Data integration for 3D structure determination.

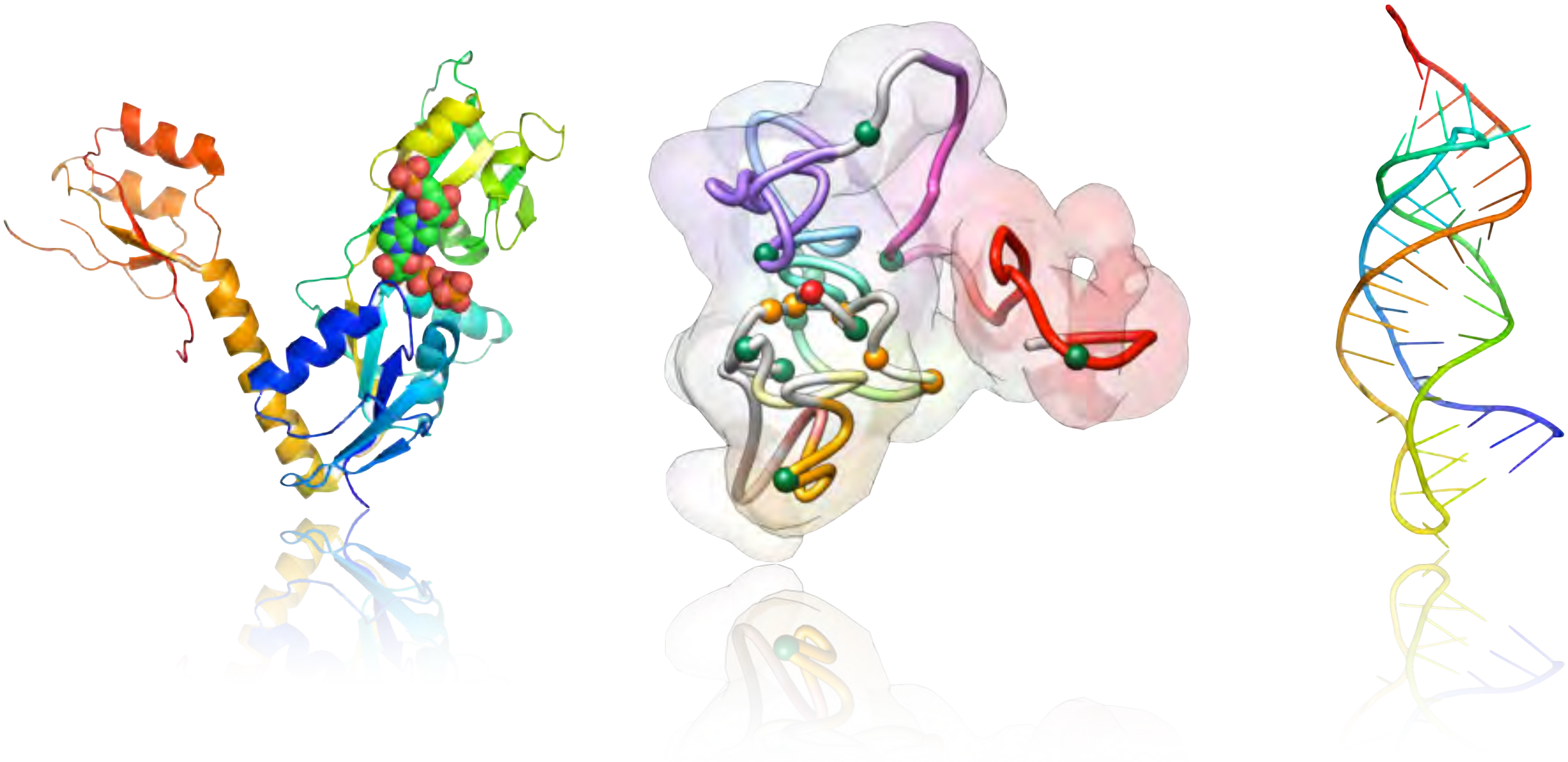
Marc A. Marti-Renom
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

ICREA

cnag CRG Centre for Genomic Regulation

Structural Genomics Group

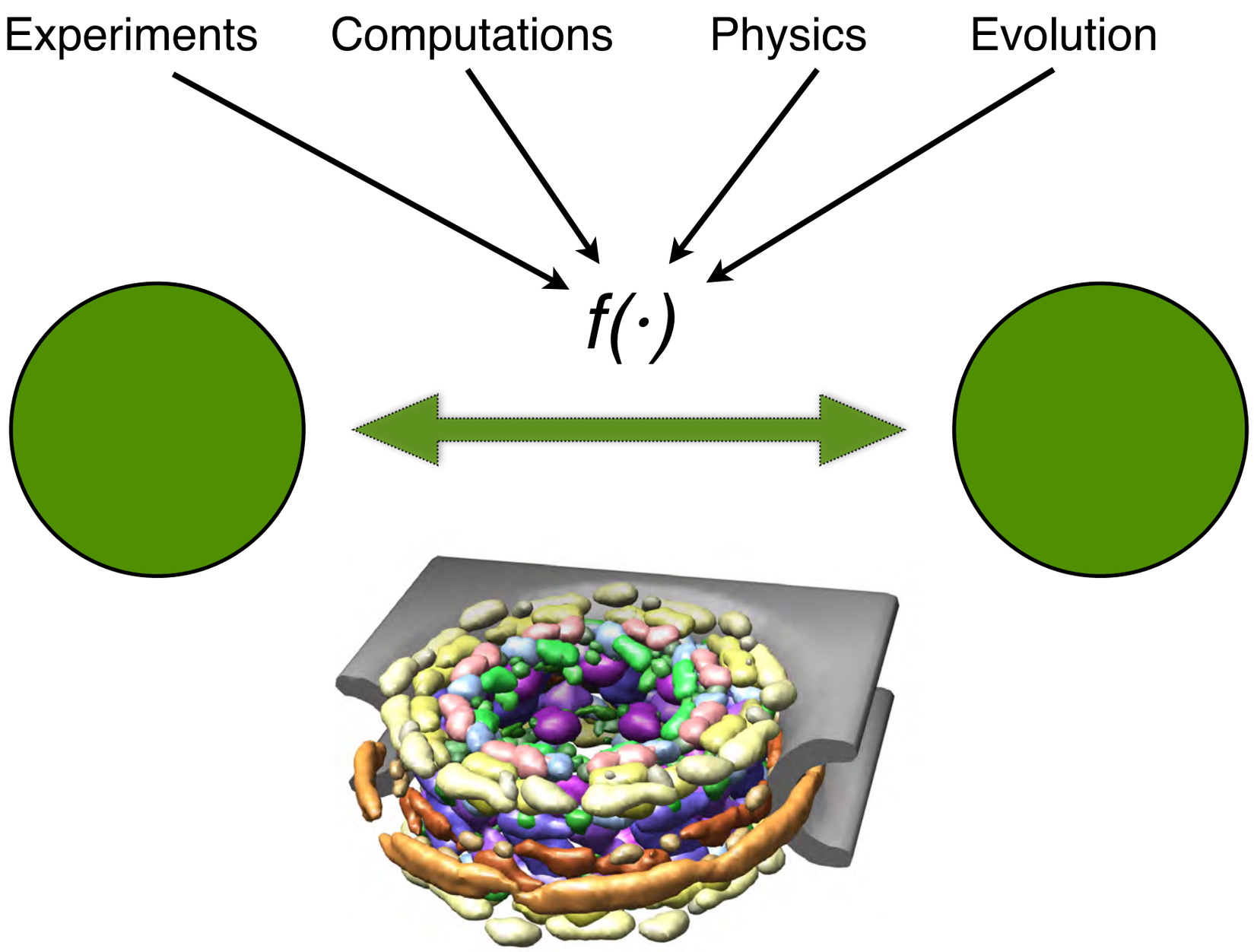
http://www.marciuslab.org



cnag CRG Centre for Genomic Regulation

Integrative Modeling Platform

http://www.integrativemodeling.org



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.

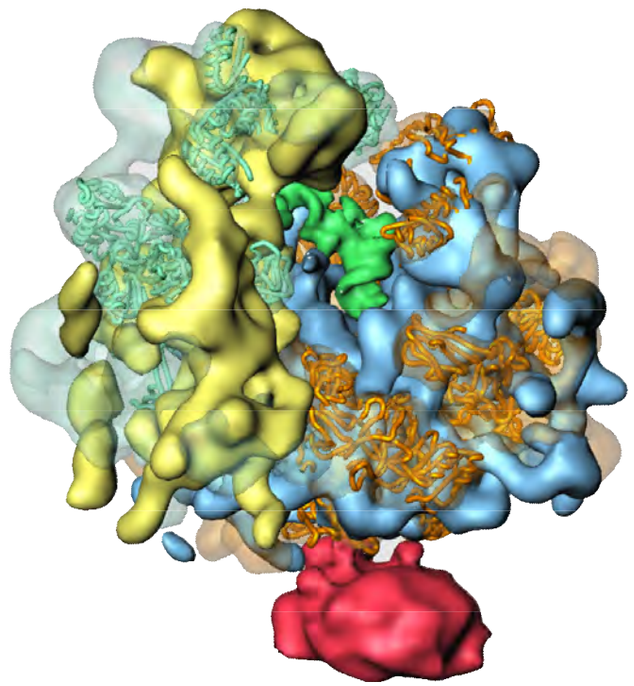
Stage 3: Finding Models That Score Well. The search for models that score well is performed using any of a variety of sampling and optimization schemes (such as the Monte Carlo method). There may be many models that score well if the data are incomplete or none if the data are inconsistent due to errors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

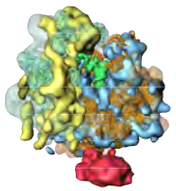
Integrative modeling iterates through these stages until a satisfactory model is built. Many iterations of the cycle may be required, given the need to gather more data as well as to resolve errors and inconsistent data.

