Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 12

Pierre-Luc Germain

ETH Zürich

Plan for today

- Debriefing on the assignment

- **Theory:**

  Chromatin conformation & related technologies

- **Practice:**
  Using long-range interactions to annotate distal sites

Debriefing on the assignments

Plot a heatmap of the methylation levels of **the genes in top 5 DMR** regions located on chr1.

A DMR can comprise several genes and surrounding regions or even no gene!

Debriefing on the assignments

Plot a heatmap of the methylation levels of **the genes in top 5 DMR** regions located on chr1.

*A DMR can comprise **several genes and surrounding regions** or even no gene!*

```
> head(subset(dmr, grepl(",", dmr$overlapping.genes)),5)[,c("no.cpgs", "overlapping.genes")]
GRanges object with 5 ranges and 2 metadata columns:
      seqnames            ranges strand |   no.cpgs      overlapping.genes
         <Rle>         <IRanges>  <Rle> | <integer>            <character>
  [1]    chr19   2250311-2253429      * |        94    AMH, MIR4321, JSRP1
  [2]    chr14 24780119-24780723      * |        38 LTB4R2, LTB4R2, LTB4..
  [3]     chr8   9008196-9009307      * |        79  RP11-10A14.4, PPP1R3B
  [4]    chr22 46448594-46450296      * |        29 RP6-109B7.5, RP6-109..
  [5]    chr19   1467273-1470809      * |       118         APC2, C19orf25
  -------
  seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

Debriefing on the assignments

Plot a heatmap of the methylation levels of **the genes in top 5 DMR** regions located on chr1.

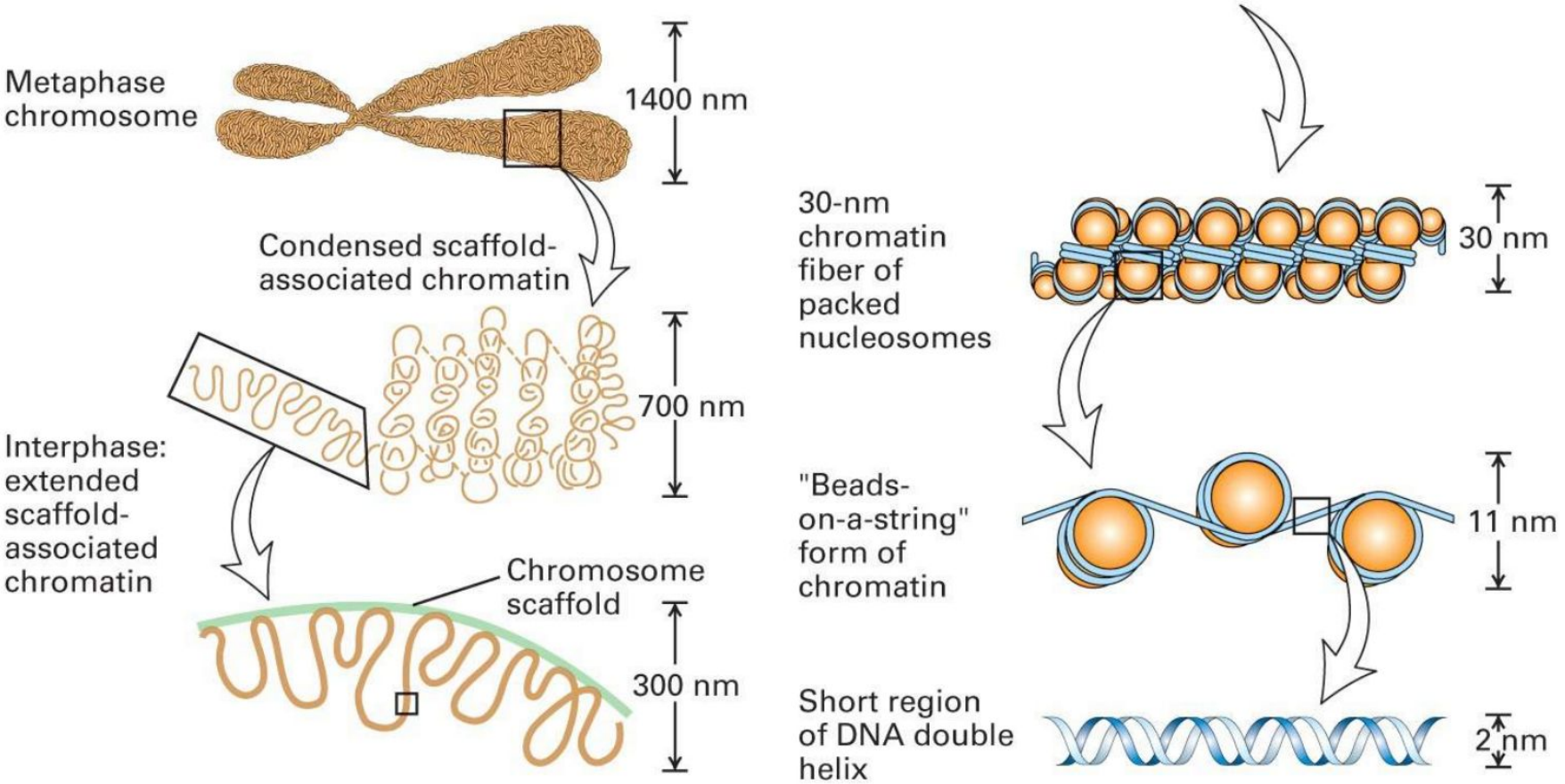
*A DMR can comprise several genes and surrounding regions or even **no gene**!*

```
> head(subset(dmr, is.na(dmr$overlapping.genes)), 5)[,c("no.cpgs", "overlapping.genes")]
GRanges object with 5 ranges and 2 metadata columns:
      seqnames              ranges strand |   no.cpgs overlapping.genes
         <Rle>           <IRanges>  <Rle> | <integer>       <character>
  [1]     chr1 240160779-240161545      * |        47              <NA>
  [2]     chr2 225307000-225307498      * |        42              <NA>
  [3]    chr18   13135892-13137131      * |        52              <NA>
  [4]    chr19   48076361-48076802      * |        23              <NA>
  [5]     chr1 209404930-209405333      * |        36              <NA>
  -------
  seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

Debriefing on the assignments

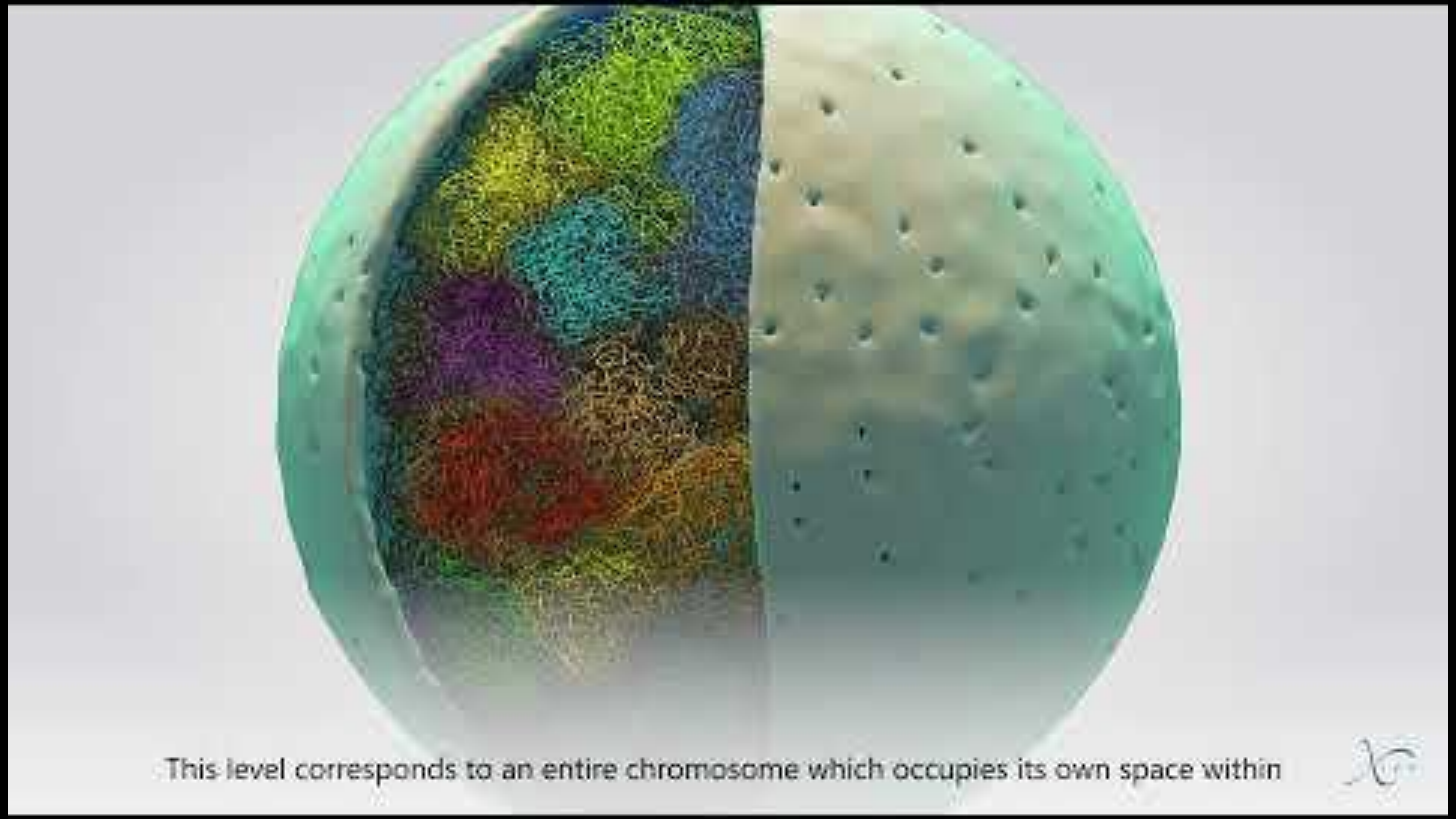
Plot a heatmap of the methylation levels of **the genes in top 5 DMR** regions located on chr1.

```r
# get the gene coordinates
ah <- AnnotationHub()
ensdb <- ah[["AH109336"]]
genesChr1 <- genes(ensdb, columns=c("gene_seq_start", "gene_seq_end", "gene_name"))
genesChr1 <- genesChr1[genesChr1@seqnames=="1"]
seqlevelsStyle(genesChr1) <- "UCSC"

# get the genes within the top DMRs
dmrRanges <- readRDS("dmr.rds")
dmrRanges[order(dmrRanges$min_smoothed_fdr)]
genesDmr <- dmrRanges[!is.na(dmrRanges$overlapping.genes)]
dmrGenes <- genesChr1[genesChr1$gene_name %in% genesDmr$overlapping.genes]

# get methylation levels of genes within DMRs
metGenes <- bsseq::getMeth(bs, regions=dmrGenes, what="perRegion")
```



