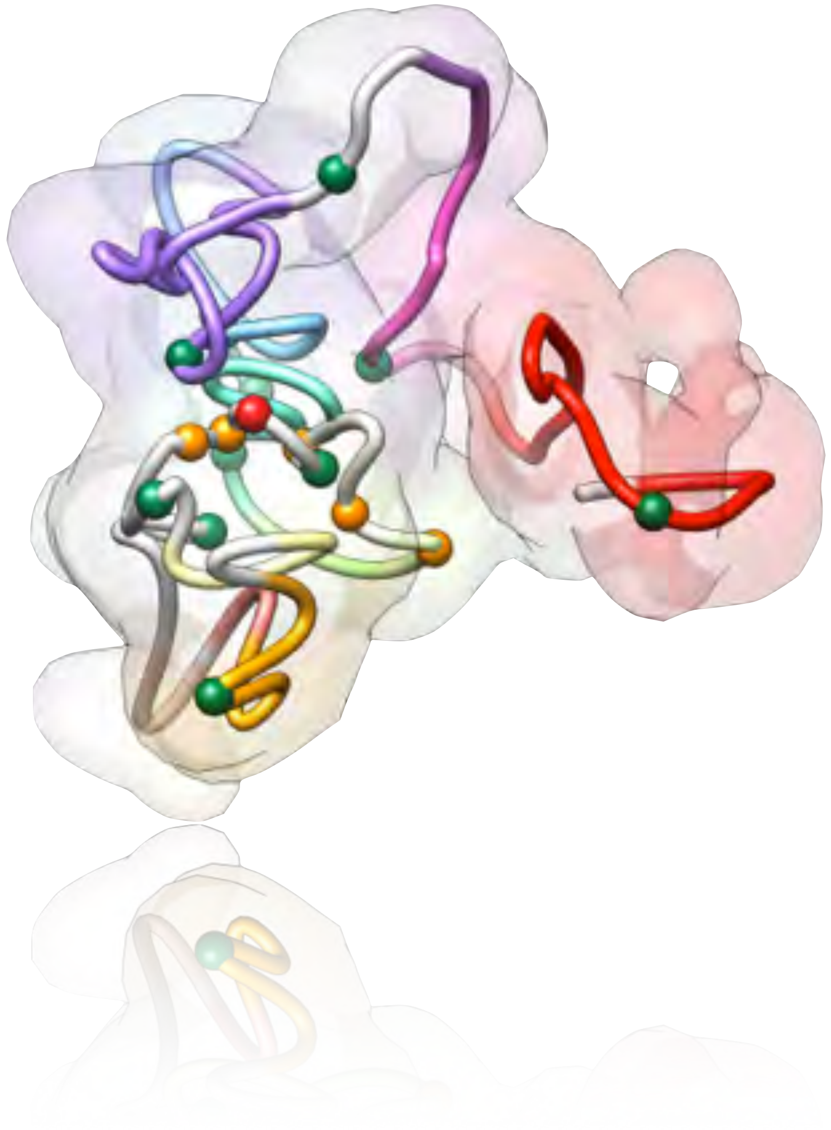
Data integration for 3D structure determination.

Marc A. Marti-Renom
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

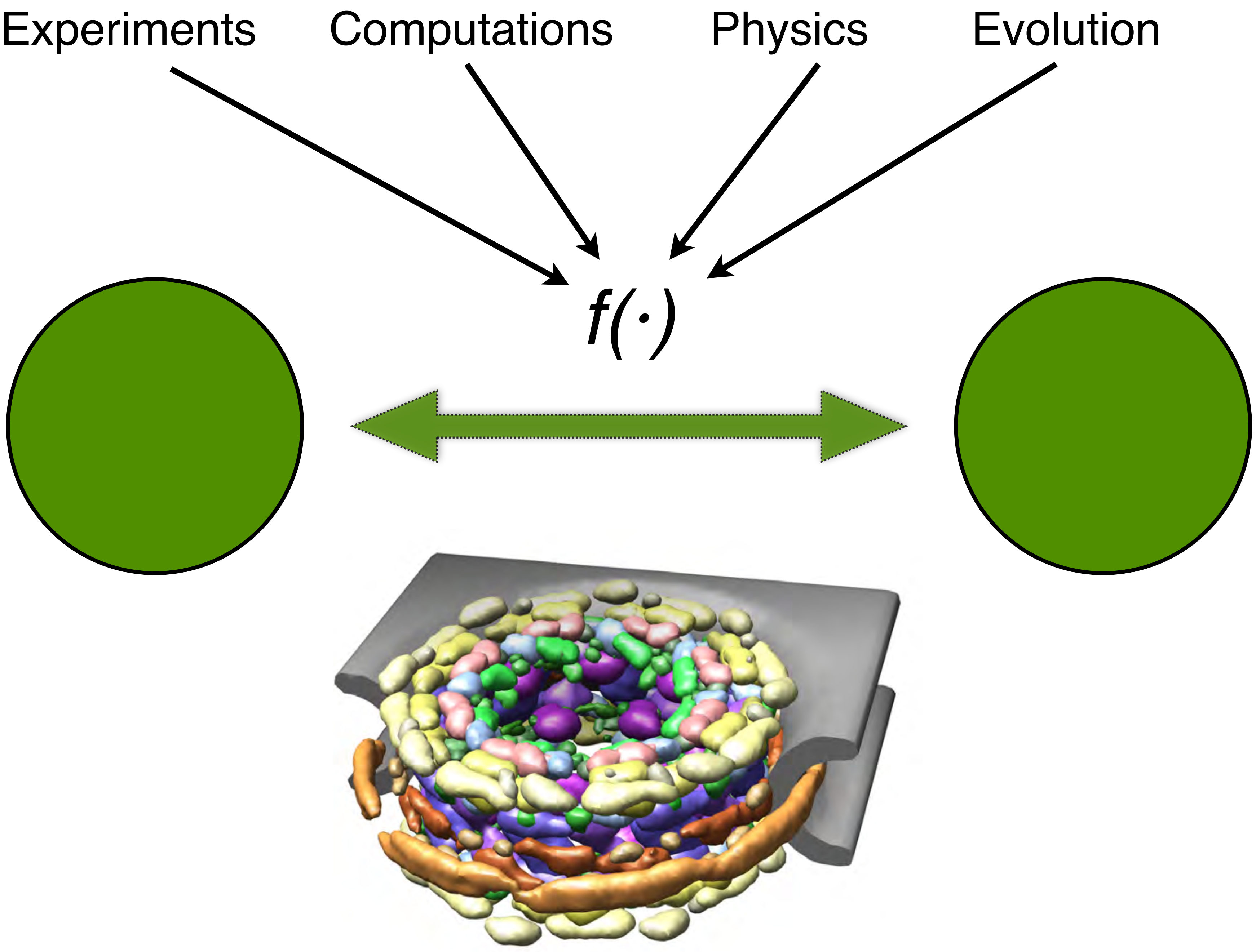
Structural Genomics Group

http://www.marciuslab.org



Integrative Modeling Platform

http://www.integrativemodeling.org



Experiments

Computations

Physics

Evolution

$f(\cdot)$

From: Russel, D. et al. PLOS Biology 10, e1001244 (2012).

Stages

Stage 1: Gathering Information. Information is collected in the form of data from wet lab experiments, as well as statistical tendencies such as atomic statistical potentials, physical laws such as molecular mechanics force fields, and any other feature that can be converted into a score for use to assess features of a structural model.

Stage 2: Choosing How To Represent And Evaluate Models. The resolution of the representation depends on the quantity and resolution of the available information and should be commensurate with the resolution of the final models: different parts of a model may be represented at different resolutions, and one part of the model may be represented at several different resolutions simultaneously. The scoring function evaluates whether or not a given model is consistent with the input information, taking into account the uncertainty in the information.

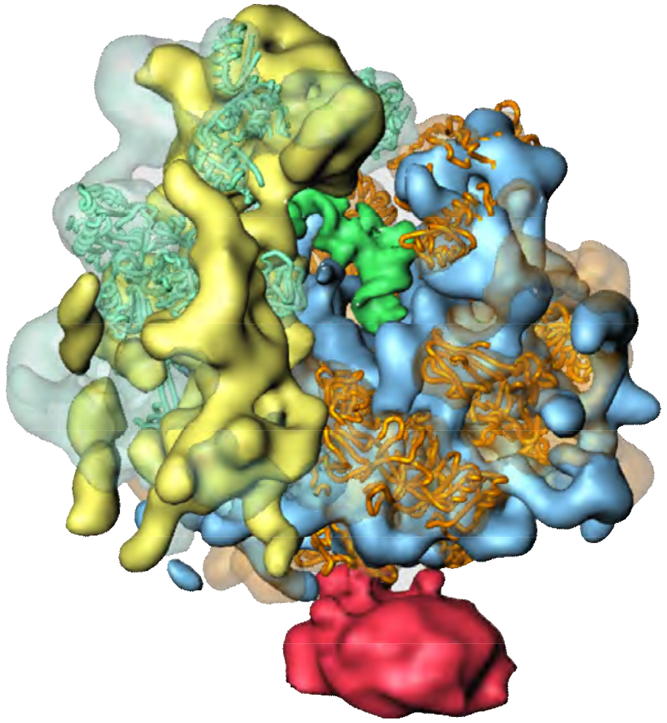
Stage 3: Finding Models That Score Well. The search for models that score well is performed using any of a variety of sampling and optimization schemes (such as the Monte Carlo method). There may be many models that score well if the data are incomplete or none if the data are inconsistent due to errors or unconsidered states of the assembly.

Stage 4: Analyzing Resulting Models and Information. The ensemble of good-scoring models needs to be clustered and analyzed to ascertain their precision and accuracy, and to check for inconsistent information. Analysis can also suggest what are likely to be the most informative experiments to perform in the next iteration.

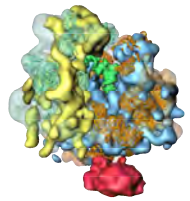
Integrative modeling iterates through these stages until a satisfactory model is built. Many iterations of the cycle may be required, given the need to gather more data as well as to resolve errors and inconsistent data.

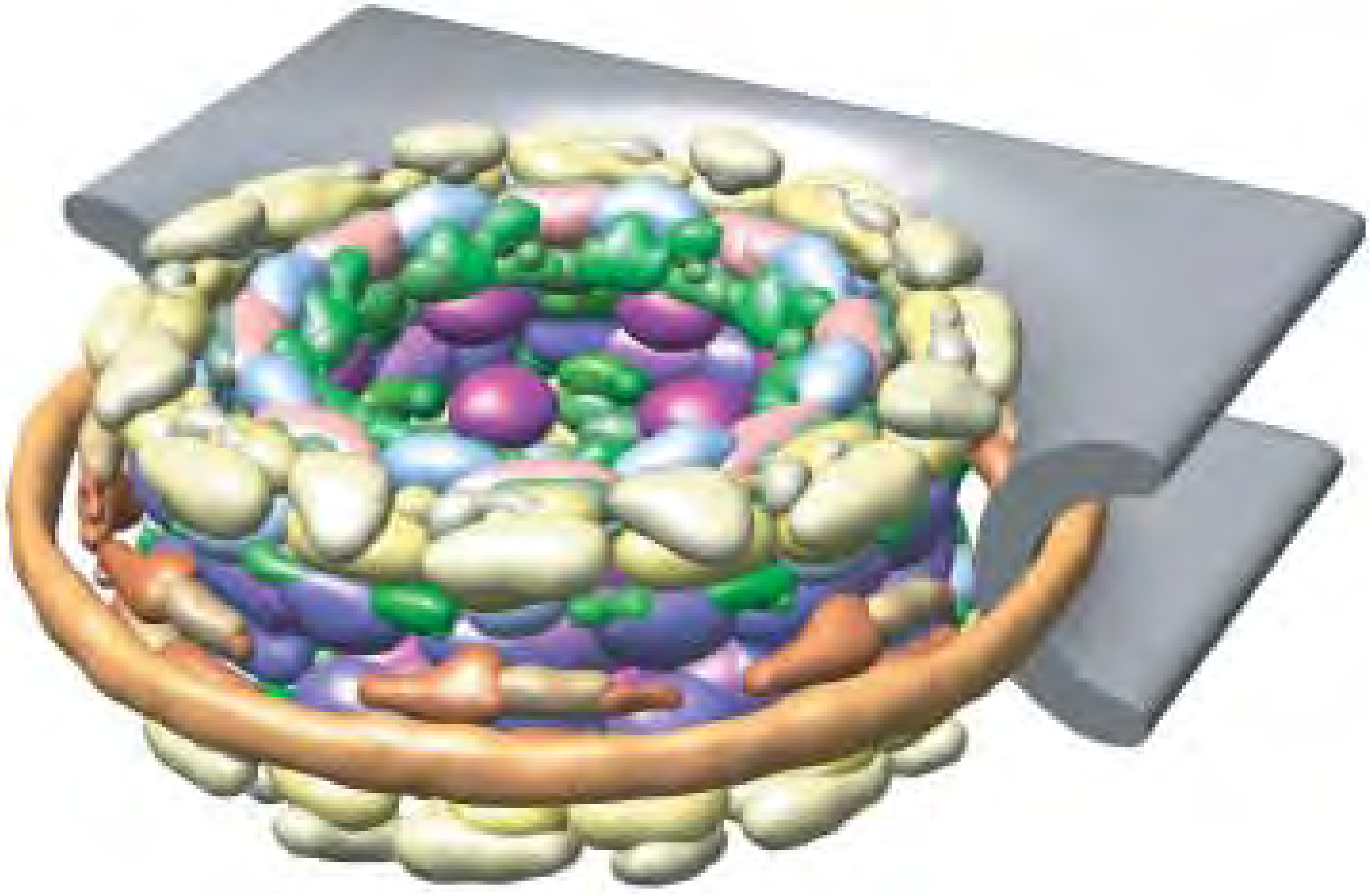
Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

Data Integration

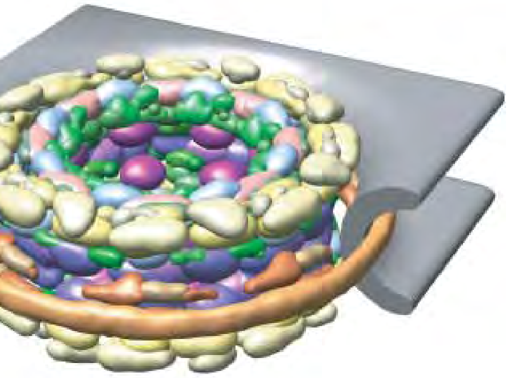


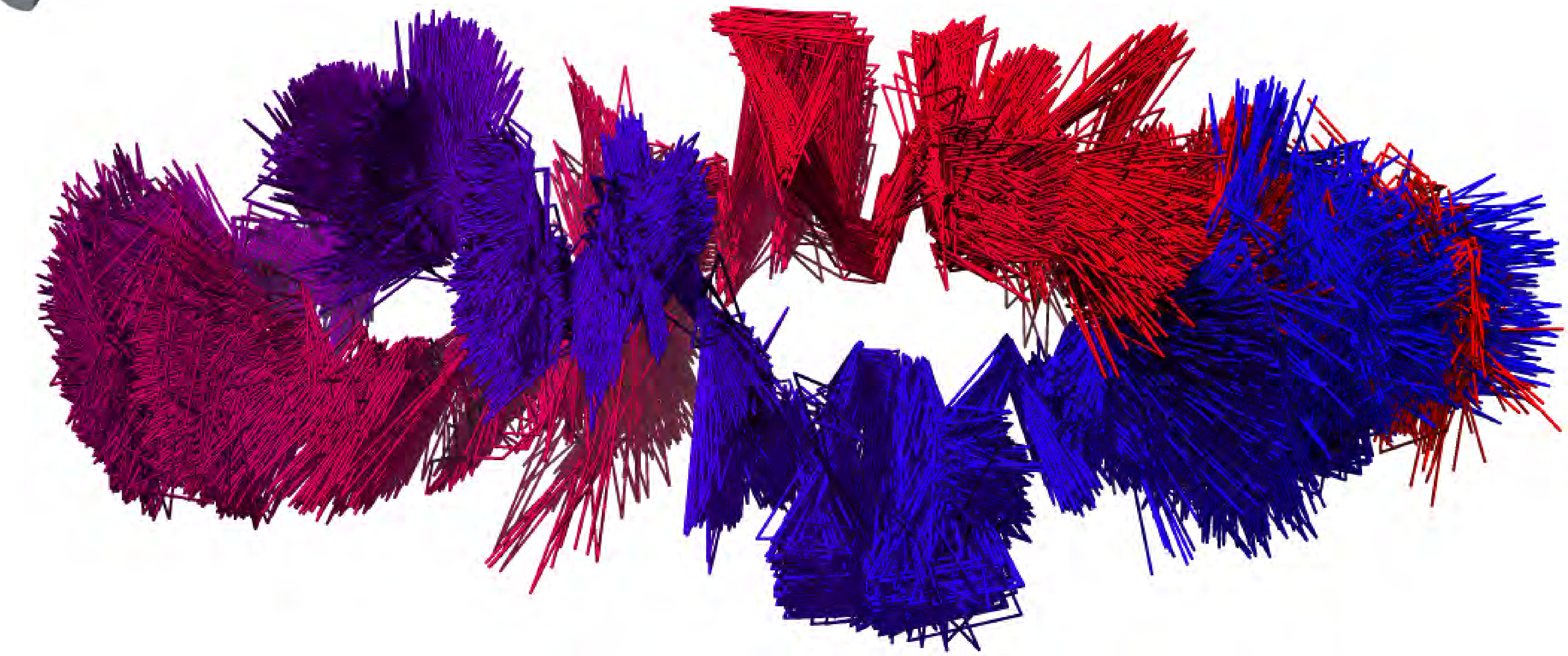
Data Integration



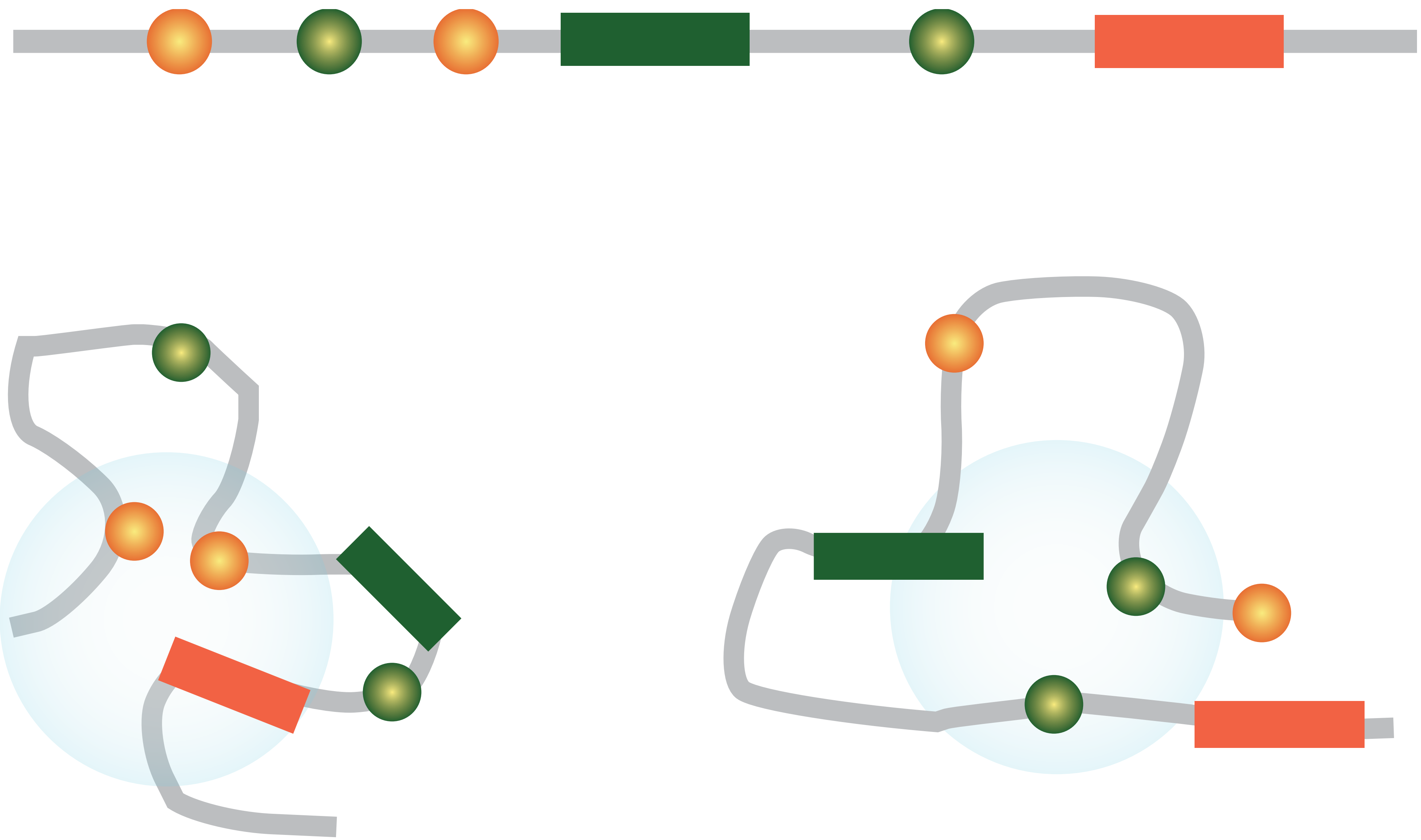


Data Integration



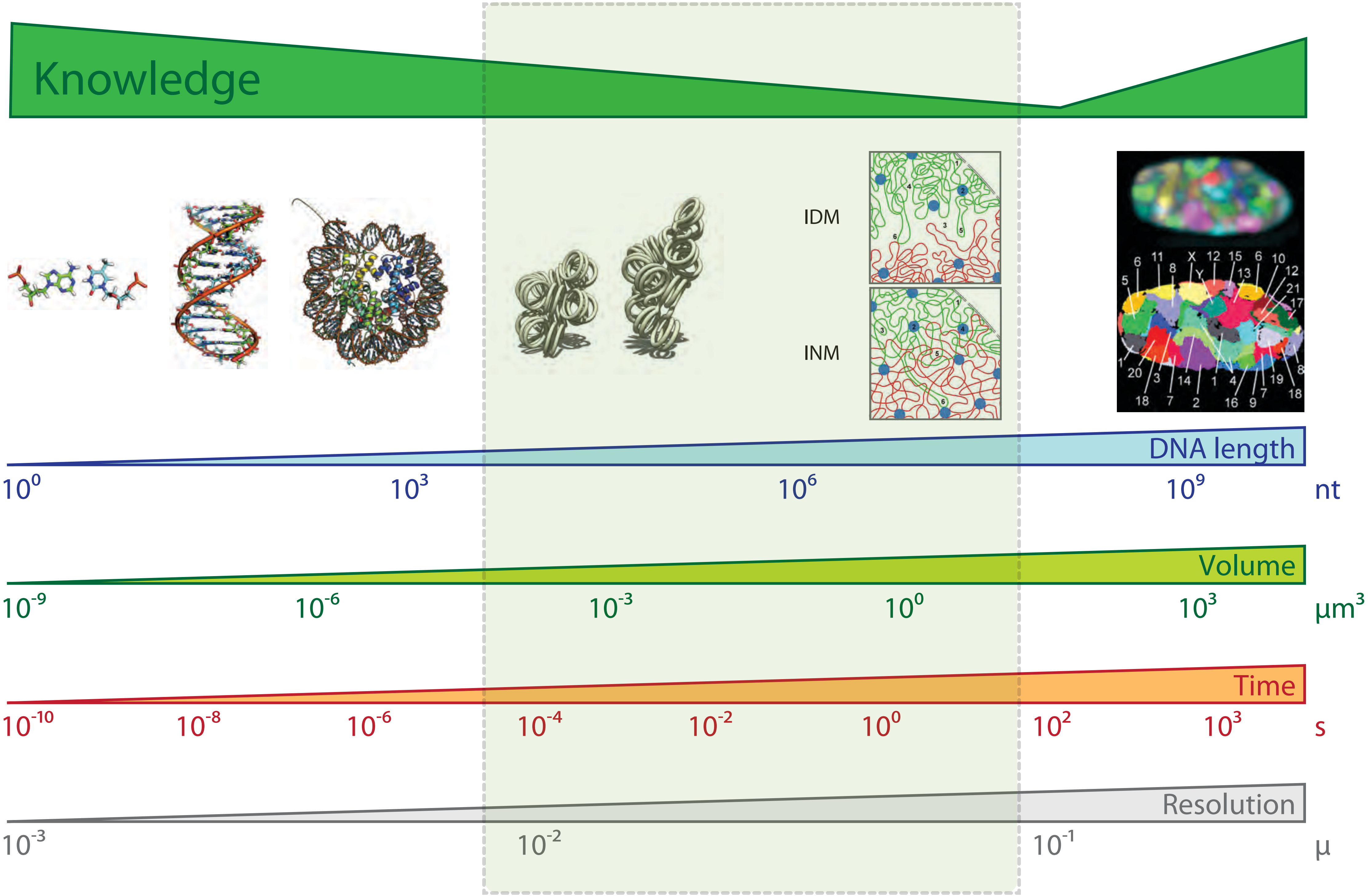


Complex genome organization



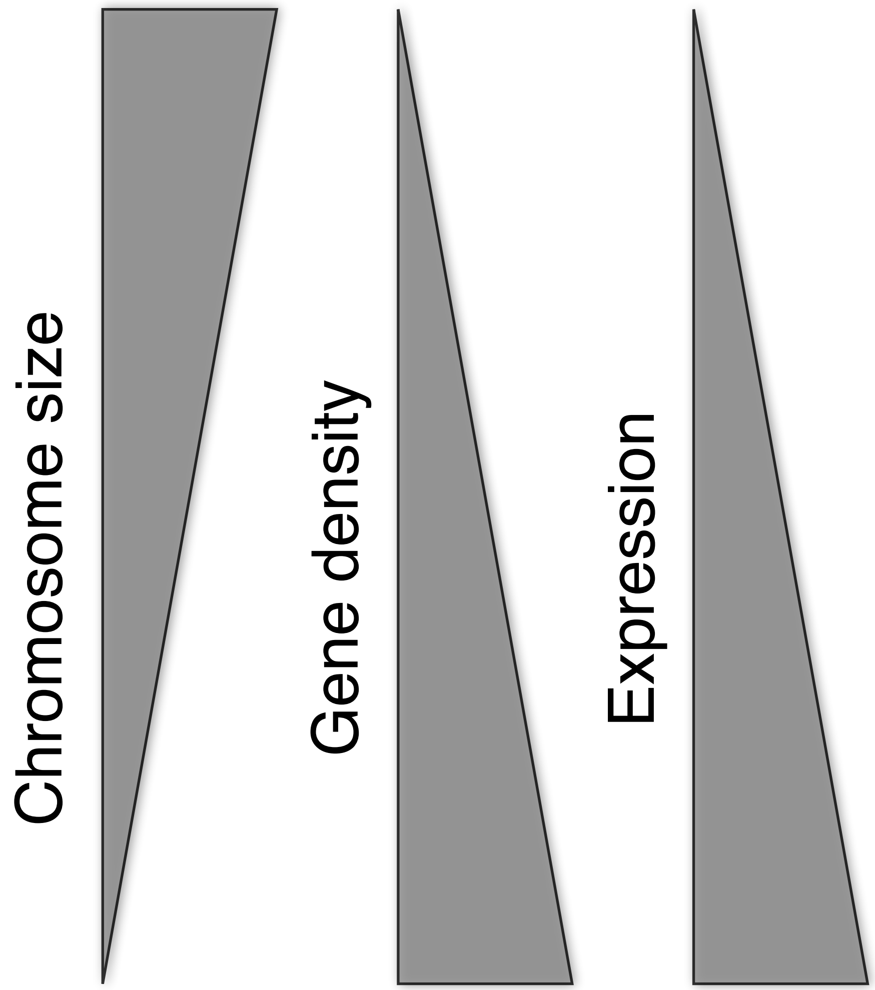
Resolution Gap

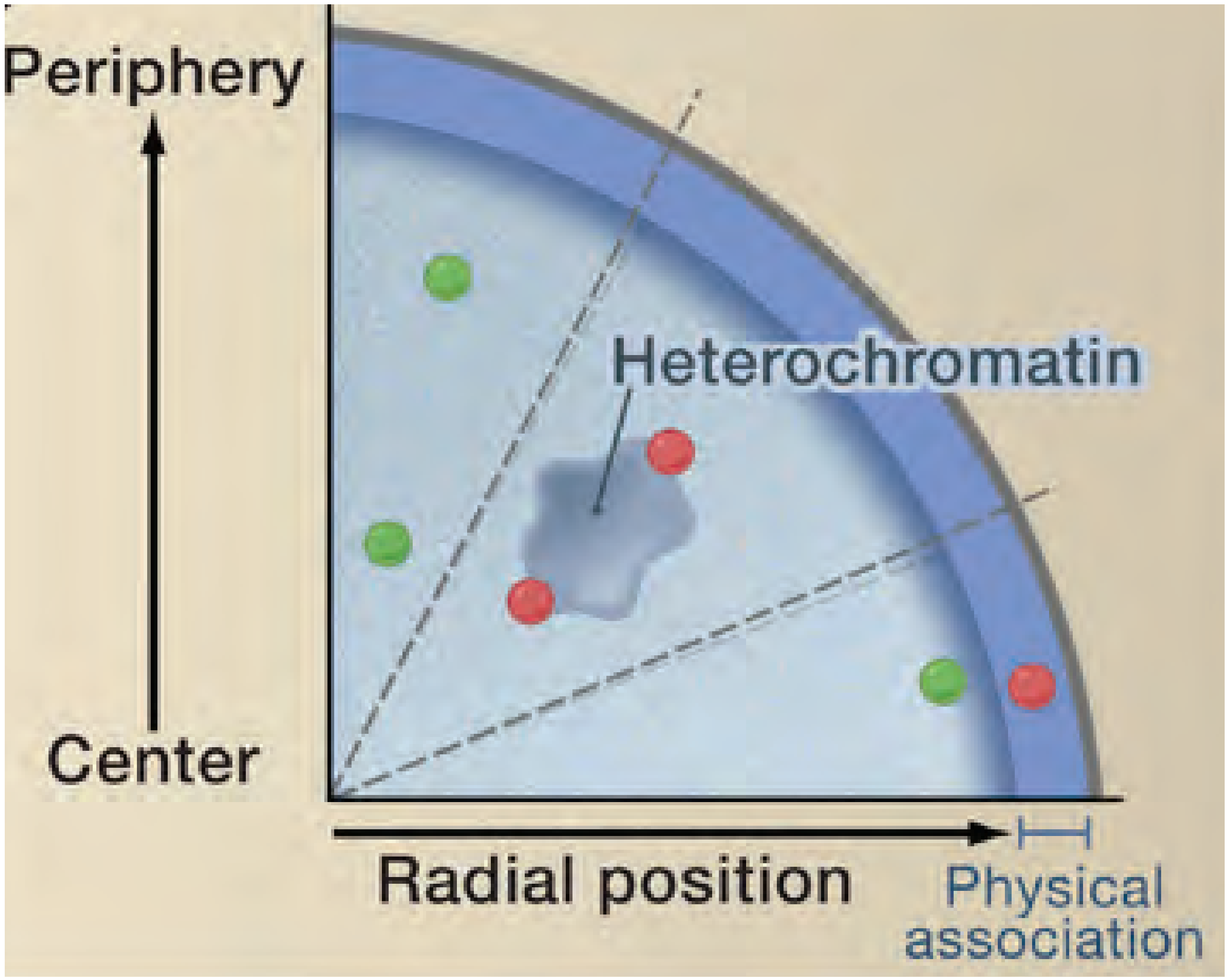
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)



