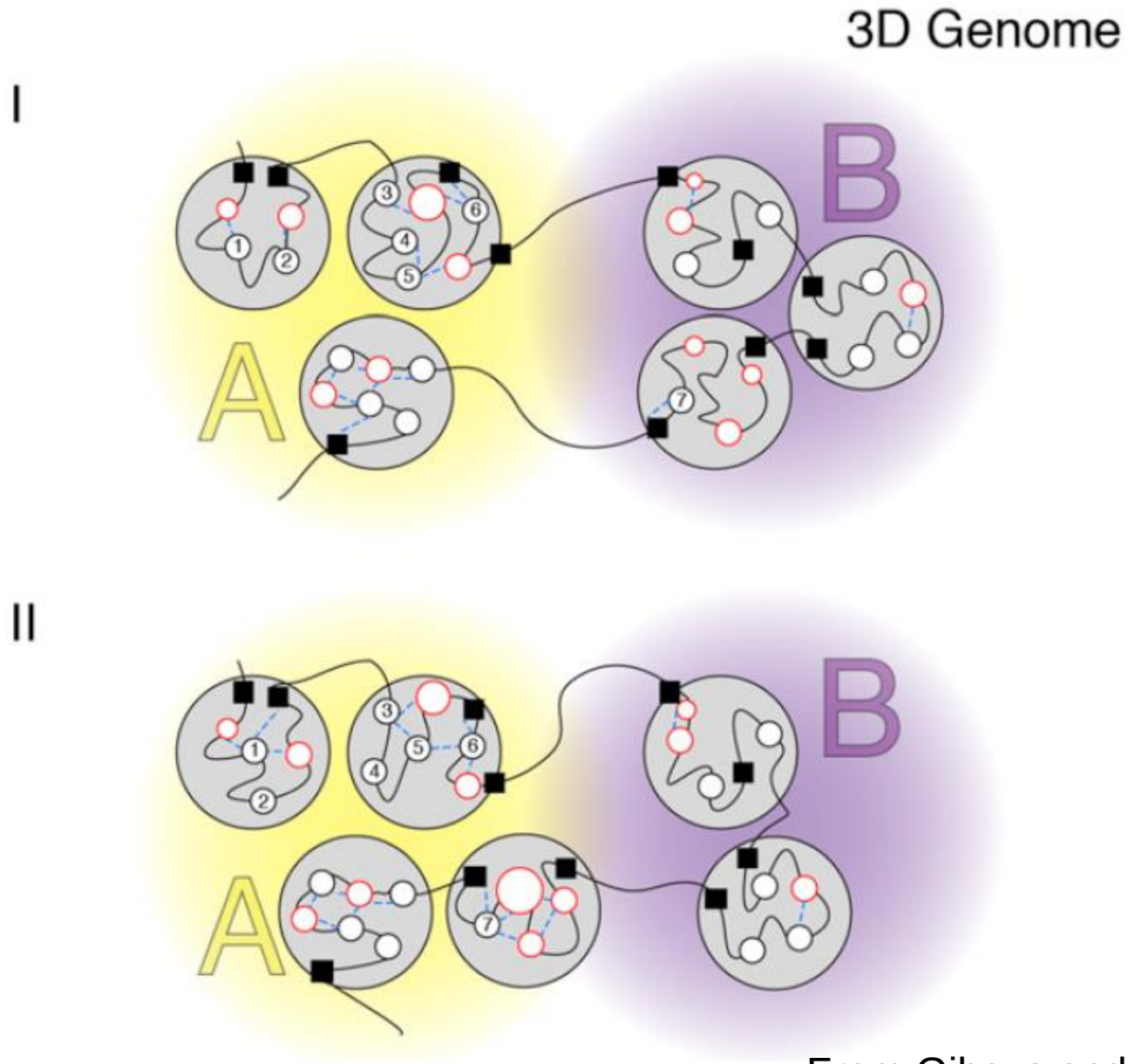
TADs' size: between 0.1 to 1Mb

2. TADs are well conserved between species



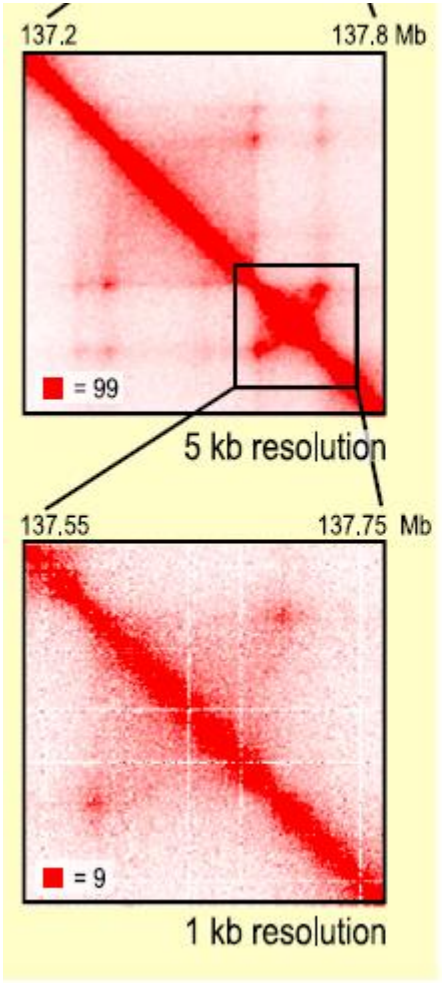
Dixon et al. 2012

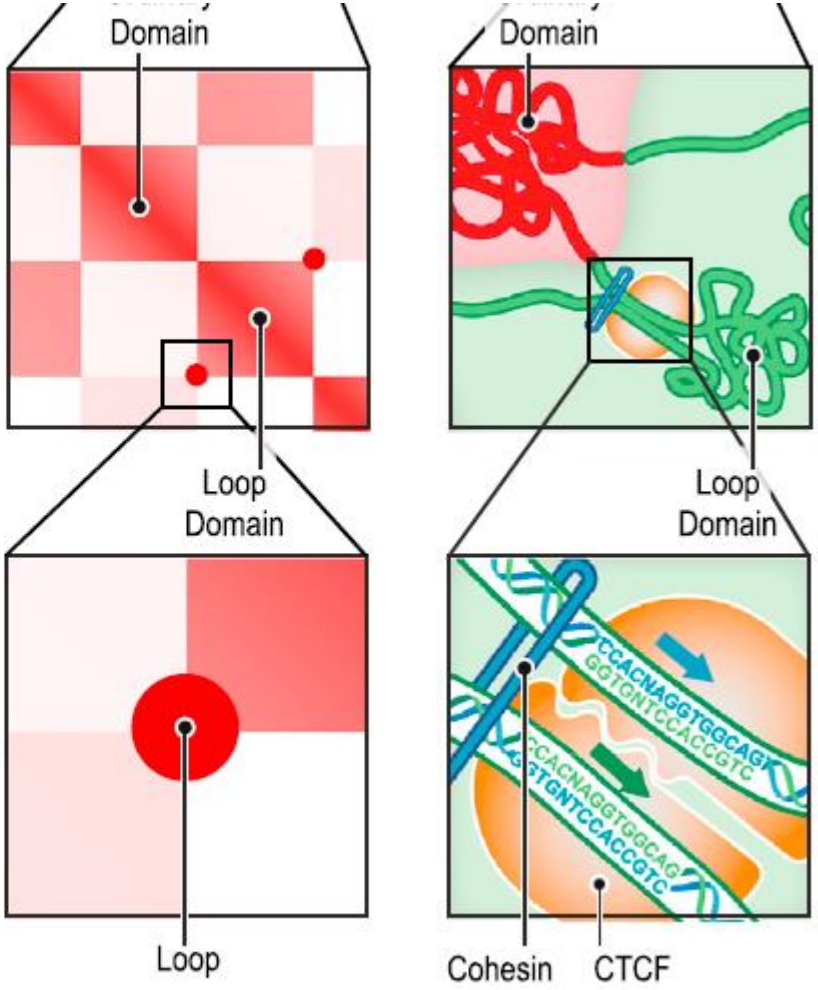
2. Balance between TADs and compartments



