2310640 Genomics and Systems Biology

Chromatin organization

April 25th, 2022

Sira Sriswasdi, Ph.D.
Computational Molecular Biology Group
Faculty of Medicine, Chulalongkorn University

Today's contents

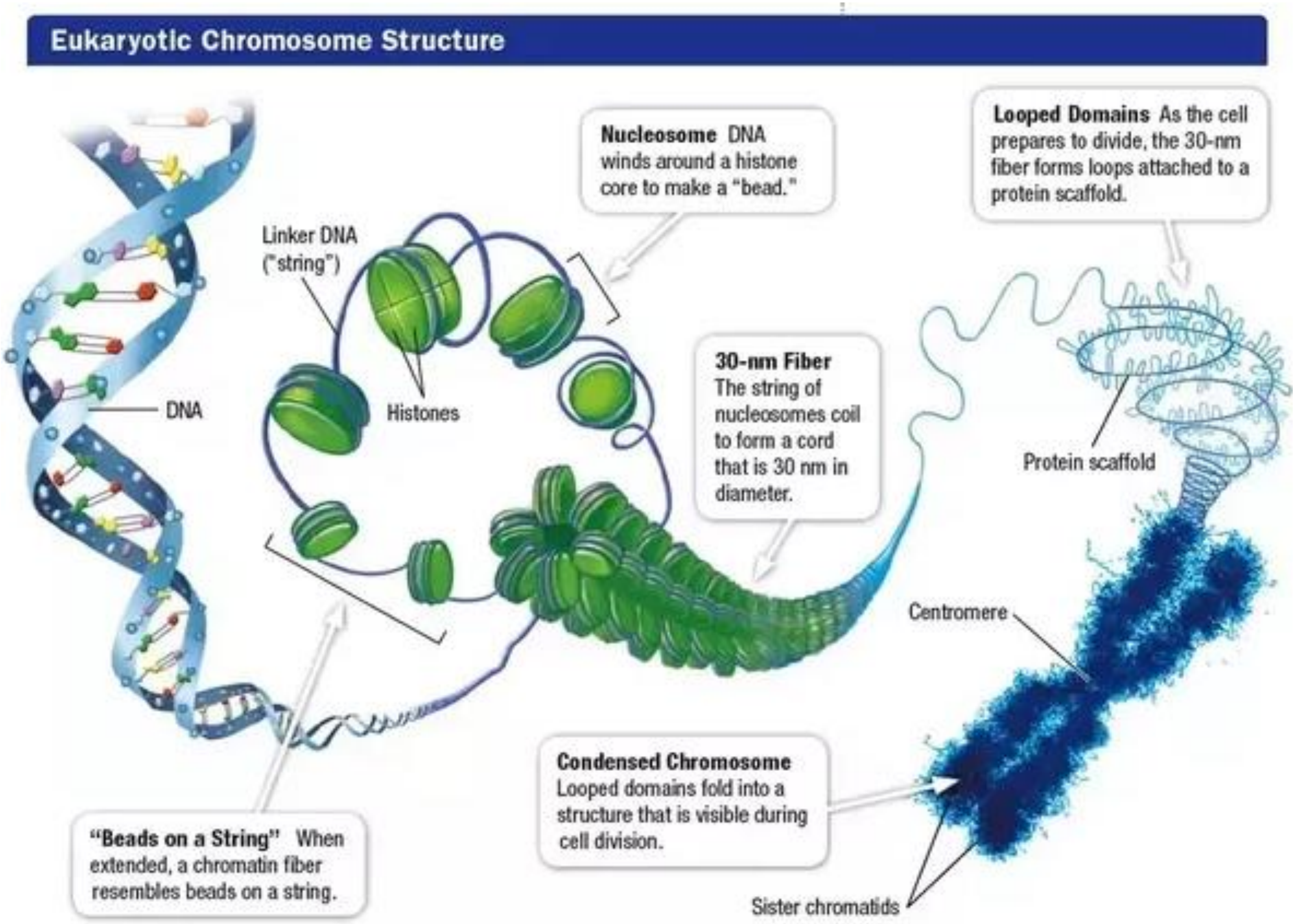
- Roles of chromatin organization in gene regulation
  - Packaging: heterochromatin and euchromatin
  - Protein (polymerase, TF, enhancer, etc.) occupation
  - Clustering of co-regulated loci

- Technology and data analysis
  - DNA methylation
  - Chromatin accessibility
  - Protein occupancy and histone mark
  - Enhancer-target interaction
  - Chromatin 3D structure

- Building blocks of chromatin structure
  - Topologically associating domain

Why study chromatin organization?

Chromatin packaging



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

Mutation sites affected by chromatin state

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

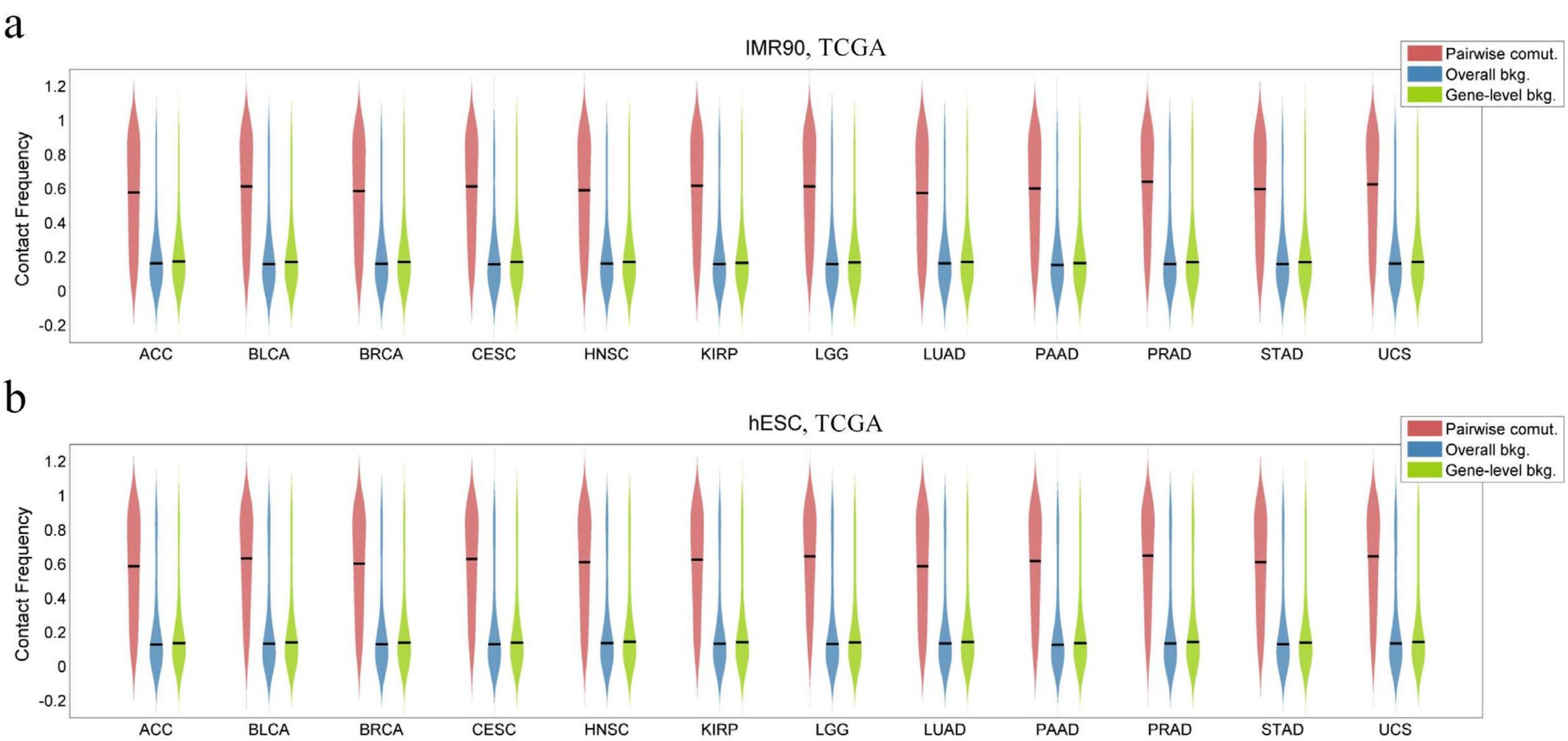
A genome-wide scan for correlated mutations detects macromolecular and chromatin interactions in *Arabidopsis thaliana*

Laura Perlaza-Jiménez, Dirk Walther

Nucleic Acids Research, Volume 46, Issue 16, 19 September 2018, Pages 8114–8132, https://doi.org/10.1093/nar/gky576

Published: 09 July 2018

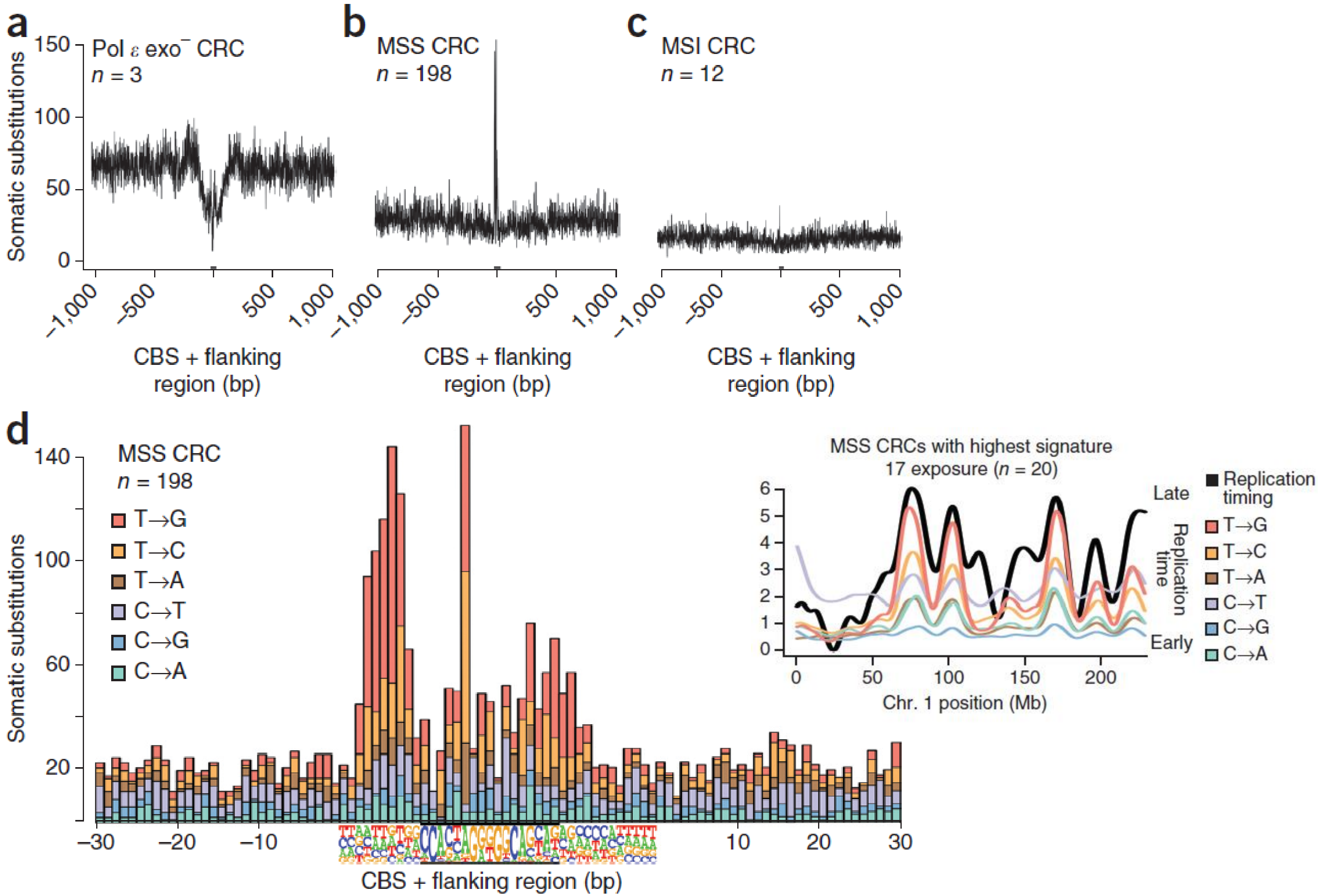
Cancer mutations are clustered in 3D



Shi *et al*. Scientific Reports (2016)

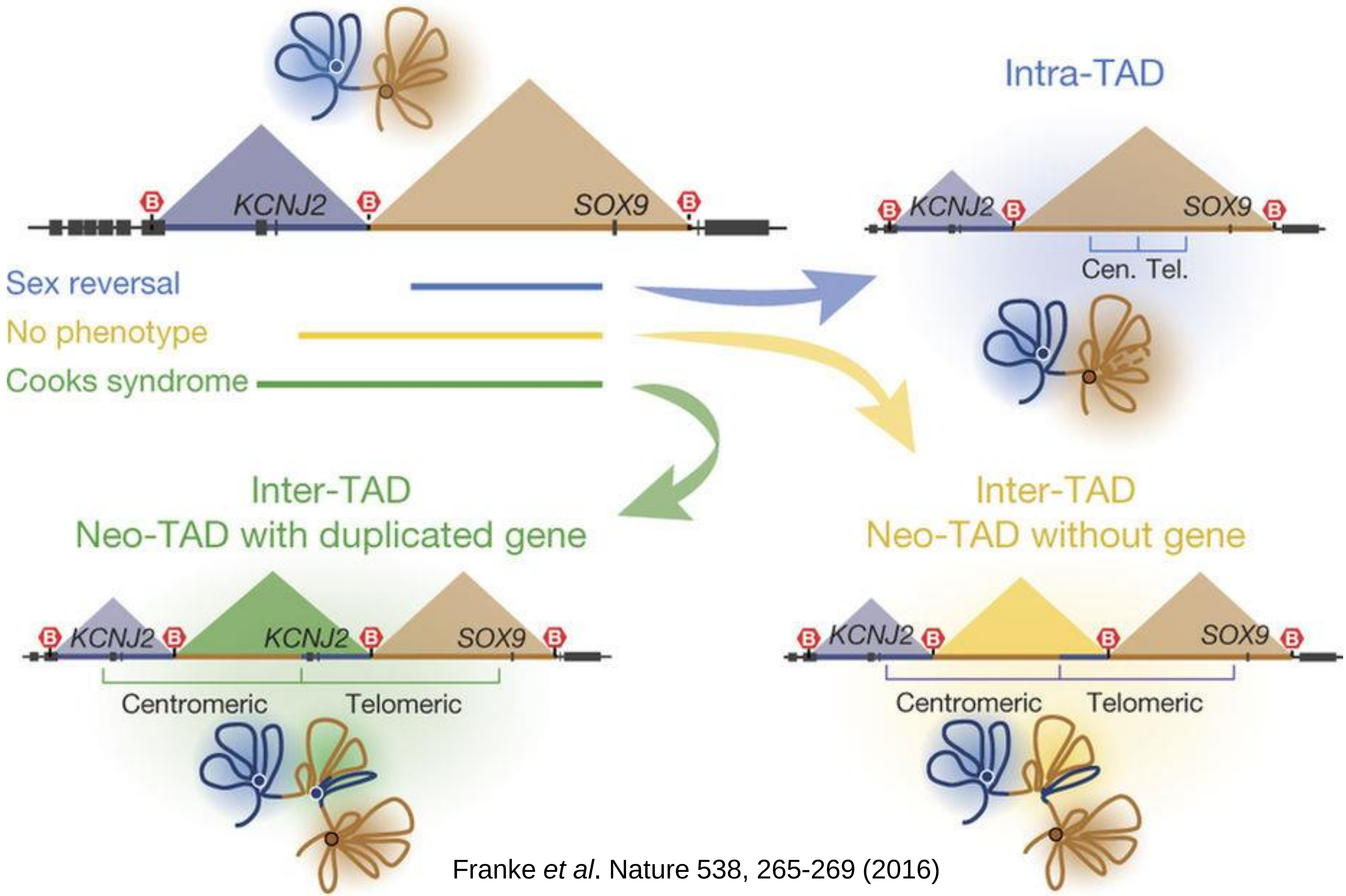
- Red = co-mutated genomic loci
- Blue & green = random pairs of loci

Cancer subtype with destabilized chromatin



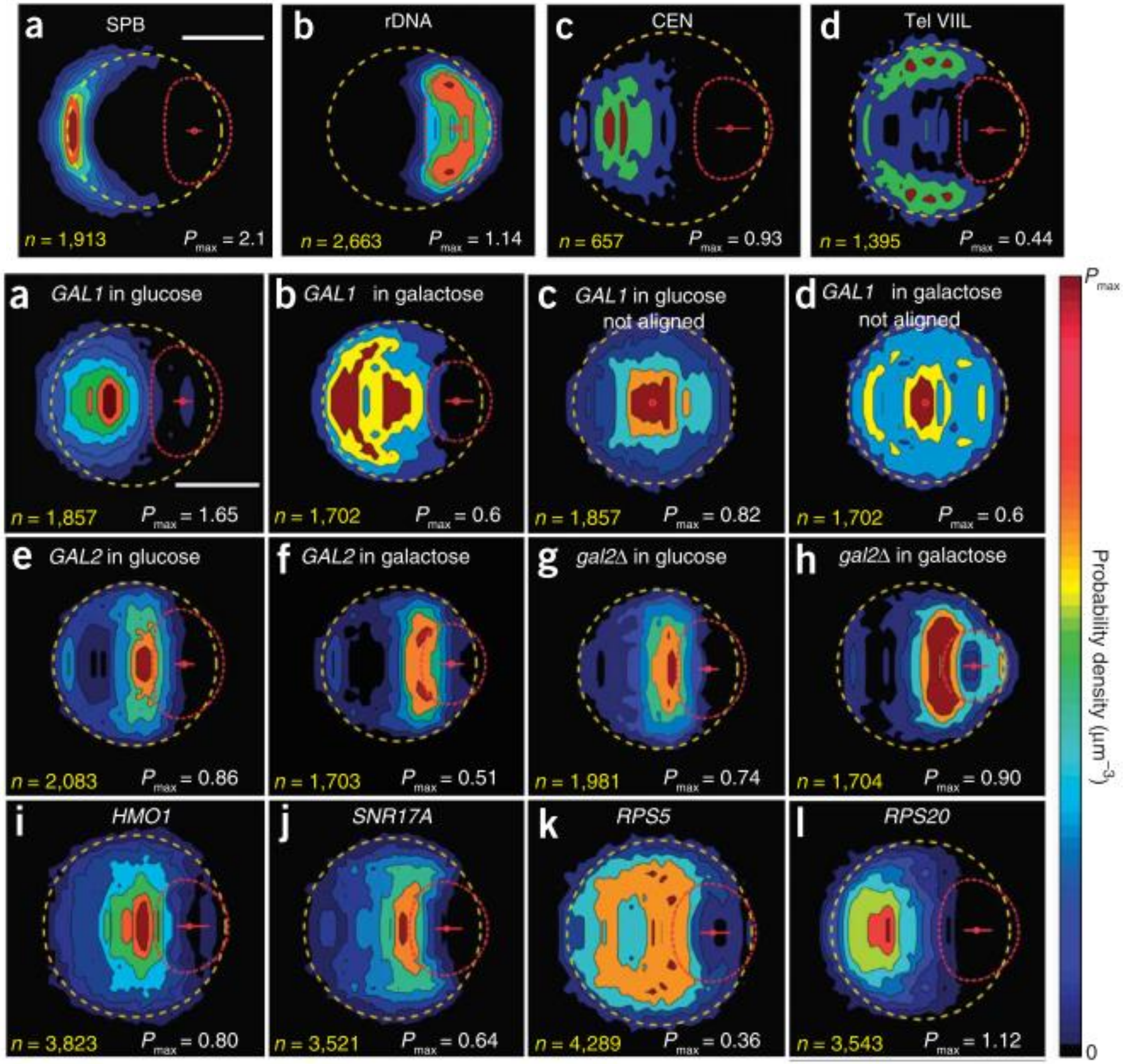
Katainen *et al*. Nature Genetics (2015)

Cooks syndrome



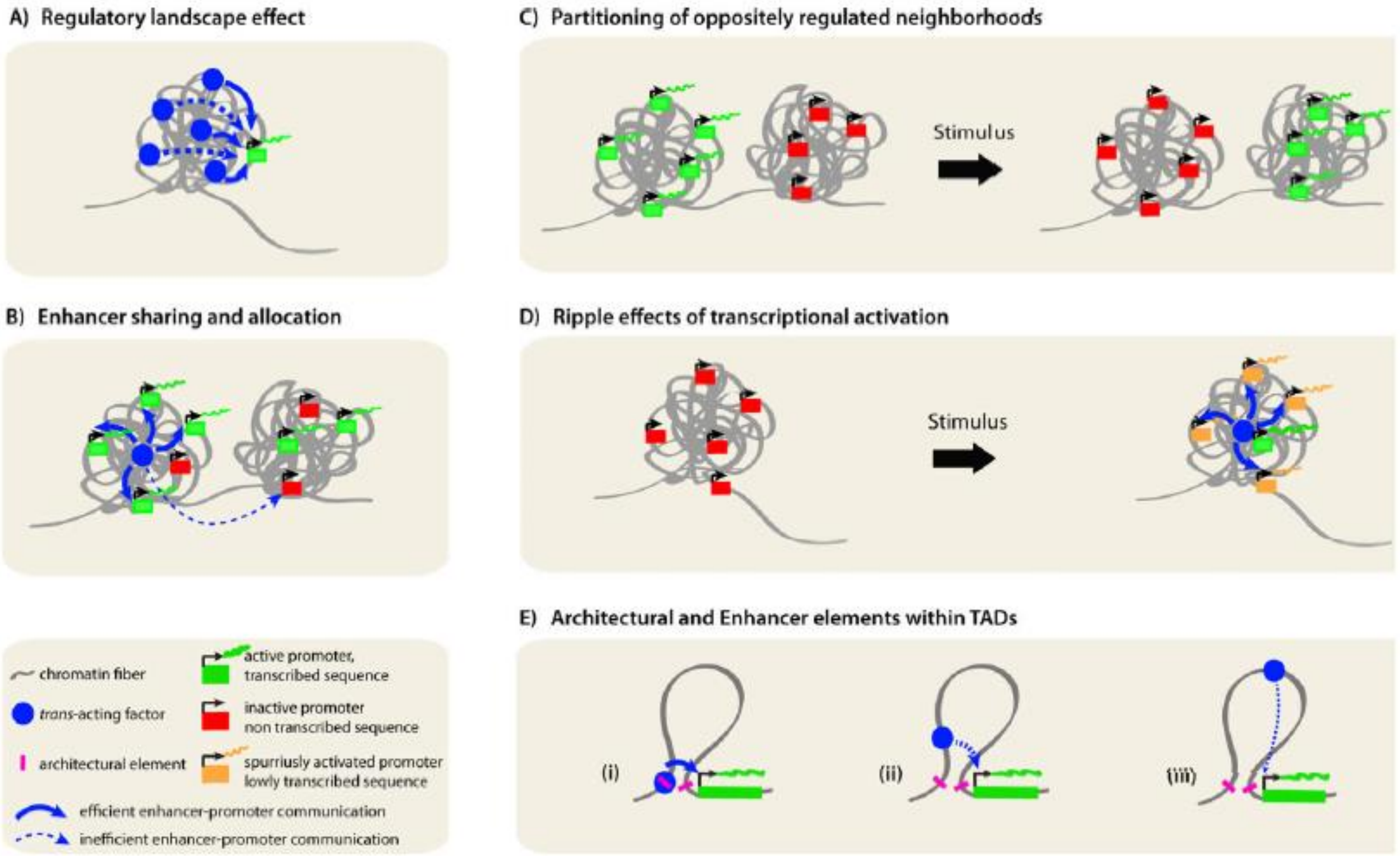
Franke *et al*. Nature 538, 265-269 (2016)

Localization of genes in nucleus



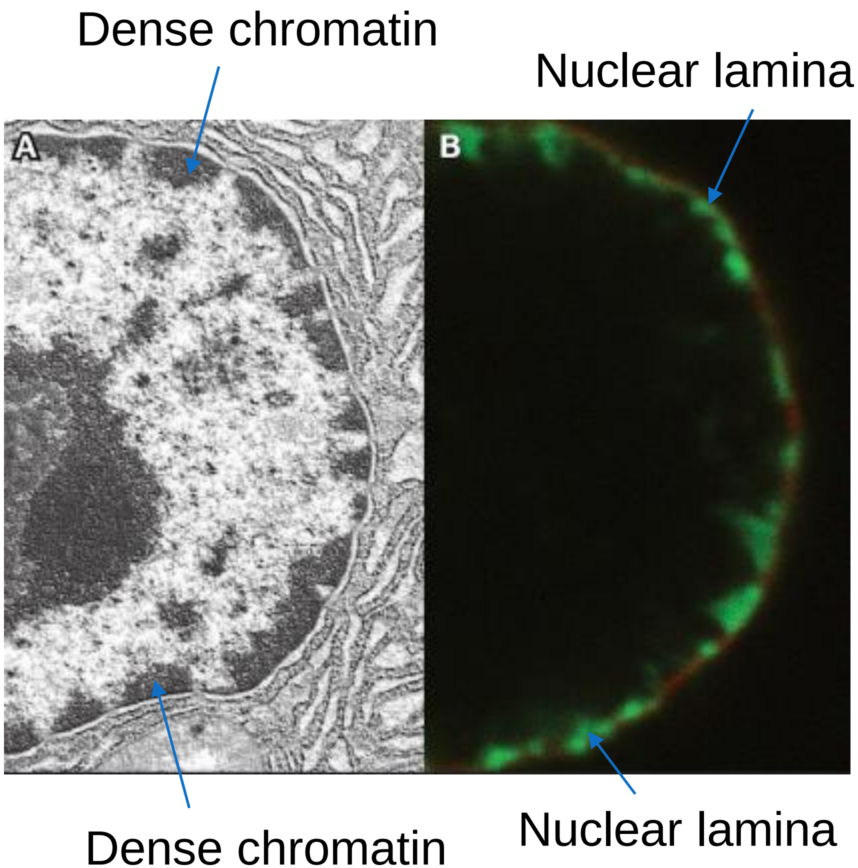
Source: Berger *et al*. Nature Methods (2008)