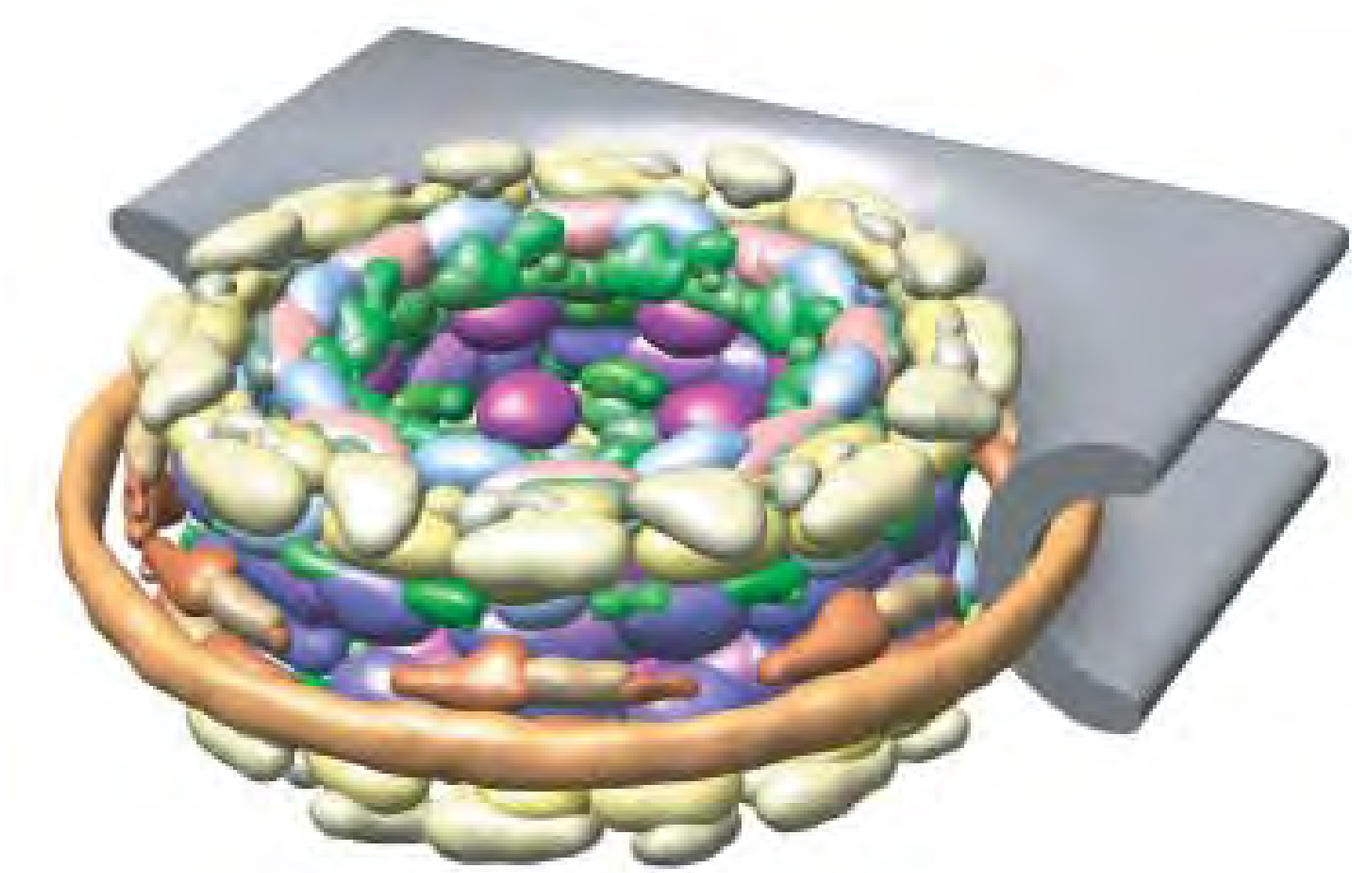
Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

