Using statistical profiling to decipher hidden chromatin contacts resulting from repeated sequences

26/11/2024

Sébastien Gradit
Spatial Regulation of Genomes - Genomes & Genetics - UMR3525

Thesis supervisor: Axel Cournac

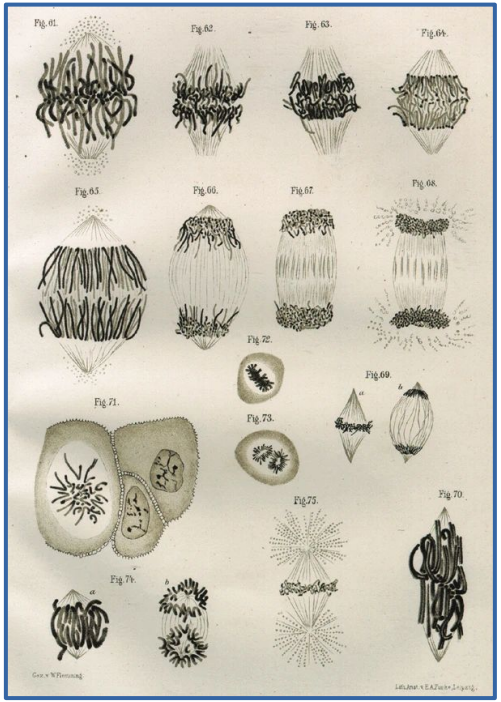
INCEPTION

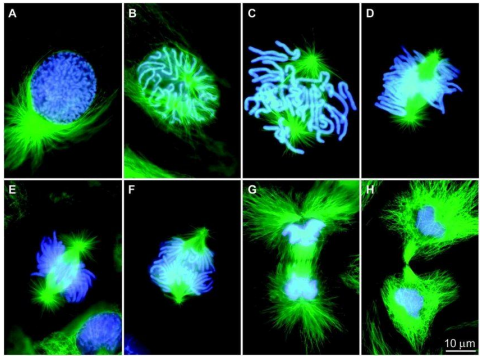
anr

INSTITUT PASTEUR

Early observation of genomes organisation

Lilium corceum drawings





Human cell

3.10^9 bp = 2m

yeast

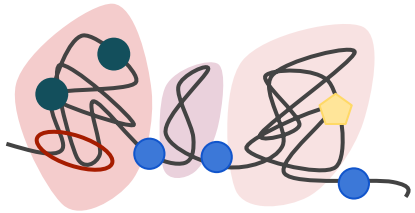
12.10^6 bp = 8 mm

DNA folding is a highly organized process necessary to fit DNA in nucleus

Zellsubstanz, Kern und Zelltheilung, Flemming, 1882
Risler, 2011

Chromosome folding influences or regulates dynamic processes?

Genome folding



Biological functions

- Gene transcription
- DNA Repair ; Replication
- Chromosome segregation

ADN

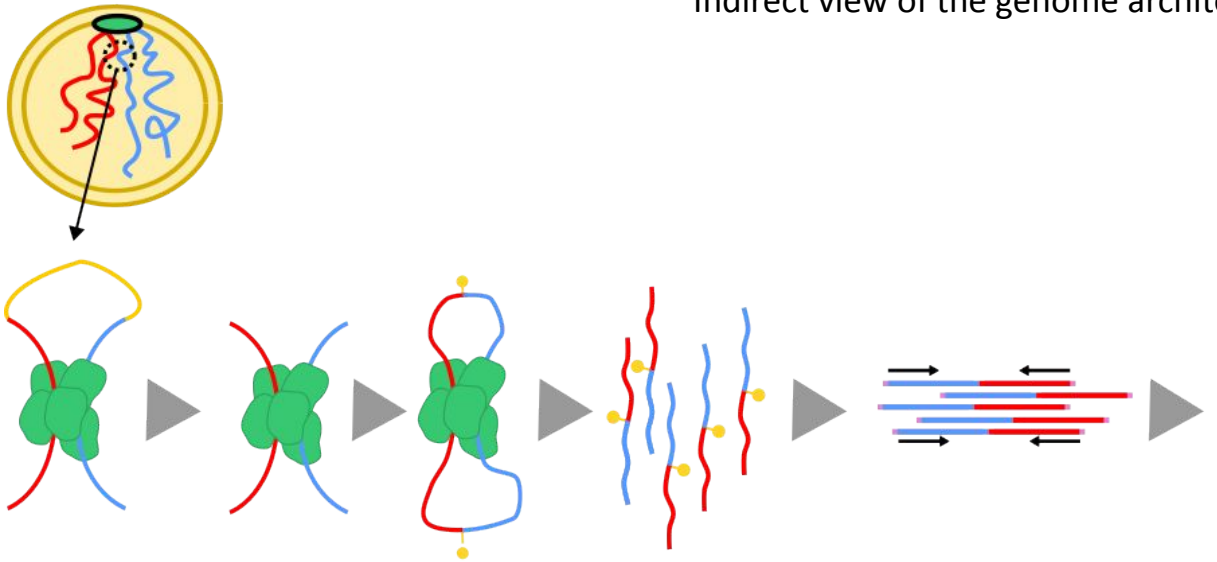
DNA binding proteins

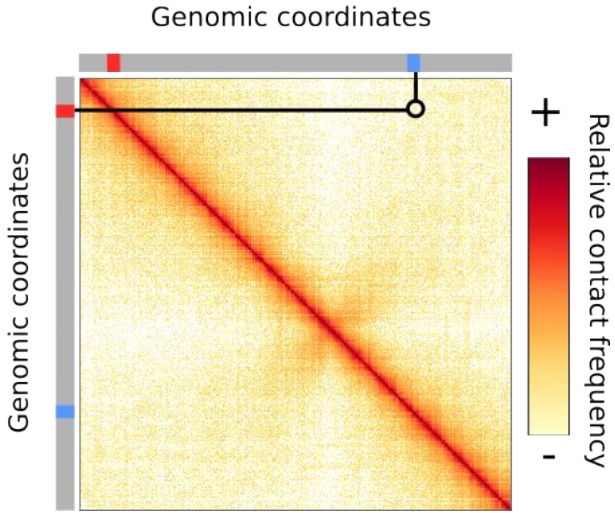
Physical constraints (supercoiling, etc.)

Cause or consequence ?

Hi-C to capture genomes 3D structures

Indirect view of the genome architecture

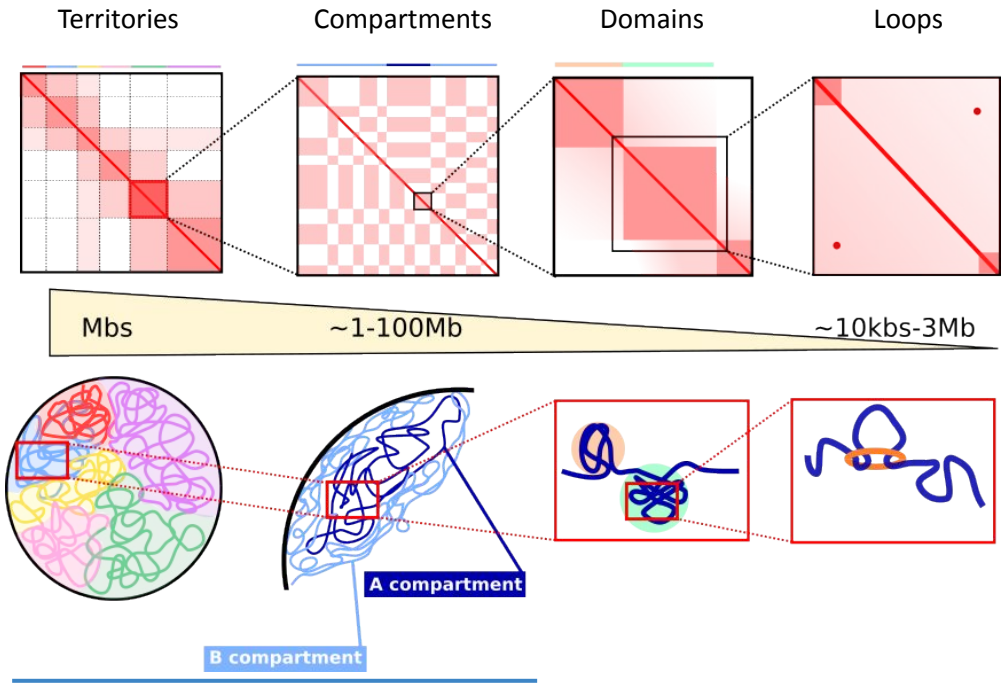




Adapted from Lieberman-Aiden, et al. 2009

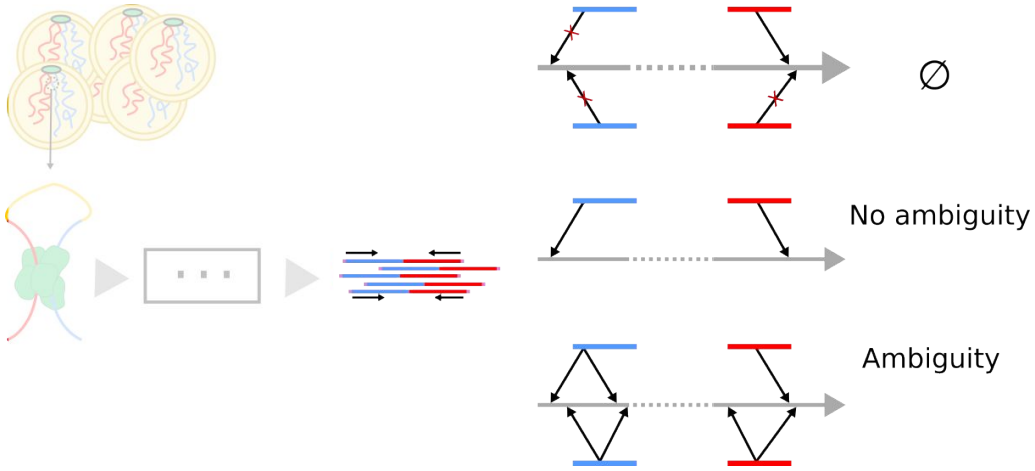
Chromosome Configurations: Influencing Genome Organization