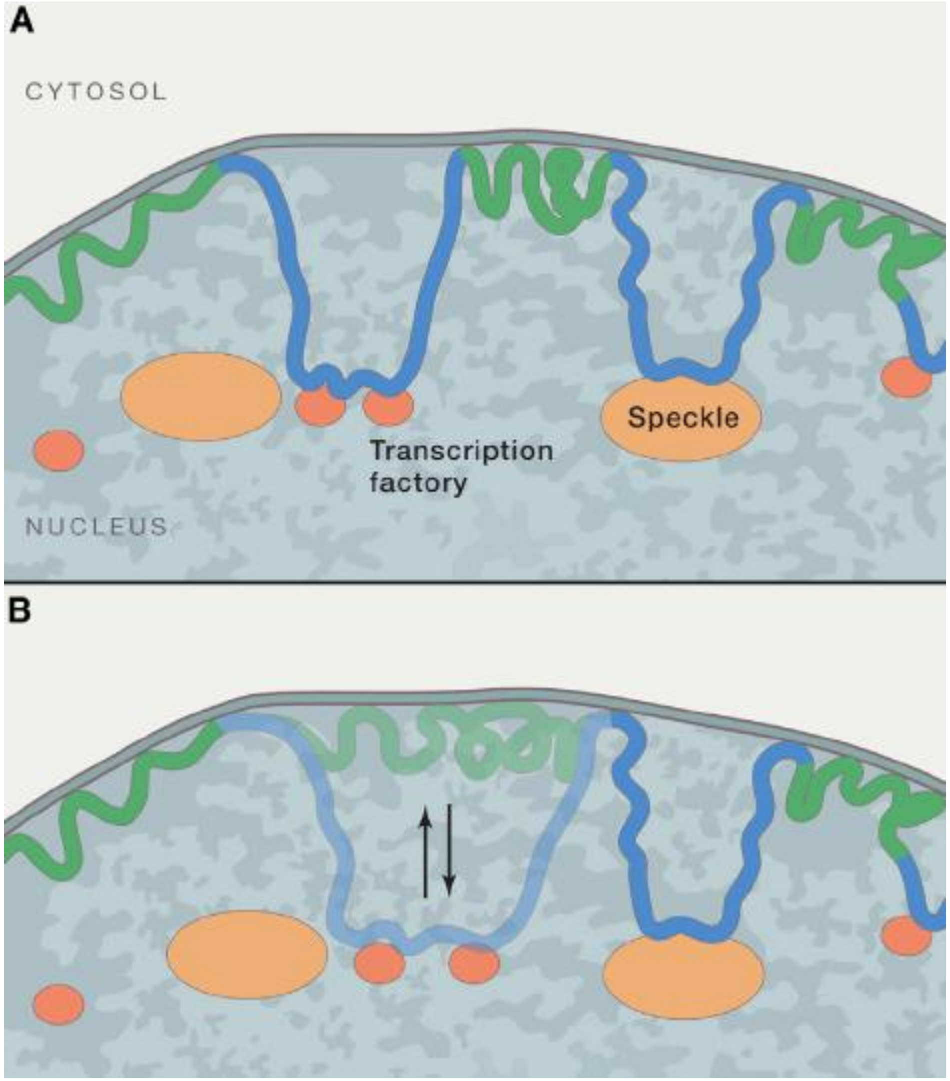
Efficient gene cluster regulation



Nora *et al*. Bioessay (2013)

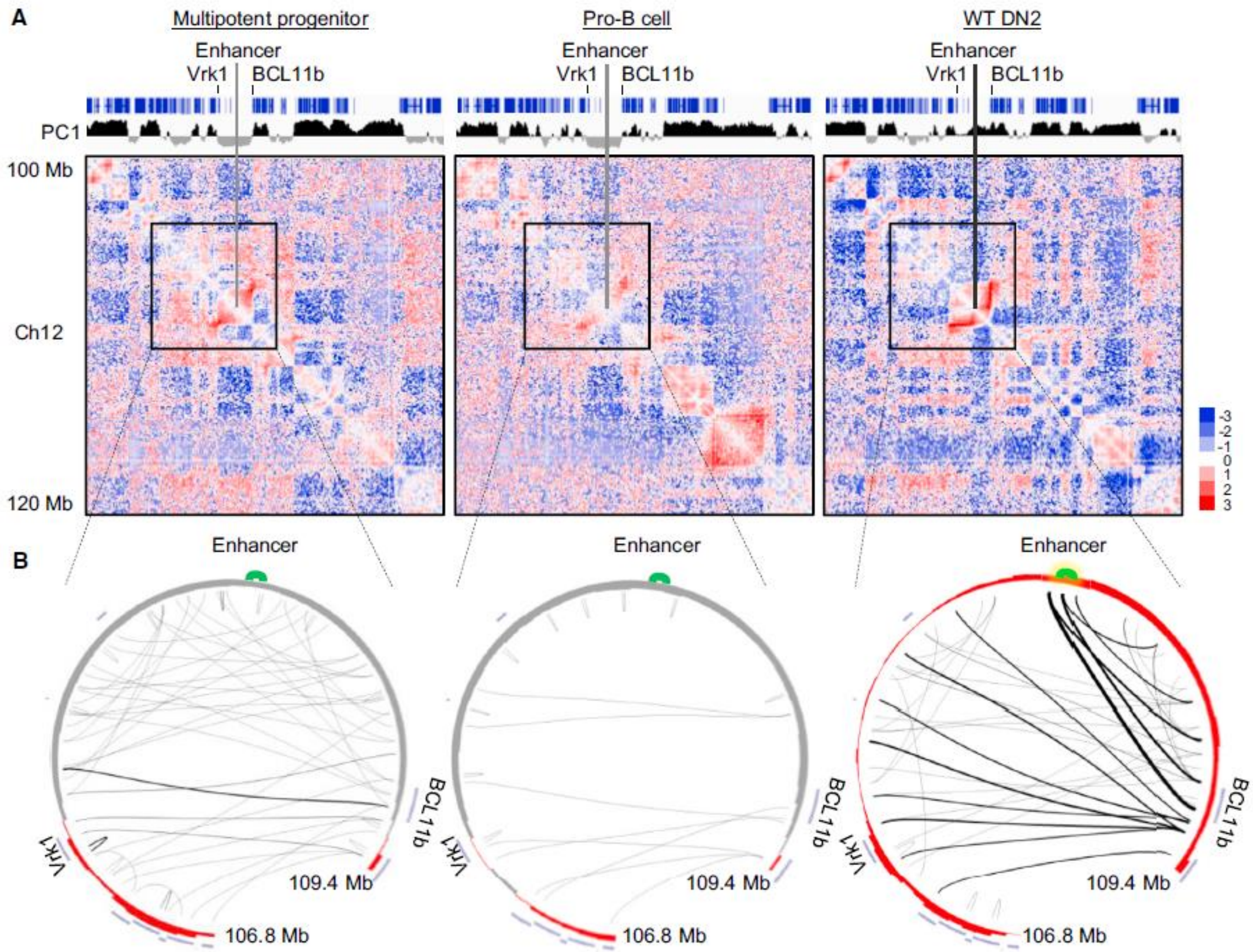
Heterochromatin associates with lamina





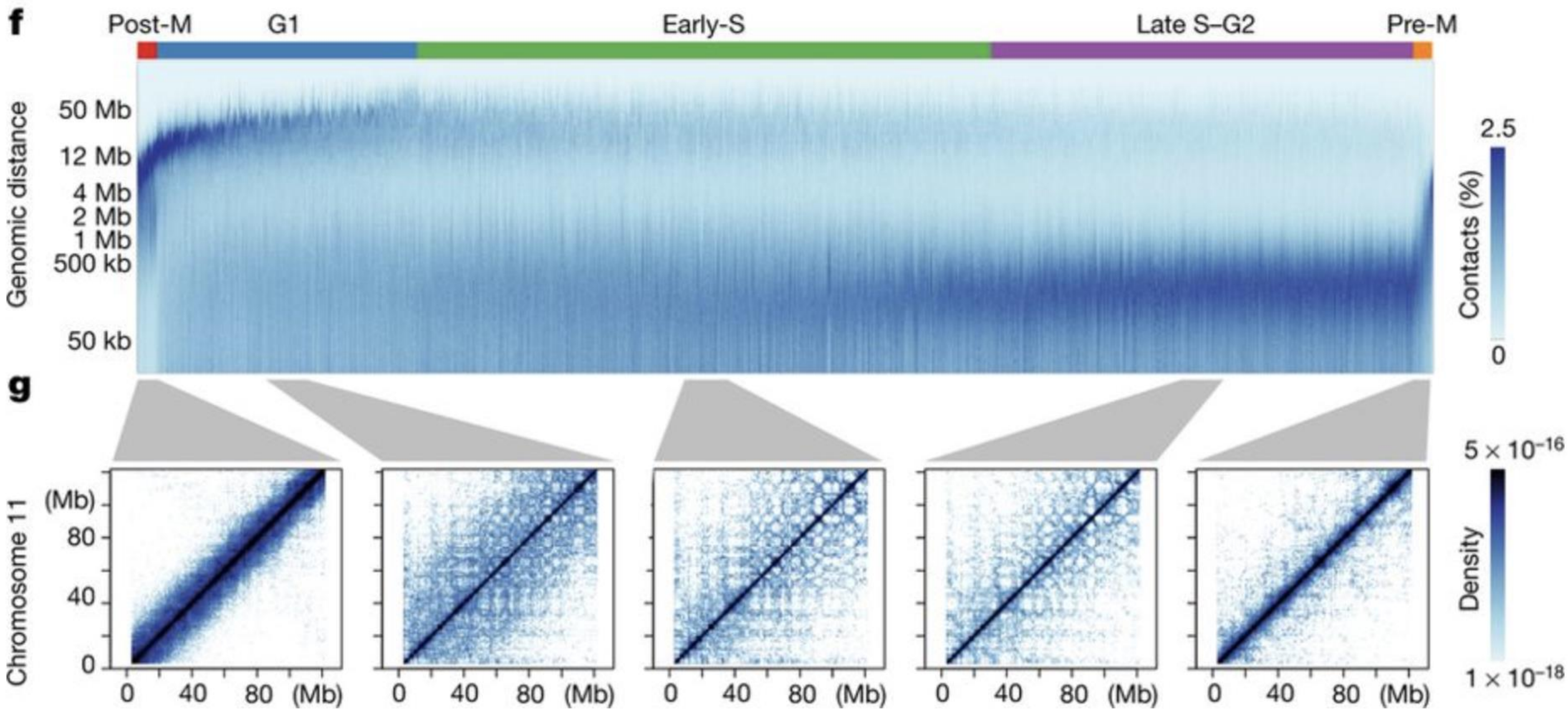
Steensel and Belmont. Cell (2017)

Chromatin restructuring during development



Isoda *et al*. Cell (2017)

Chromatin restructuring during cell cycle

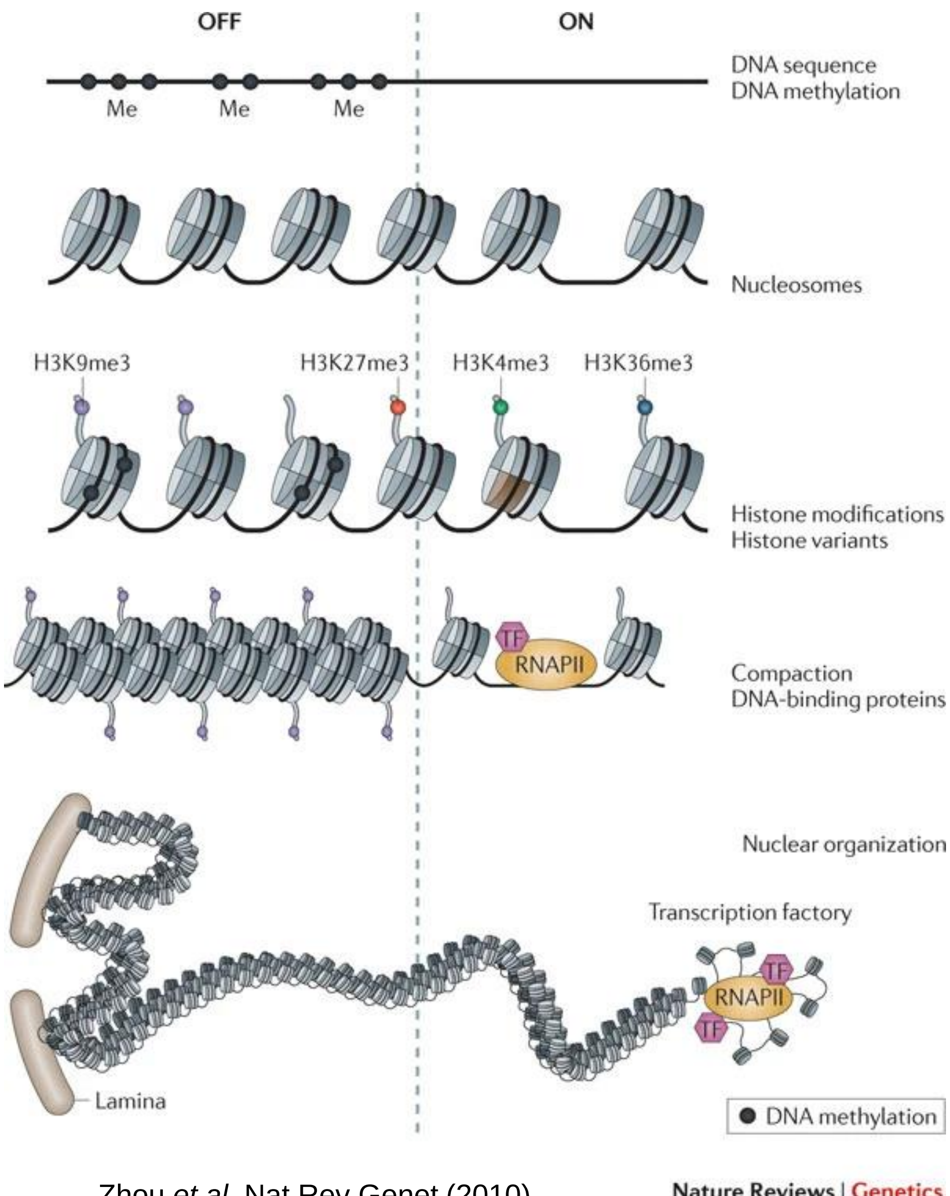


Nagano *et al*. Nature (2017)

- Un-packing for duplication and re-packing when done

Technology for querying chromatin organization

Layers of chromatin-based regulation



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

- DNA methylation
  - Bisulfite sequencing
- Accessibility
  - ATAC-seq
  - DNase-seq
- Histone mark and protein occupation
  - ChIP-seq
- 3D structure and interactions
  - Hi-C
  - ChIA-PET

Bisulfite sequencing

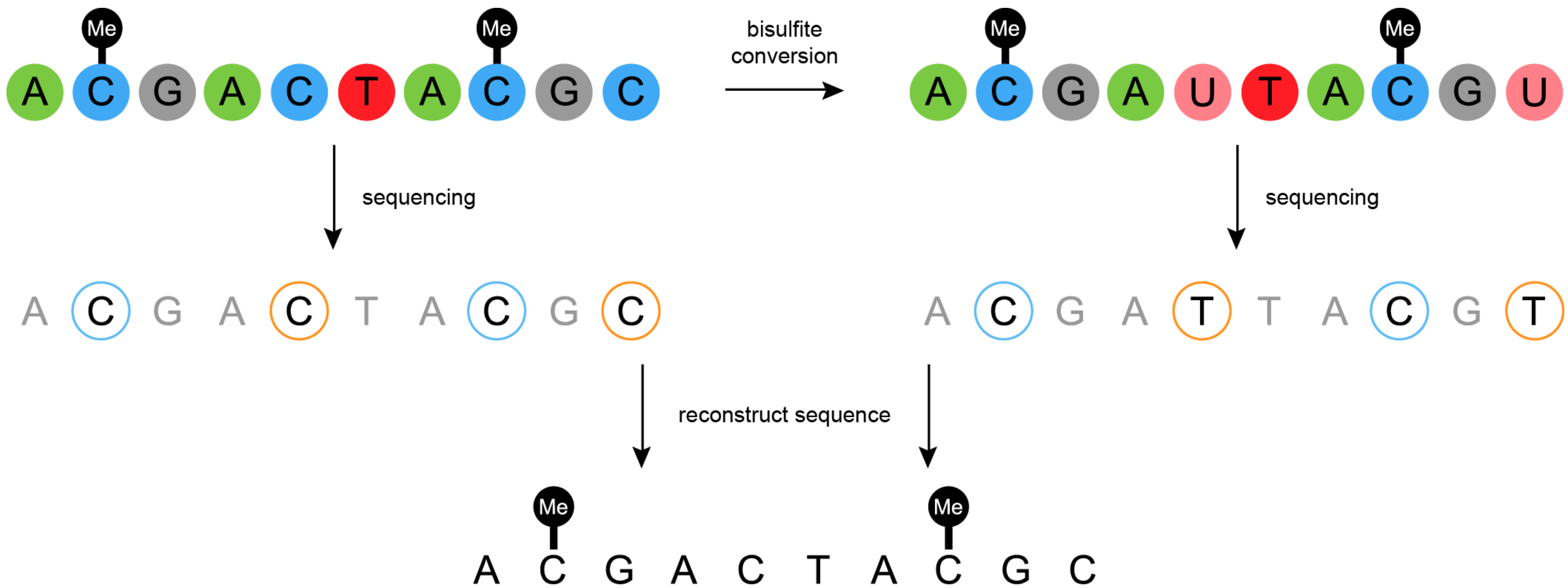


Image from http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis

- Detect methylated cytosine (Cm)
- Methylation represses transcription

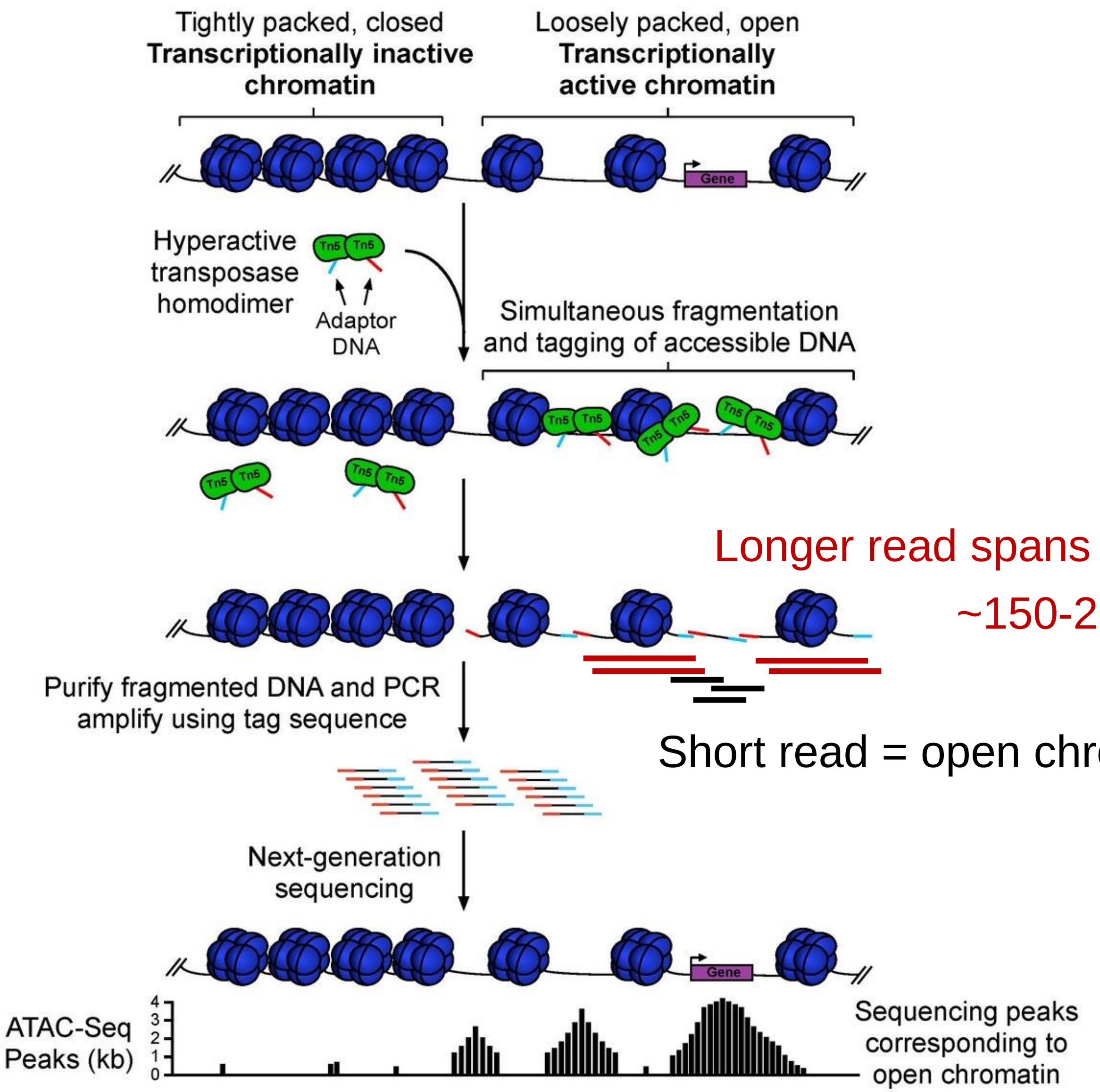
Alignment of bisulfite converted reads

- **Key assumption**: Most C's in genome are unmethylated
  - Most C's in bisulfite converted reads will turn into T's

- **Strategy 1**: Convert C's in reference genome to T's
  - Small number of C's in bisulfite converted reads → mismatches
  - Use traditional aligner such as "bowtie"
  - <u>Example</u>: Bismark

- **Strategy 2**: Adjust traditional sequence alignment's scoring criteria for C-T mismatch
  - P(random C-T mismatch) ~ P(C in genome) x P(T in reads)
  - P(C-T mismatch due to bisulfite) ~ P(C in genome) x P(non-methylated)
  - <u>Example</u>: Last

ATAC-seq



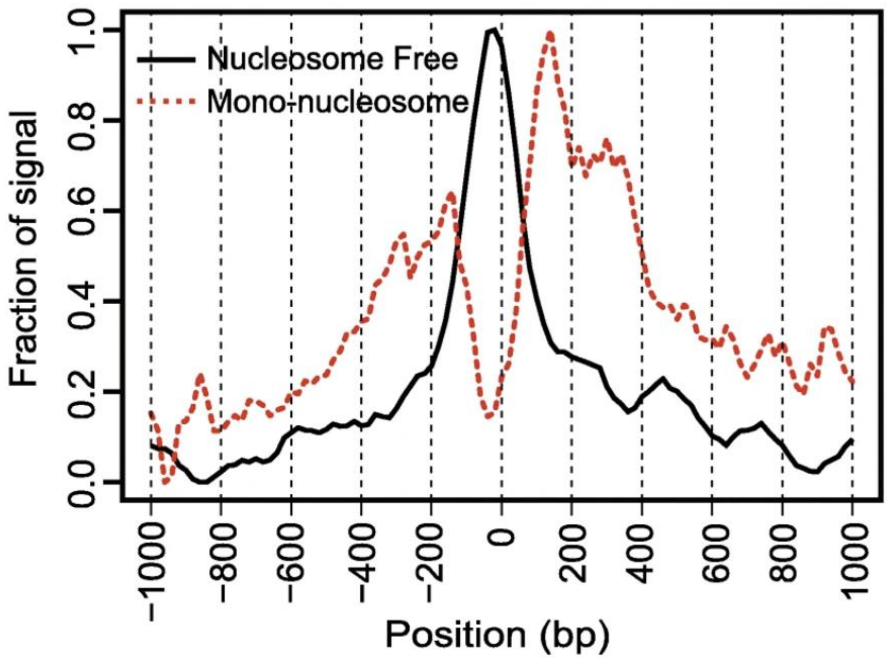
Longer read spans nucleosome

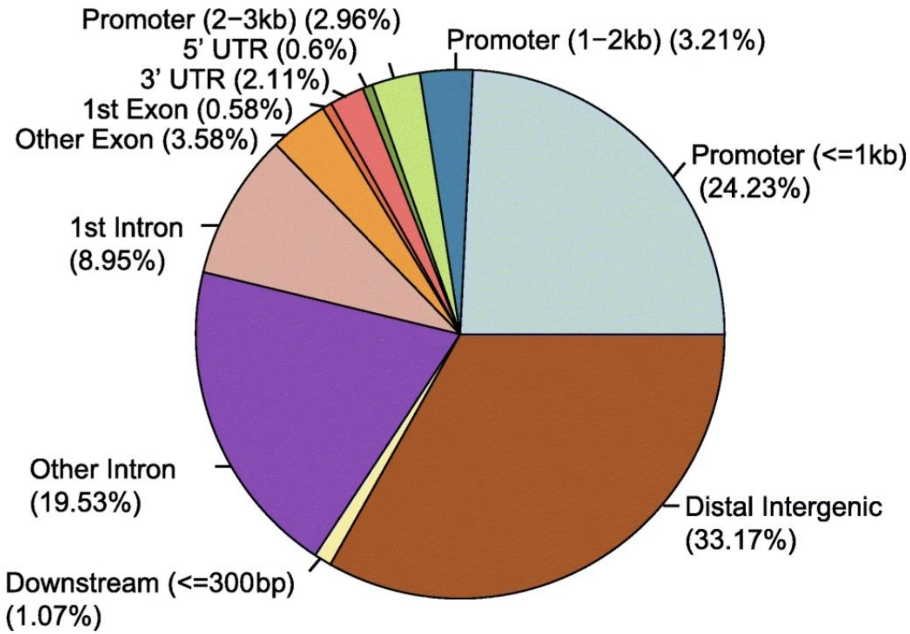
~150-200 bp

Short read = open chromatin

Image from wikipedia

Additional QC for location-based NGS





Yan *et al*. Genome Biology (2020)