From Gibcus and Dekker 2013

3. TADs boundaries

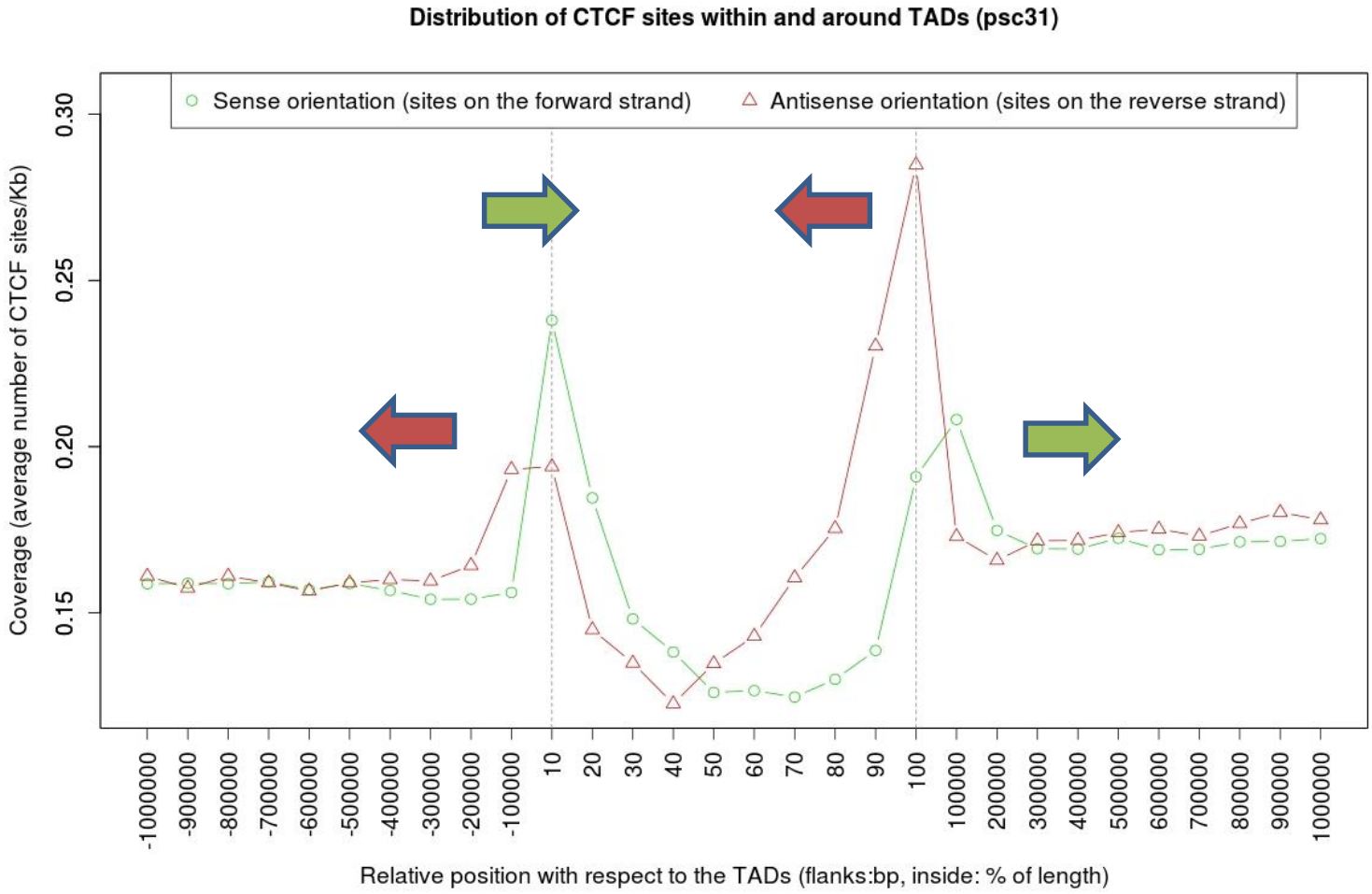




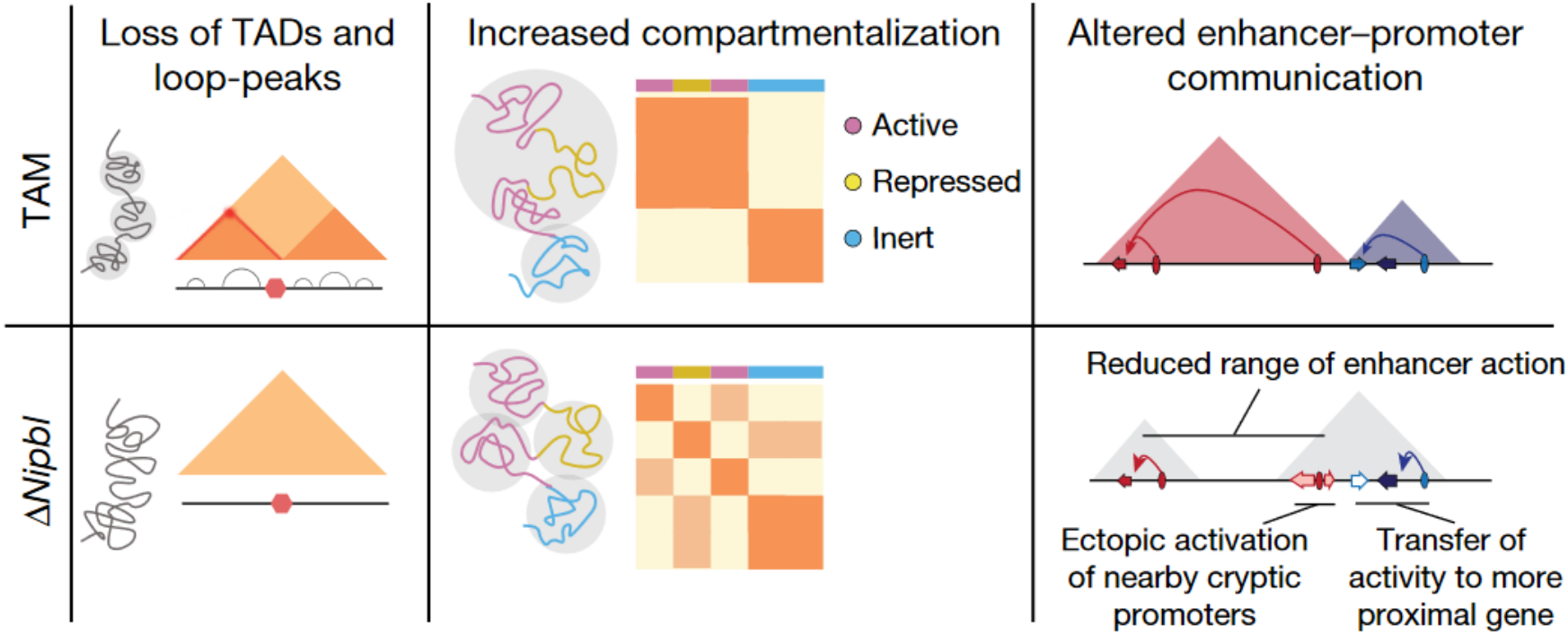
From Rao et al. 2014

3. TADs boundaries are enriched for CTCF binding sites

Comparative analysis of TADs with *in silico* predicted CTCF binding sites (whole genome)