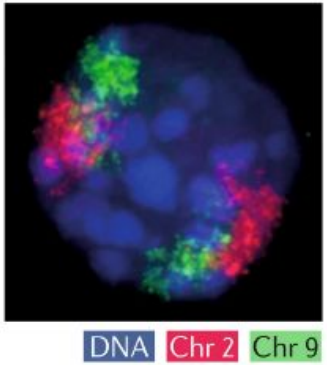
Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

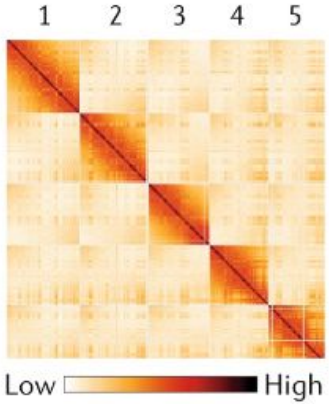
3D organization of the genome

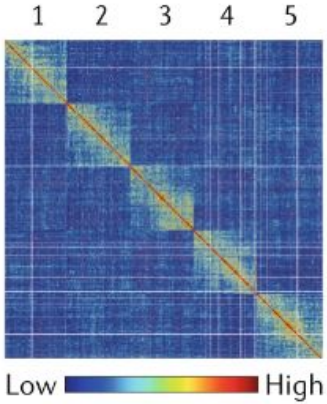


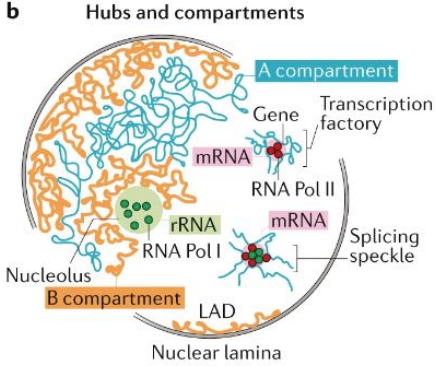
https://youtu.be/Pl44JjA--2k

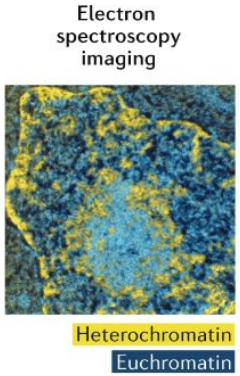
The nucleus is organized into chromosome territories







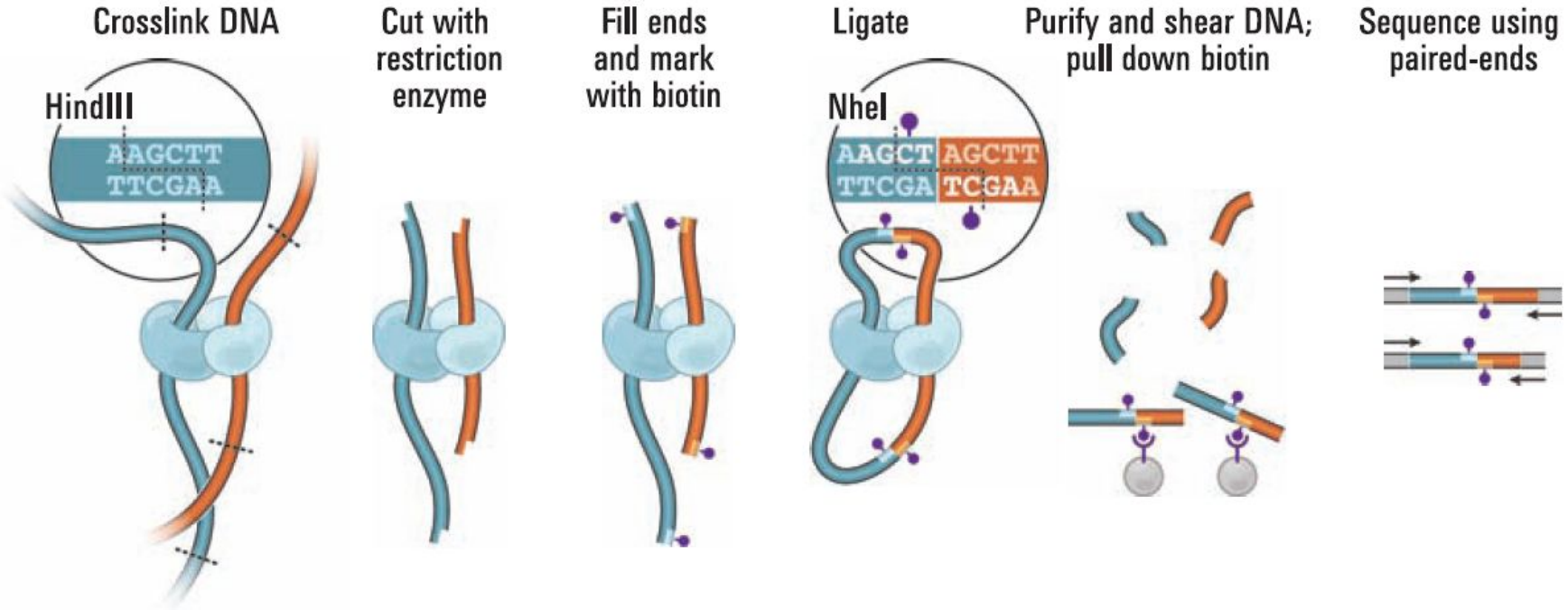




Chromosomes are split into active (A) and inactive (B) compartments

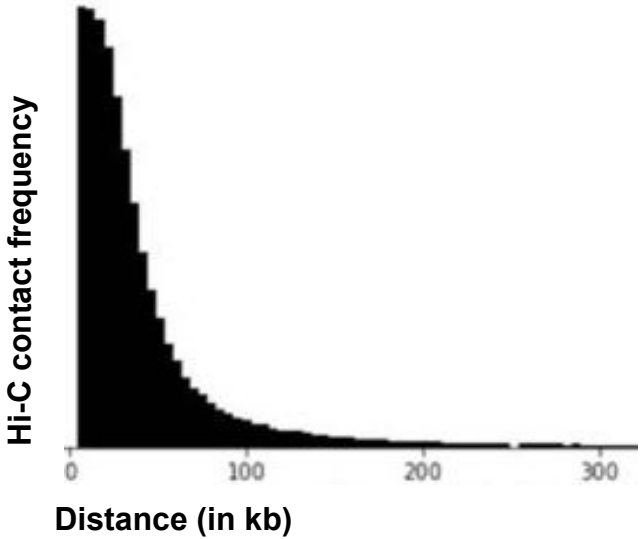
(Adapted from Kempfer and Pombo 2020)

