3000788 Intro to Comp Molec Biol

Lecture 20: Chromatin organization

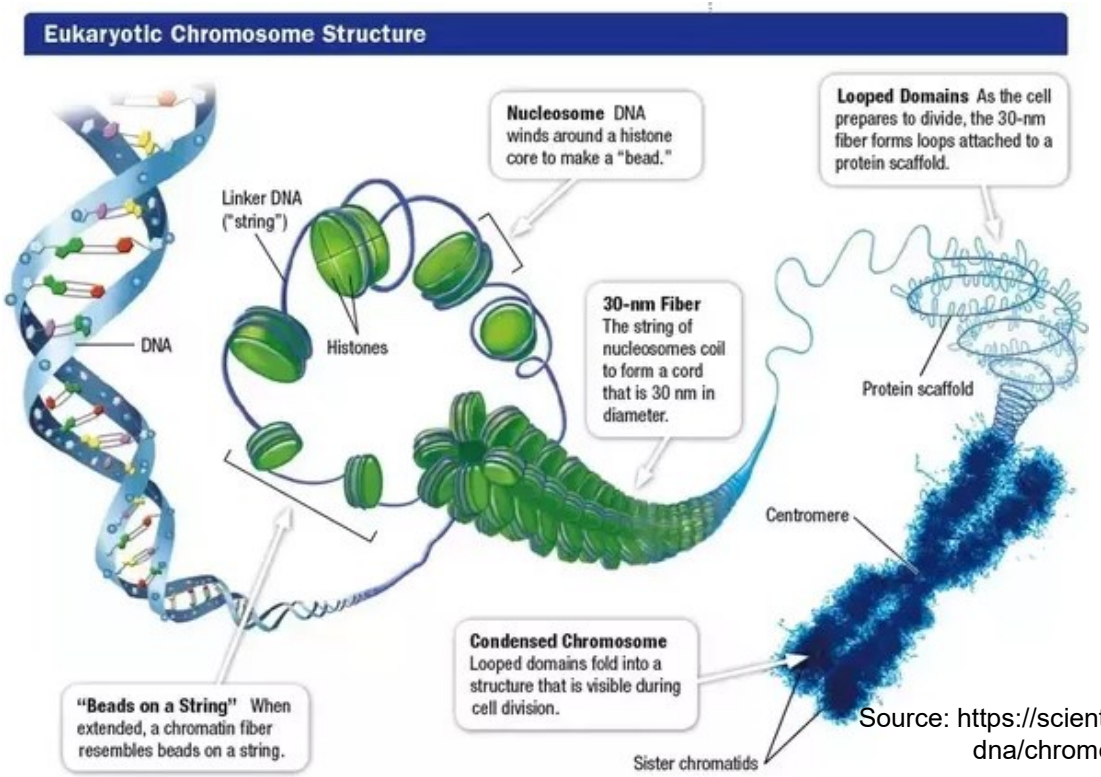
October 26, 2023

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

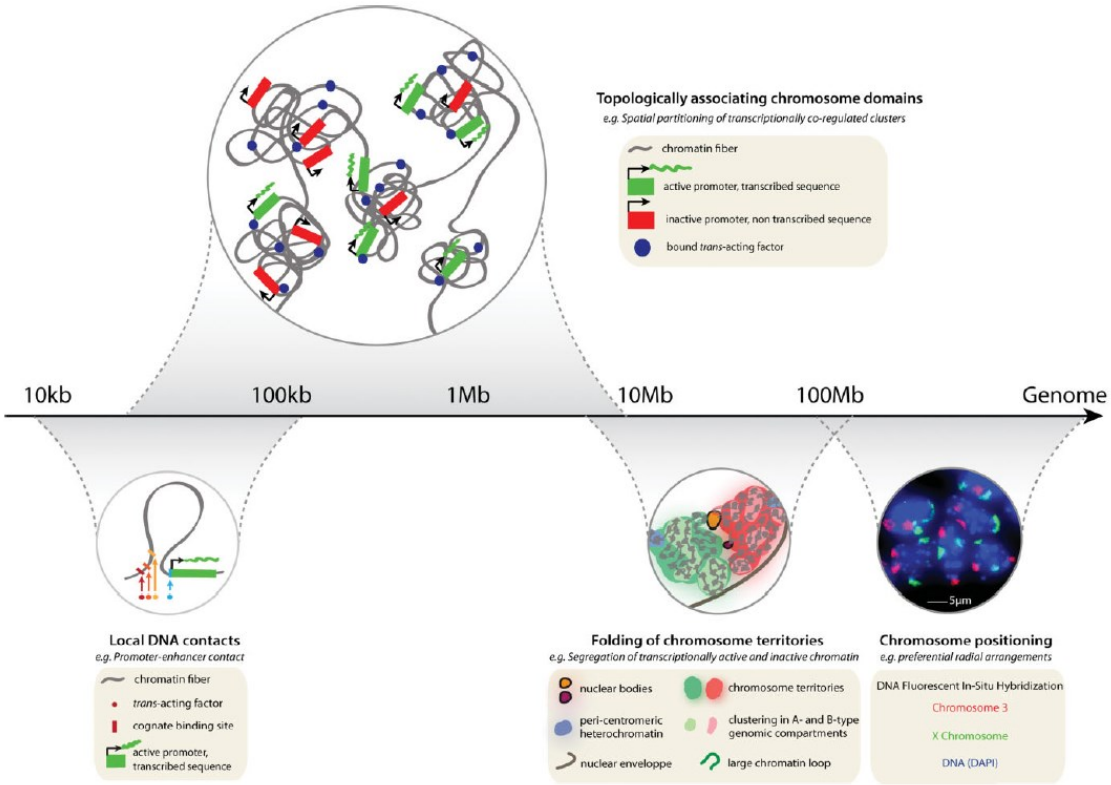
Chromatin organization

Chromatin folding for packaging



Source: https://scientificallytalking.com/what-is-dna/chromosome-structure/

Hierarchical organization of chromatin



- Organized chromatin is easy to split during cell division
- Topologically associating domain (TAD) = local chromatin folds
- Enhancer looping

Source: Nora *et al*. Bioessay (2013)

Mutational hotspots on chromatin

Article | Open Access | Published: 20 October 2016

Chromatin accessibility contributes to simultaneous mutations of cancer genes

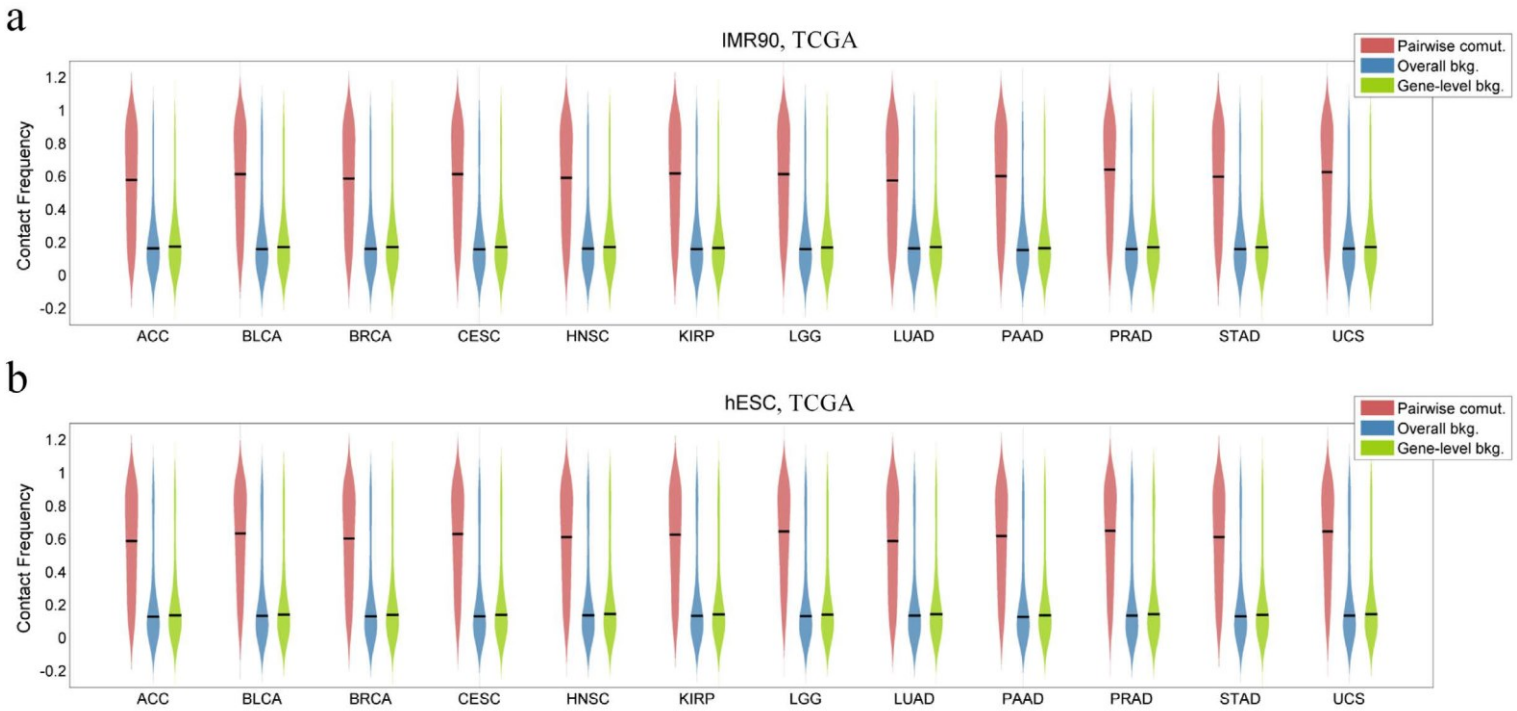
Yi Shi, Xian-Bin Su, Kun-Yan He, Bing-Hao Wu, Bo-Yu Zhang & Ze-Guang Han

A genome-wide scan for correlated mutations detects macromolecular and chromatin interactions in *Arabidopsis thaliana*

Laura Perlaza-Jiménez, Dirk Walther

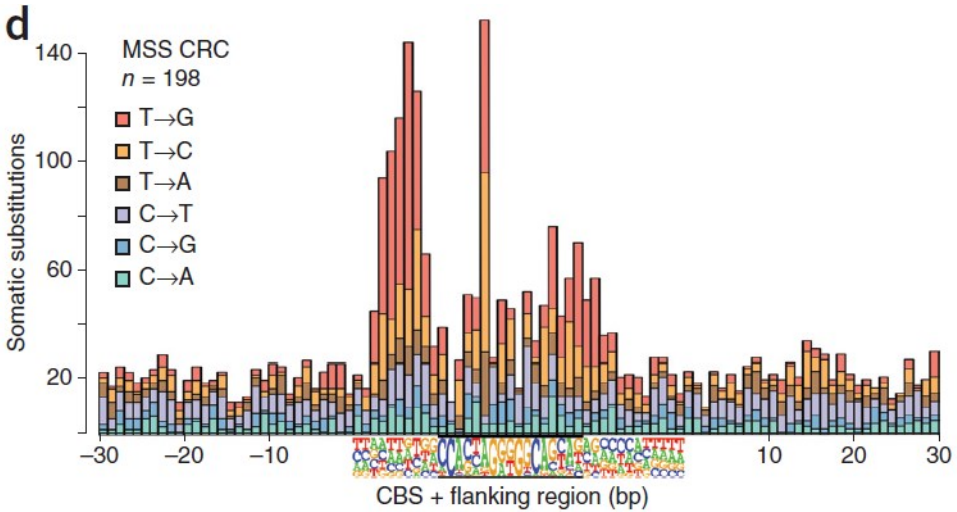
Nucleic Acids Research, Volume 46, Issue 16, 19 September 2018, Pages 8114–8132,

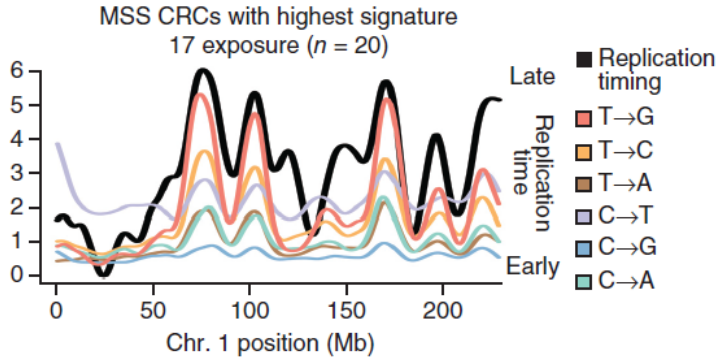
Co-mutations on 3D chromatin



Shi *et al*. Scientific Reports (2016)

Cancer subtype with destabilized chromatin

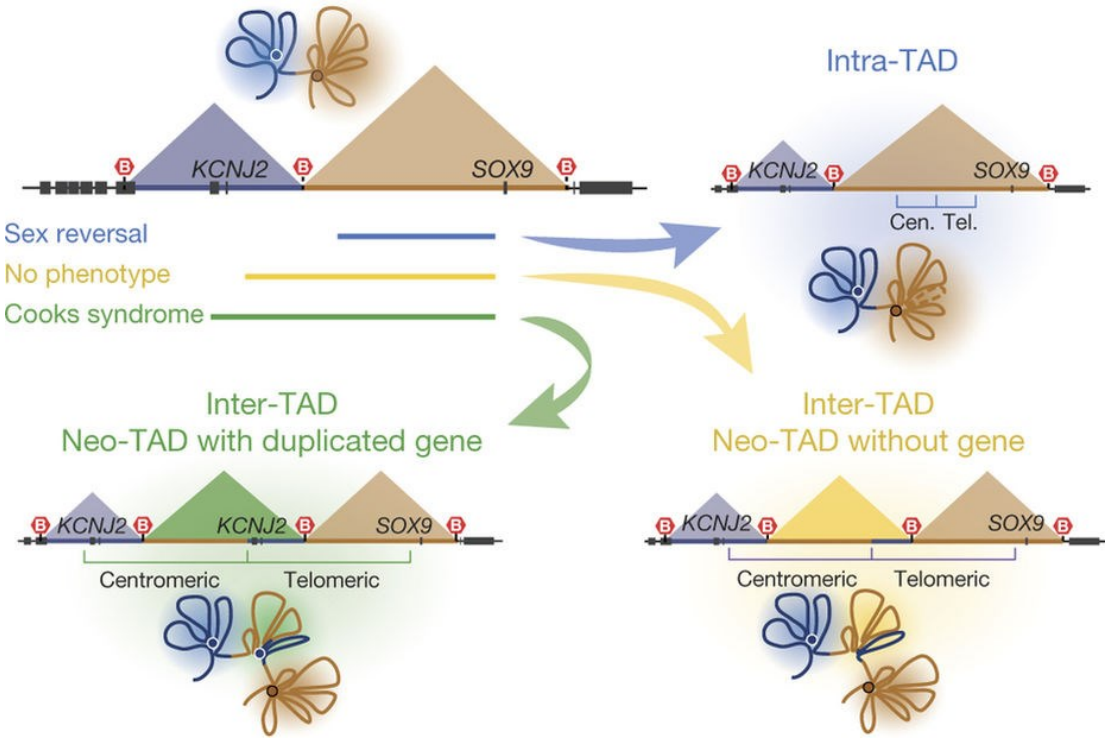




Katainen *et al*. Nature Genetics (2015)

- CBS = CTCF binding site on chromosome to regulate the folding
- Mutations on CBS disrupt chromatin folding
  - Widespread dis-regulations of gene expression

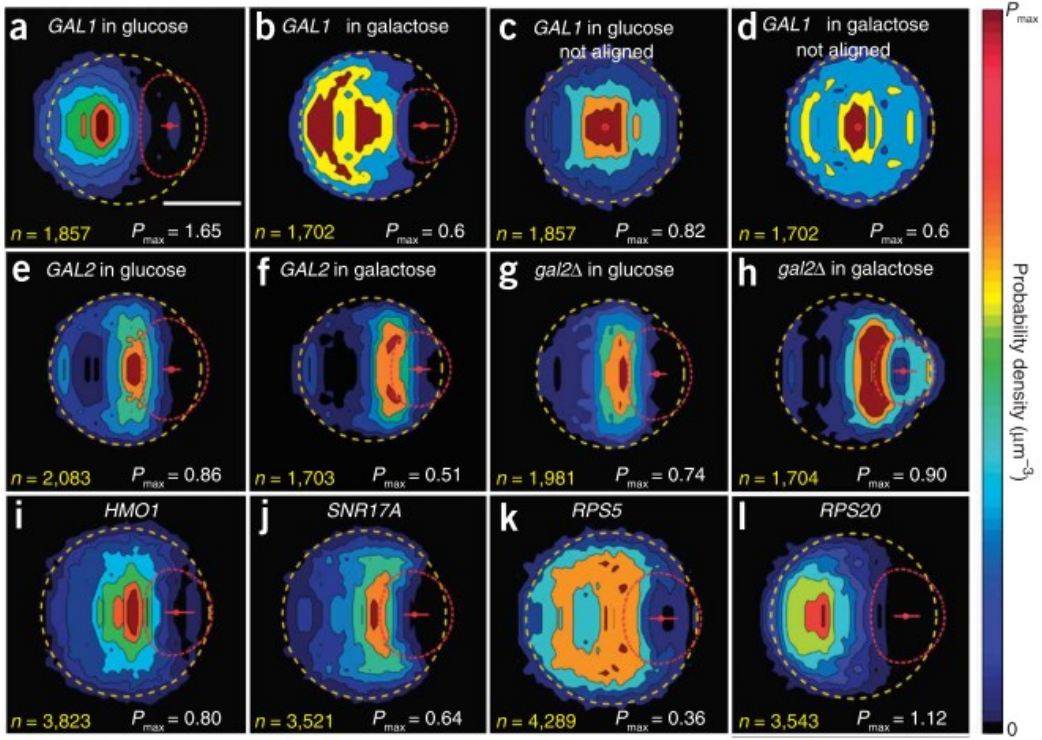
Disease due to chromatin organization

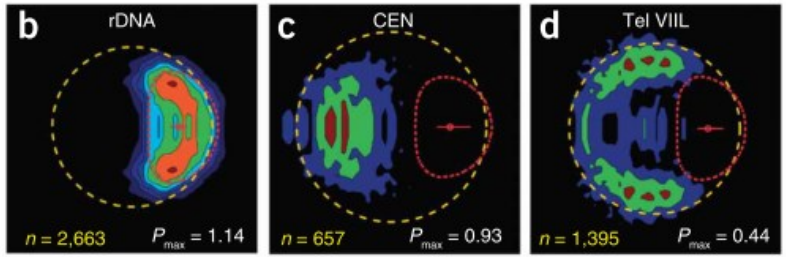


- Cooks syndrome
- Caused by a tandem duplication that forms a new local structure
- TAD = topologically associating domain
  - Local folding of chromatin