- Short reads (nucleosome-free) are usually enriched around transcription start sites (TSS) and within 1kb of promoters
- Longer reads should be enriched near TSS

Chromatin immunoprecipitation

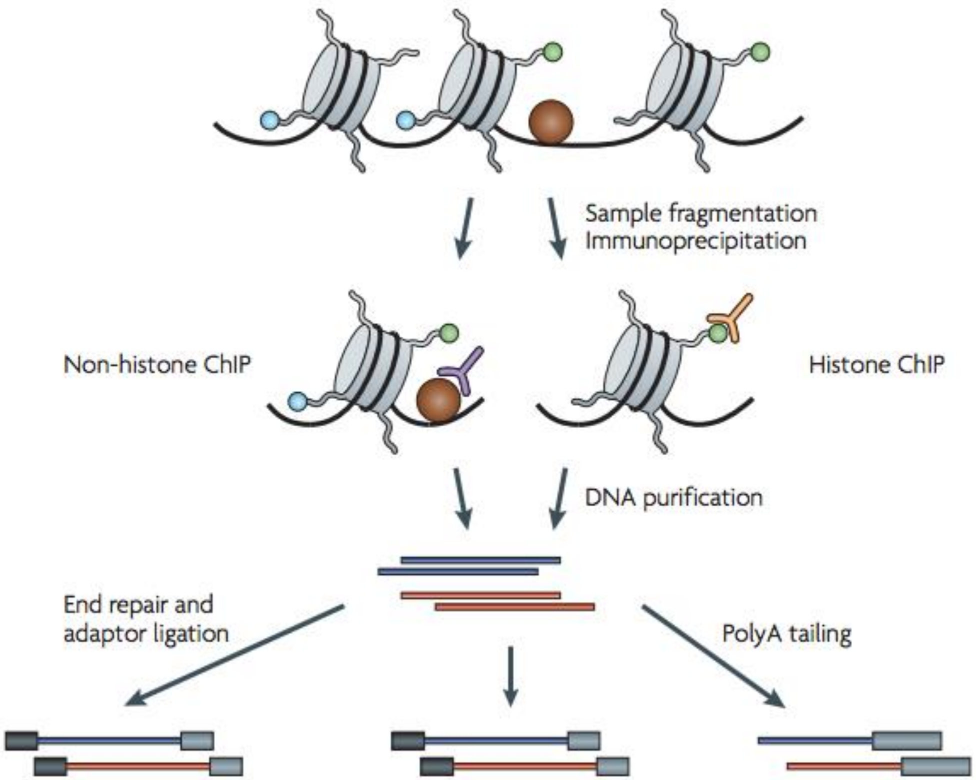
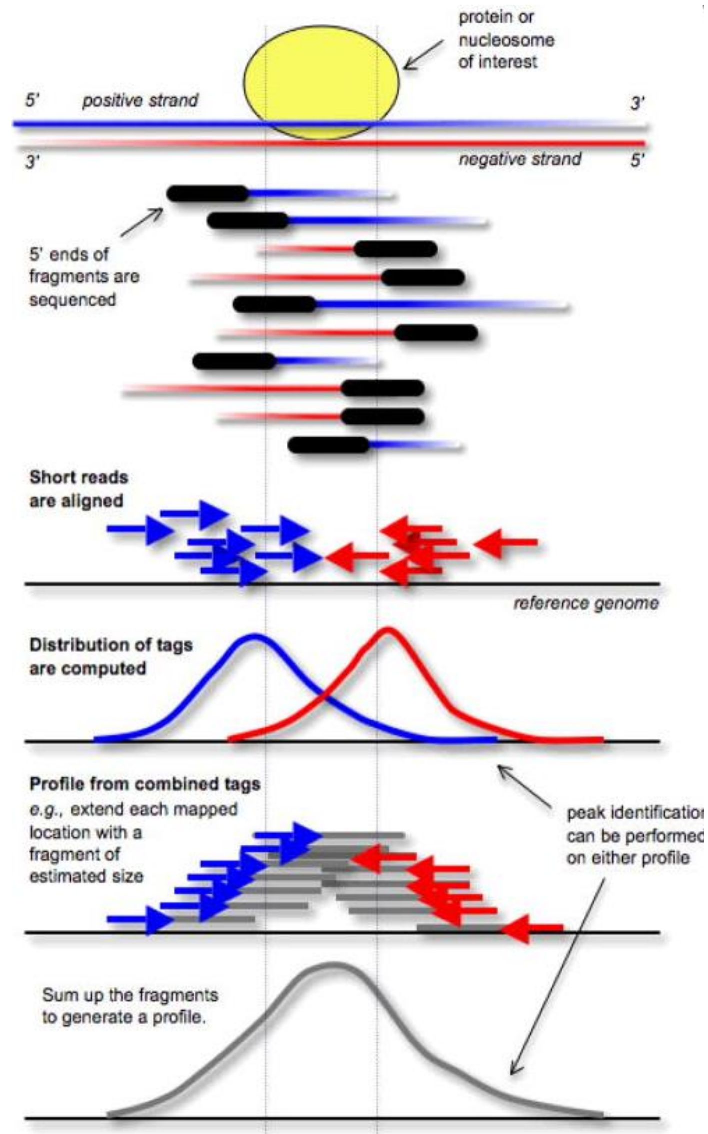


Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

- Select protein-bound or histone-bound DNA for sequencing



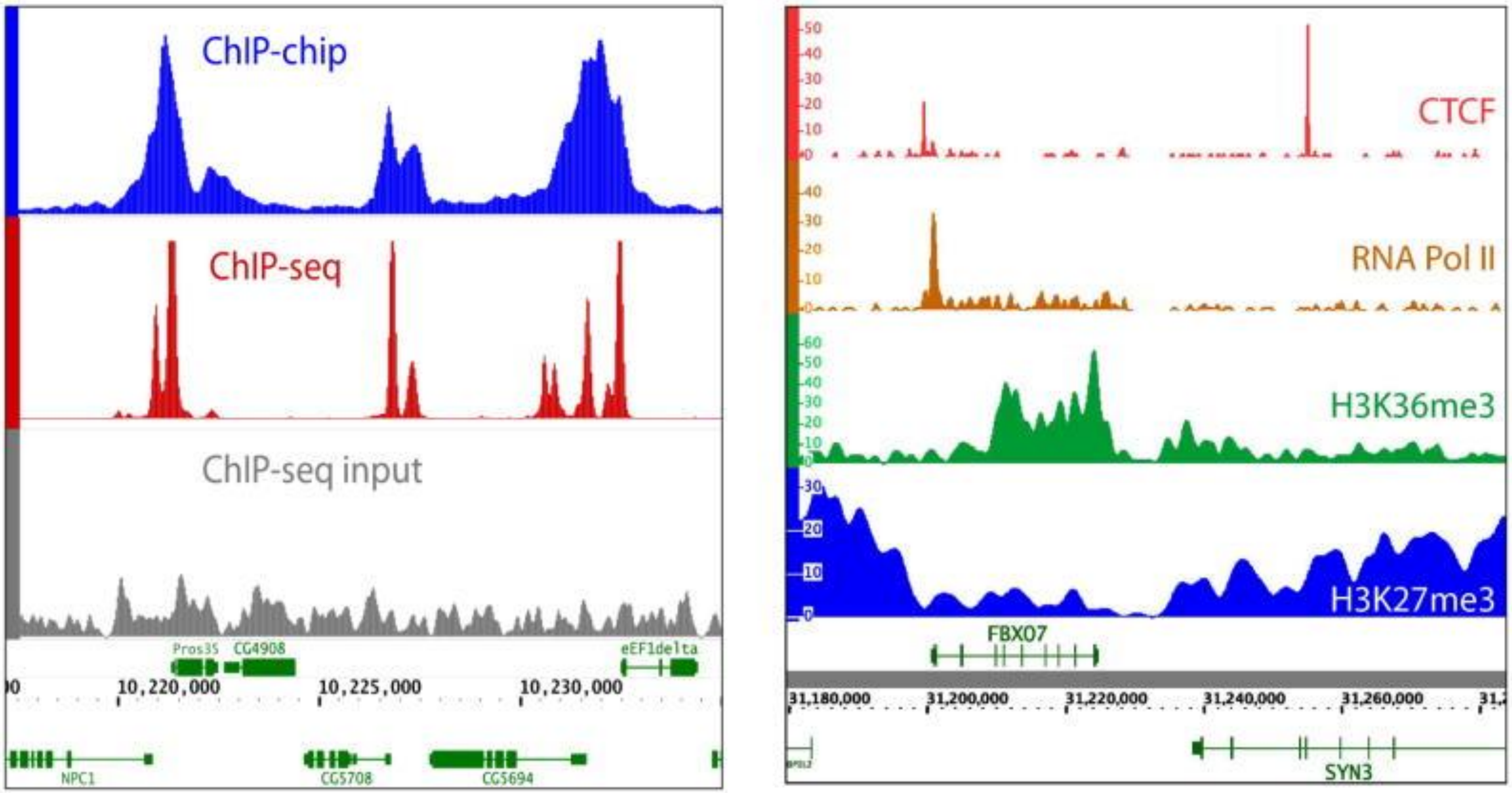
Park *et al.* Nat Rev Genet 10:669-680 (2009)

Layers of gene expression regulation

- RNA-seq shows changes in expression but not the mechanisms
- Chromatin analysis provides clues into the mechanisms
- ChIP-seq shows if changes in expression coincide with local TF binding & histone modifications
- ATAC-seq shows if changes in expression coincide with opening & closing of local chromatin

Analysis of chromatin binding data

ChIP-seq data



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

- Input = non-IP sample
- Protein (e.g., CTCF, RNA Pol II) binds to discrete sites
- Histone mark spans broad areas

Peak calling

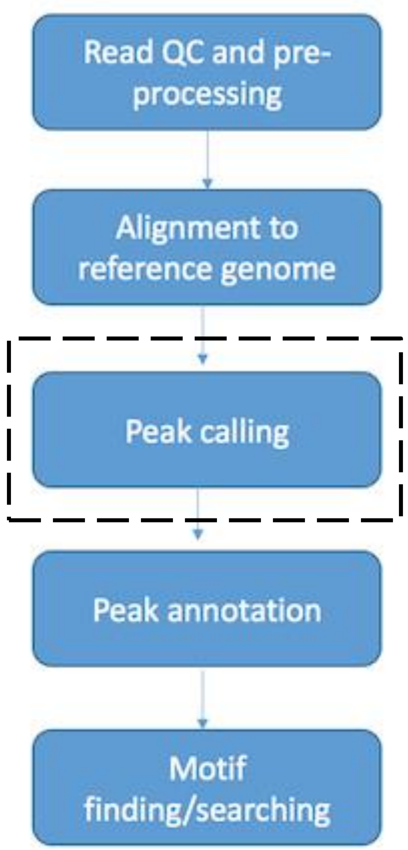
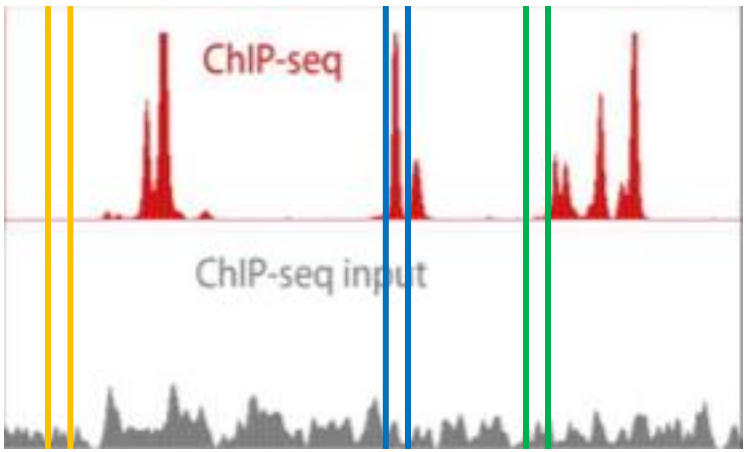


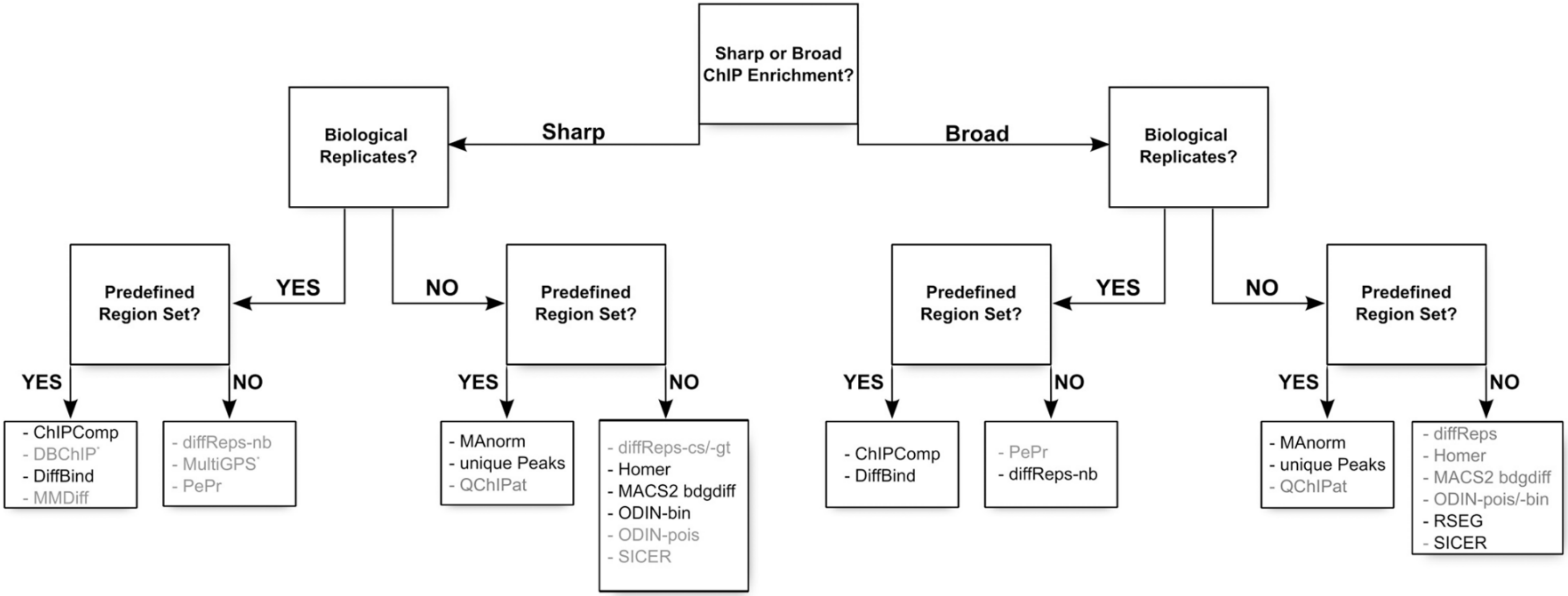
Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html



Genomic Position

- Number of reads in each window ~ Poisson distribution
- **Null hypothesis**: estimated λ from non-IP "input" sample
- λ varies across genomic regions
- P-value = $\frac{\lambda^n e^{-\lambda}}{n!}$, n = observed reads, λ estimated from the same window

ChIP-seq analysis pipelines



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

- Different algorithms for analysis of broad (histone) and sharp (protein) signals
- Predefined region set = called peaks

Peak annotation

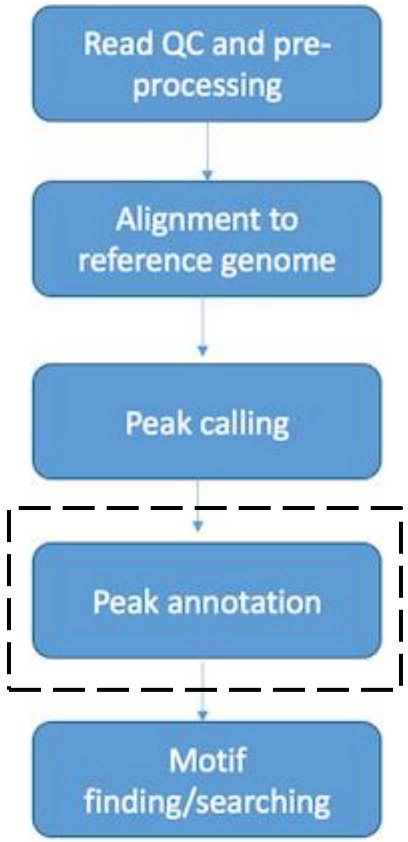


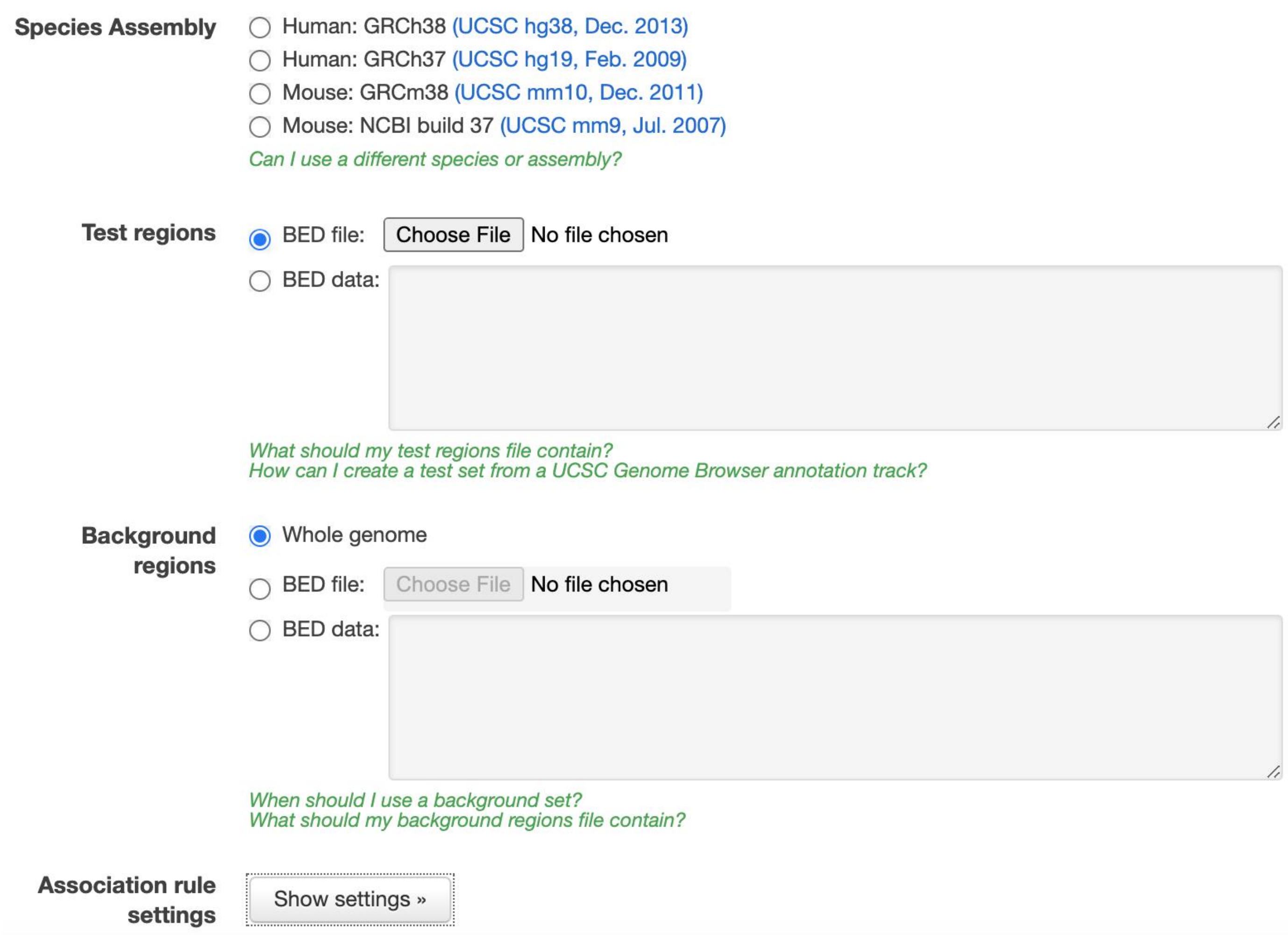
Image from http://informatics.fas.harvard.edu/chip-seq-workshop.html

- A ChIP-seq peak may be near multiple genes and/or genomic features
- Which ones of them are affected by this protein binding or histone mark?
  - Nearest gene?
  - All genes within 10,000 bp?
  - All genes within 5 nm in 3D structure?
- Different assumptions for TF, histone mark, enhancer, etc.

Functional enrichment for ChIP-seq

GREAT improves functional interpretation of *cis*-regulatory regions

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



Assign genomic region to gene(s)

Associating genomic regions with genes

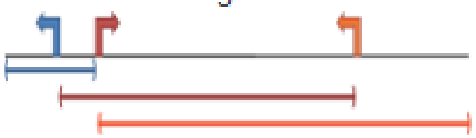
GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two step process. First, every gene is assigned a regulatory domain. Then, each genomic region is associated with all genes whose regulatory domain it overlaps.

◉ 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.