Sequencing-based conformation capture



(Lieberman-Aiden et al. 2009)

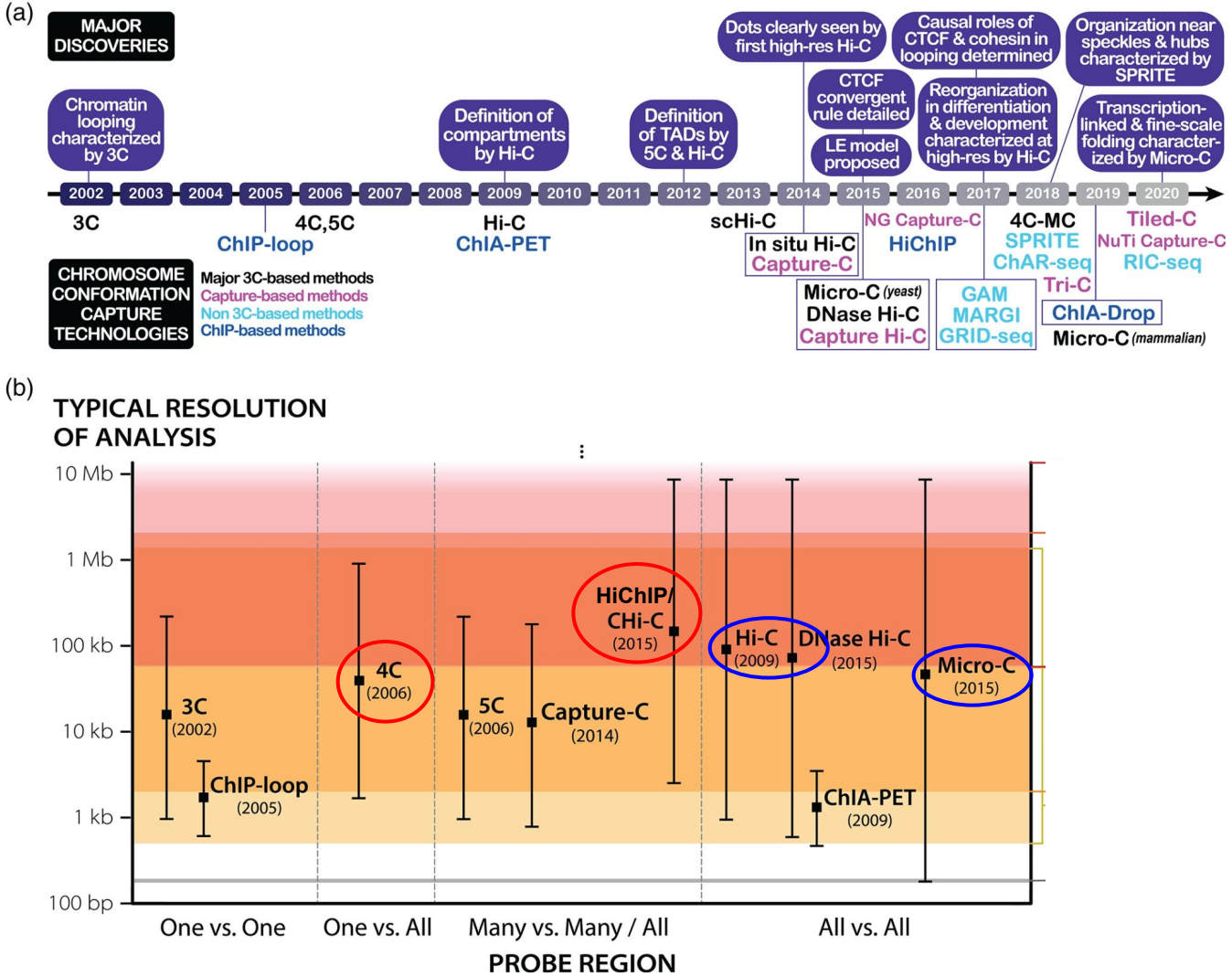
Sequencing-based conformation capture



Under the null hypothesis of random 3D organization, we expect to have more contacts from DNA regions that are closer to each other.

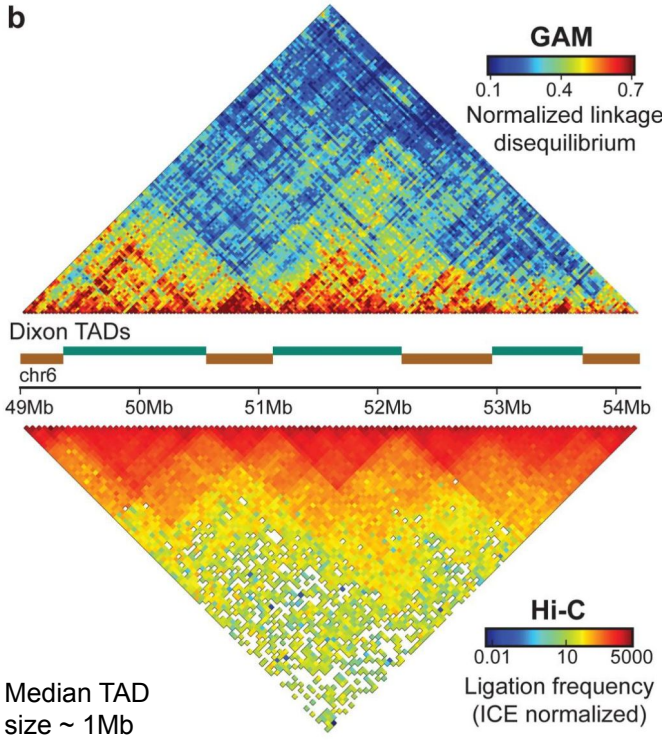
This is why, when identifying interactions, methods look for an excess over this null hypothesis

Sequencing-based conformation capture



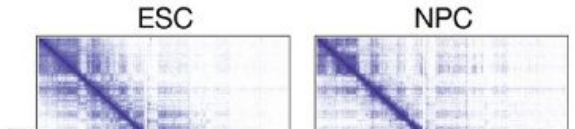
(Adapted from Goel and Hansen 2020)

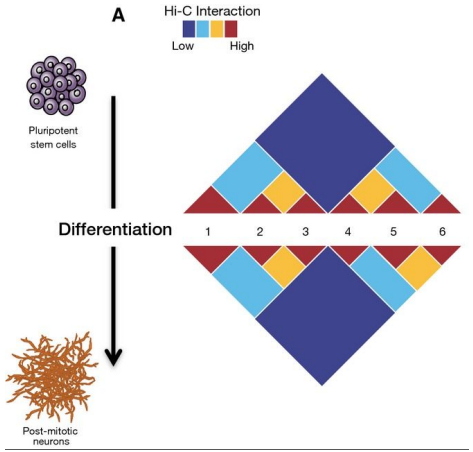
Chromosomes are organized into topologically associated domains (TADs)



Median TAD size ~ 1Mb

(Beagrie et al., 2017)