Franke *et al*. Nature 538, 265-269 (2016)

Chromain and gene territories

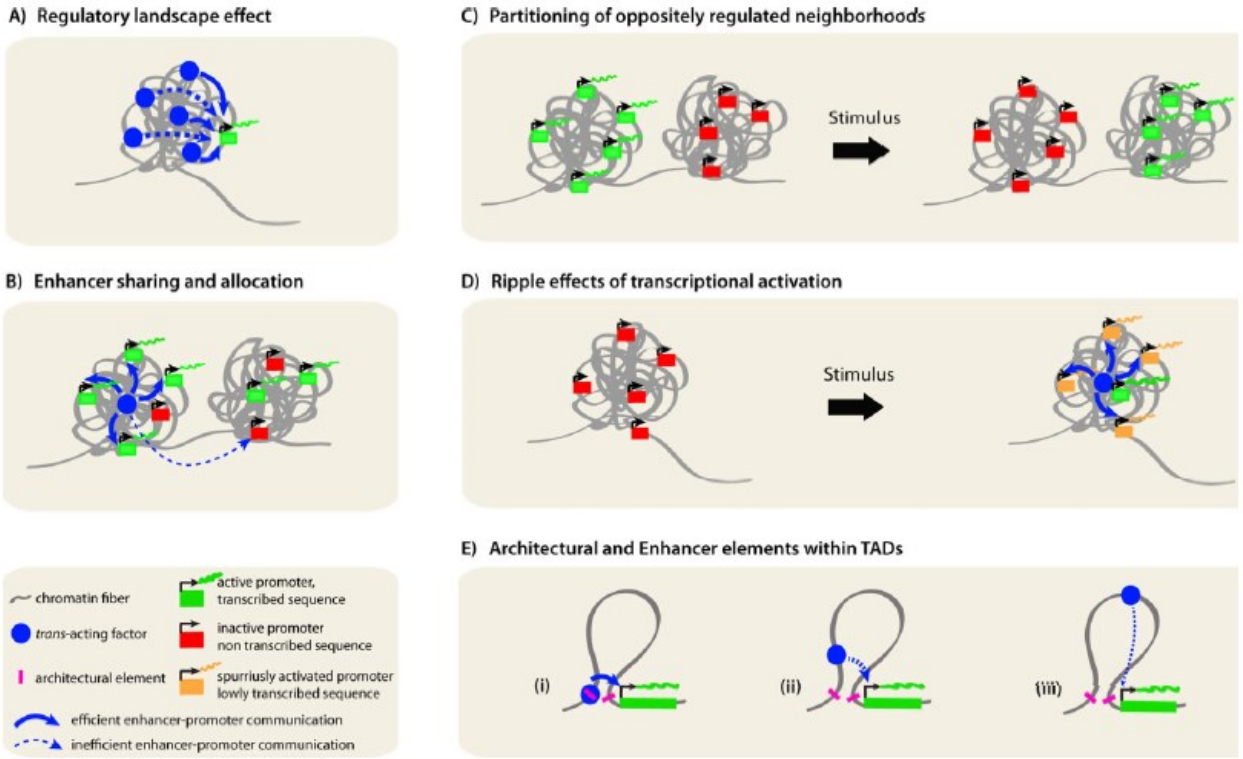




- Chromatin features, such as rDNA, centromere, and telomere, and some genes occupy specific regions in the nucleus

Source: Berger *et al*. Nature Methods (2008)

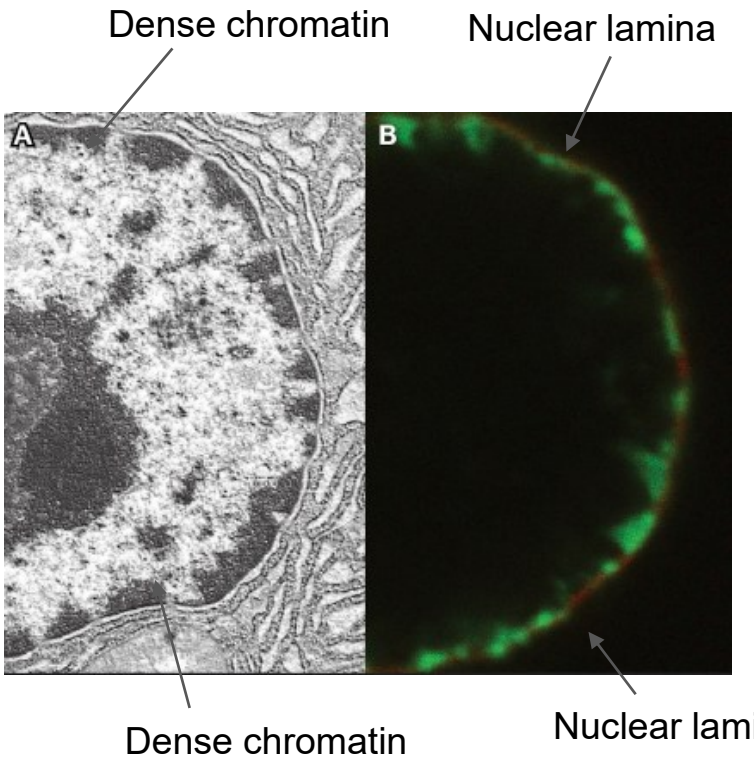
Biological implications of gene territories

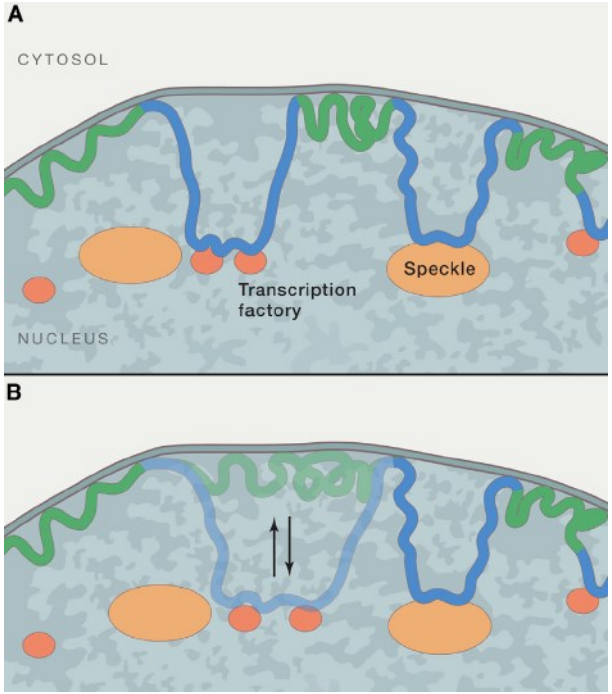


- Localization of co-regulated genes
- Sharing of transcription factors and enhancer

Nora *et al*. Bioessay (2013)

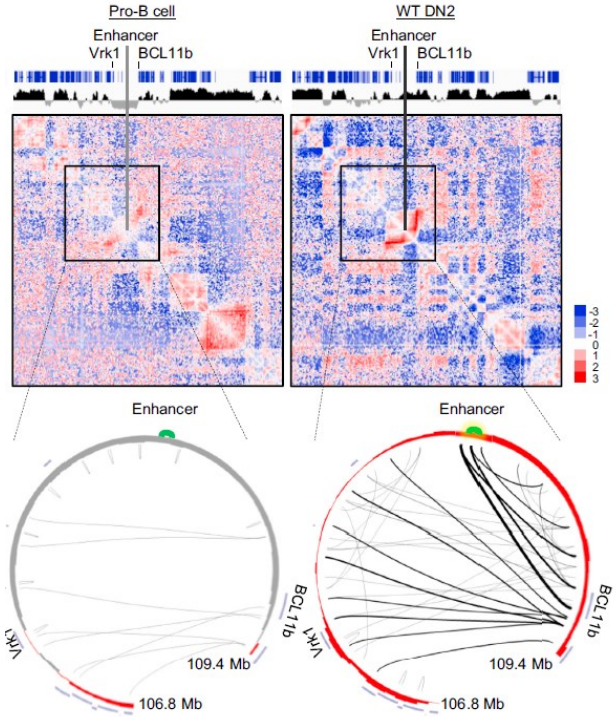
Heterochromatin associates with nuclear lamina



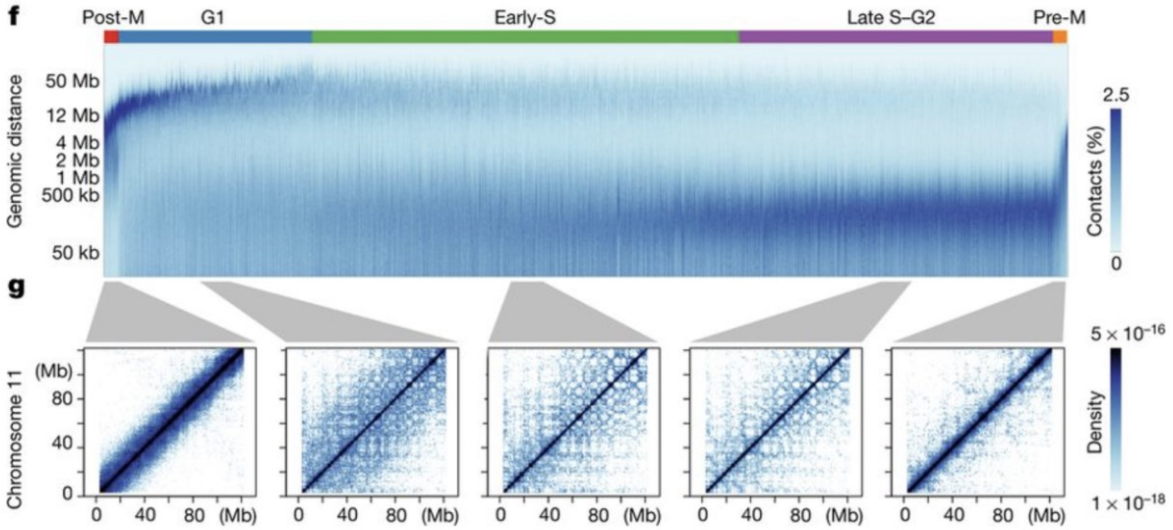


Steensel and Belmont. Cell (2017)

Restructuring of chromatin during development



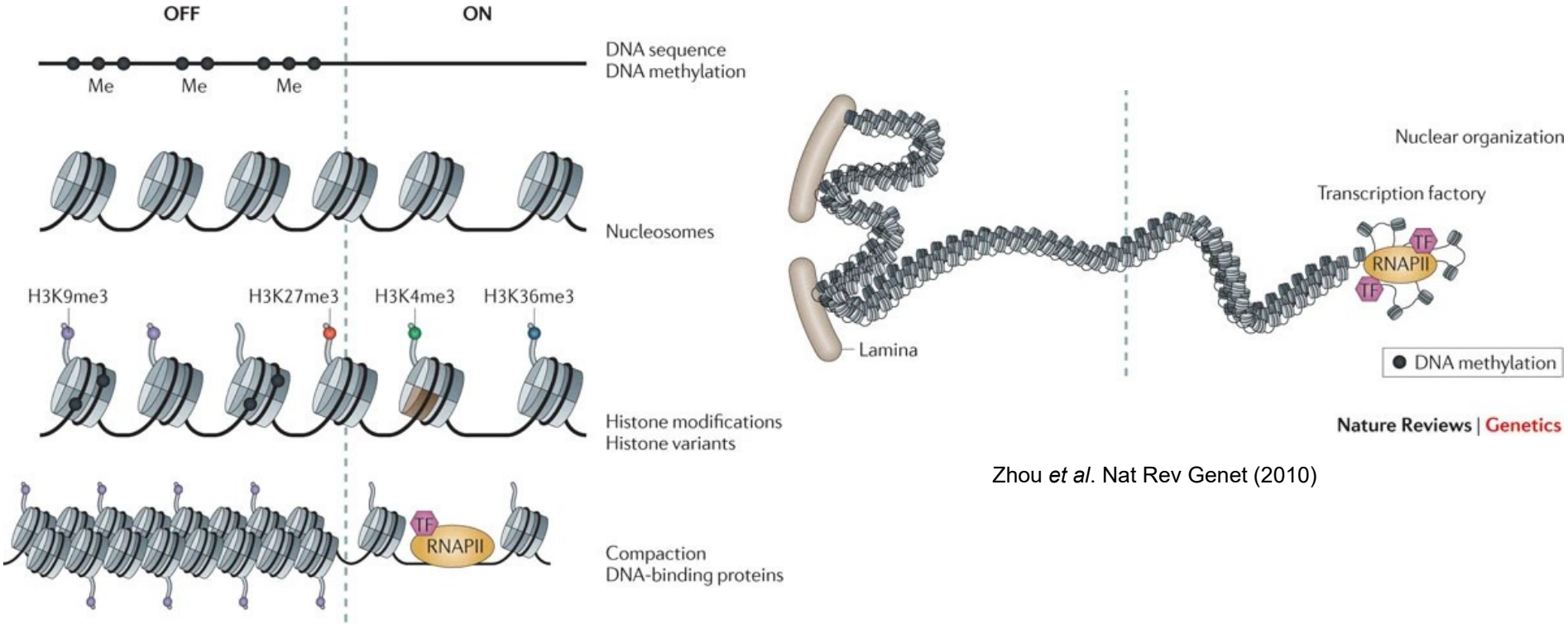
Isoda *et al*. Cell (2017)



Nagano *et al*. Nature (2017)

Chromatin assays

Chromatin and epigenomics assays



Zhou *et al*. Nat Rev Genet (2010)

Understanding gene regulation through chromatin

- Transcriptomics tell us whether a gene becomes activated or repressed
  - But not necessarily the mechanism

- Many gene regulatory mechanisms involve chromatin
  - DNA methylation **recruits repressive proteins** and marks heterochromatin
  - Transcription factors **occupation of enhancers and promoters**
  - Transcription activation requires **chromatin accessibility**
  - Looping of enhancer to gene loci

Bisulfite sequencing analysis

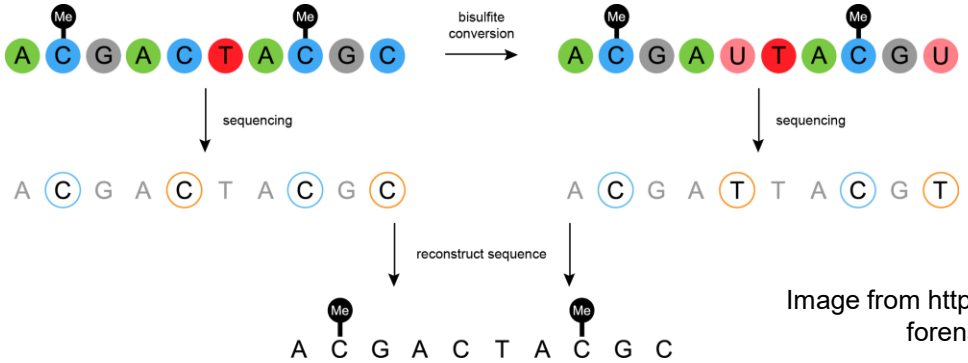
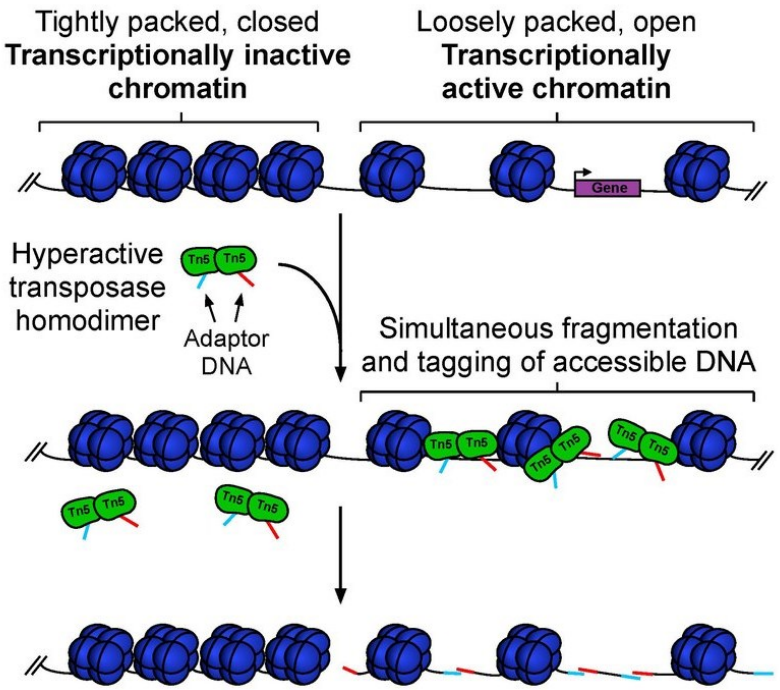
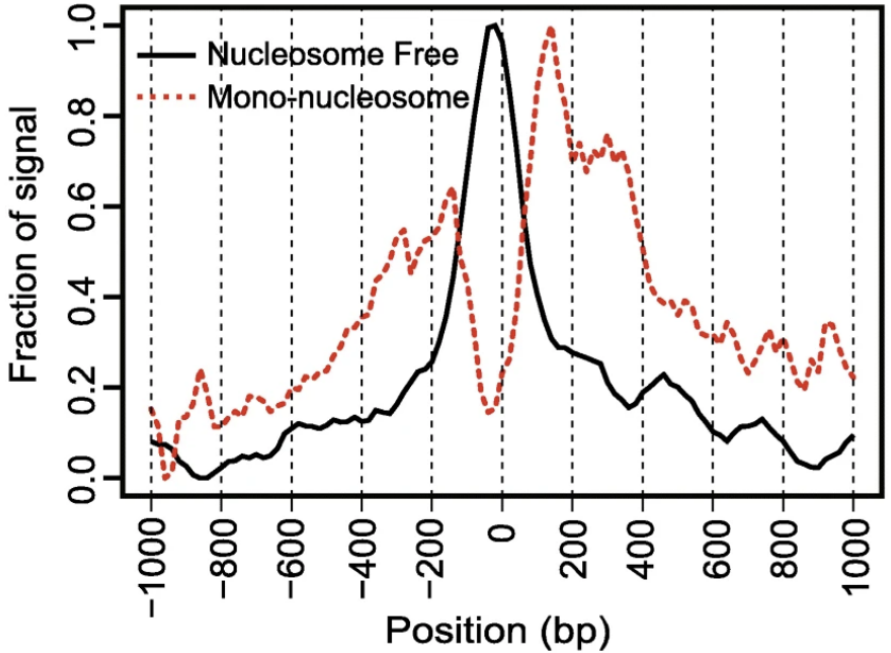


Image from http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

- Strategy 1: Convert C in reference genome to T
  - Use regular aligner + compare match to C and T versions
- Strategy 2: Adjust mismatch score for C-T
  - P(random C-T mismatch) scales with P(T in read)
  - P(C-T bisulfite) scales with P(non-methylated)