Radial organization of the genome

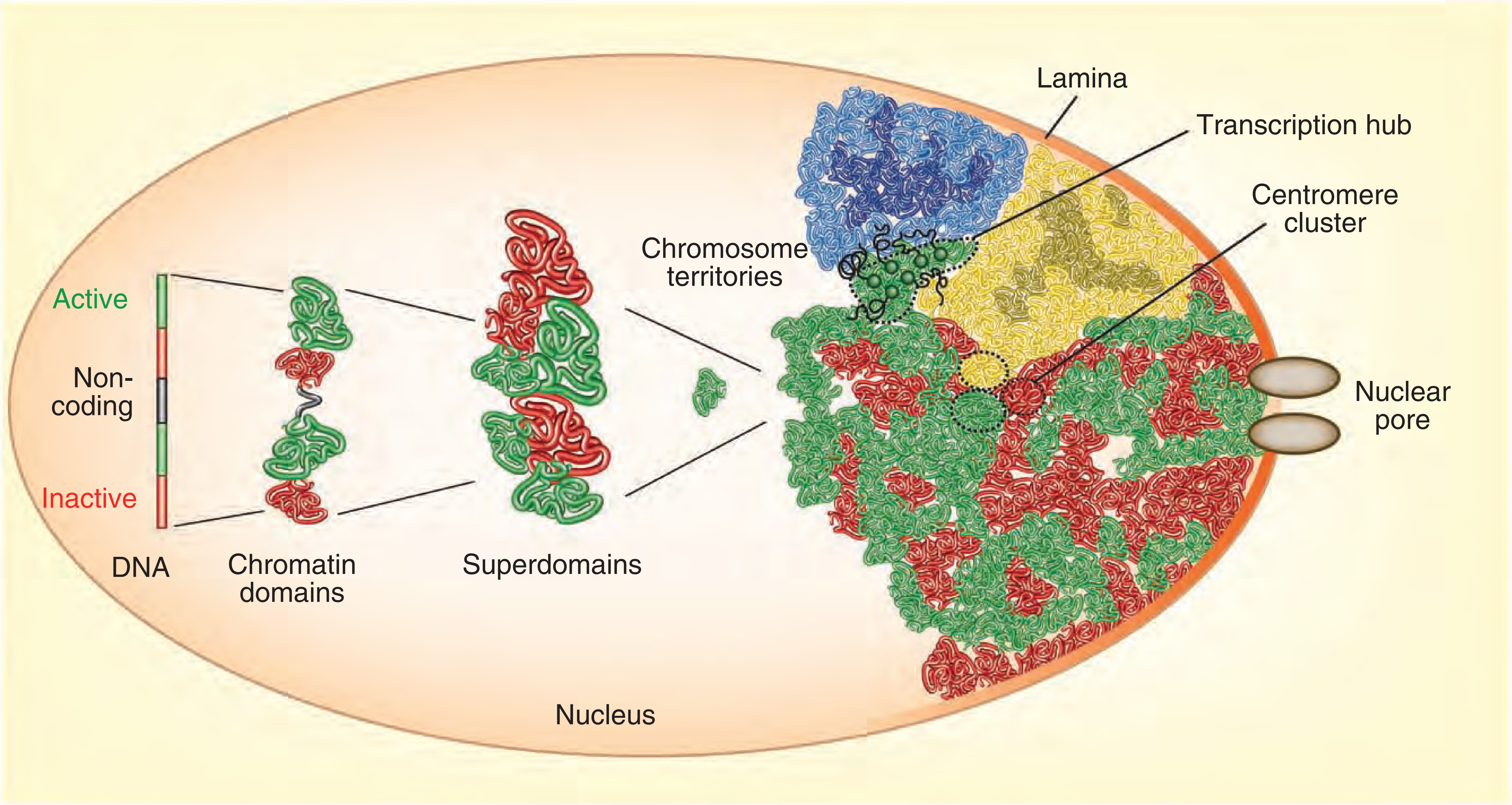
Takizawa, T., Meaburn, K. J. & Misteli, T. The meaning of gene positioning. Cell 135, 9–13 (2008).





Complex genome organization

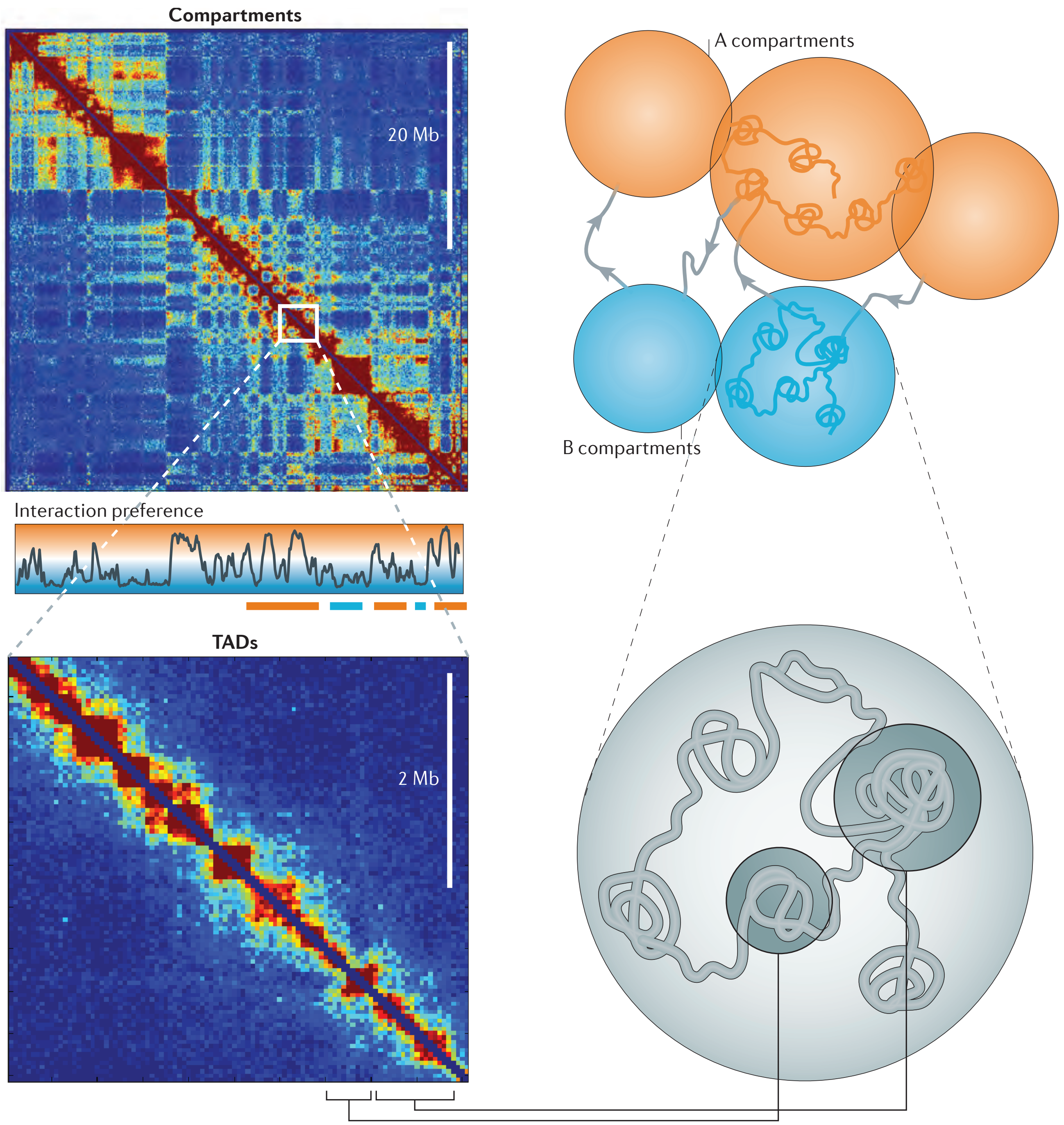
Cavalli, G. & Misteli, T. Functional implications of genome topology. Nat Struct Mol Biol 20, 290–299 (2013).



Marina Corral

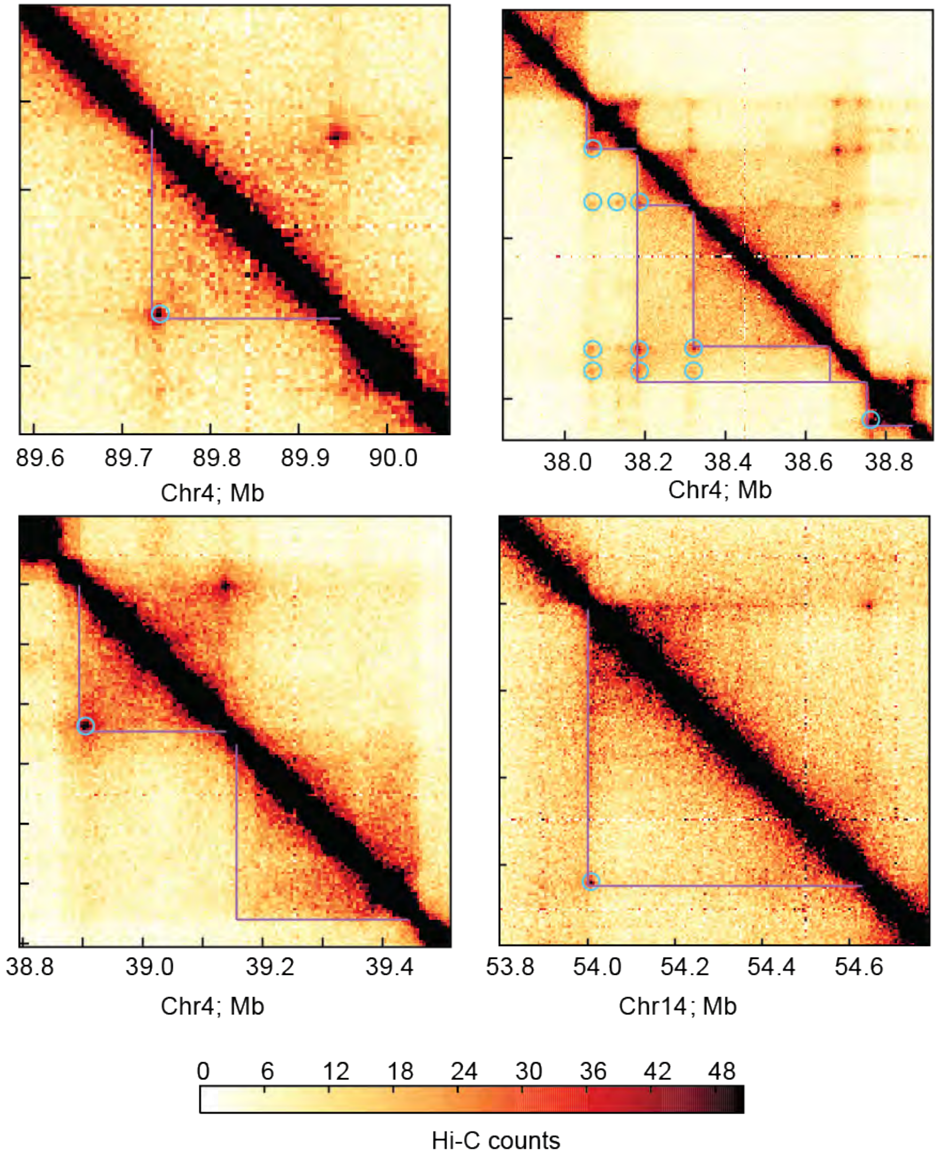
A/B Compartments and TADs

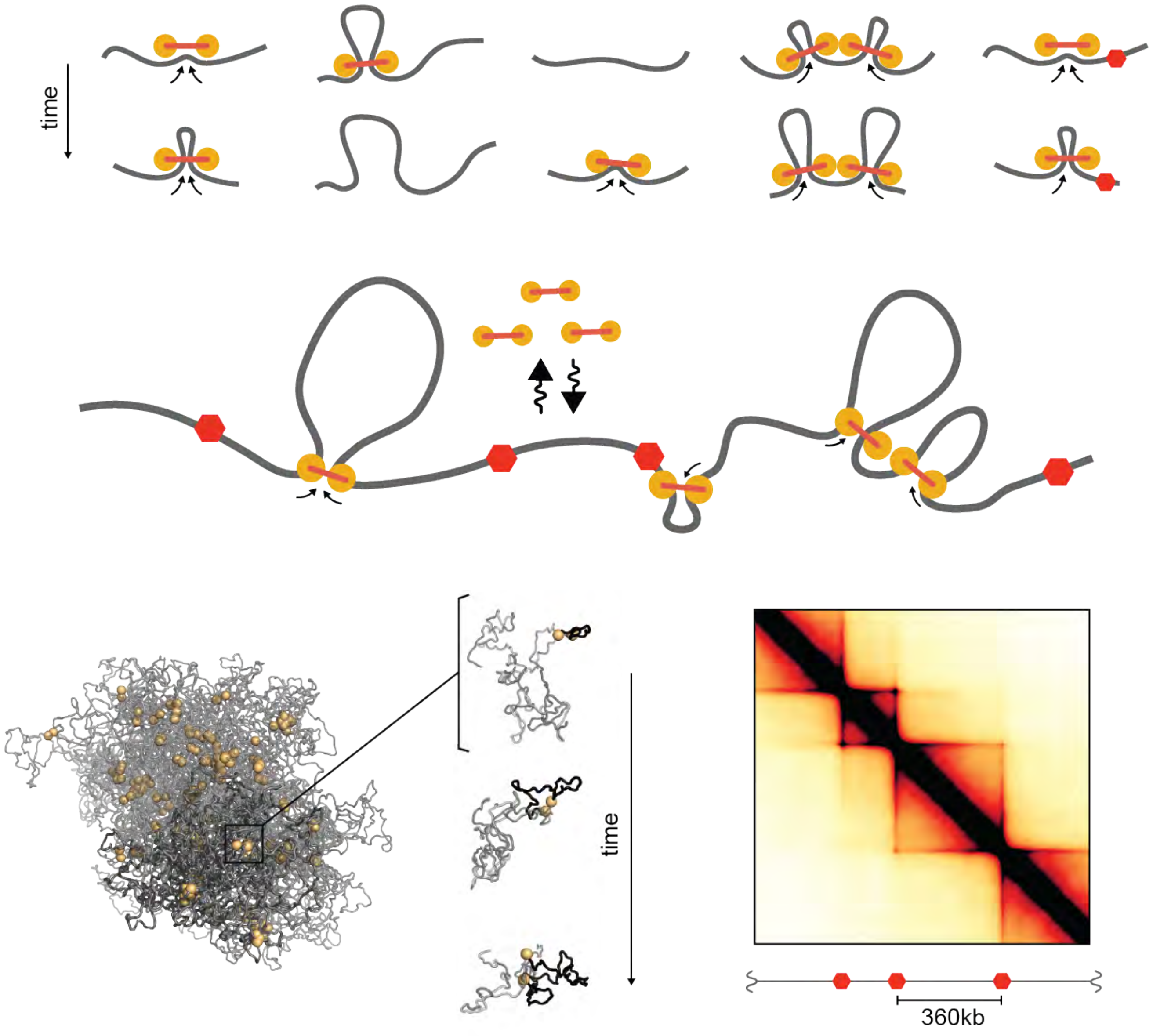
Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390–403 (2013).



Loop-extrusion as a driving force

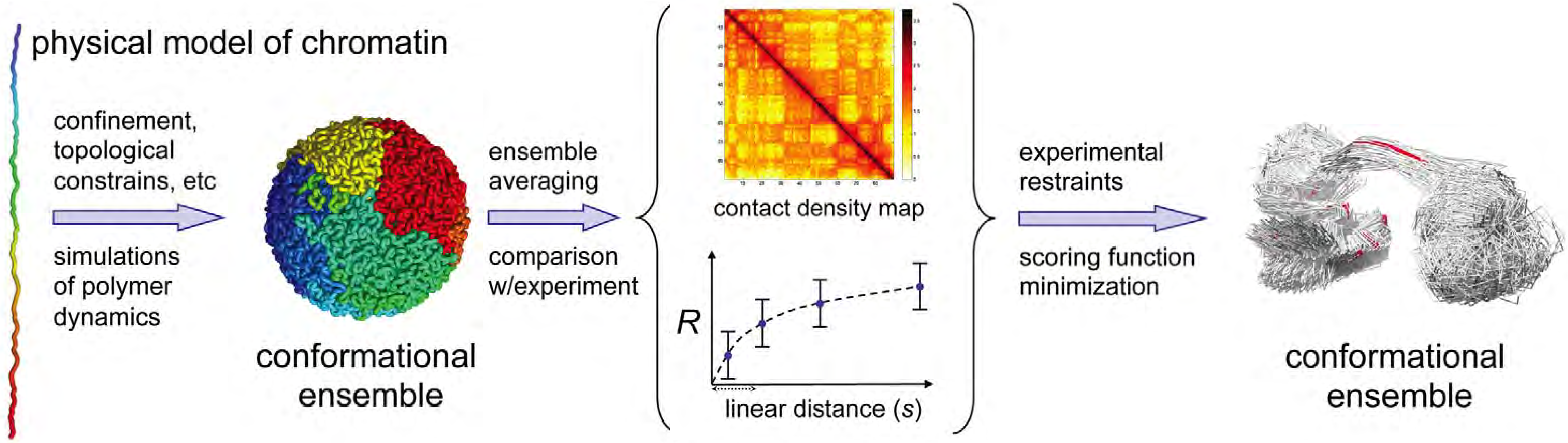
Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2015). Formation of Chromosomal Domains by Loop Extrusion. bioRxiv.





Modeling Genomes

Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)



physical model of chromatin

confinement, topological constrains, etc

simulations of polymer dynamics

conformational ensemble

ensemble averaging

comparison w/experiment

contact density map

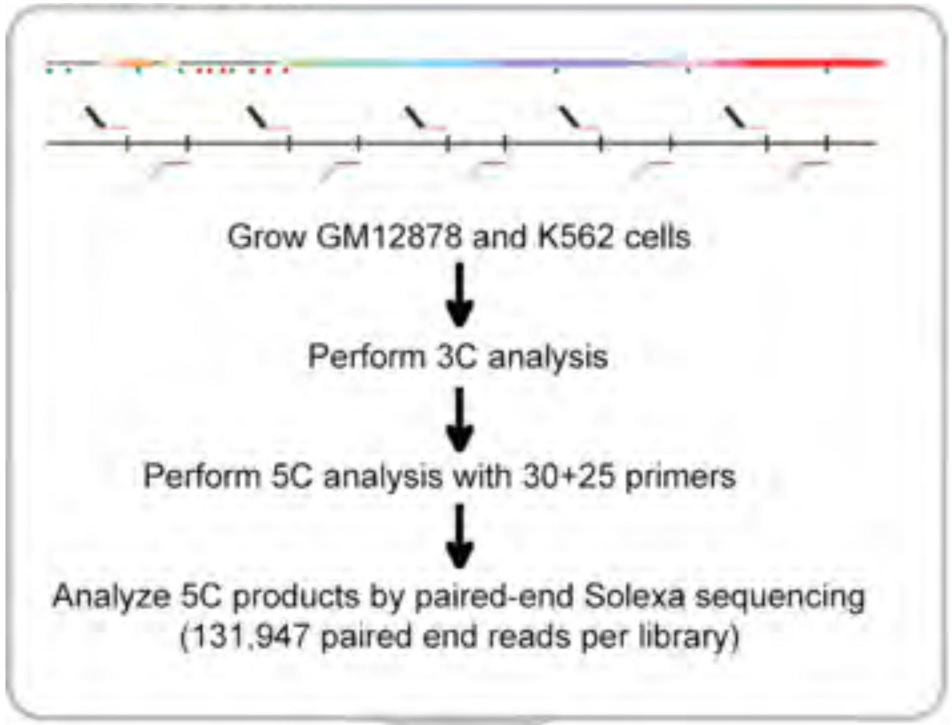
linear distance (*s*)

experimental restraints

scoring function minimization

conformational ensemble

Experiments

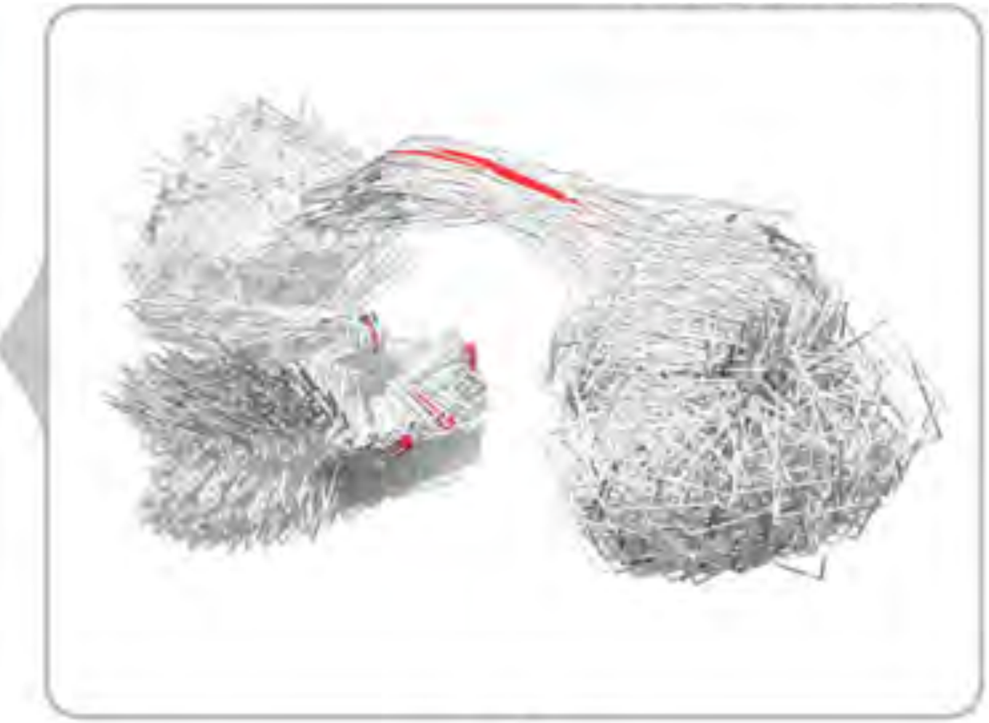


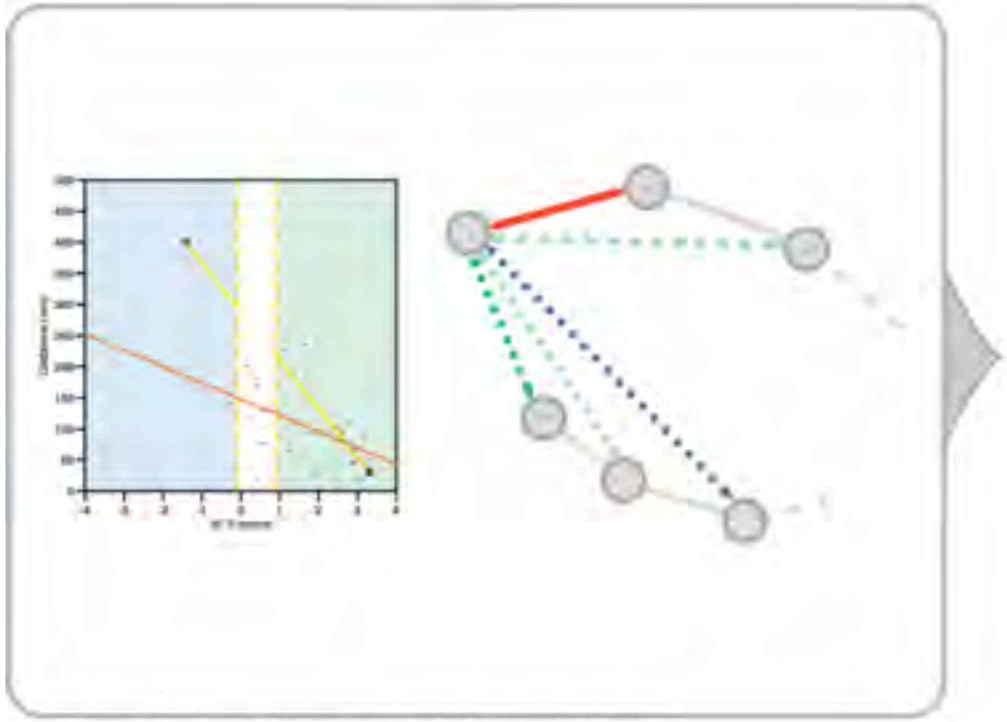
Grow GM12878 and K562 cells

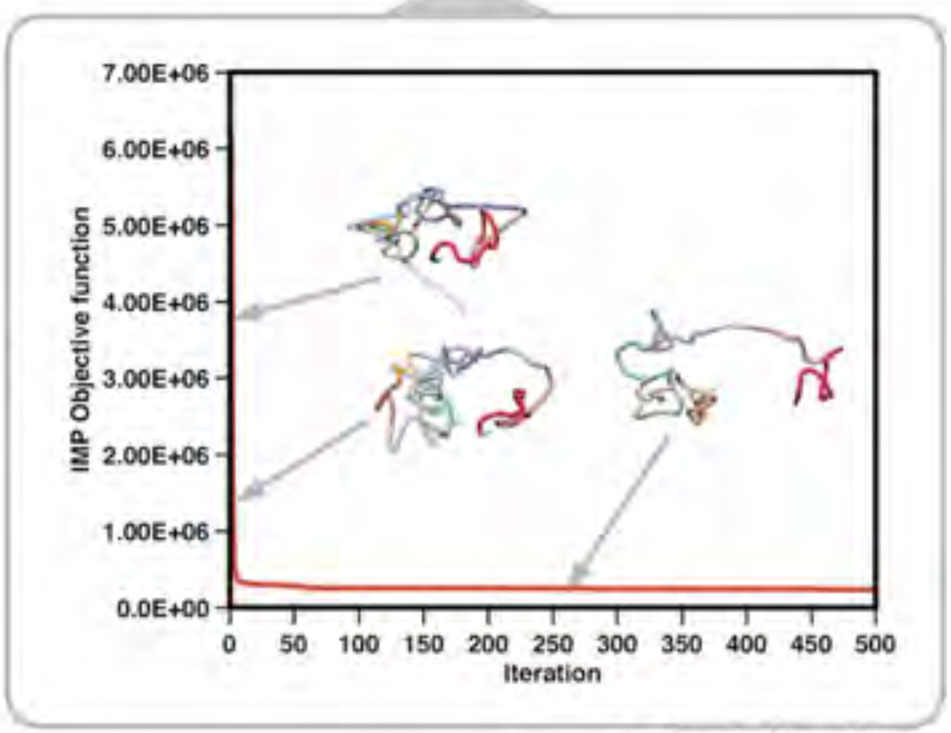
Perform 3C analysis

Perform 5C analysis with 30+25 primers

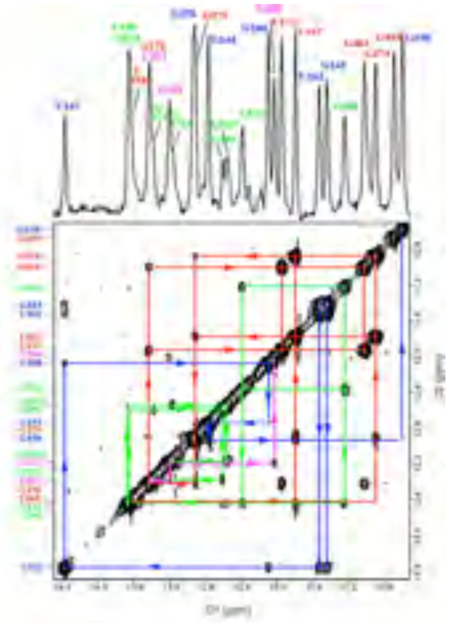
Analyze 5C products by paired-end Solexa sequencing (131,947 paired end reads per library)