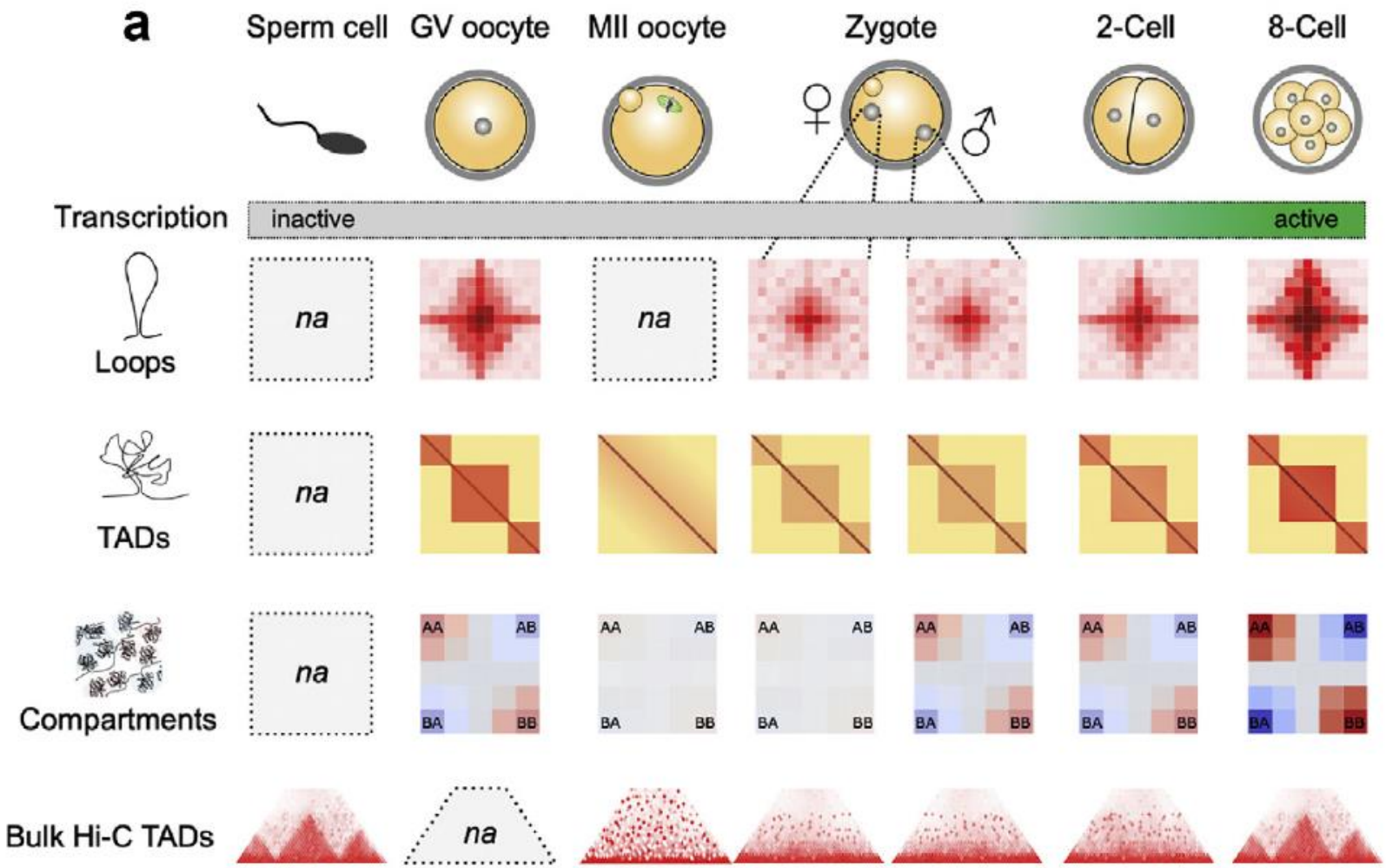


3. TADs boundaries required CTCF and cohesin



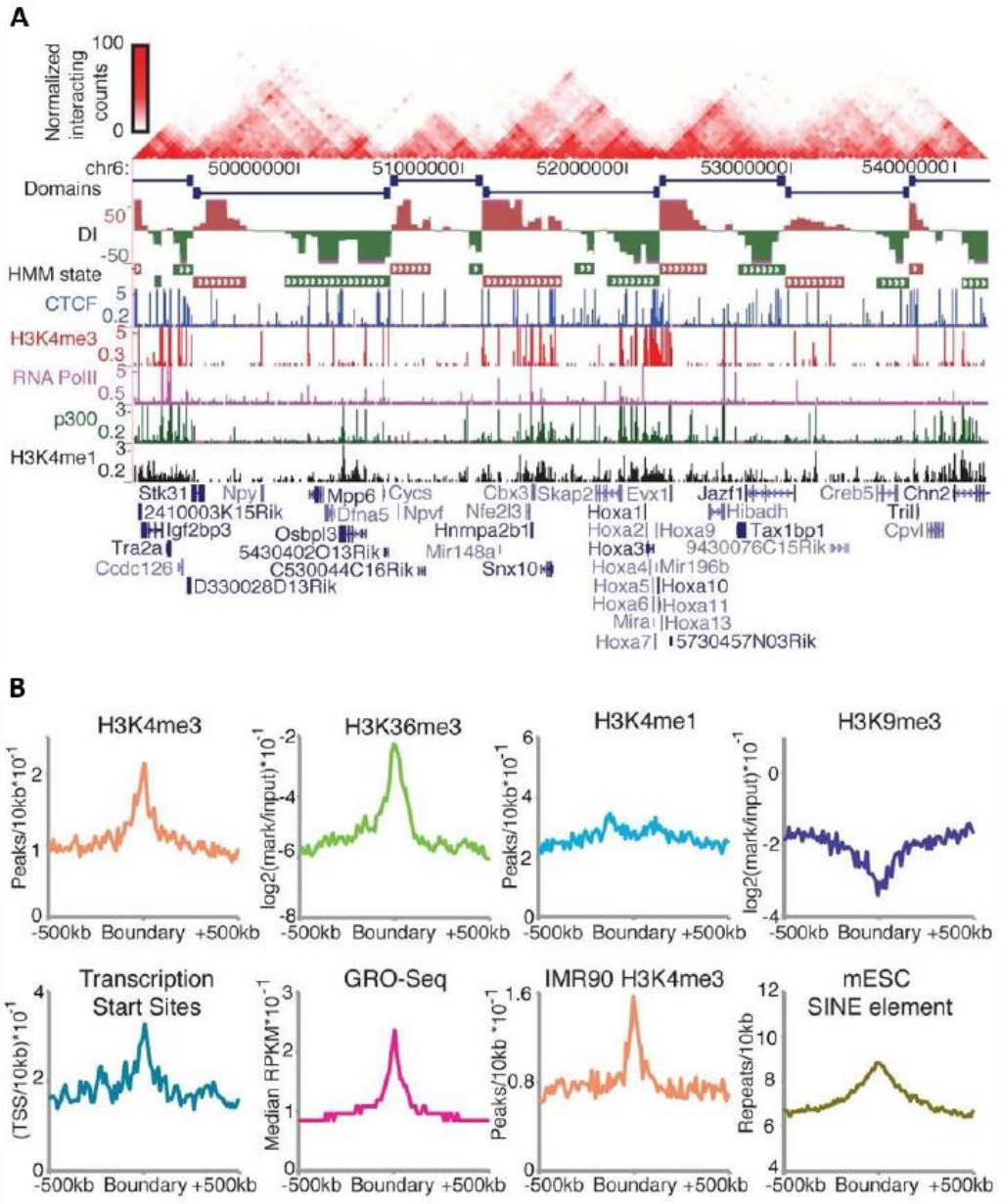
Schwarzer et al. 2017

3. TADs boundaries appear after ZGA



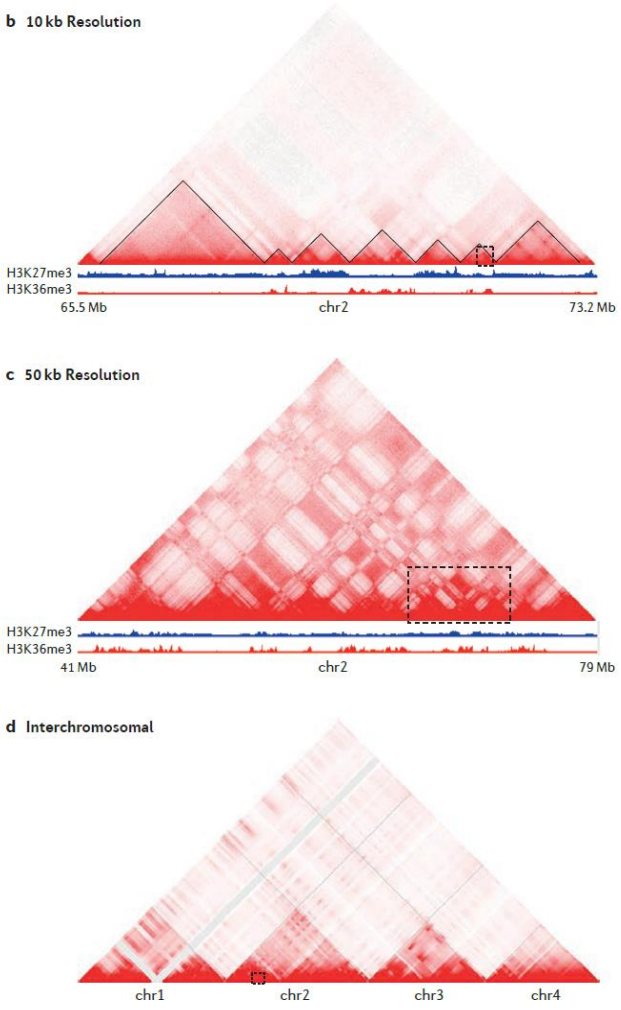
From Vallot and Tachibana 2020

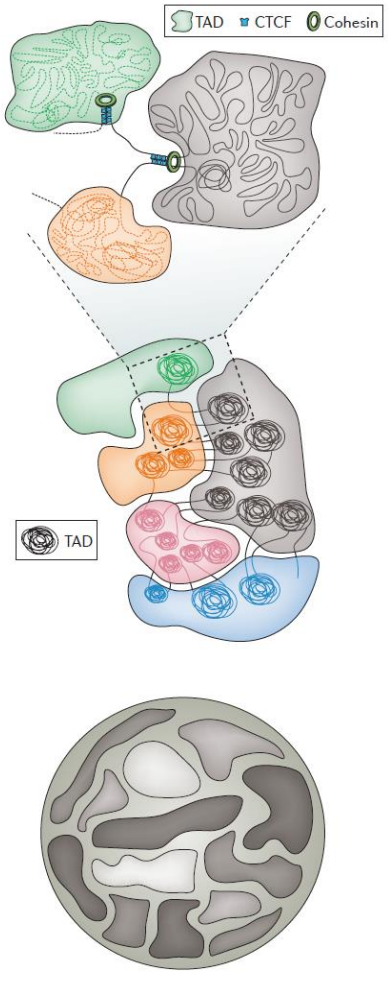
3. TADS boundaries are labeled by histone marks