ATAC-seq





Yan *et al*. Genome Biology (2020)

ChIP-seq and peak calling

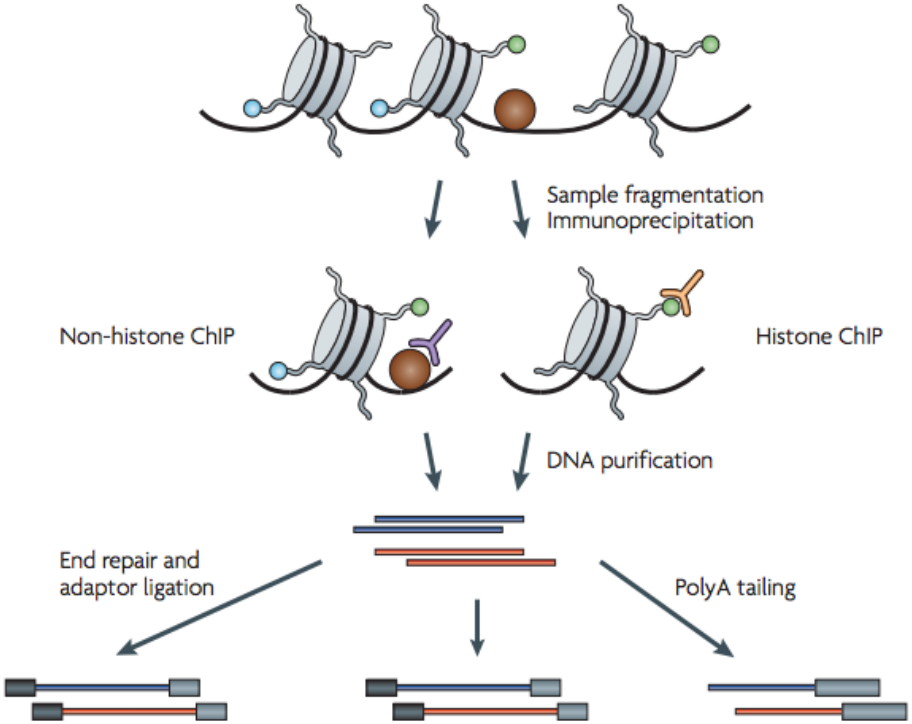
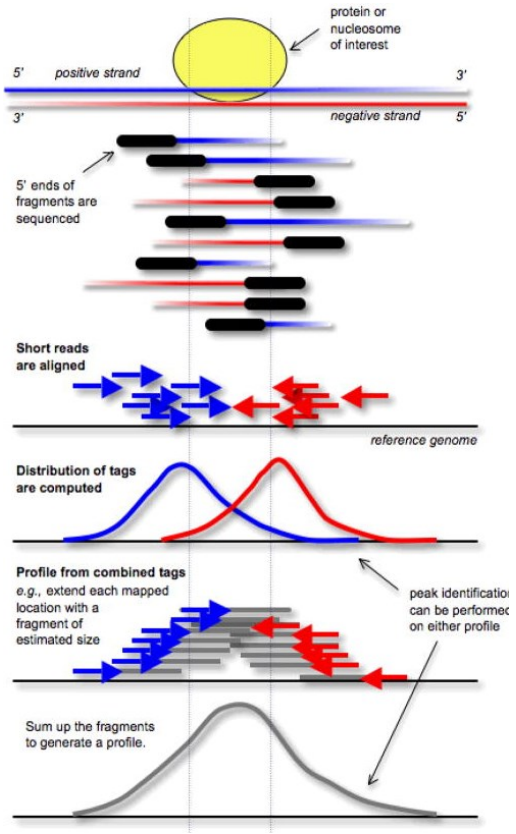


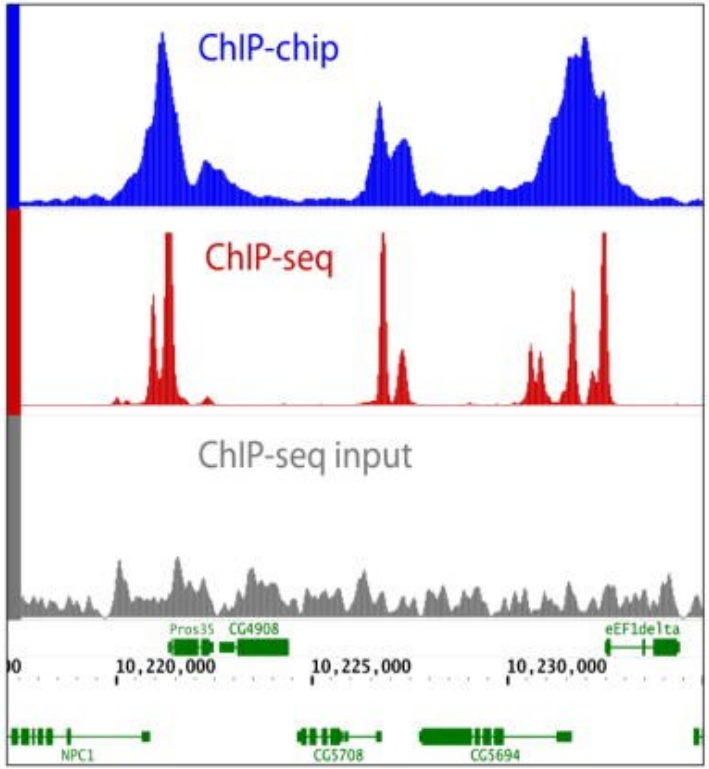
Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

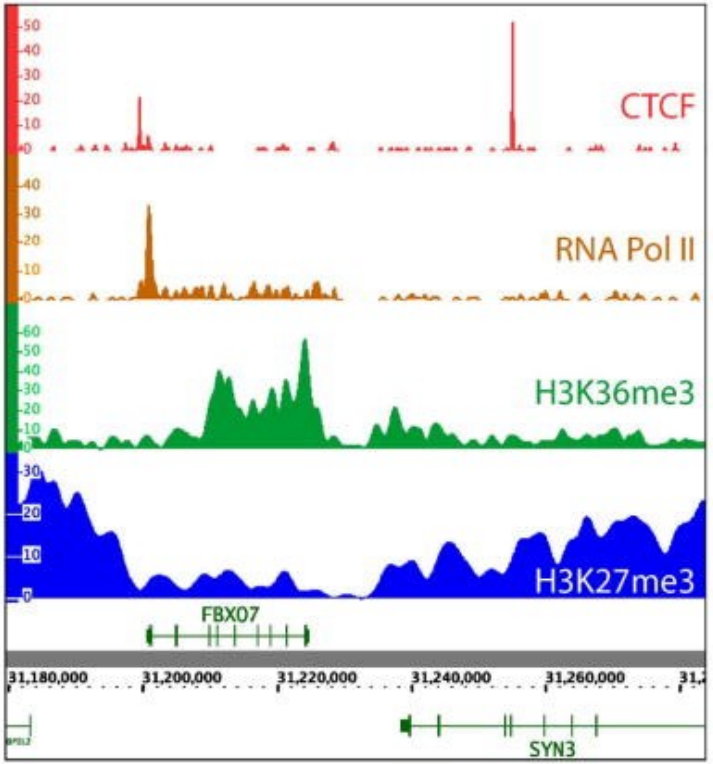


Park *et al.* Nat Rev Genet 10:669-680 (2009)

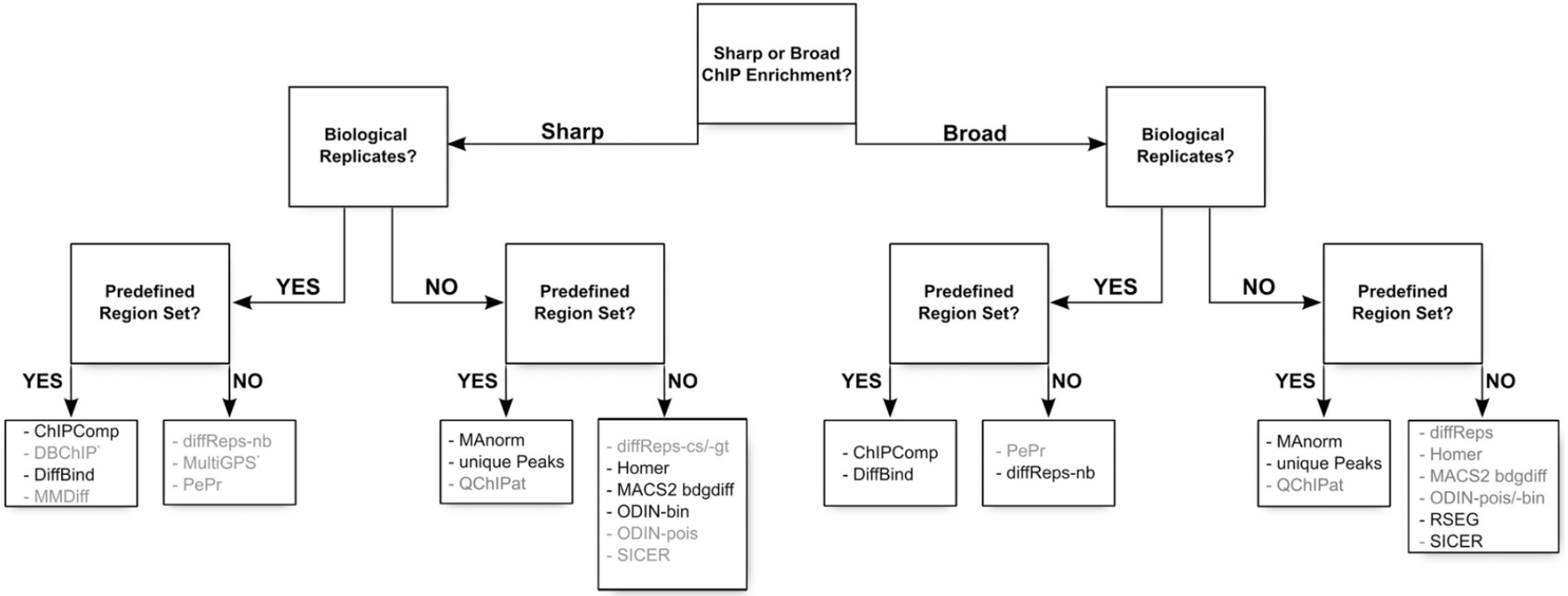
Broad and sharp peaks

Park *et al.* Nat Rev Genet 10:669-680 (2009)





Different tools for broad vs sharp peaks



Steinhauser *et al.* Brief in Bioinfo 17:953-966 (2016)

Poisson model for peak calling

- Number of reads at each genomic locus ~ Poisson(λ)
  - λ varies and represent the background number of reads
  - Estimate λ from control samples (no immunoprecipitation)

- P-value for observing n reads at a locus with λ reads in the control
  - $\sum_{k=n}^{N} \frac{\lambda^k e^{-\lambda}}{k!}$

- Can we use other dataset's controls to reduce cost?
  - Maybe, but will need several to capture the randomness
  - Will lose some power

ChIP-seq analysis

- Starts off like any sequencing dataset

- Peak calling and annotation
  - Linking peaks to genes
    - Nearest gene?
    - All genes within 2000 bp?
  - Functional enrichment

- Motif search
  - Find common DNA patterns across peaks
  - Possible TF binding sites

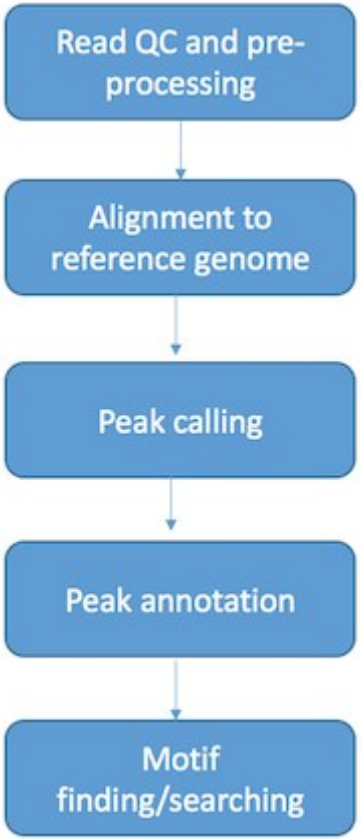


Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

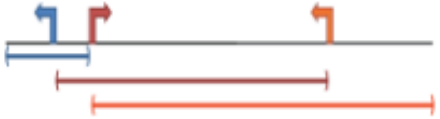
Peak annotation

(•) Basal plus extension

Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

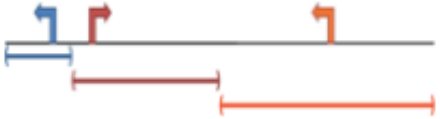
( ) Two nearest genes



within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction.

( ) Single nearest gene

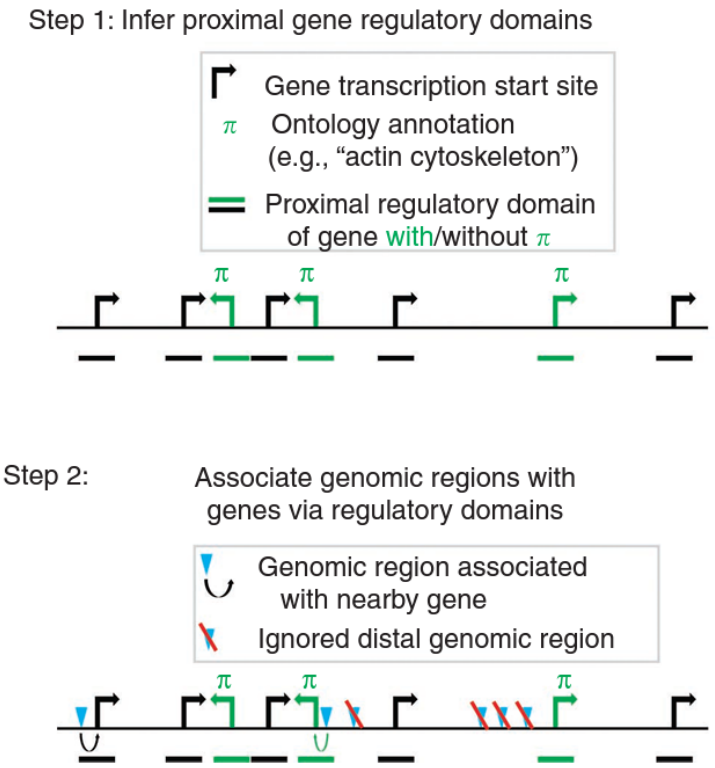


within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

Gene Transcription Start Site (TSS)

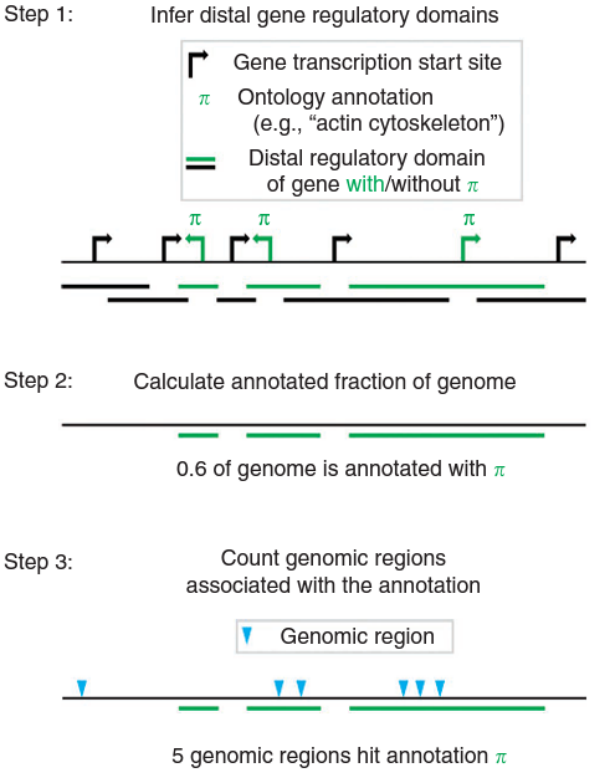
Overrepresentation analysis



- Are peaks located near genes with certain functionality?
- Use peak locations instead of differential expression
- Hypergeometric distribution

McLean *et al*. Nat Biotech, 28(5):495-501 (2010)

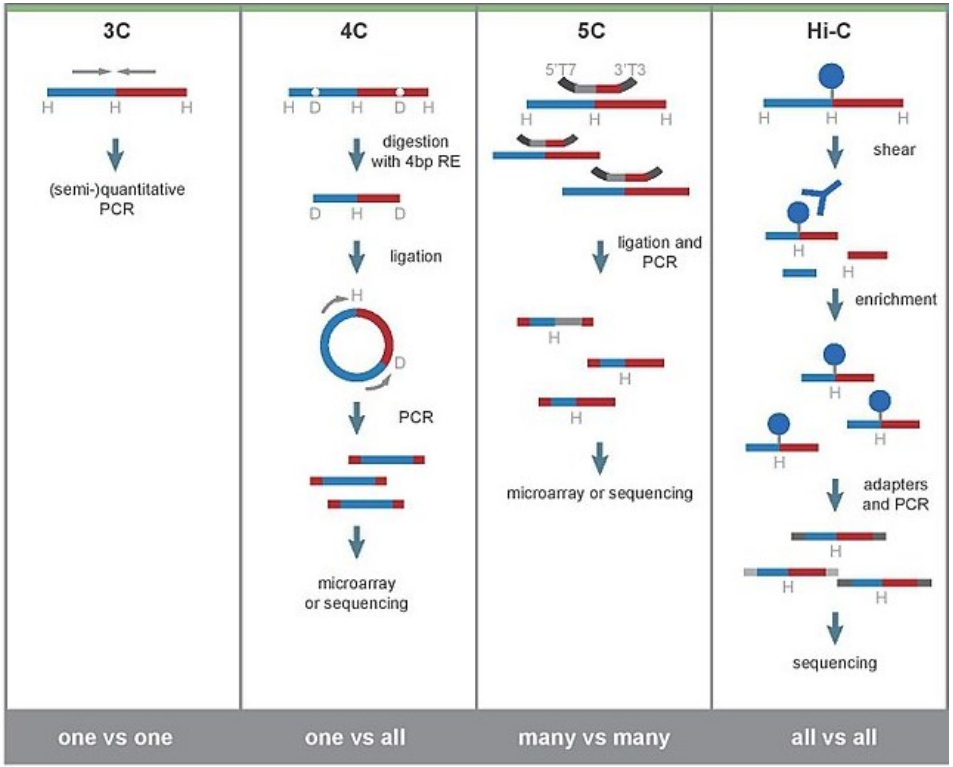
Overrepresentation analysis with binomial model