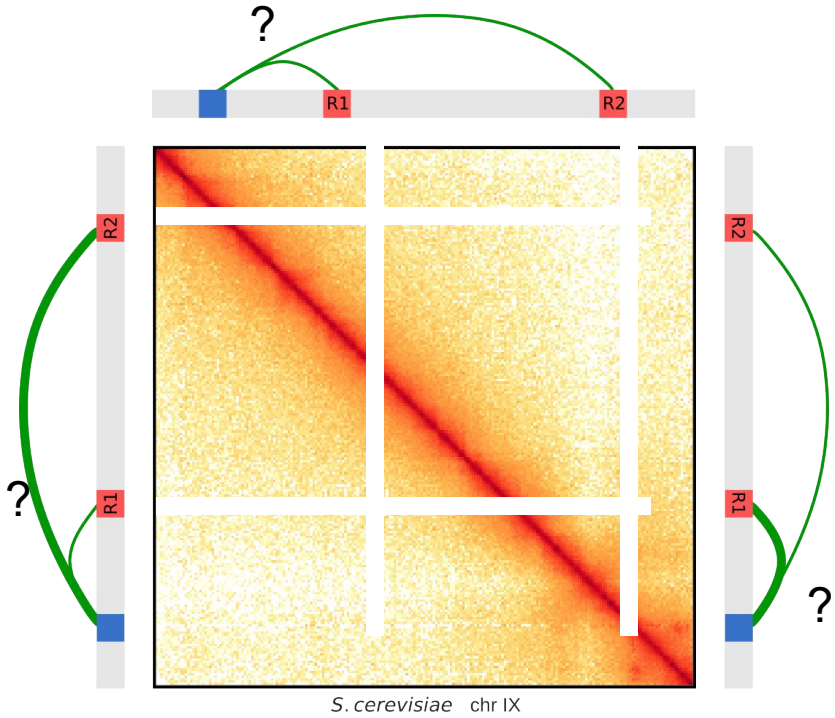
How these different spatial structures correlates with biological processes?



- Cell cycle phase dependence
- Response to stresses
- Quiescence / replicating?
- Viral infection?

Limitations of Traditional Hi-C



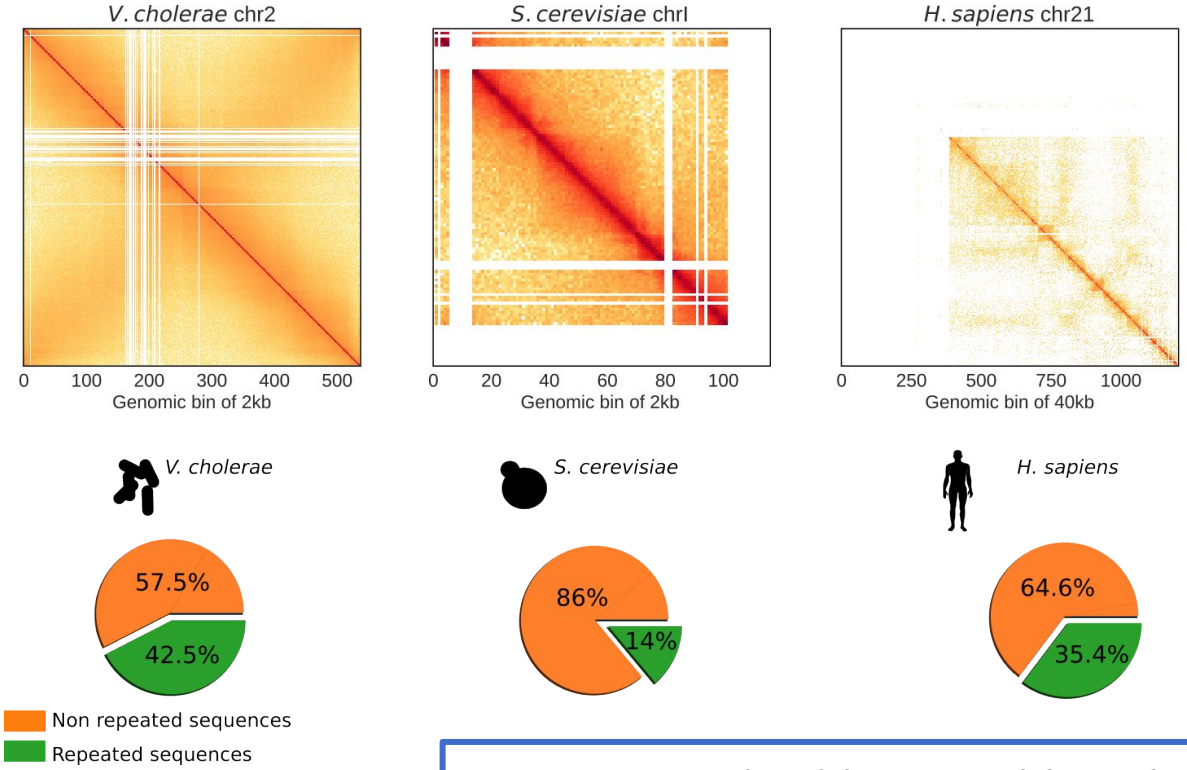


- Pair-end sequencing produces short reads ~ 50 bp
- Genomes carry repeated regions

  → Some reads can have multiple genomic positions

Statistical approach for ambiguous contacts prediction?

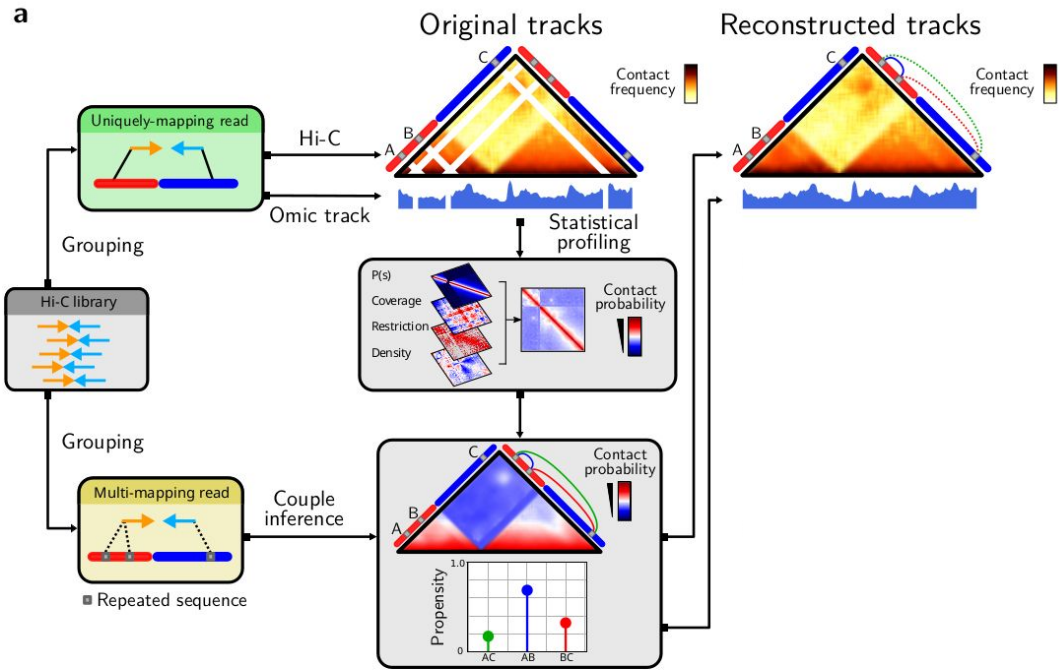
Thesis Objectives



Can a statistical modeling approach be used to infer missing information in Hi-C data and accurately reconstruct the 3D interactions of repeated sequences?

Hicberg: Reconstruction of genomic signals from repeated elements

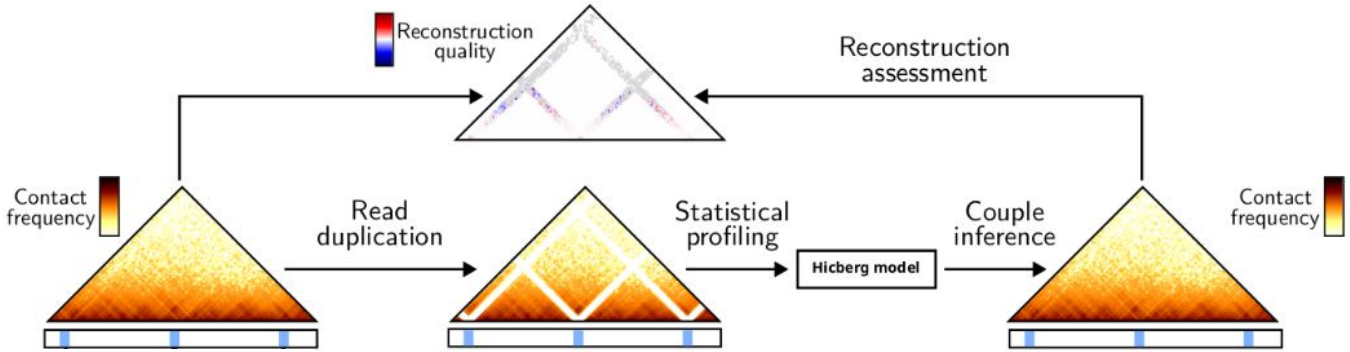
Hicberg: a novel tool for (3D) genomics



- **Hicberg:** Reconstruction of Hi-C data, and more!
- **On-the-fly statistical profiling:** extraction of statistical tendencies from unambiguous data.
- **Probability Mass Function estimation:** Guides accurate placement of ambiguous reads.