Dixon et al. 2012

Functional annotations of spatial organization of genomes





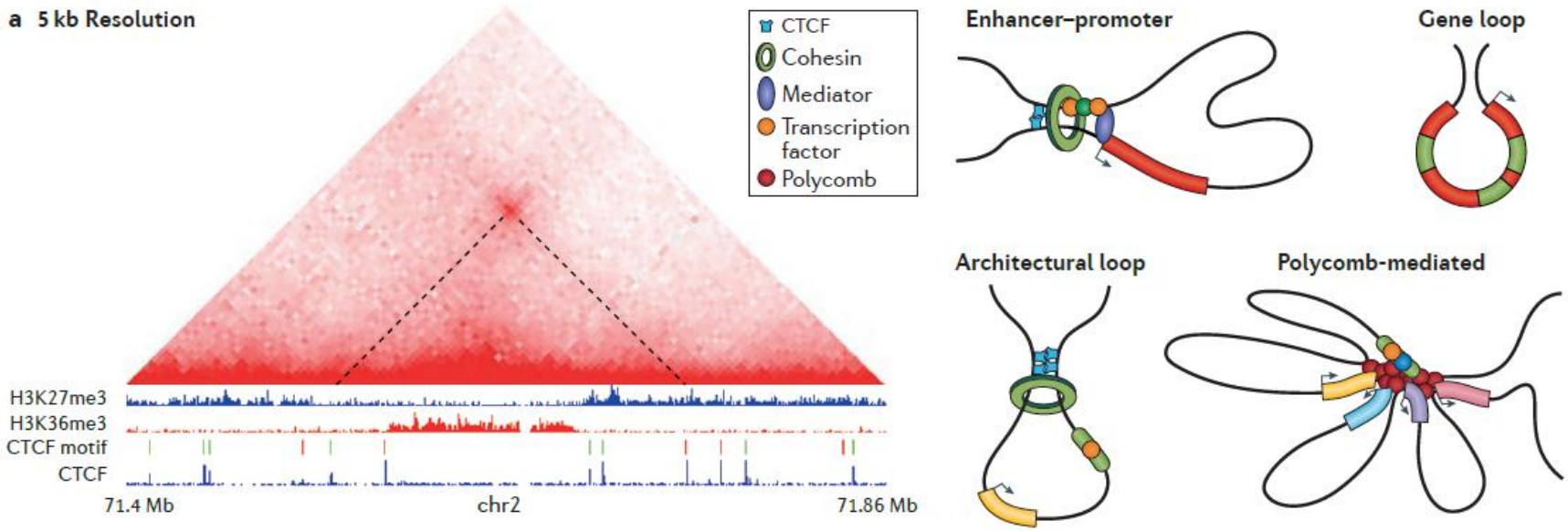
Loops CTCF: TADs' border
(~10-20kb)

Topologically Associated Domains (TADs)
(~0.5Mb)

A & B Compartment
(~5Mb)

From Bonev & Cavalli 2016

4. Could we get spatial annotations at the kb resolution ?



Getting such resolution is sequencing and ressources consuming.
For one mammalian sample, at 1kb resolution, more than 5G PE reads are needed.

4. Getting spatial annotations at high resolution

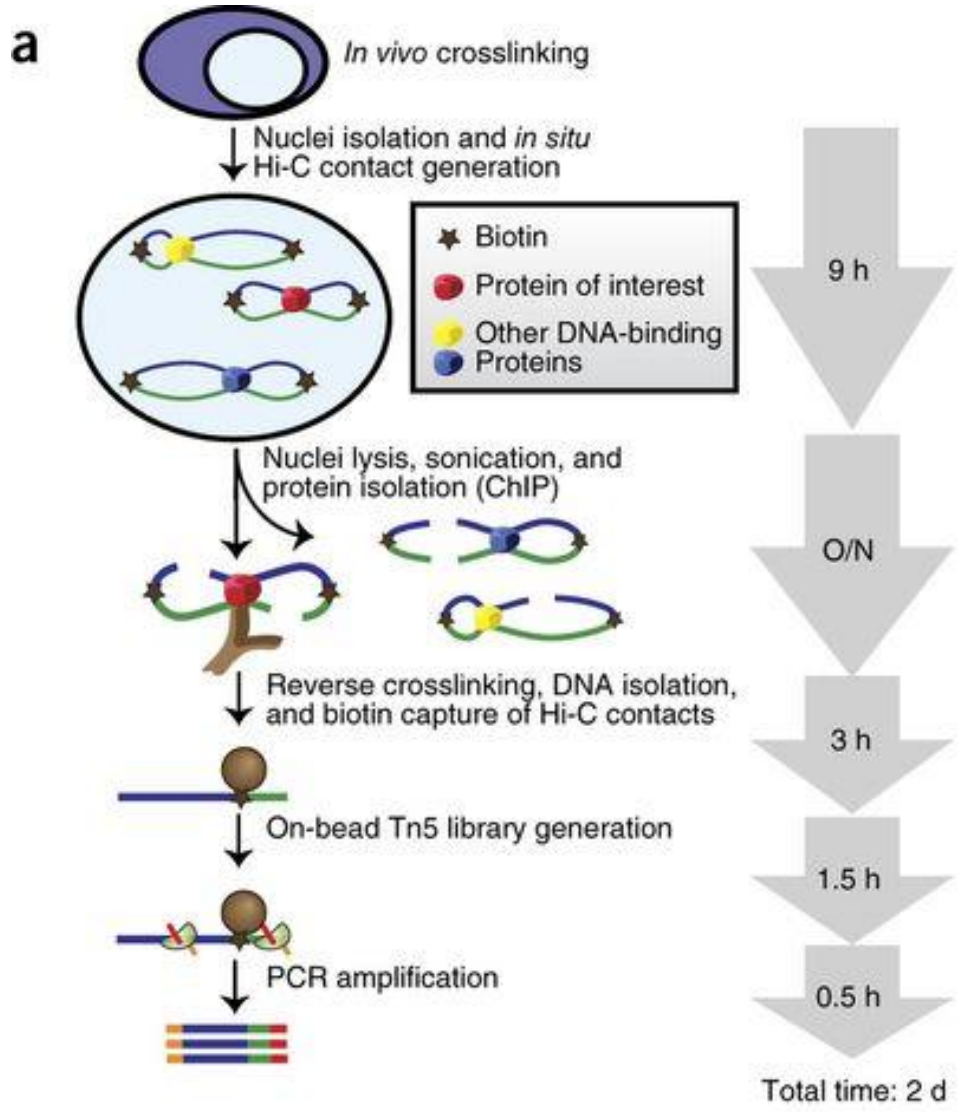
How to improve Hi-C resolution ?

By enriching in the information you need without sequencing all the information you are not interested in !!

Two different way:

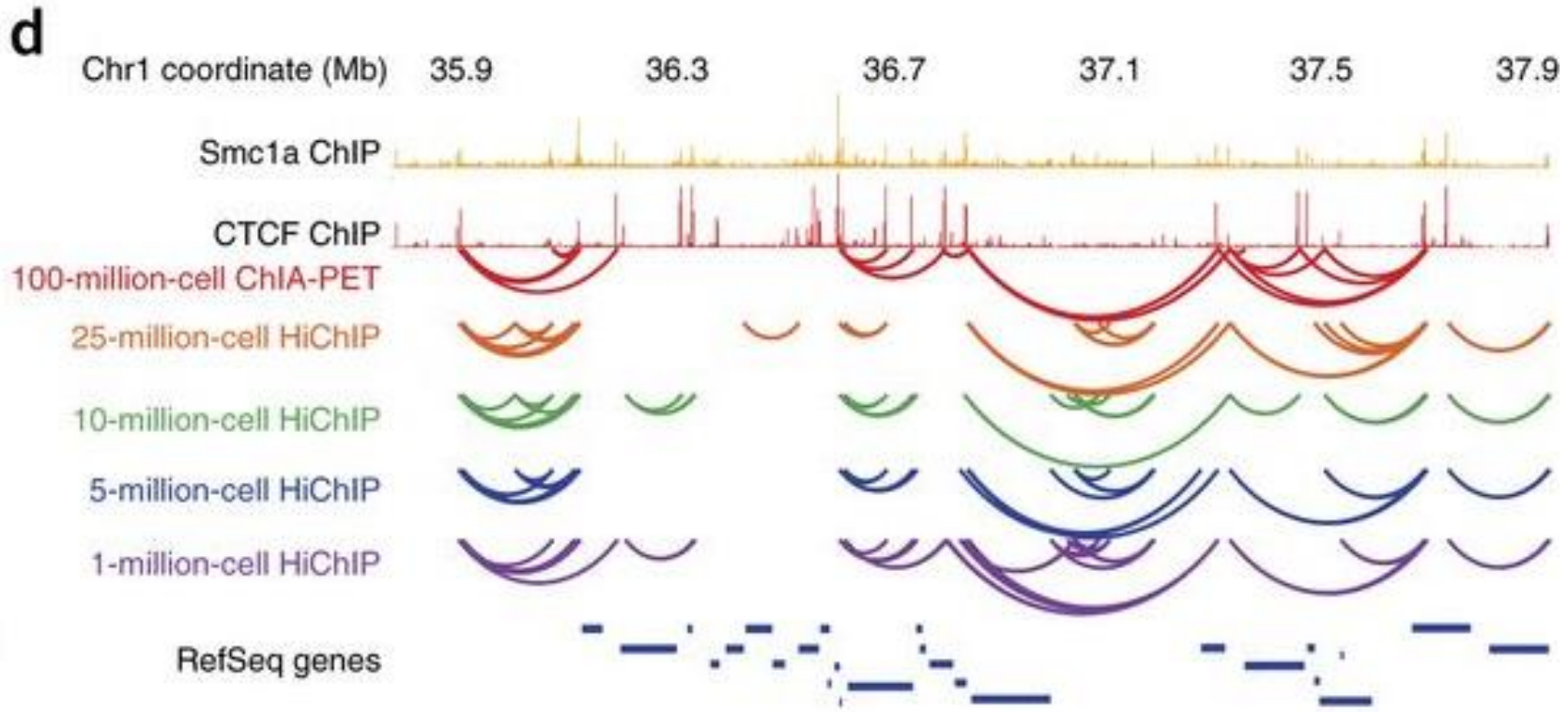
- Enriched for chromatin proteins – DNA complex of interests
- Capture genomic regions of interest