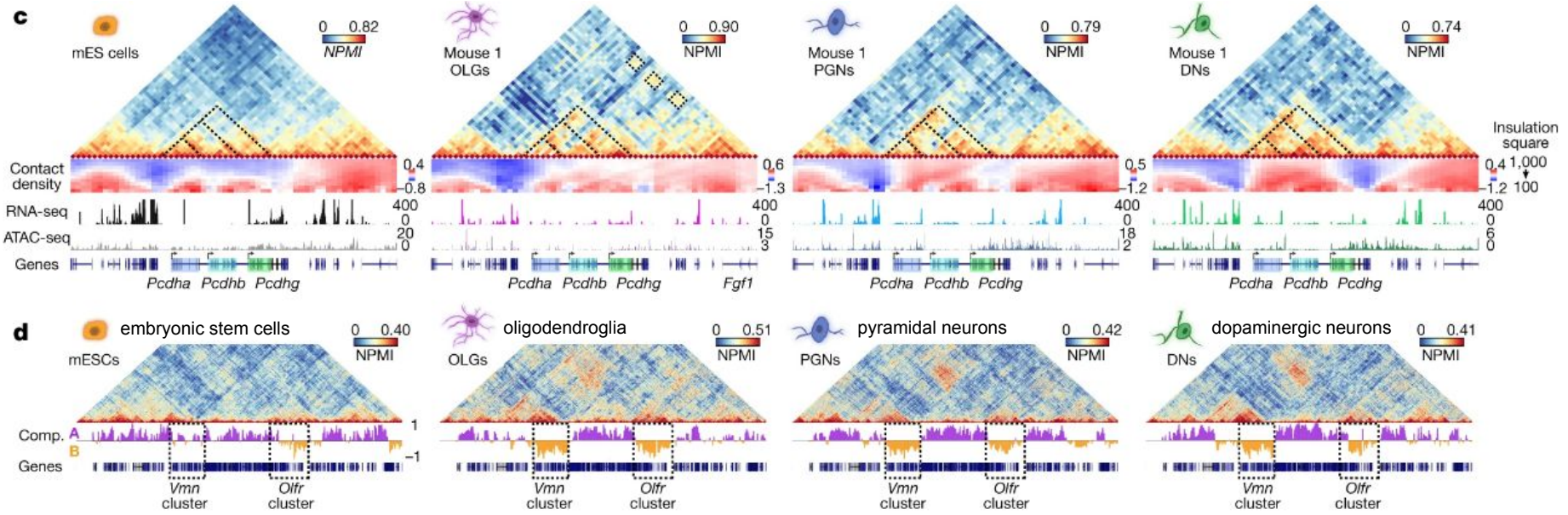




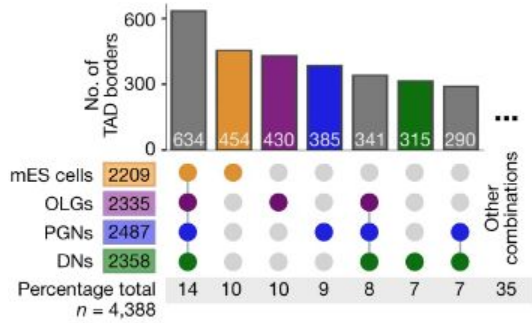
(Fraser et al., 2015)

...that are rather stable across cell types

TADs rearrangement across cell types



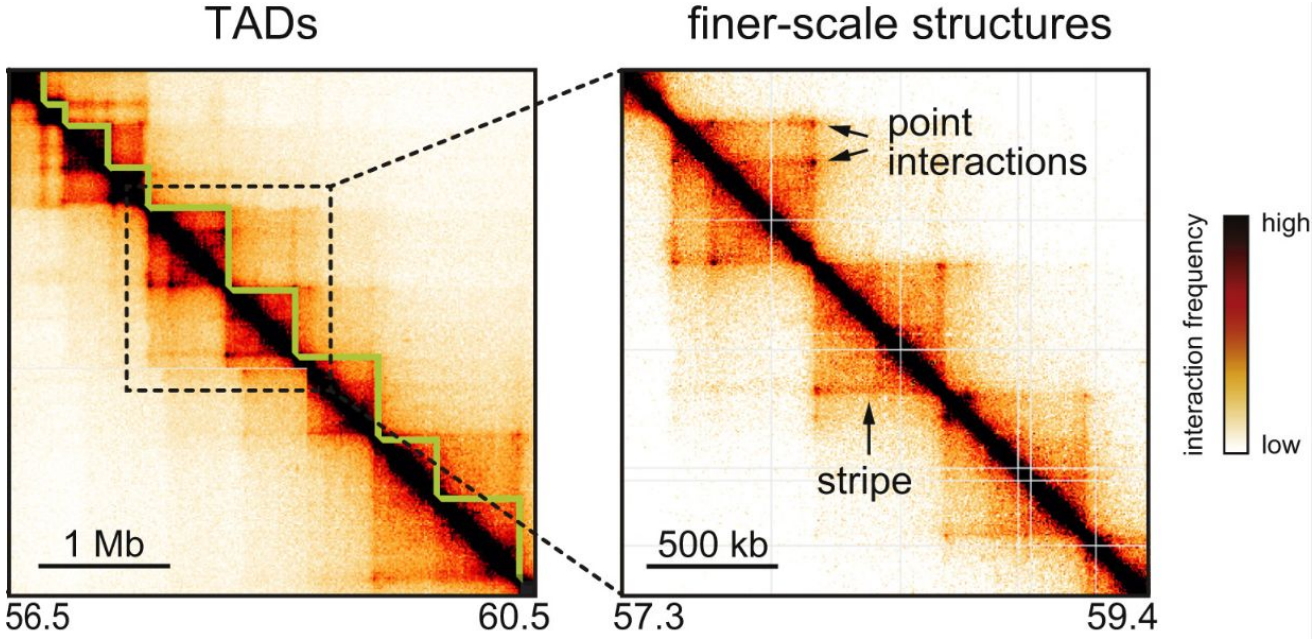
Comparison of TADs borders:



(Winick-Ng et al., *Nature* 2021)

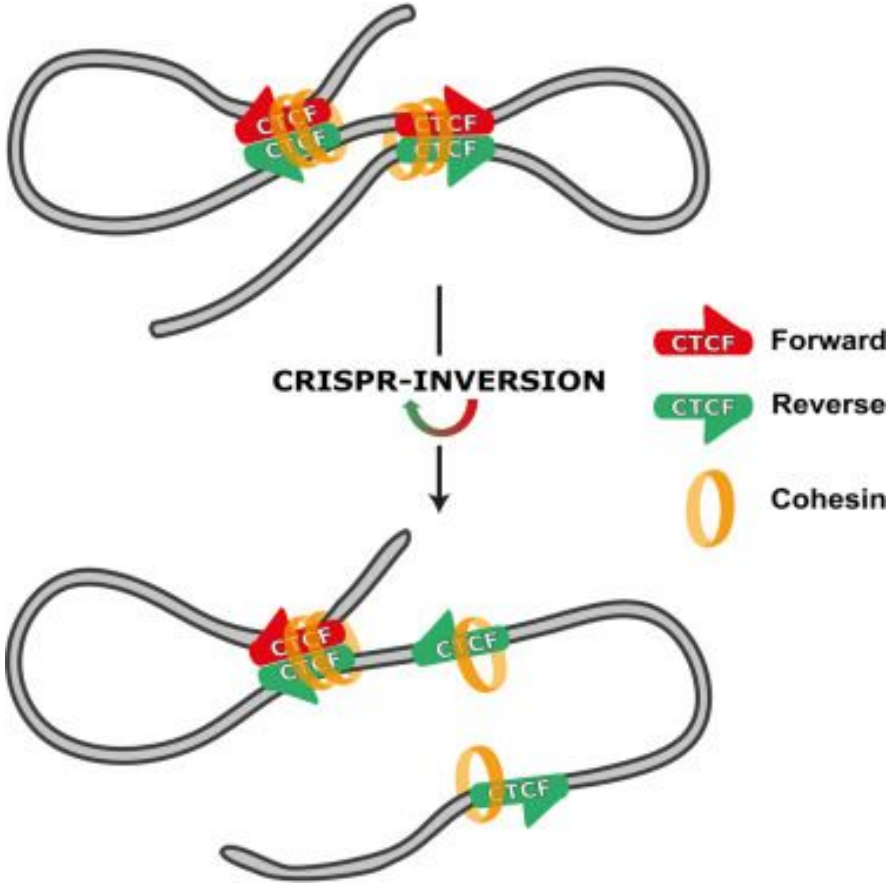
TAD boundaries are defined by very stable point interactions…

…which represent CTCF binding sites



(Adapted from McCord, Kaplan and Giorgetti, Mol Cell 2020)

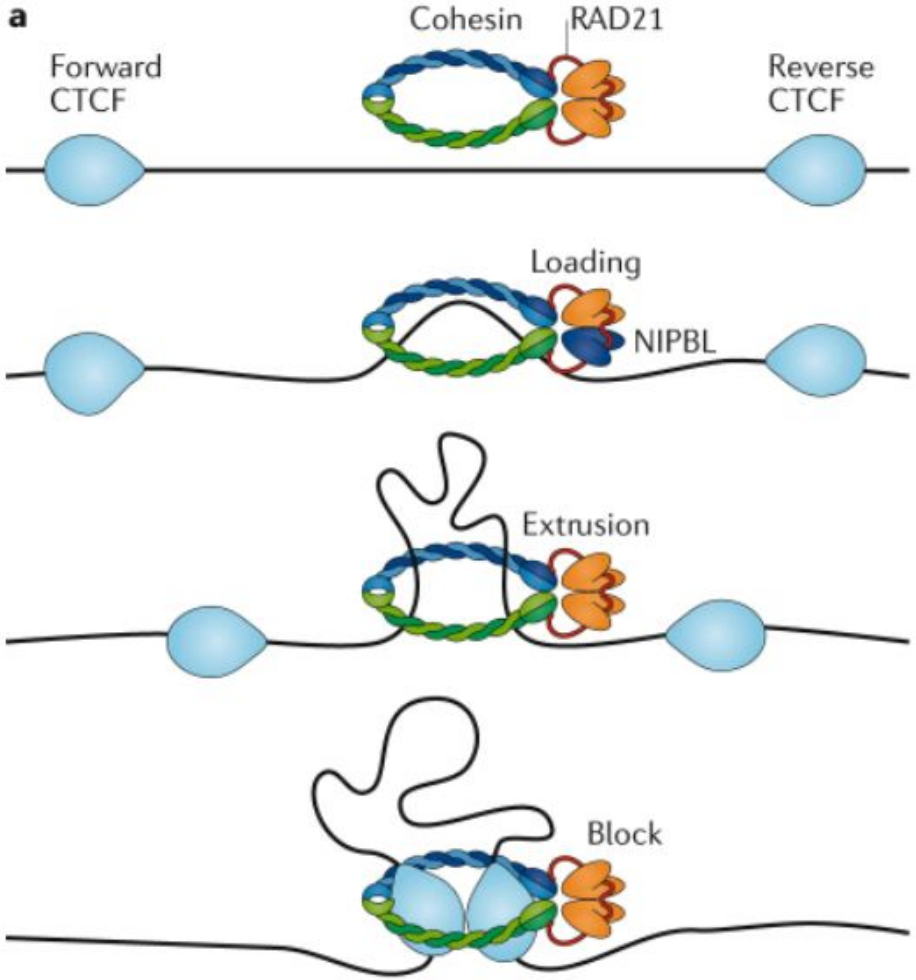
CTCF forms **convergent** dimers at loops