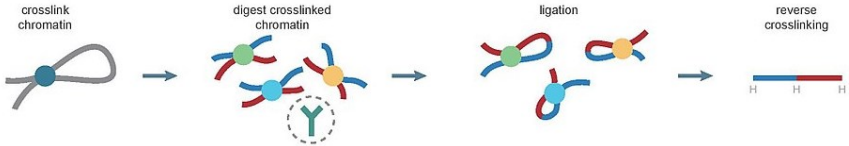


- Assign function to genomic regions surrounding the genes
  - Fraction of genome = success rate
- Identify number of peaks located in the annotated regions
  - Number of success
- Binomial distribution

Long-range chromatin interaction

Chromatin conformation capture





- Induce cross-link between proximal chromatins
- Generate hybrid DNA fragments
- Identify by high-throughput sequencing

Chromatin structure resolution

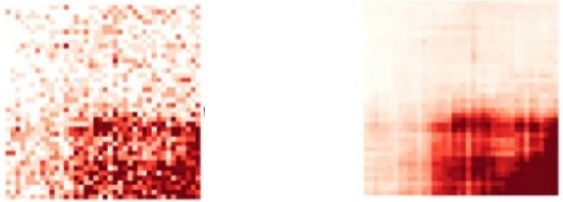
Article | Open Access | Published: 21 February 2018

Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus

Yan Zhang, Lin An, Jie Xu, Bo Zhang, W. Jim Zheng, Ming Hu, Jijun Tang & Feng Yue

Nature Communications **9**, Article number: 750 (2018) | Cite this article

Low-resolution High-resolution



Zhang *et al.*, Nat Comm (2018)

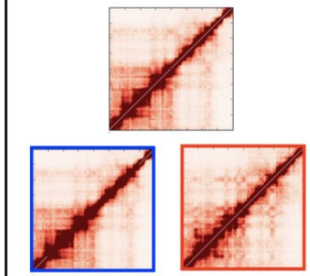
- Studying chromatin structure with Hi-C = filling an *N* x *N* matrix with read counts
- Requite N^2 read depths
- Consolidate genomes into bins
  - 10kb, 20kb, ..., 100kb
  - Larger bins = require less reads

QC for Hi-C data

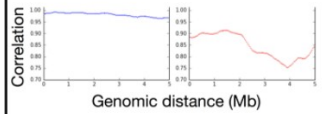
Source: github.com/kundajelab/3DChromatin_ReplicateQC

HiCRep

Transformation: 2D mean filter



Comparison: weighted sum of correlation coefficients stratified by distance



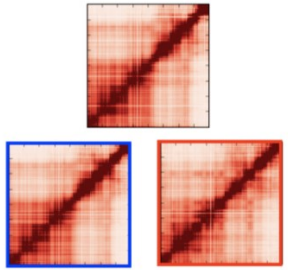
Reproducibility score:

$$\sum_d w_d \cdot \rho_d$$

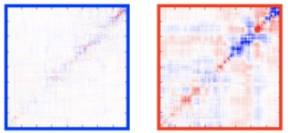
weight correlation

GenomeDISCO

Transformation: smoothing using graph diffusion



Comparison: difference in smoothed contact maps

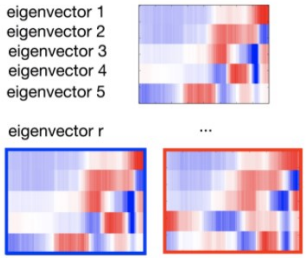


Reproducibility score:

$$\frac{\sum_i \sum_j |rw(A)_{ij} - rw(B)_{ij}|}{\#\ genomic\ bins}$$

HiC-Spector

Transformation: eigen-decomposition of Laplacian



Comparison: weighted difference of eigenvectors

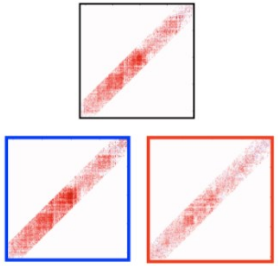
$$S_d(A,B) = \sum_{i=0}^{r-1} \|v_i^A - v_i^B\|$$

Reproducibility score:

$$\left(1 - \frac{1}{r}\frac{S_d}{l}\right) \qquad l = \sqrt{2}$$

QuASAR-Rep

Transformation: correlation matrix of distance-based contact enrichment

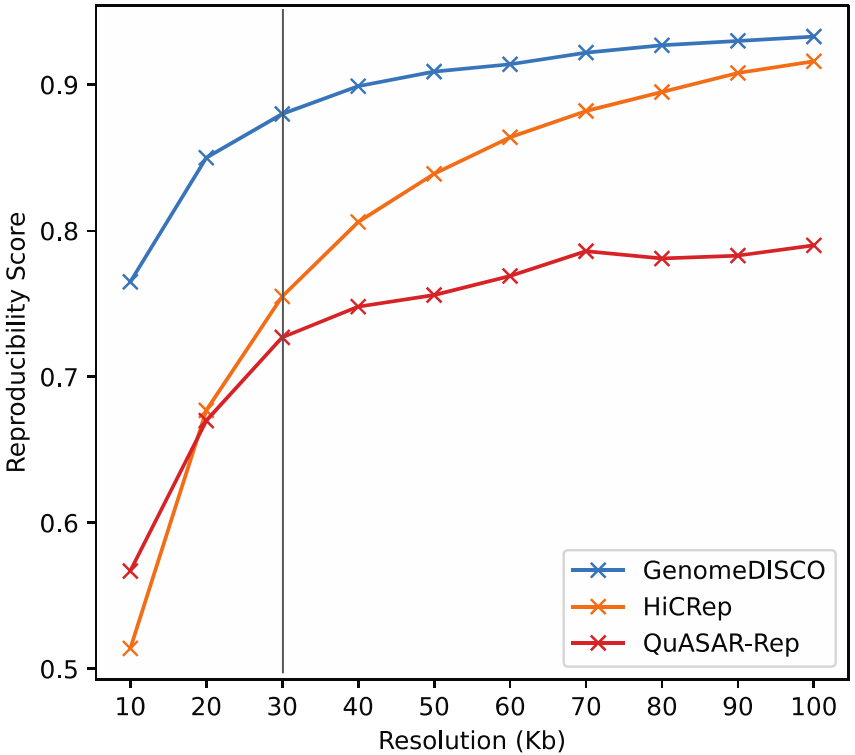


Comparison: compute correlation of values in the 2 transformed matrices

Reproducibility score:

$$Pearson\ correlation\ (\ quasar(A),\ quasar(B)\ )$$

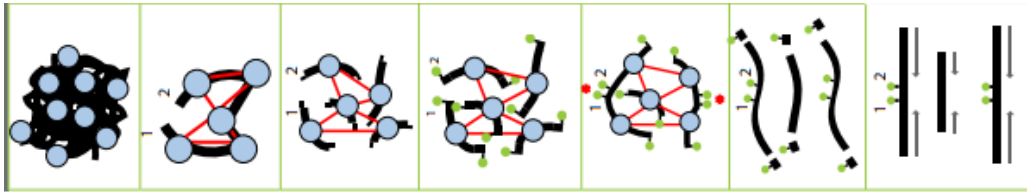
Finding appropriate resolution



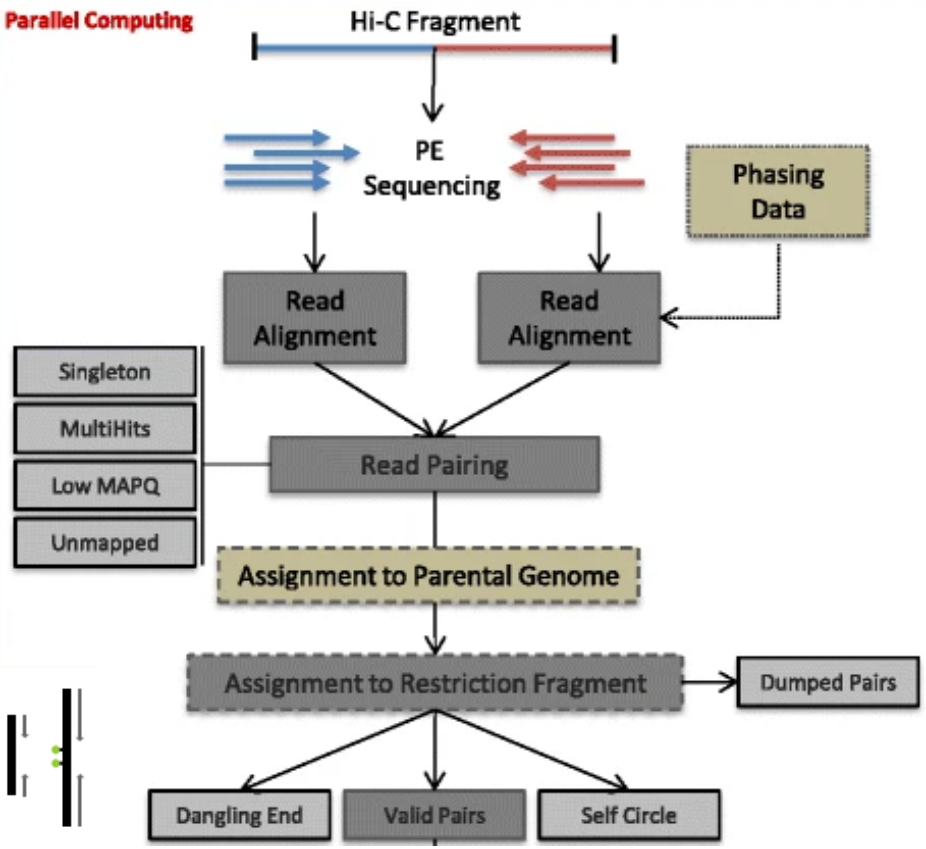
- Low read depth at high resolution → high variance across bootstraps
- Reducing resolution beyond a certain point → no benefit
- Typically split data in half

Hi-C data analysis

- Starts off as typical
- Filtering for **valid reads** that capture two regions of the genome

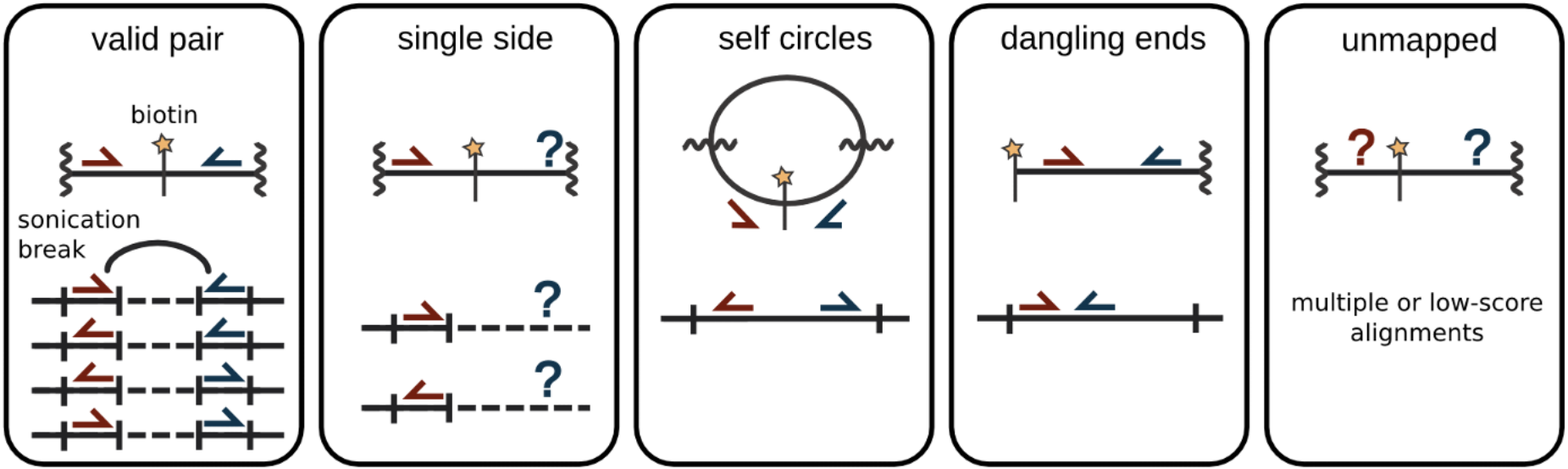


Source: Dovetail Genomics Hi-C pamphlet



Servant et al. Genome Biology (2015)

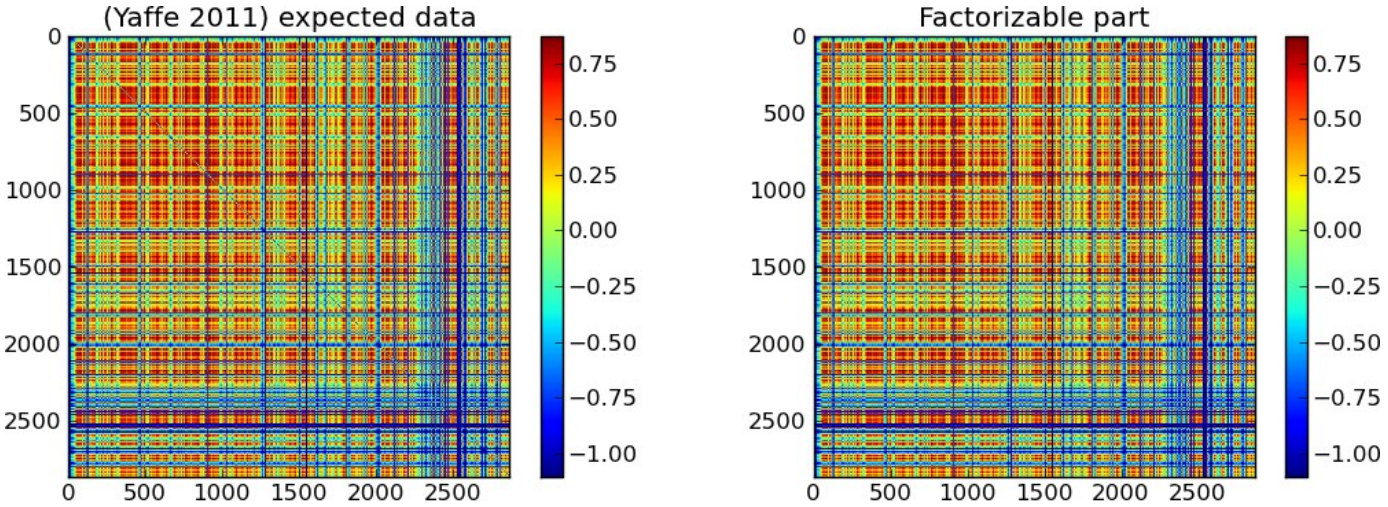
Valid fragments vs unwanted products



Imakaev et al. Nature Methods (2012)

- Look for read pairs whose orientation point away from restriction sites

Normalizing read count bias

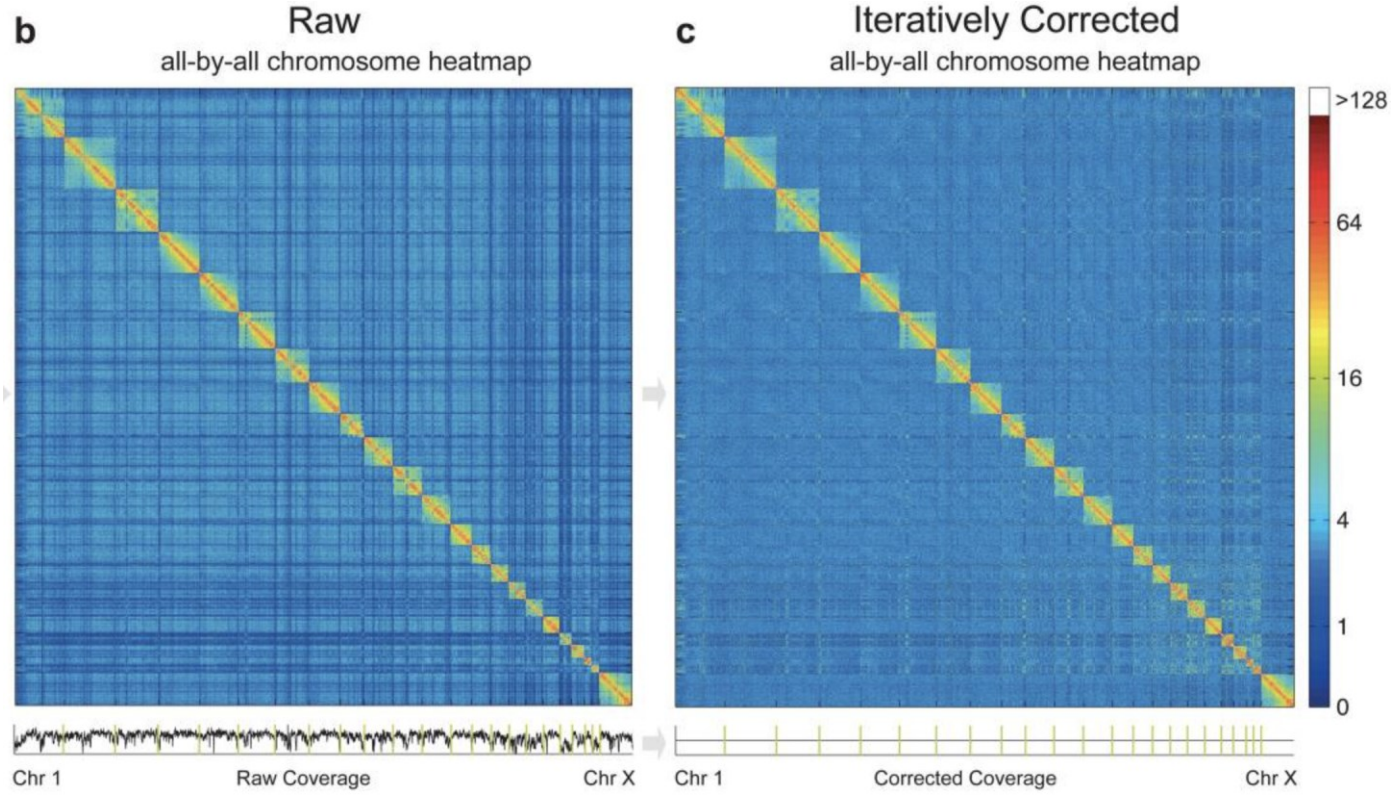


Imakaev et al. Nature Methods (2012)

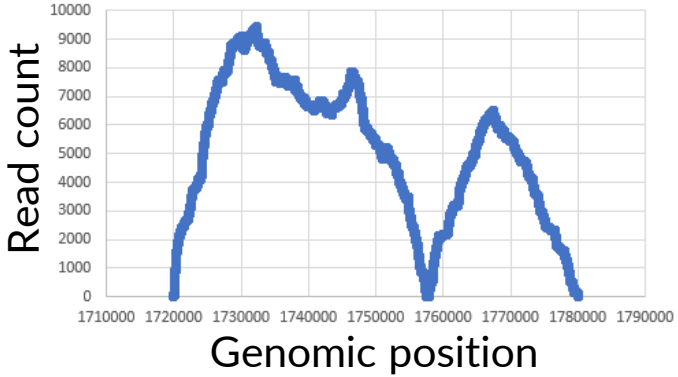
- Total bias for interaction between loci A and B = **bias at A** x **bias at B**
  - P(getting DNA fragment from a genomic region by chance)