Data Integration

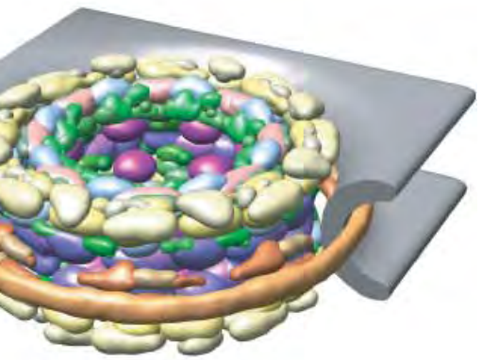


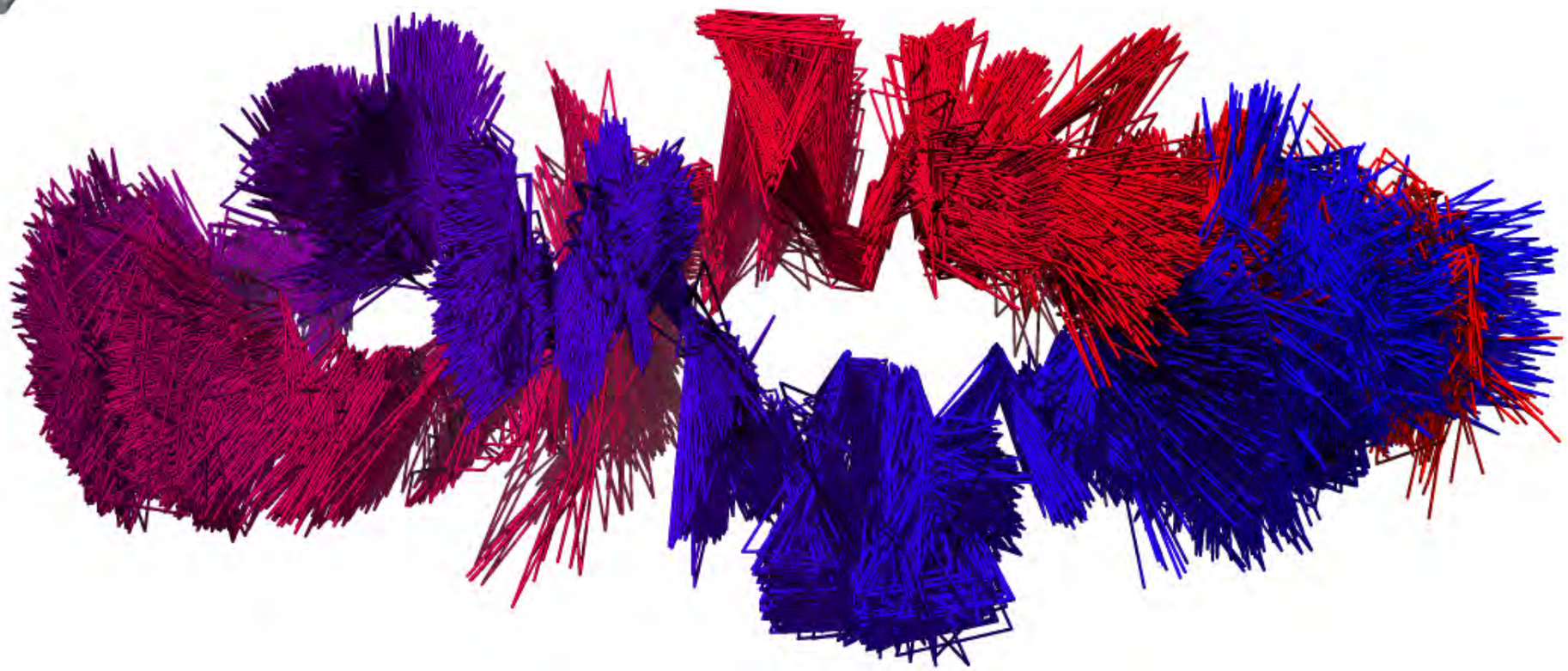
Data Integration



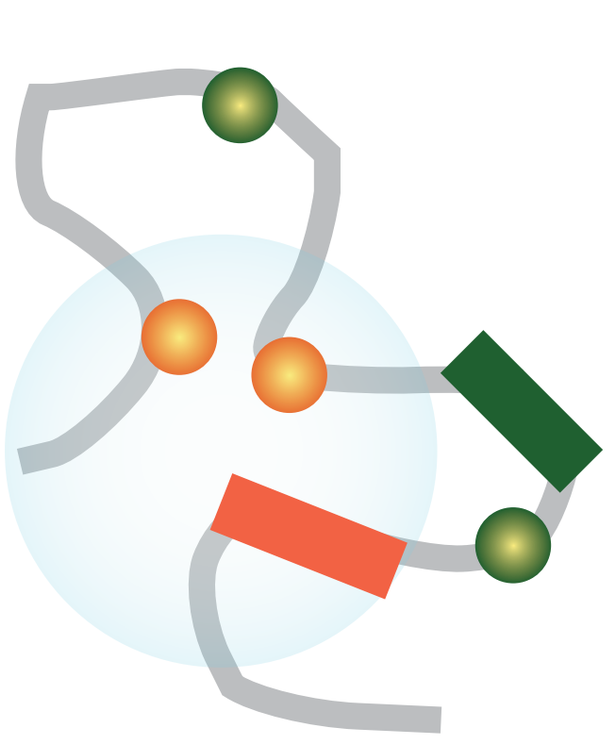


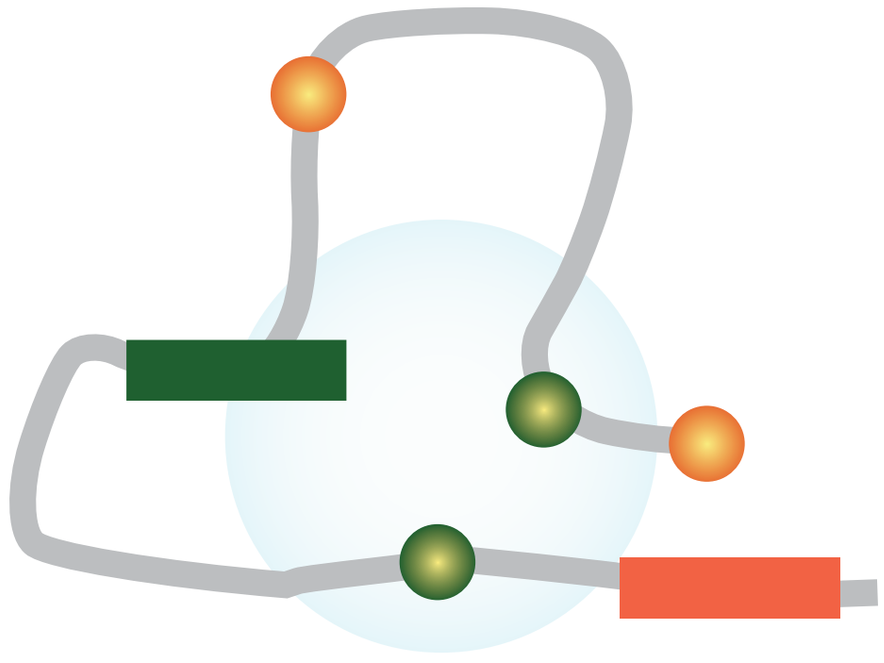
Data Integration





Complex genome organization

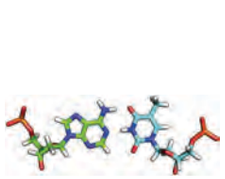


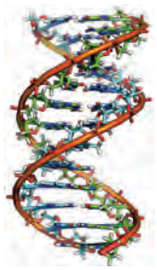


Resolution Gap

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

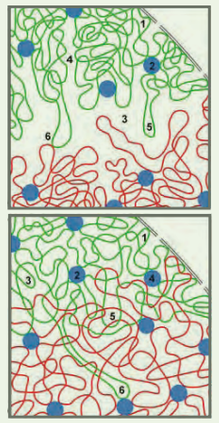
Knowledge

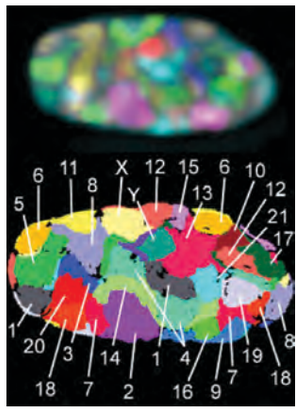




IDM

INM





DNA length

10^0 10^3 10^6 10^9 nt

Volume

10^{-9} 10^{-6} 10^{-3} 10^0 10^3 μm^3

Time

10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^0 10^2 10^3 s

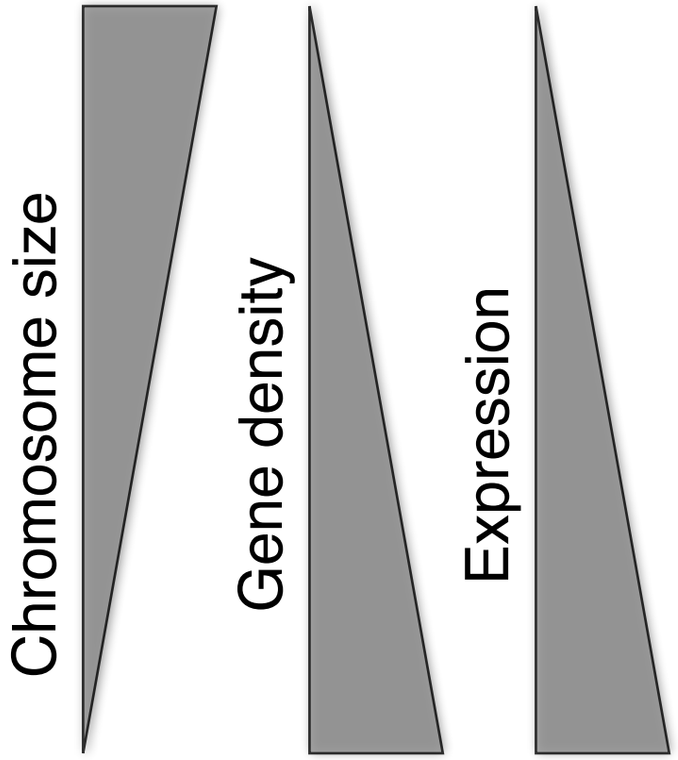
Resolution

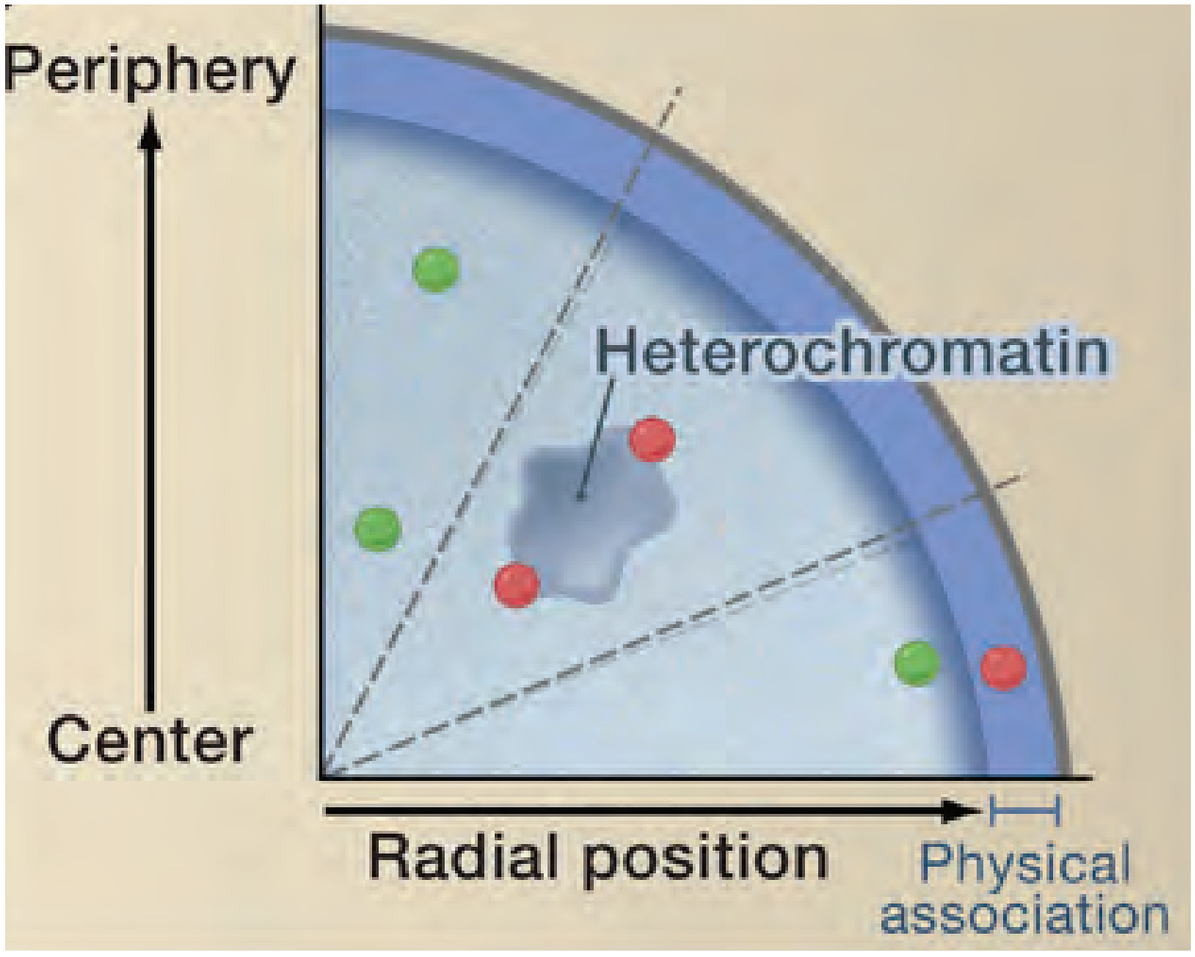
10^{-3} 10^{-2} 10^{-1} μ

cnag CRG Centre for Genomic Regulation

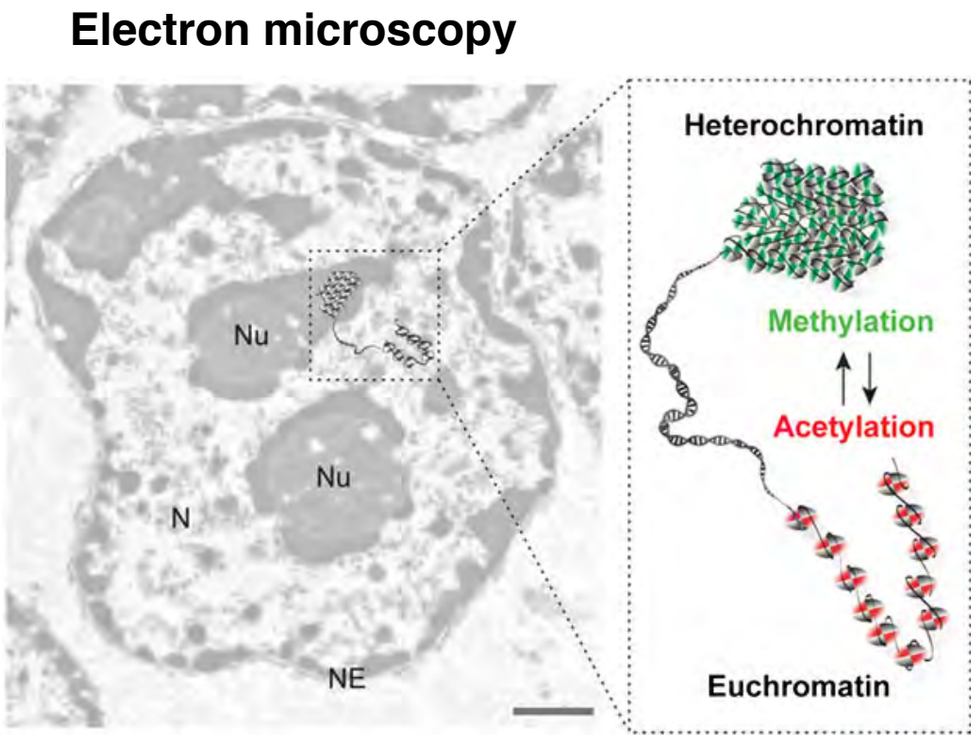
Level I: Radial genome organization

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





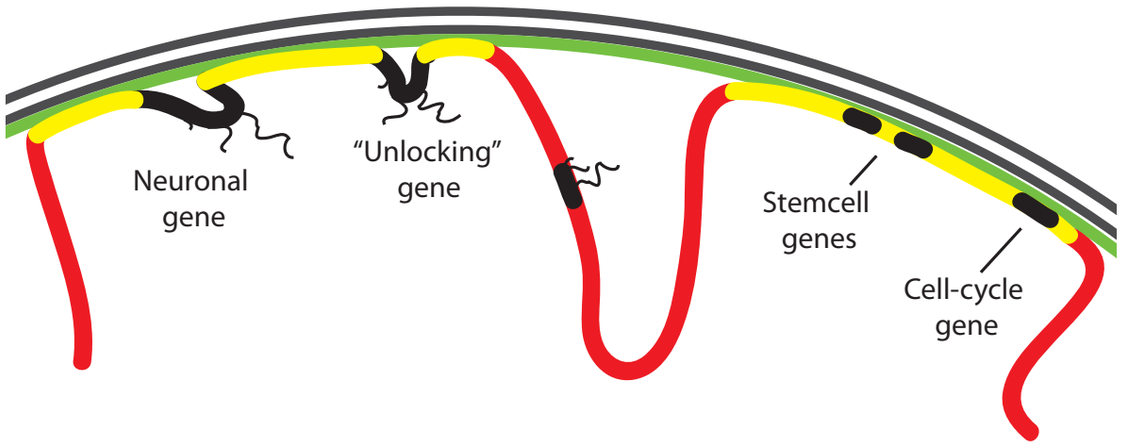
Level II: Euchromatin vs heterochromatin