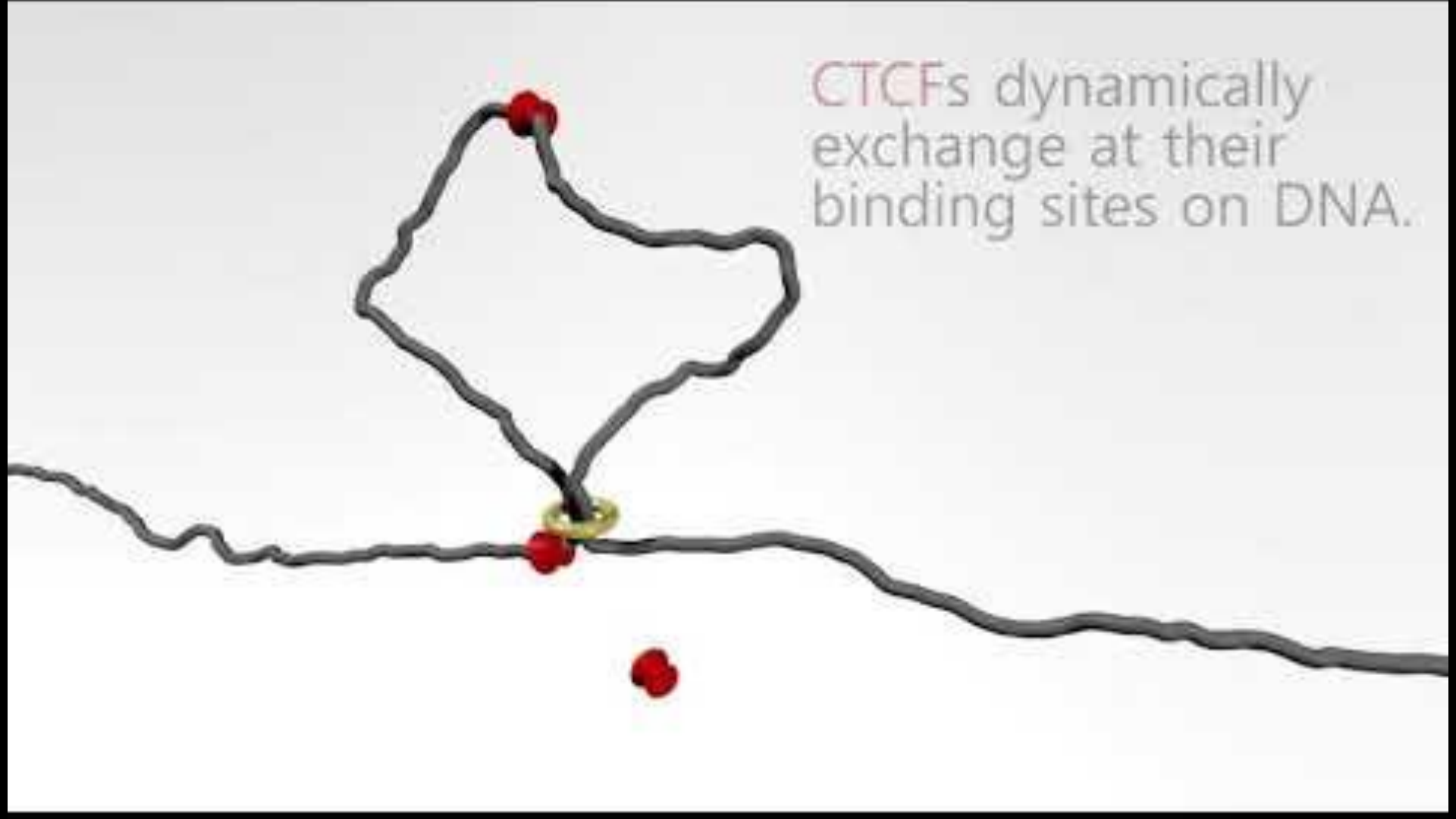
(Adapted from de Wit et al., Mol Cell 2015)

The loop extrusion model

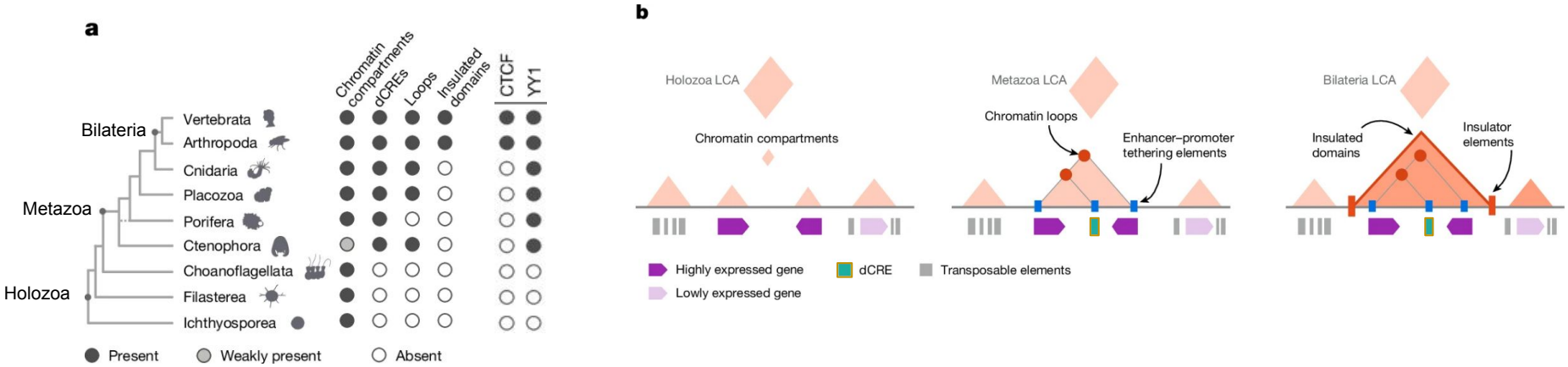
CTCF dimers form loops
by blocking Cohesin

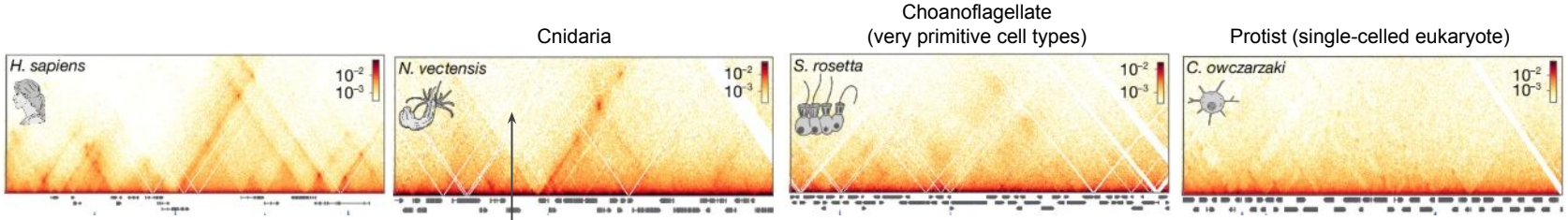


(Adapted from
Rowley and Corces, Nat Rev Gen 2018)


CTCFs dynamically
exchange at their
binding sites on DNA.