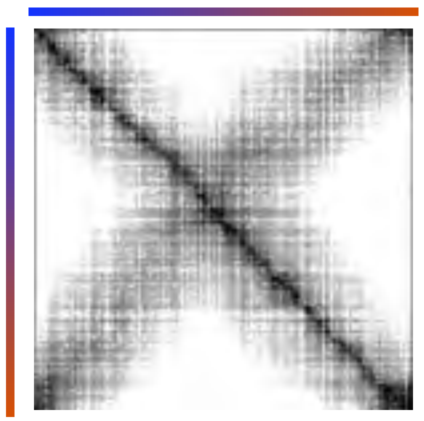




Computation

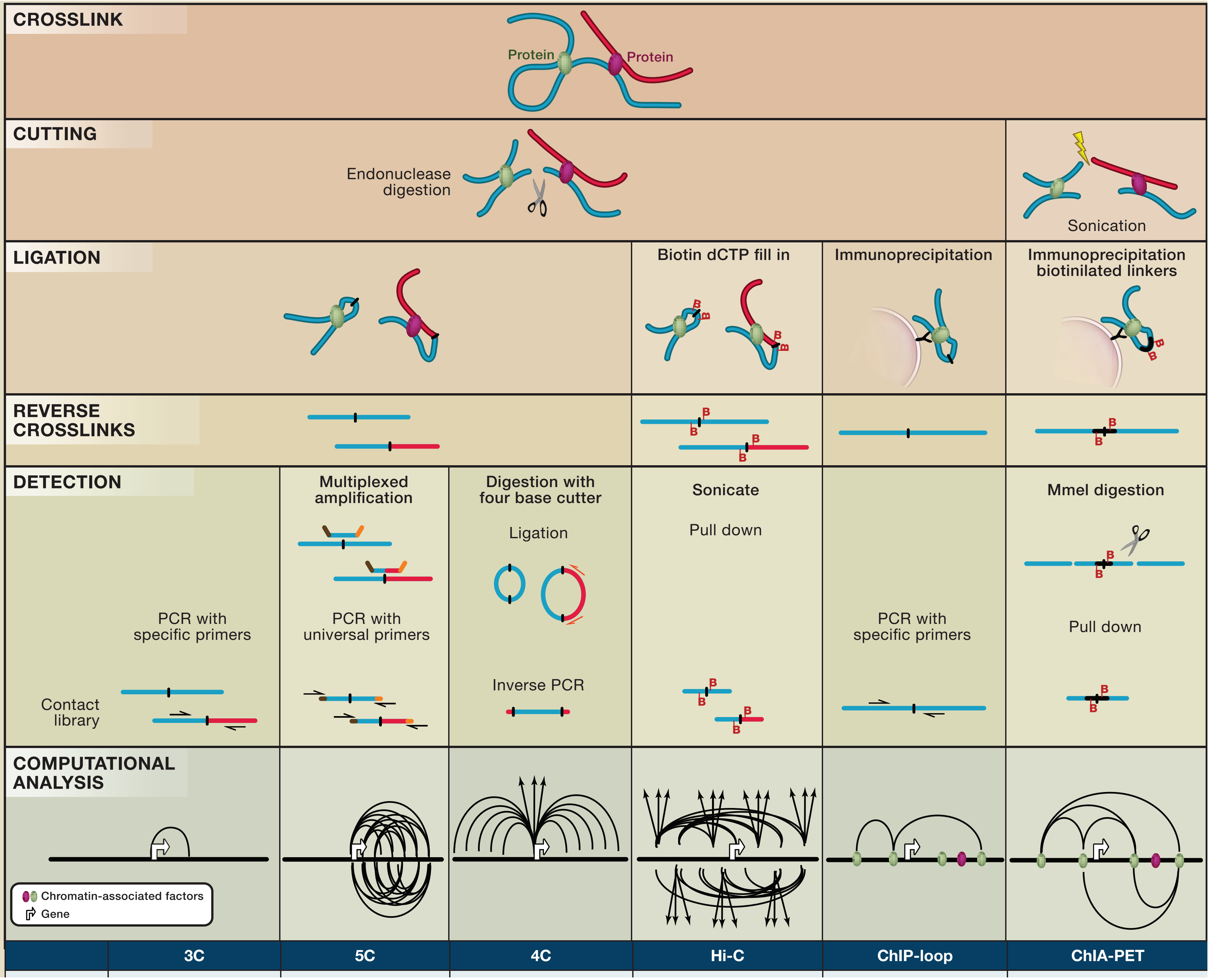


Biomolecular structure determination
2D-NOESY data



Chromosome structure determination
5C data

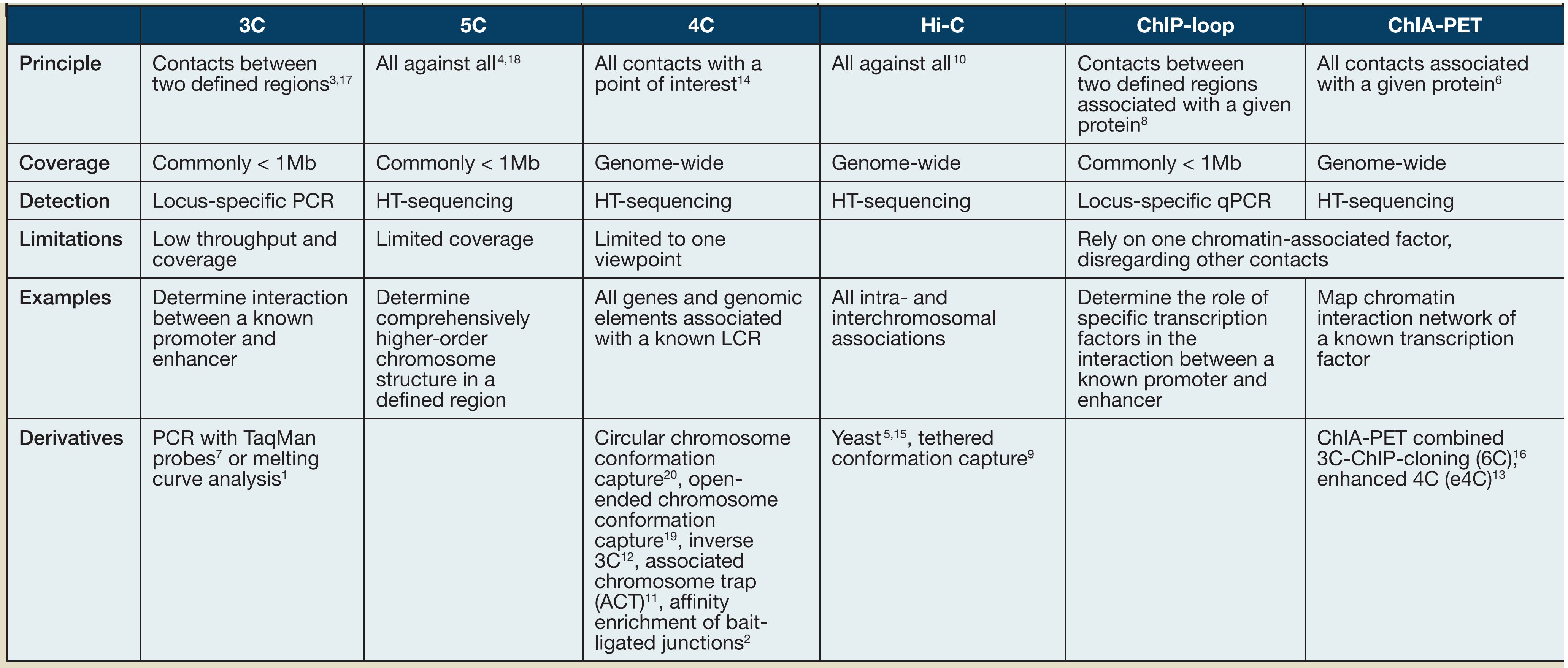
Chromosome Conformation Capture



	3C	5C	4C	Hi-C	ChIP-loop	ChIA-PET
CROSSLINK	Protein Protein					
CUTTING	Endonuclease digestion					Sonication
LIGATION				Biotin dCTP fill in	Immunoprecipitation	Immunoprecipitation biotinilated linkers
REVERSE CROSSLINKS						
DETECTION	PCR with specific primers Contact library	Multiplexed amplification PCR with universal primers	Digestion with four base cutter Ligation Inverse PCR	Sonicate Pull down	PCR with specific primers	Mmel digestion Pull down
COMPUTATIONAL ANALYSIS						

Hakim, O., & Misteli, T. (2012). SnapShot: Chromosome Confirmation Capture. Cell, 148(5), 1068–1068.e2.

Chromosome Conformation Capture

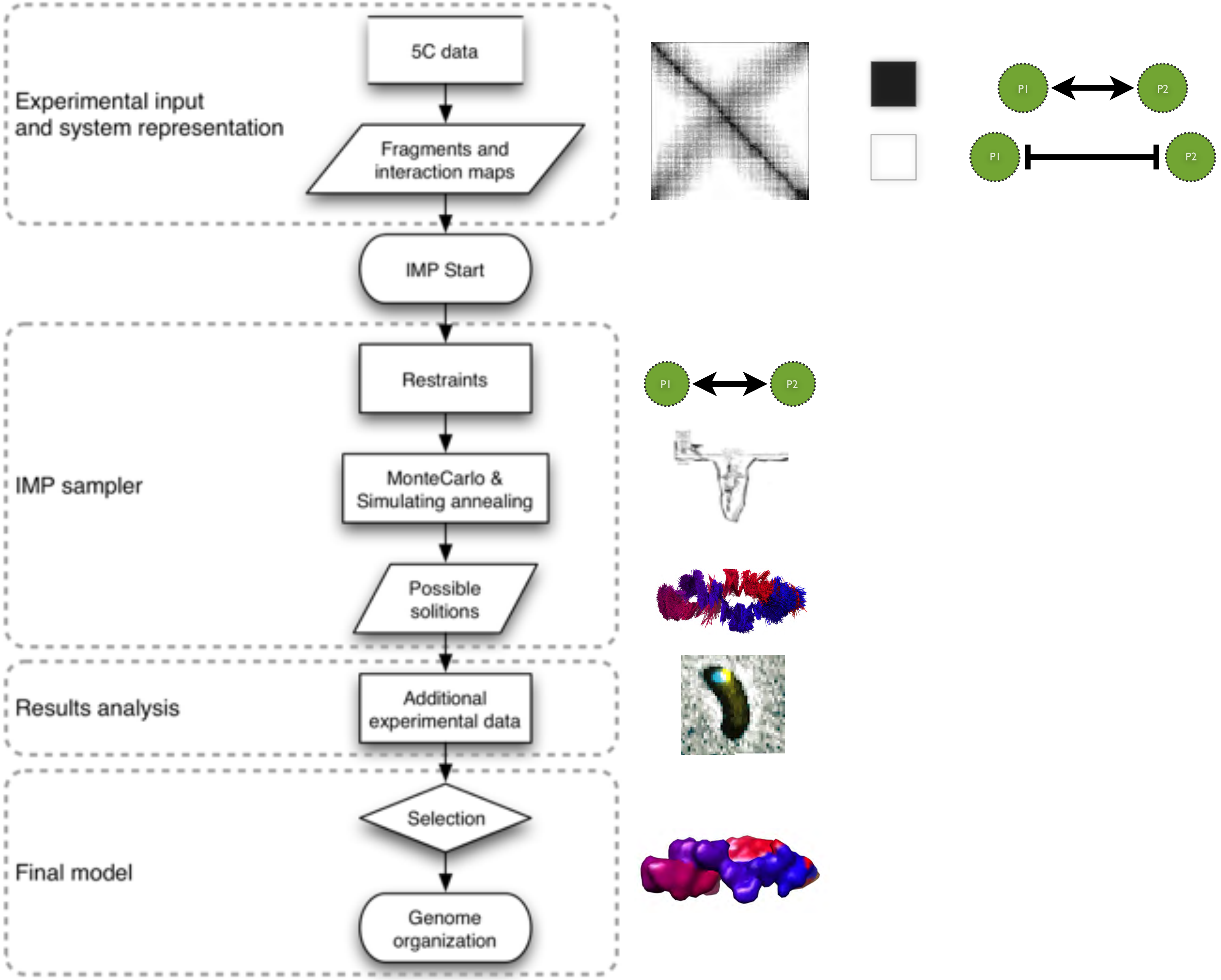


	3C	5C	4C	Hi-C	ChIP-loop	ChIA-PET
Principle	Contacts between two defined regions[3,17]	All against all[4,18]	All contacts with a point of interest[14]	All against all[10]	Contacts between two defined regions associated with a given protein[8]	All contacts associated with a given protein[6]
Coverage	Commonly < 1Mb	Commonly < 1Mb	Genome-wide	Genome-wide	Commonly < 1Mb	Genome-wide
Detection	Locus-specific PCR	HT-sequencing	HT-sequencing	HT-sequencing	Locus-specific qPCR	HT-sequencing
Limitations	Low throughput and coverage	Limited coverage	Limited to one viewpoint		Rely on one chromatin-associated factor, disregarding other contacts	
Examples	Determine interaction between a known promoter and enhancer	Determine comprehensively higher-order chromosome structure in a defined region	All genes and genomic elements associated with a known LCR	All intra- and interchromosomal associations	Determine the role of specific transcription factors in the interaction between a known promoter and enhancer	Map chromatin interaction network of a known transcription factor
Derivatives	PCR with TaqMan probes[7] or melting curve analysis[1]		Circular chromosome conformation capture[20], open-ended chromosome conformation capture[19], inverse 3C[12], associated chromosome trap (ACT)[11], affinity enrichment of bait-ligated junctions[2]	Yeast[5,15], tethered conformation capture[9]		ChIA-PET combined 3C-ChIP-cloning (6C);[16] enhanced 4C (e4C)[13]

Hakim, O., & Misteli, T. (2012). SnapShot: Chromosome Confirmation Capture. Cell, 148(5), 1068–1068.e2.

Modeling 3D Genomes

Baù, D. & Marti-Renom, M. A. Methods 58, 300–306 (2012).



Experimental input and system representation

5C data

Fragments and interaction maps

IMP Start

IMP sampler

Restraints

MonteCarlo & Simulating annealing

Possible solitions

Results analysis

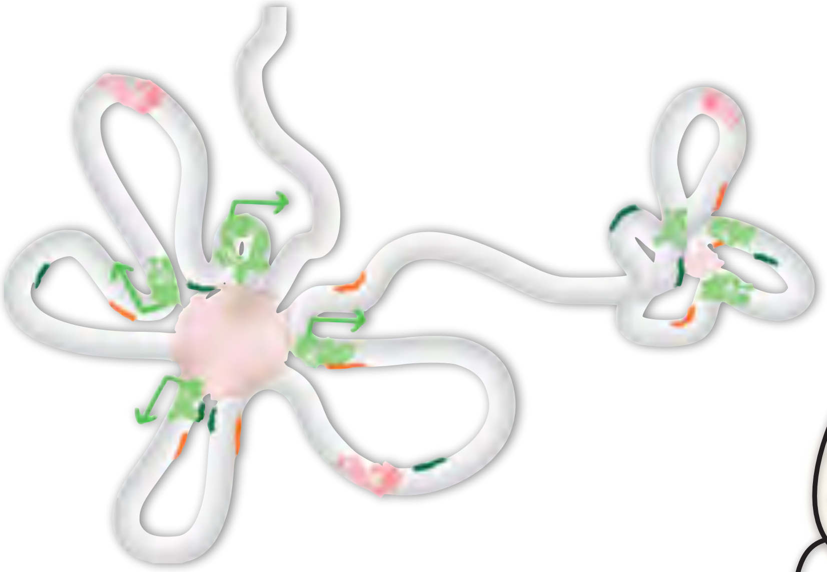
Additional experimental data

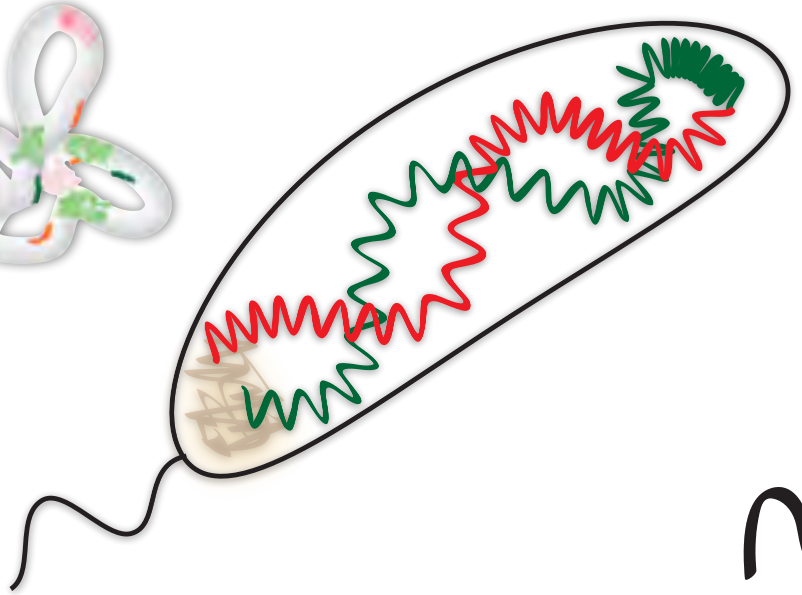
Final model

Selection

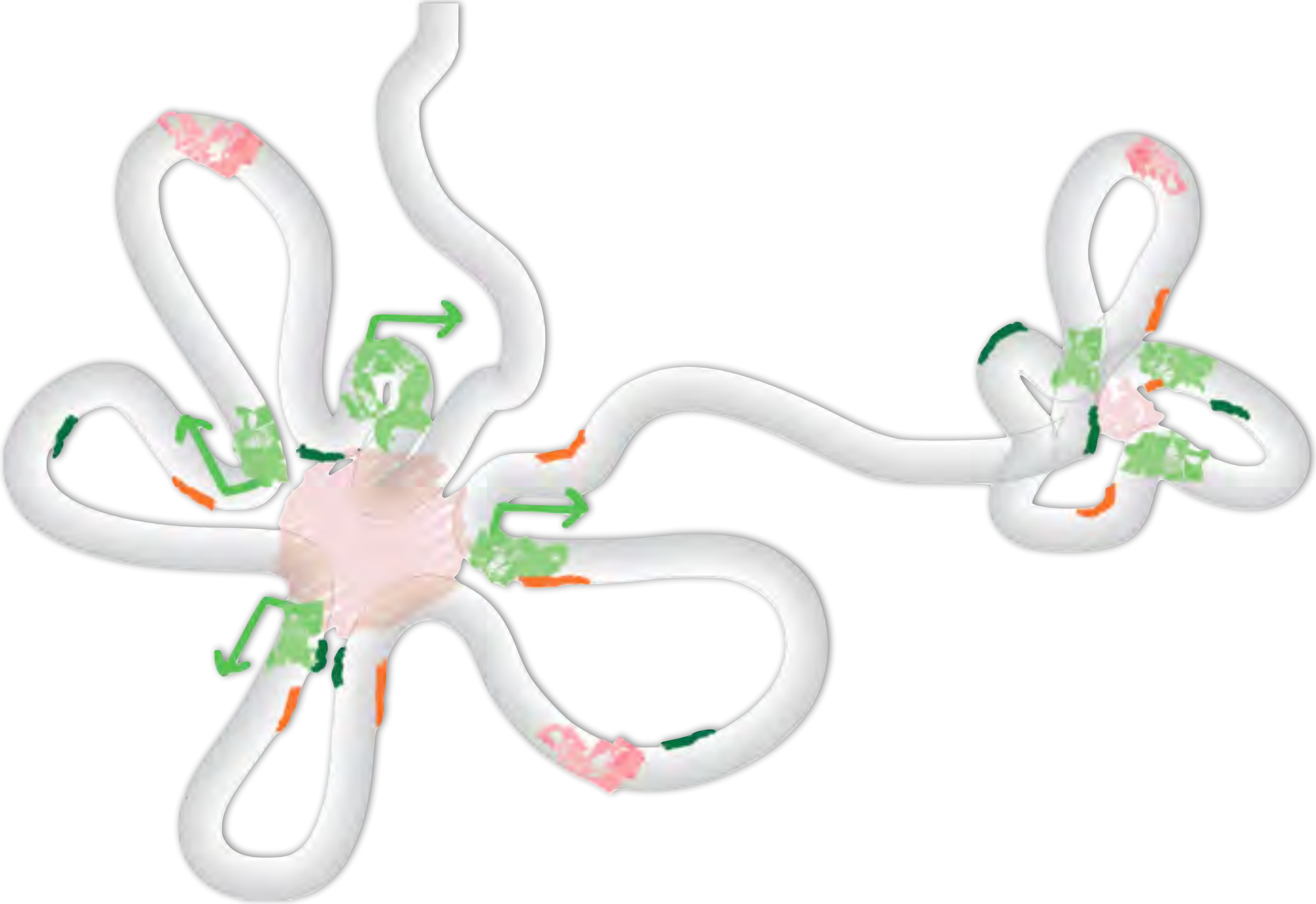
Genome organization

Examples...



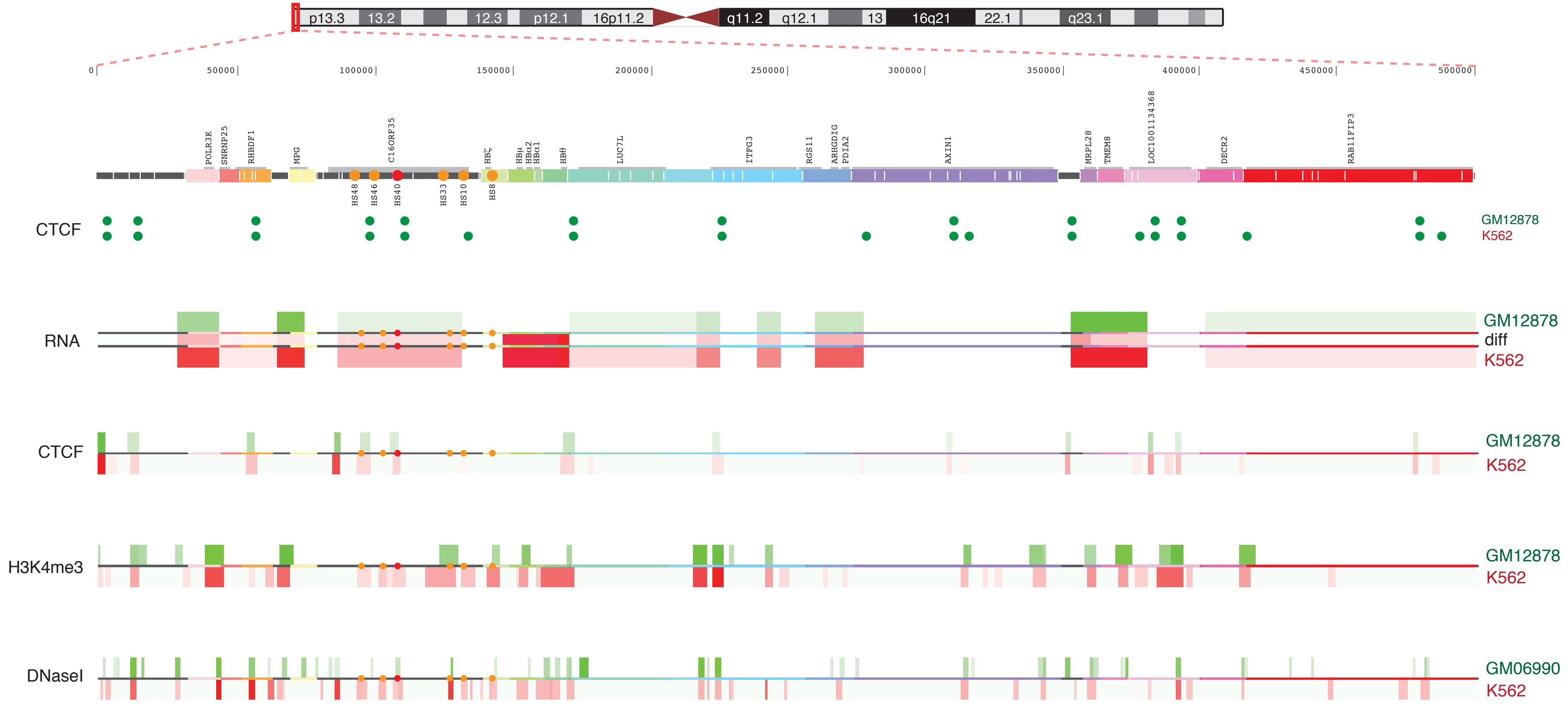


Human α-globin domain



Human α-globin domain

ENm008 genomic structure and environment

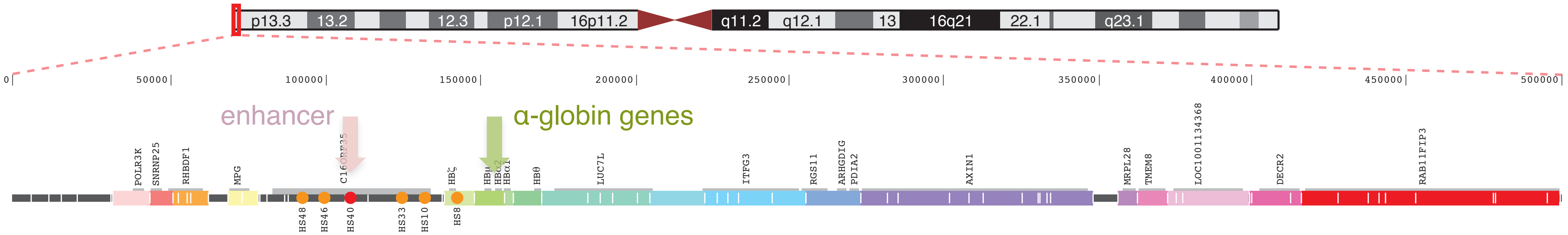


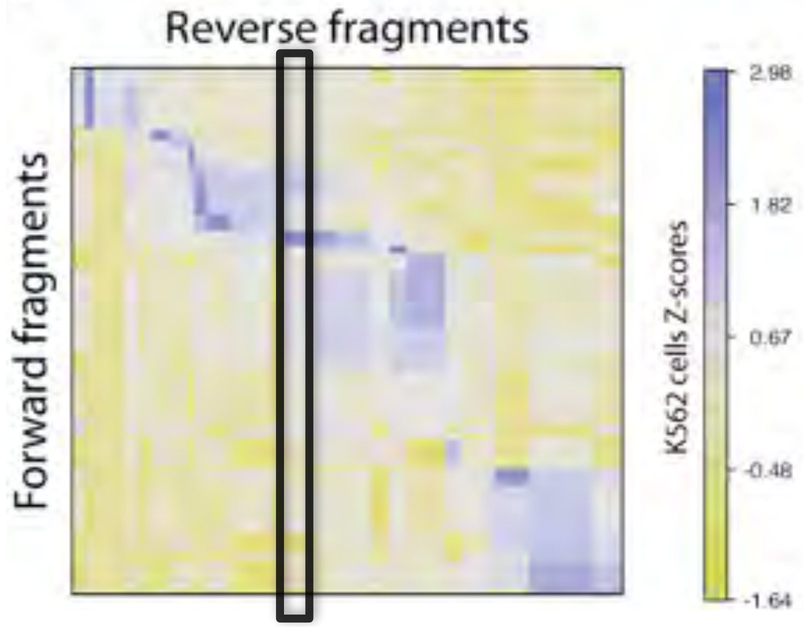
The ENCODE data for ENm008 region was obtained from the UCSC Genome Browser tracks for: RefSeq annotated genes, Affymetrix/CSHL expression data (Gingeras Group at Cold Spring Harbor), Duke/NHGRI DNaseI Hypersensitivity data (Crawford Group at Duke University), and Histone Modifications by Broad Institute ChIP-seq (Bernstein Group at Broad Institute of Harvard and MIT).

ENCODE Consortium. Nature (2007) vol. 447 (7146) pp. 799-816

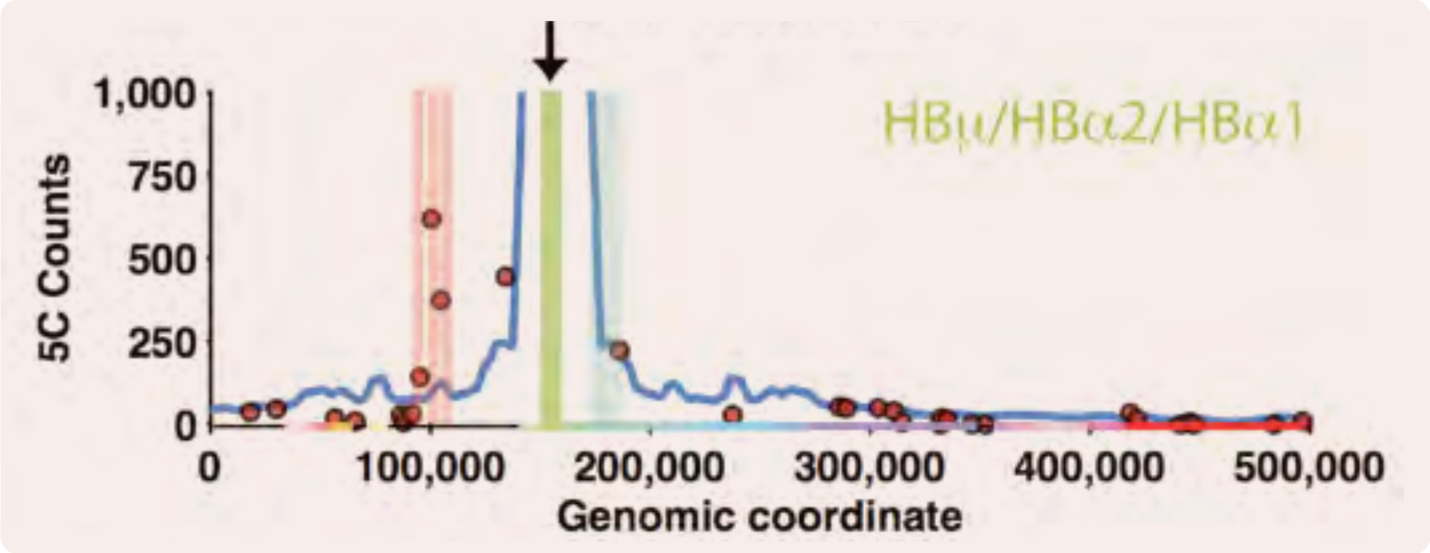
Human α-globin domain

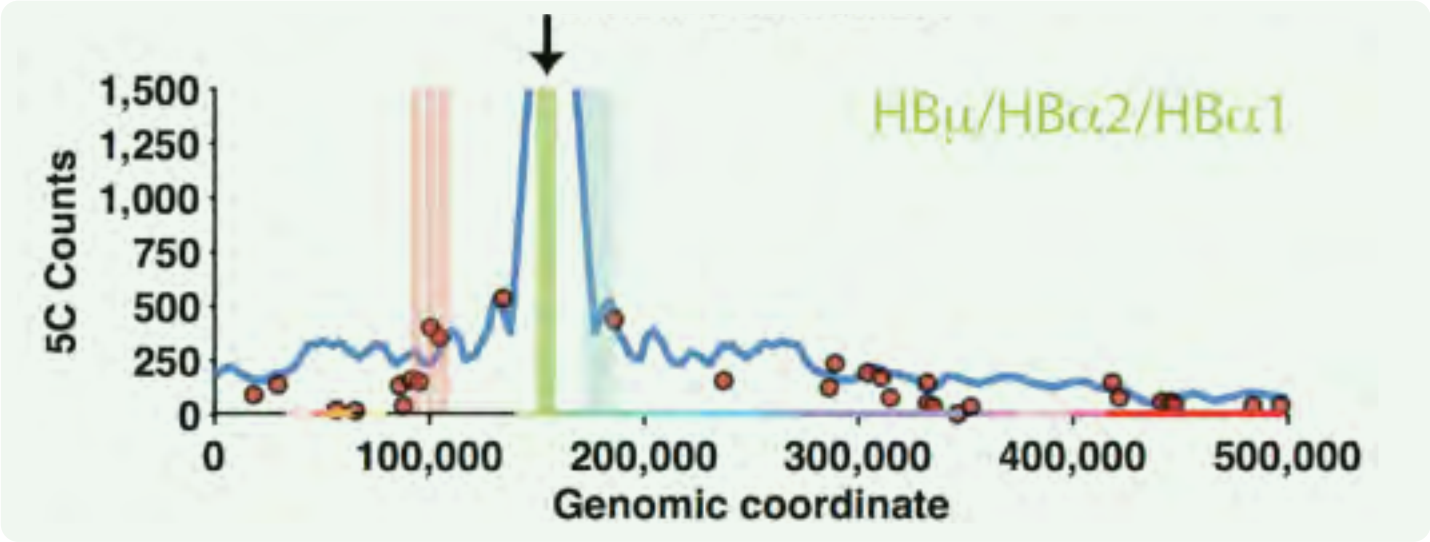
ENm008 genomic structure and environment