Euchromatin:
chromatin that is located away from the nuclear lamina, is generally less densely packed, and contains actively transcribed genes

Heterochromatin:
chromatin that is near the nuclear lamina, tightly condensed, and transcriptionally silent

Level III: Lamina-genome interactions

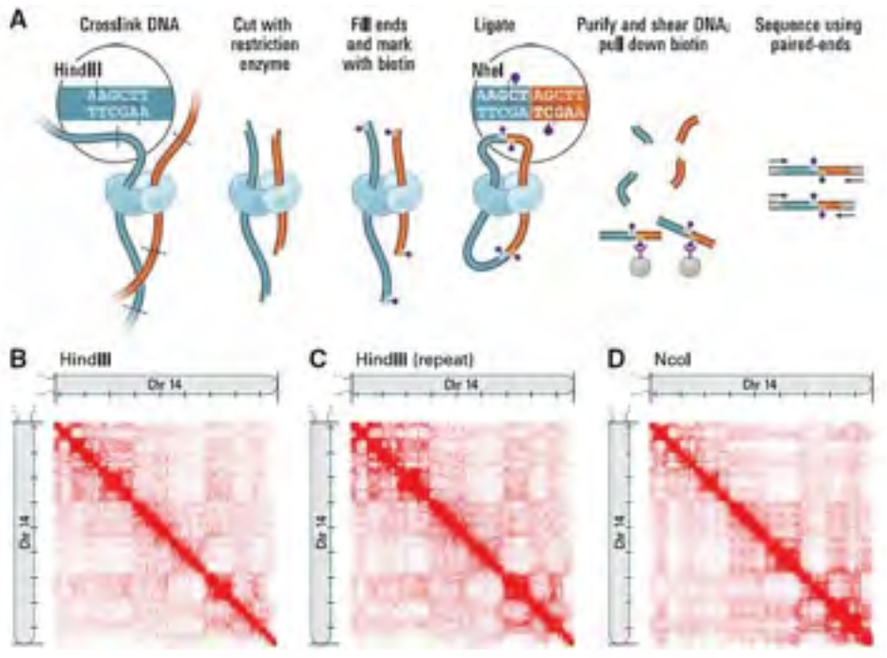


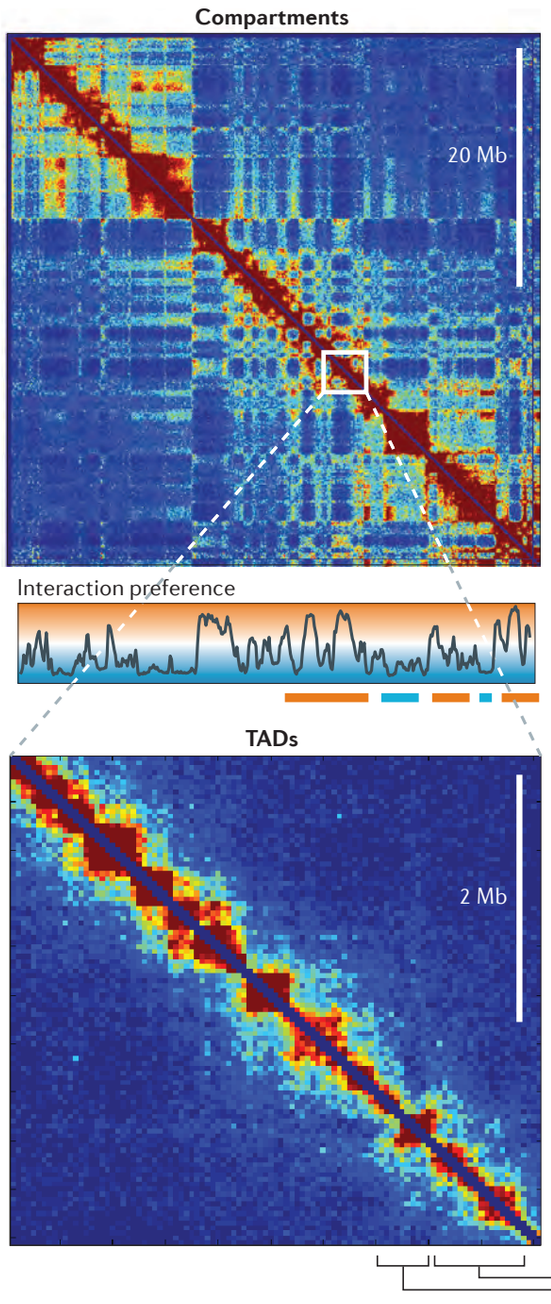
Most genes in Lamina Associated Domains are transcriptionally silent, suggesting that **lamina-genome interactions** are widely involved in the control of **gene expression**

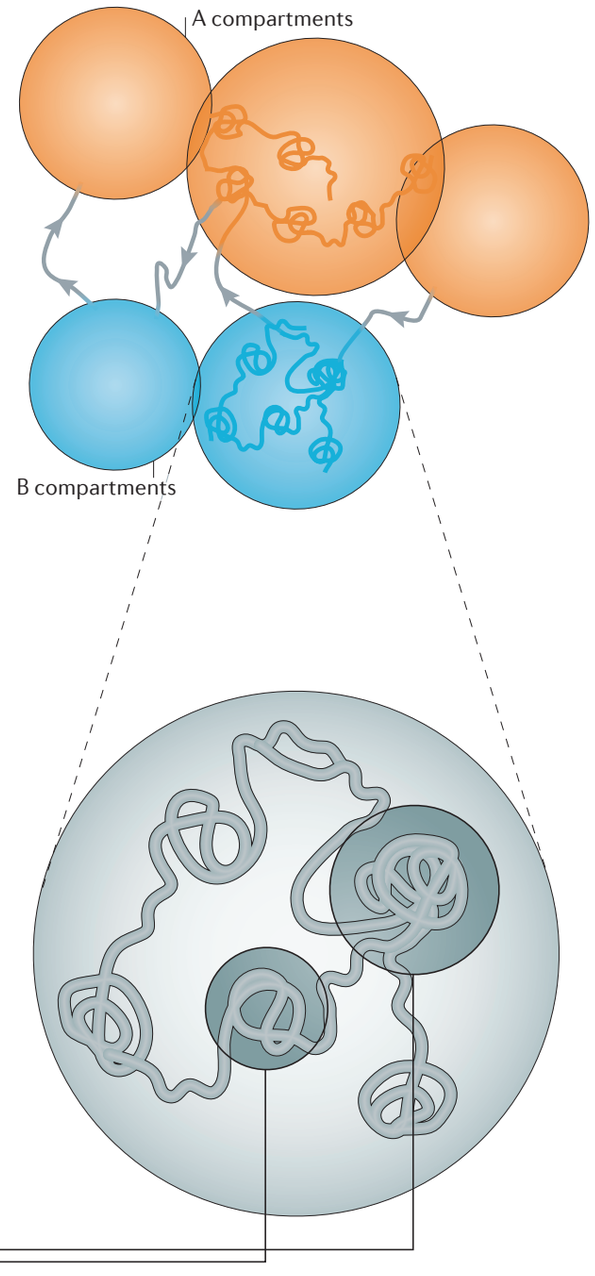
Adapted from Molecular Cell 38, 603-613, 2010

Level IV: Higher-order organization

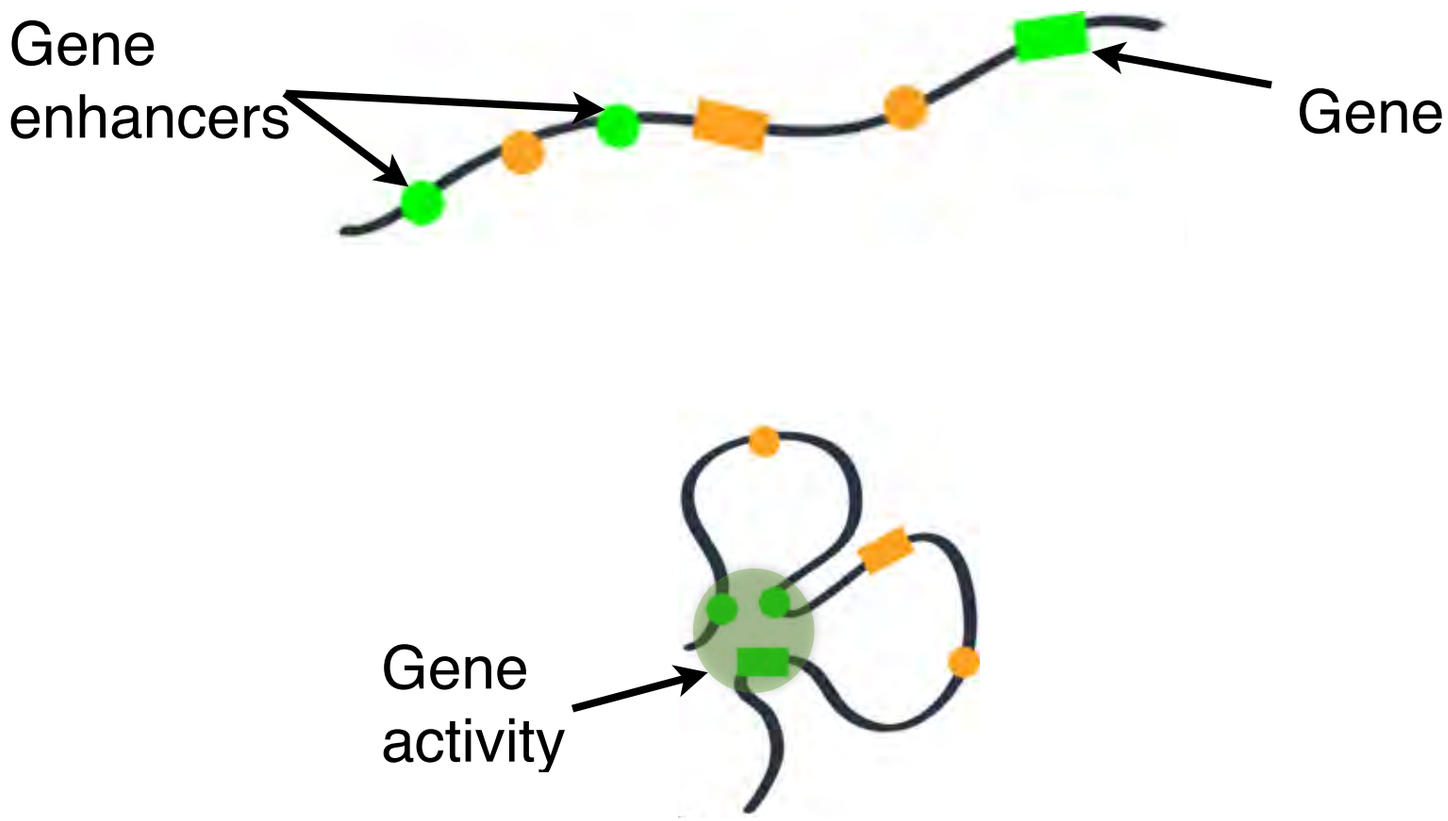
Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390–403 (2013).