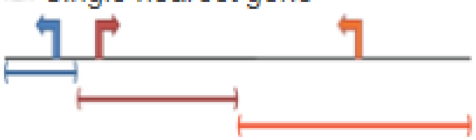
○ Two nearest genes



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



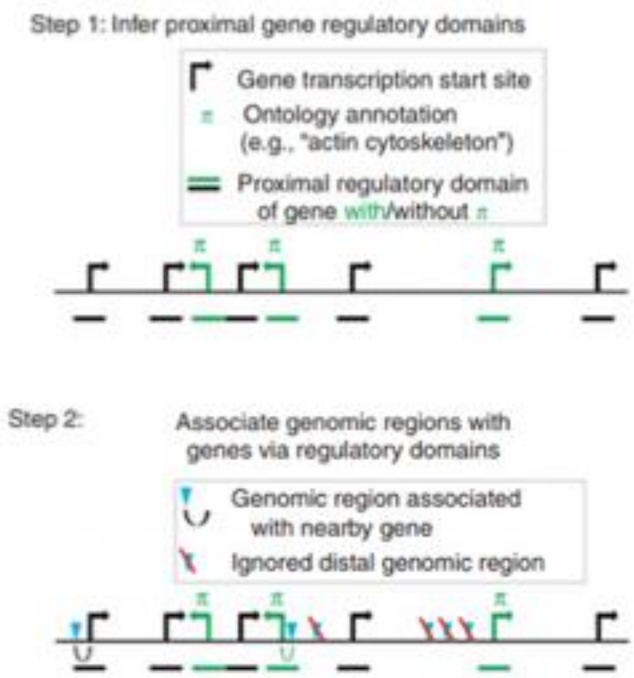
within 1000.0 kb

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)

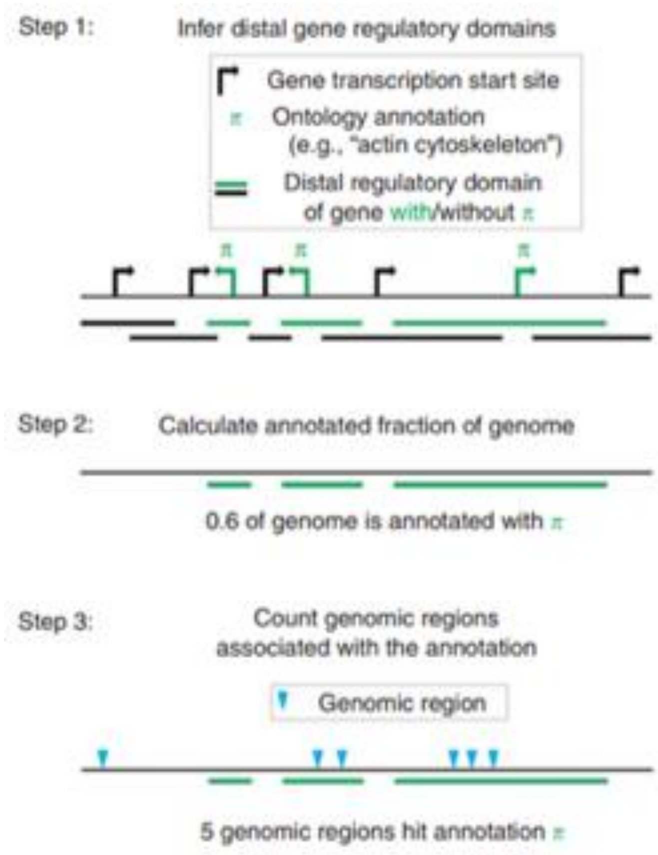
☑ Include curated regulatory domains *What are curated regulatory domains?*

Significant testing for gene-based model



- Assign ChIP-seq peaks in each region to nearby genes
- Count number of genes that are associated with specific function and also located near ChIP-seq peaks
- Calculate p-value using hypergeometric distribution model

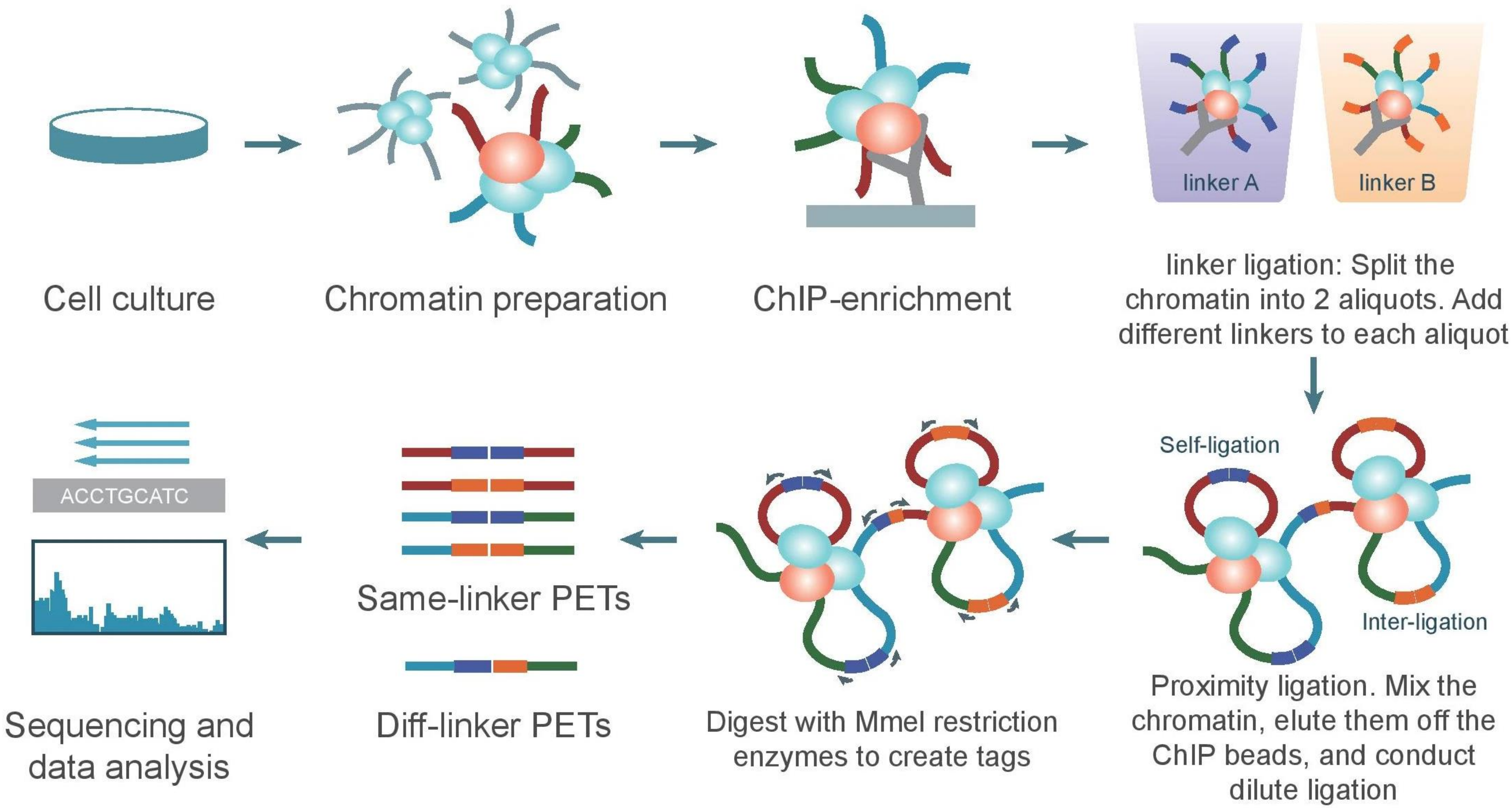
Significant testing for area-based model



- Assign each genomic region to nearby genes
- Calculate fraction *p* of genome that is associated with specific function
- Count the number of ChIP-seq peaks that fall into this region
- Calculate p-value using binomial distribution model

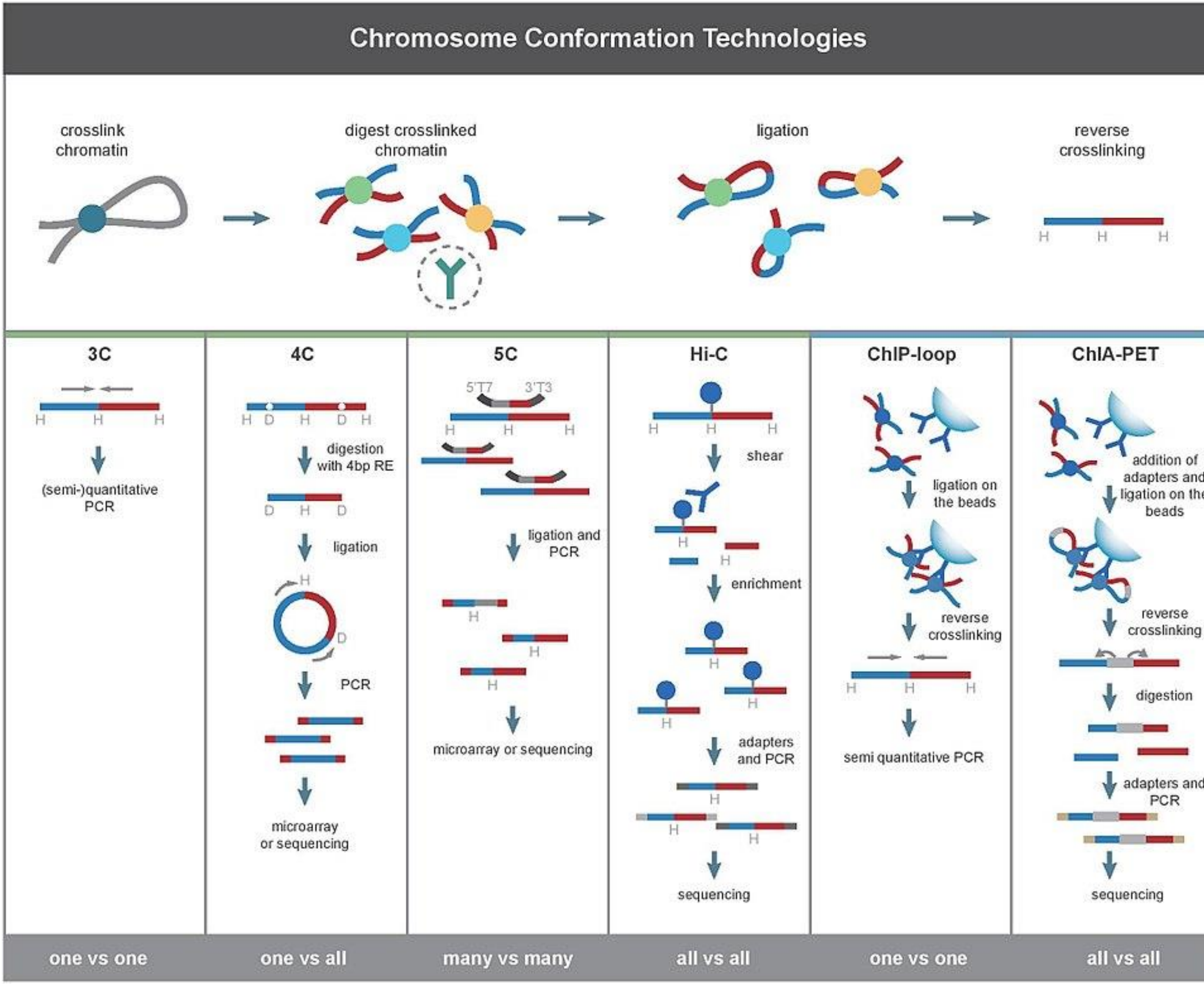
Long-range interactions

From ChIP-seq to ChIA-PET



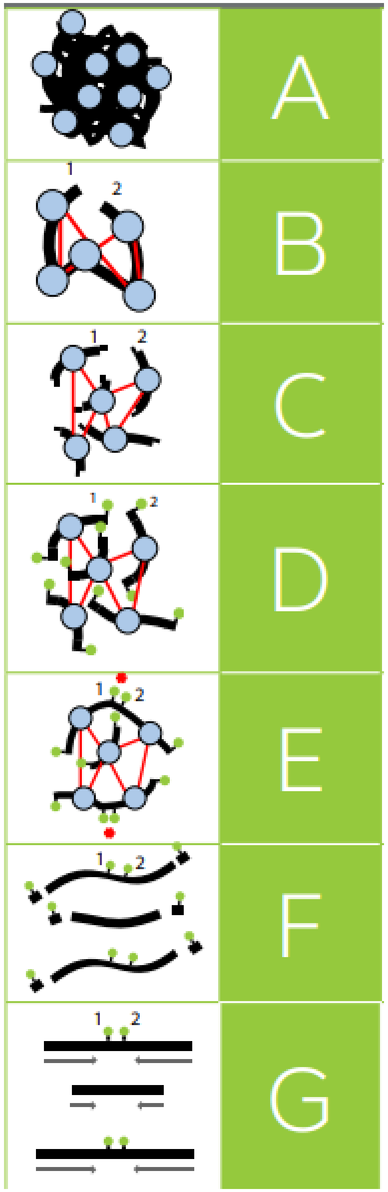
Li *et al.* BMC Genomics 15:S11 (2014)

- ChIP-seq tells us LOCAL signals but not DISTAL effects
- Enhancer binds one location but operate on far-away genes



Source: wikipedia

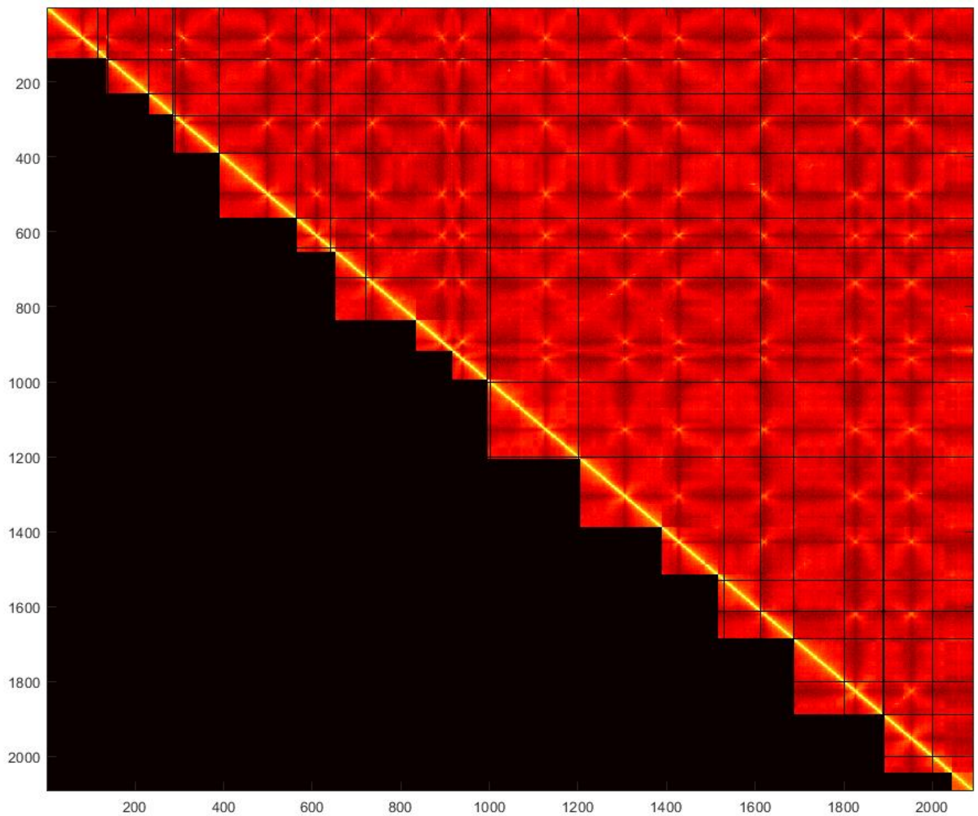
Hi-C protocol



- Endogenous chromatins
- Crosslinking of proximal chromatins
- Restriction endonuclease digestion
- Biotinylation
- Proximity ligation
- Crosslinking reversal
- Paired-end sequencing

Source: Dovetail Genomics Hi-C pamphlet

Hi-C resolution



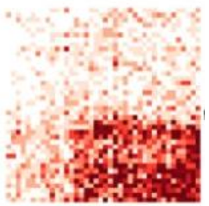
Article | Open Access | Published: 21 February 2018

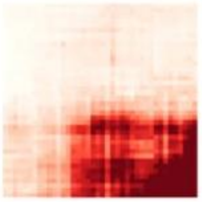
Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus

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



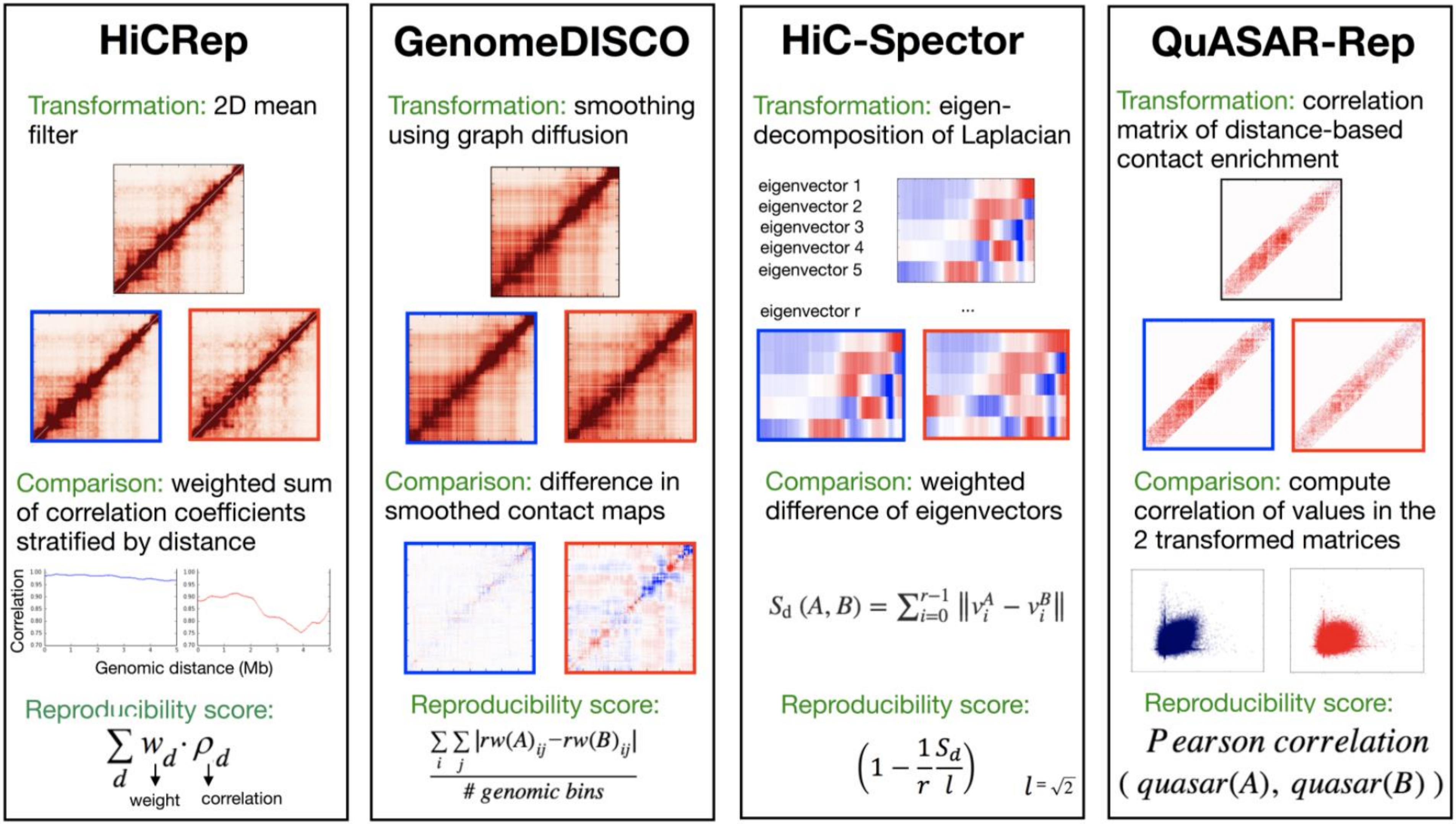


Zhang *et al.*, Nat Comm (2018)

- High-resolution = small bp per bin
  - Requires high read count → otherwise there won't be enough read to estimate contact frequency
  - Typically use step size of 10kb → [10kb, 20kb, …, 100kb]

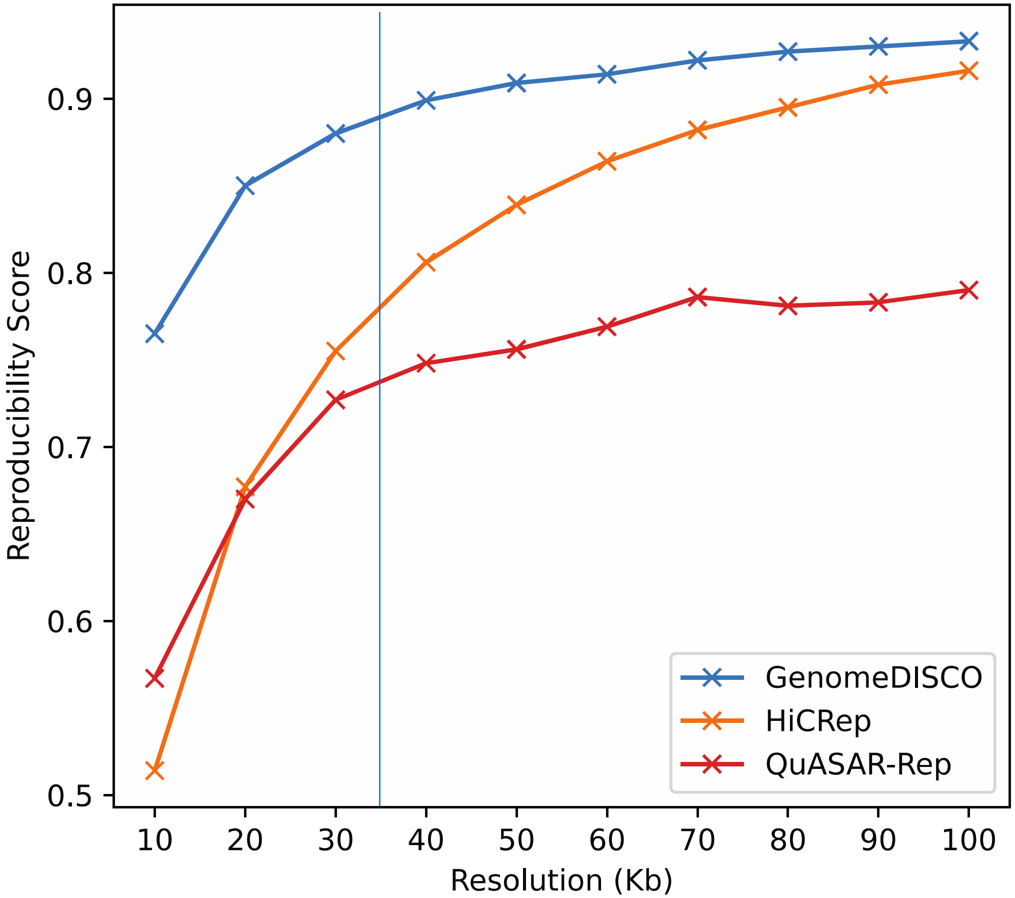
Quality assessment

Source: github.com/kundajelab/3DChromatin_ReplicateQC



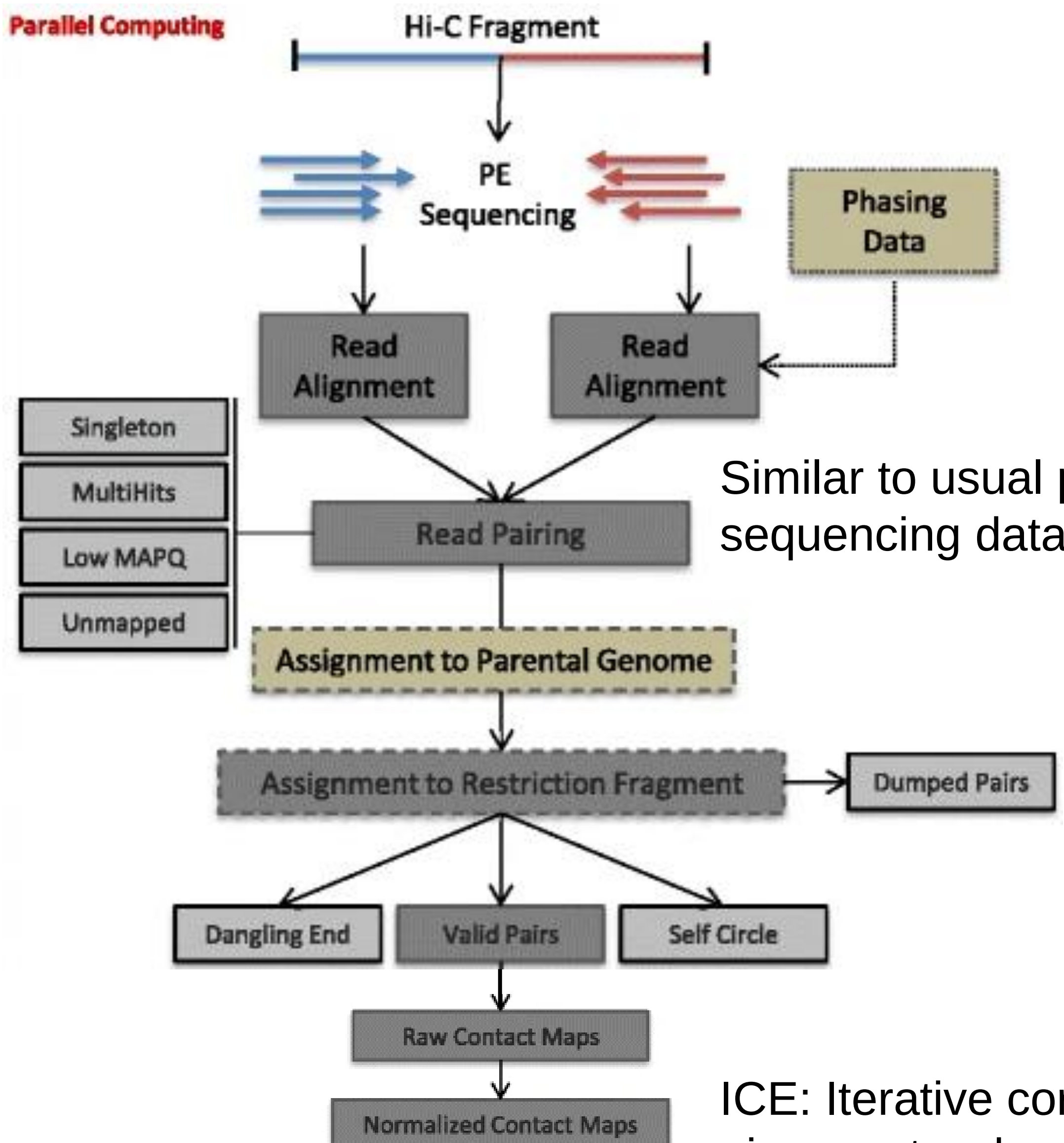
- Appropriate resolution = high reproducibility

Quality assessment



- Analyze Hi-C data at various resolution
- Then calculate reproducibility score
  - Look for resolution where the score plateau (~30-40 kb here)