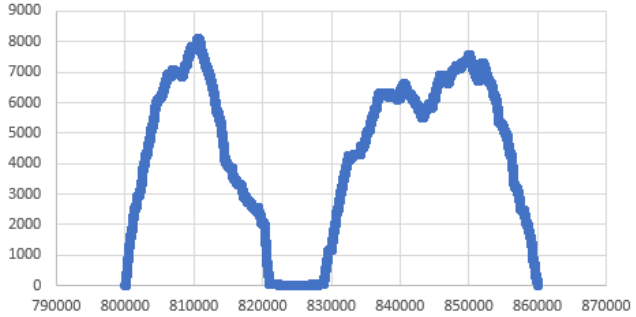
Iterative correction and eigenvector decomposition (ICED)

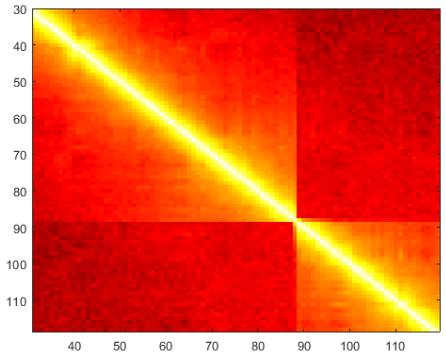
Imakaev et al. Nature Methods (2012)

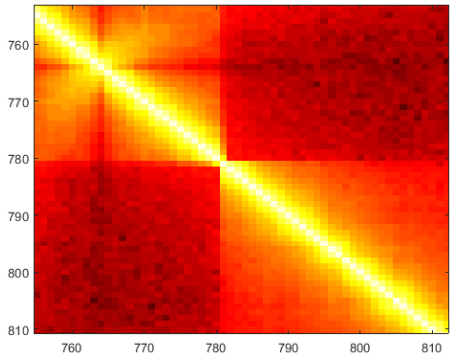


Hi-C identifies misassembled genomic region



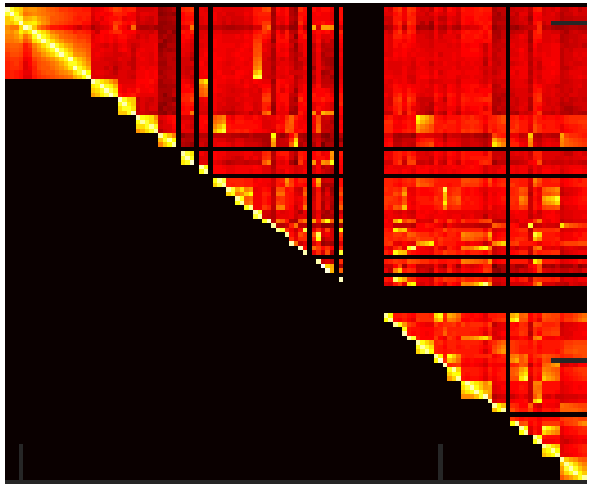


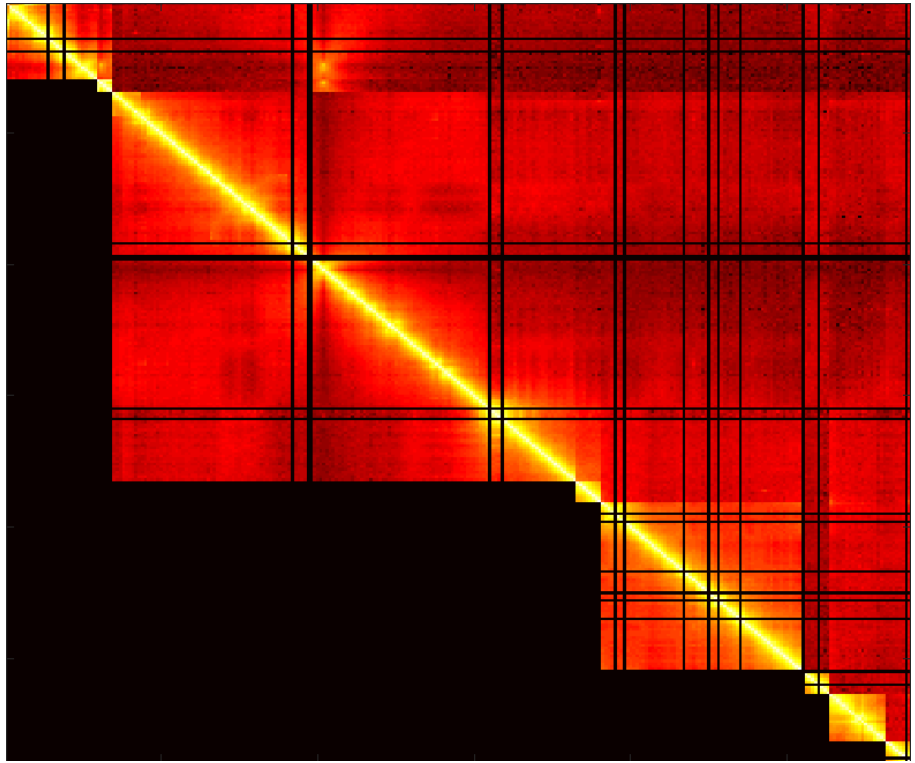




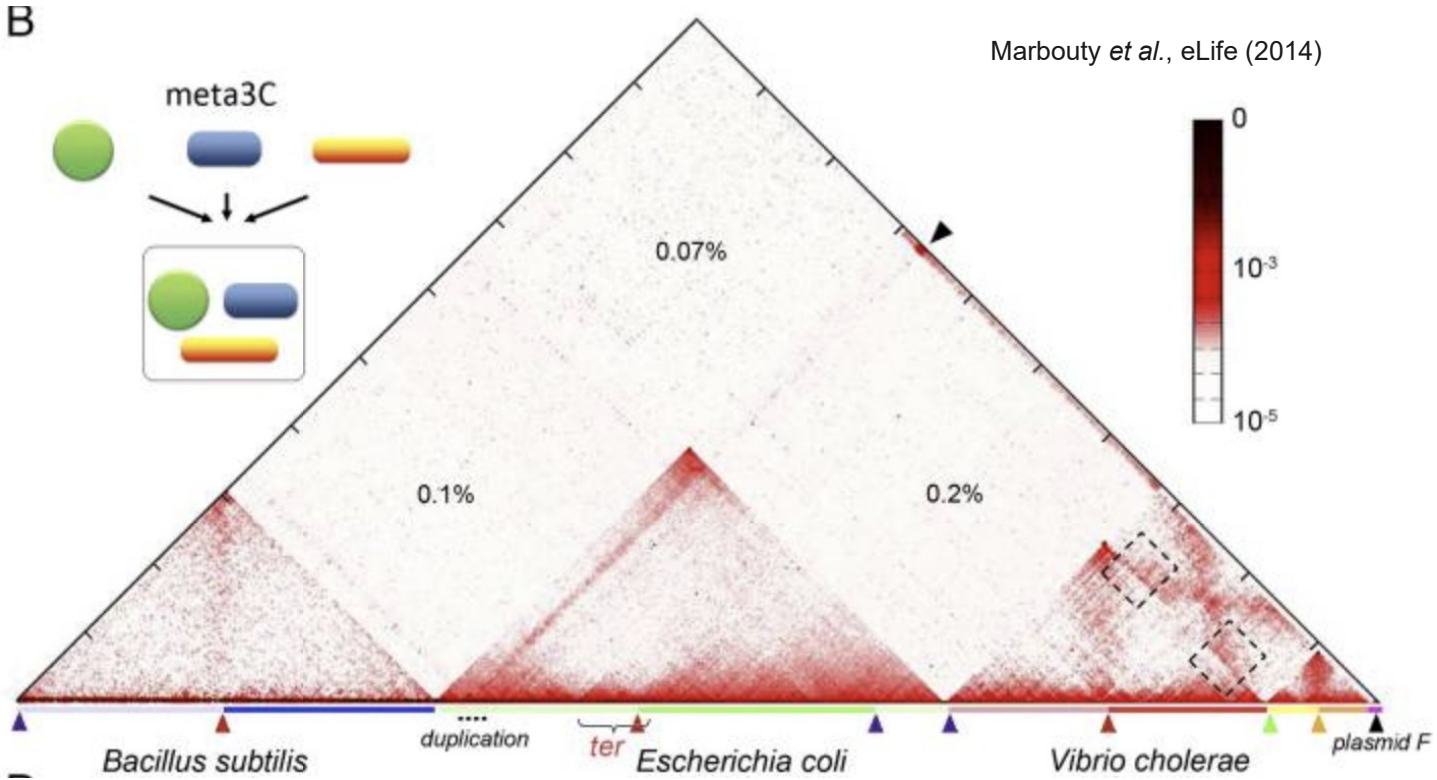
Hi-C guides re-assembly

Many small contigs





Hi-C guides metagenomic analysis

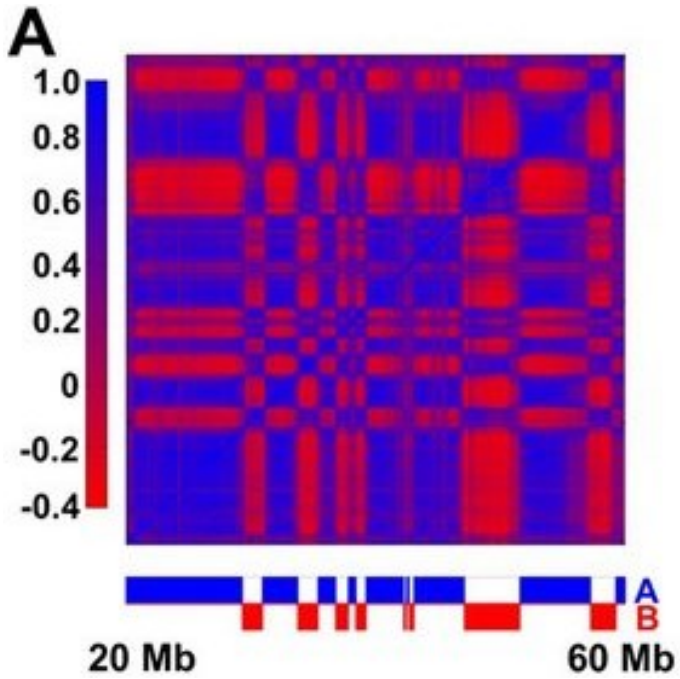


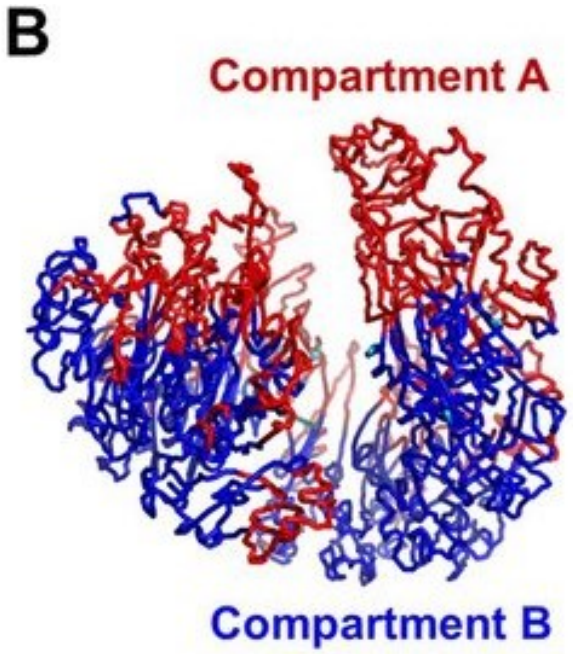
Marbouty *et al.*, eLife (2014)

A/B compartment and topologically associating domain (TAD)

A/B compartment

Xie at al. Scientific Reports (2017)

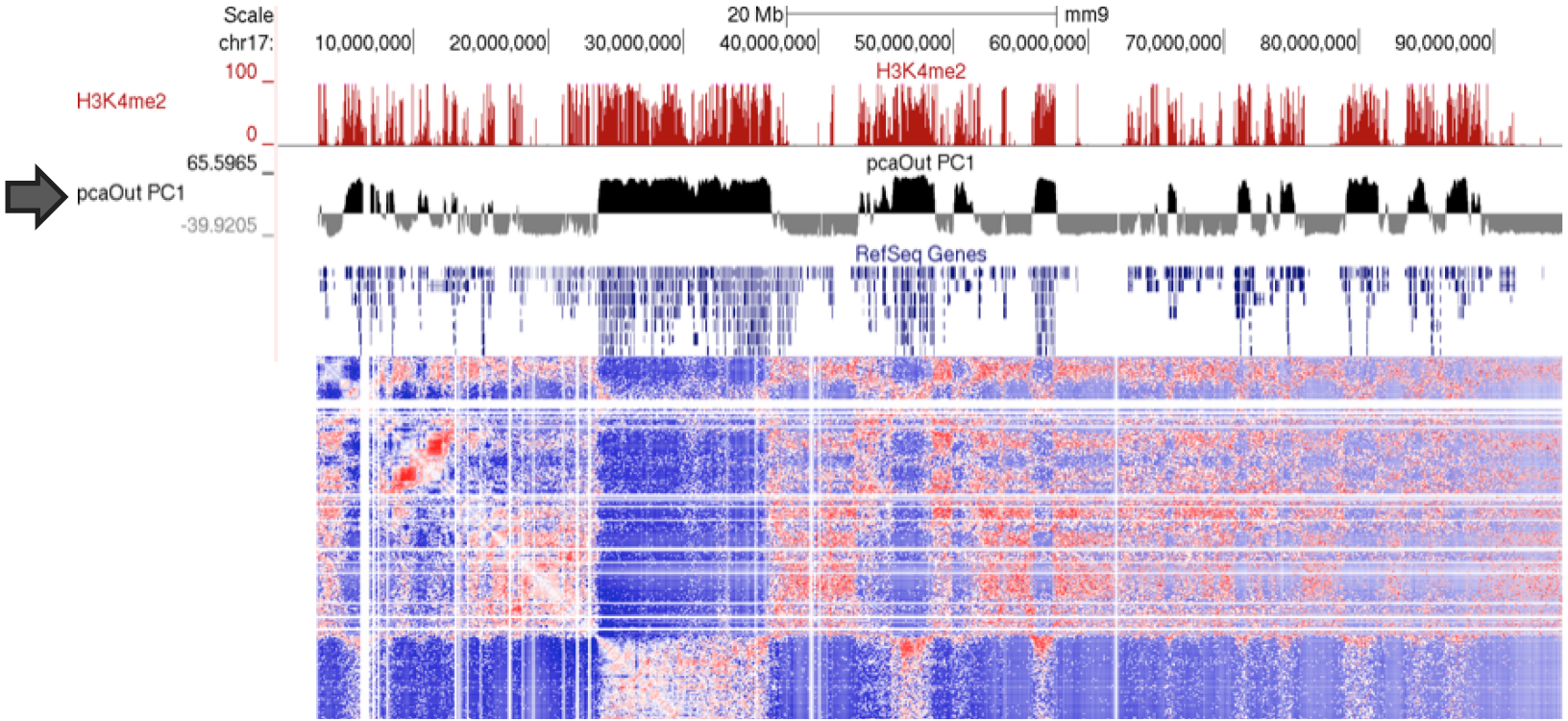




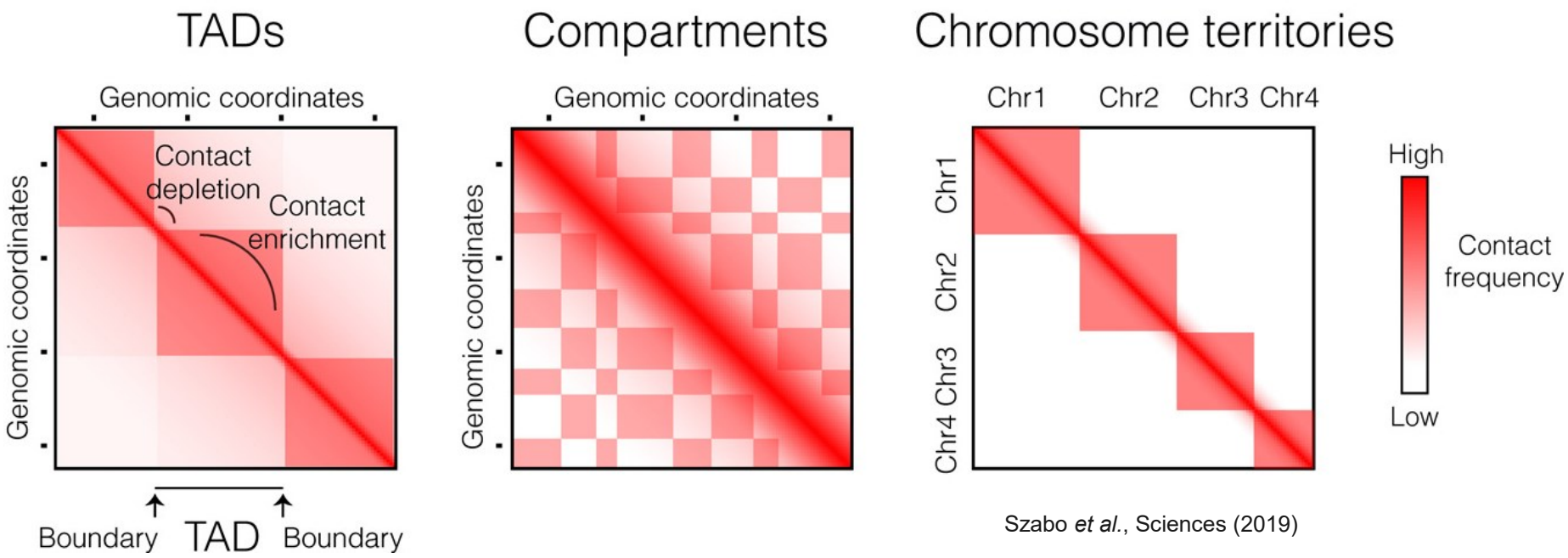
- Caused by localization of euchromatin and heterochromatin