4. Capturing local interactions: PLAC-seq and Hi-ChIP



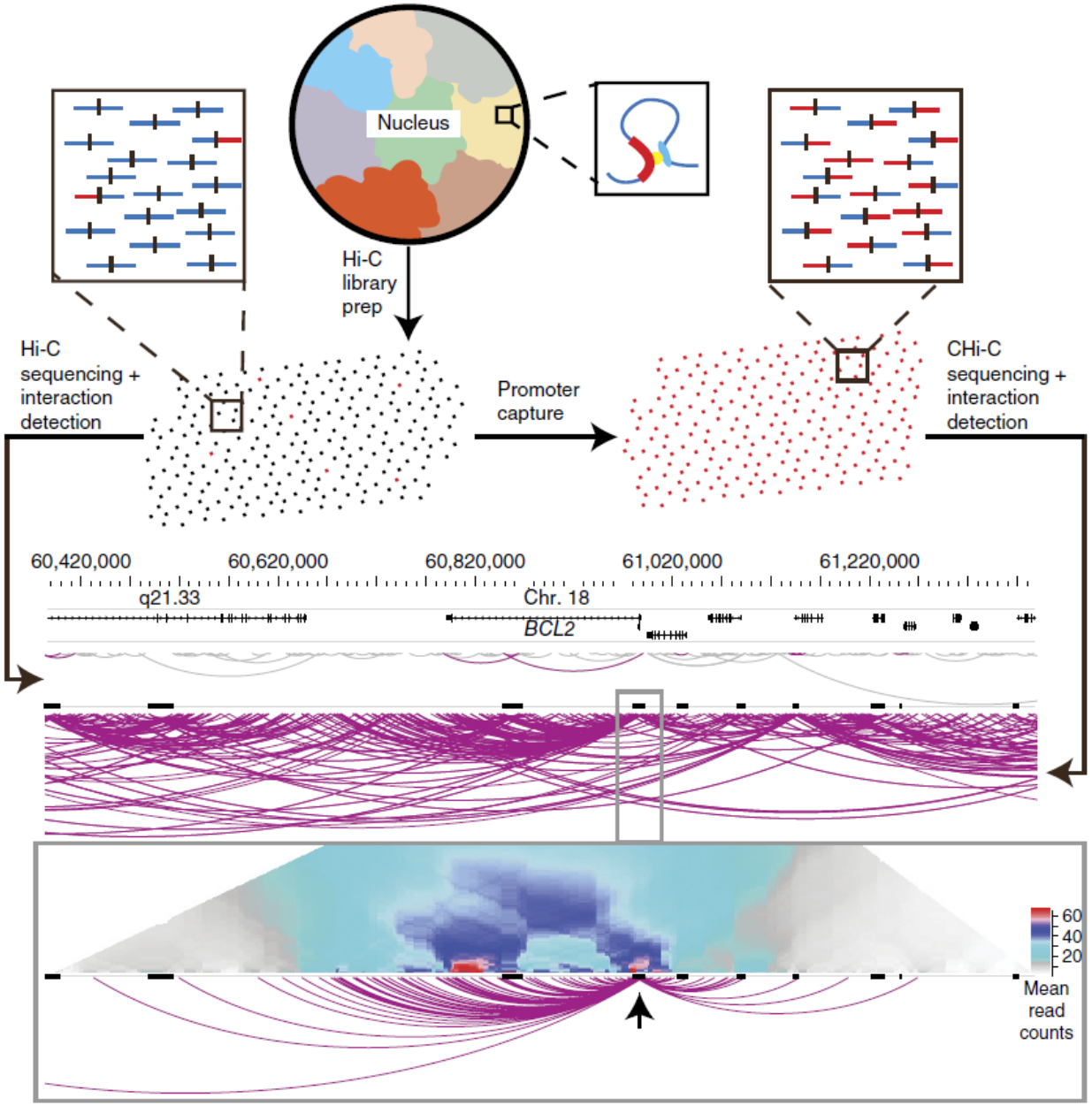
Mumbach et al. 2016

4. Capturing local interactions: PLAC-seq and Hi-ChIP



Mumbach et al. 2016

4. Capturing local interactions: Capture HiC



Mifsud et al. 2015

4. Capturing local interaction to analyse the link between Genotype to Phenotype

Resource

Cell

Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters

Javierre et al. 2017

ARTICLE

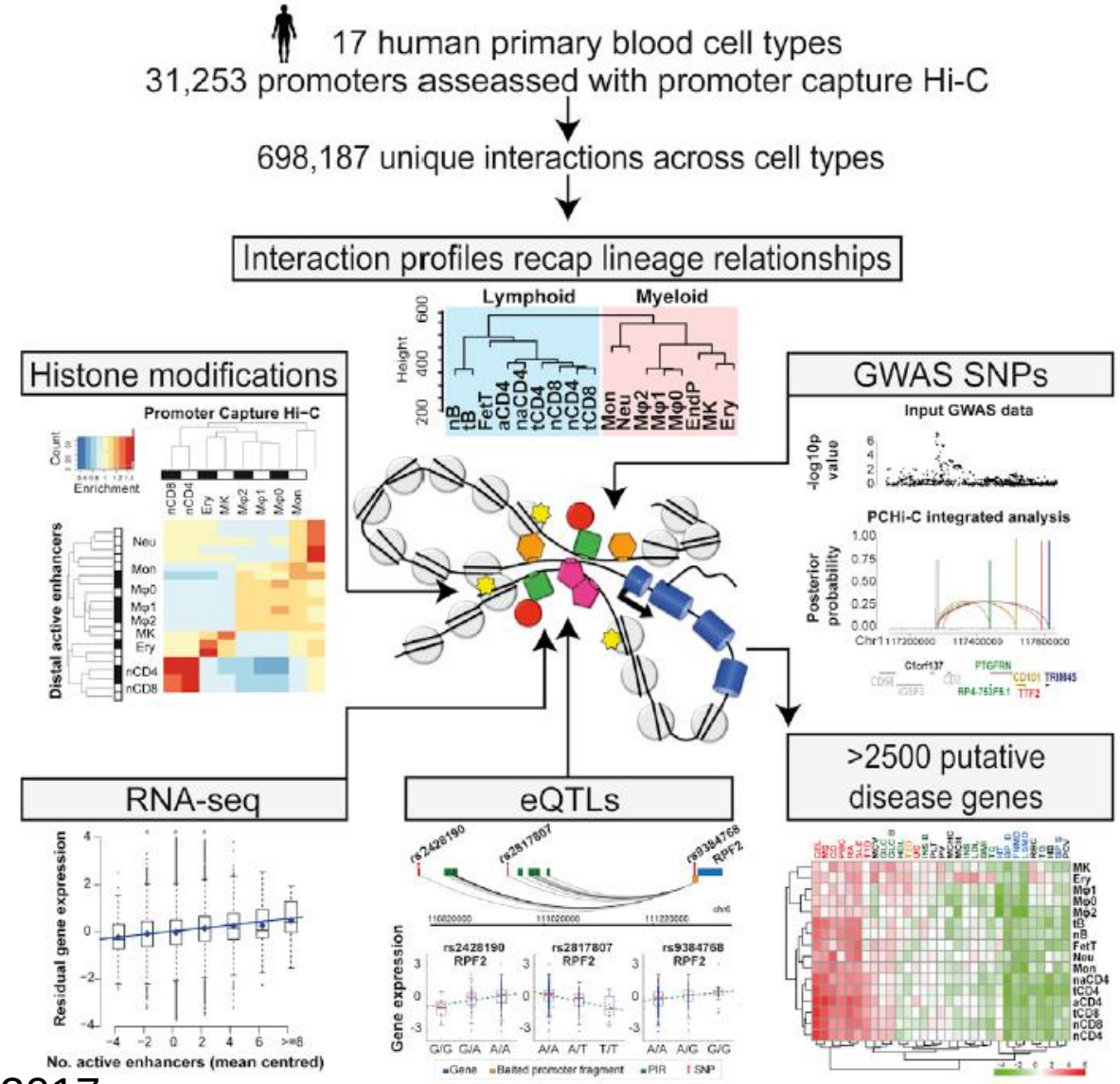
DOI: 10.1038/s41467-018-03411-9 OPEN