K562 cells:
α-globin genes active

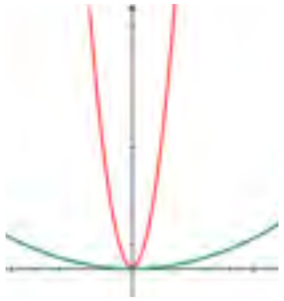




Representation

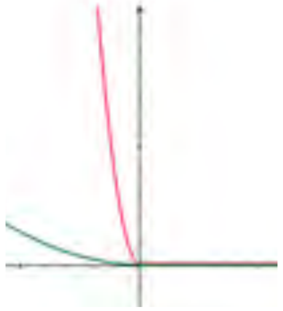
Harmonic

$$H_{i,j} = k\left(d_{i,j} - d_{i,j}^{0}\right)^2$$



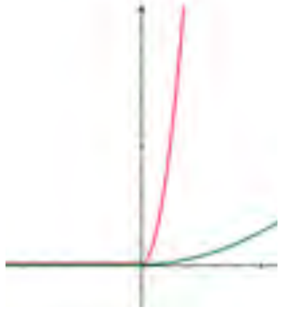
Harmonic Lower Bound

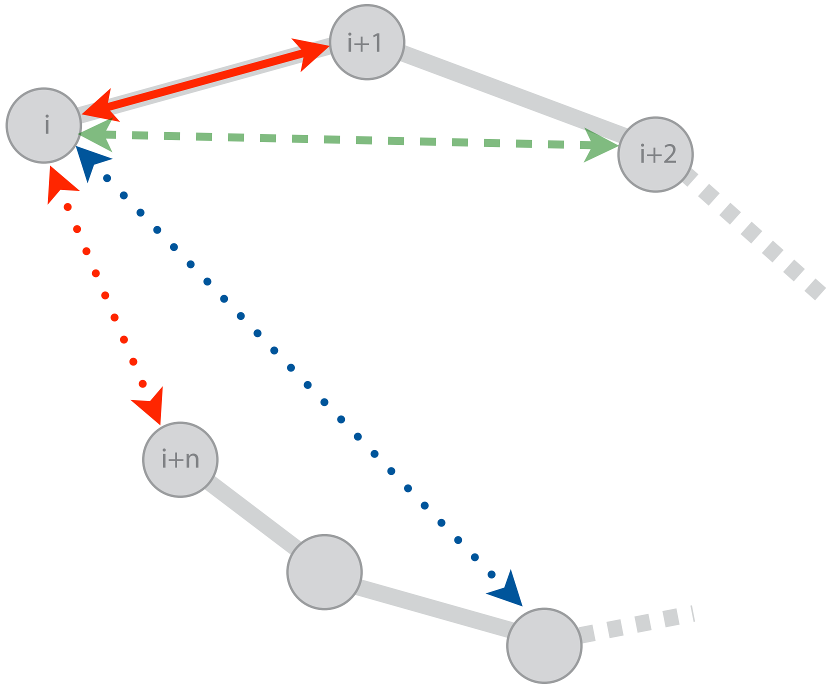
$$\begin{cases} if \;\; d_{i,j} \le d_{i,j}^{0}; & lbH_{i,j} = k\left(d_{i,j} - d_{i,j}^{0}\right)^2 \\ if \;\; d_{i,j} > d_{i,j}^{0}; & lbH_{i,j} = 0 \end{cases}$$



Harmonic Upper Bound

$$\begin{cases} if \;\; d_{i,j} \ge d_{i,j}^{0}; & ubH_{i,j} = k\left(d_{i,j} - d_{i,j}^{0}\right)^2 \\ if \;\; d_{i,j} < d_{i,j}^{0}; & ubH_{i,j} = 0 \end{cases}$$

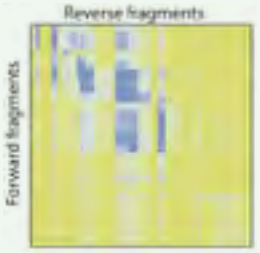


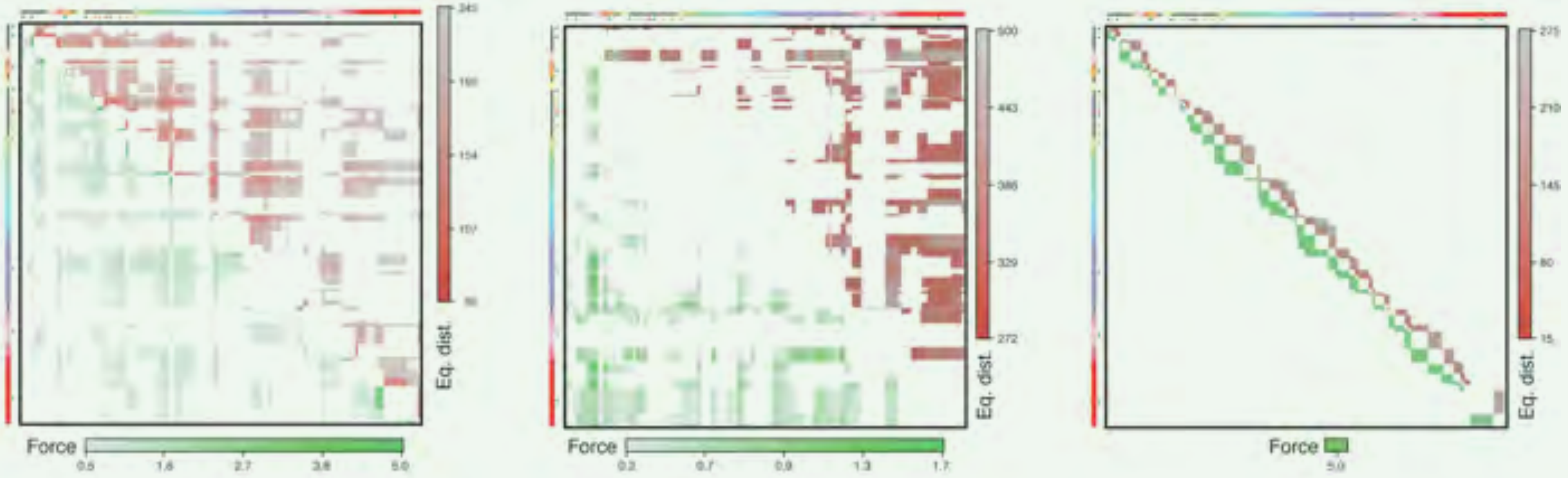


Scoring

GM12878

70 fragments
1,520 restraints





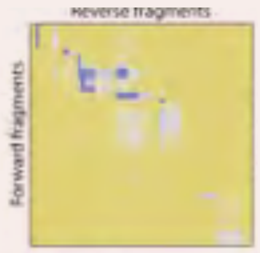
Harmonic

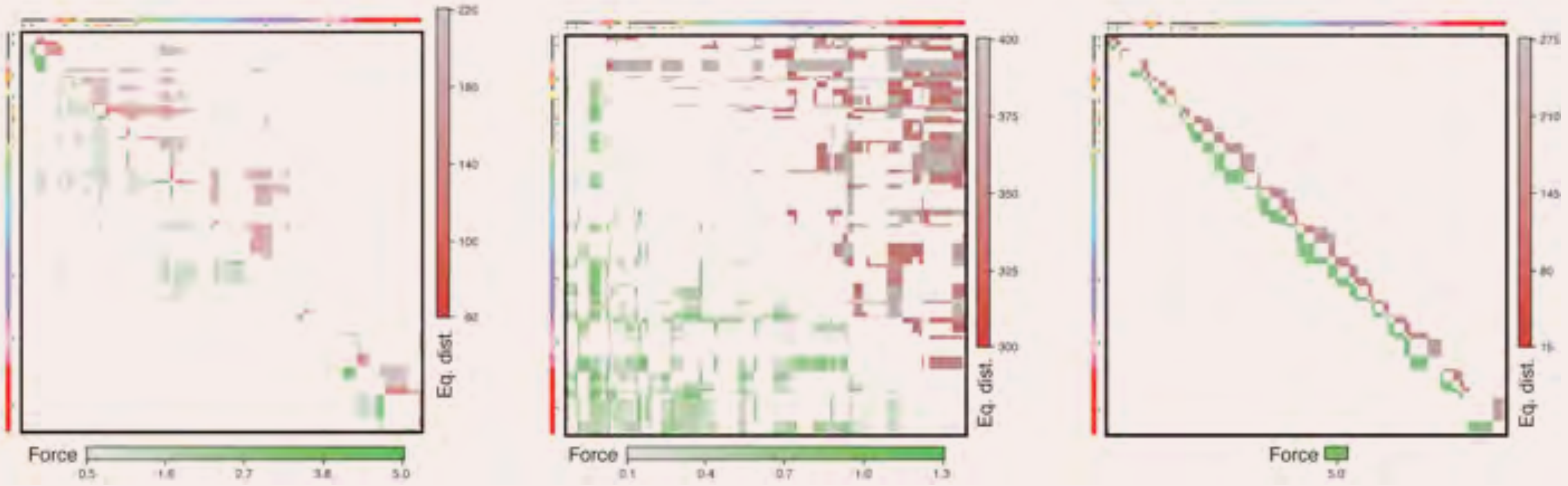
Harmonic Lower Bound

Harmonic Upper Bound

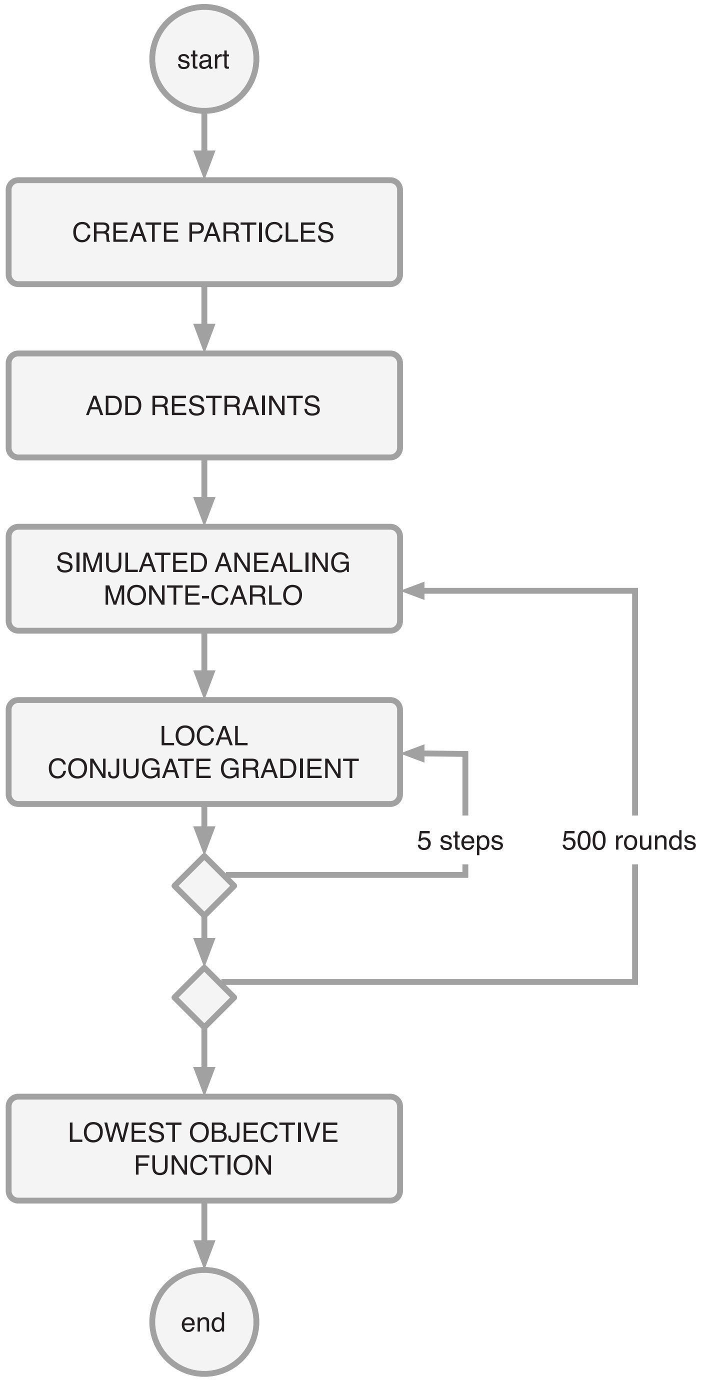
K562

70 fragments
1,049 restraints





Optimization


start

CREATE PARTICLES

ADD RESTRAINTS

SIMULATED ANEALING MONTE-CARLO

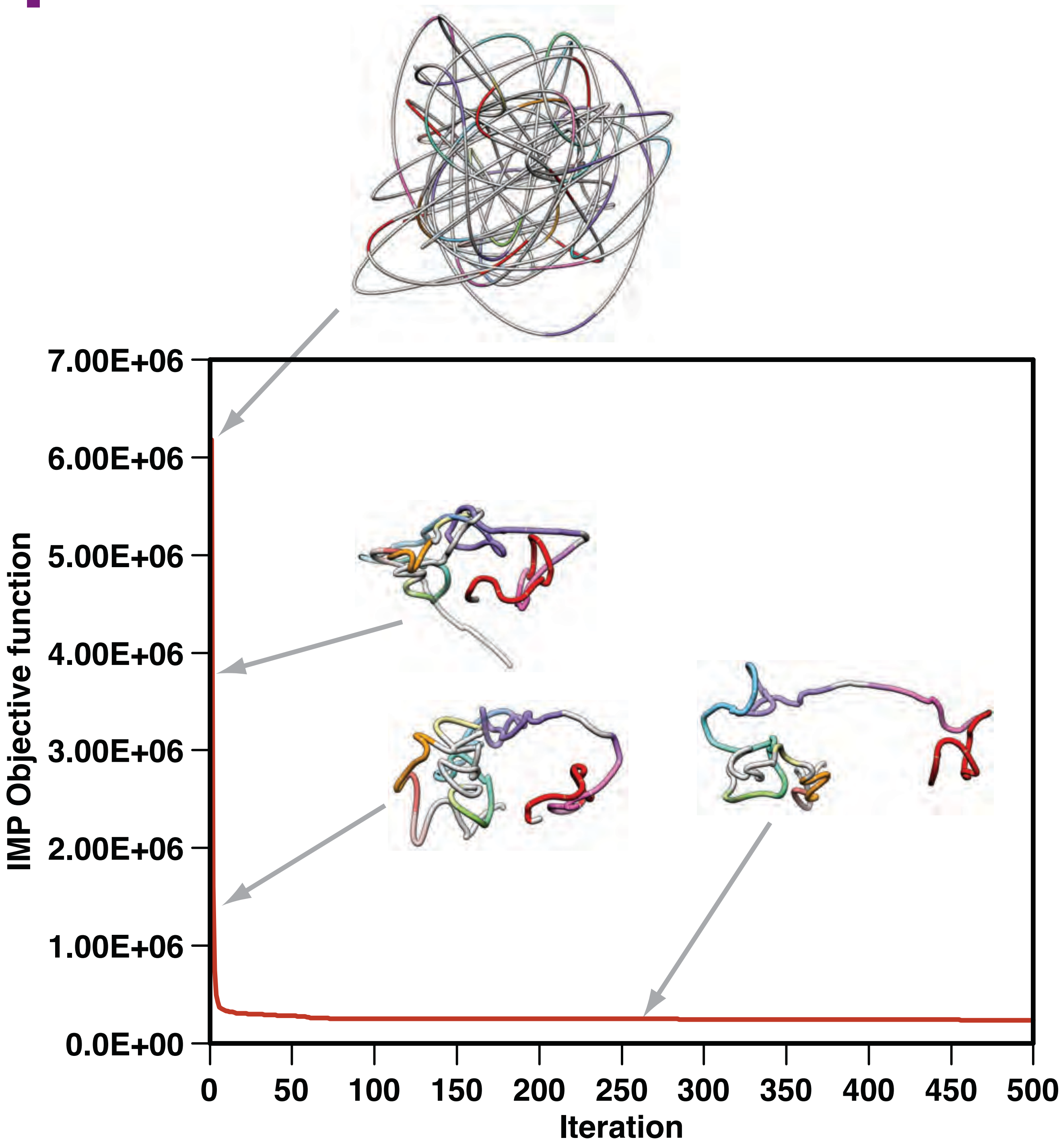
LOCAL CONJUGATE GRADIENT

5 steps

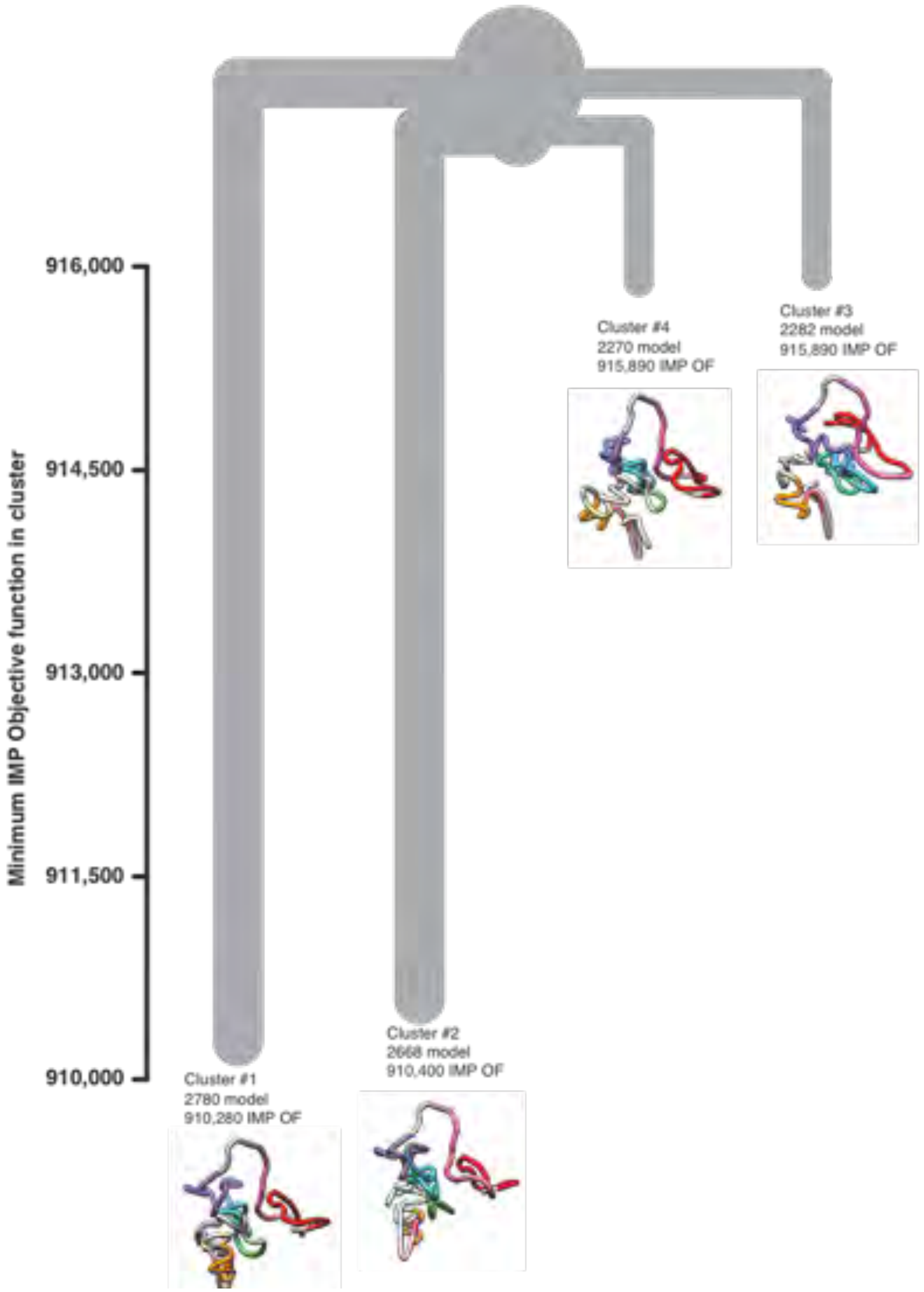
500 rounds

LOWEST OBJECTIVE FUNCTION

end

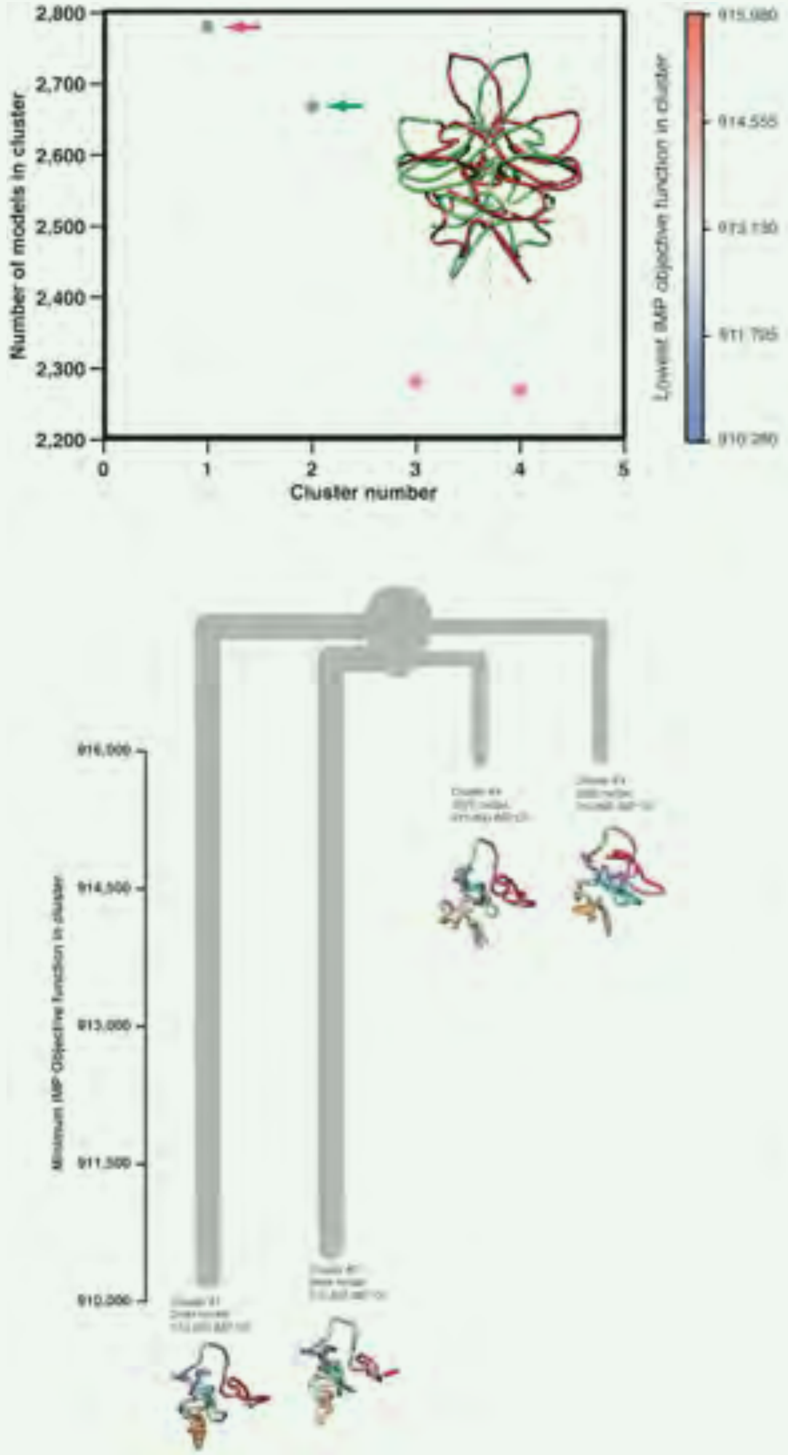


Clustering

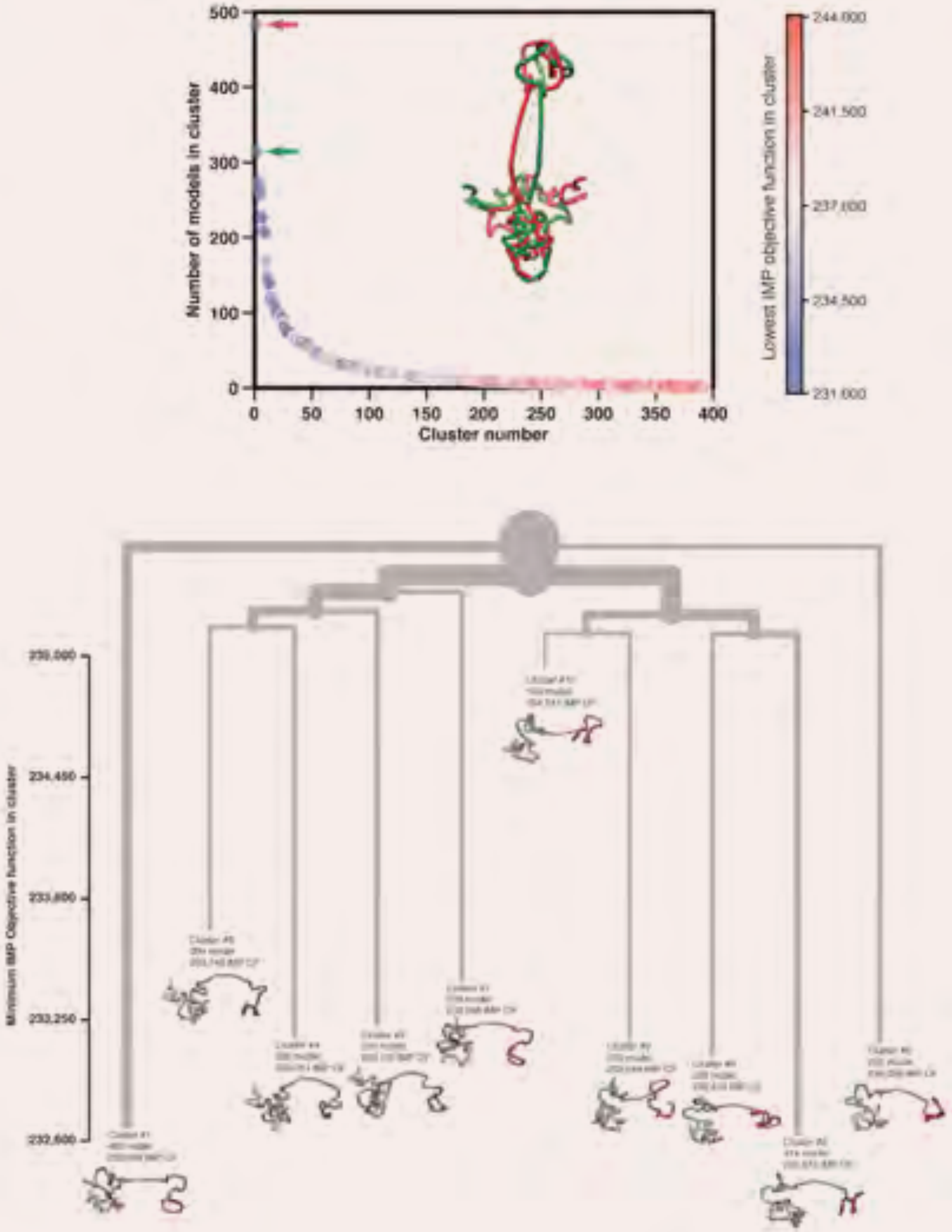

916,000
914,500
913,000
911,500
910,000
Minimum IMP Objective function in cluster
Cluster #4
2270 model
915,890 IMP OF
Cluster #3
2282 model
915,890 IMP OF
Cluster #1
2780 model
910,280 IMP OF
Cluster #2
2668 model
910,400 IMP OF