Level V: Chromatin loops



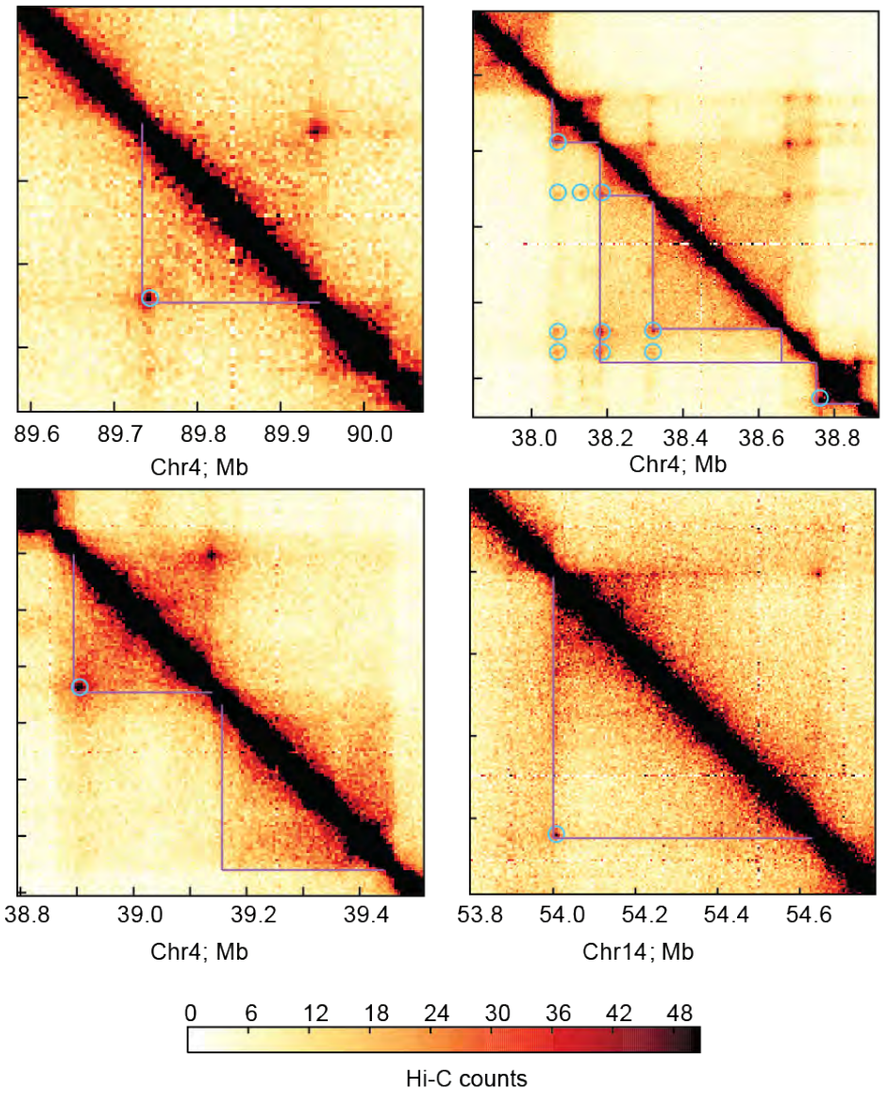
Loops bring distal genomic regions in close proximity to one another

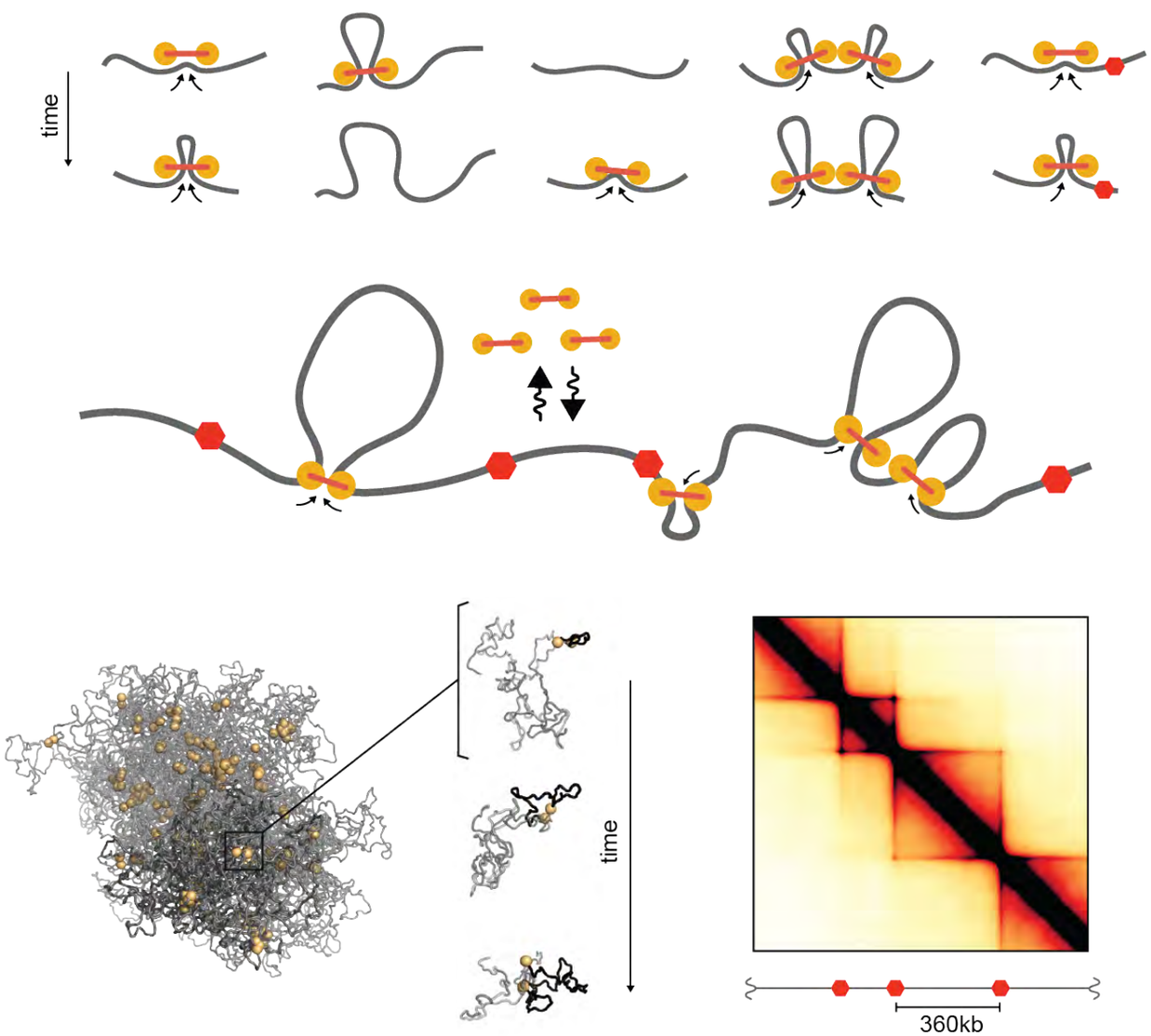
This in turn can have profound effects on gene transcription

Enhancers can be thousands of kilobases away from their target genes in any direction (or even on a separate chromosome)