Gradit et al., in prep.

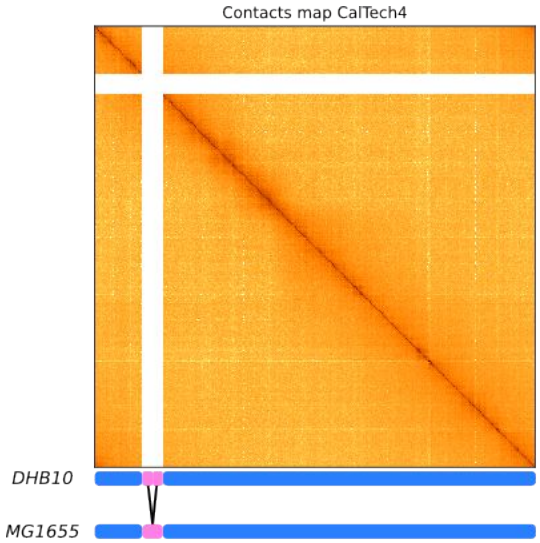
Benchmarking & In-silico Evaluation

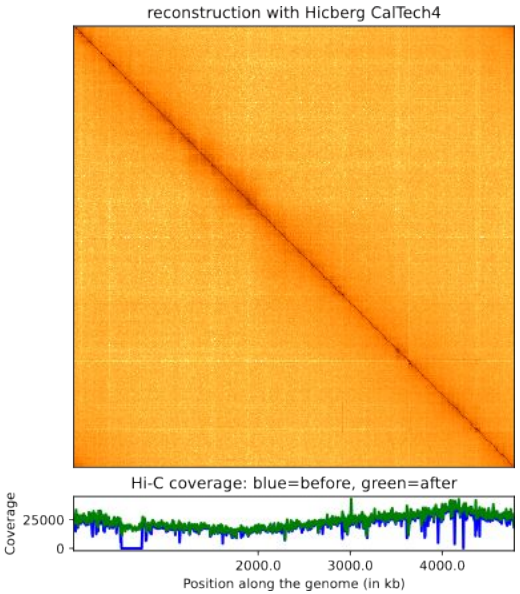


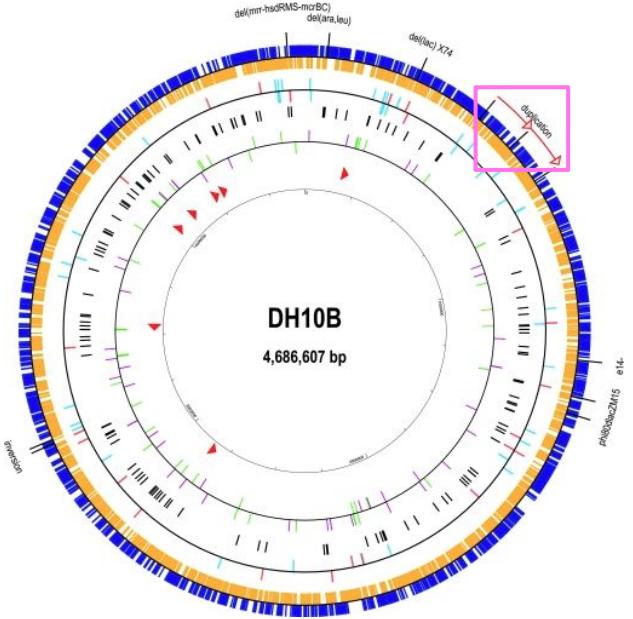
- **Simulated** repeated sequences by duplicating unambiguous reads.
- **Robust** across various repeat sizes, numbers, and spacing.
- **Polymeric** behavior component is key for accurate inference.

Does Hicberg perform as well in real-context?

Biological validation



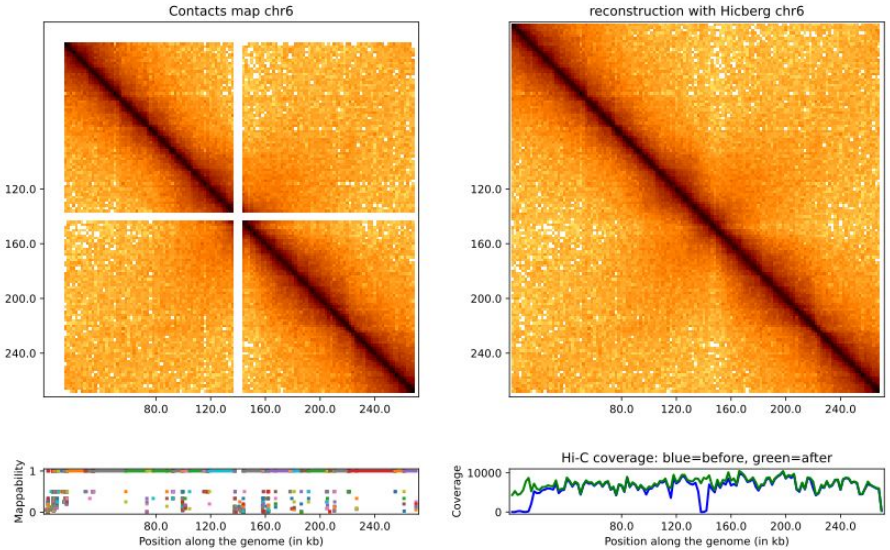




- Accurate reconstruction of the complex 3D structure of *E. coli* carrying tandem duplication.
- Hicberg captured expected DNA polymer behavior, demonstrating versatility and power.

Adapted from Durfee, et al. 2008

Biological validation



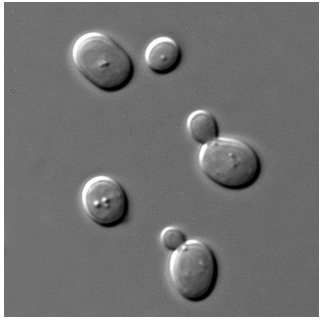
- Accurate reconstruction of the complex 3D structure of *S. cerevisiae* chr VI carrying repeats.
- Hicberg captured expected DNA polymer behavior, demonstrating versatility and power.
- Hicberg recovered expected mean coverage despite low theoretical mappability

Unveiling Hidden Connections: Hicberg in action

How do repetitive elements shape the 3D genome and influence its function?

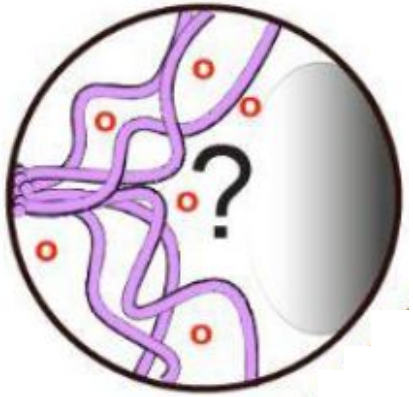
1) *Study of Ty1 Retrotransposons: 3D Organization and Cohesin Loading*

2) *Study of rDNA Contact Dynamics in Response to Stress*

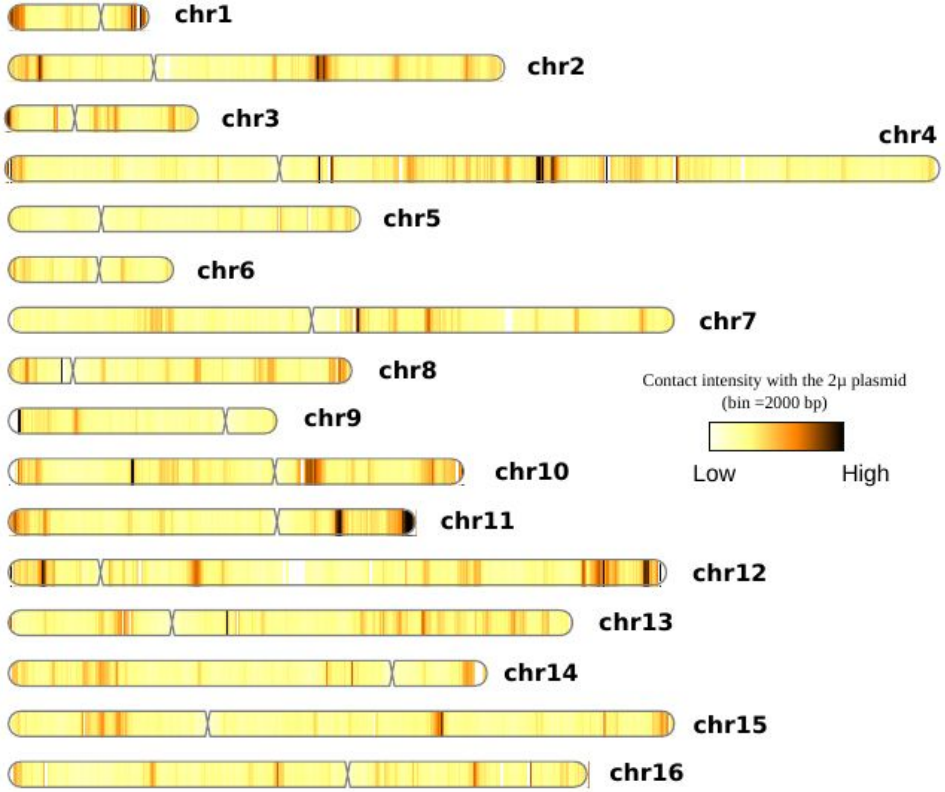


Beyond Parasitic Elements: Ty1 Retrotransposons as Organizers of Nuclear Architecture