Not just one solution

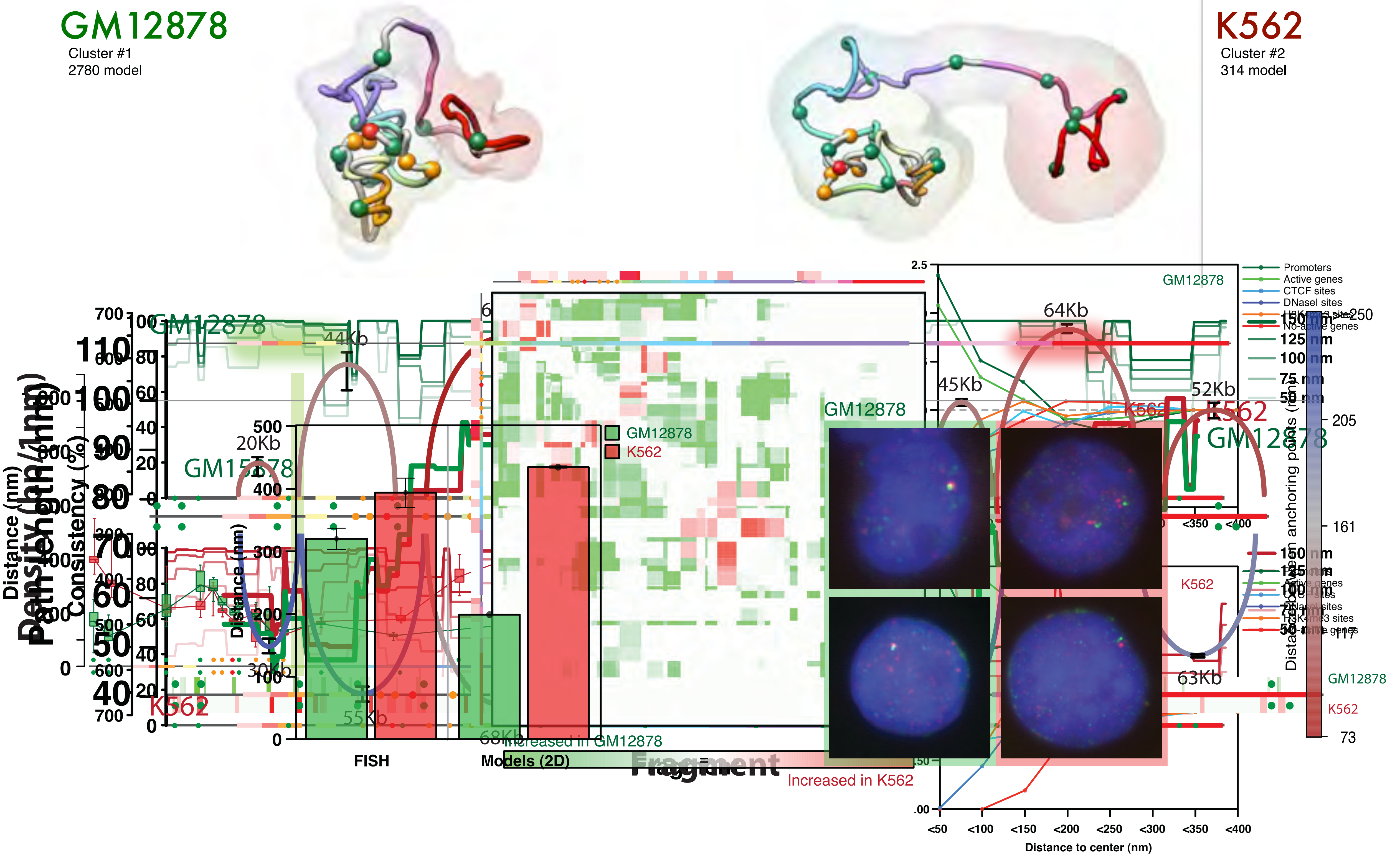
GM12878



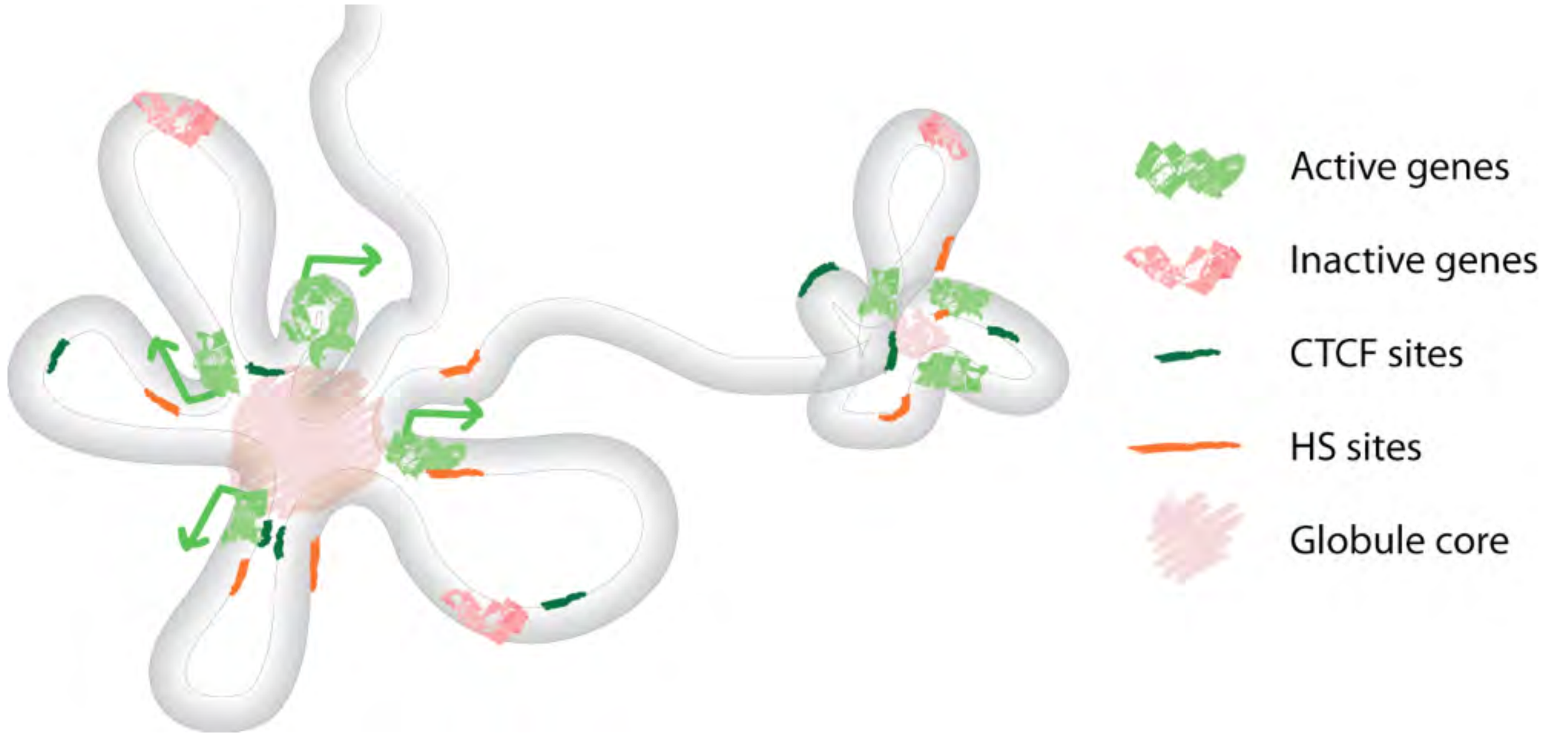
K562

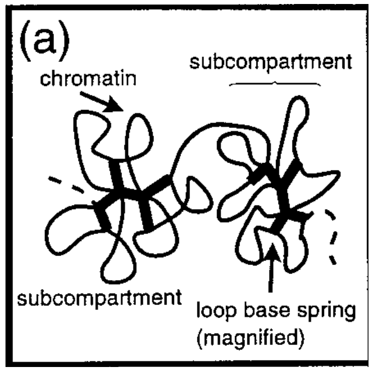


Reg[illegible]ents

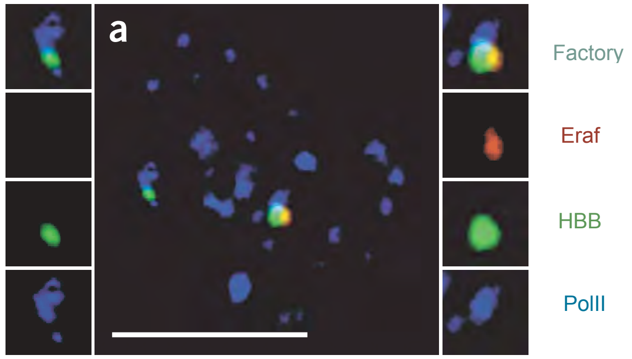

GM12878
Cluster #1
2780 model
K562
Cluster #2
314 model
GM12878
K562
44Kb
20Kb
30Kb
55Kb
68Kb
64Kb
45Kb
52Kb
63Kb
FISH
Models (2D)
Distance (nm)
Consistency (%)
Density (bp/1nm)
Path length (nm)
Fragment
Increased in GM12878
Increased in K562
Promoters
Active genes
CTCF sites
DNaseI sites
H3K4me3 sites
No-active genes
150 nm
125 nm
100 nm
75 nm
50 nm
Distance to center (nm)
<50
<100
<150
<200
<250
<300
<350
<400
Distance between anchoring points (nm)
250
205
161
117
73
0
100
200
300
400
500
600
700
20
40
60
80
100
110
90
70
50
40
.00
2.5

The “Chromatin Globule” model

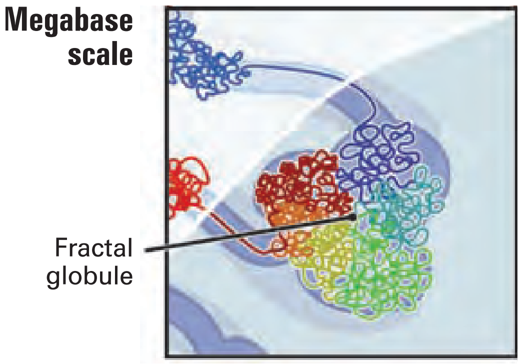




Münkel et al. JMB (1999)



Osborne et al. Nat Genet (2004)

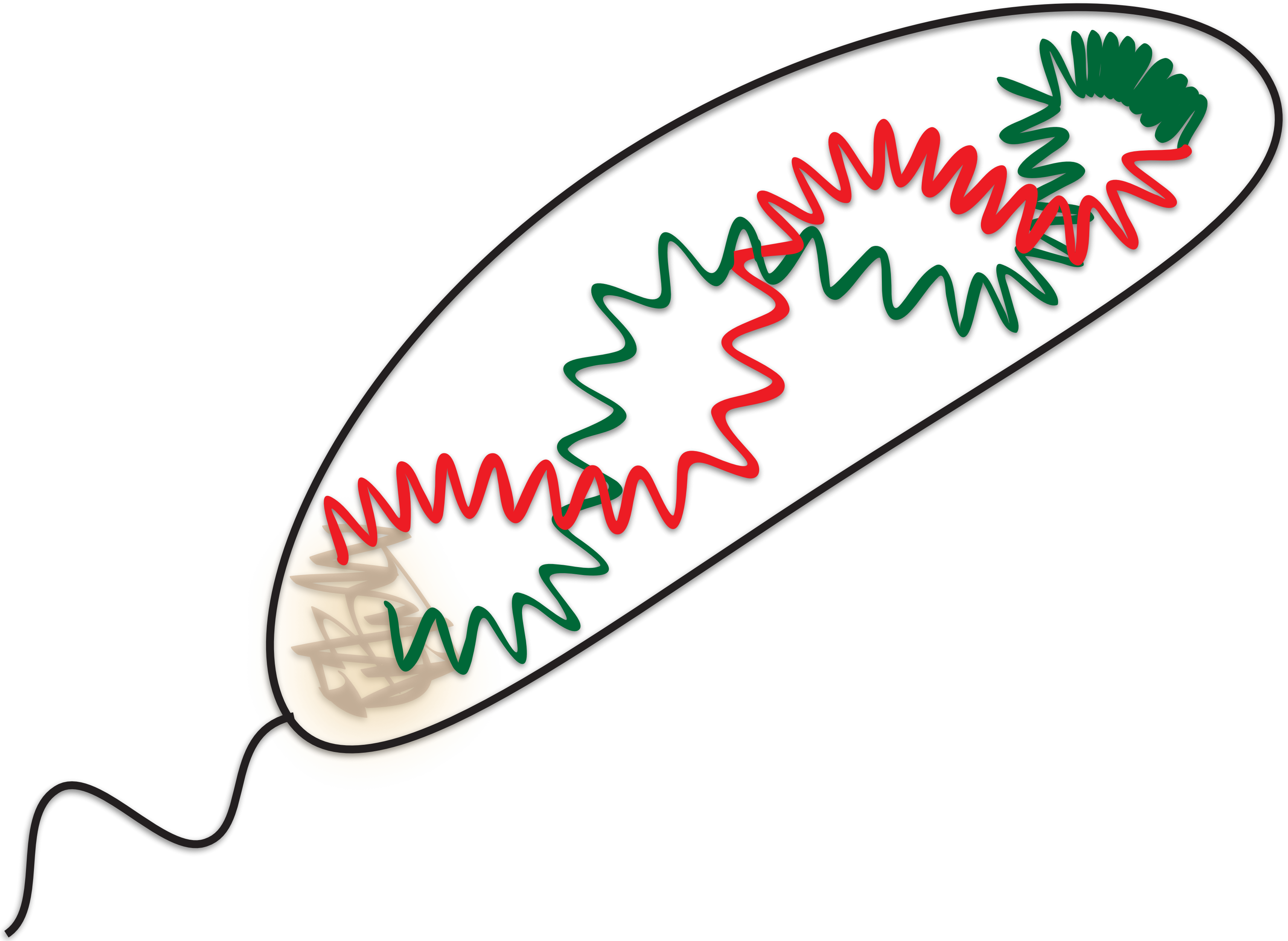


Lieberman-Aiden et al. Science (2009)

D. Baù *et al.* **Nat Struct Mol Biol** (2011) 18:107-14

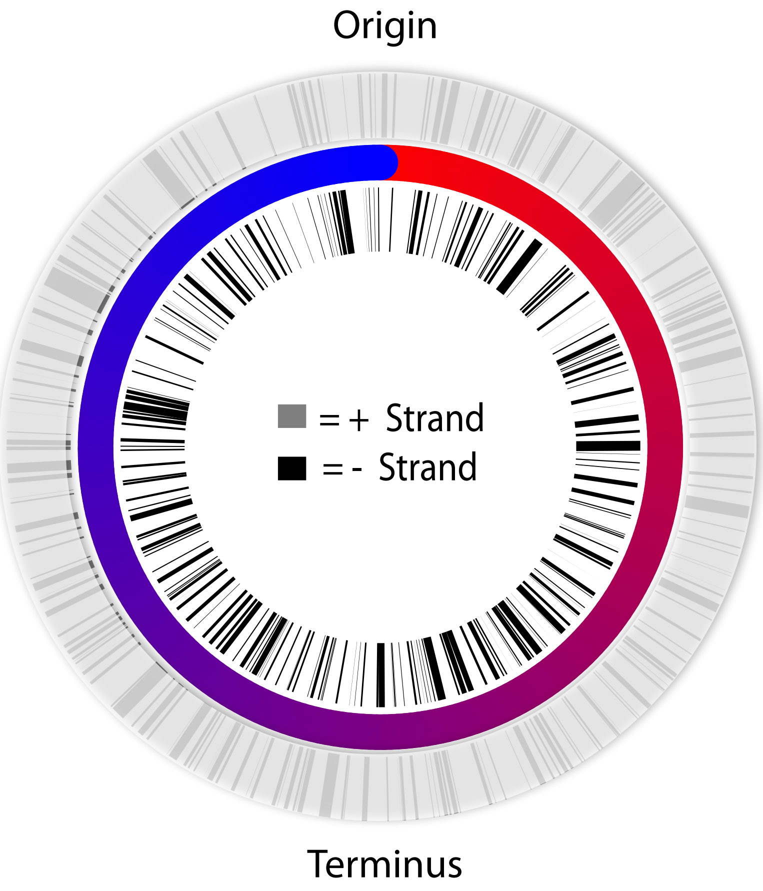
A. Sanyal *et al.* **Current Opinion in Cell Biology** (2011) 23:325–33.