The evolution of chromatin structure is tied to multicellular life





TADs and loops, but poor isolation

TAD-like structure, but no clear loops

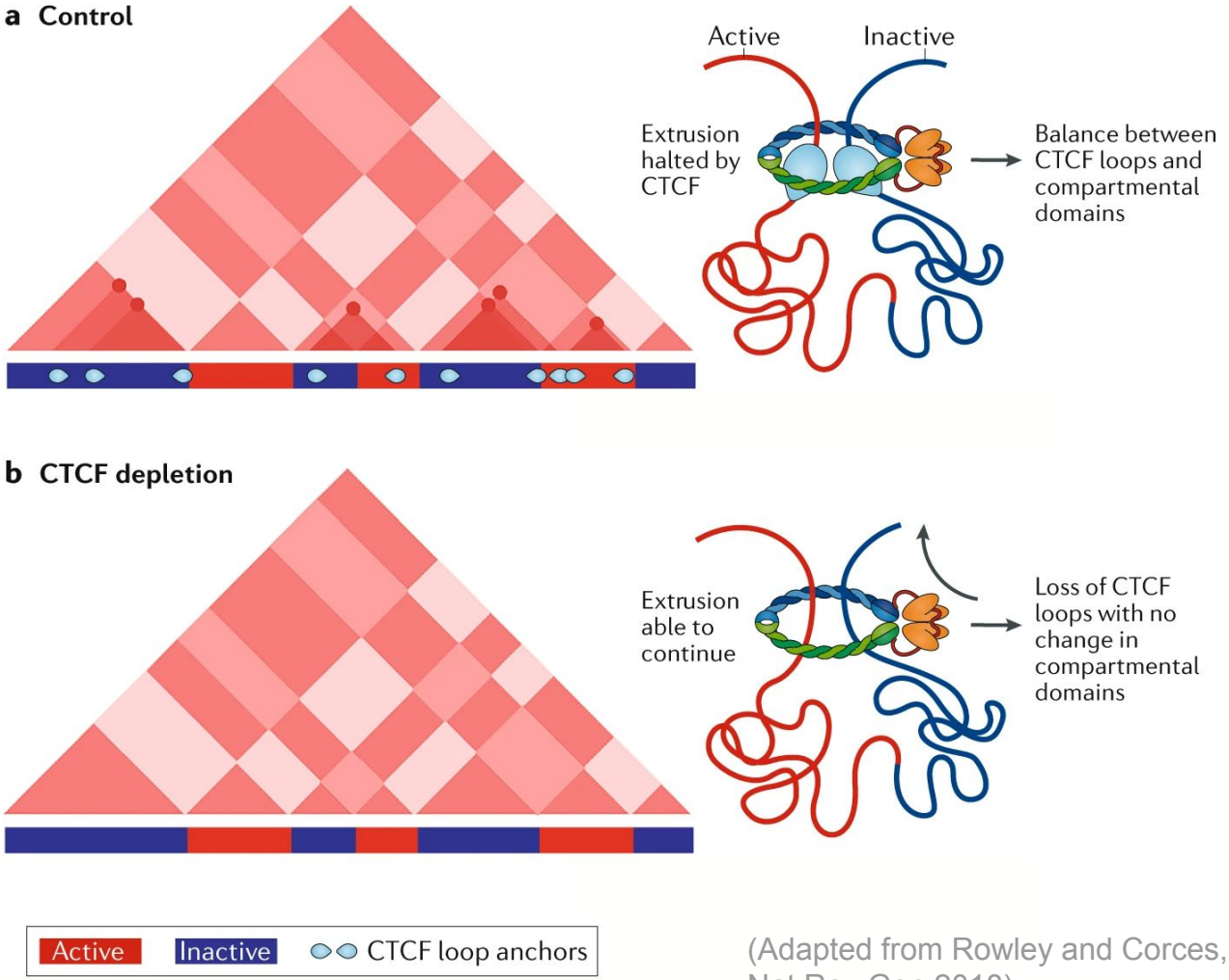
Lack of loops and clear TAD structure

(Adapted from Kim et al., *Nature* 2025)

CTCF depletion disrupts especially TAD-internal structures

presumed to include promoter-enhancer loops

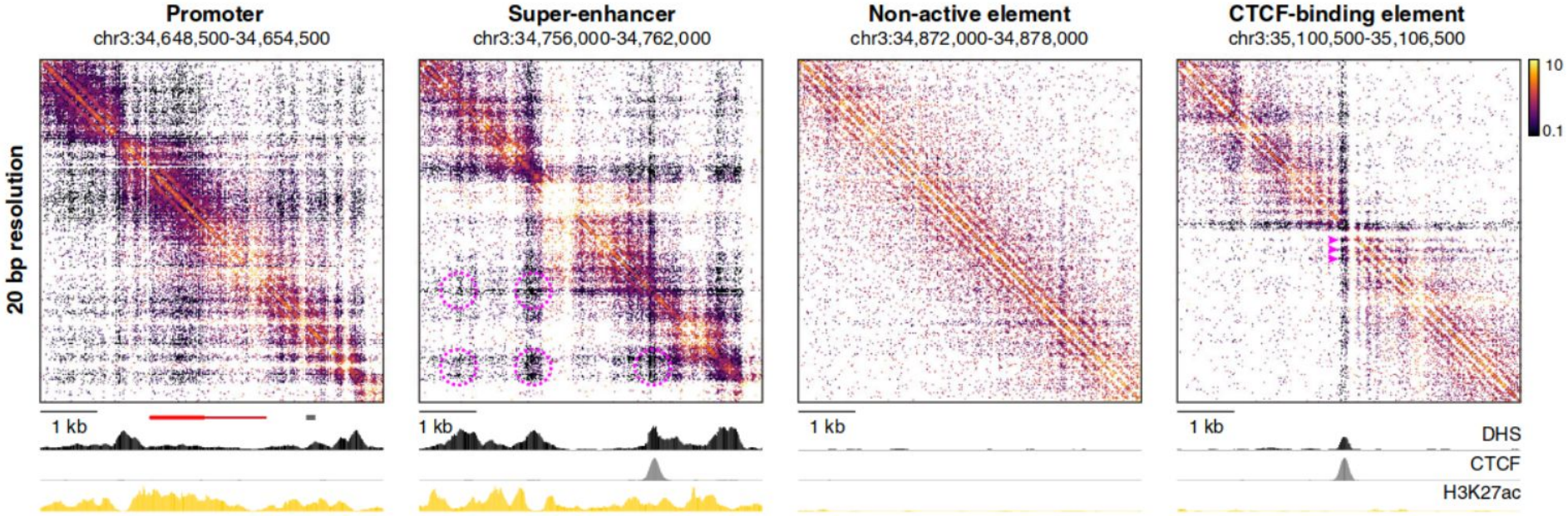
however…



(Adapted from Rowley and Corces, Nat Rev Gen 2018)

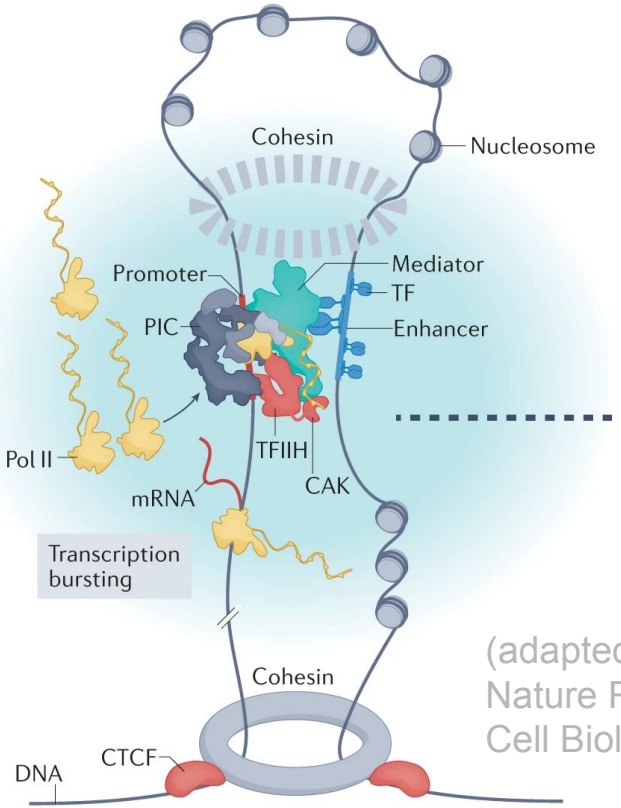
When going at very high resolutions, similar contact patterns appear, that are however not associated with CTCF

The structures however appear correlated with DHS, suggesting that at this fine-grained level the it is determined by (most likely combinations of) other factors



(20bp-resolution contact maps, adapted from Aljahani et al., *Nat Comm* 2022)