Application : *S. cerevisiae* TYs as new 2µ preferencial contact spots



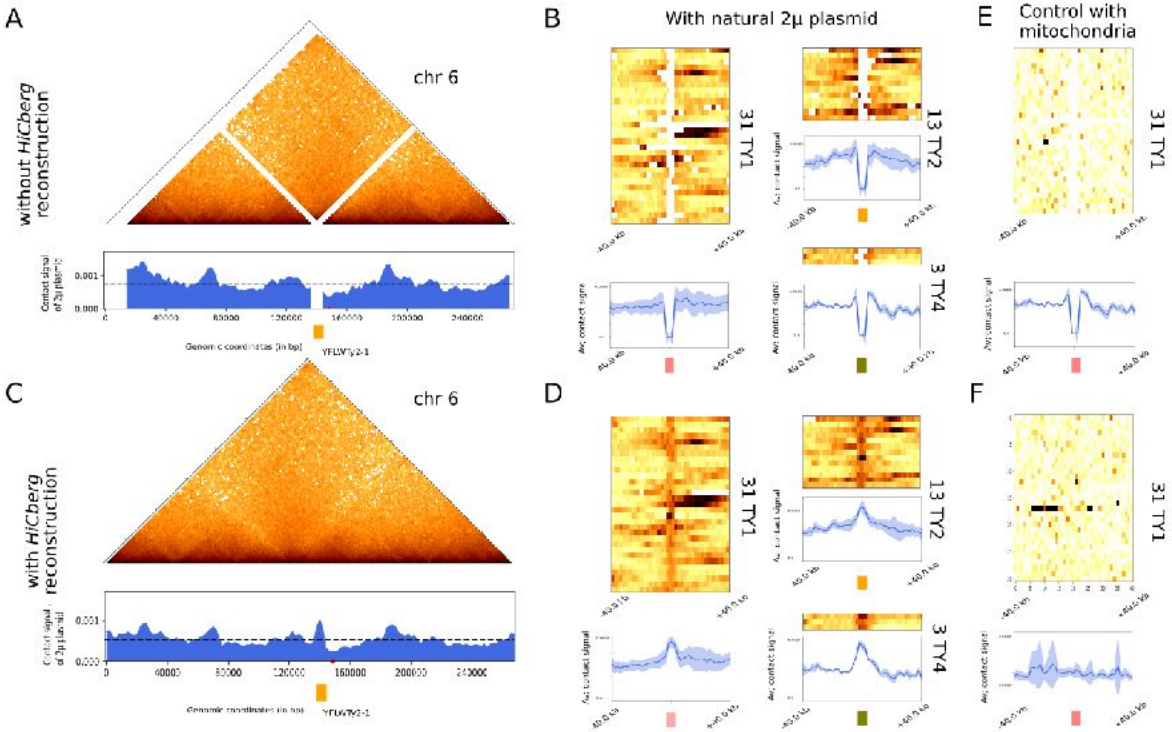
Presence of natural 2µ plasmid in *S. cerevisiae*



Detection of ~ 75 hotspots of contact

Adapted from Girard, et al. 2023

Application : *S. cerevisiae* TYs as new 2µ preferencial contact spots

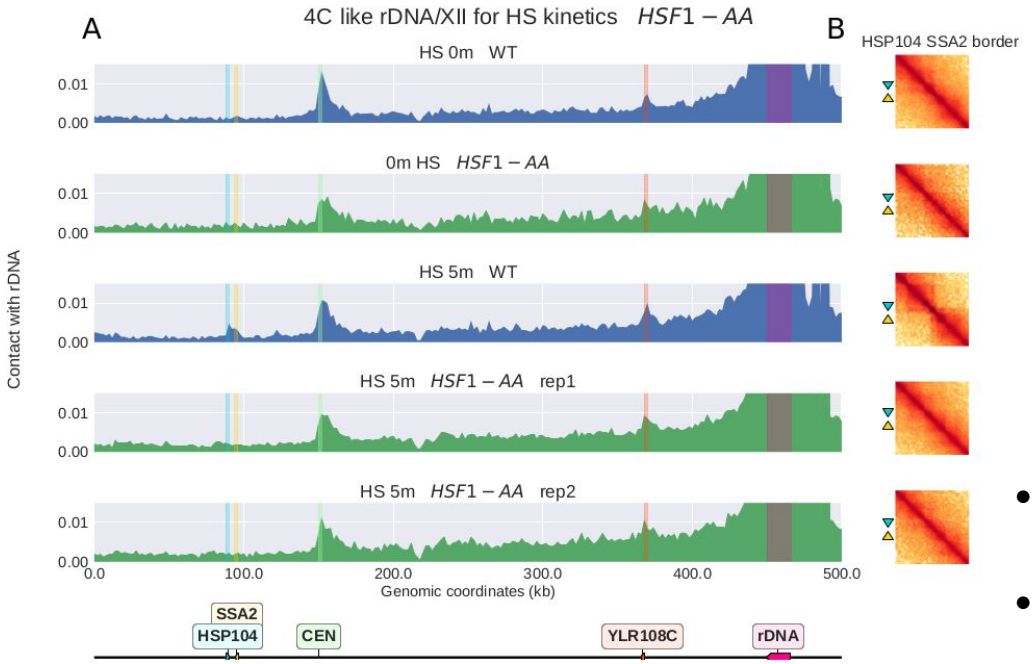


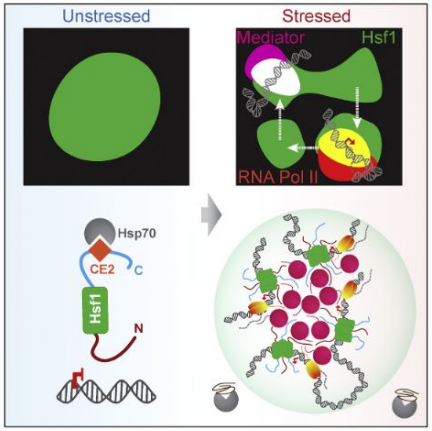
Retrotransposons (TYs) appear to be **new hotspots** of *2µ plasmid*

Gradit et al., in prep.

Unveiling the rDNA Network: Key Players and Choreographers of its Spatial Organization

Application : *S. cerevisiae* TYs as new 2µ preferencial contact spots





- **Hsf1 Mutant:** Transient rDNA-HSP104/SSA2 interaction and border pattern are lost.
- **Dual Requirement:** Both transcription and Hsf1 are essential for this stress-induced structure.
- **Mechanism:** Supports Chowdary's hypothesis of micro-aggregates, not loop extrusion, for contact formation.

Adapted from Chowdary, et al. 2022

HiC-BERG - Conclusion & Perspectives

- Hicberg showed **good performances** both in **simulated** and in **biological** data
- Several genomic features have been unveiled in *S. cerevisiae:*
    - 2µ plasmid **hotspots of contacts** of yeast retrotransposons
    - **New rDNA** interaction during **heat shock**, involving *Hsf1* protein complex

→ First link about spatial organization of rDNA and **stress**

Régulation Spatiale des Génomes

Romain Koszul
Axel Cournac
Martial Marbouty
Agnès Thierry
Hélène Bordelet
Jacques Serizay
Pauline Larrous
Devon Conti
Justine Groseille
Manon Perrot
Maëlys Delouis
Corina Pascal