Capture Hi-C identifies putative target genes at 33 breast cancer risk loci

Joseph S. Baxter[1], Olivia C. Leavy[2,3], Nicola H. Dryden[1], Sarah Maguire[1], Nichola Johnson[1], Vita Fedele[1], Nikiana Simigdala[1], Lesley-Ann Martin[1], Simon Andrews[4], Steven W. Wingett[4], Ioannis Assiotis[5], Kerry Fenwick[5], Ritika Chauhan[1], Alistair G. Rust[1], Nick Orr[1], Frank Dudbridge[2,3], Syed Haider[1] & Olivia Fletcher[1]

2018

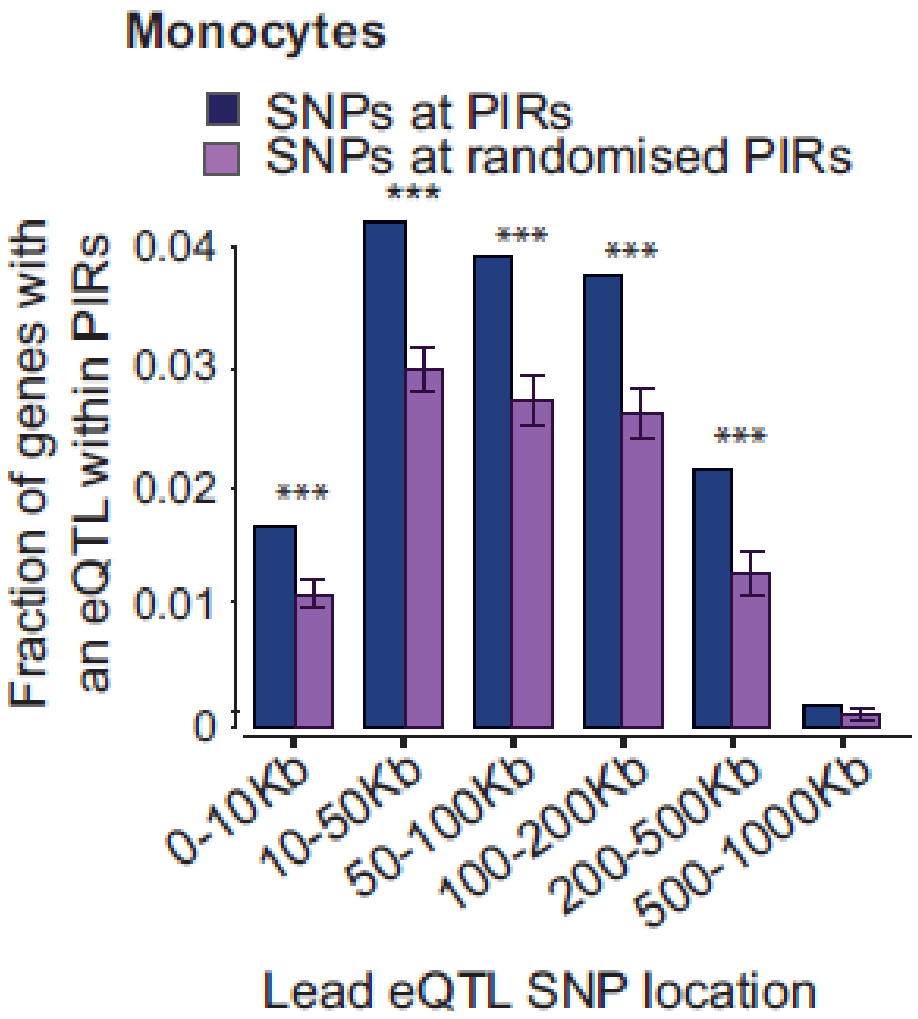
6. Capturing local interaction to analyse the link between Genotype to Phenotype