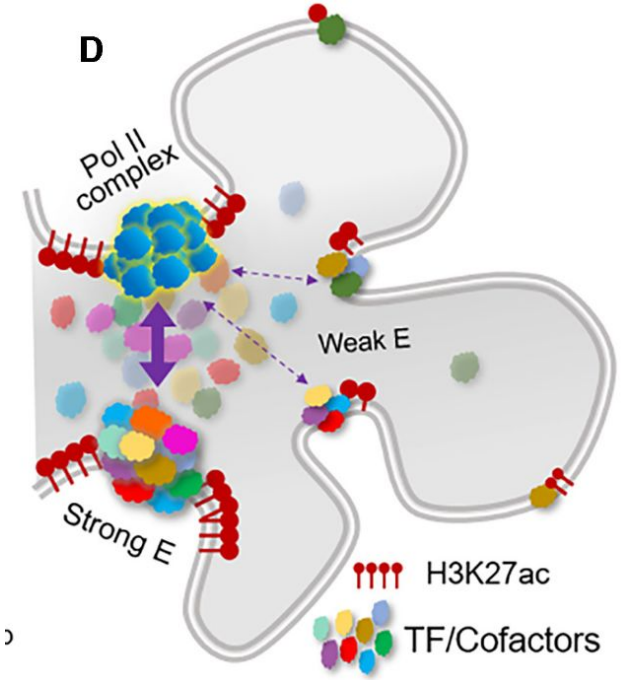
According to the prevailing model, enhancers physically interact with target TSS through combinations of proteins (in particular *Mediator*)



(adapted from Richter et al., Nature Reviews Molecular Cell Biology 2022)

In both views, chromatin loops create local neighborhoods in which these interactions can take place

An alternative view rests on diffusion between regulatory elements

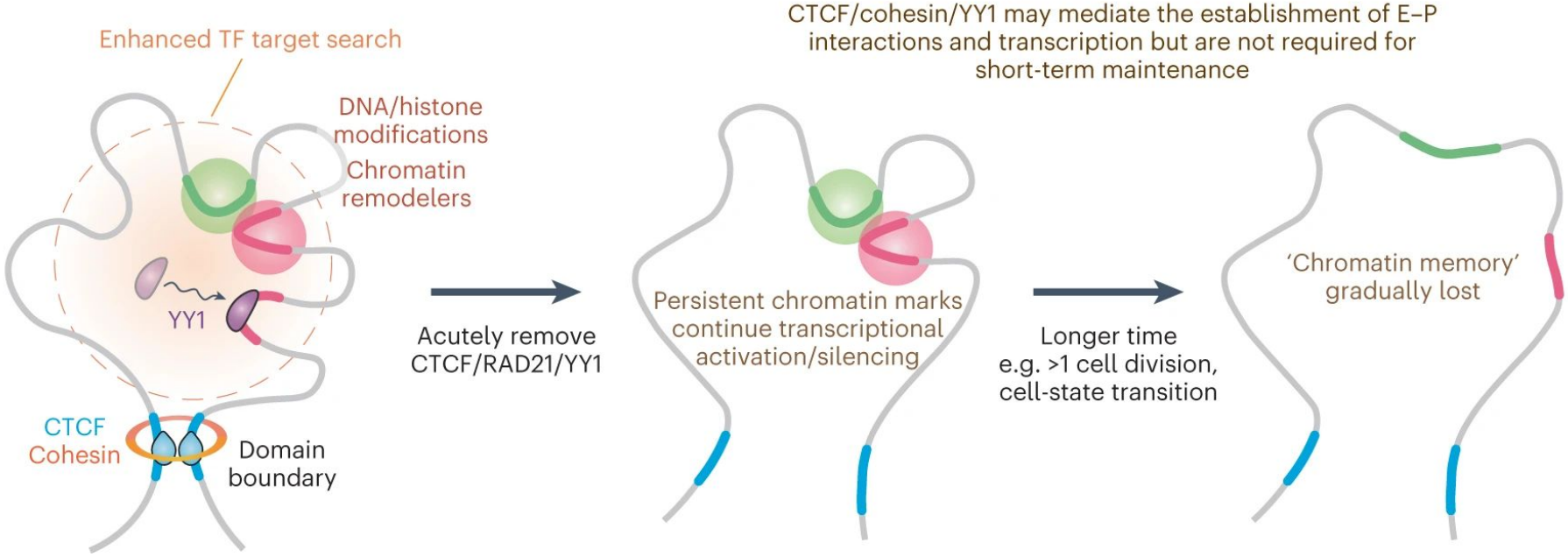


(adapted from Zhu et al., Nucleic Acids Research 2021; see also Galouzis and Furlong, Current Opinion in Cell Biology 2022)

Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1

Tsung-Han S. Hsieh, Claudia Cattoglio, Elena Slobodyanyuk, Anders S. Hansen, Xavier Darzacq ✉ & Robert Tjian ✉

Nature Genetics **54**, 1919–1932 (2022) | Cite this article



Proximity is not sufficient for enhancer-mediated activation

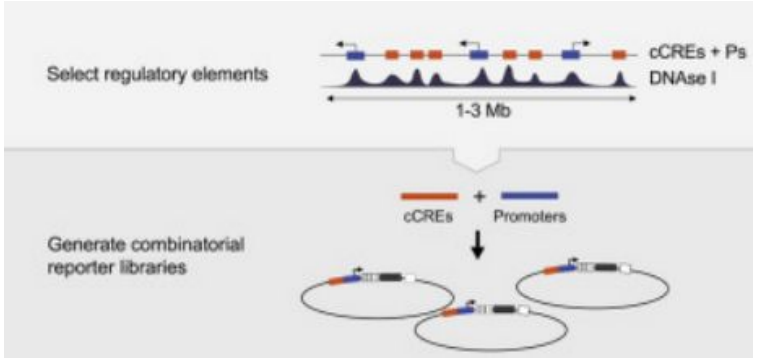
Molecular Cell

Volume 82, Issue 13, 7 July 2022, Pages 2519-2531.e6

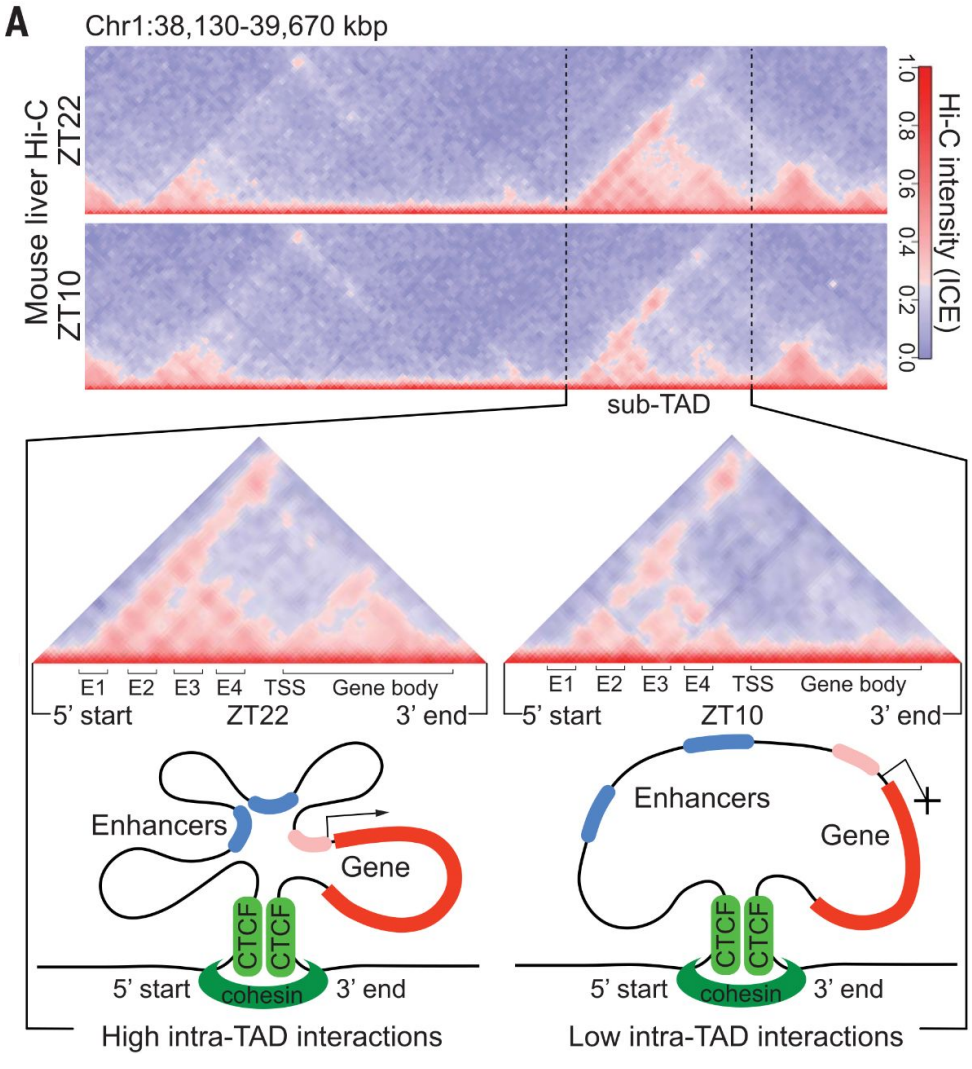
Article

Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome

Miguel Martinez-Ara [1], Federico Comoglio [1 2], Joris van Arensbergen [1 3], Bas van Steensel [1 4]