Hi-C analysis

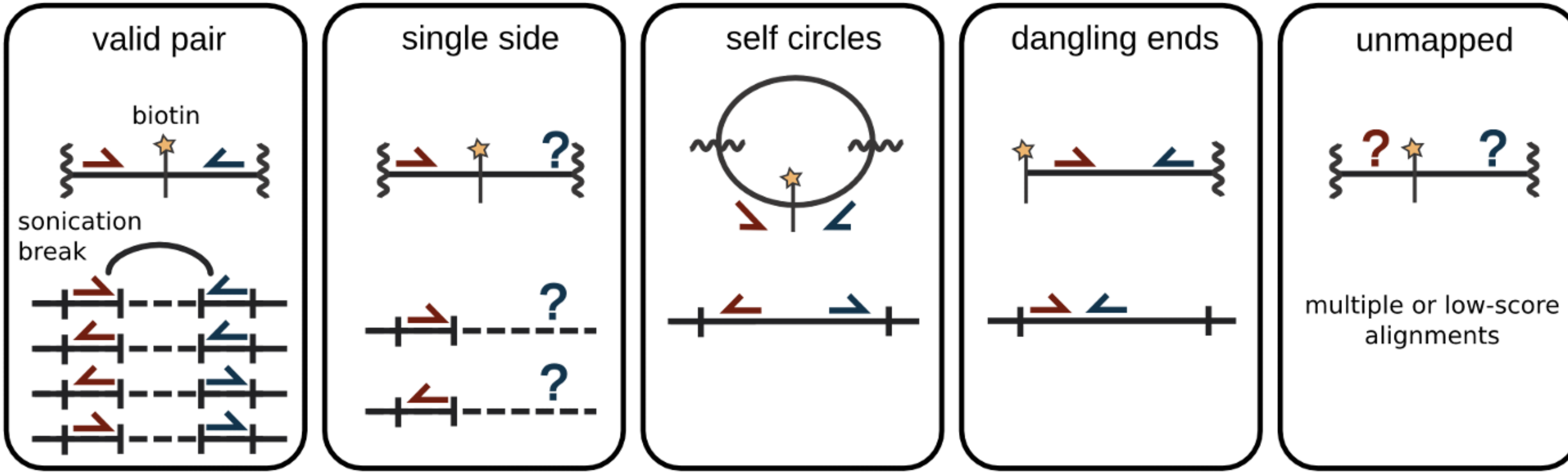


Similar to usual pair-ended sequencing data analysis

ICE: Iterative correction and eigenvector decomposition

Servant et al. Genome Biology (2015)

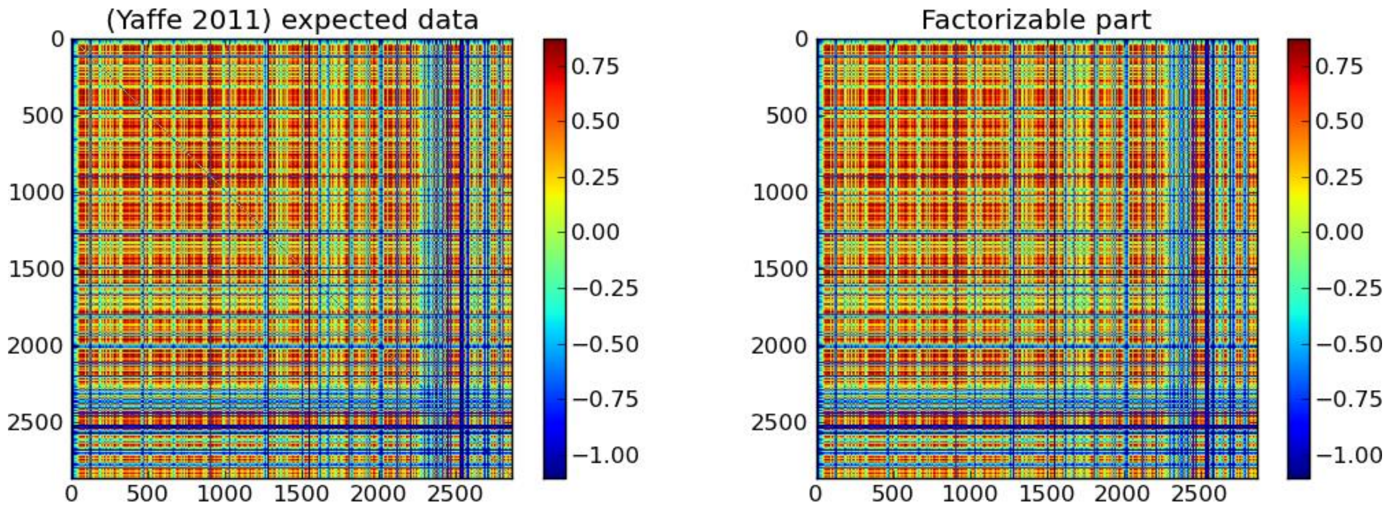
Hi-C analysis



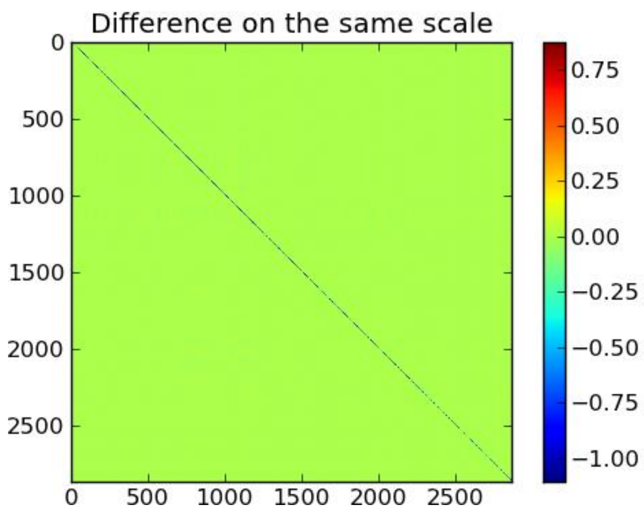
Imakaev et al. Nature Methods (2012)

- Biotin site = restriction enzyme motif
- Orientation of paired reads with respect to expected restriction enzyme digestion sites can tell us whether the reads come from valid pair or not

Bias factorization



Imakaev et al. Nature Methods (2012)



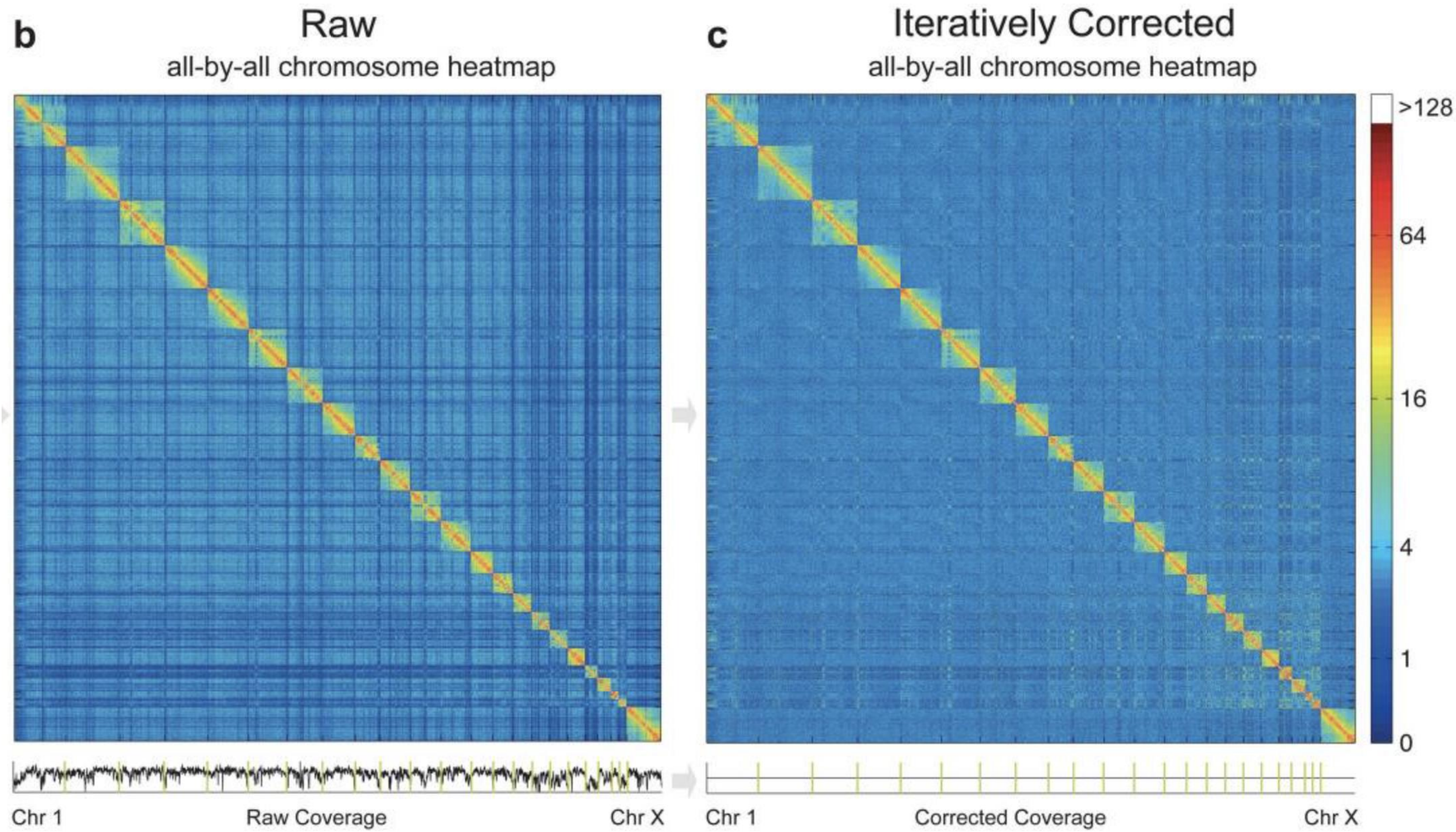
$$Obs_{i,j} = True_{i,j} \cdot Bias_{i,j}$$
$$= True_{i,j} \cdot Bias_i \cdot Bias_j$$

$$\sum_{i,j} True_{i,j} = 1$$

$$\sum_{i} Bias_i = 1$$

- Total bias (GC content, chromatin state, etc.) at interaction between loci A and B ~ product of biases at A and B

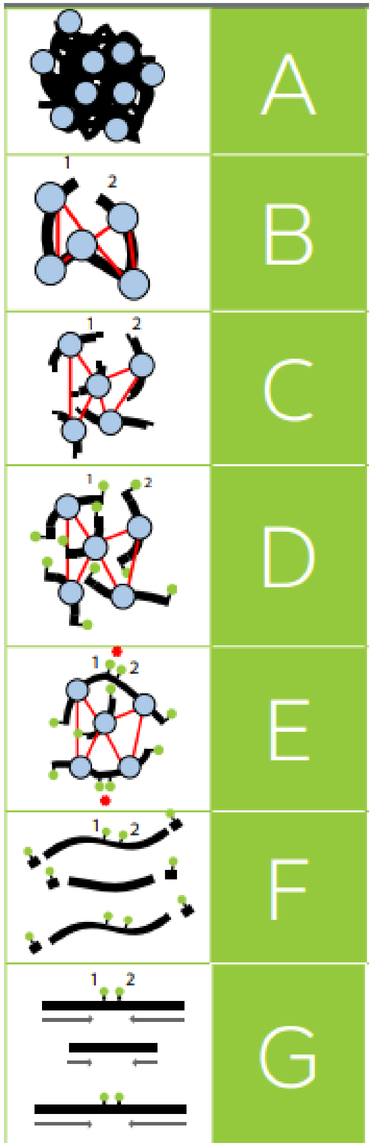
ICE correction for Hi-C data

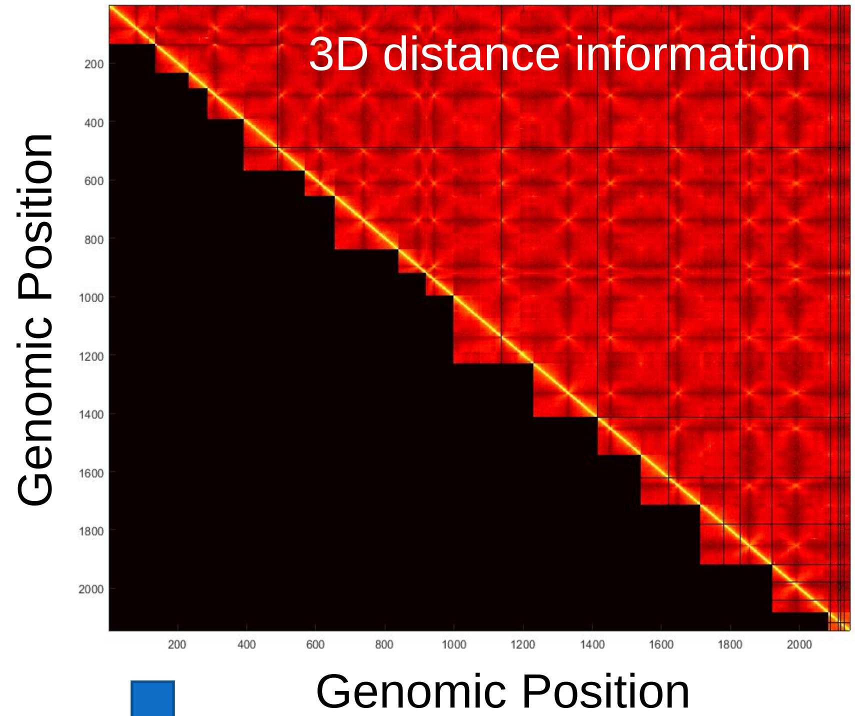


Imakaev et al. Nature Methods (2012)

Hi-C guided genome assembly

Hi-C data guides genome scaffolding

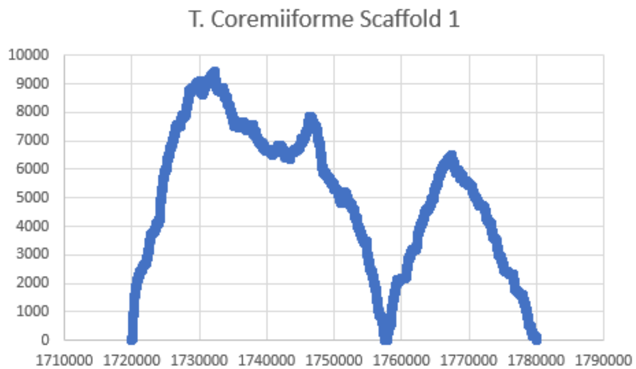


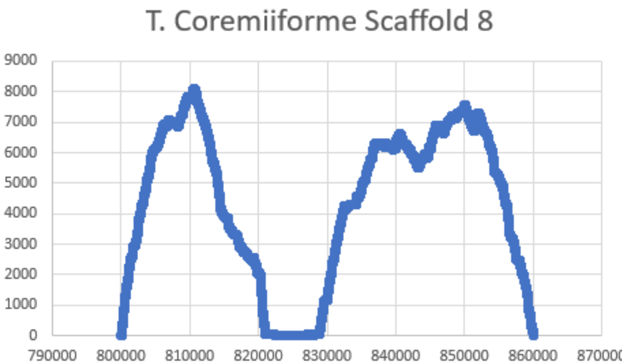


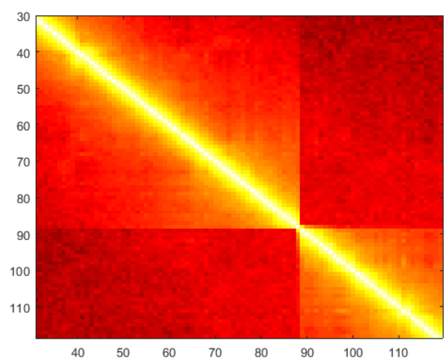
Refined Scaffolding

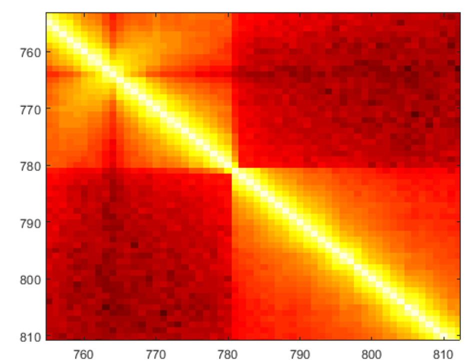
Species	Original Scaffold	New Scaffold	% of total bp
T. asahii	36	13	96.02
T. faecale	29	9	99.69
T. coremiiforme	190	16	98.33
T. inkin	18	8	99.35
T. ovoides	89	13	99.17

Detection of mis-scaffolding









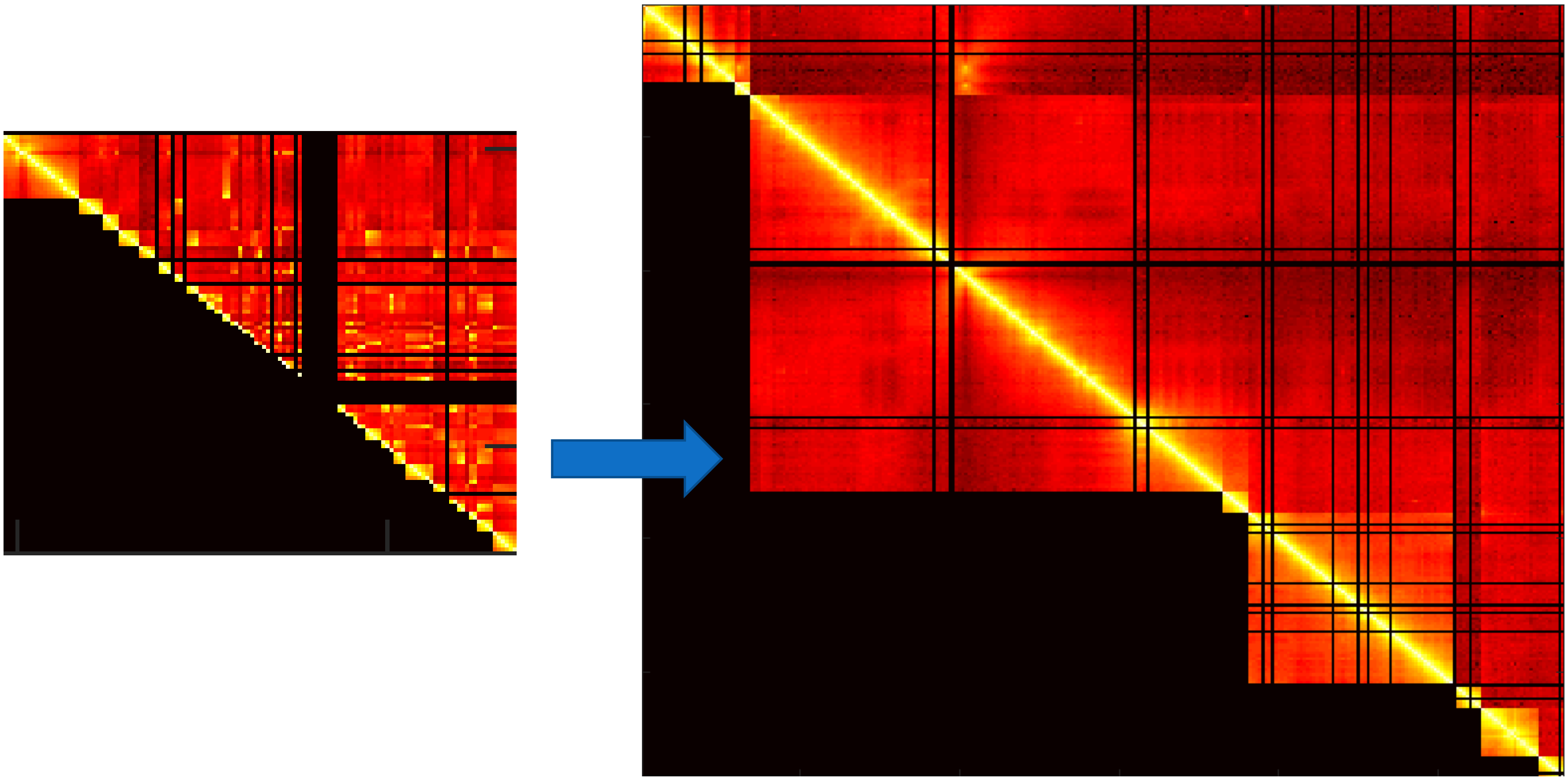
- Genome scaffolding for unknown organism relies on paired-end read information and merging of overlapping reads
- Hi-C data provide definite evidence because adjacent loci must be very close by in 3D

Hi-C assisted scaffolding

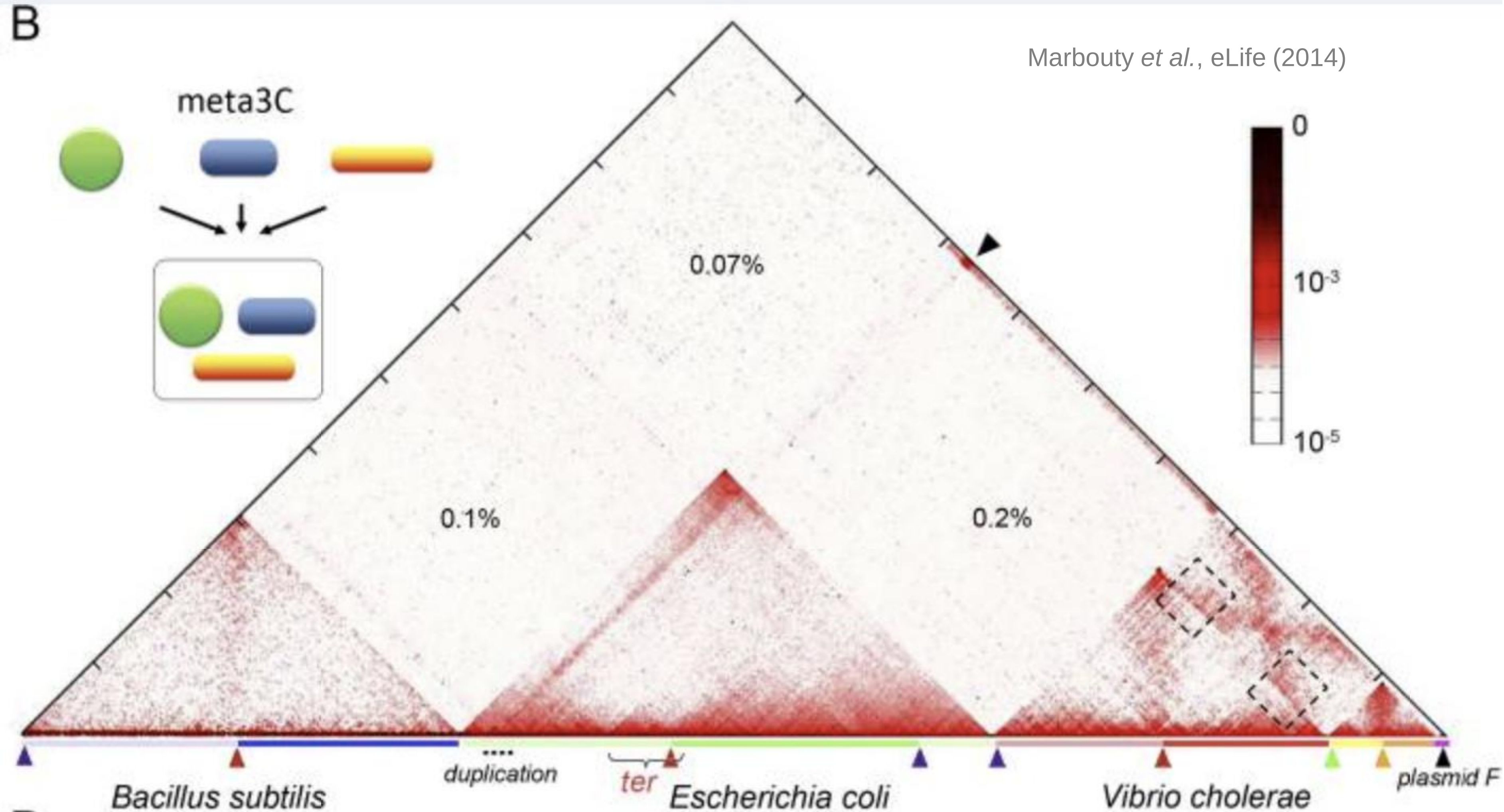
Scaffolding of long read assemblies using long range contact information

Jay Ghurye, Mihai Pop, Sergey Koren, Derek Bickhart & Chen-Shan Chin

BMC Genomics **18**, Article number: 527 (2017)