A/B compartment identification with PCA

Source: homer.ucsd.edu/homer/interactions/HiCpca.html



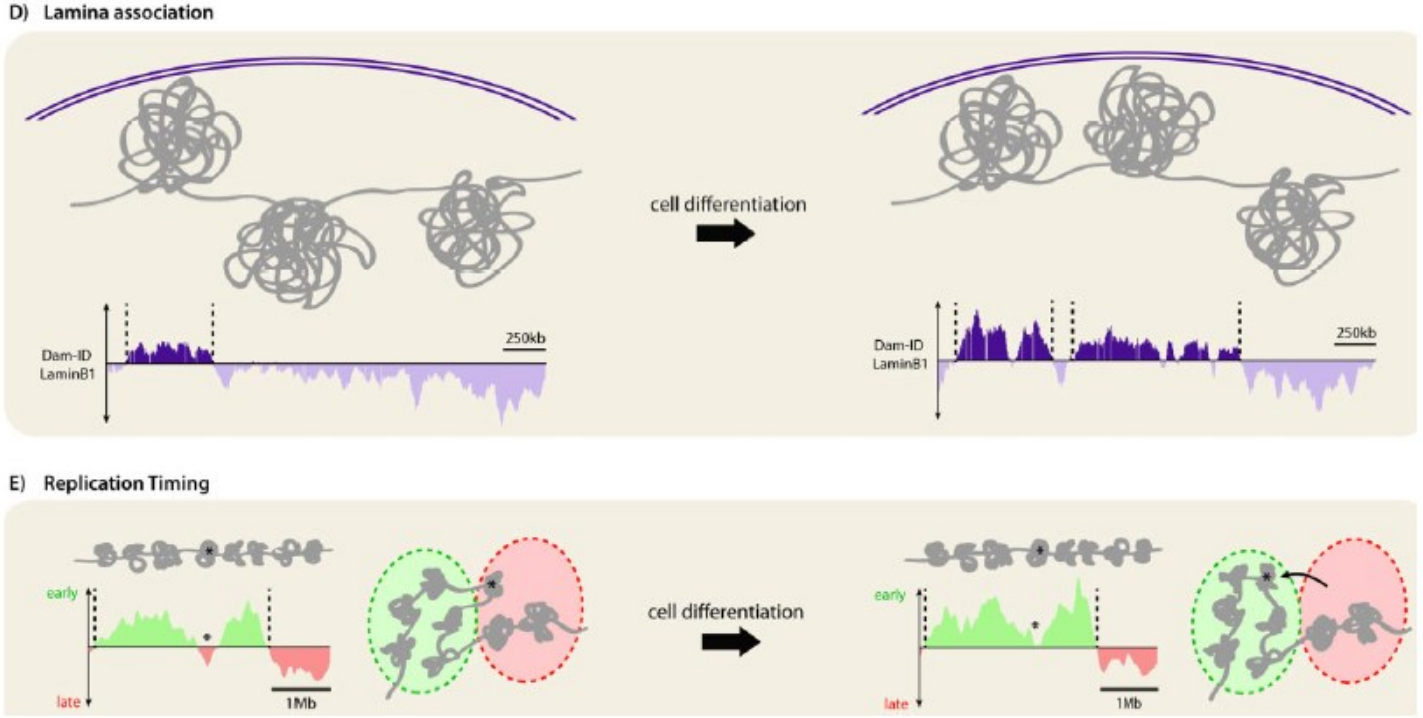
Hierarchical organization of chromatin



Szabo *et al.*, Sciences (2019)

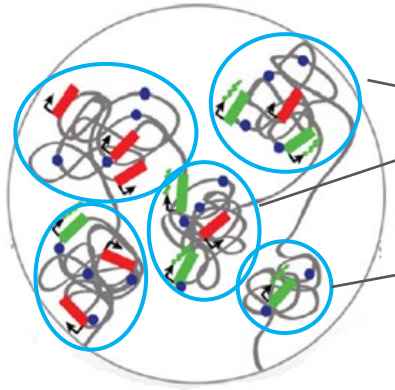
- TAD size ranges from 100kb to 10Mb

TAD = unit of chromatin structure

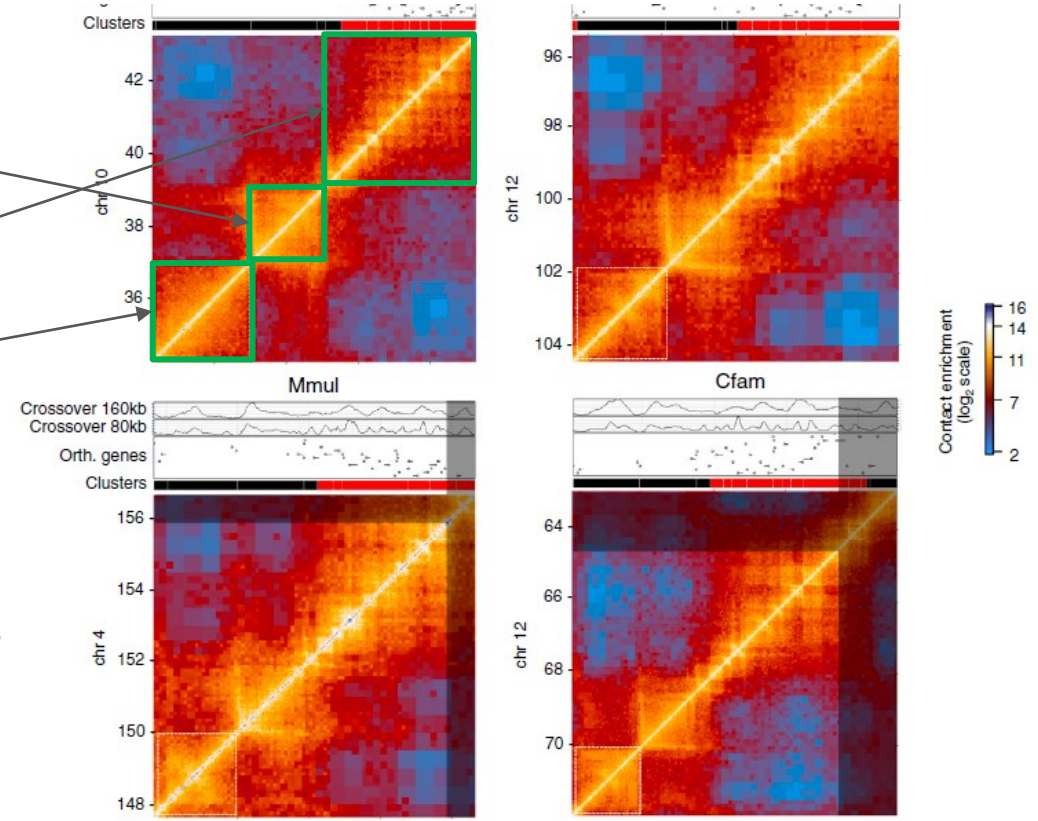


Source: Nora *et al*. Bioessay (2013)

Evolutionary conservation of TAD



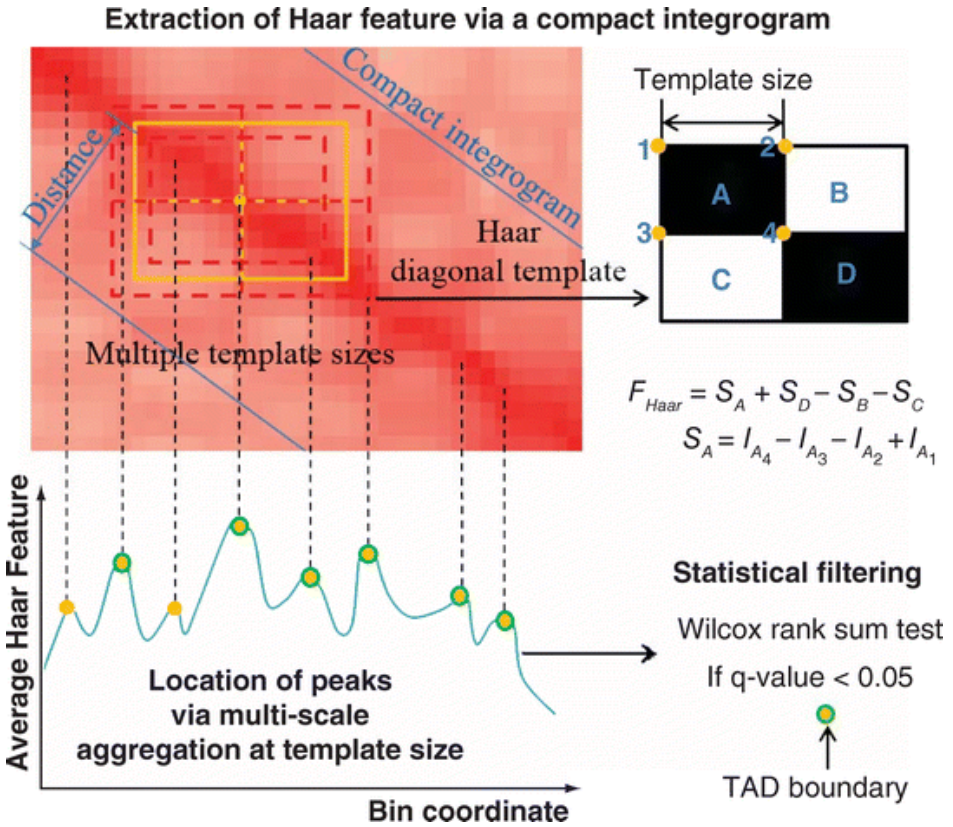
Rudan *et al*. Cell Reports (2015)



- Structural information is (partly) encoded in the DNA sequence

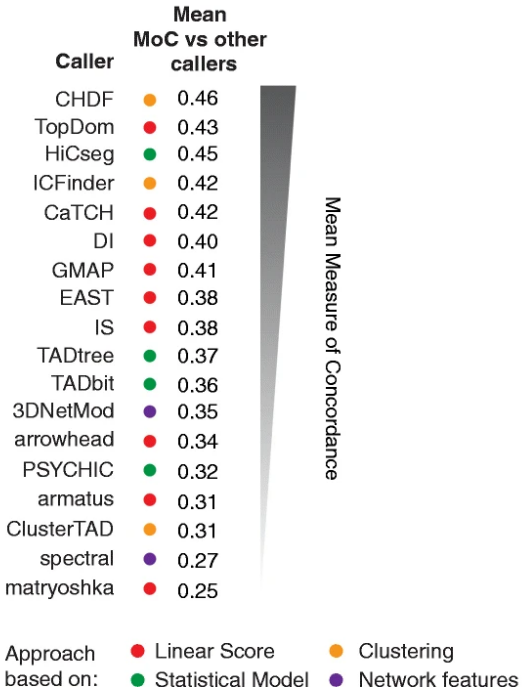
TAD calling

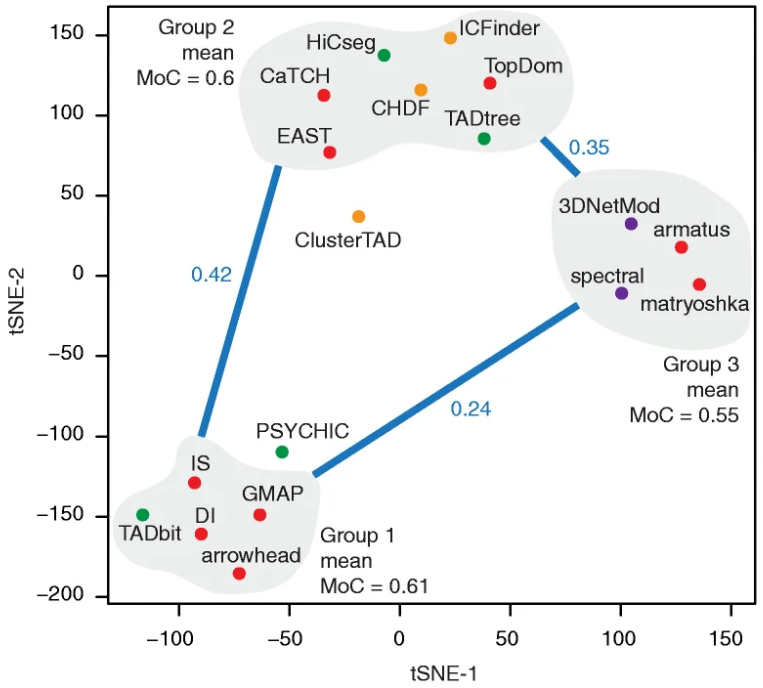
- TAD boundary = location where interaction difference between A+D and B+C is maximized
- Peak detection
- Mann-Whitney U test

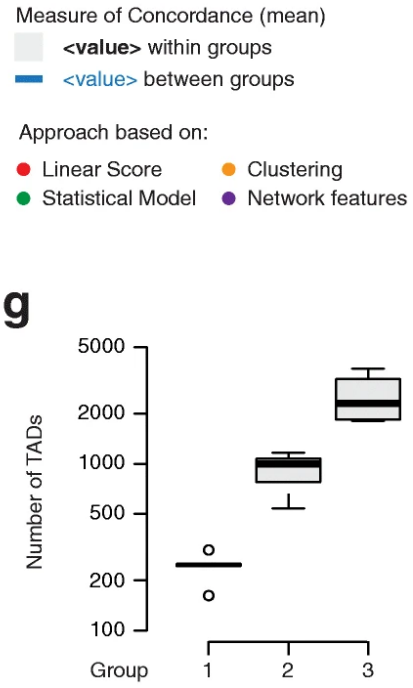


Lyu *et al.*, Biotechniques (2020)

Variety of TAD calling



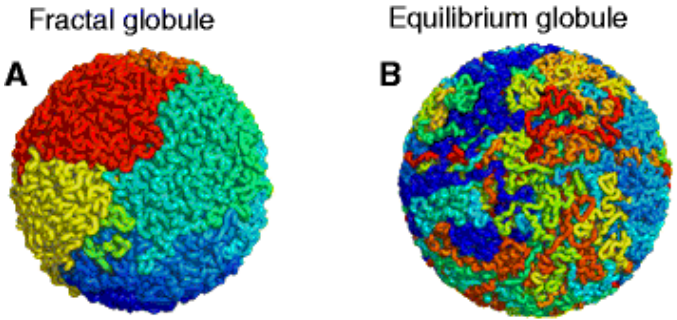




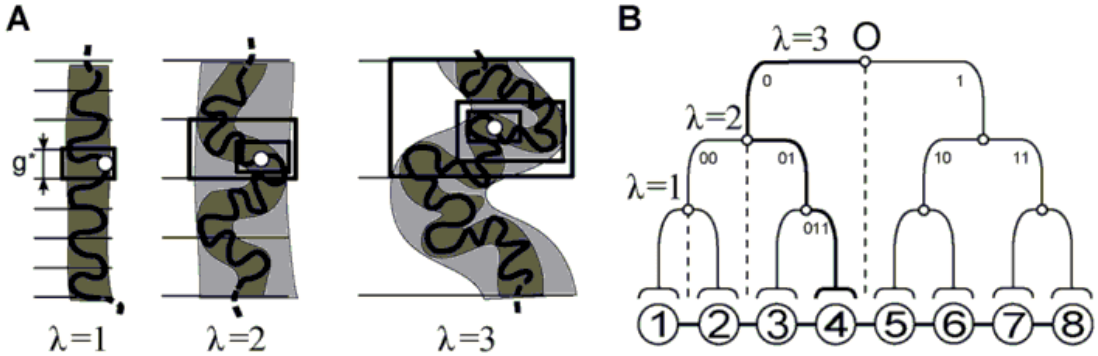
Zufferey *et al.*, Genome Biology (2018)

Chromatin structural modeling

Coarse early models



Mirny. Chromosome Research (2011)

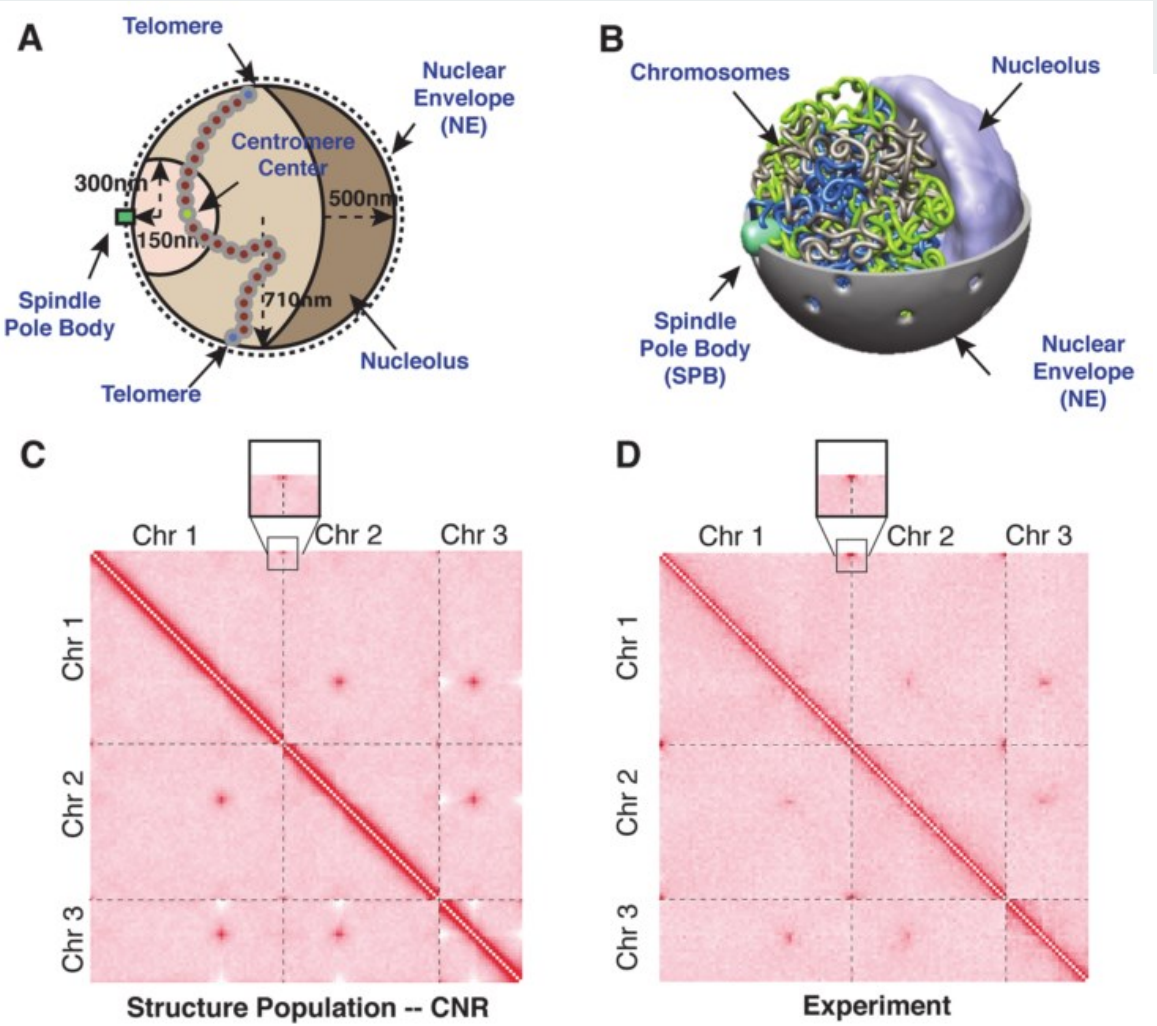


Nazarov *et al*. Soft Matter (2014)