Level V: Loop-extrusion as a driving force

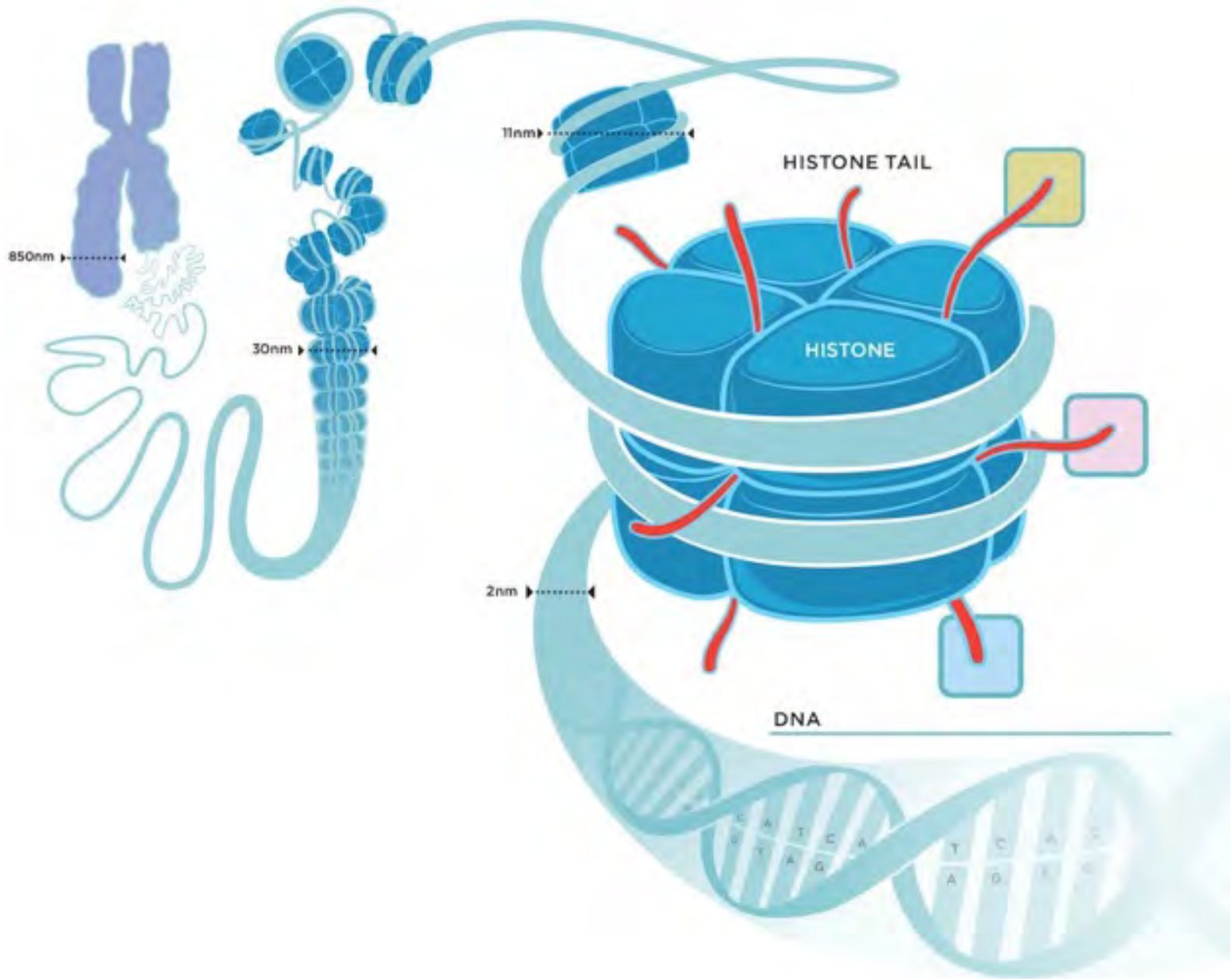
Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2015).
Formation of Chromosomal Domains by Loop Extrusion. bioRxiv.





Level VI: Nucleosome

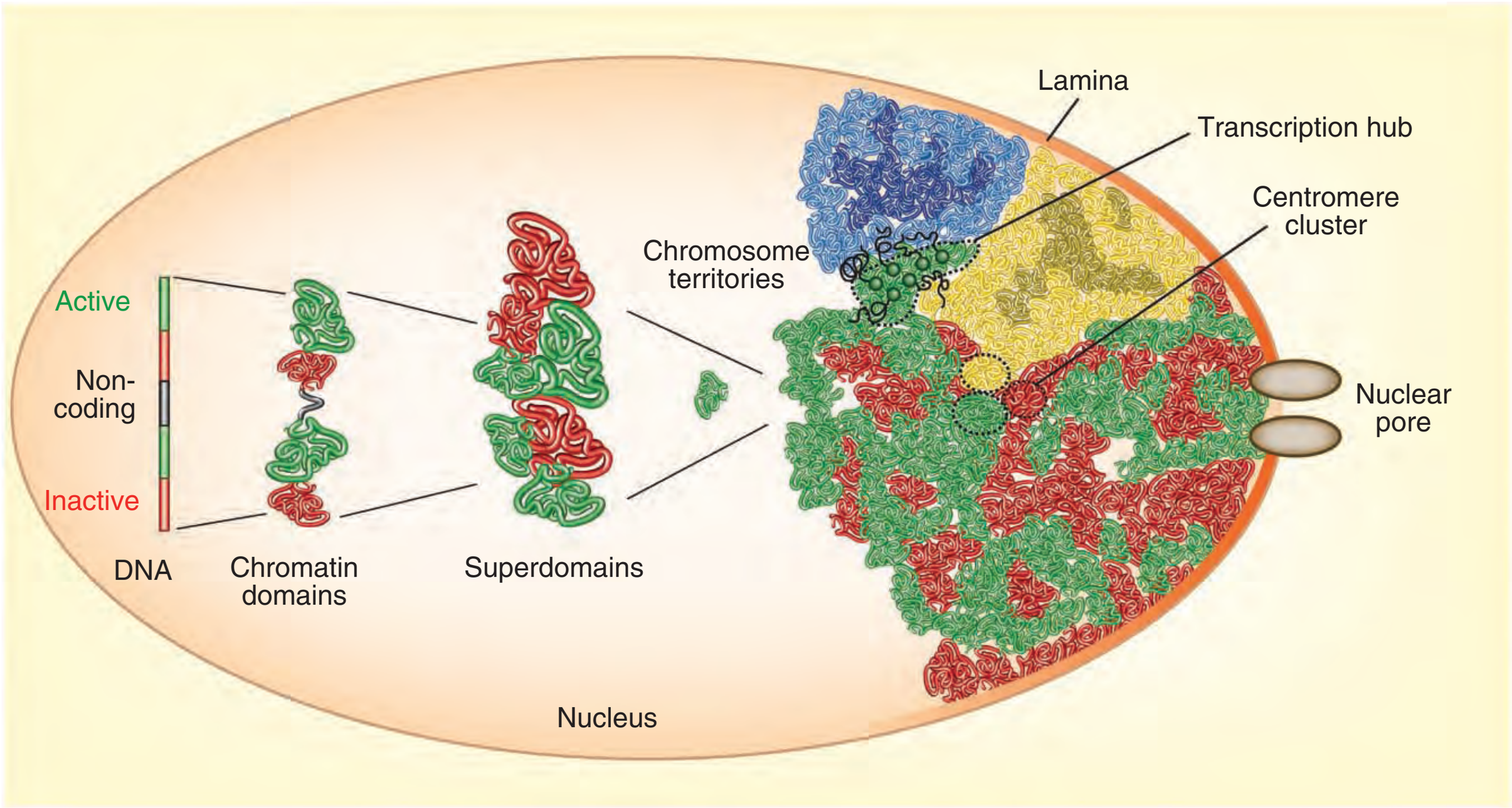
Chromosome **Chromatin fibre** **Nucleosome**



Adapted from Richard E. Ballermann, 2012

Complex genome organization

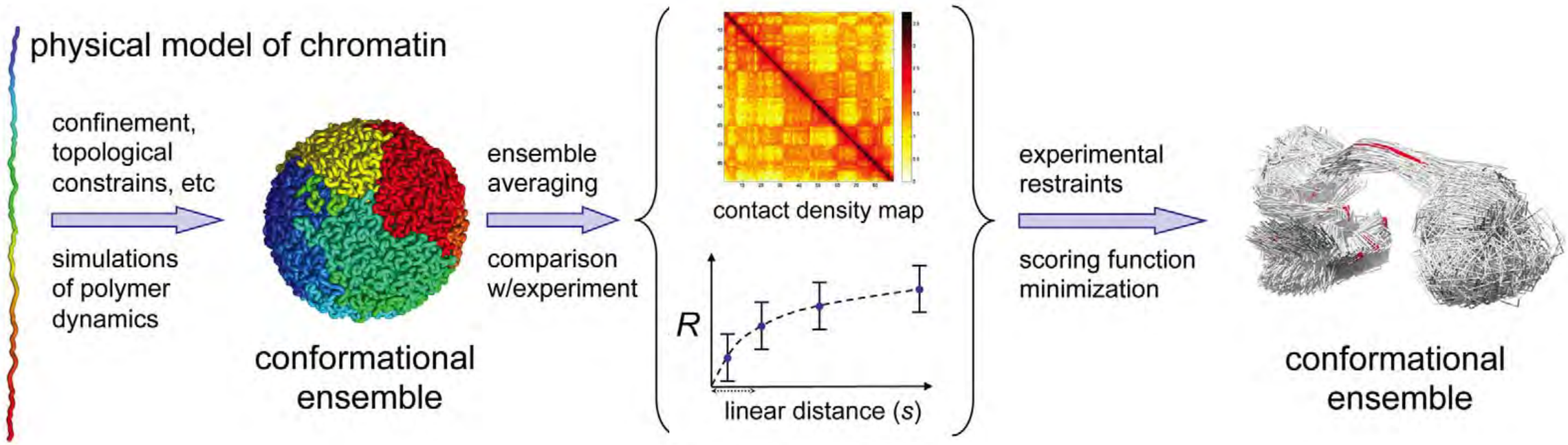
Cavalli, G. & Misteli, T. Functional implications of genome topology. Nat Struct Mol Biol 20, 290–299 (2013).