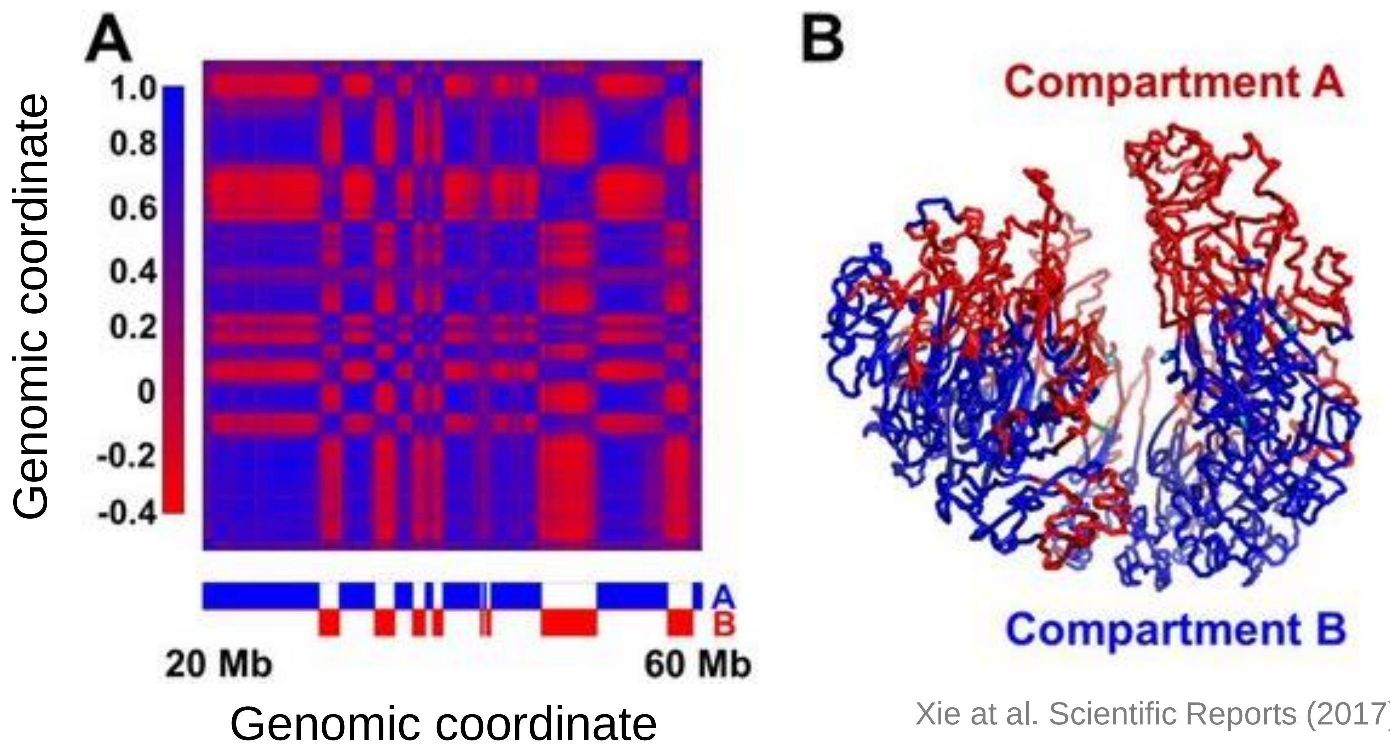
Hi-C data guide metagenome deconvolution



- Hi-C reads link two regions in the same genome

Hierarchical organization of chromatin

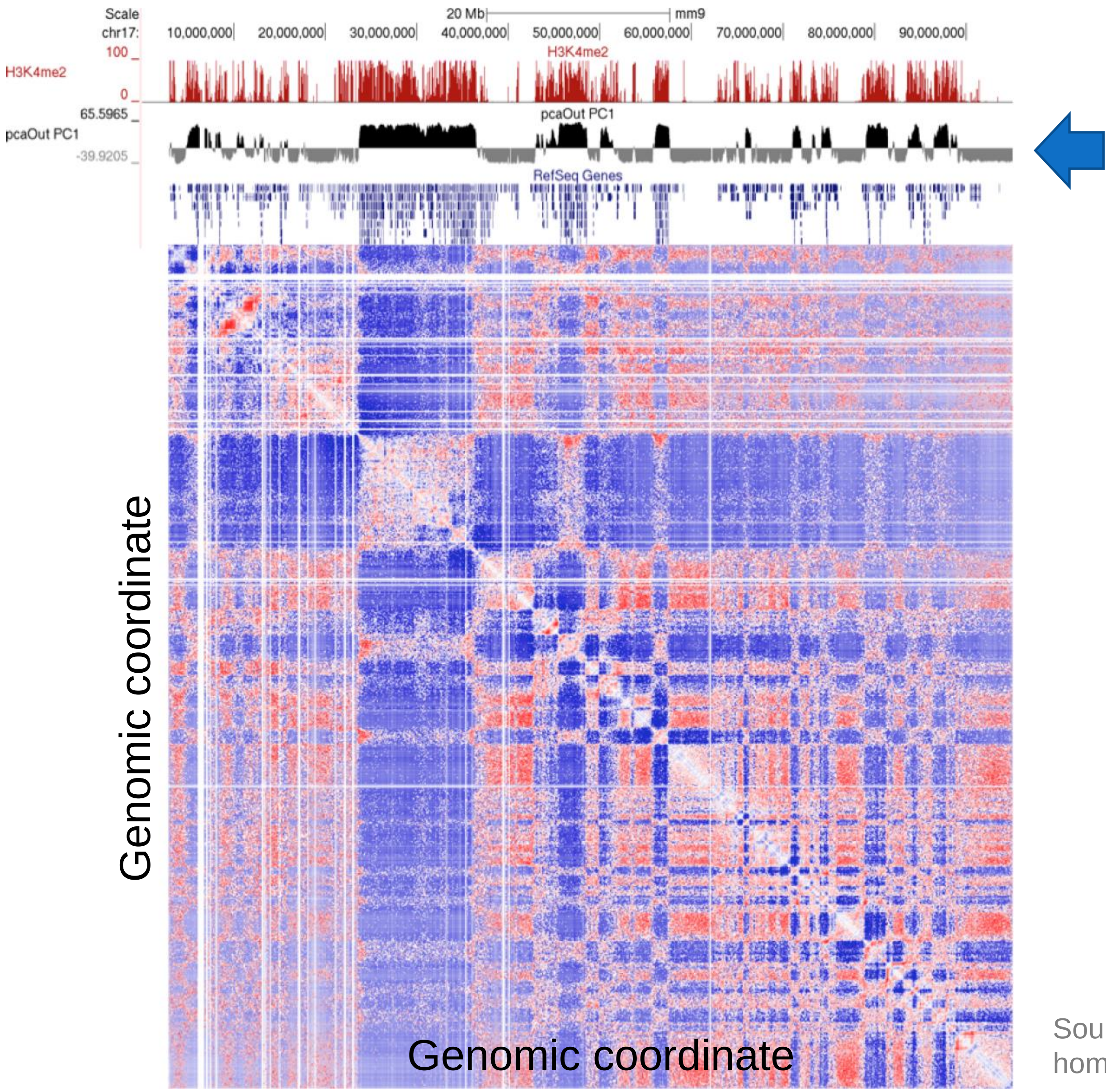
Chromatin segregate into two compartments



Xie at al. Scientific Reports (2017)

- Genomic loci associate together in plaid-like pattern
- In many system, this patterns correspond to open and close chromatins
  - Also called A and B compartments

PCA identifies A/B compartments



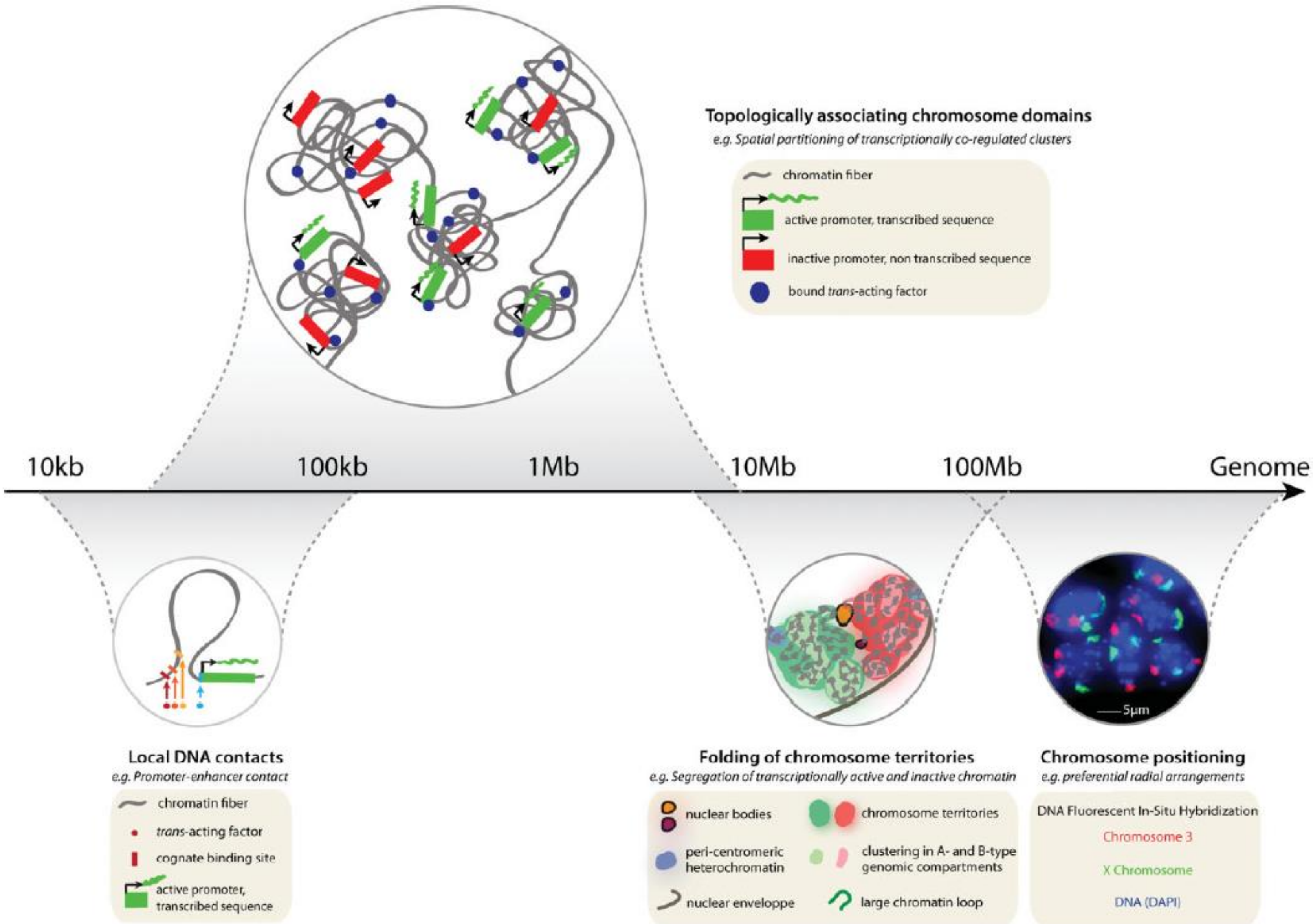
A/B compartment structure corresponds to +/- sign of PCA analysis of Hi-C data

Not always PC1

Should visualize the data and check correlation to select appropriate PCA component

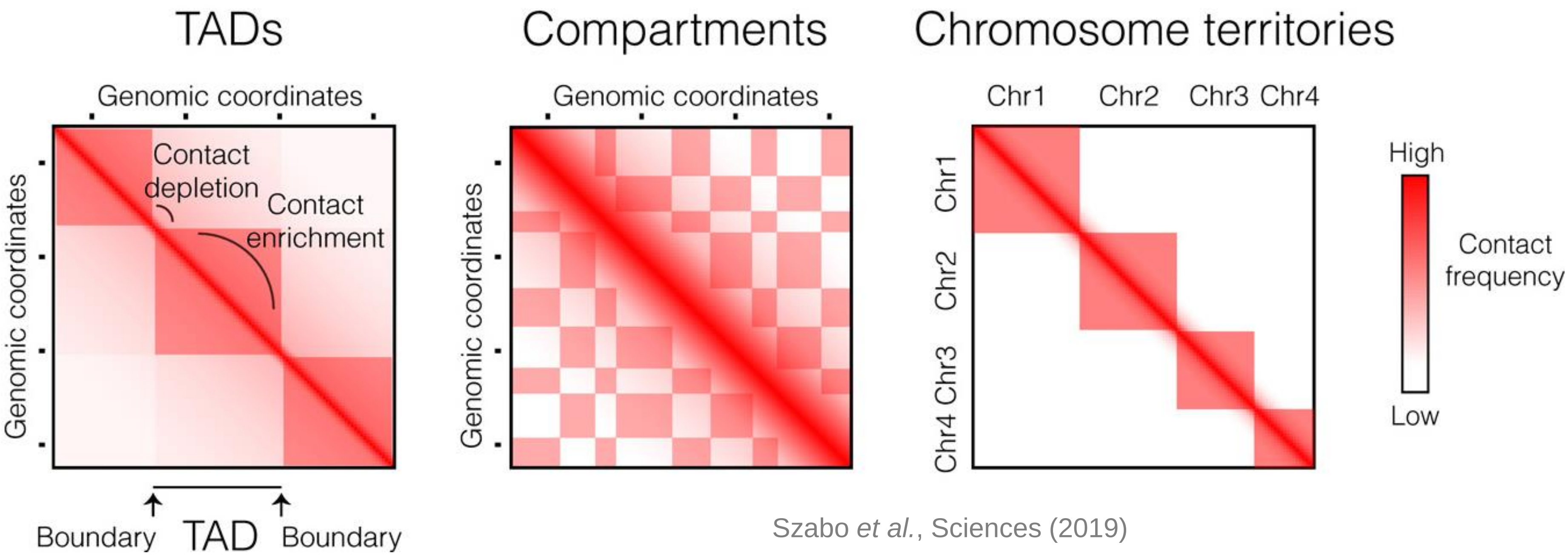
Source: homer.ucsd.edu/homer/interactions/HiCpca.html

Hierarchical organization of chromatin



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

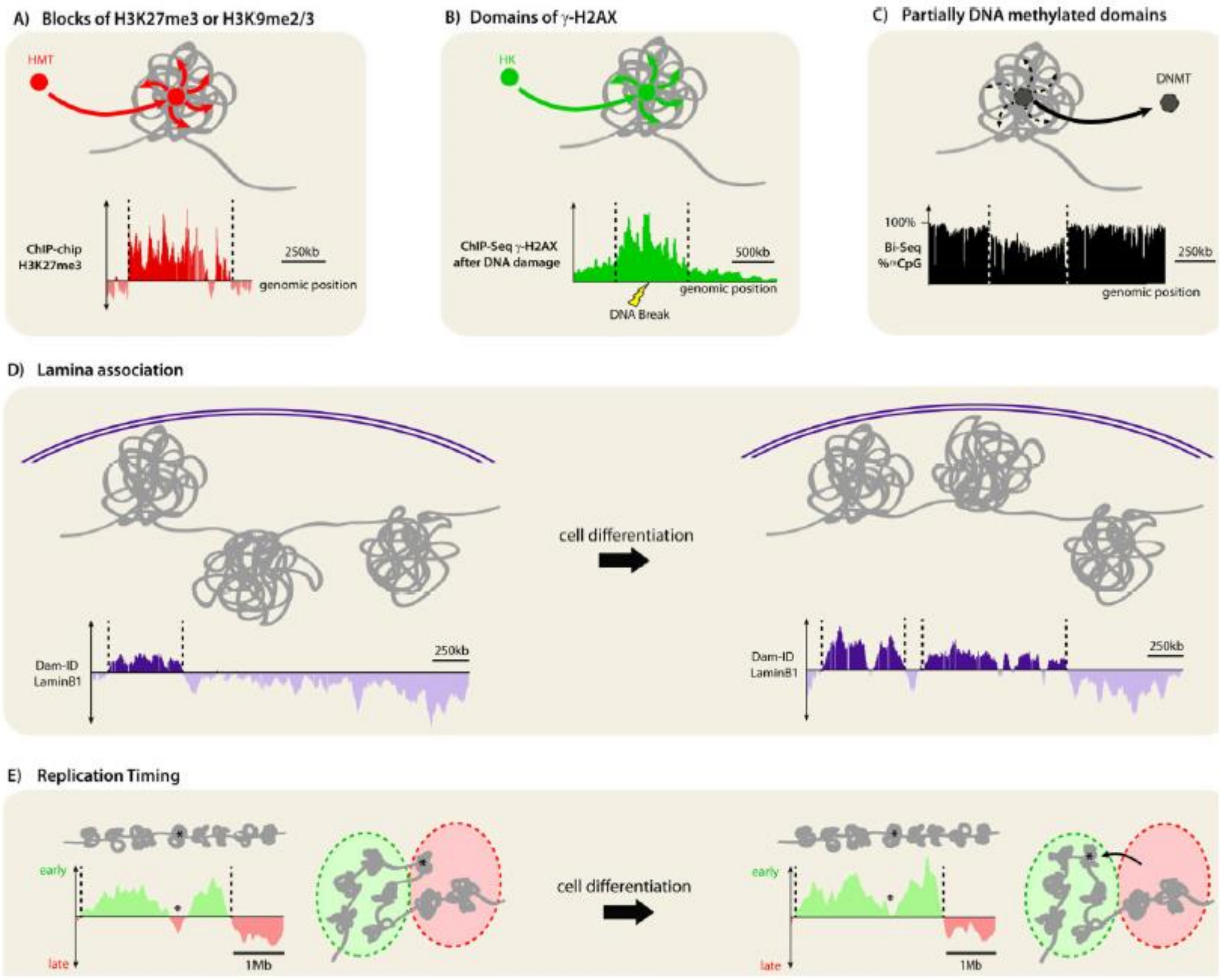
Hierarchical organization of chromatin



Szabo *et al.*, Sciences (2019)

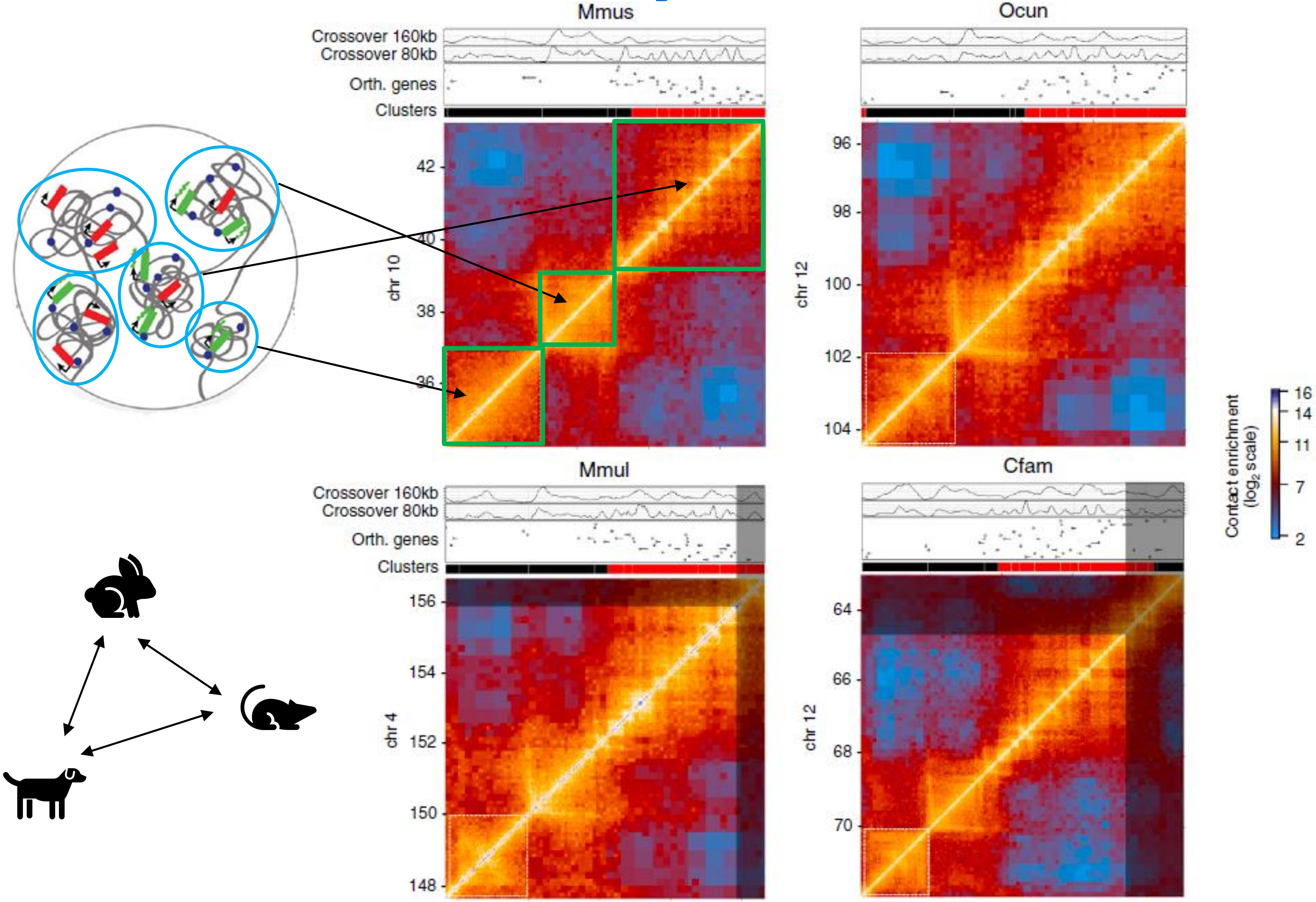
- Chromosome territory
- Active-inactive chromatin compartment
  - Also called A/B compartment
- Topologically associated domain (TAD)
  - 100kb – 10Mb

Topologically associating domains (TAD)



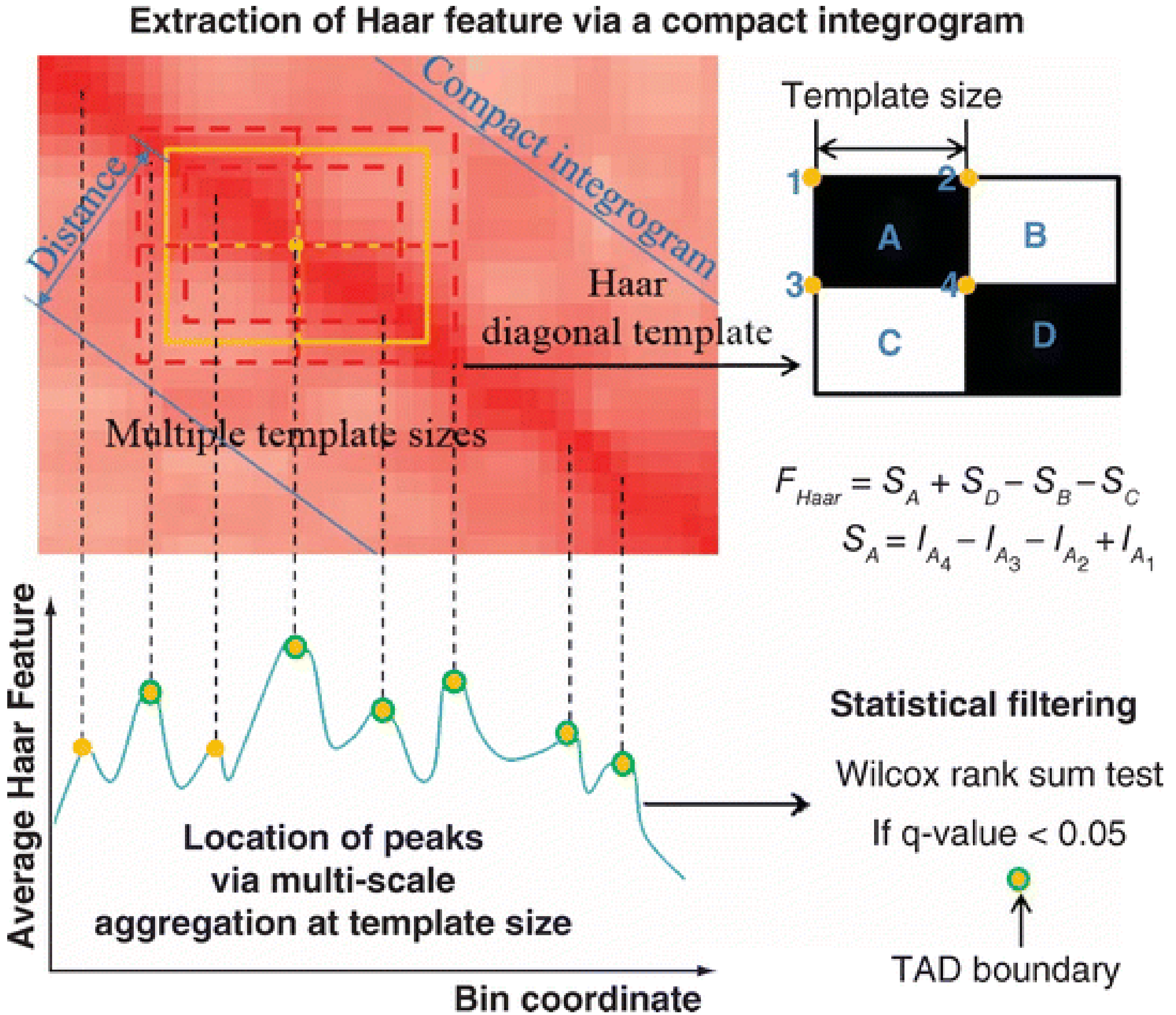
Nora *et al*. Bioessay (2013)

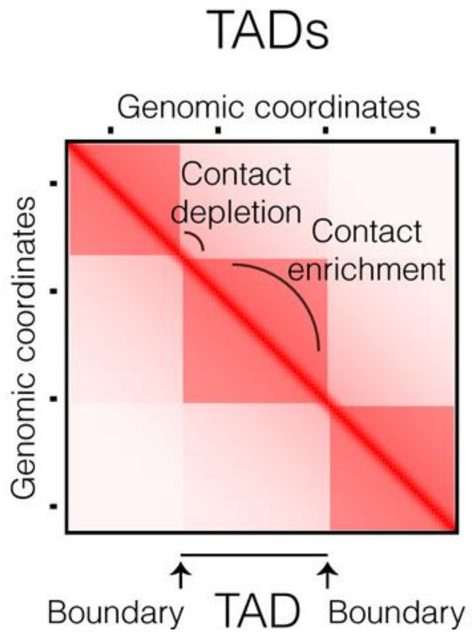
TADs are evolutionarily conserved



Rudan *et al*. Cell Reports (2015)