Javierre et al. 2017

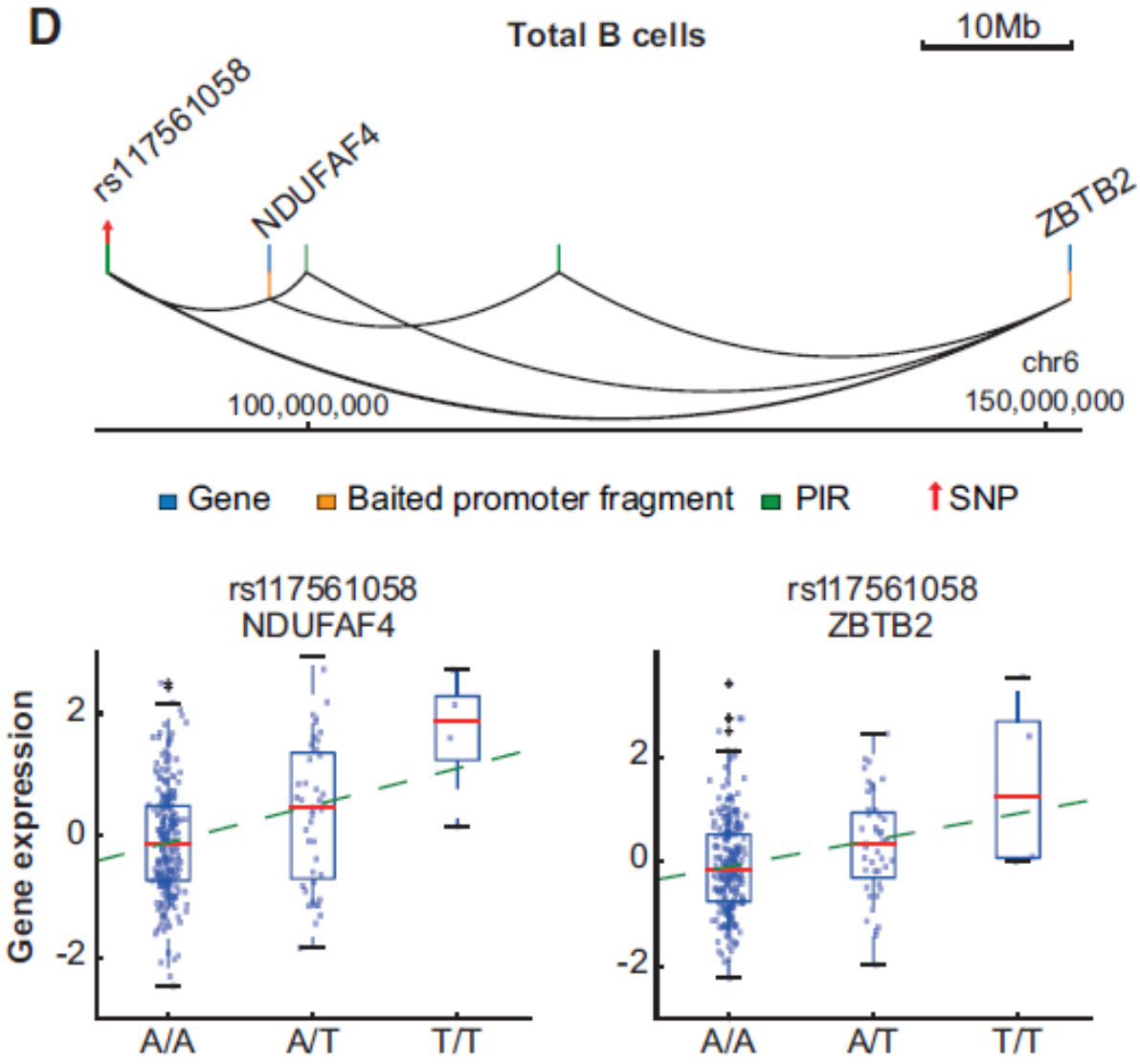
6. Capturing local interaction to analyse the link between Genotype to Phenotype

SNPs are enriched at promoter interacting regions



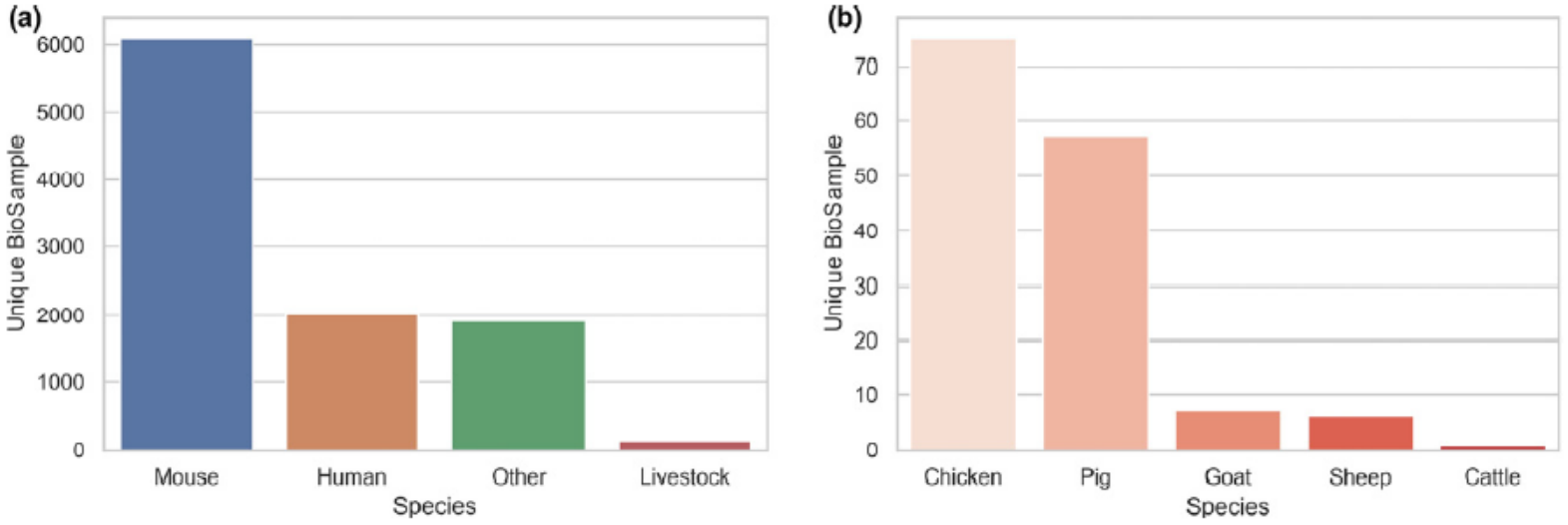
6. Capturing local interaction to analyse the link between Genotype to Phenotype

SNPs at PIRs are associated with differential gene expression



6. Capturing local interaction to analyse the link between Genotype to Phenotype

For livestock species and animal breeding ?



MacPhillamy et al. 2021

To date, really few datasets available (missing high resolution data).

Conclusions

Hi-C is a revolutionary technology allowing to decipher the function of spatial genome organization.

It provides key information, from basic knowledge to a wide range of applications.