Marina Corral

Modeling Genomes

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



Experiments

Grow GM12878 and K562 cells

↓

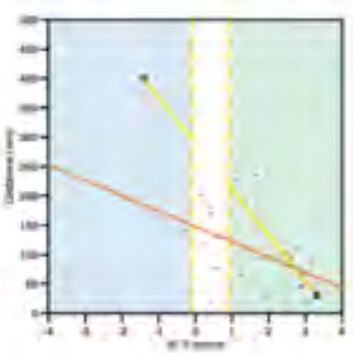
Perform 3C analysis

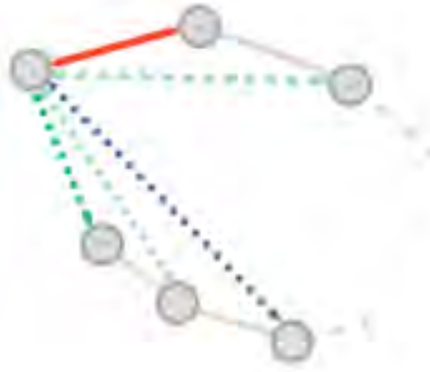
↓

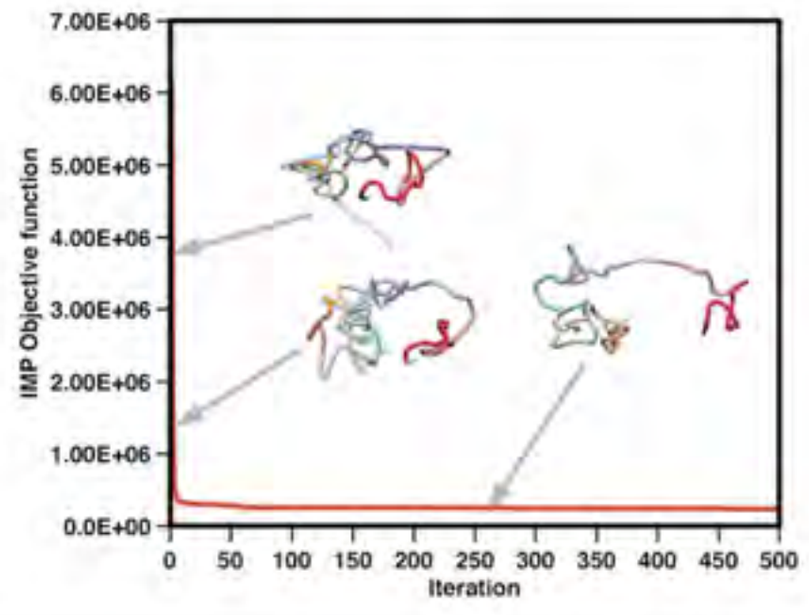
Perform 5C analysis with 30+25 primers

↓

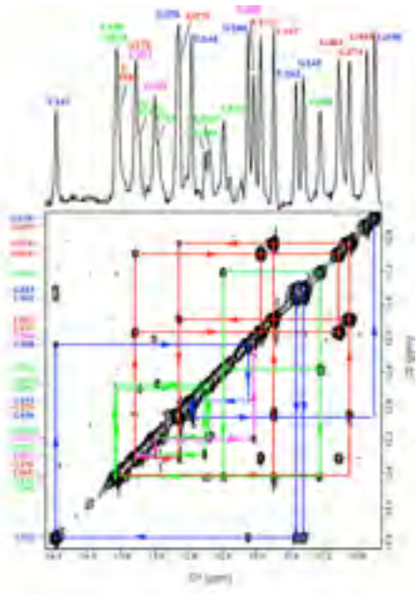
Analyze 5C products by paired-end Solexa sequencing
(131,947 paired end reads per library)





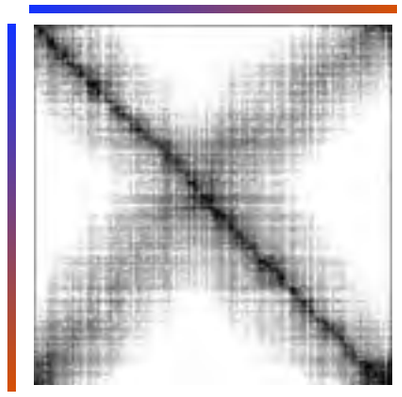


Computation



Biomolecular structure determination

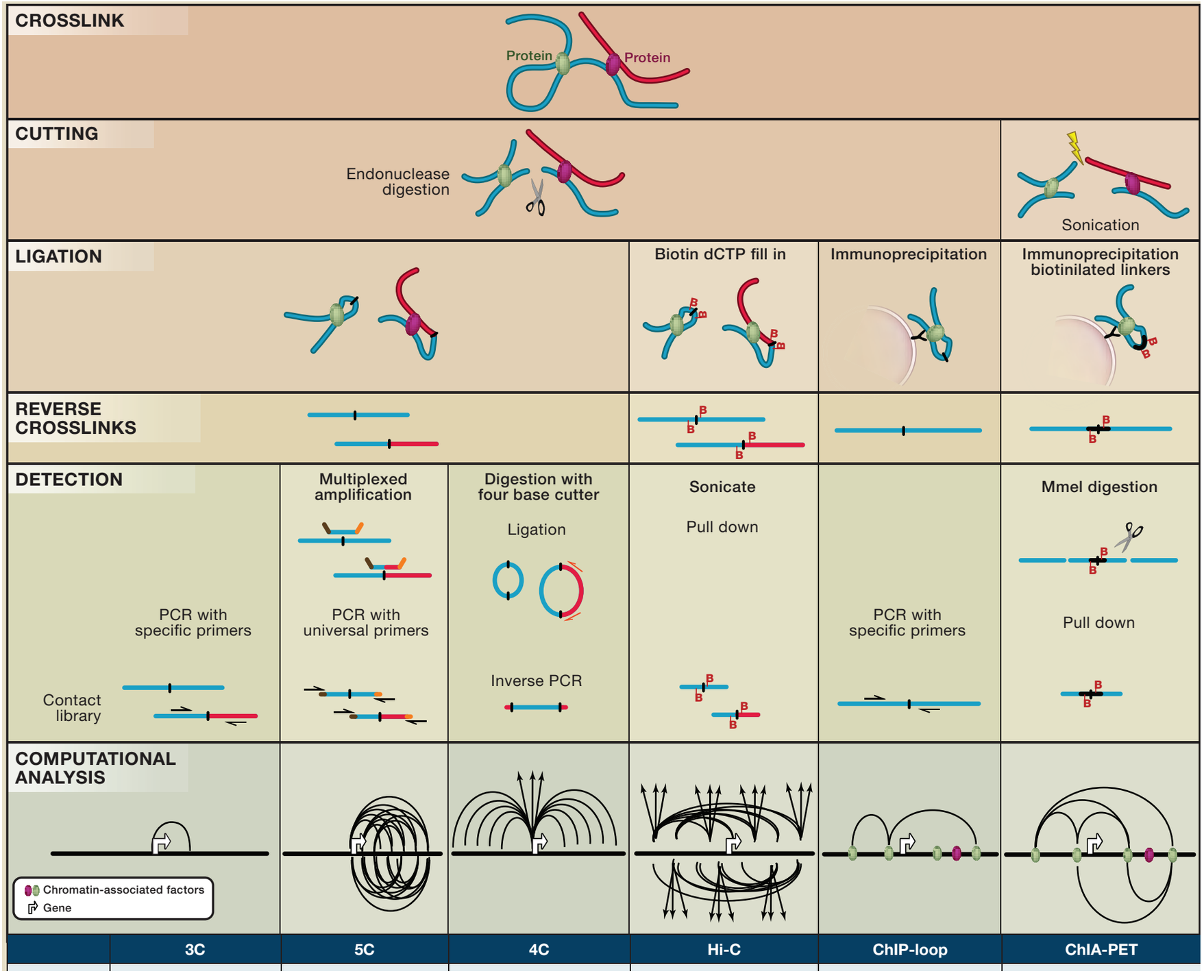
2D-NOESY data



Chromosome structure determination

5C data

Chromosome Conformation Capture



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.

Capture-C

Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment

Jim R Hughes[1], Nigel Roberts[1], Simon McGowan[2], Deborah Hay[1], Eleni Giannoulatou[2], Magnus Lynch[1], Marco De Gobbi[1], Stephen Taylor[2], Richard Gibbons[1] & Douglas R Higgs[1]

Gene expression during development and differentiation is regulated in a cell- and stage-specific manner by complex networks of intergenic and intragenic *cis*-regulatory elements whose numbers and representation in the genome far exceed those of structural genes. Using chromosome conformation capture, it is now possible to analyze in detail the interaction between enhancers, silencers, boundary elements and promoters at individual loci, but these techniques are not readily scalable. Here we present a high-throughput approach (Capture-C) to analyze *cis* interactions, interrogating hundreds of specific interactions at high resolution in a single experiment. We show how this approach will facilitate detailed, genome-wide analysis to elucidate the general principles by which *cis*-acting sequences control gene expression. In addition, we show how Capture-C will expedite identification of the target genes and functional effects of SNPs that are associated with complex diseases, which most frequently lie in intergenic *cis*-acting regulatory elements.

It is now possible to rapidly map the positions of many *cis*-regulatory sequences (promoters, enhancers, silencers and boundary elements) across the genome by analyzing chromatin structure, histone modifications and the binding of transcription factors and cofactors[1–4]. Detailed studies of a relatively small number of individual genes have revealed unexpected levels of complexity in their interactions with *cis*-regulatory sequences. Expression of a single gene is often controlled by multiple regulatory elements that may lie tens to hundreds of kilobases upstream or downstream of their targets. In addition, the regulatory elements controlling one gene may lie within the introns of another, unrelated gene. Detailed studies of individual loci have revealed some mechanistic insights into how *cis*-acting elements regulate gene expression, but because only a few loci have been studied in detail, general principles have not yet emerged. Understanding the mechanisms underlying long-range regulation of gene expression is of critical importance in molecular medicine, as it was recently shown that most SNPs that are associated with complex diseases lie within distal *cis*-regulatory elements and presumably alter the timing or levels of expression of their target genes in specific cell types. Therefore, a major challenge in mammalian genetics is to develop high-throughput techniques to link specific *cis*-regulatory elements to their cognate genes and determine how these interactions and their associated variants influence gene expression during development and differentiation.

It has been shown that when *cis*-acting sequences influence gene expression, they may physically interact with the promoter(s) they regulate. The resulting physical contacts can be detected by various protocols, which are generally referred to as chromosome conformation capture or 3C. These techniques involve digestion and re-ligation of fixed chromatin followed by quantification of ligation junctions, which reflect the frequencies of interaction[5,6].

Several adaptations of the original 3C method (4C, 5C, HiC and ChIA-PET) have been developed to assay interactions across the genome but are currently unsuitable for linking *cis*-acting sequences with gene promoters both at high resolution and in a high-throughput manner[7–10]. To map enhancer-promoter interactions in detail using chromosome conformation capture requires a resolution of ~1–2 kb because most *cis*-acting sequences are in the range of hundreds of basepairs in length and may be closely clustered. Some versions of the 4C-seq method can map interactions at this resolution, but they only interrogate a single region of interest at a time[10]. However any given cell type contains thousands of active promoters and even greater numbers of potential *cis*-acting sequences[4]. Also, typically the number of significant disease-associated genome-wide association study (GWAS) variants affecting these elements are numbered in the thousands[11]. Hence, using 4C-seq to interrogate each active element and its associated variants individually is laborious and prohibitively expensive. Clearly, this represents a major bottleneck in our ability to investigate both the normal regulation of genes and the effects of sequence variants.