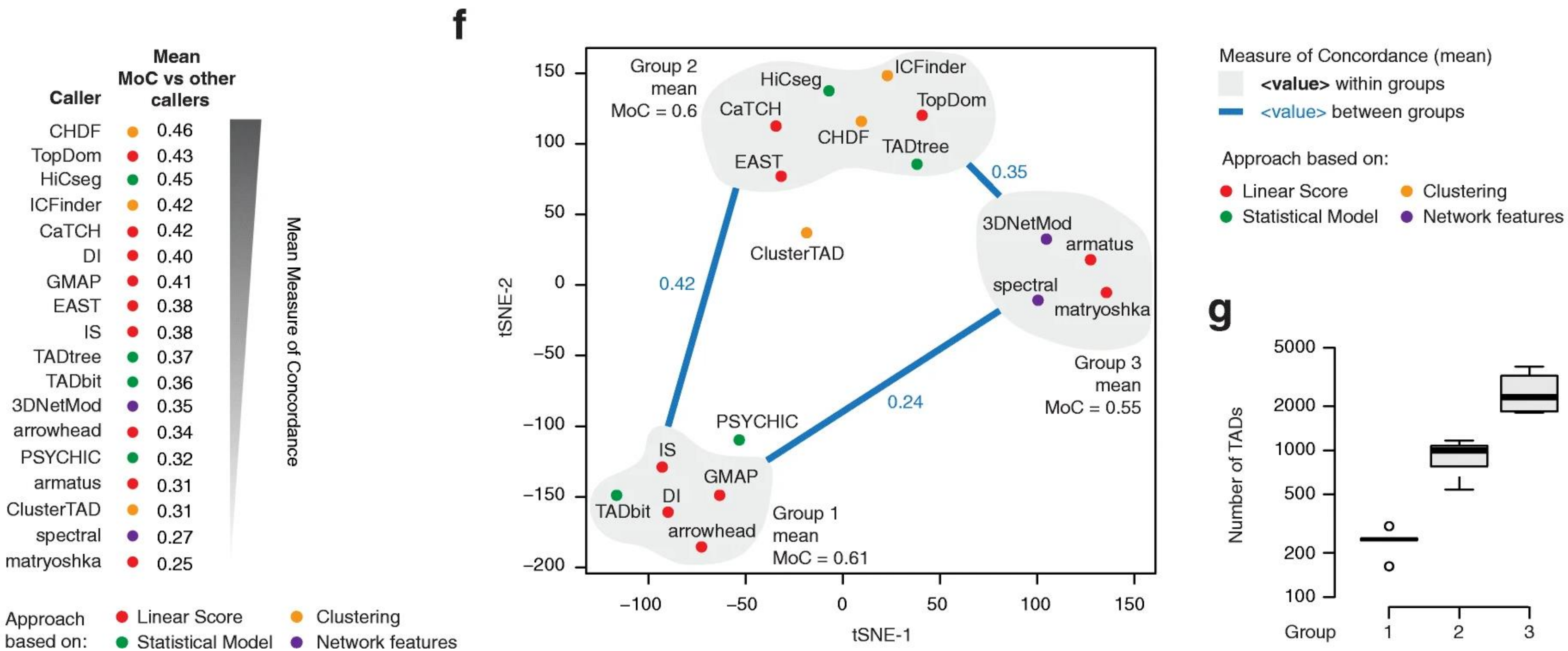
TAD identification





Lyu *et al.*, Biotechniques (2020)

Various TAD analysis tools

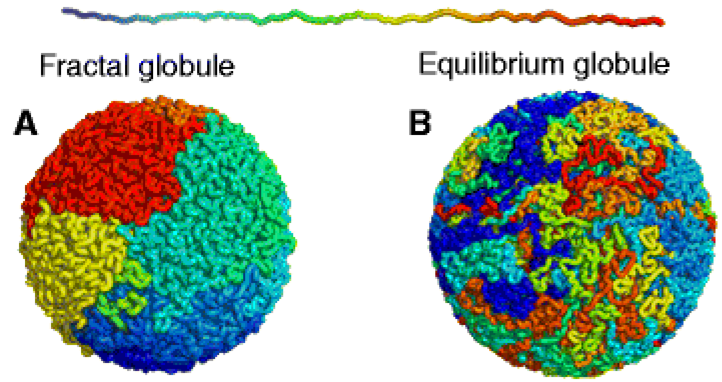


Zufferey *et al.*, Genome Biology (2018)

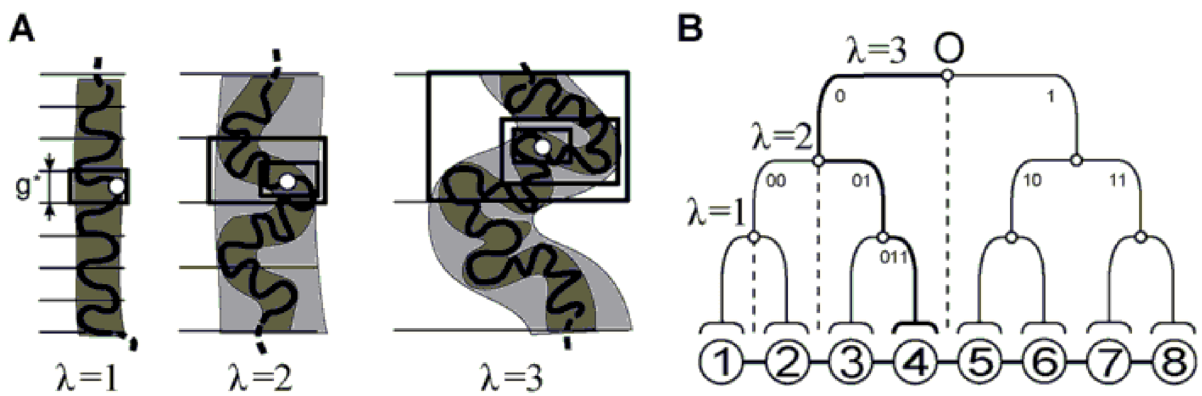
- Because Hi-C data were generated from thousands of cells, observed TAD structure is the average over multiple TAD conformations

Chromatin folding simulation

Early models of chromatin folding



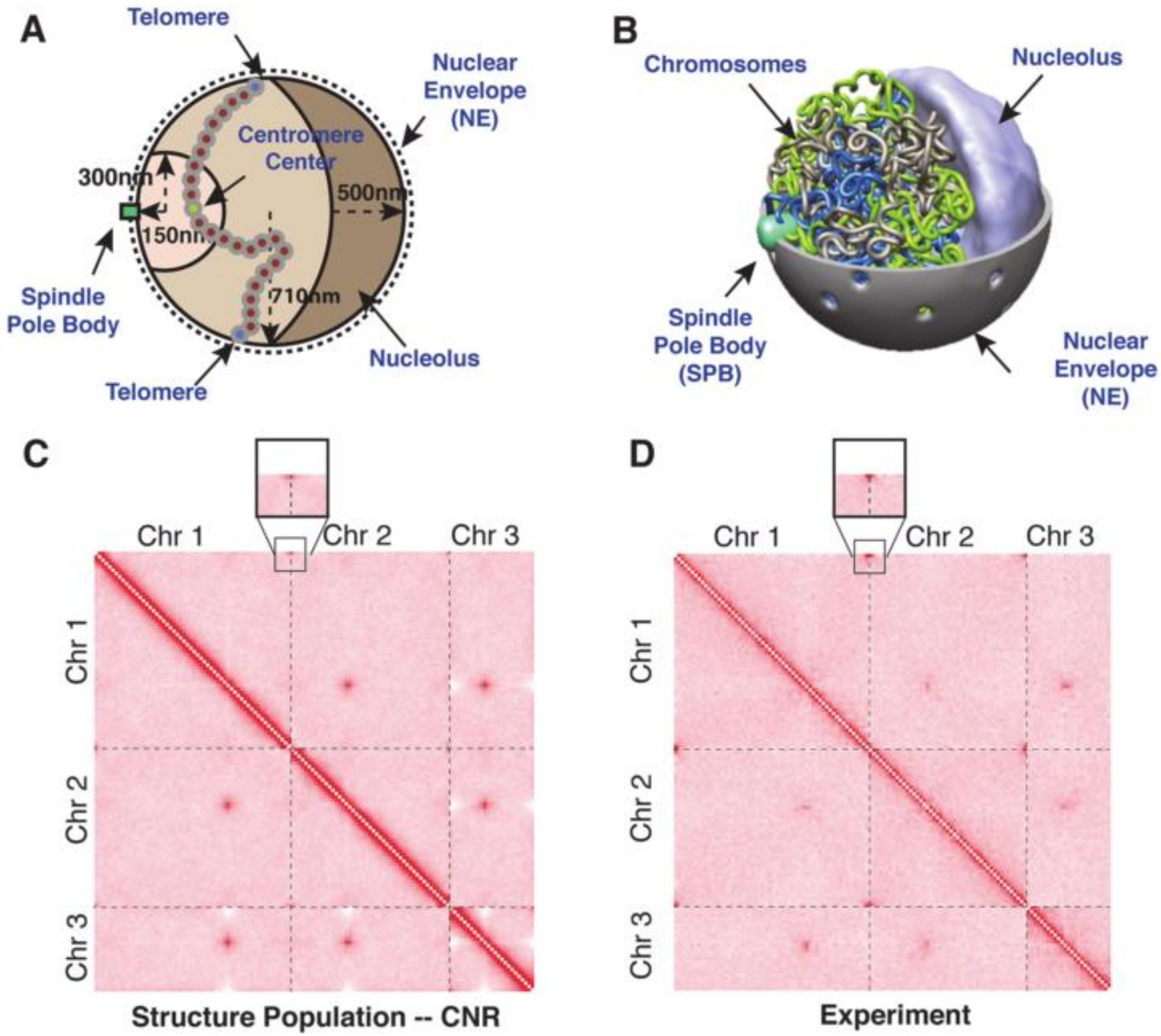
Mirny. Chromosome Research (2011)



Nazarov *et al*. Soft Matter (2014)

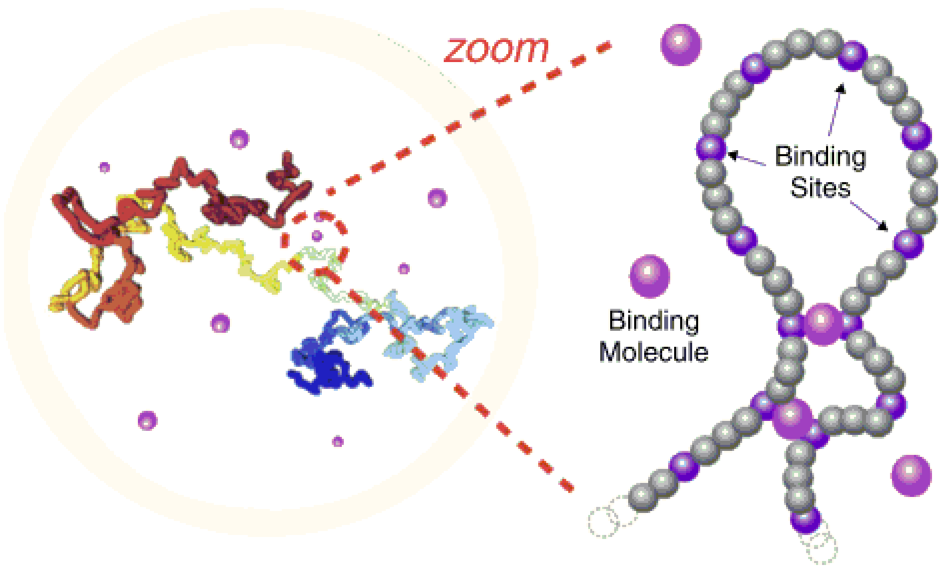
- Early models with simple global physical properties could explain the organization of chromosome to some extent

Modeling chromatin packaging

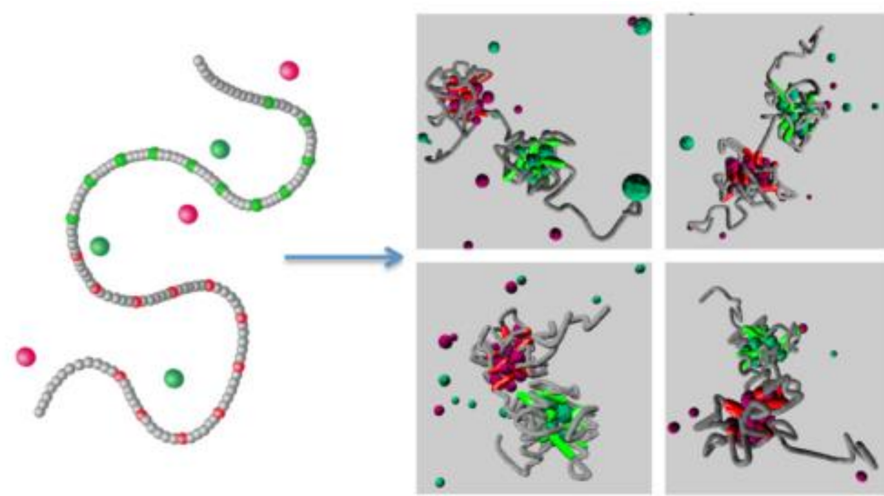


Gong *et al*. PLoS ONE (2015)

Strings and binders switch model



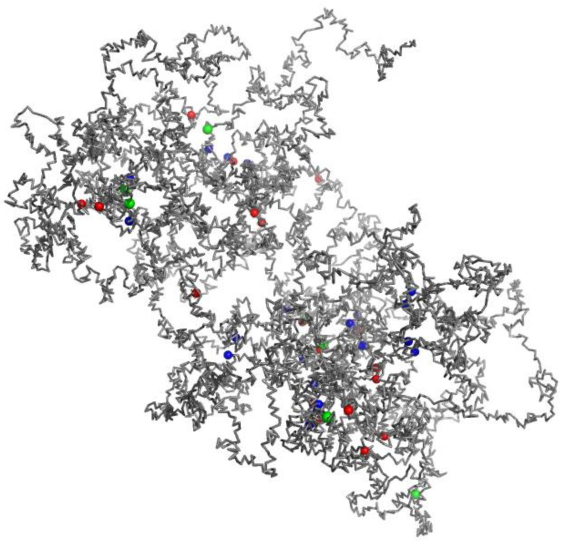
Nicodemi and Pombo. Curr Opin Cell Biol (2014)



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

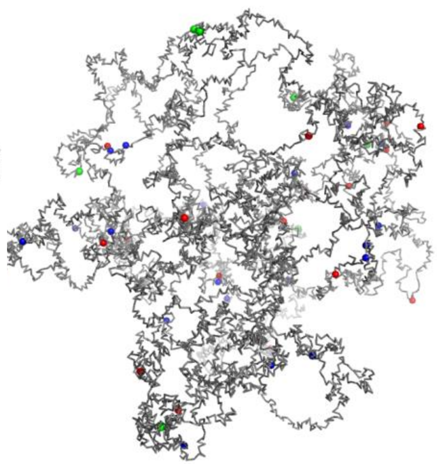
- Protein binds to two loci on genome
- Requires multiple protein-chromatin specificity to form multiple stable TADs

Modeling chromatin as polymer

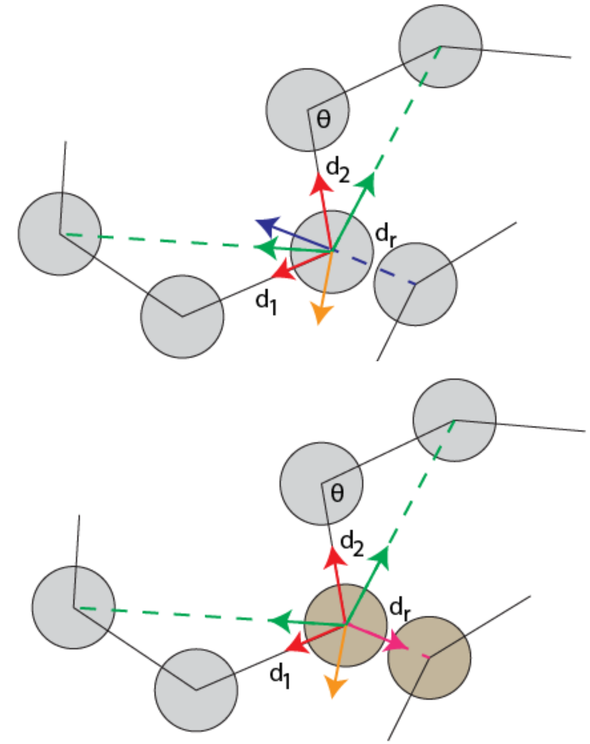


With Protein

Without Protein



1 Mb section of a chromosome
represented by a 7000-unit polymer chain

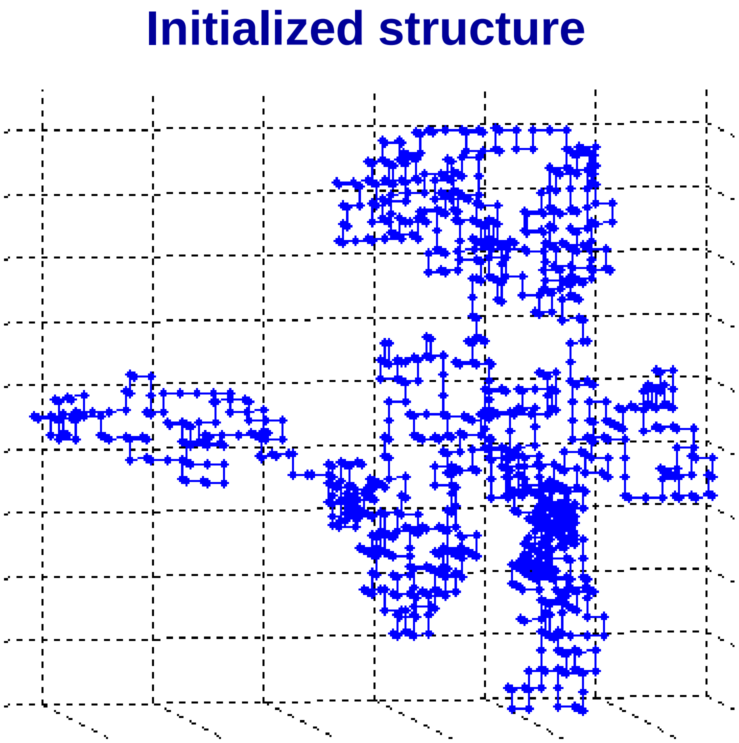


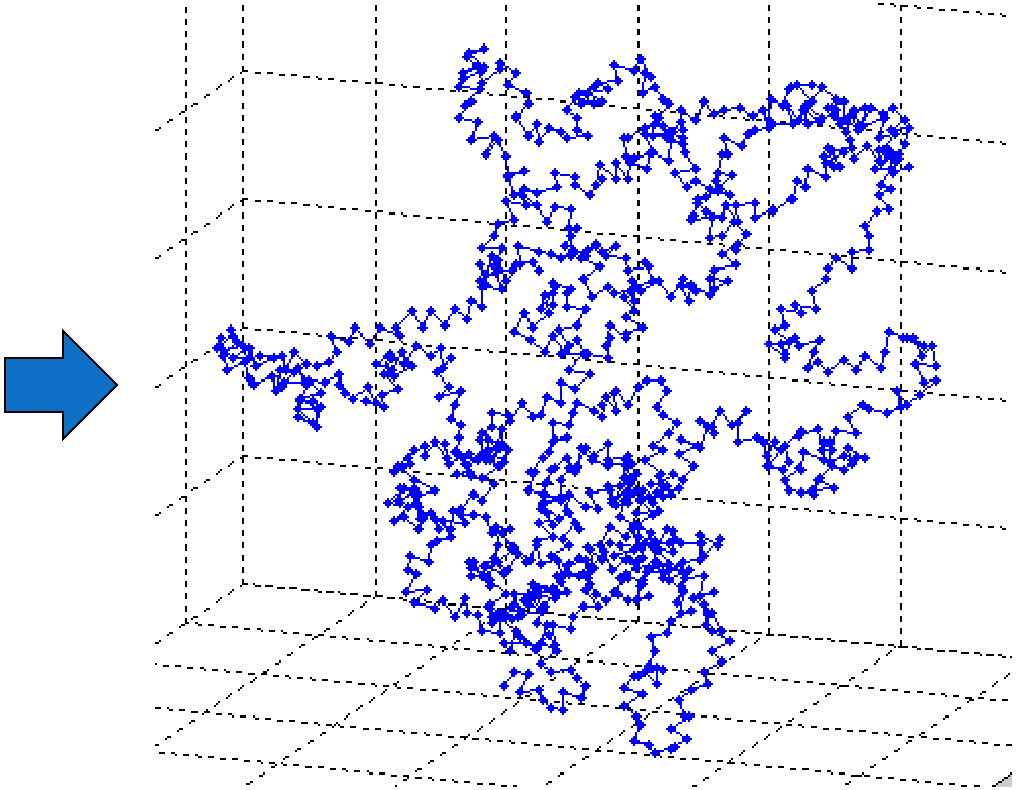
Regular unit

Unit containing protein binding site

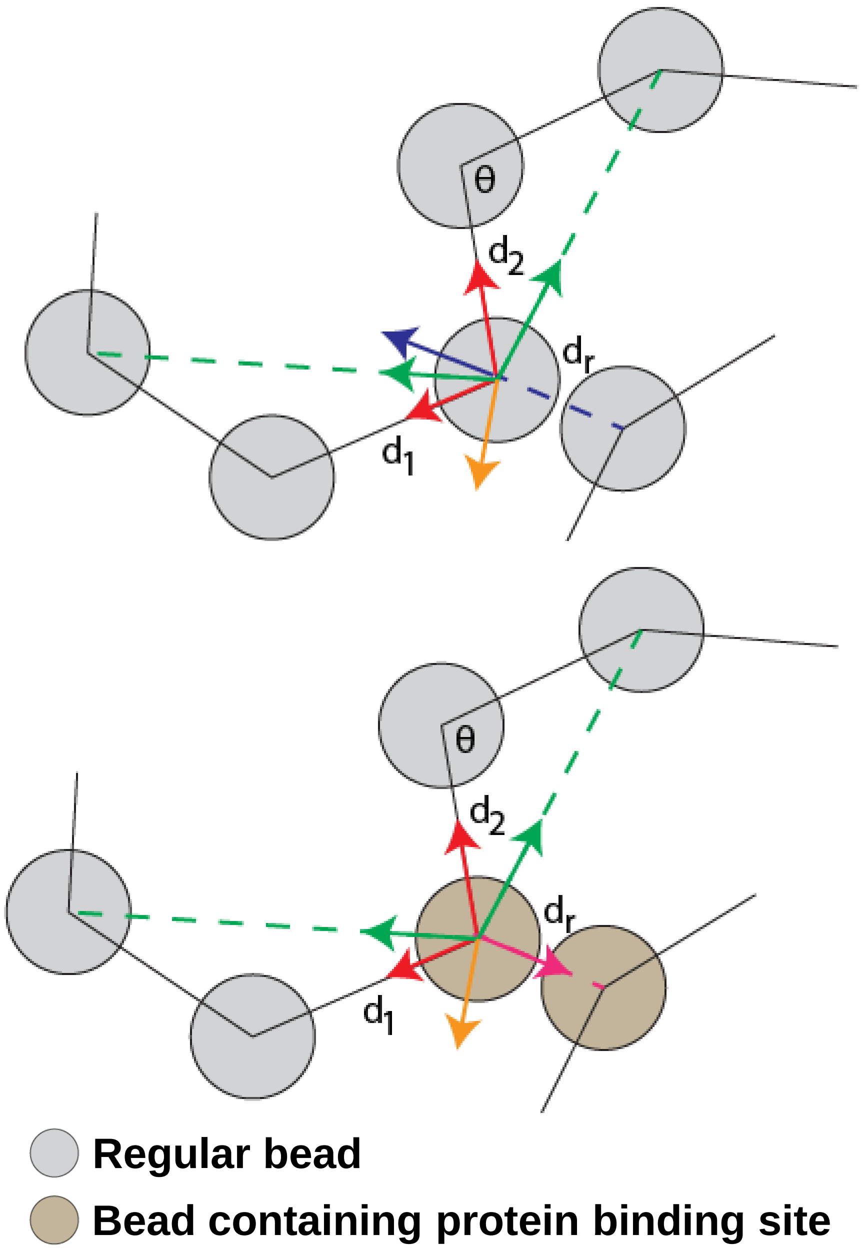
- A monomer unit can be 1 nucleotide, 1 nucleosome, or larger genomic segment
- More monomers means more computational cost!