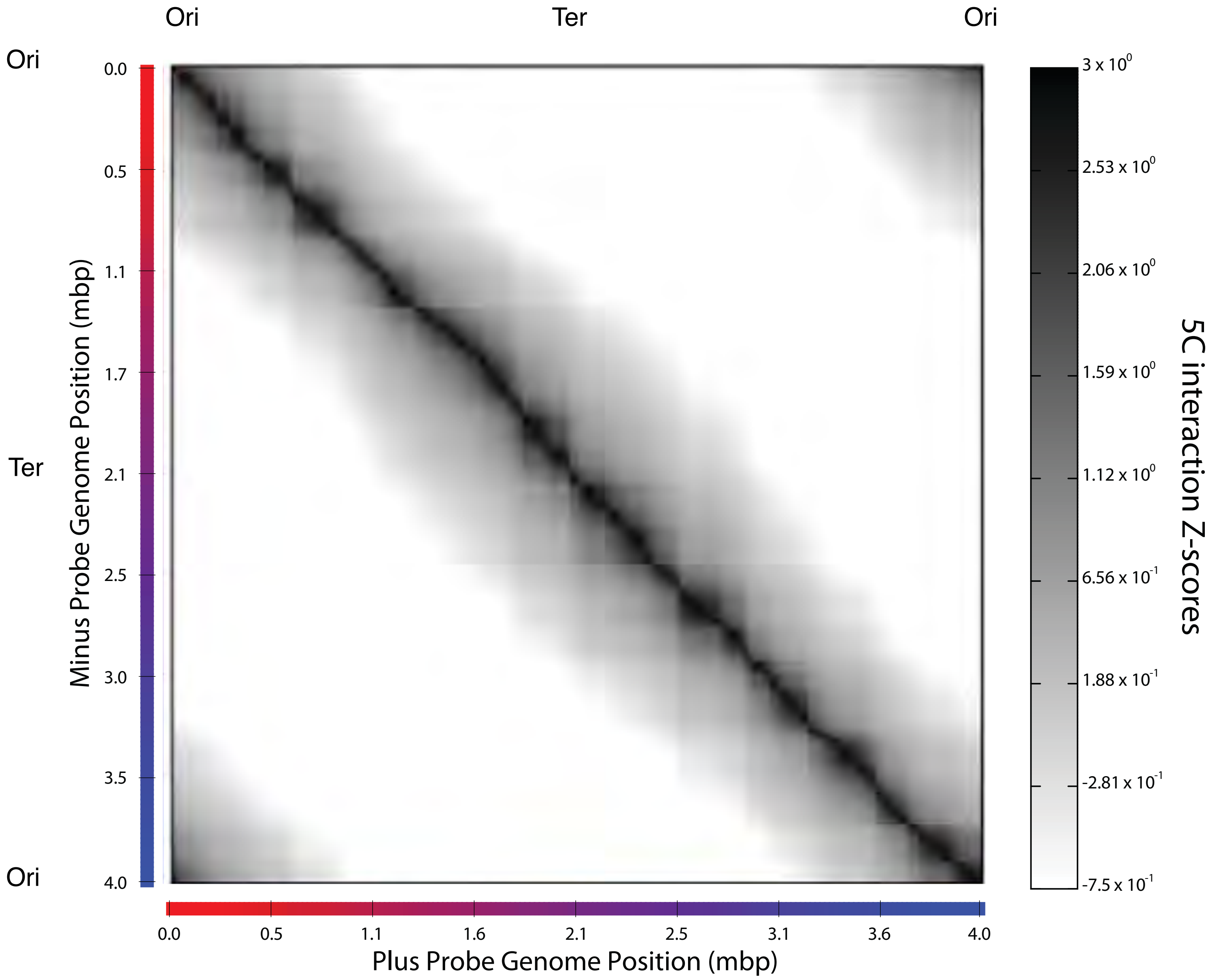
Caulobacter crescentus genome



The 3D architecture of Caulobacter Crescentus

4,016,942 bp & 3,767 genes





169 5C primers on + strand

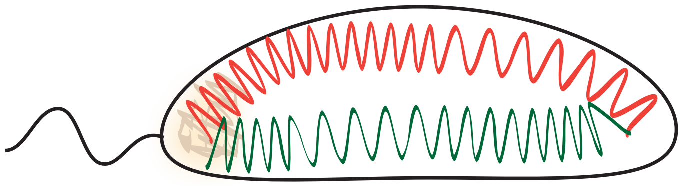
170 5C primers on – strand

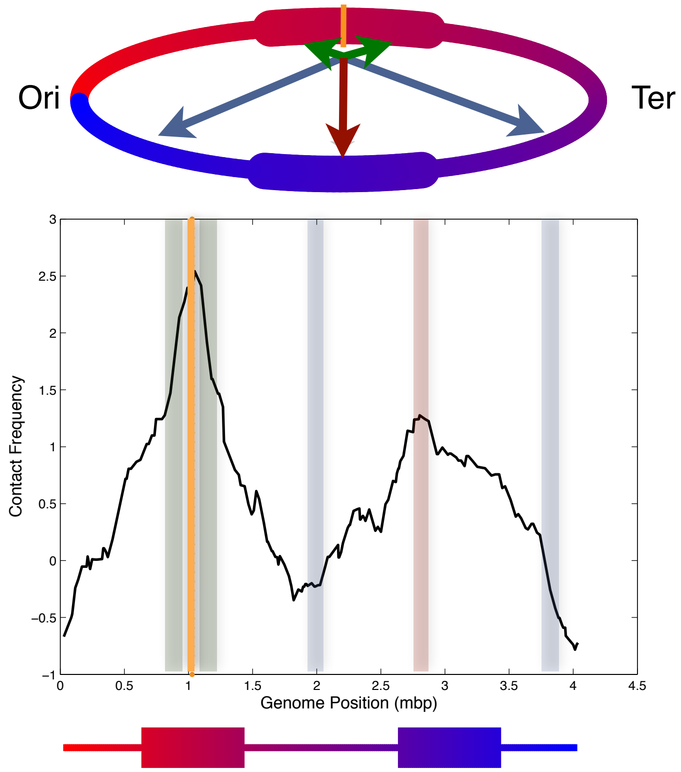
28,730 chromatin interactions

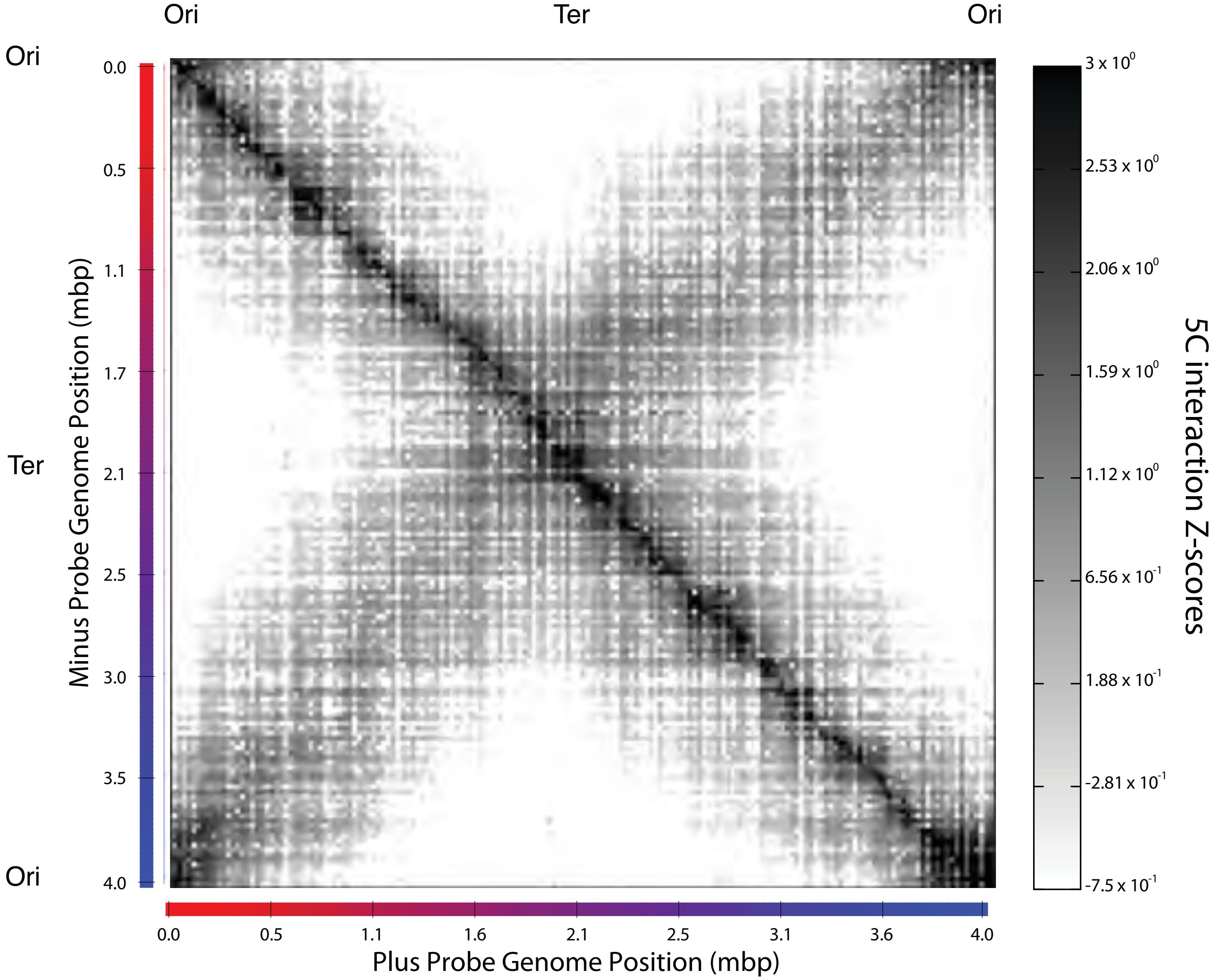
~13Kb

5C interaction matrix

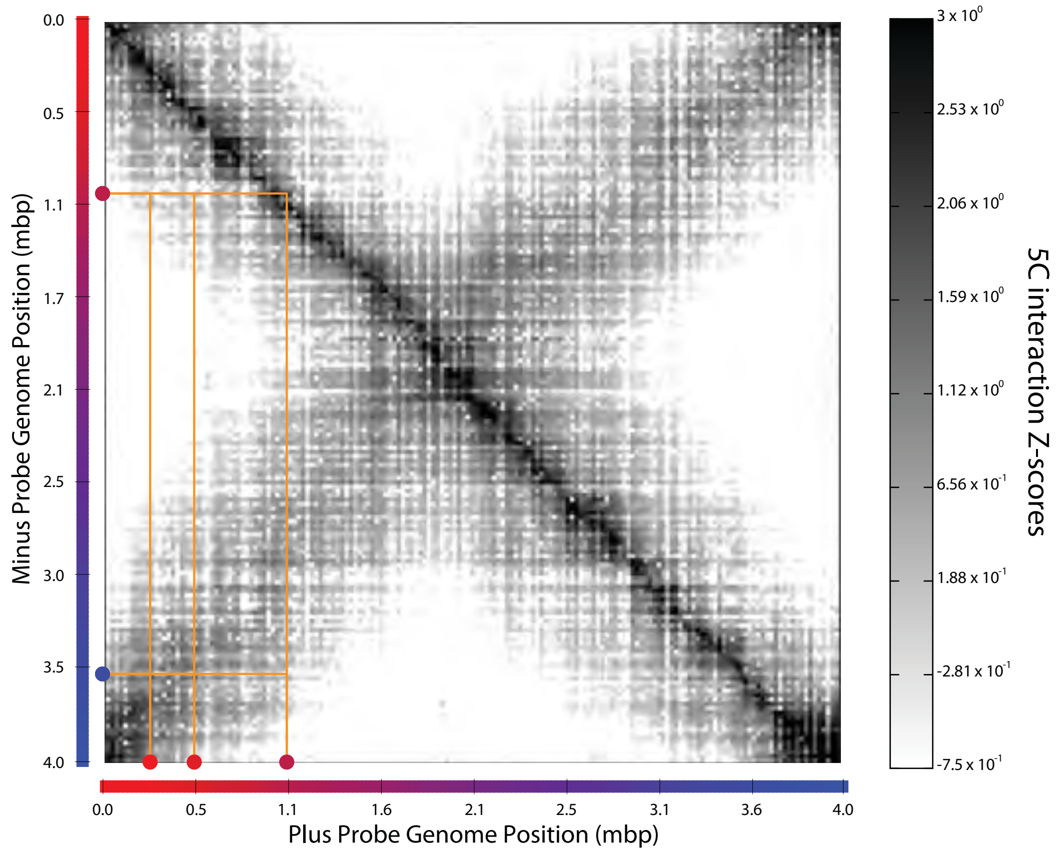
ELLIPSOID for Caulobacter cresentus



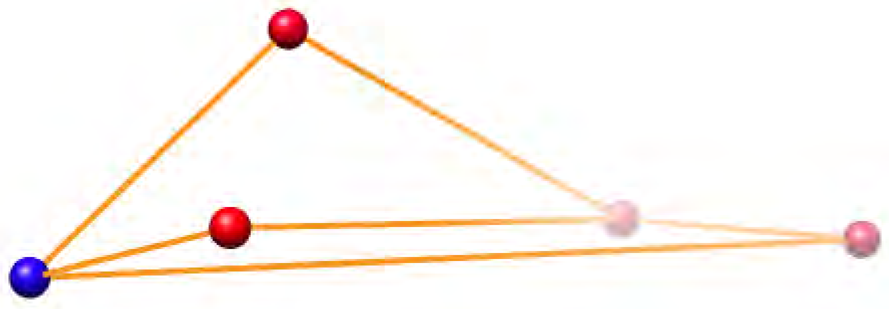


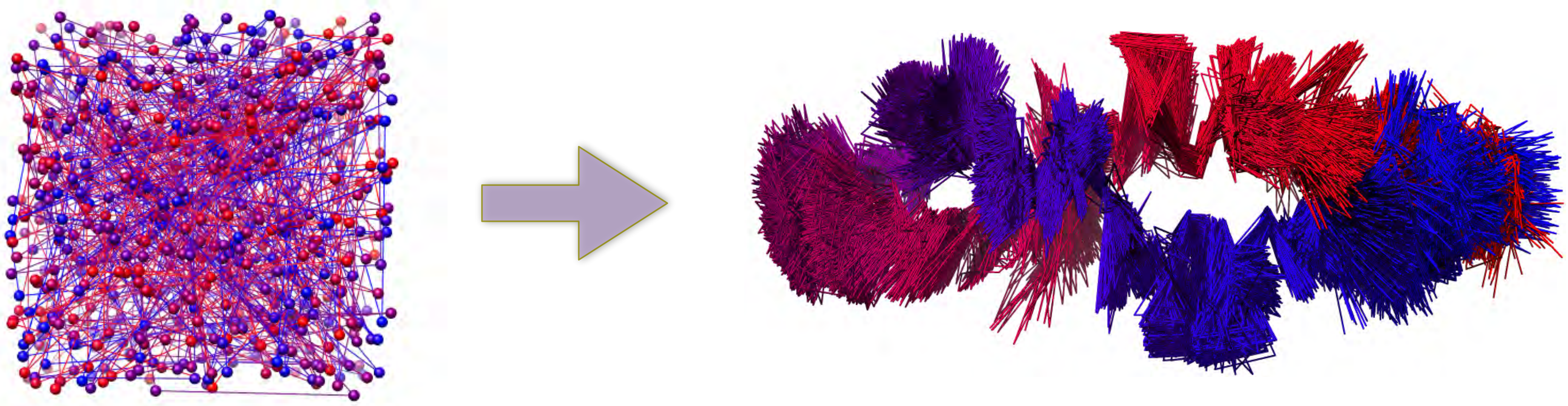


3D model building with the 5C + IMP approach



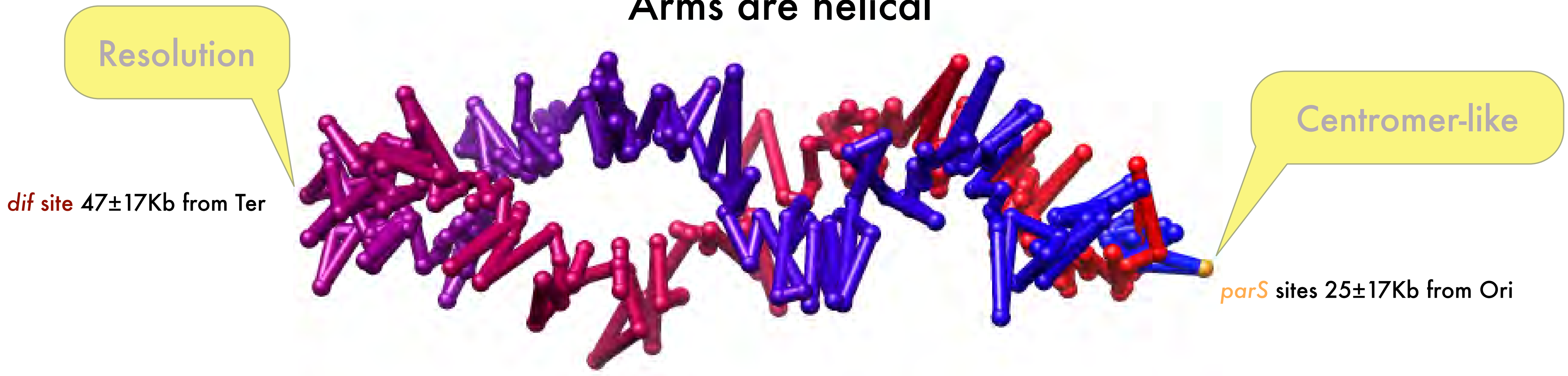
339 mers

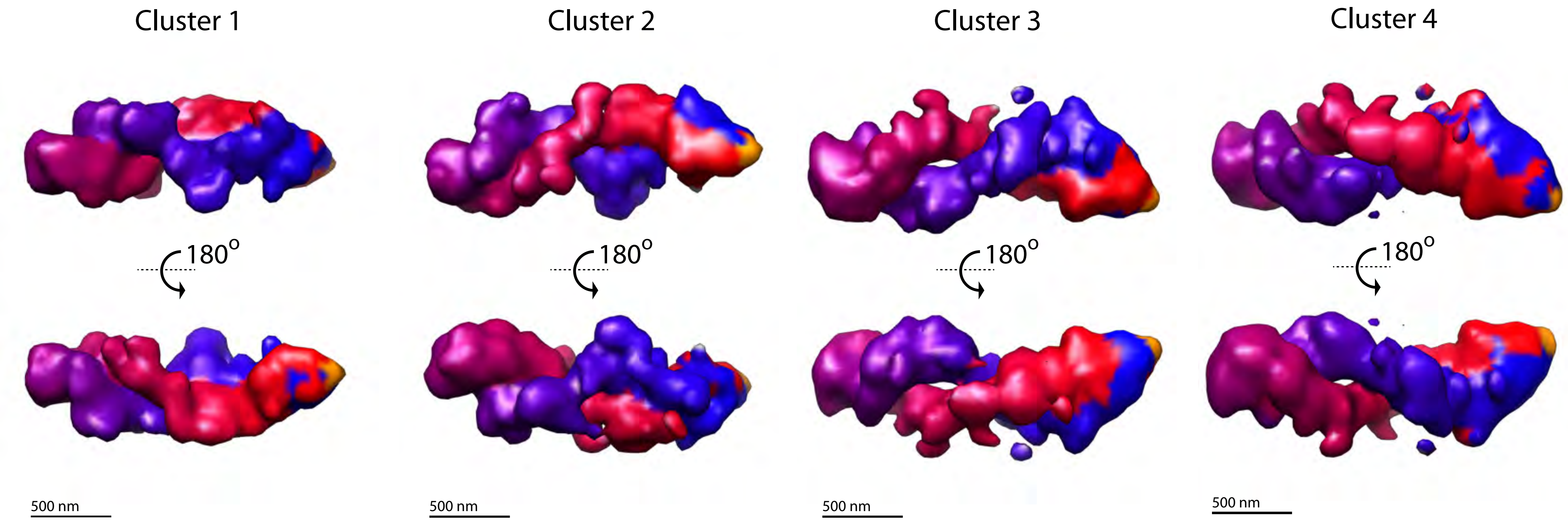




Genome organization in Caulobacter crescentus

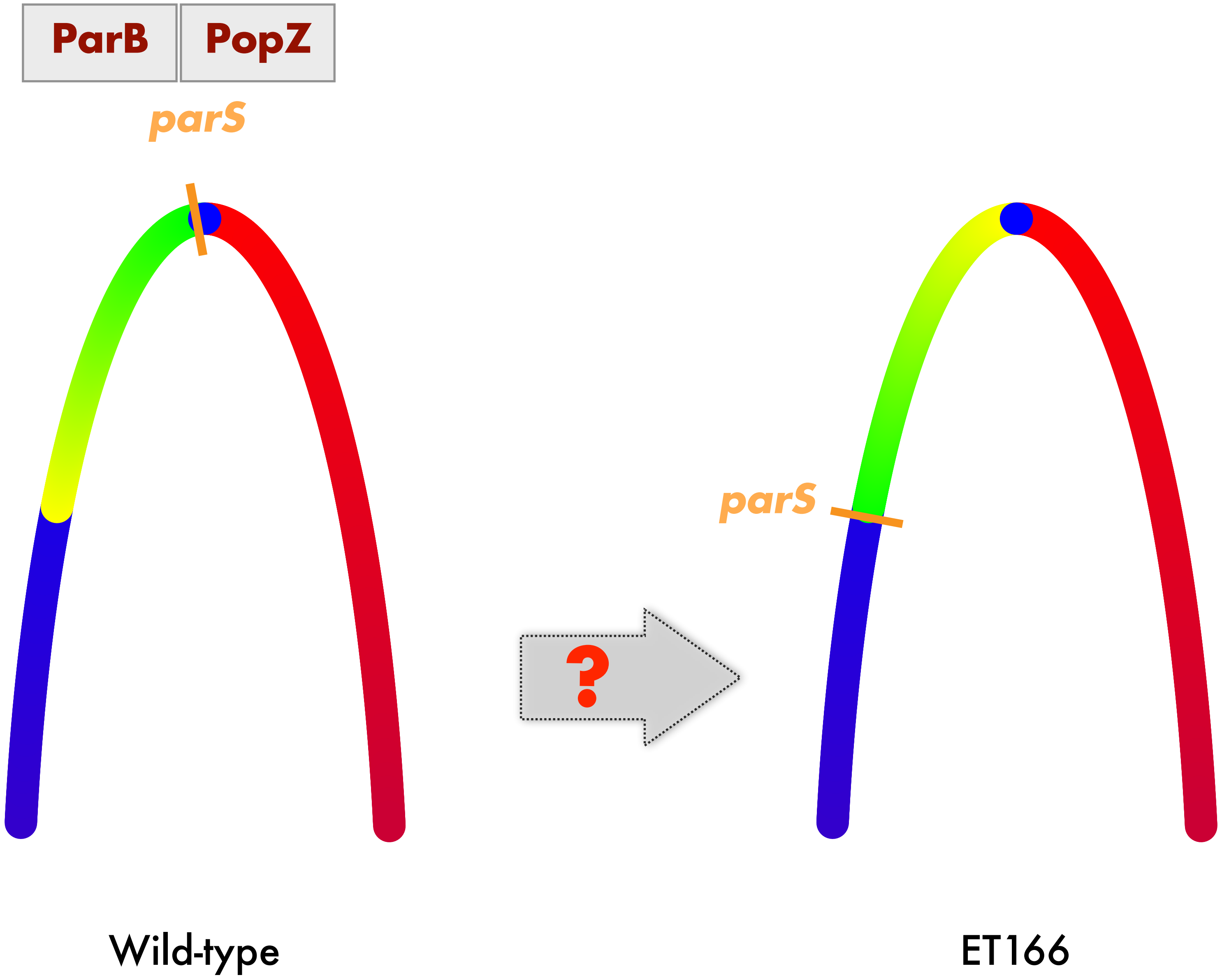
Arms are helical





MIRRORS!

Moving the parS sites 400 Kb away from Ori



ParB

PopZ

parS

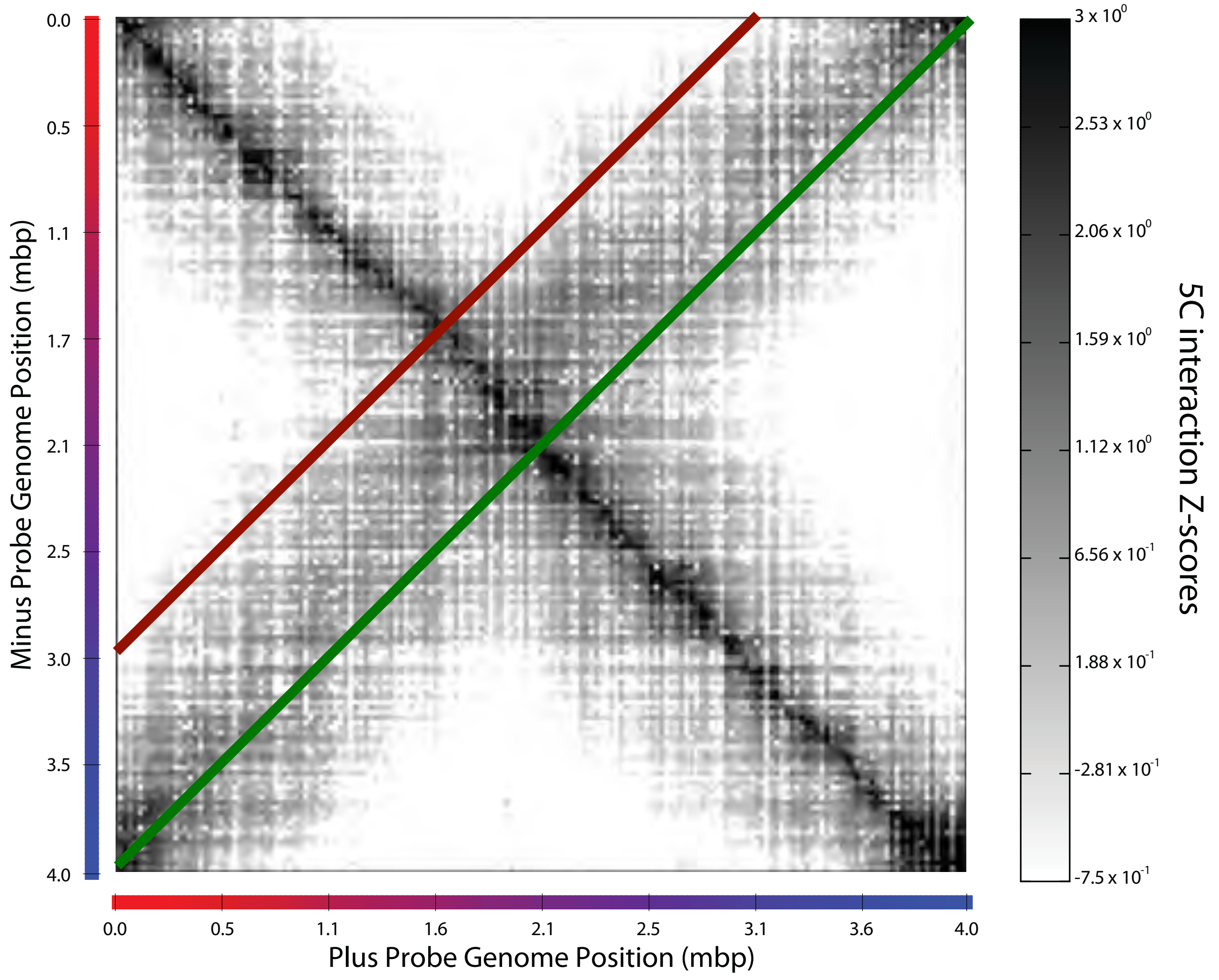
?

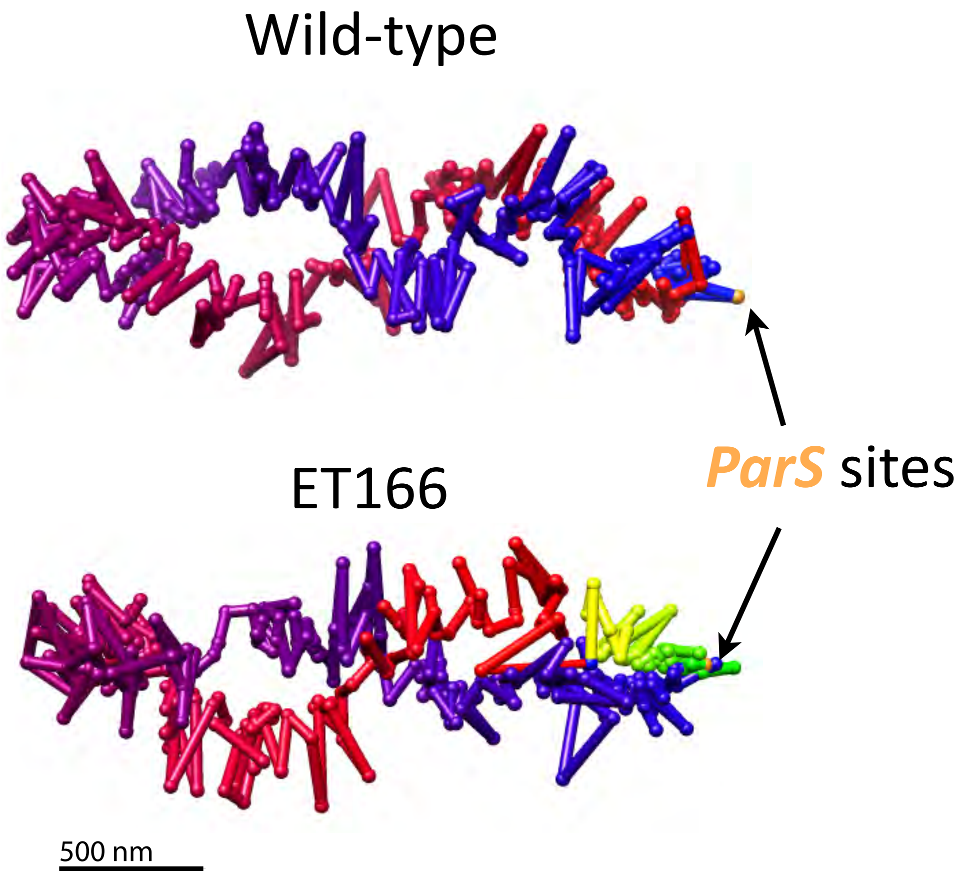
parS

Wild-type

ET166

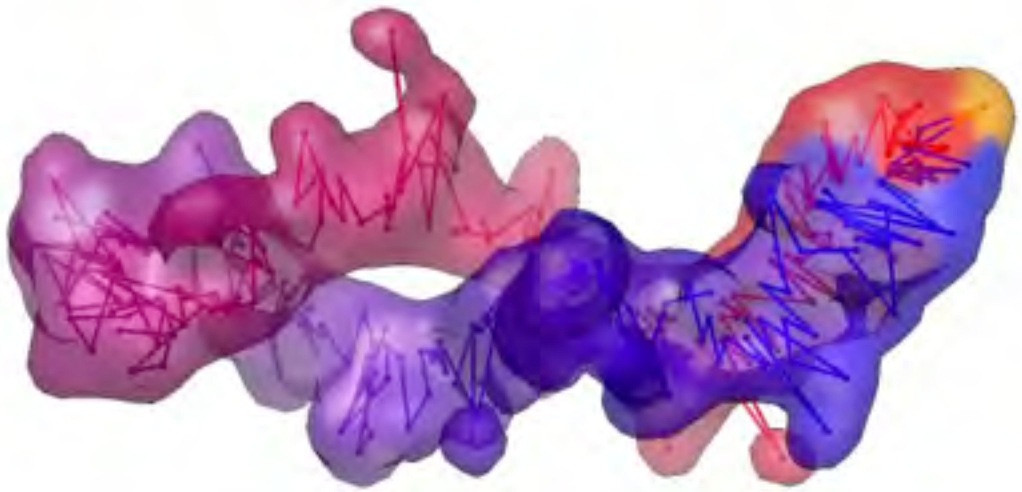
Moving the parS sites results in whole genome rotation!

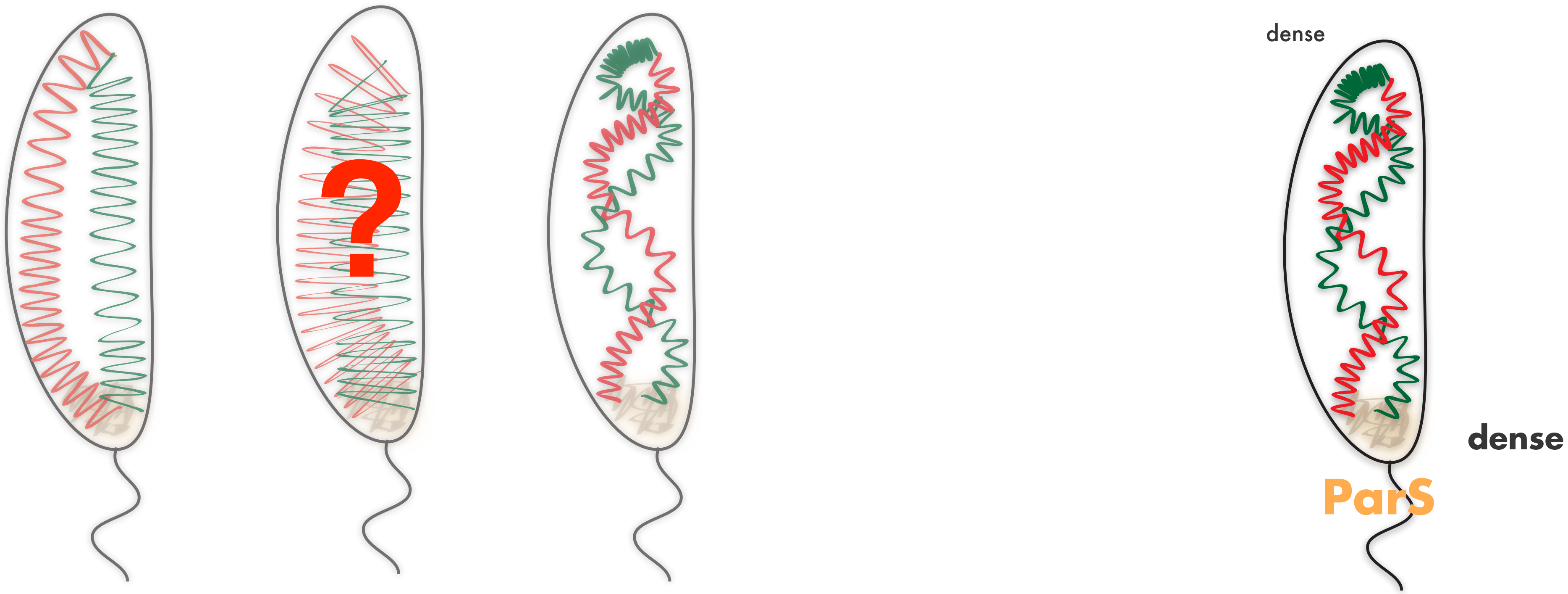




Arms are **STILL** helical

Genome architecture in Caulobacter



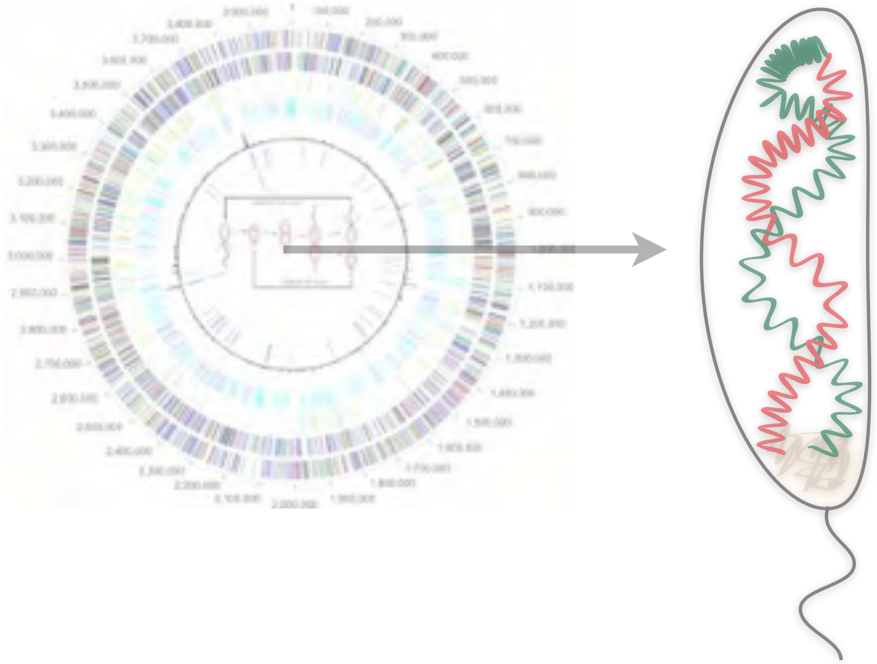


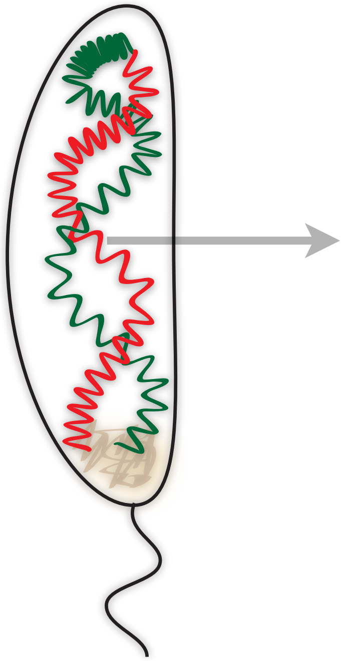
M.A. Umbarger, *et al.* **Molecular Cell** (2011) **44**:252–264

From Sequence to Function

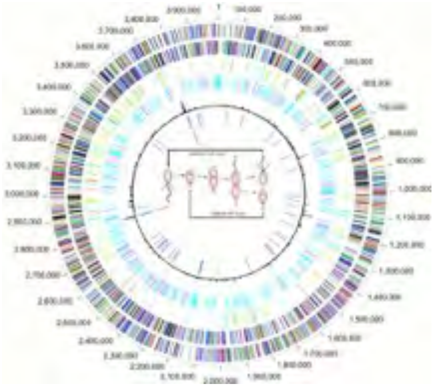
5C + IMP

Technology





Hypothesis

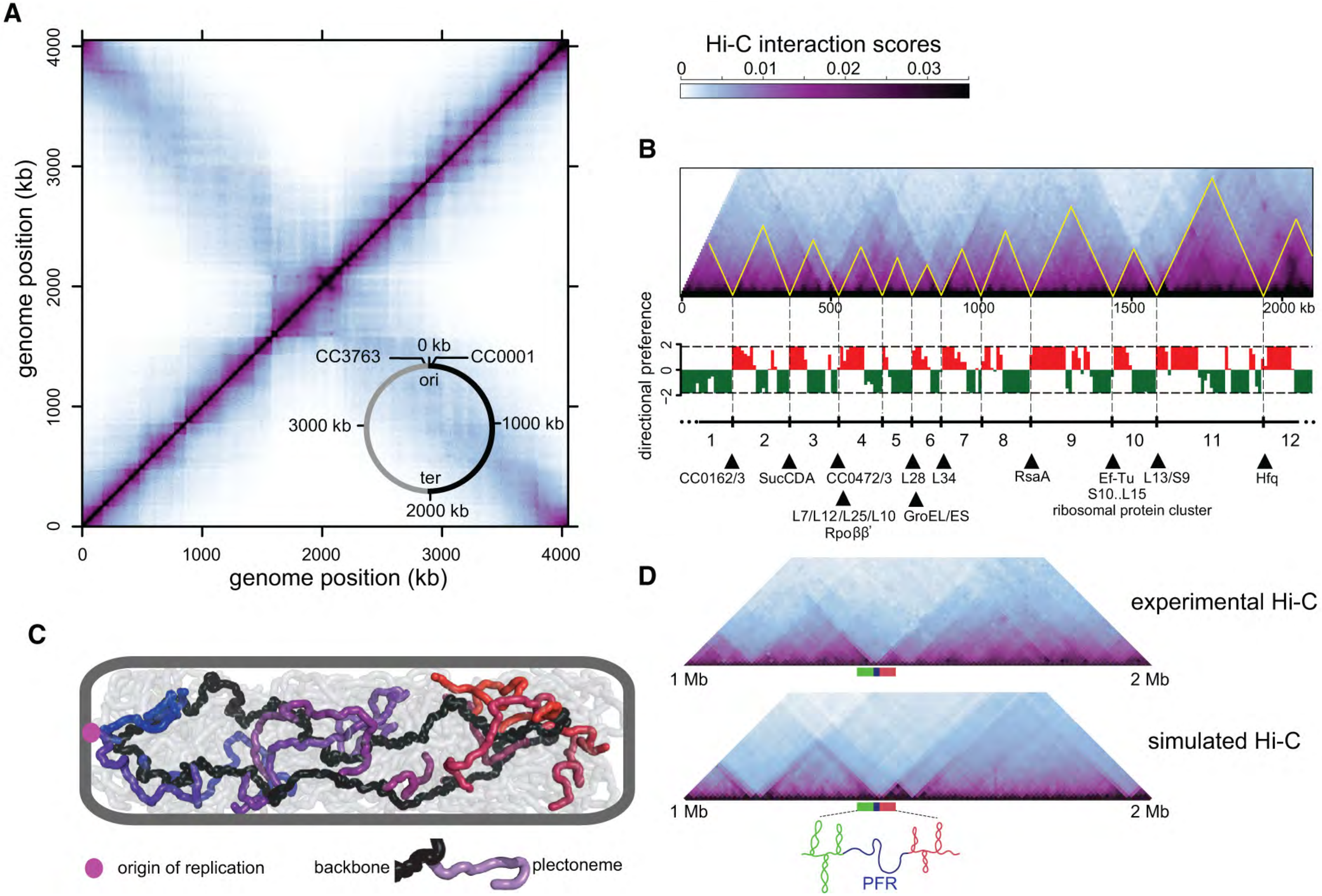


Function!

D. Baù and M.A. Marti-Renom **Chromosome Res** (2011) 19:25-35.

Bacteria has also TADs (CIDs)

Le, T. B. K., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. Science (New York, NY), 1242059

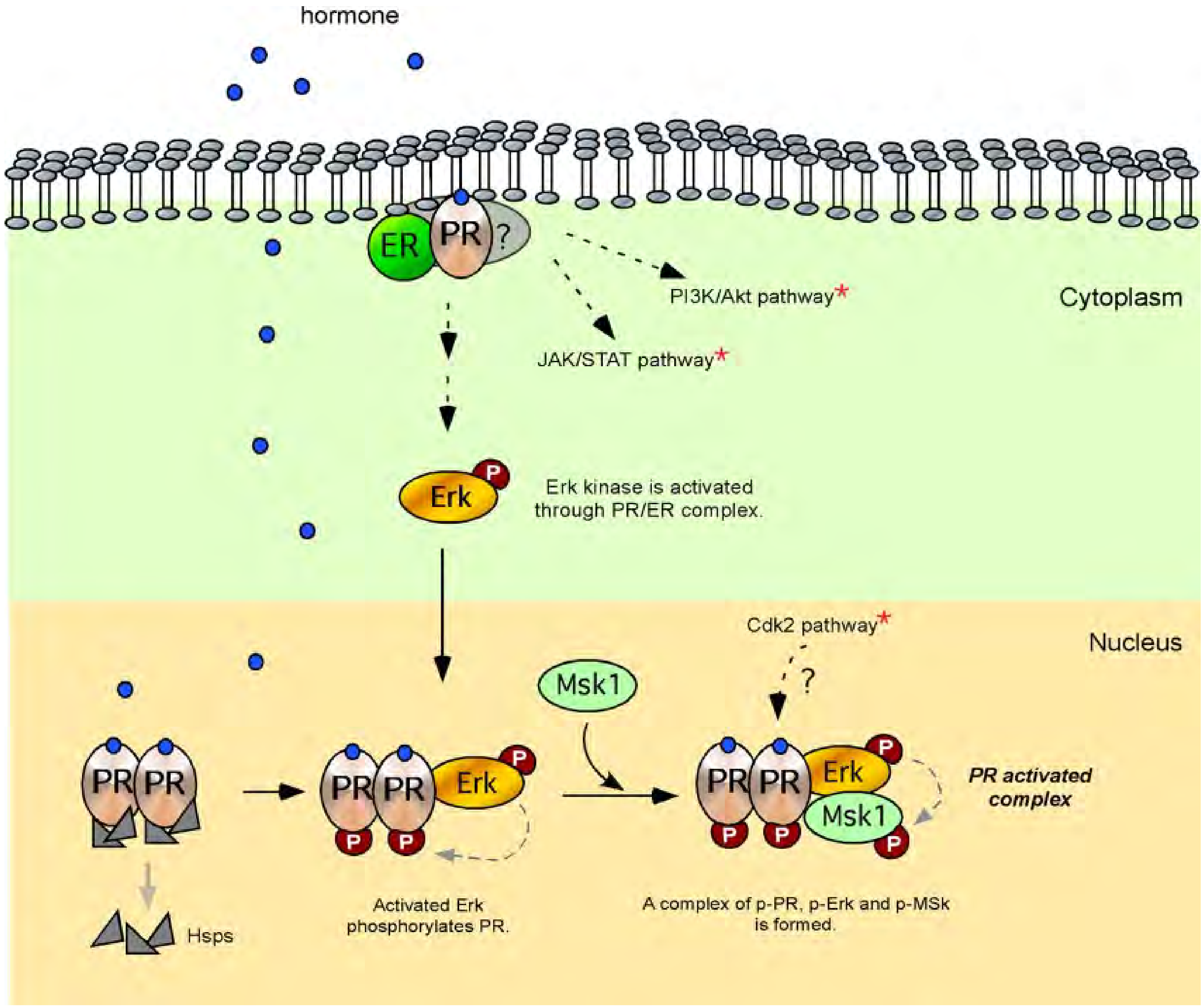


On TADs and hormones

Davide Baù

François le Dily

Progesterone-regulated transcription in breast cancer



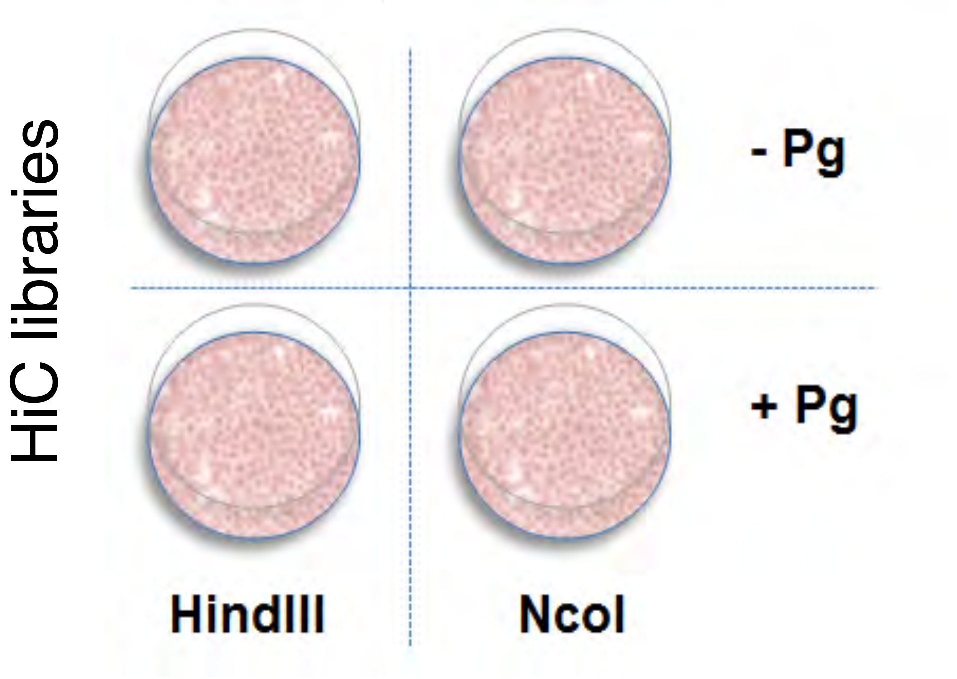
Vicent *et al* 2011, Wright *et al* 2012, Ballare *et al* 2012

> 2,000 genes **Up**-regulated

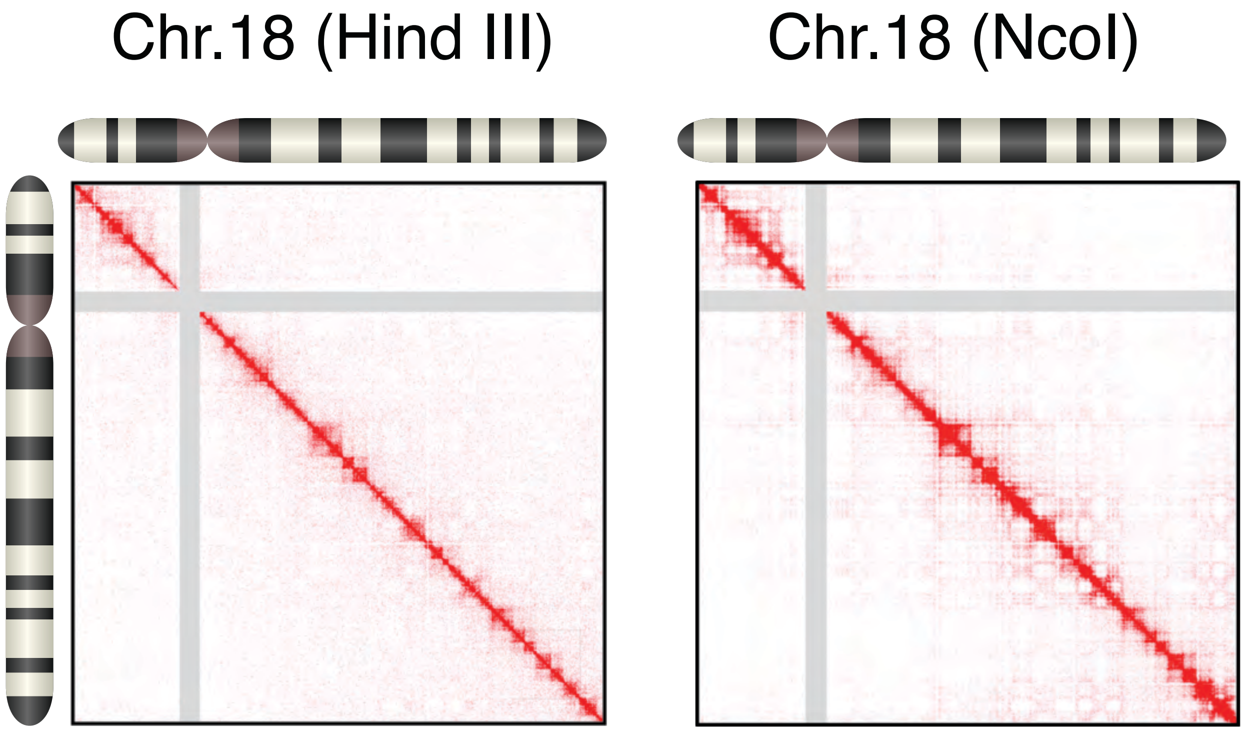
> 2,000 genes **Down**-regulated

Regulation in 3D?