- Average properties of chromatin as a very long polymers
- Can explain chromosome territory

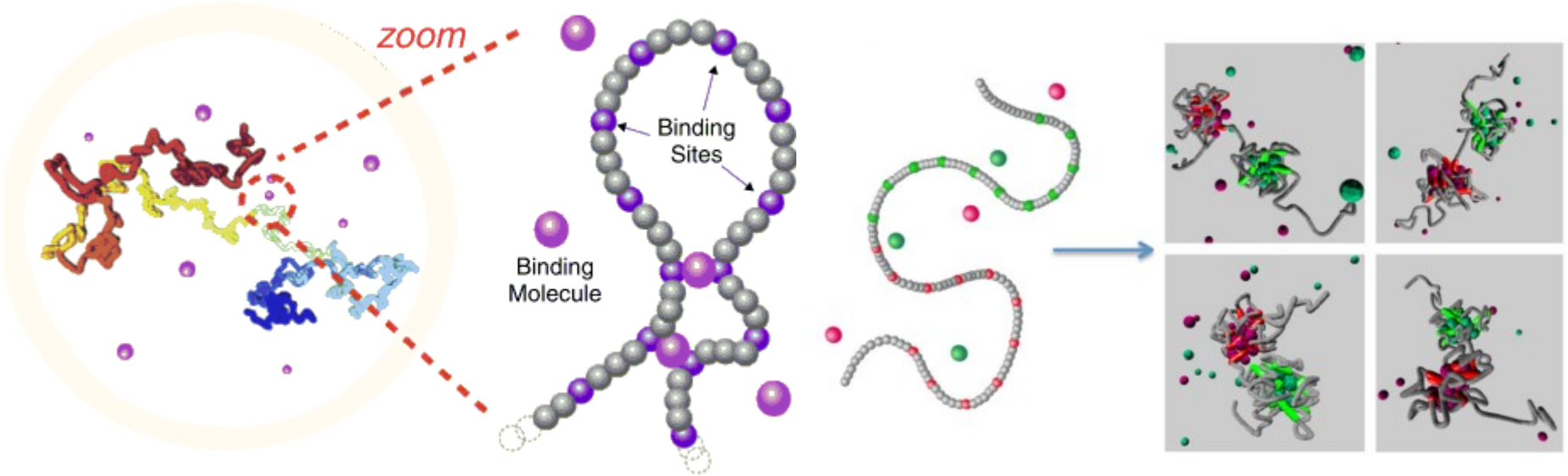
Polymer simulation

- Control the positions of centromeres, telomeres, and rDNA
- Let the rest behave like a flexible polymer chain
- No interaction



Gong *et al*. PLoS ONE (2015)

Strings and binders switch model

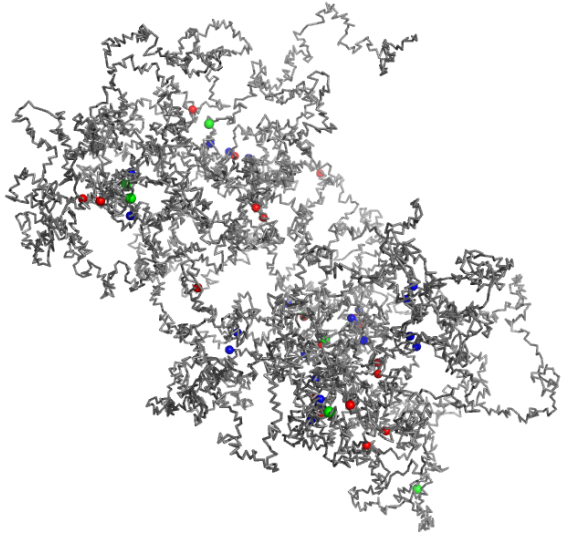


Nicodemi and Pombo. Curr Opin Cell Biol (2014)

Barbieri, M. et al. PNAS (2012).

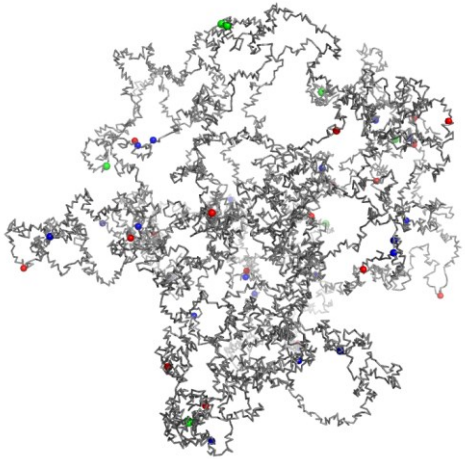
- Accommodate protein binding at specific loci

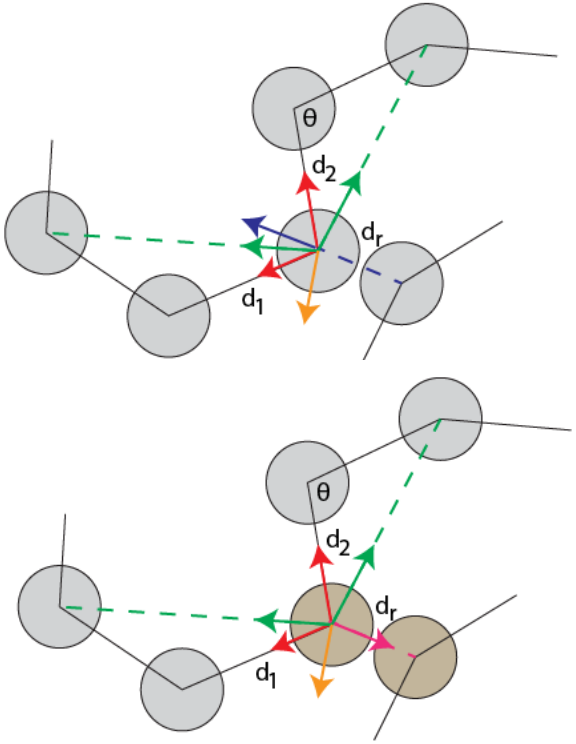
Strings and binders simulation



With Protein

Without Protein



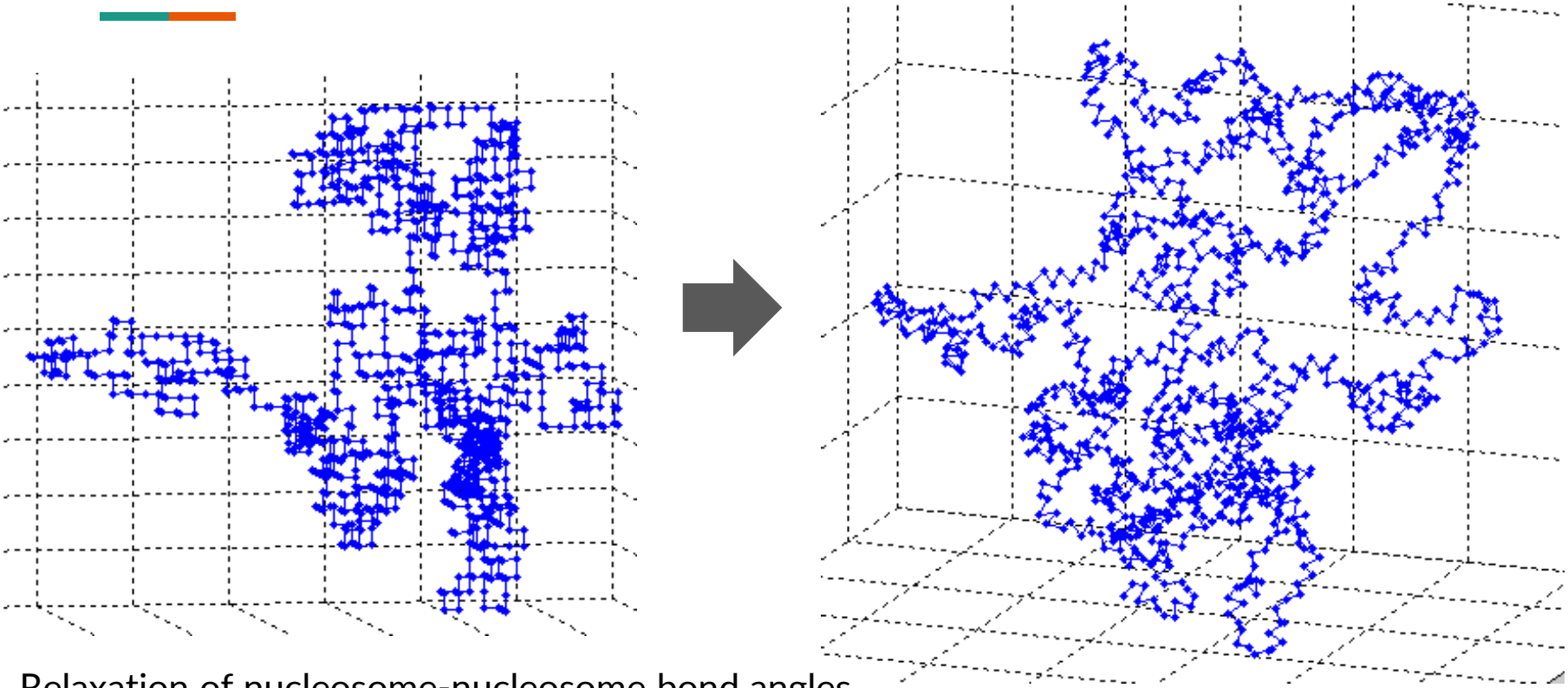


Regular unit

Protein binding site

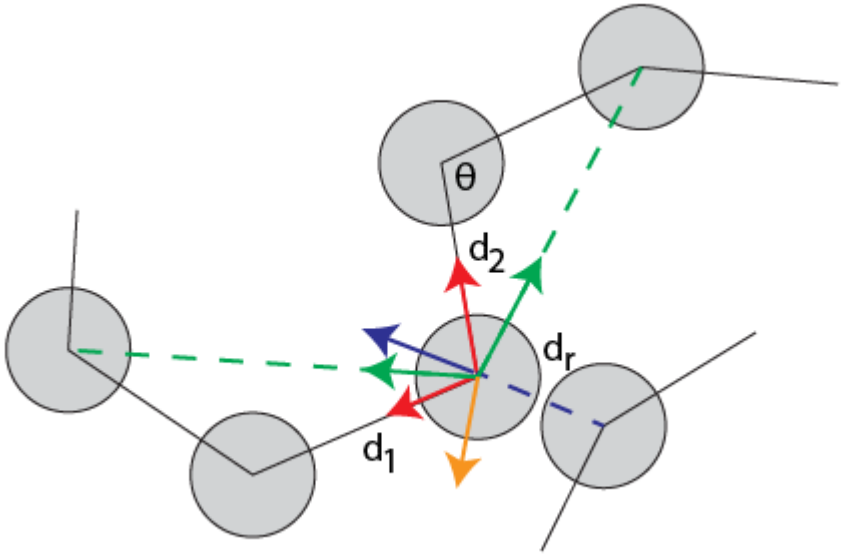
1 Mb represented by a 7000-unit polymer chain
Each unit = 150 bp ~ 1 nucleosome

Initialization with self-avoiding random walk



Relaxation of nucleosome-nucleosome bond angles

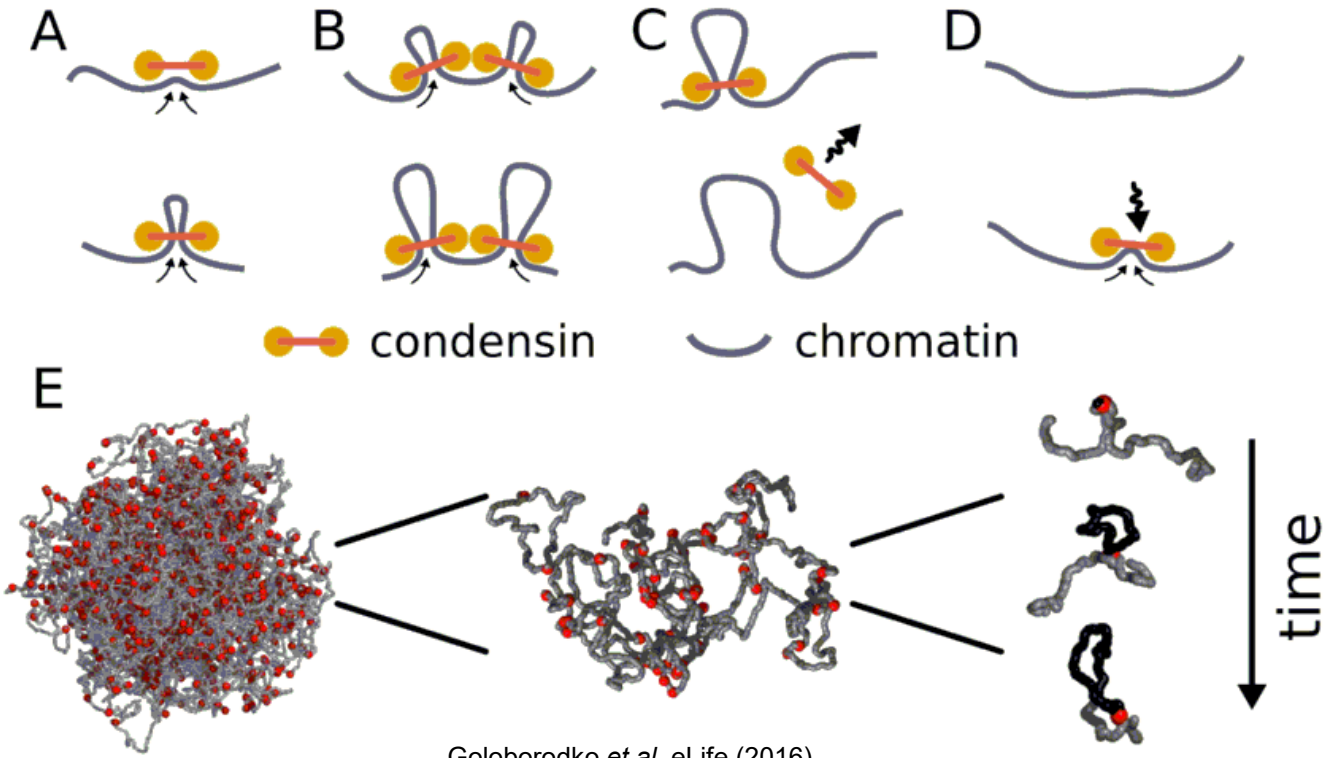
Physical forces involved



- Entropic (diffusional) = constant force with random direction
- Tension (bead-bead) = spring force between adjacent beads
- Repulsion = dispersion force between close-by beads
- Attraction = artificial force to control the distribution of θ

Loop extrusion model

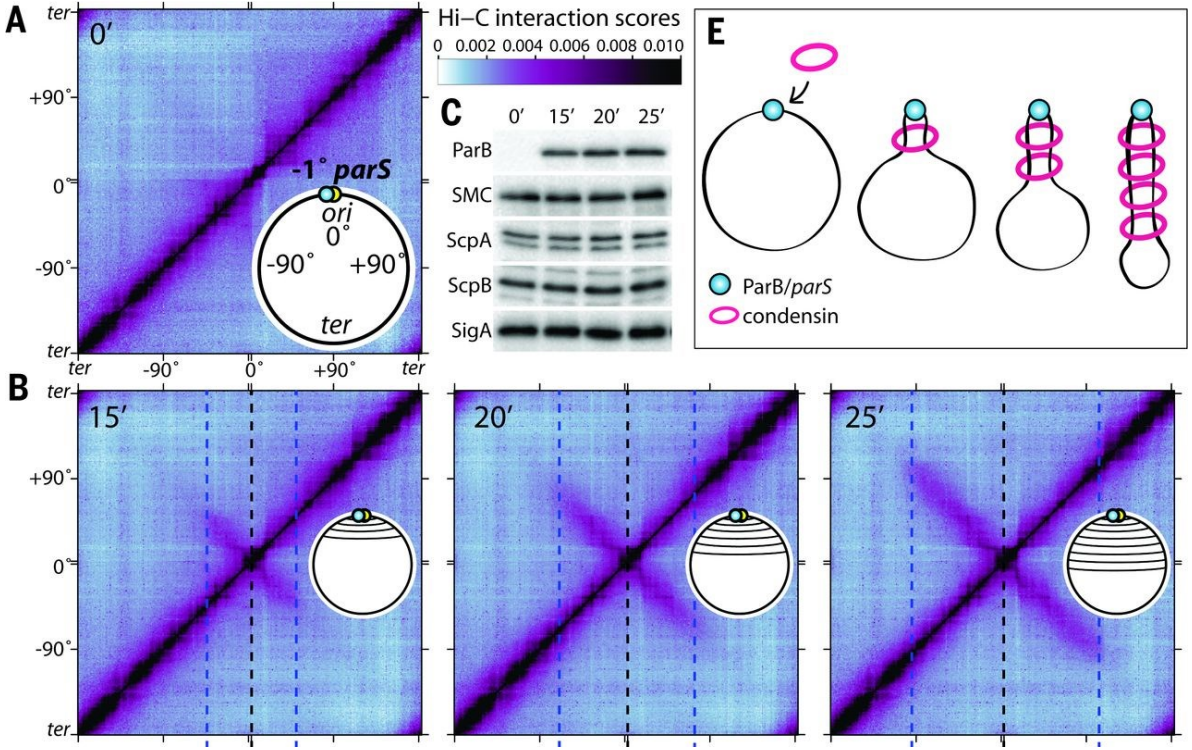
Loop extrusion by proteins



Goloborodko *et al*. eLife (2016)