Anciens membres

Cyril Matthey-Doret
Amaury Bignaud
Léa Meneu
Samuel Ortion

Collaborators

Bouk Wim De Jong

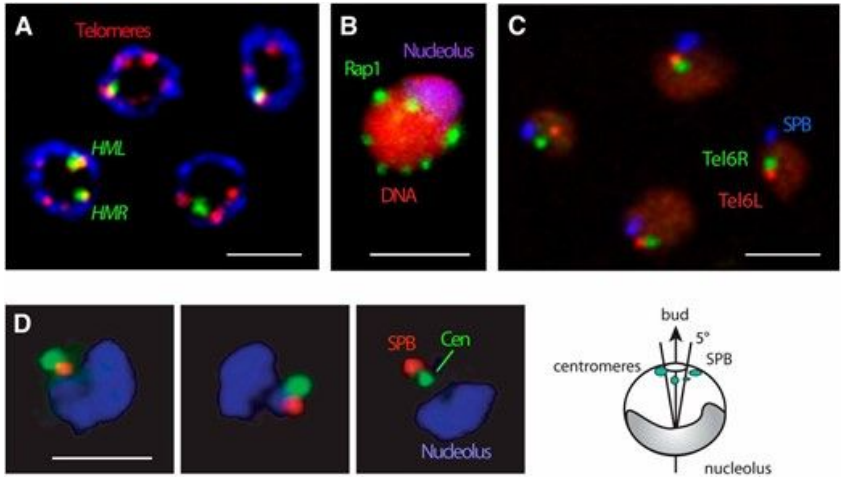
INCEPTION

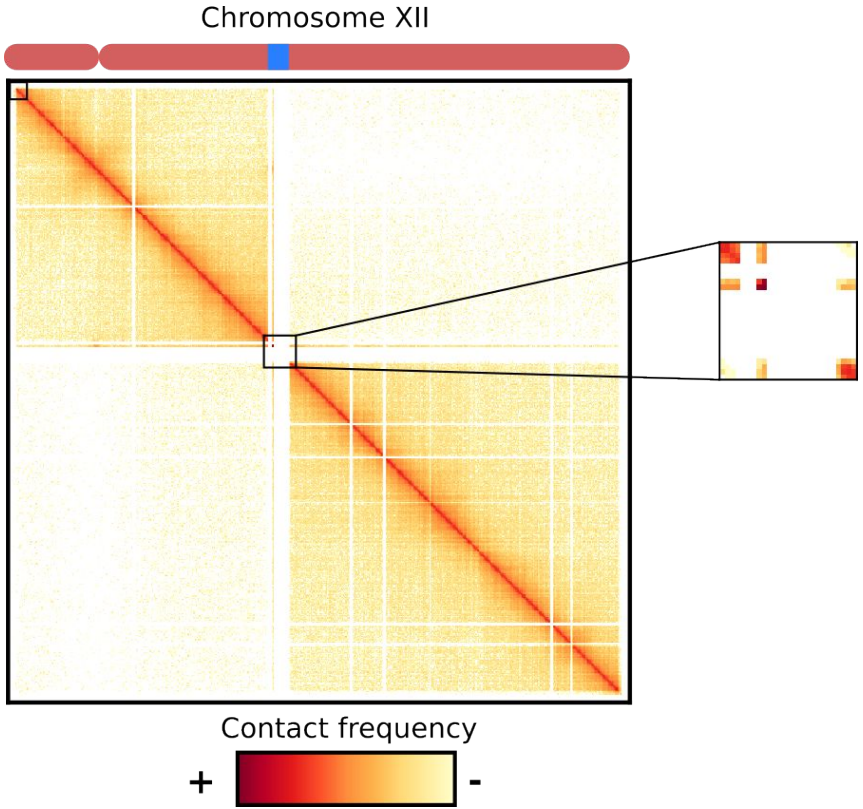
anr

INSTITUT PASTEUR

SORBONNE UNIVERSITÉ

rDNA: A Landmark in the Nuclear Landscape

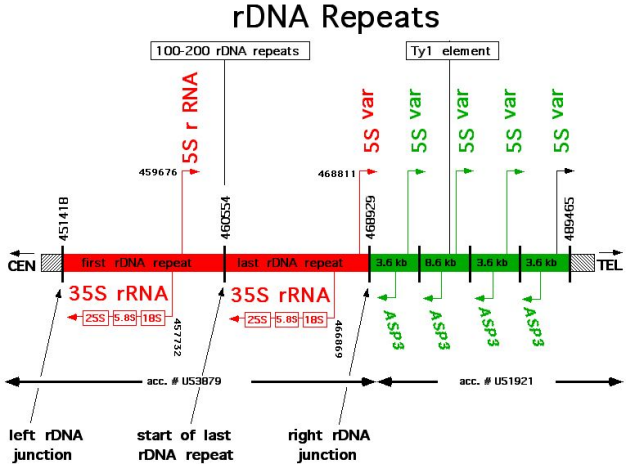




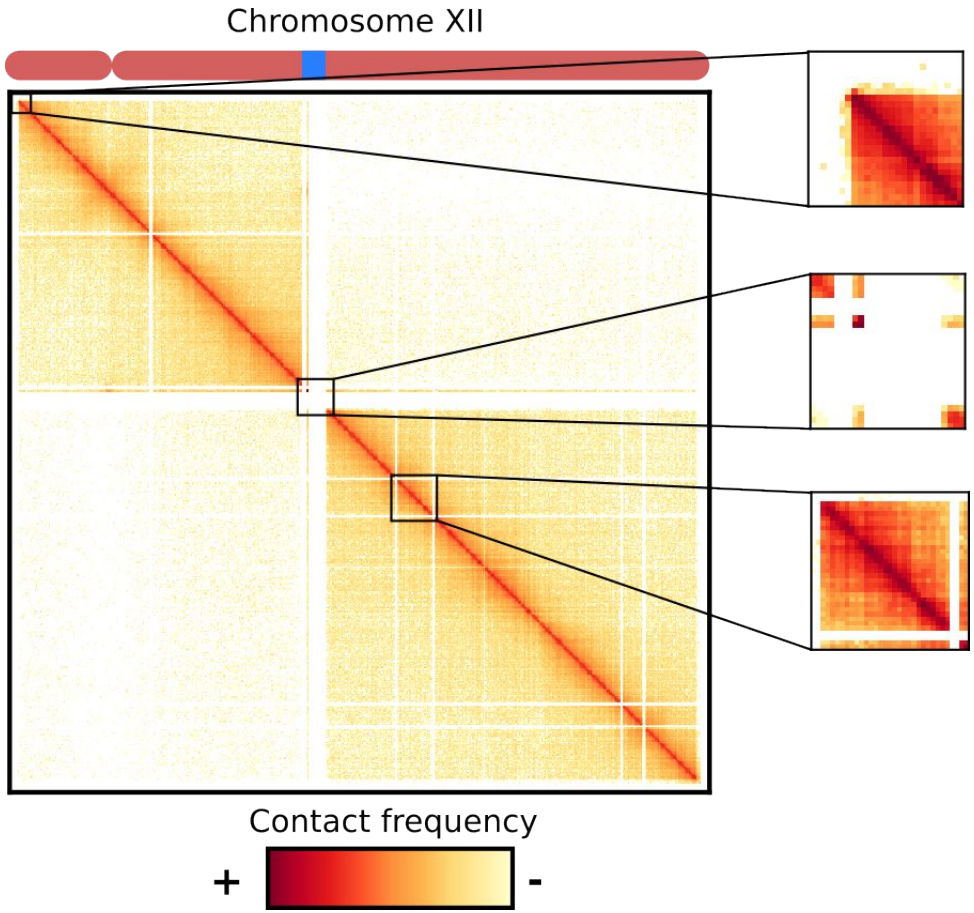
- rDNA as an organizational feature within the nucleus
- Ribosomes biogenesis
- Often exhibits interactions with nuclear landmarks (telomere, centromeres)

Adapted from Taddei, et al. 2012

The Challenge of Repeated Sequences

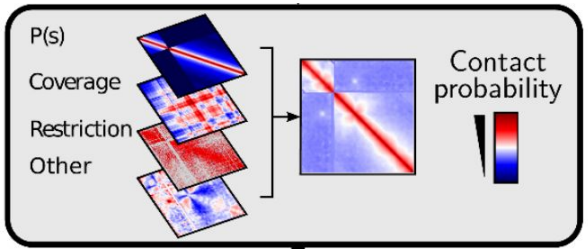


Why are repetitive sequences, particularly challenging to study using traditional Hi-C?