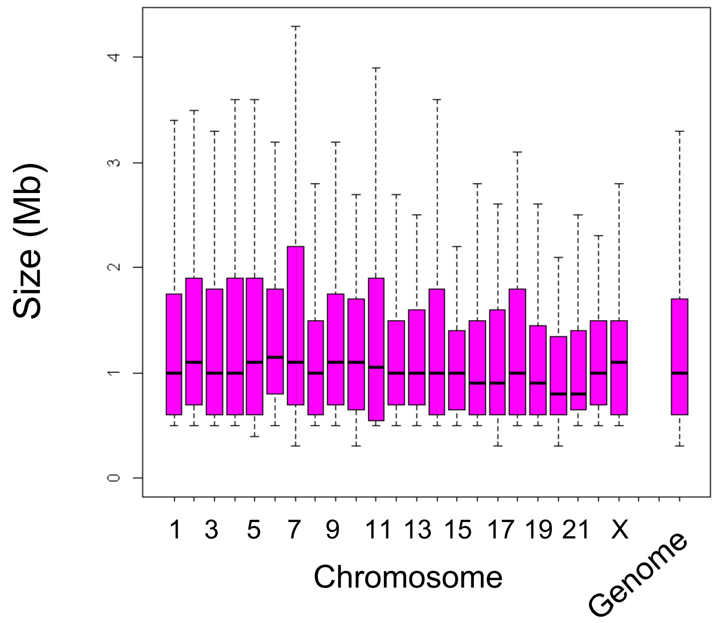
Experimental design



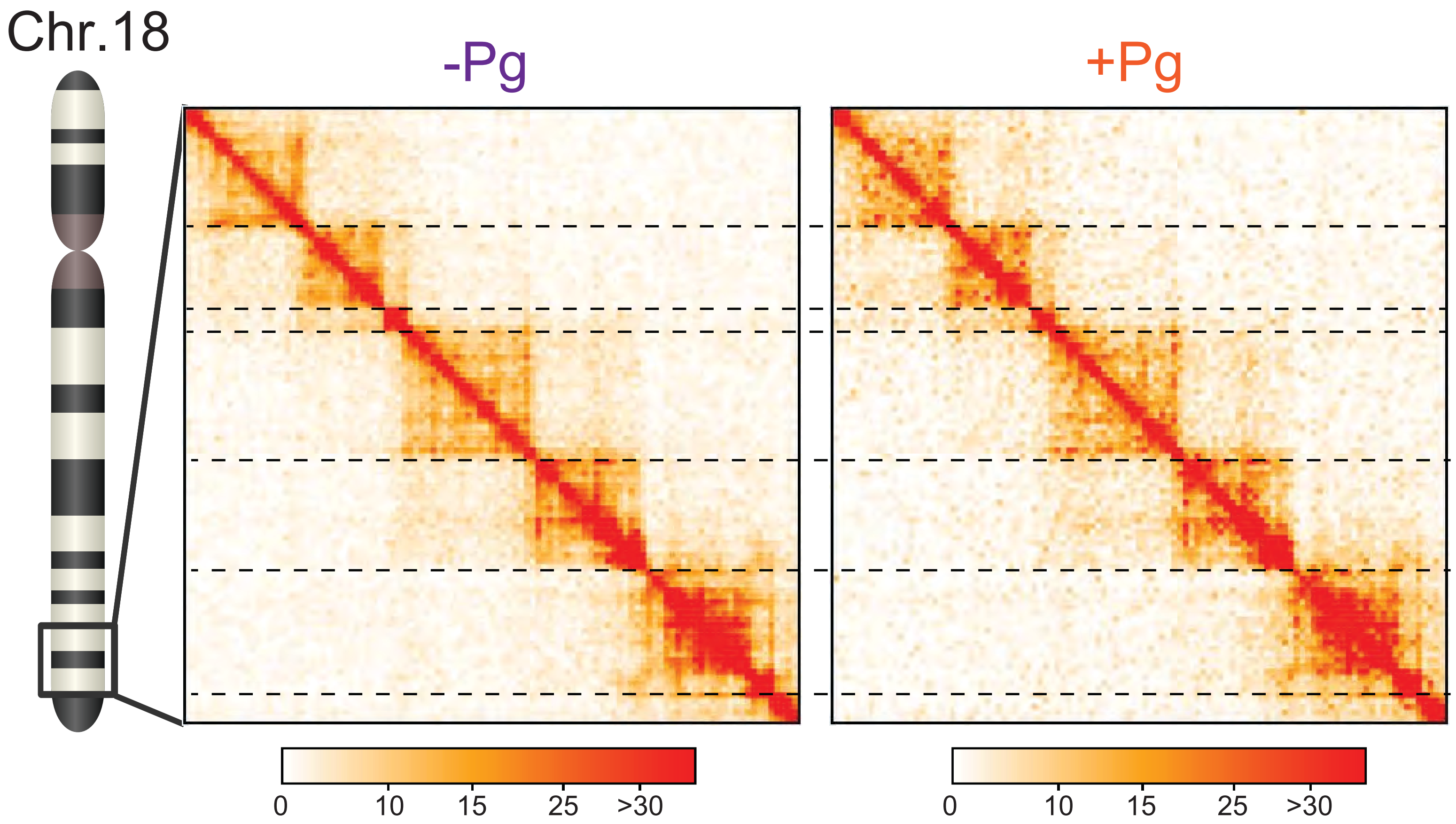
ChIP-Seq
RNA-Seq
Hi-C

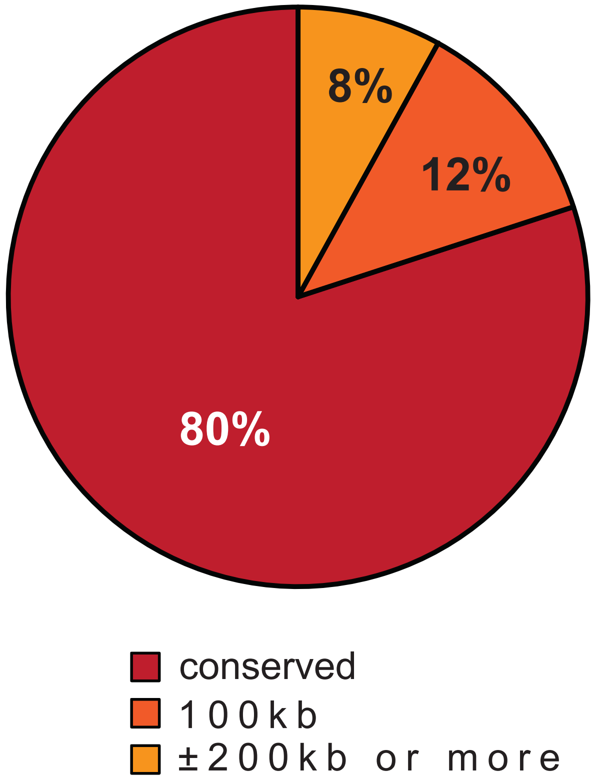


Are there TADs? how robust?

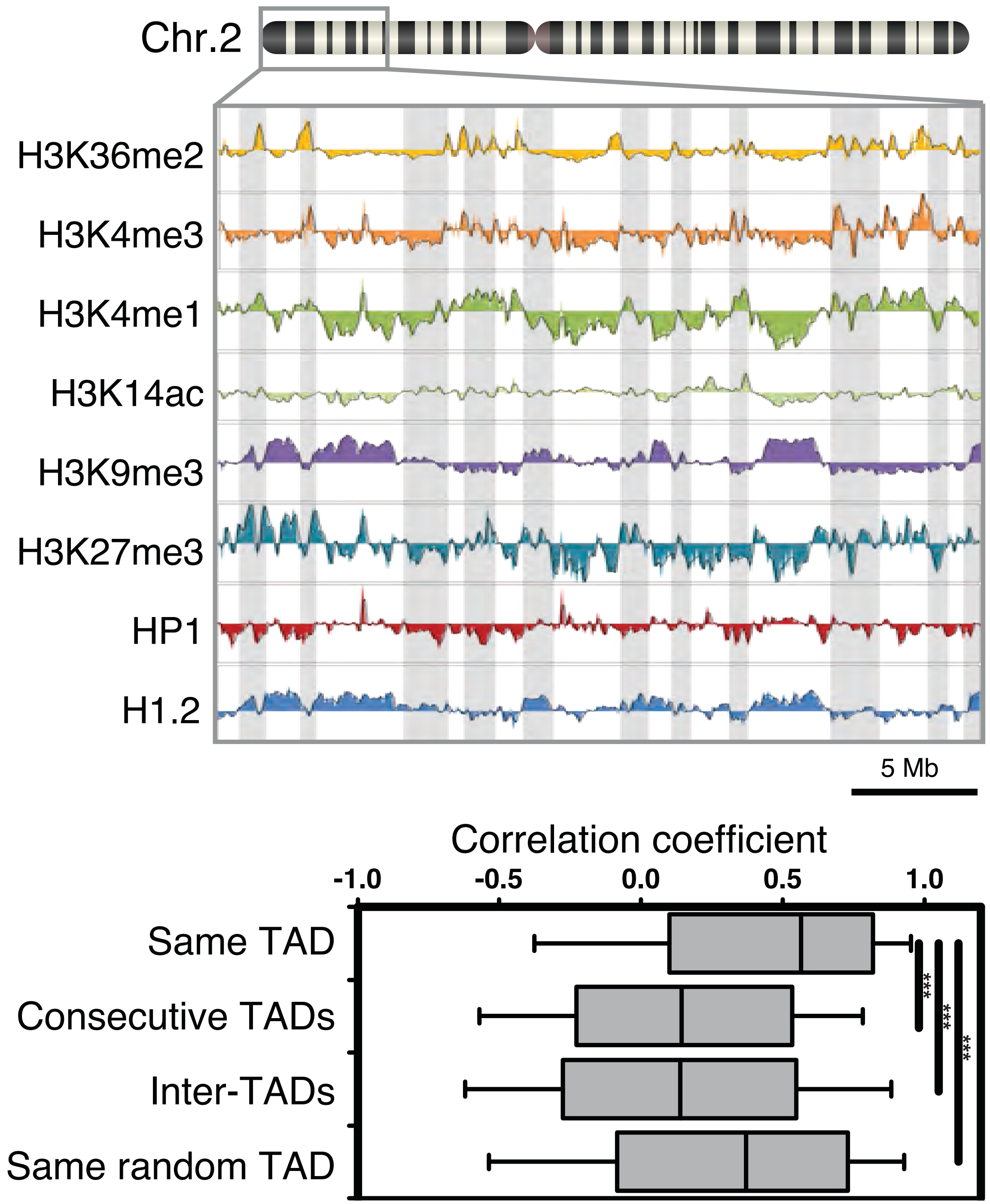


>2,000 detected TADs

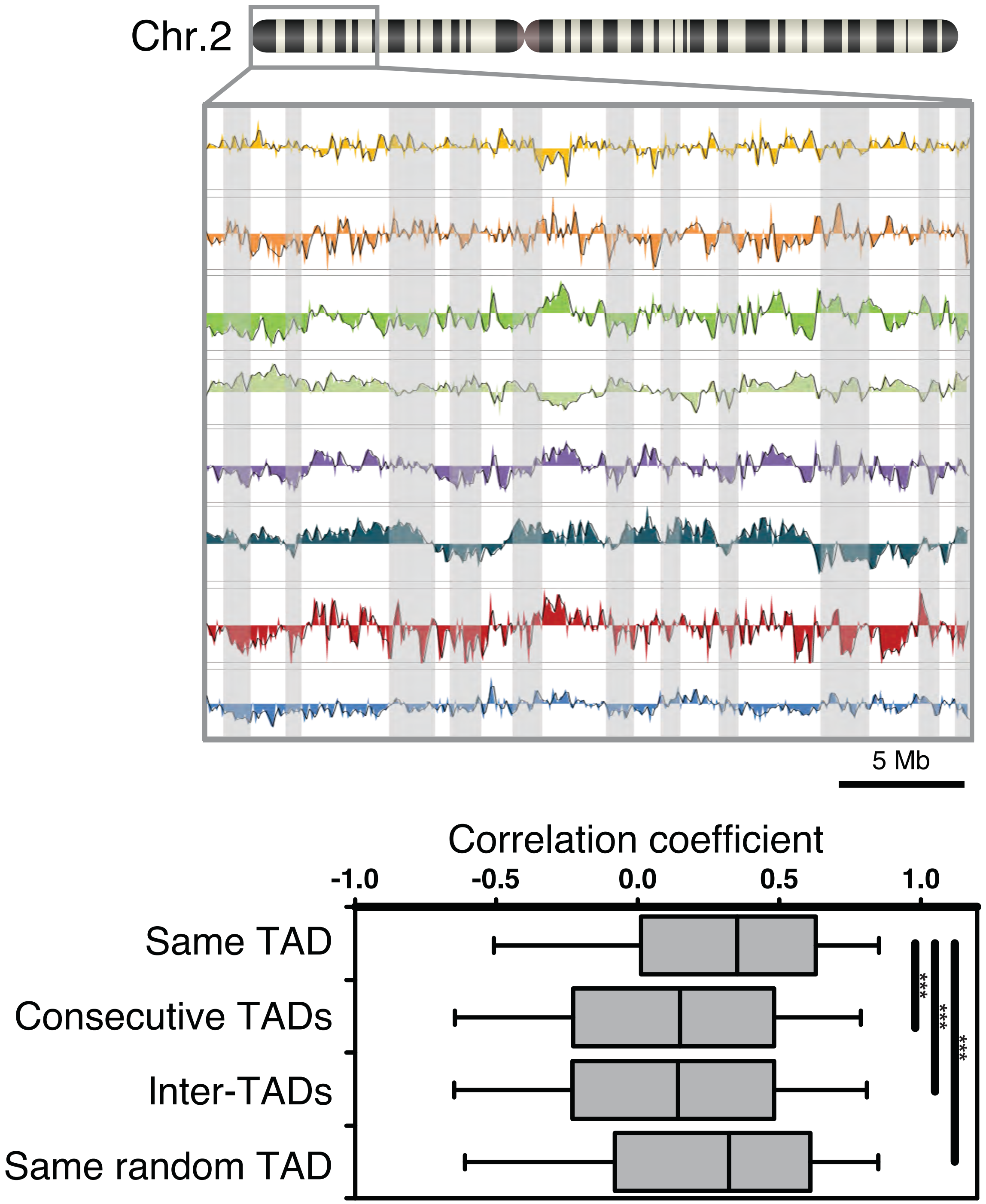




Are TADs homogeneous?

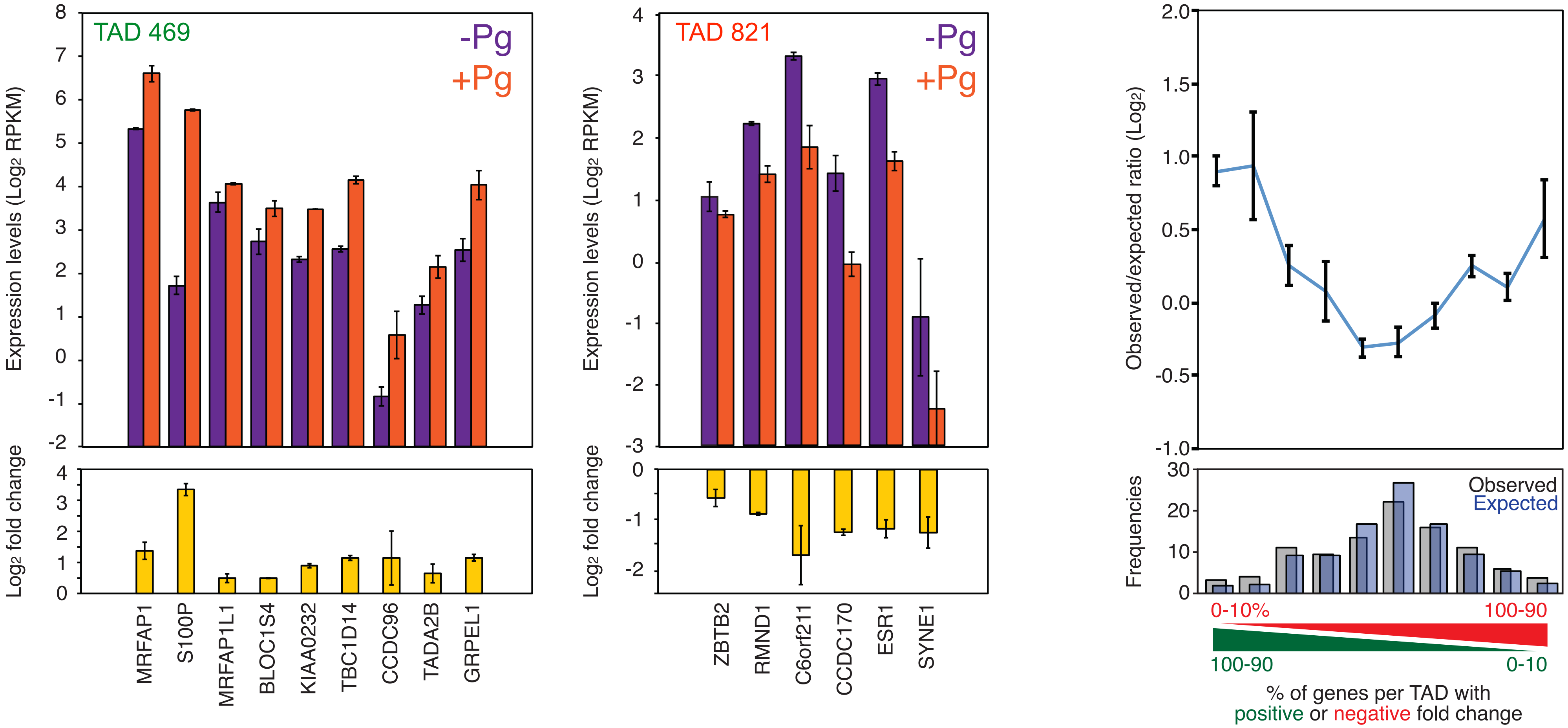


-Pg

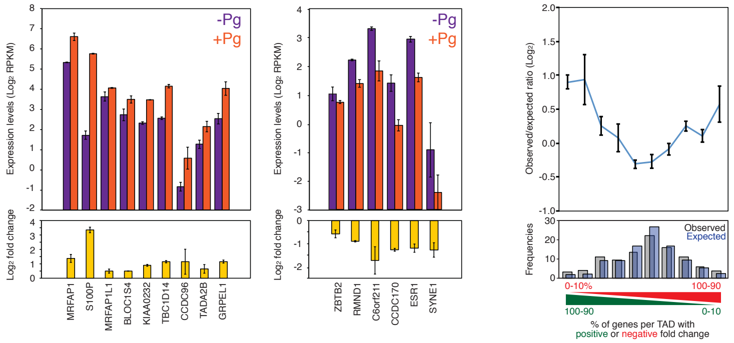


+Pg/-Pg

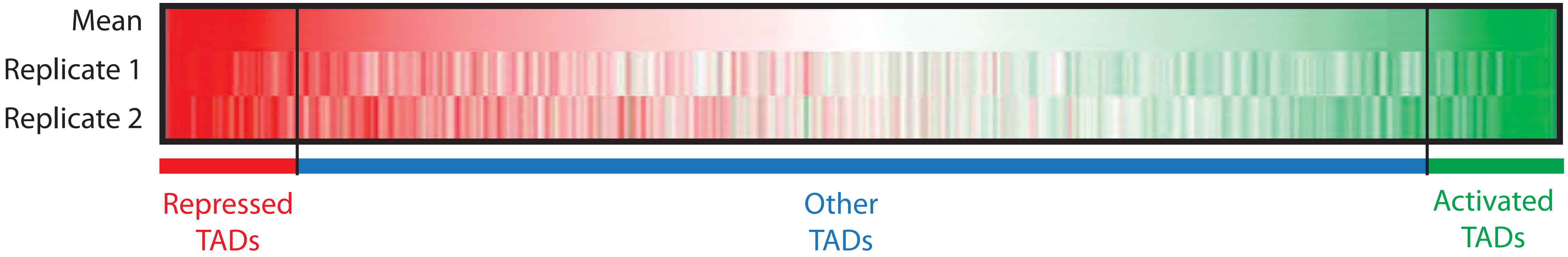
Do TADs respond differently to Pg treatment?

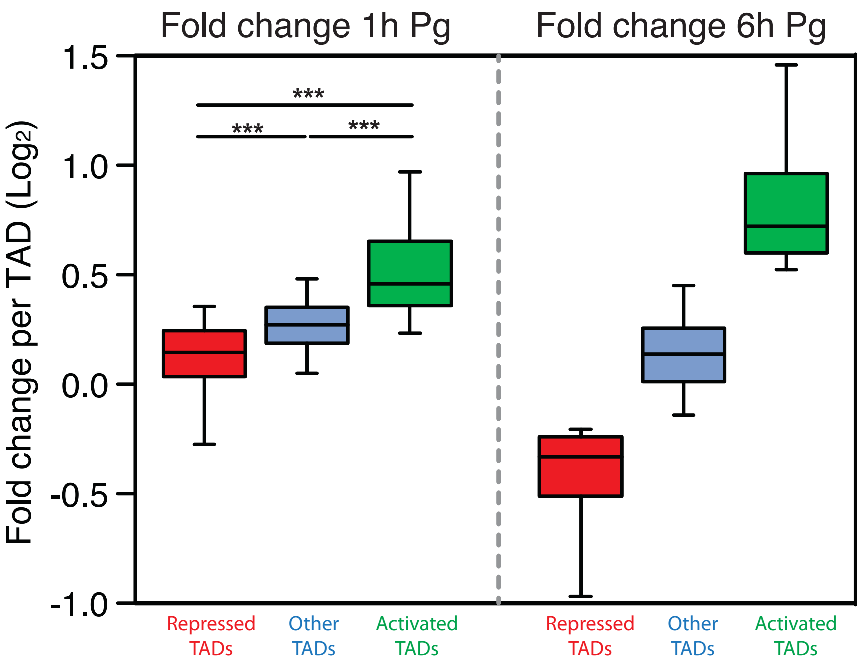


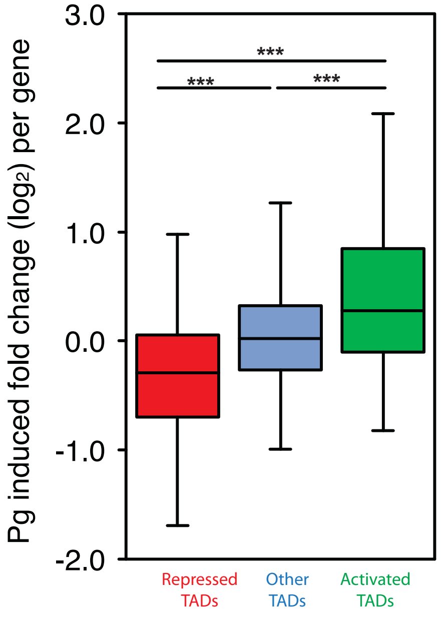
Do TADs respond differently to Pg treatment?

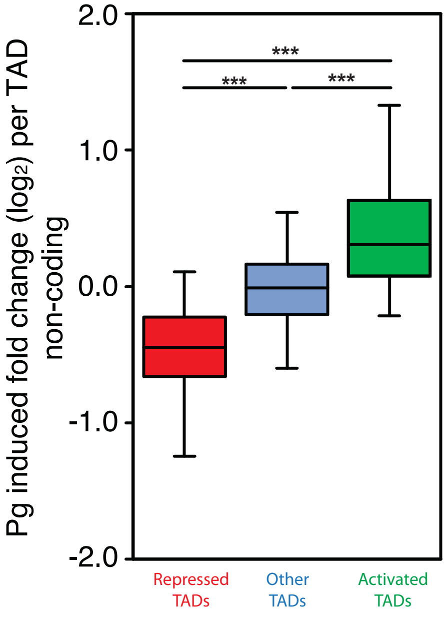


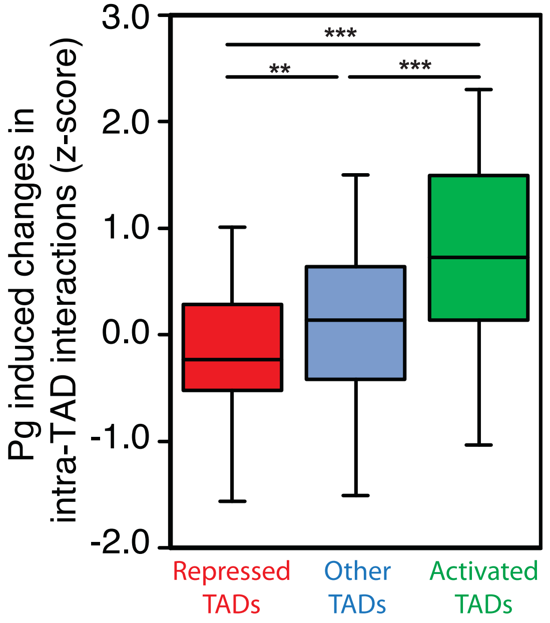
Pg induced fold change per TAD (6h)