- Even when enhancers are made to come in proximity to a promoter, **not all enhancers can regulate all promoters**
- E-P compatibility (i.e. the specificity of the interactions) appears to be dictated by TF interactions between the two domains



Changes in chromatin contact occur within stable TADs/subTADs

Here an example driven by the circadian rhythm (mouse liver, ZT22=5am ZT10=5pm)

(Kim et al., Science 2018)

A few take-home messages

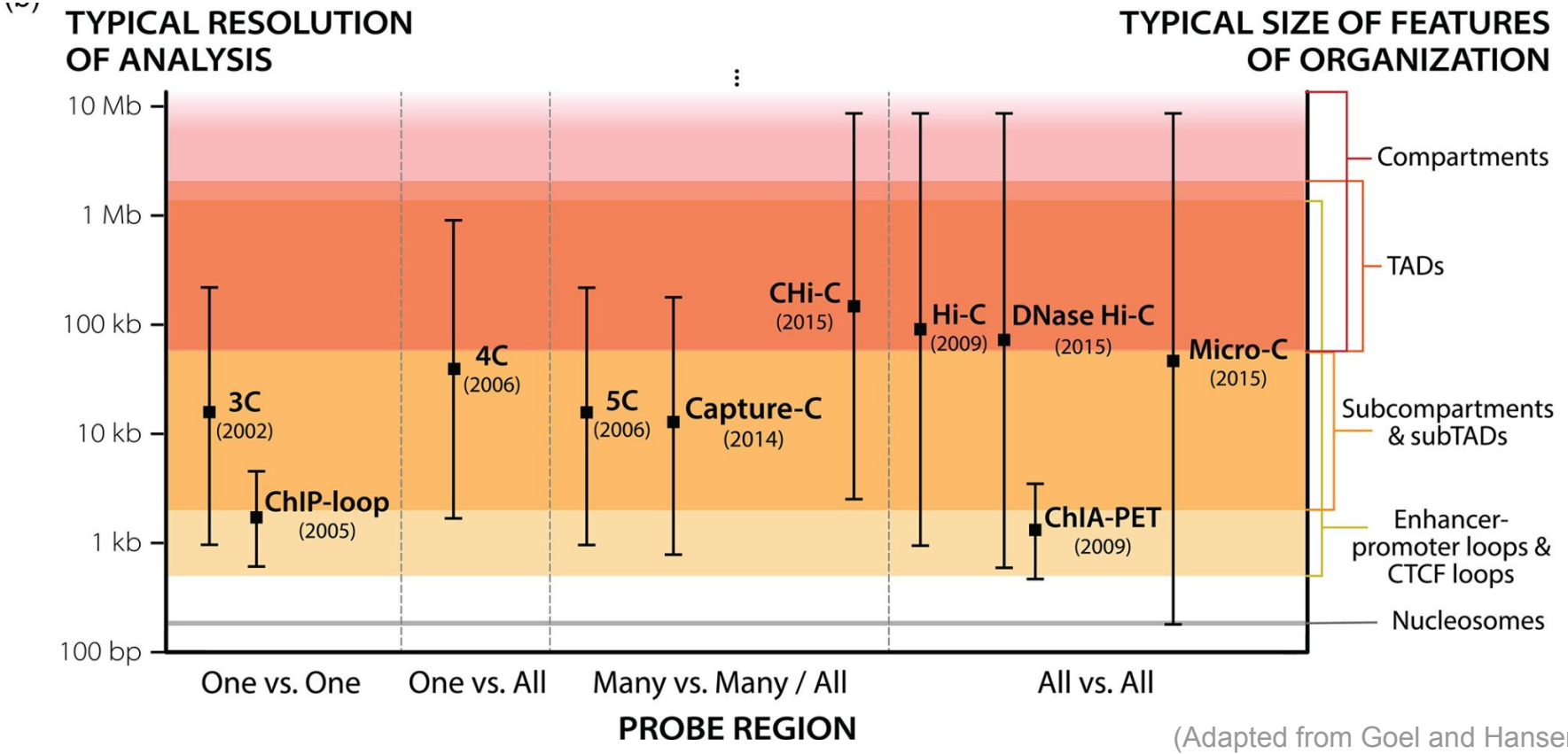
- the nucleus is organized into chromosome territories,
  and chromosomes are organized into active/inactive compartments

- chromosomes are further organized into TADs that are
  - largely stable across cellular states and
  - are delimited by CTCF dimers (loops)

- chromatin loops bring regions of chromatin in proximity,
  - enabling the formation of enhancer-promoter interactions,
  - but are not required for the short-term *maintenance* of these interactions

- Functional enhancer-promoter interactions appear determined by TF interactions, made possible by loops

Sequencing-based conformation capture



(Adapted from Goel and Hansen 2020)

(See also Hun Lee, Wu and Rhie, Epigenetics & Chromatin 2022 for a review)

Data analysis - some references

Method & analysis refs

- Hi-C:
  - Juicer
- HiChIP:
  - Dovetail doc
  - FitHiChIP

Main output & formats

- Contact maps
- TADs/ compartments:
  - bedpe (?)
- Interactions:
  - bedpe / pairix

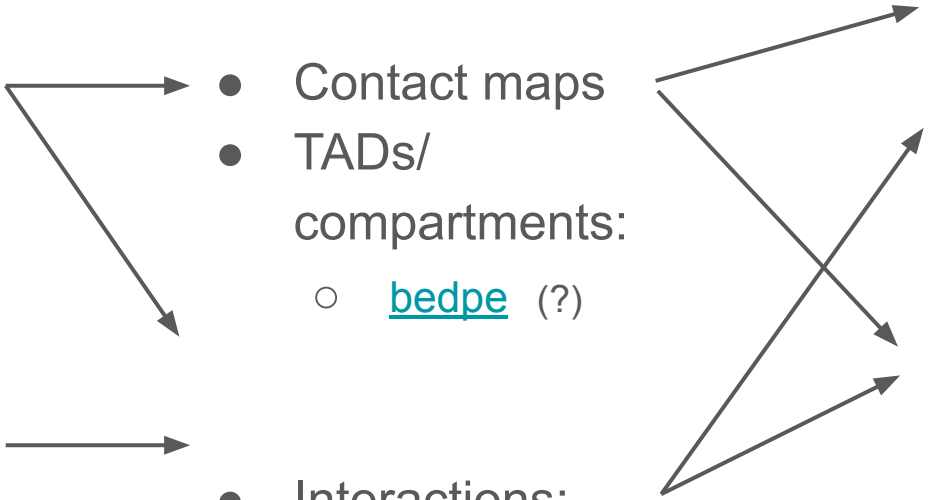
Visualization in R

HiCBricks

plotgardener

Data structure in R

interactionSet



Practical

Enhancer-promoter interactions: what to do with distal regulatory elements?

- Enhancers within <1kb of a TSS nearly always regulate that TSS (Sanyal et al., Nature 2012; Morris et al., Science 2023),
- Beyond 2.5kb, more than half of the interactions are not with the closest gene (Sanyal et al., Nature 2012).

Enhancer-promoter interactions: what to do with distal regulatory elements?

- Several types of information can be used to link such distal regulatory regulatory elements to target genes:
  - **distance**
  - chromatin **conformation** data (e.g. Fulco et al., Nat Gen 2019; Salviato et al., NAR 2021) → requires very high-depth data for the cell-type of interest
  - expression Quantitative-Trait-Loci (**eQTLs**) (e.g. Fishilevich et al., Database 2017) → depends on well-represented genetic variation profiled in the right cell types
  - statistical **association** (e.g. correlation) between signals (e.g. accessibility or histone modification) at the distal enhancer and gene expression (e.g. Hait et al., NAR 2022)
    → requires a lot of data with sufficient variation → single-cell multi-omic data!

Enhancer-promoter interactions: what to do with distal regulatory elements?

- For the assignment, you'll work with something easy to get and to use, namely :

  Generic (i.e. non-celltype-specific) enhancer-target predictions based on both association and conformation,

  produced by Salviato et al., NAR 2021

Assignment

- Choose a transcription factor (e.g. p300), and obtain peaks from ENCODE (ChIP-seq in a human context!)
- Subset to those peaks that have a predicted distal target(s) using Salviato et al. (2021)
    - You can download a GRanges of those interactions at https://ethz-ins.org/content/hg38.SalviatoDistalEnhancerTargets.GR.rds
- Label peaks with the genes it is in 3D contact with (henceforth called ‘predicted target(s)’)
- Find the nearest TSS for each peak
- Split the peaks (or genes) into those for which the peak is upstream or downstream of the predicted target
- For each of the two sets, report:
    - 1) in what proportion of the cases the predicted target is the closest gene?
    - 2) plot a distribution of the distances between the peak and the (TSS of the) predicted target gene
- Hints:
    - beware not to count, when calculating proportions, peaks that don’t have interactions with any TSS!