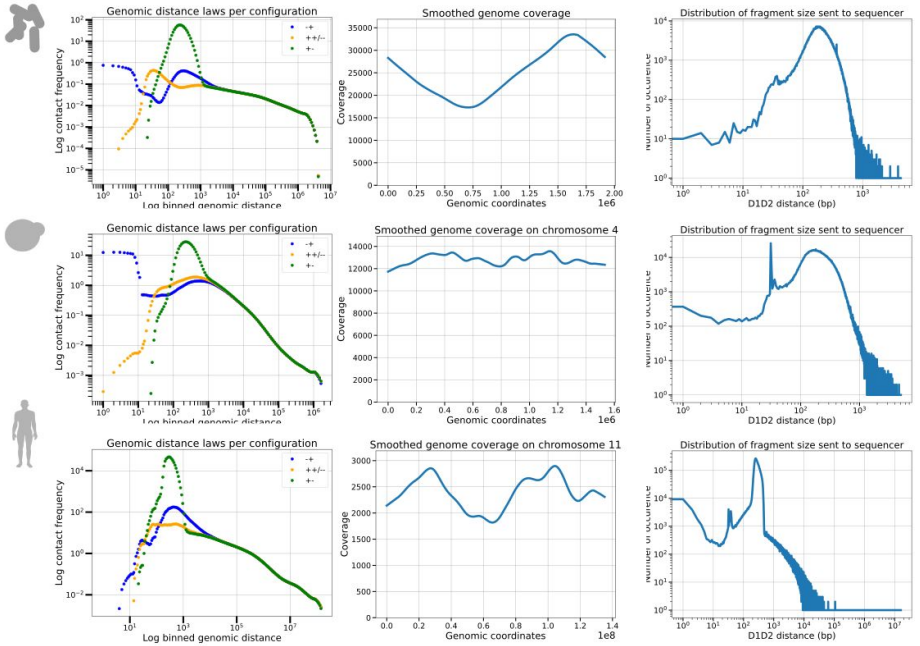


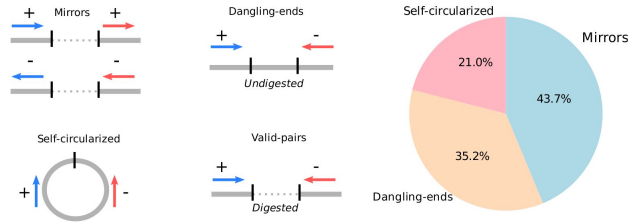
Adapted from Johnston, et al. 1997

Hicberg: a novel tool for (3D) genomics

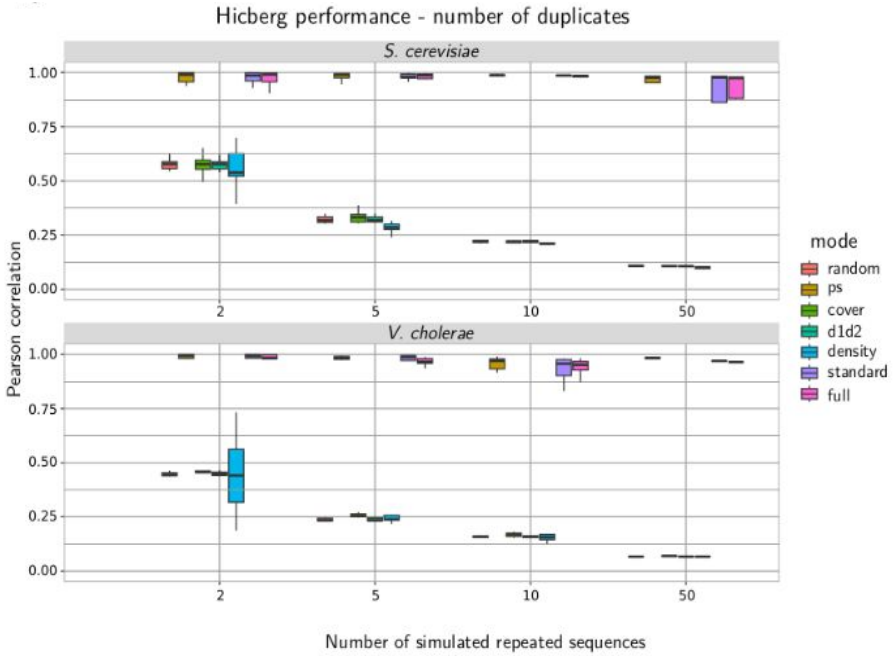


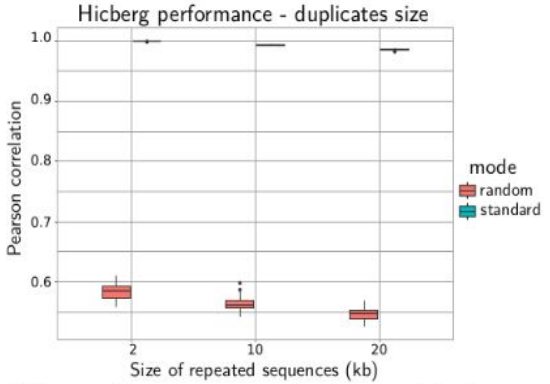
- **Polymer Behavior:** Captures DNA flexibility with 3 species/condition-specific sub-laws.
- **Coverage Bias:** Accounts for uneven read distribution; higher coverage increases interaction probability.
- **Restriction Site Proximity:** Accounts for fragment length biases.

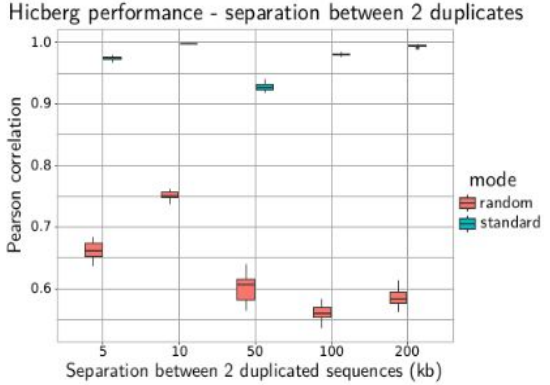




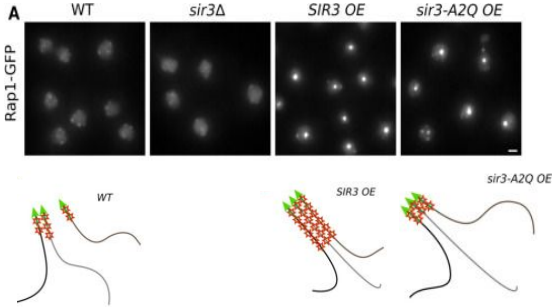
In-silico Evaluation

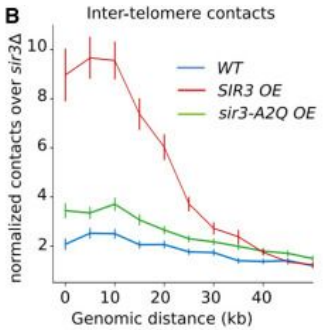




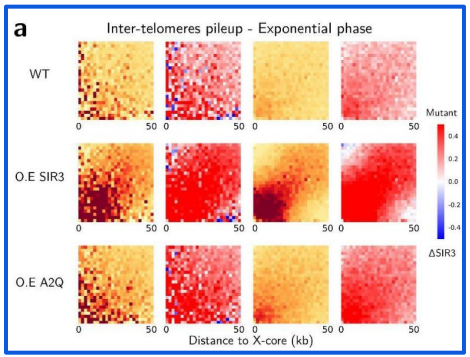


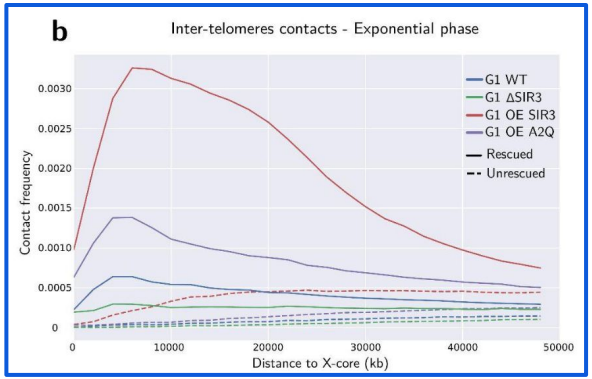
Biological validation





Hicberg to re-analyze yeast Hi-C data on telomere clustering under SIR3 over-expression studied with a diffusion-based method (Serpentine)



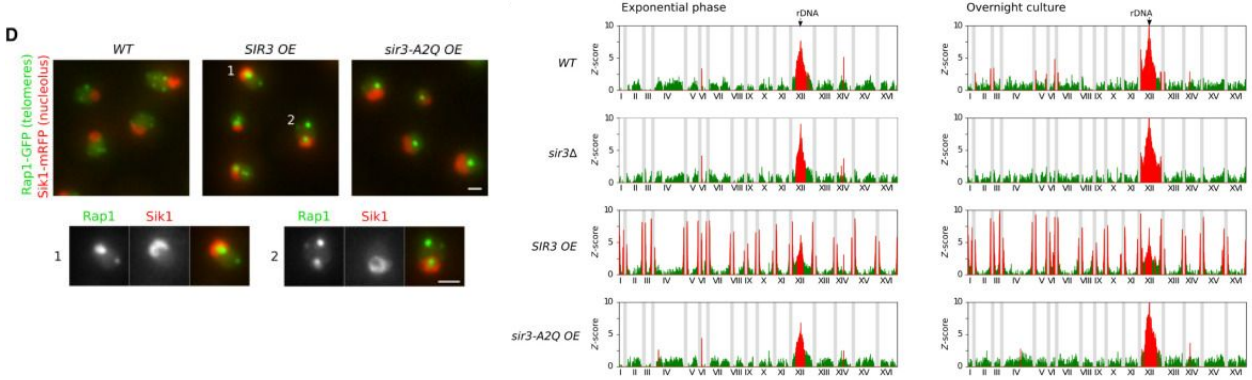


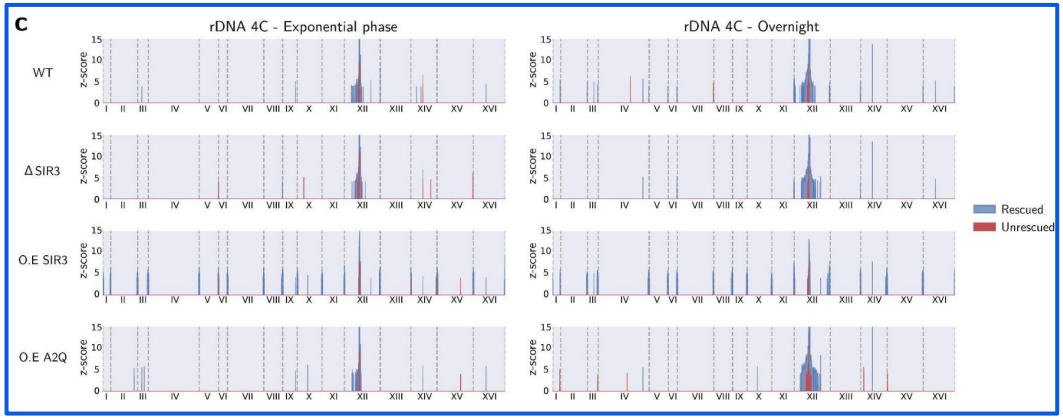
Hicberg reconstruction: Accurately reproduced Ruault's findings:

- SIR3 overexpression → telomere hyper-clustering.
- Altered SIR3 (A2Q) → reduced clustering.