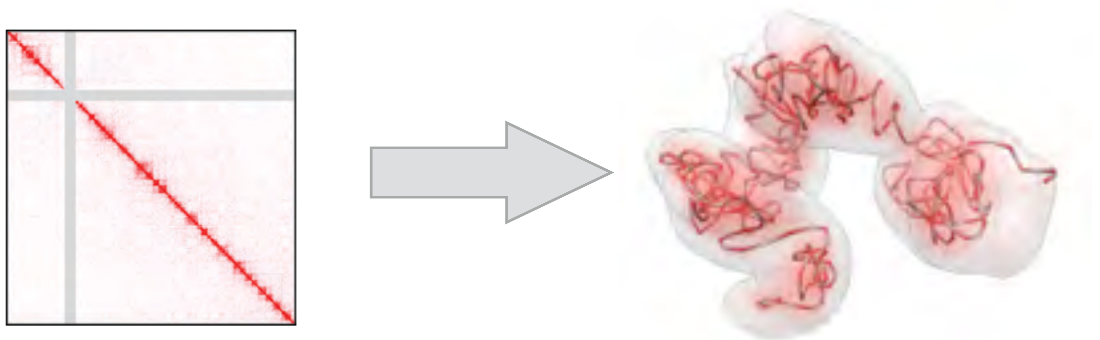


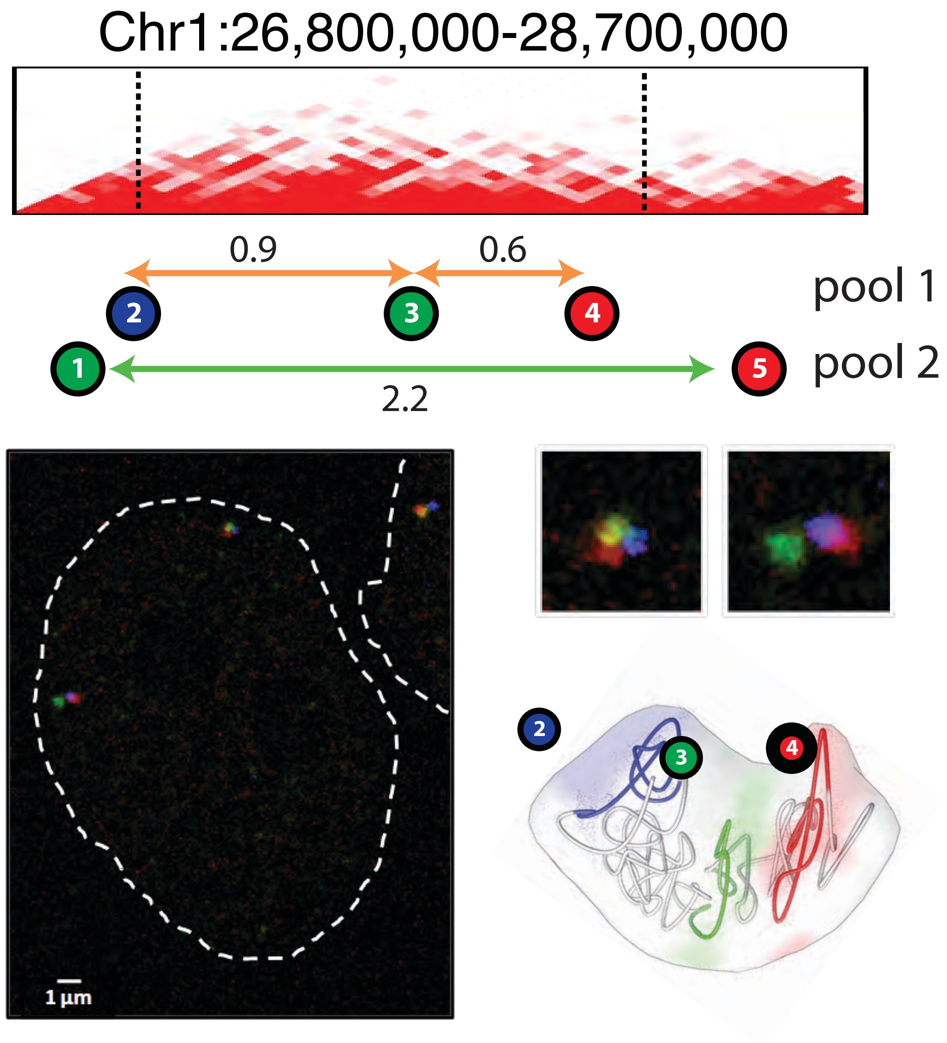


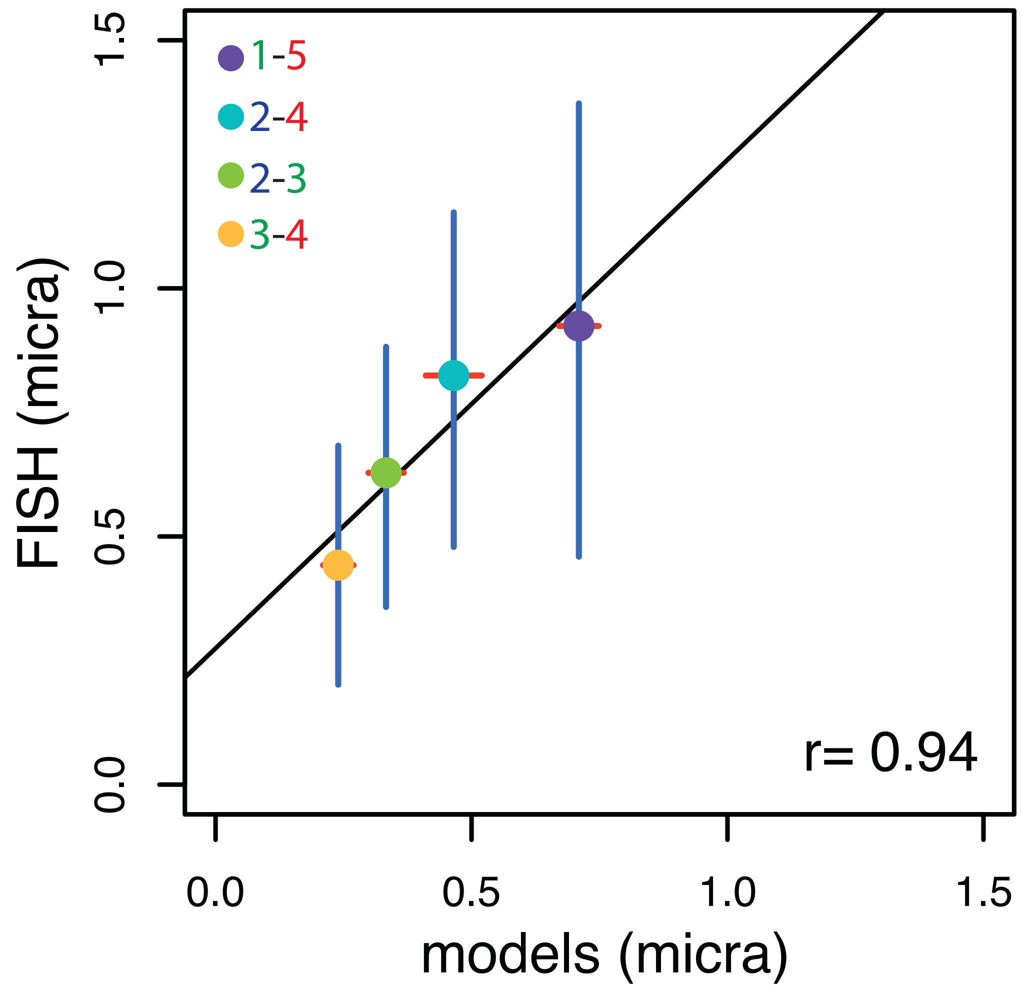




Modeling 3D TADs



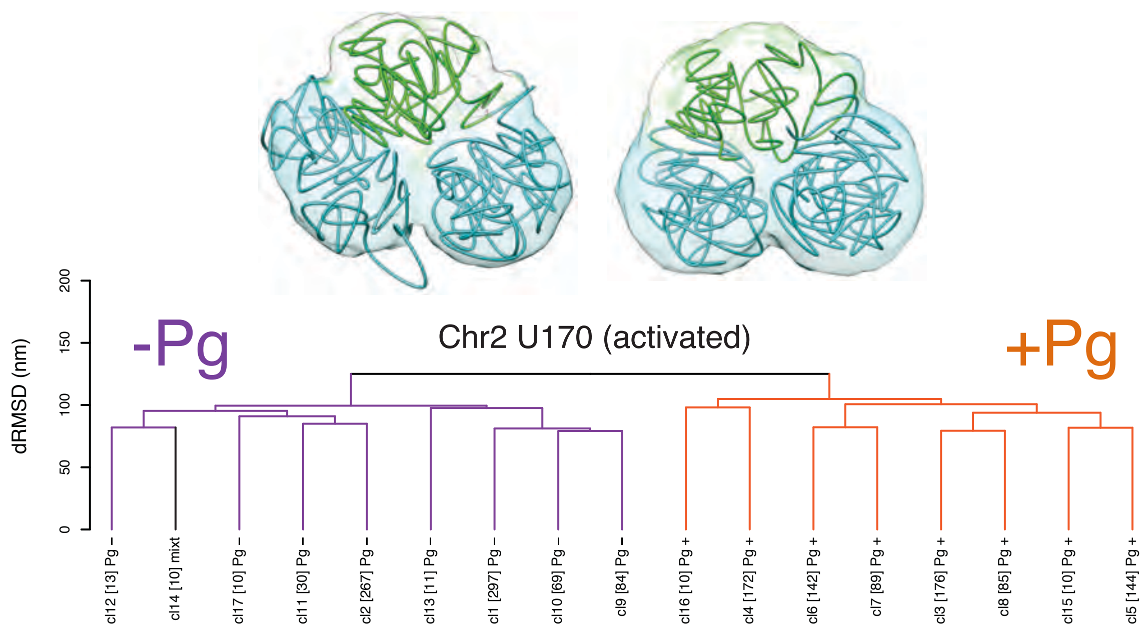




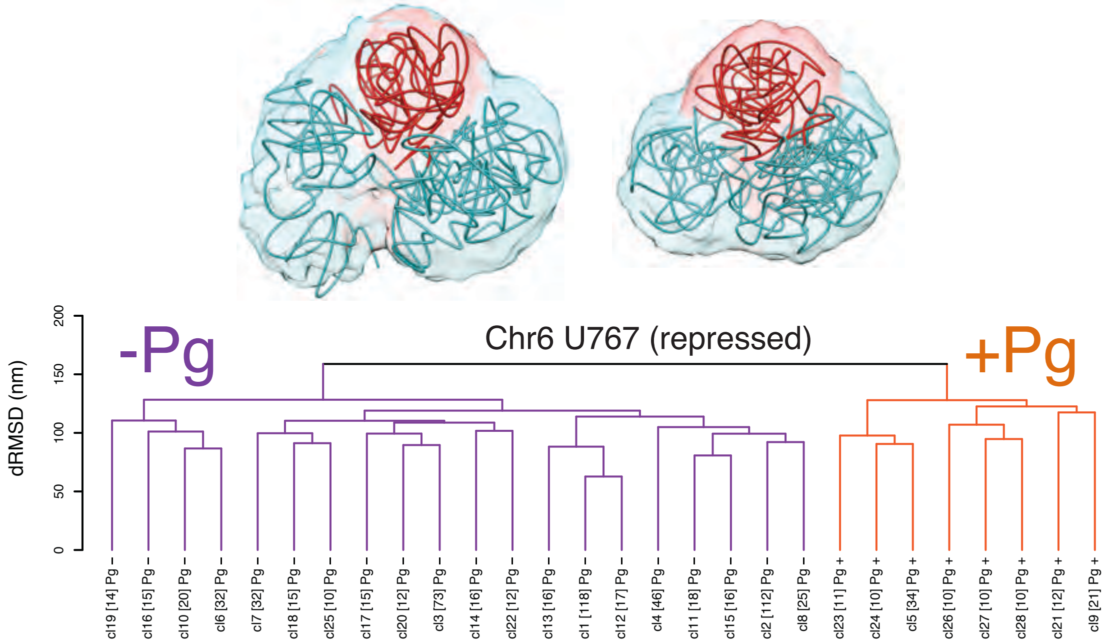
61 genomic regions containing 209 TADs covering 267Mb

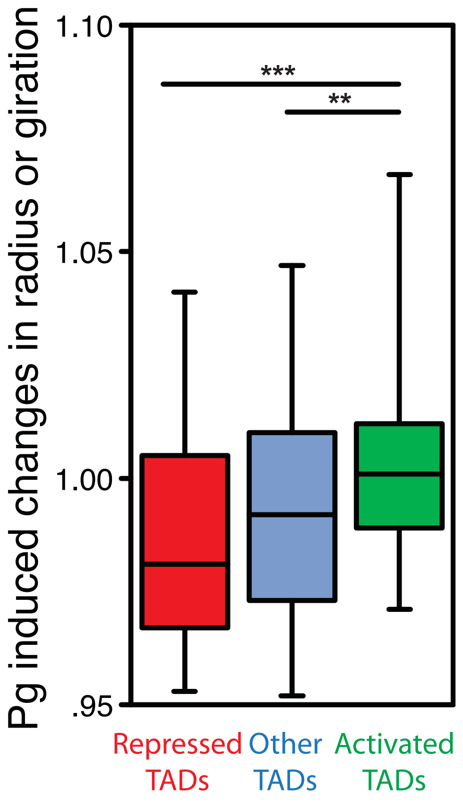
How TADs respond structurally to Pg?

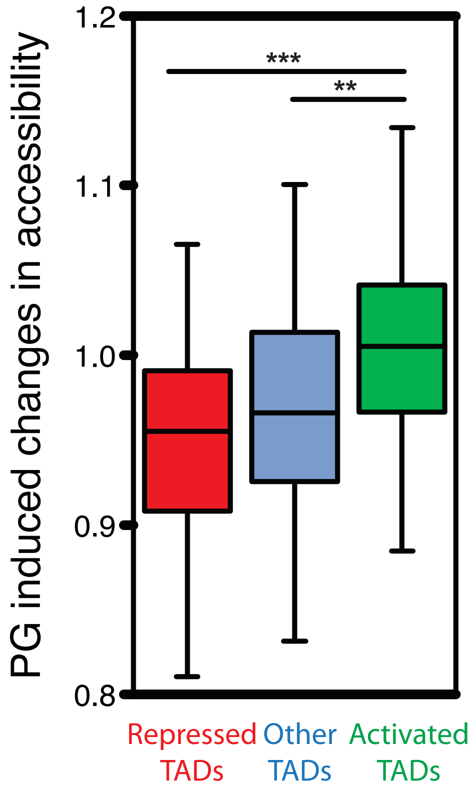
Chr2:9,600,000-13,200,000

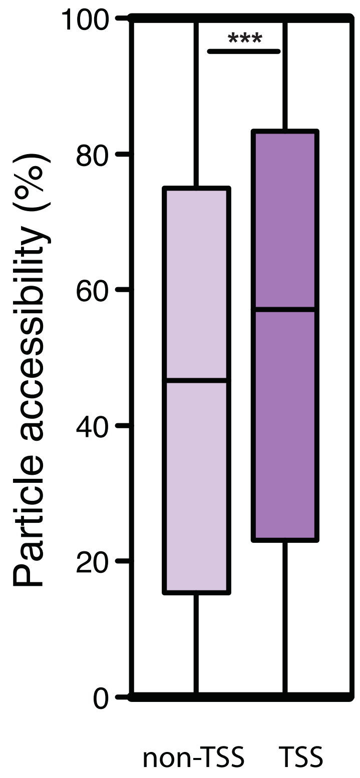


Chr6:71,800,000-76,500,000

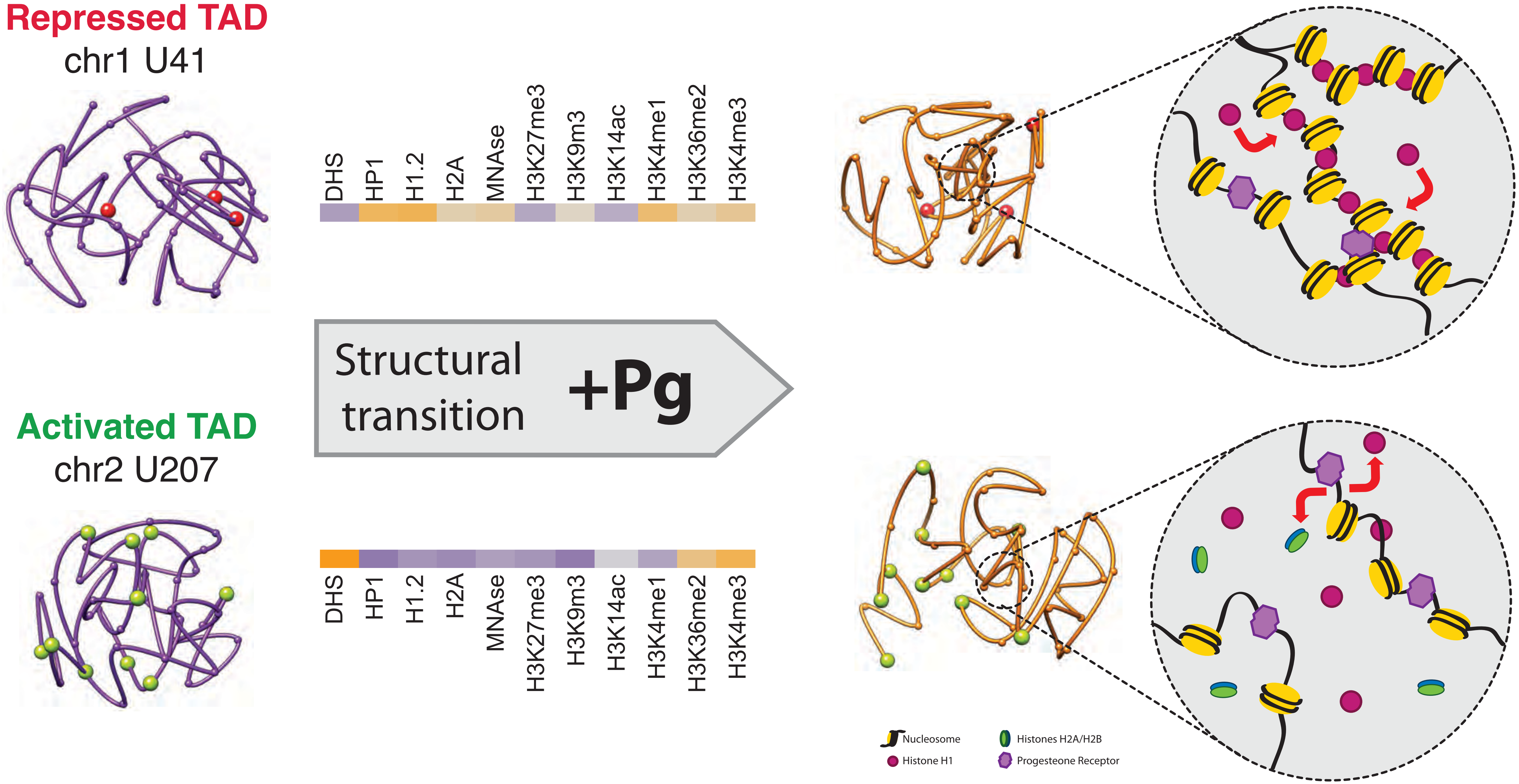








Model for TAD regulation



Repressed TAD
chr1 U41

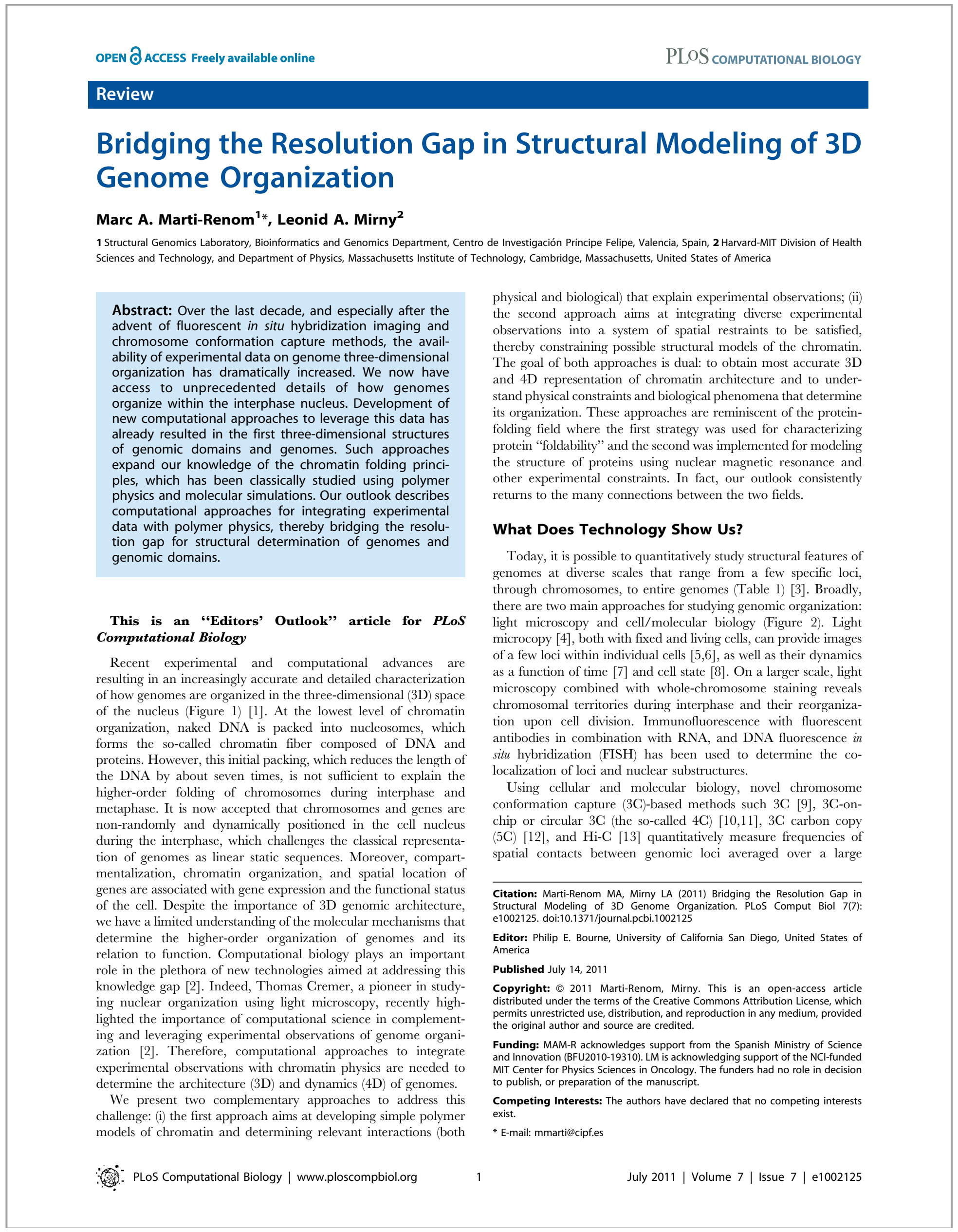
Activated TAD
chr2 U207

Structural transition +Pg

Nucleosome
Histones H2A/H2B
Histone H1
Progesteone Receptor

PLoS CB Outlook

Marti-Renom MA, Mirny LA (2011) PLoS Comput Biol 7(7): e1002125.



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization

Marc A. Marti-Renom[1]*, **Leonid A. Mirny**[2]

[1] Structural Genomics Laboratory, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain, [2] Harvard-MIT Division of Health Sciences and Technology, and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract: Over the last decade, and especially after the advent of fluorescent *in situ* hybridization imaging and chromosome conformation capture methods, the availability of experimental data on genome three-dimensional organization has dramatically increased. We now have access to unprecedented details of how genomes organize within the interphase nucleus. Development of new computational approaches to leverage this data has already resulted in the first three-dimensional structures of genomic domains and genomes. Such approaches expand our knowledge of the chromatin folding principles, which has been classically studied using polymer physics and molecular simulations. Our outlook describes computational approaches for integrating experimental data with polymer physics, thereby bridging the resolution gap for structural determination of genomes and genomic domains.

This is an ''Editors' Outlook'' article for *PLoS Computational Biology*

Recent experimental and computational advances are resulting in an increasingly accurate and detailed characterization of how genomes are organized in the three-dimensional (3D) space of the nucleus (Figure 1) [1]. At the lowest level of chromatin organization, naked DNA is packed into nucleosomes, which forms the so-called chromatin fiber composed of DNA and proteins. However, this initial packing, which reduces the length of the DNA by about seven times, is not sufficient to explain the higher-order folding of chromosomes during interphase and metaphase. It is now accepted that chromosomes and genes are non-randomly and dynamically positioned in the cell nucleus during the interphase, which challenges the classical representation of genomes as linear static sequences. Moreover, compartmentalization, chromatin organization, and spatial location of genes are associated with gene expression and the functional status of the cell. Despite the importance of 3D genomic architecture, we have a limited understanding of the molecular mechanisms that determine the higher-order organization of genomes and its relation to function. Computational biology plays an important role in the plethora of new technologies aimed at addressing this knowledge gap [2]. Indeed, Thomas Cremer, a pioneer in studying nuclear organization using light microscopy, recently highlighted the importance of computational science in complementing and leveraging experimental observations of genome organization [2]. Therefore, computational approaches to integrate experimental observations with chromatin physics are needed to determine the architecture (3D) and dynamics (4D) of genomes.

We present two complementary approaches to address this challenge: (i) the first approach aims at developing simple polymer models of chromatin and determining relevant interactions (both physical and biological) that explain experimental observations; (ii) the second approach aims at integrating diverse experimental observations into a system of spatial restraints to be satisfied, thereby constraining possible structural models of the chromatin. The goal of both approaches is dual: to obtain most accurate 3D and 4D representation of chromatin architecture and to understand physical constraints and biological phenomena that determine its organization. These approaches are reminiscent of the protein-folding field where the first strategy was used for characterizing protein ''foldability'' and the second was implemented for modeling the structure of proteins using nuclear magnetic resonance and other experimental constraints. In fact, our outlook consistently returns to the many connections between the two fields.

What Does Technology Show Us?

Today, it is possible to quantitatively study structural features of genomes at diverse scales that range from a few specific loci, through chromosomes, to entire genomes (Table 1) [3]. Broadly, there are two main approaches for studying genomic organization: light microscopy and cell/molecular biology (Figure 2). Light microcopy [4], both with fixed and living cells, can provide images of a few loci within individual cells [5,6], as well as their dynamics as a function of time [7] and cell state [8]. On a larger scale, light microscopy combined with whole-chromosome staining reveals chromosomal territories during interphase and their reorganization upon cell division. Immunofluorescence with fluorescent antibodies in combination with RNA, and DNA fluorescence *in situ* hybridization (FISH) has been used to determine the co-localization of loci and nuclear substructures.

Using cellular and molecular biology, novel chromosome conformation capture (3C)-based methods such 3C [9], 3C-on-chip or circular 3C (the so-called 4C) [10,11], 3C carbon copy (5C) [12], and Hi-C [13] quantitatively measure frequencies of spatial contacts between genomic loci averaged over a large

Citation: Marti-Renom MA, Mirny LA (2011) Bridging the Resolution Gap in Structural Modeling of 3D Genome Organization. PLoS Comput Biol 7(7): e1002125. doi:10.1371/journal.pcbi.1002125

Editor: Philip E. Bourne, University of California San Diego, United States of America

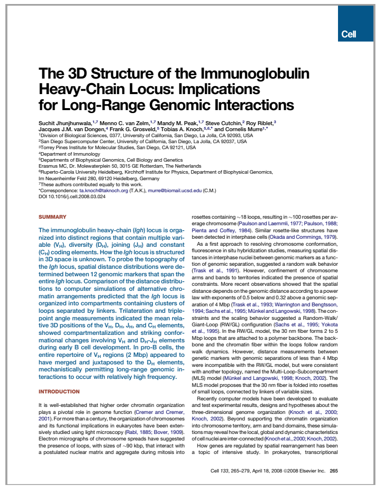
Published July 14, 2011

Funding: MAM-R acknowledges support from the Spanish Ministry of Science and Innovation (BFU2010-19310). LM is acknowledging support of the NCI-funded MIT Center for Physics Sciences in Oncology. The funders had no role in decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

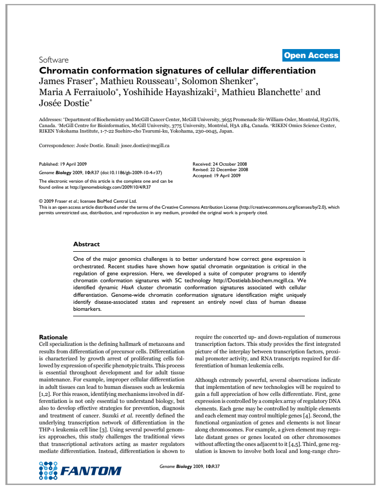
* E-mail: mmarti@cipf.es

MURRE
Cell (2008) **133**:265-79



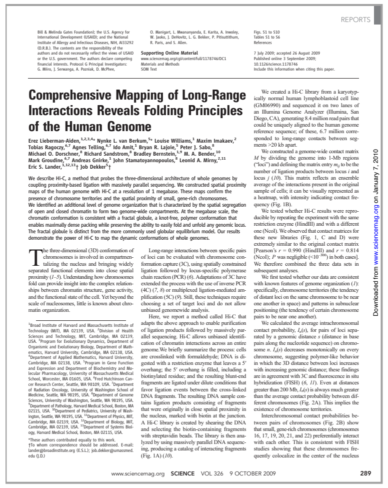
The 3D Structure of the Immunoglobulin Heavy-Chain Locus: Implications for Long-Range Genomic Interactions

DOSTIE/BLANCHETTE
Genome Biol (2009) **10**: R37



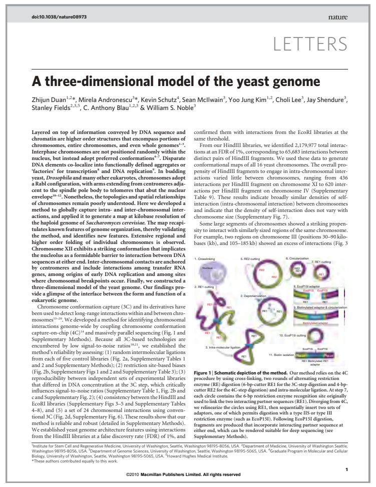
Chromatin conformation signatures of cellular differentiation

DEKKER/LANDER/MIRNY
Science (2009) **326**:289-93



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

NOBLE
Nature (2010) **465**: 363-7

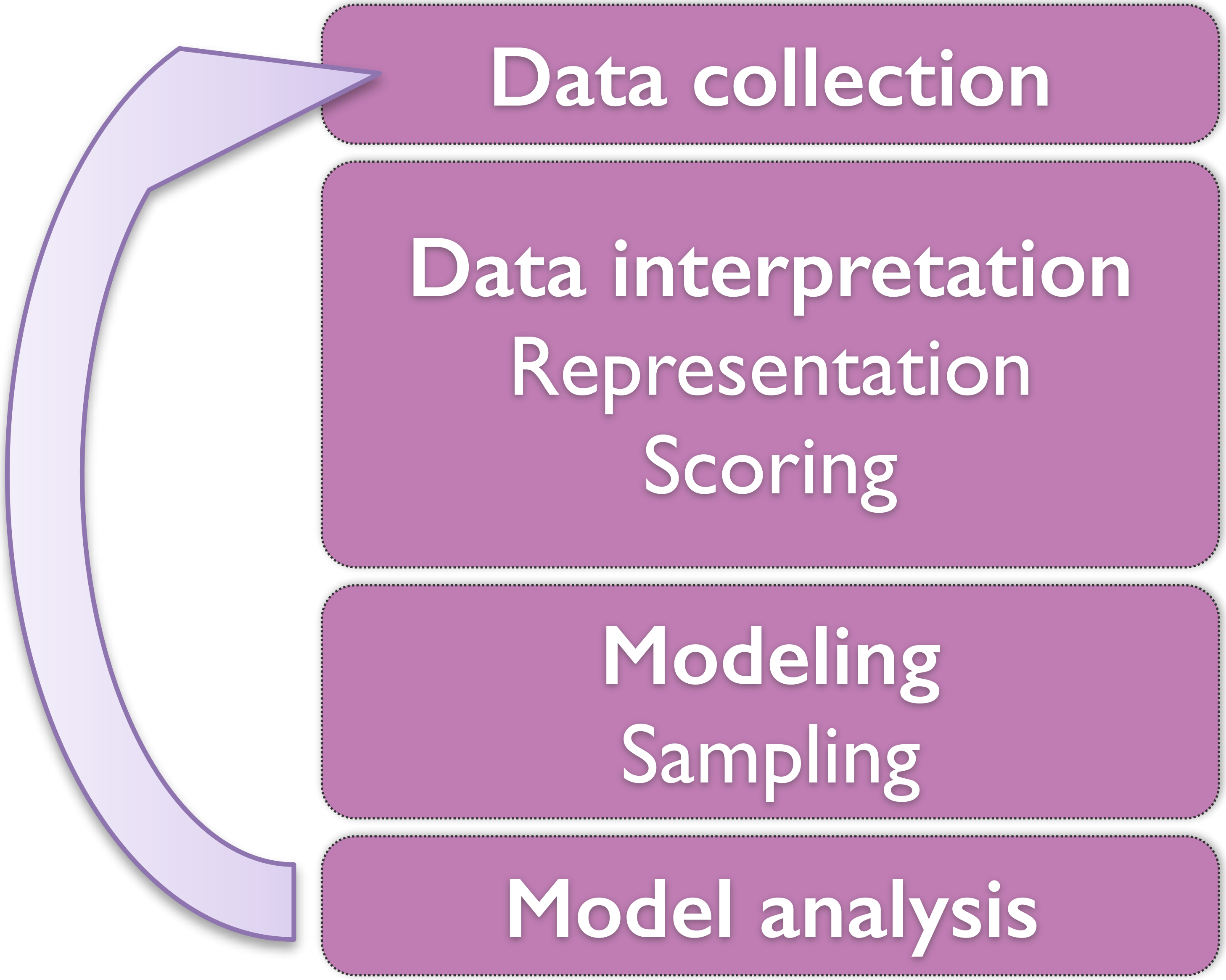


A three-dimensional model of the yeast genome

DEKKER/MARTI-RENOM
NSMB (2011) **18**:107-14

The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules

Take home message



Data collection

Data interpretation
Representation
Scoring

Modeling
Sampling

Model analysis

Acknowledgments

http://marciuslab.org
http://3DGenomes.org
http://cnag.crg.eu