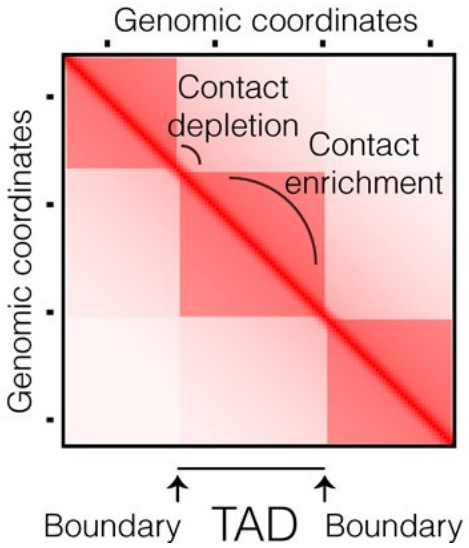
Live loop extrusion in a bacteria

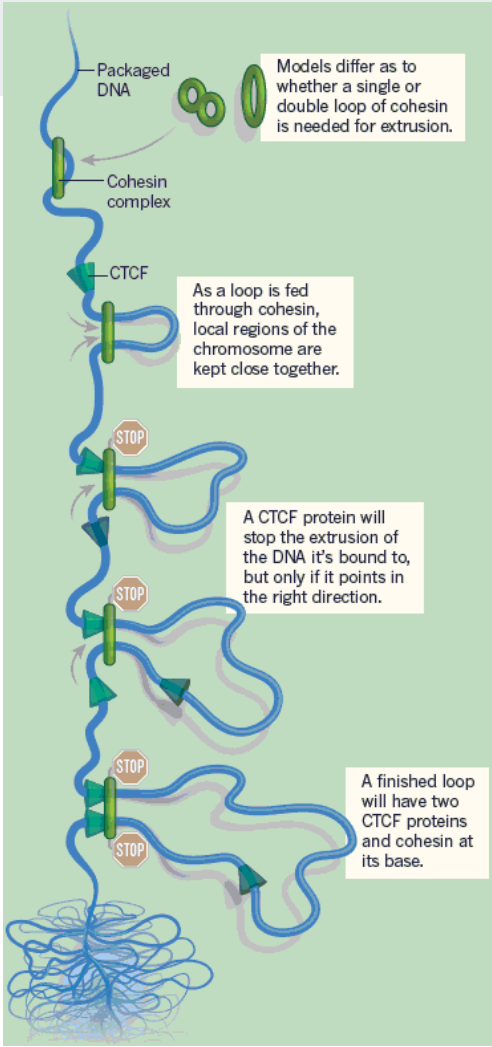
Wang *et al*. Science 355:524-527 (2017)



Loop extrusion and CTCF

- Proteins like cohesin and condensin extrude loops
- How to stop at TAD boundary?
- CTCF proteins can stop extrusion process in a specific direction
- Not every species has CTCF

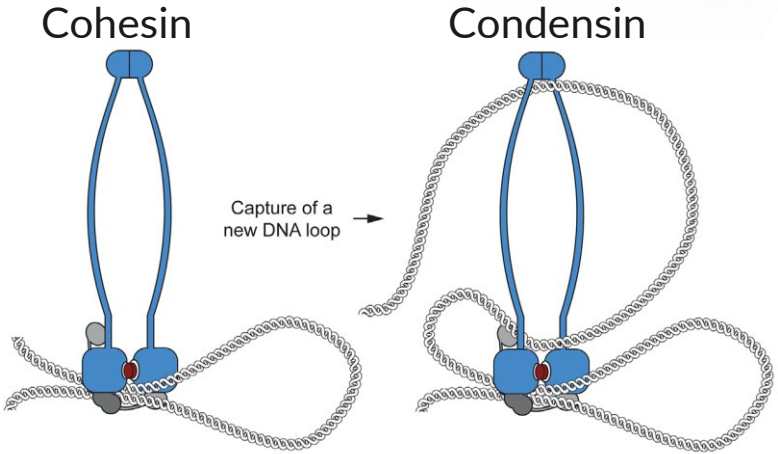




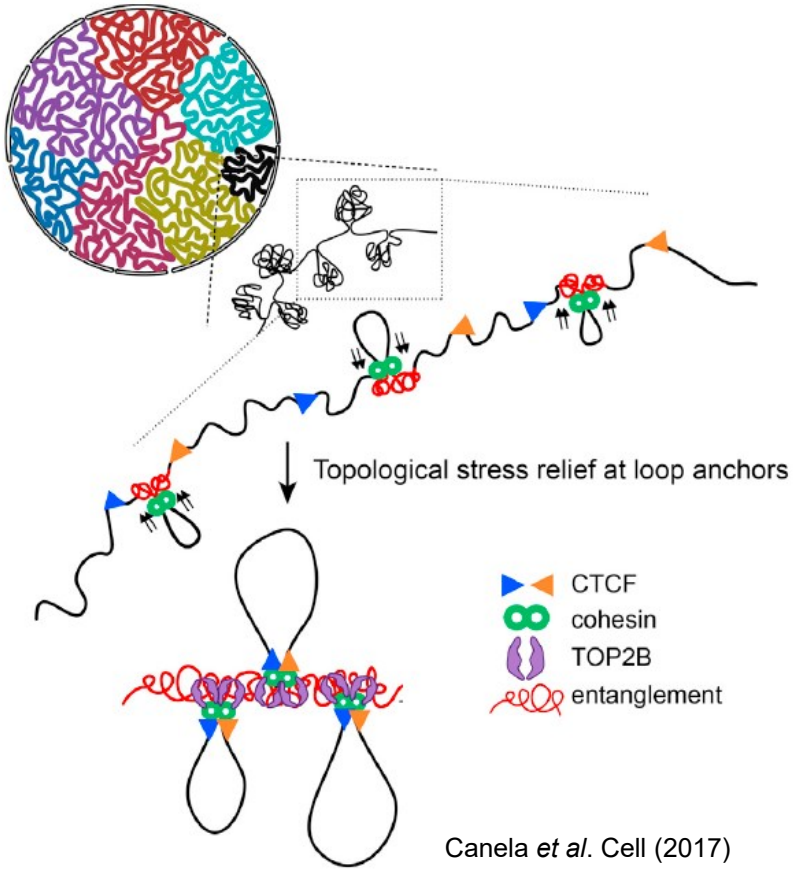
Dolgin. Science (2017)

Ingredients of TAD formation

CTCF in complex with DNA

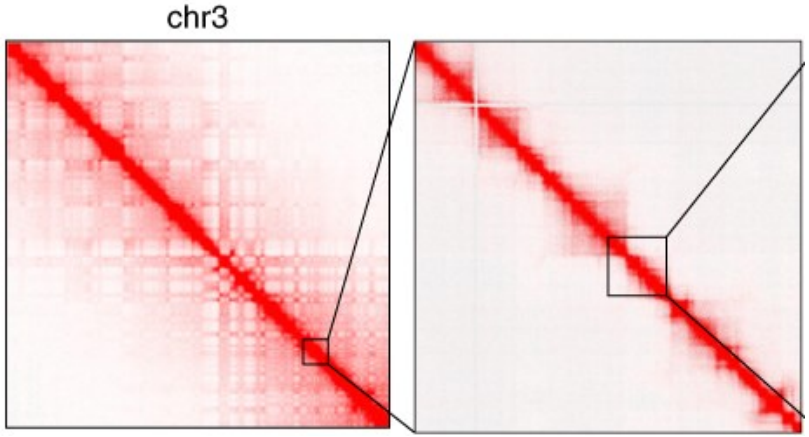


Diebold-Durand *et al*. Molecular Cell (2017)



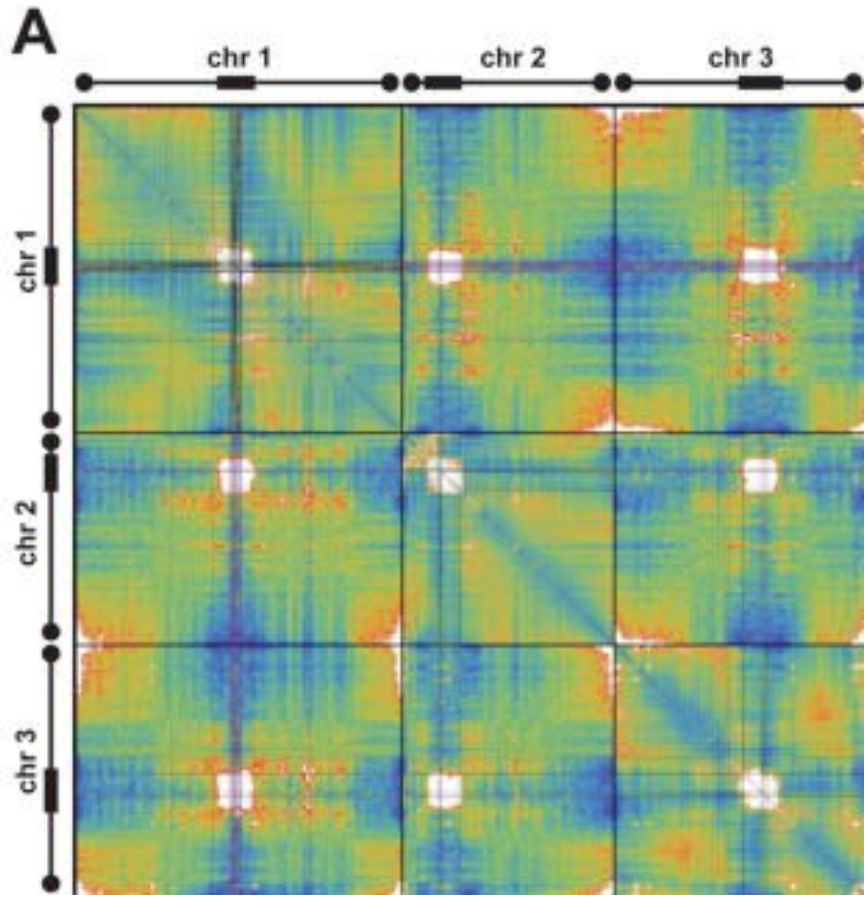
Canela *et al*. Cell (2017)

Human vs arabidopsis



Li et al. Nat Methods 16:991-993 (2019)

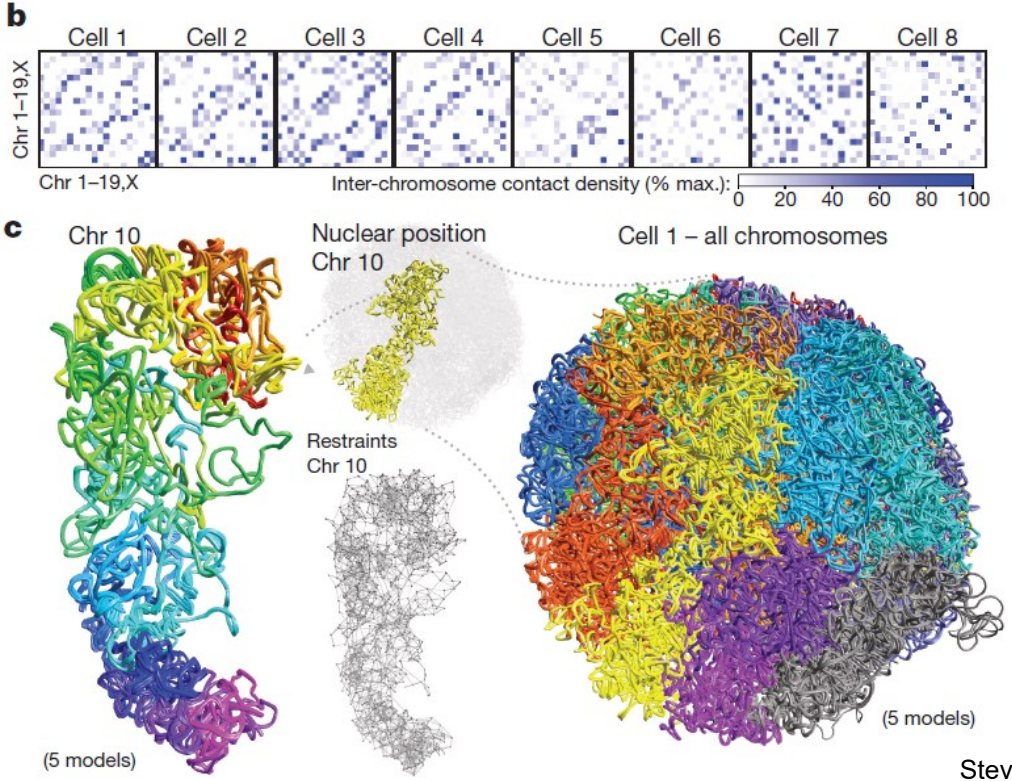
- CTCF is responsible for TADs in human genome



Feng et al. Mol Cell 55:694-707 (2014)

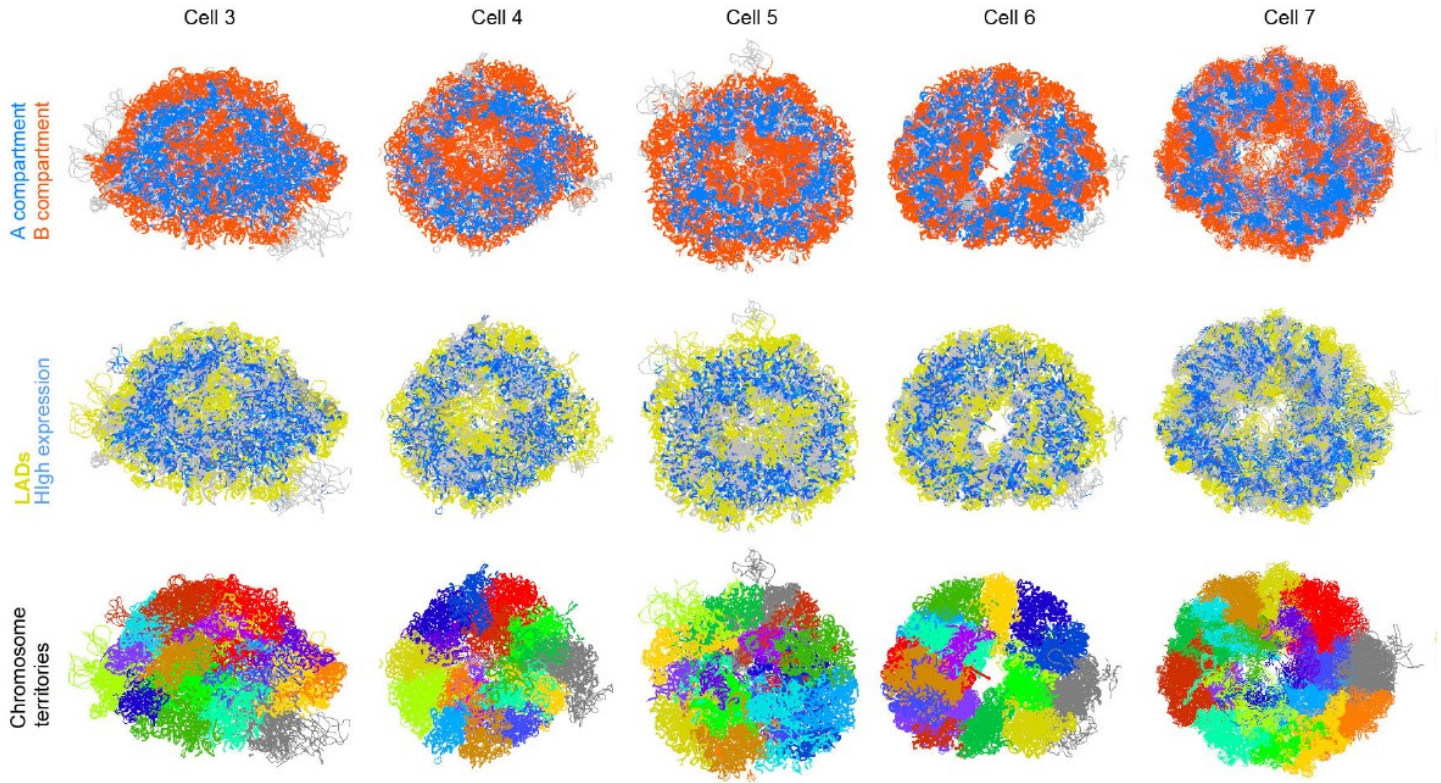
Chromatin structure is not fixed

Single-cell Hi-C

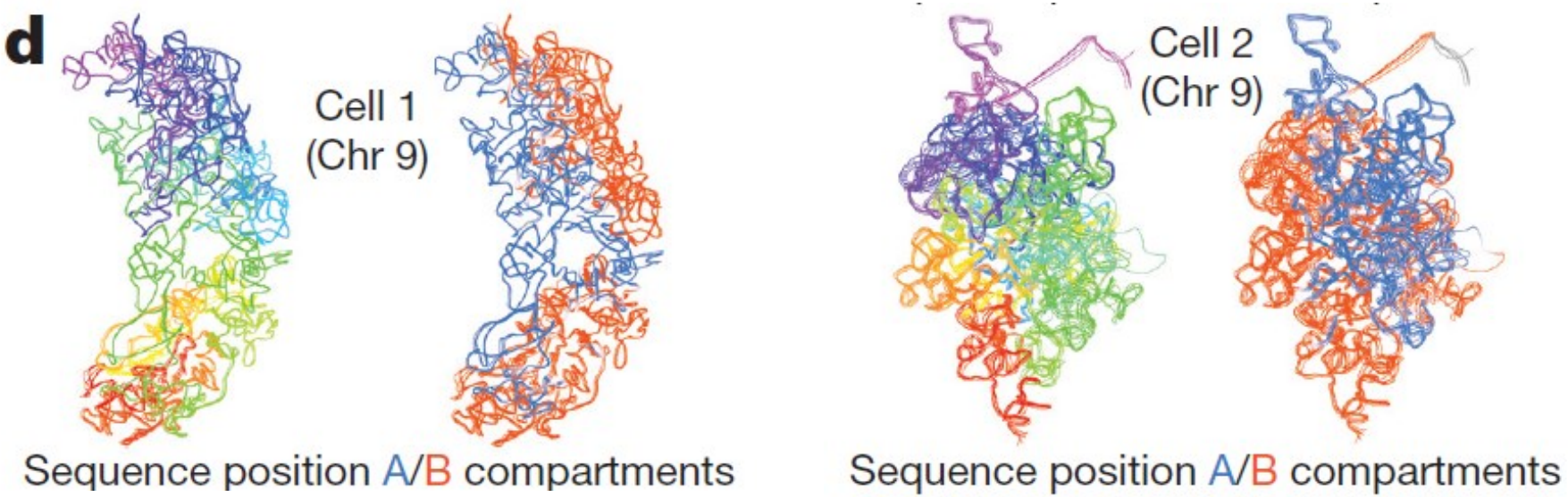


Stevens *et al*. Nature (2017)

Different structures, same organization

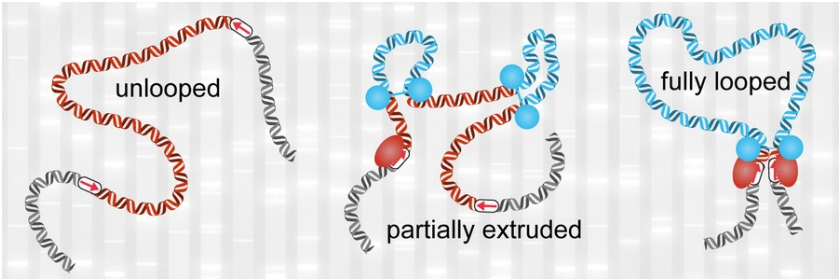


Cell-to-cell variations



Summary

- Features and functions of chromatin
  - Structural
  - Gene regulation
- Assays for probing chromatin and epigenomics
- Structural modeling of chromatin
  - Physics and proteins
- Cell-to-cell variation



https://news.mit.edu/2022/chromatin-loop-structures-gene-expression-0414

Any question?

- See you on October 30