Adapted from Ruault, et al. 2021

Biological validation





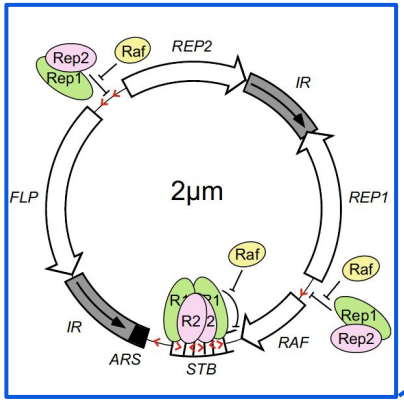
Hicberg reconstruction: Reproduced Ruault's findings:

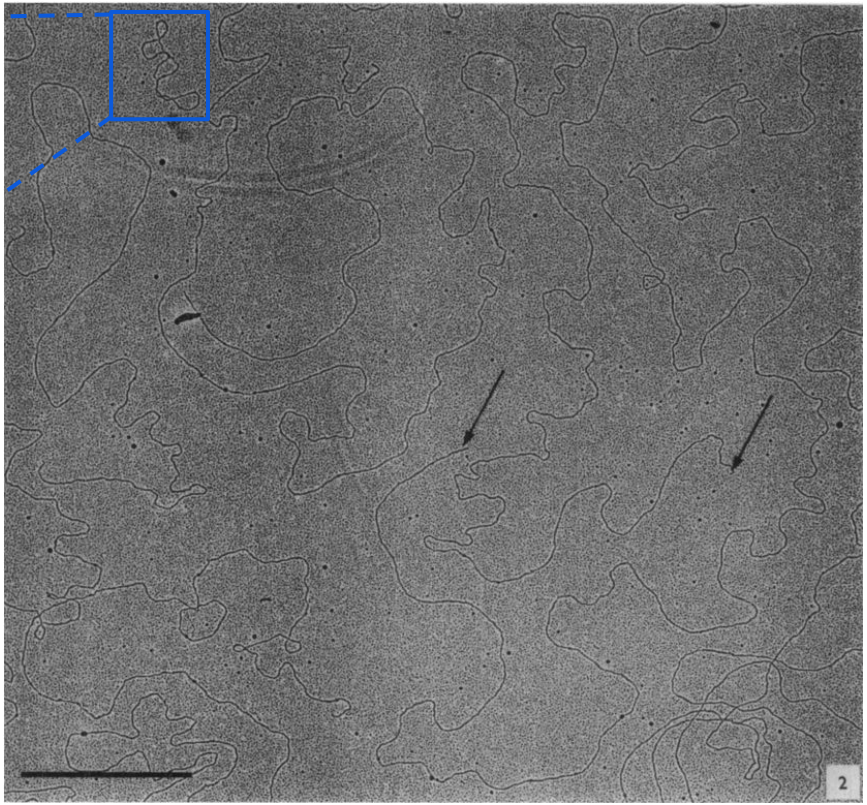
- SIR3 overexpression → increased rDNA-telomere clustering (exponential and overnight cultures).
- Altered SIR3 (A2Q) → partially reduced clustering.

Hicberg accurately captures complex, condition-specific 3D interactions in repeat rich regions, further validating its biological relevance.

Adapted from Ruault, et al. 2021

Discovery of 2µ plasmid in Saccharomyces cerevisiae





Observation in the 1970's of **enriched fraction of small circular DNA** found as different molecules species

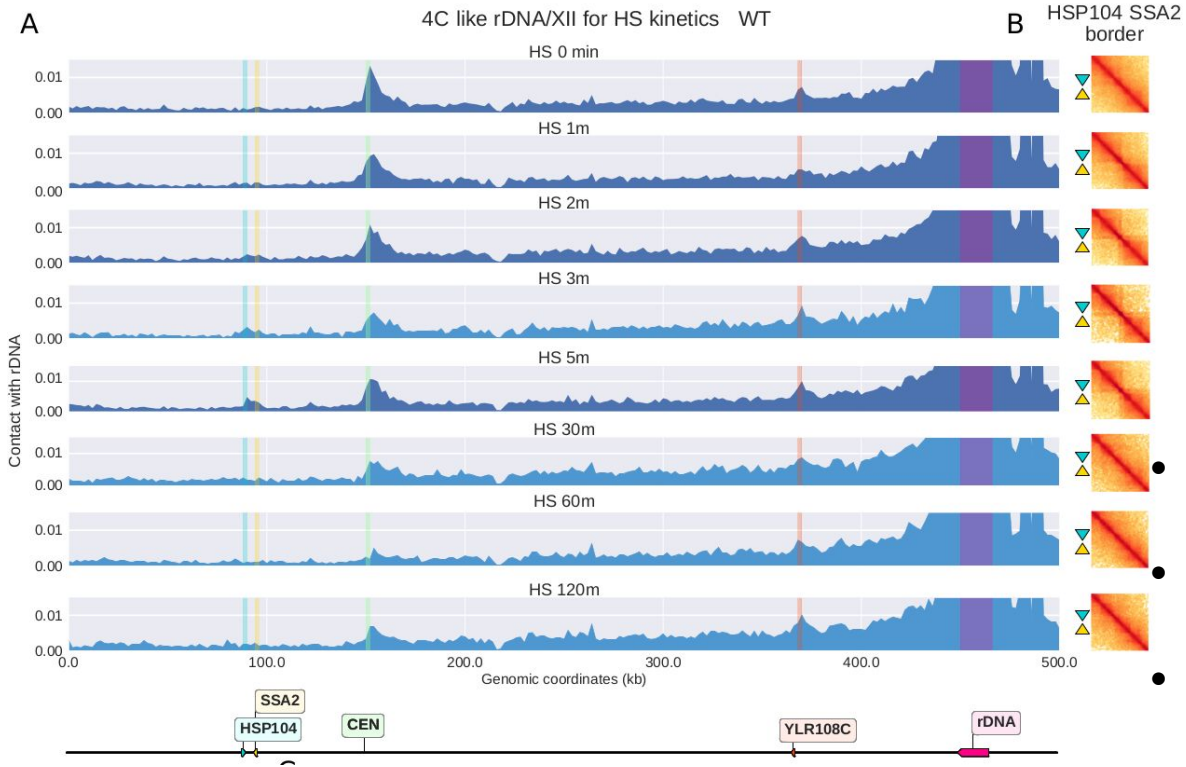
Predominant class of with **mean contour length of 1.88 ± 0.11 µm**

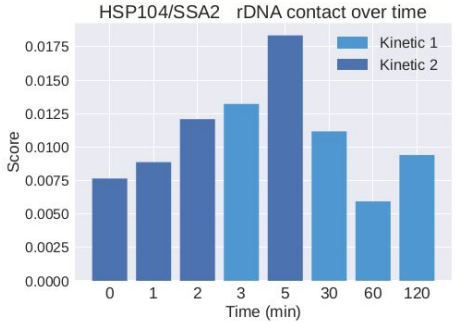
Conserved through generations

Specific locations in *S.cerevisiae* genome (long genes, low transcription spots)

Adapted from Girard, et al. 2023

rDNA-HSP104/SSA2: A Transient Tango During Heat Shock





- **rDNA in Heat Shock:** Hicberg reveals transient interaction with HSP104/SSA2 genes.
- **Dynamic contact:** Forms and dissolves during heat shock, suggesting a role in stress response.
- **Border pattern:** Emerges around HSP104/SSA2, indicating potential transcriptional activation.

HiC-BERG - Conclusion & Perspectives

- Hicberg showed good performances both in simulated and in context data
- Hicberg outperformed mHiC on the definition of reconstructed structures, and number of retrieved Hi-C reads
- Several genomic features have been unveiled (2µ hotspots on TYs, new rDNA interaction during heat shock, transcriptionally induced aggregates)
- First link about spatial organizationof rDNA and stress

Expand Applications:

- Explore diverse stress conditions in yeast.
- Apply to other organisms and datasets.
- Investigate various "OMICs" data.

Enhance Performance:

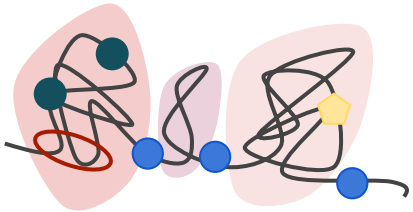
- Optimize for speed and efficiency.
- Facilitate routine use across species.

Chromosome folding influences or regulates dynamic processes?

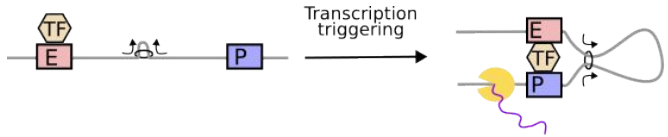
DNA binding proteins

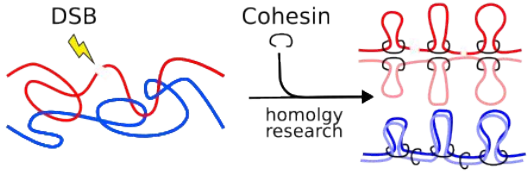
Physical constraints (supercoiling, etc.)

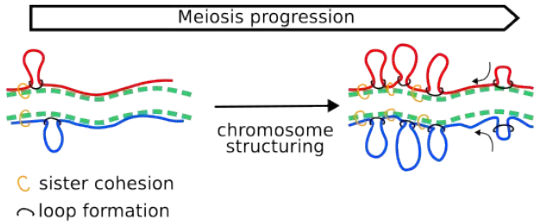
Non-coding RNA



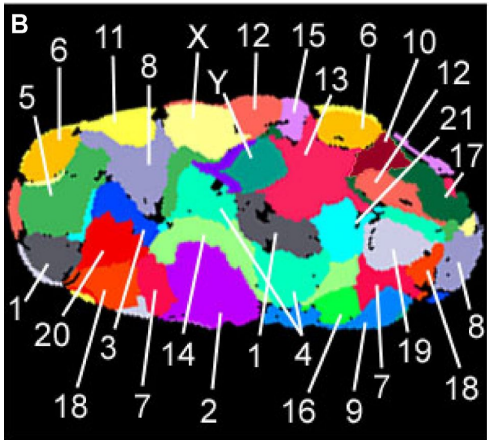
Cause or consequence ?