Self-avoiding random walk



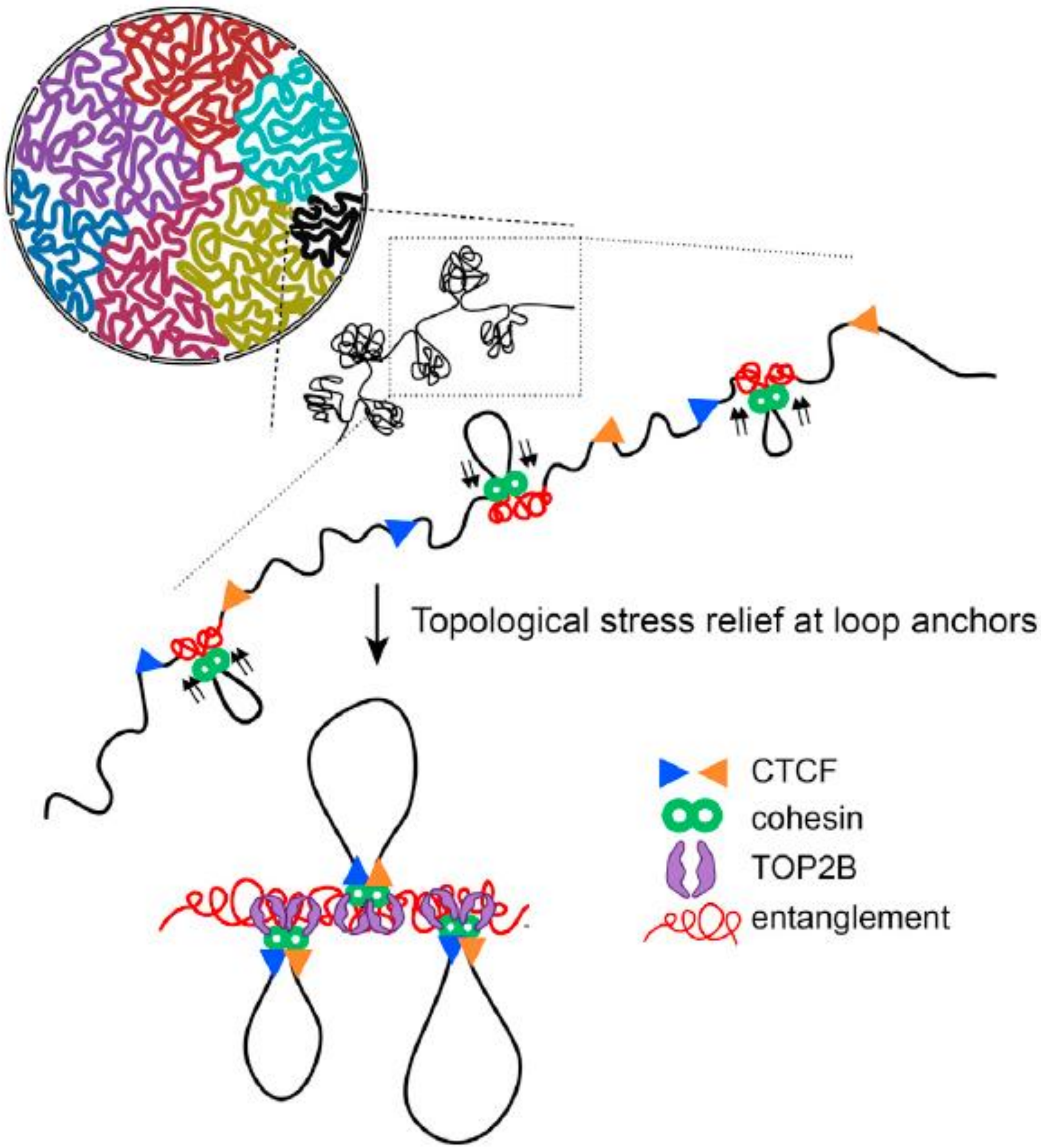


Physics of chromatin folding



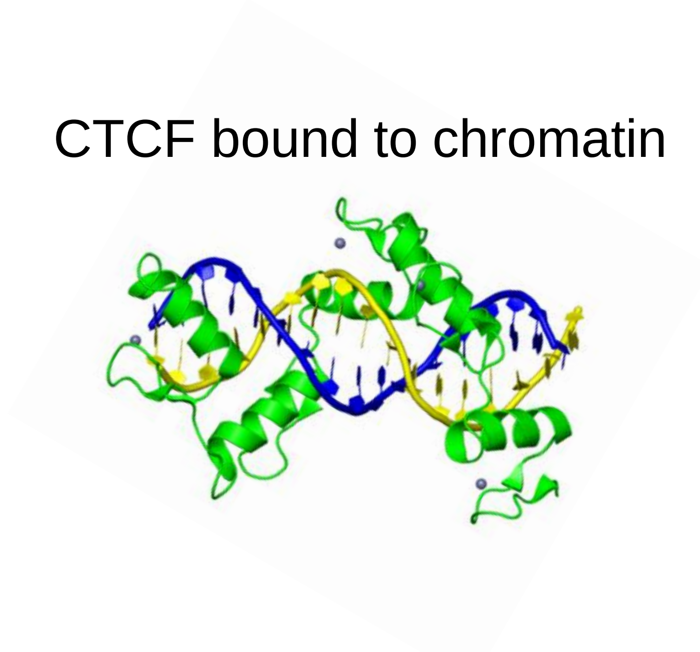
- Entropic (diffusional) = constant force with random direction
- Tension (bead-bead) = spring force between adjacent beads
- Repulsion = dispersion force between close-by beads
- Attraction = artificial force to control the distribution of θ
- Protein-mediated = spring-like force between close-by beads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

Formation of TAD

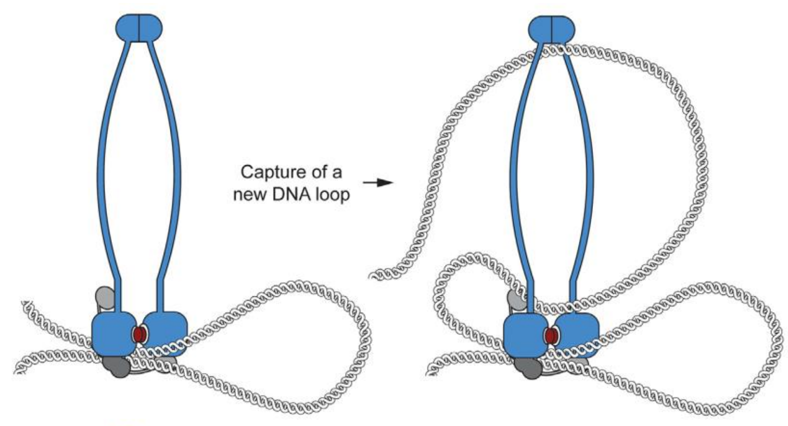


Canela *et al*. Cell (2017)

CTCF bound to chromatin

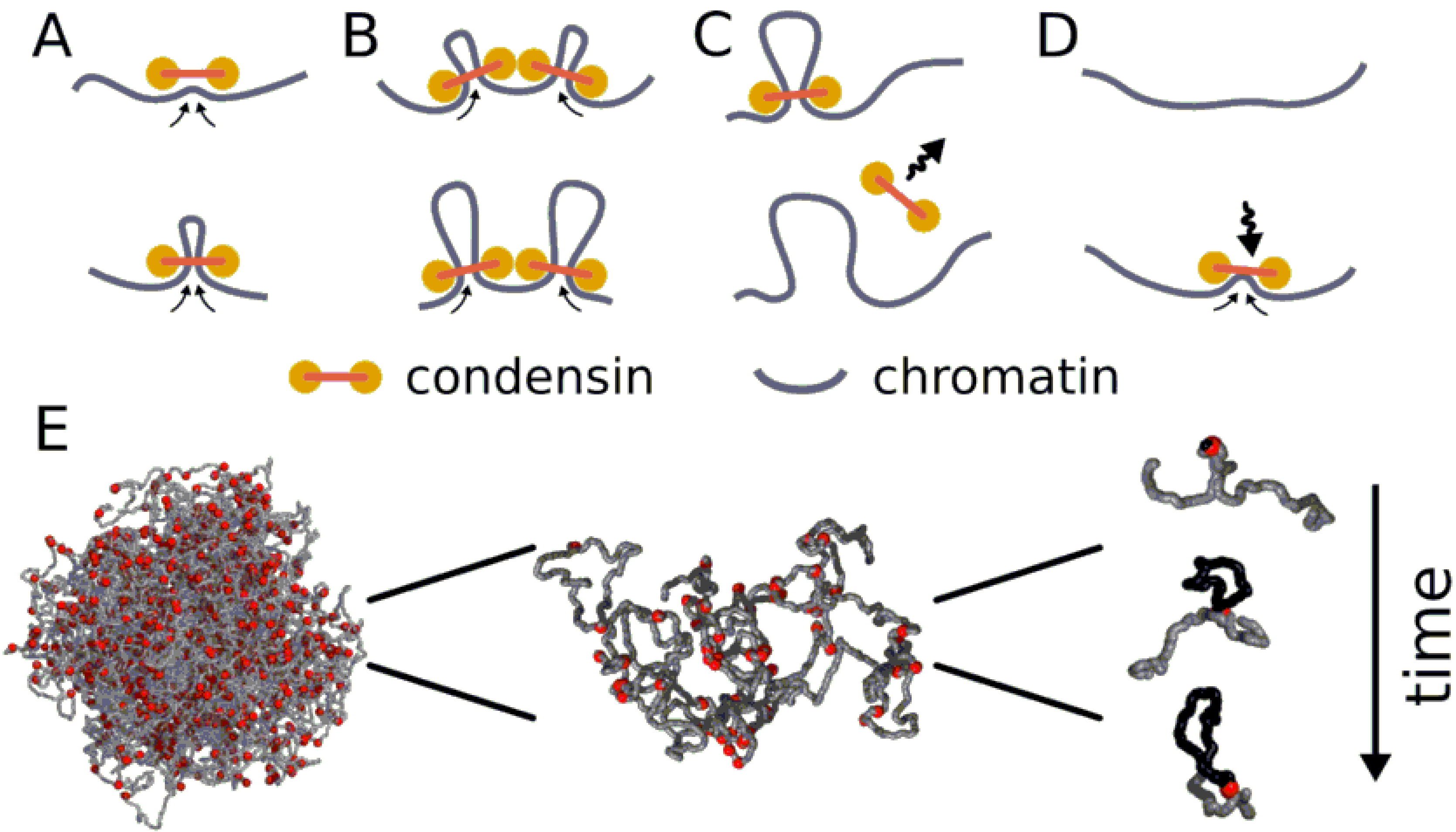


Cohesin structure



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

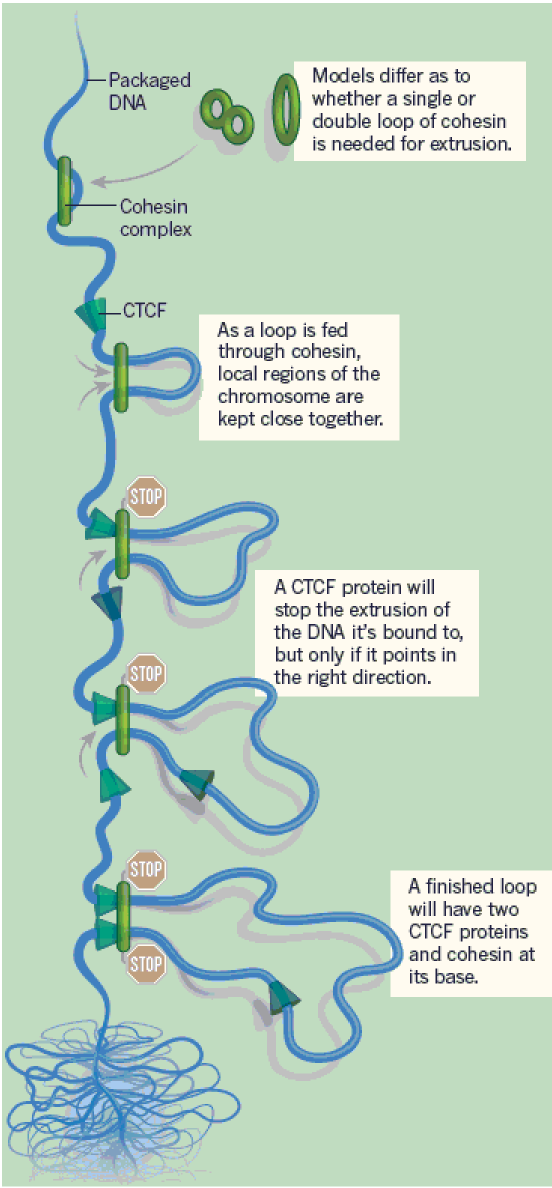
Loop extrusion model



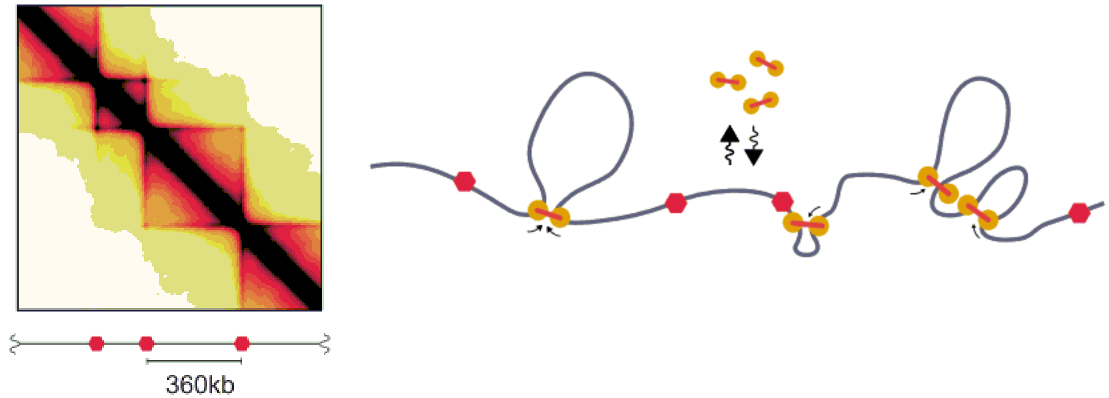
Goloborodko *et al*. eLife (2016)

- Cohesin molecules “walk” along chromosome, effectively pulling in chromosome strand to create loops

Loop extrusion model



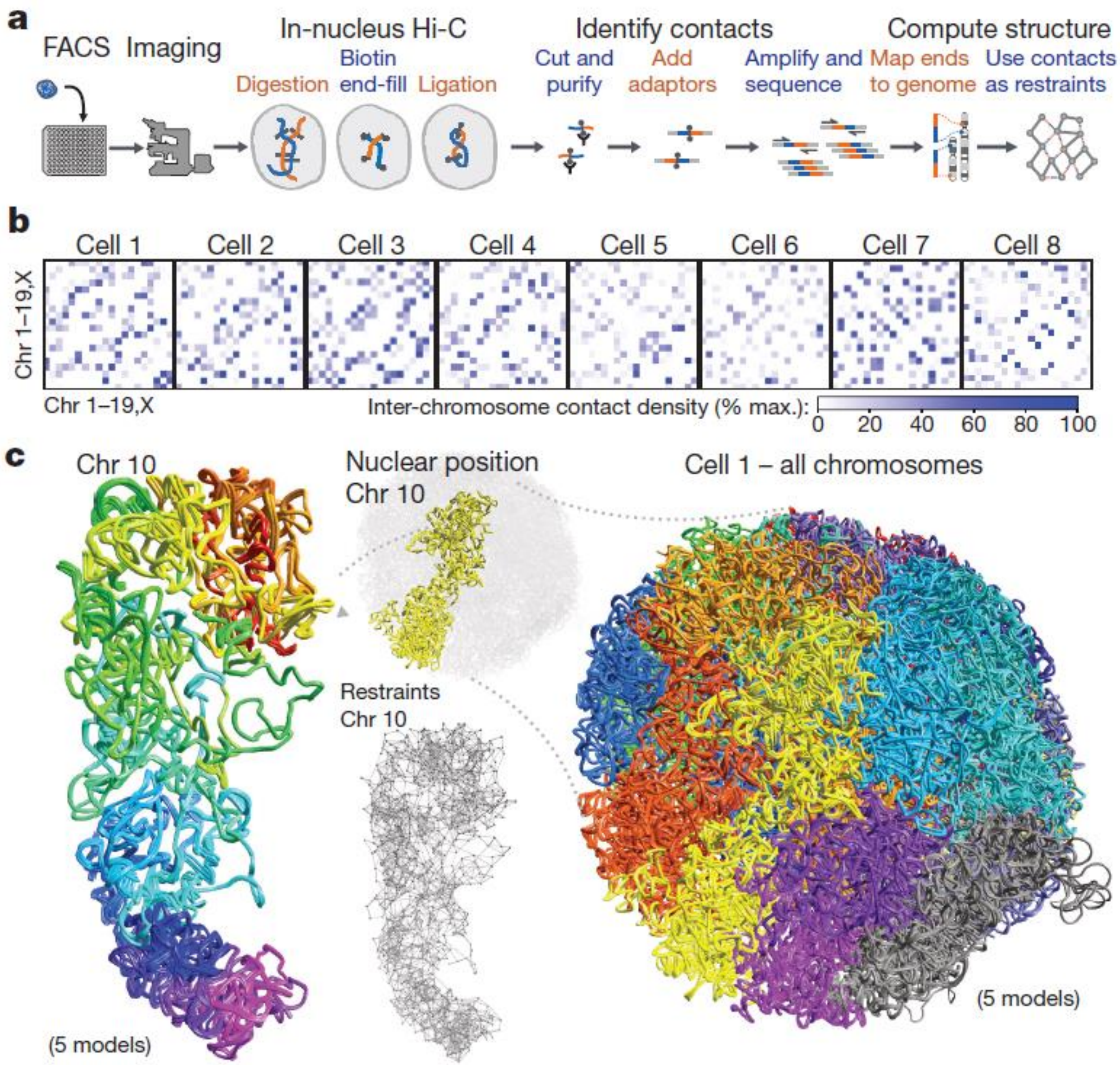
Dolgin. Science (2017)



Fudenberg *et al*. Cell Reports (2016)

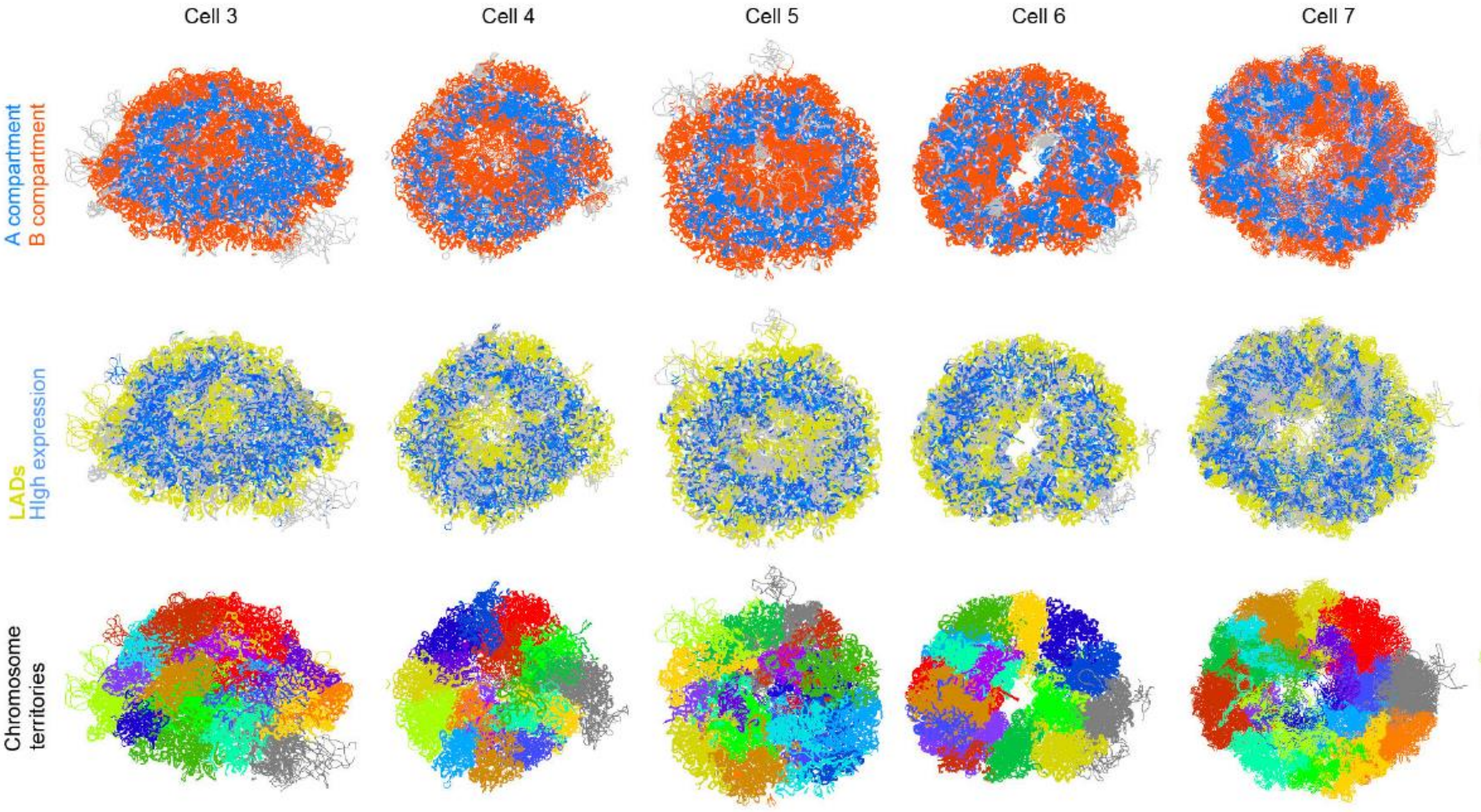
- Cohesin, condensin = walker
- CTCF = blocker
- Other proteins may be involved
- Still many open questions regarding this process

Single-cell Hi-C



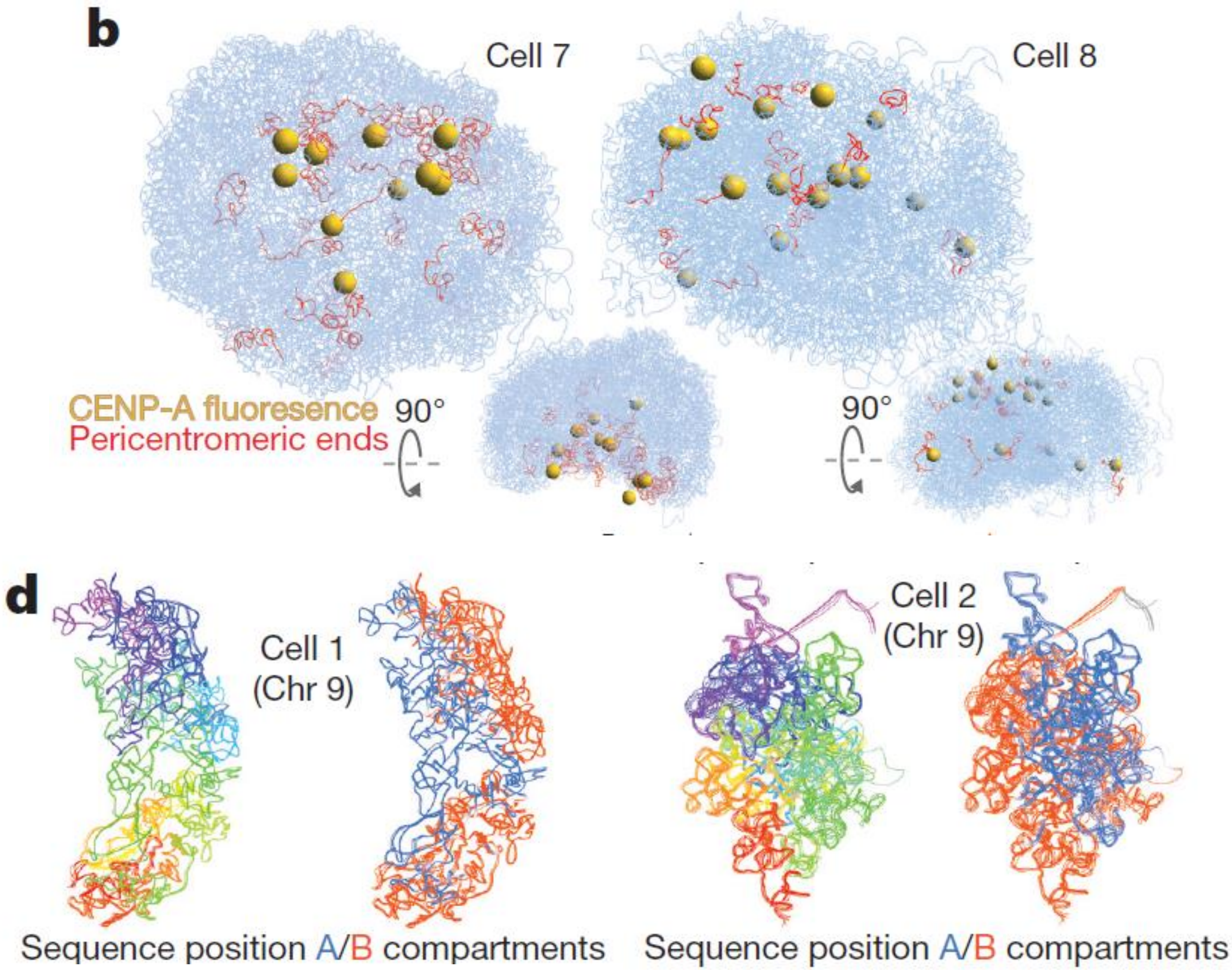
Stevens *et al*. Nature (2017)

Cell-to-cell variations



Stevens *et al*. Nature (2017)

Cell-to-cell variations



Stevens *et al*. Nature (2017)

Any question?