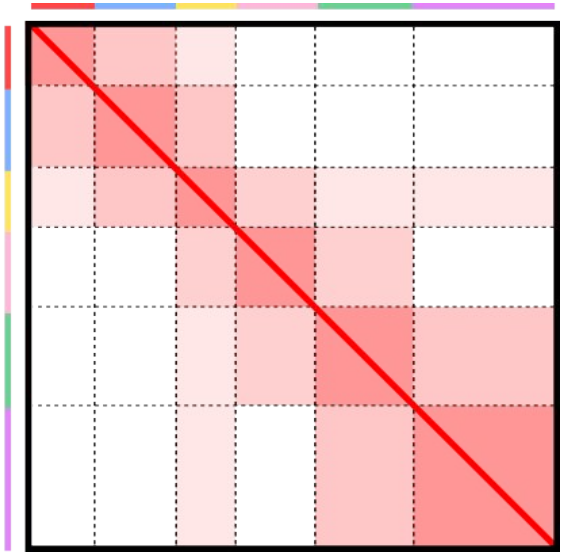






Chromosome Location in Genomes





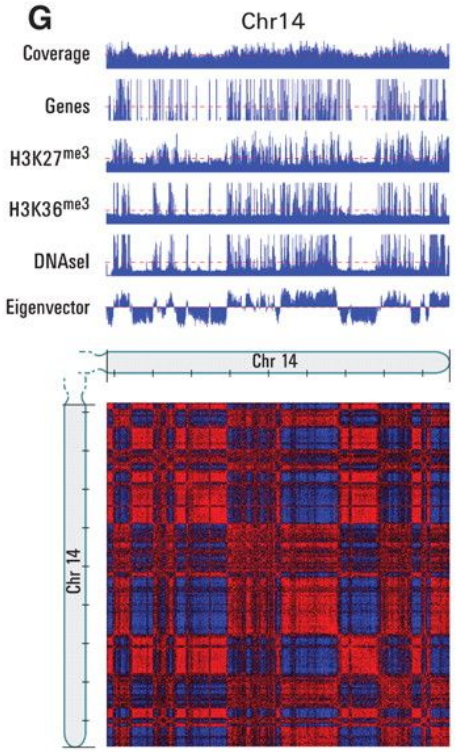
- Size/shape vary → chromosome, cell type
- Dynamic state (cellular processes, cycle, …)
- Gene rich toward interior

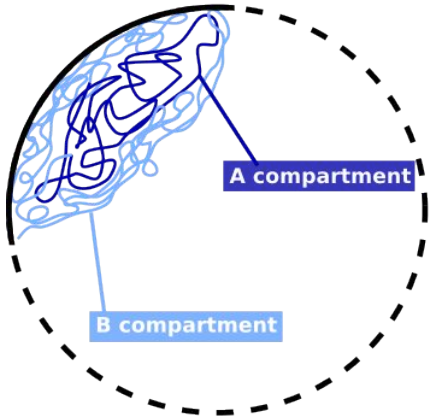
Response to biological processes:

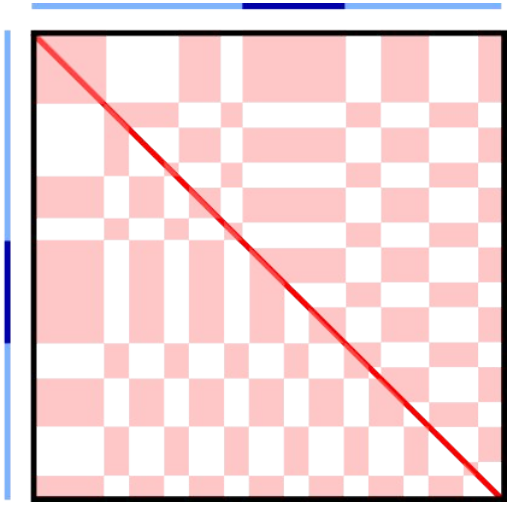
- ➢ Efficiency of DNA replication and repair
- ➢ Prevent inter-chromosomes translocations

Adapted from Bolzer, et al. 2005

From chromosomes to compartments







A compartments:

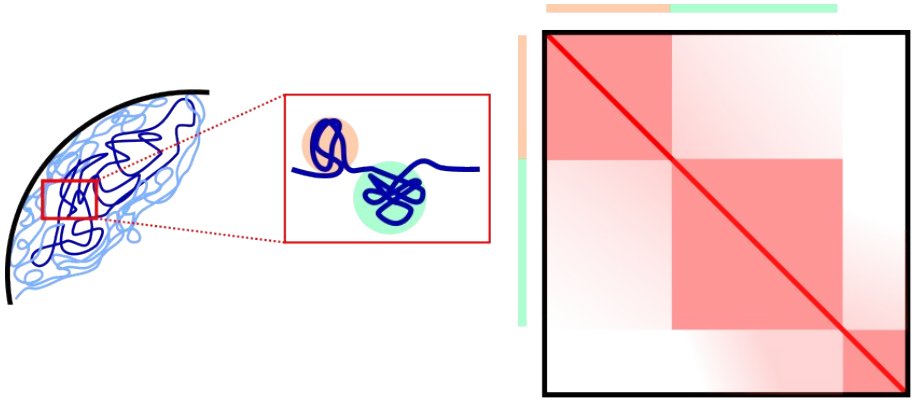
- Open chromatine
- Gene rich
- Acetylation marks
- Interior of nucleus

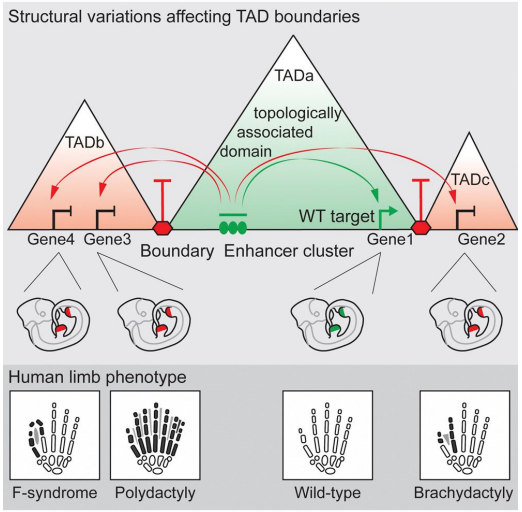
B compartments:

- Dense chromatine
- Gene devoid
- Methylation marks
- Toward lamina

Adapted from Lieberman-Aiden, et al. 2010

Topological Associated Domains





High level of self-interaction:

- Boundary regions insulator elements : separators
- Enriched for CTCF or architectural proteins

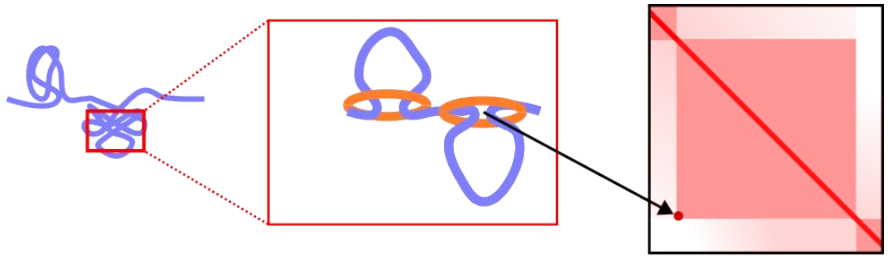
- Regularity units: appropriate enhancer/promoter
- Boundary disruption = inappropriate gene activation

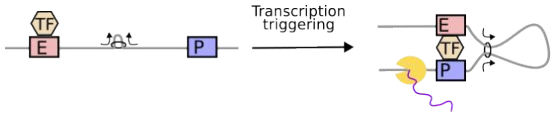
  → Disorders (development, cancer, …)

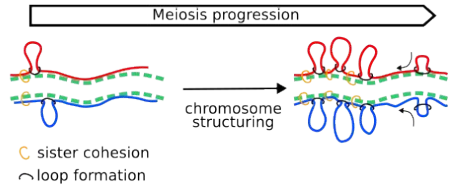
How do distant regulatory elements communicate with their target genes to control gene expression?

Adapted from Lupianez et al., 2015
Dixon et al., 2012

Chromatin loops: Bringing Elements Together







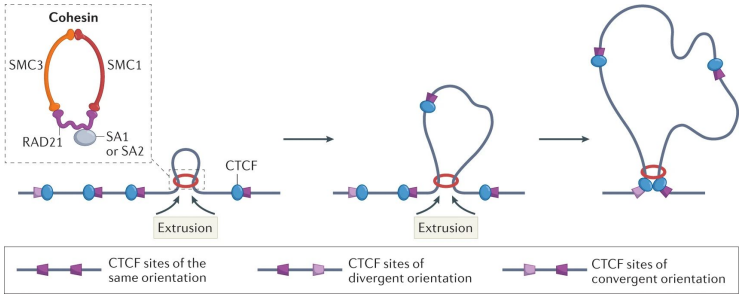
Regulatory loops

- Modulate gene expression
- Small range, gene expression precise control
- Dynamic: TF, chromatine remodelers, signaling

Structural loops

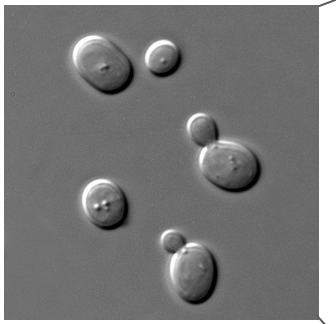
- Establish, maintain territories
- Long range, overall genome structuration
- More stable ⟶ slower timescale (cellular cycle)

Dual role, different scales but shared molecular mechanisms



Adapted from Zhang, et al. 2022

S. cerevisiae as a model for 3D genomics studies



- ~ 12 millions bp
- ~ 6000 genes
- Haploid and diploid states
- Easy to handle