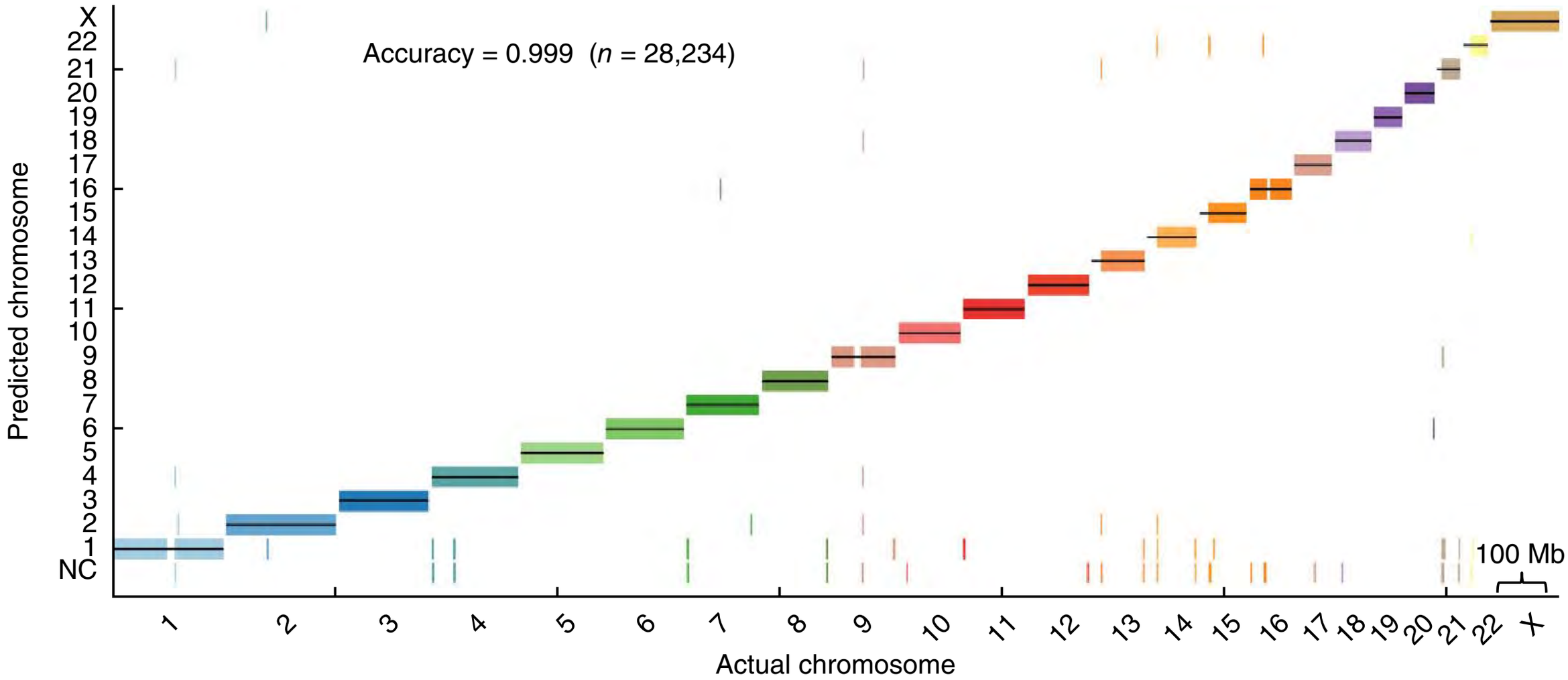
Here we present a new approach to this problem. Capture-C combines oligonucleotide capture technology (OCT), 3C and high-throughput sequencing and enables researchers to interrogate *cis* interactions at hundreds of selected loci at high resolution in a single assay. When combined with the corresponding epigenetic data that

[1]Medical Research Council (MRC) Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK. [2]Computational Biology Research Group, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, UK. Correspondence should be addressed to D.R.H. (doug.higgs@imm.ox.ac.uk) or J.R.H. (jim.hughes@imm.ox.ac.uk).

Received 24 June 2013; accepted 12 December 2013; published online 12 January 2014; doi:10.1038/ng.2871

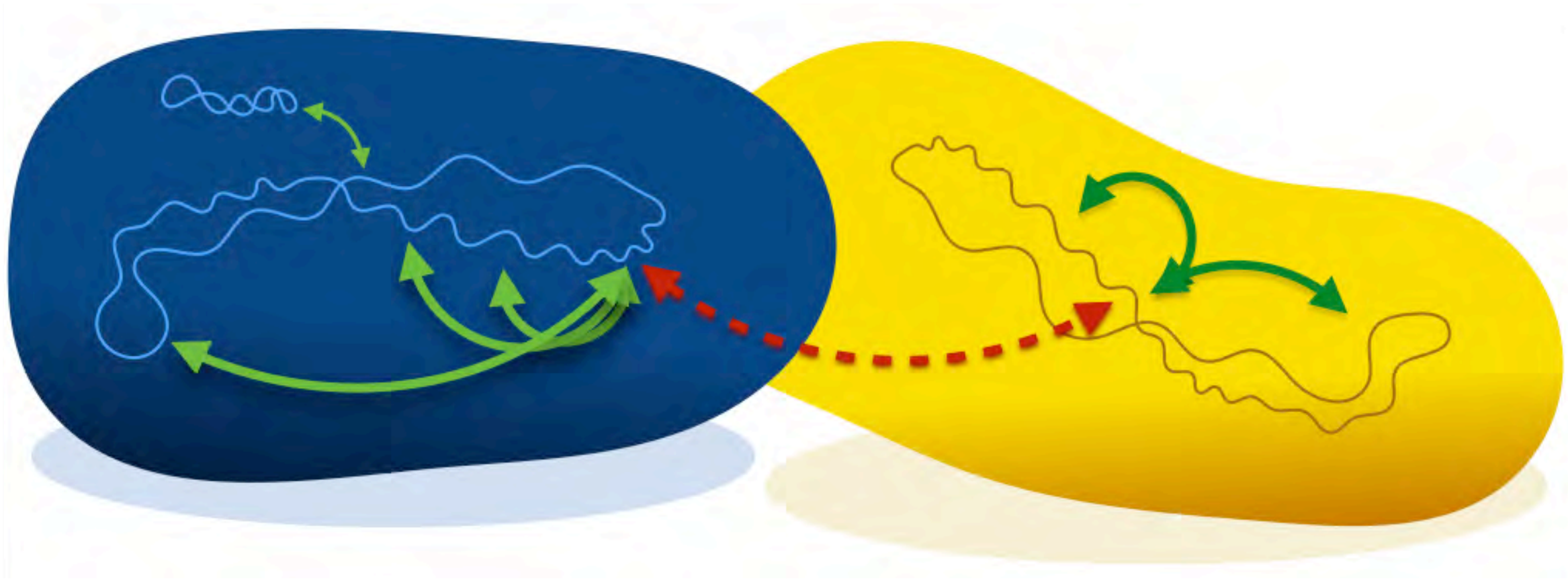
Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., et al. (2014). Analysis of hundreds of. Nature Genetics, 1–10.

Chromosome Conformation Capture for de-novo assembly



Kaplan, N., & Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature Biotechnology, 31(12), 1143–1147.

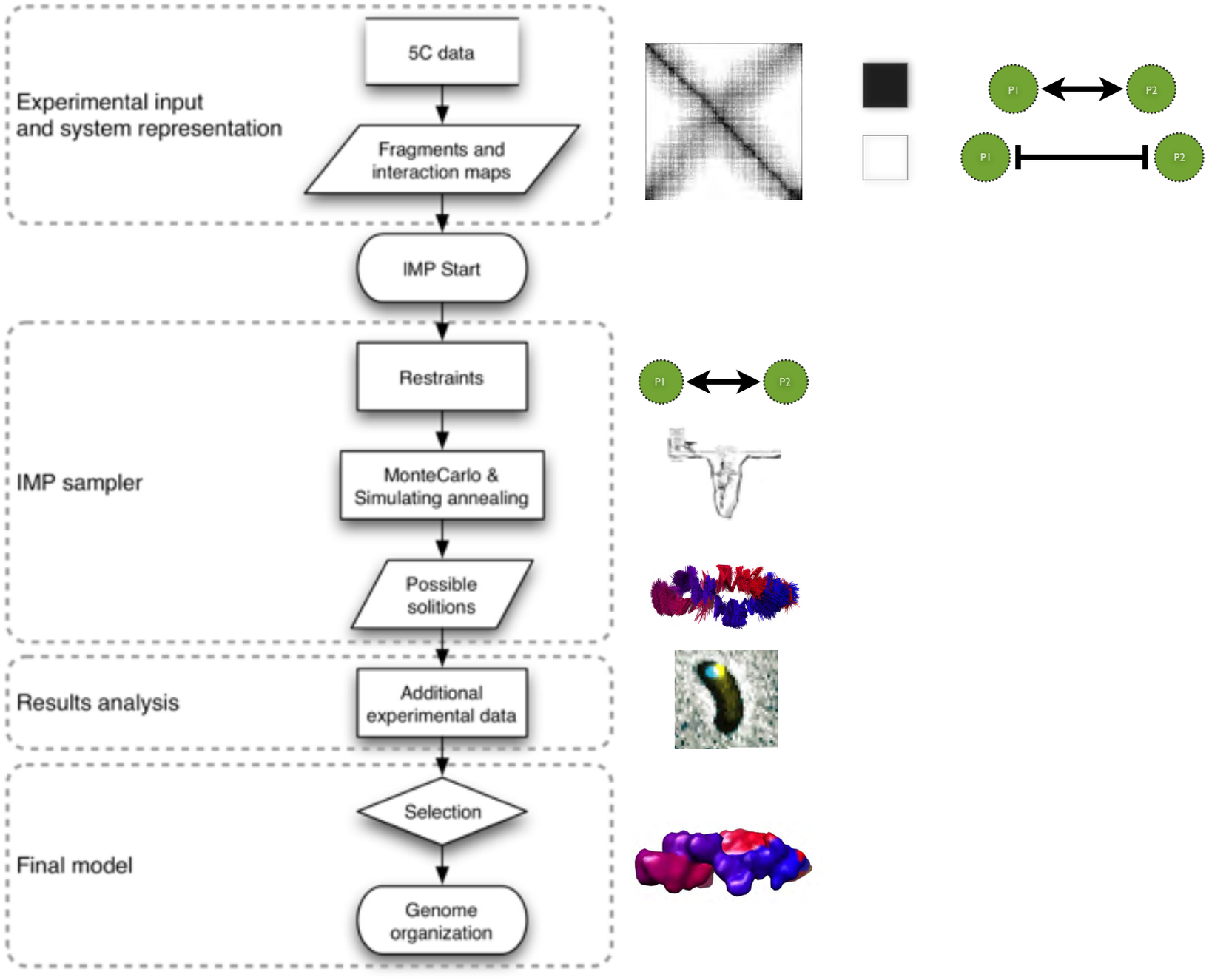
Chromosome Conformation Capture for meta genomics



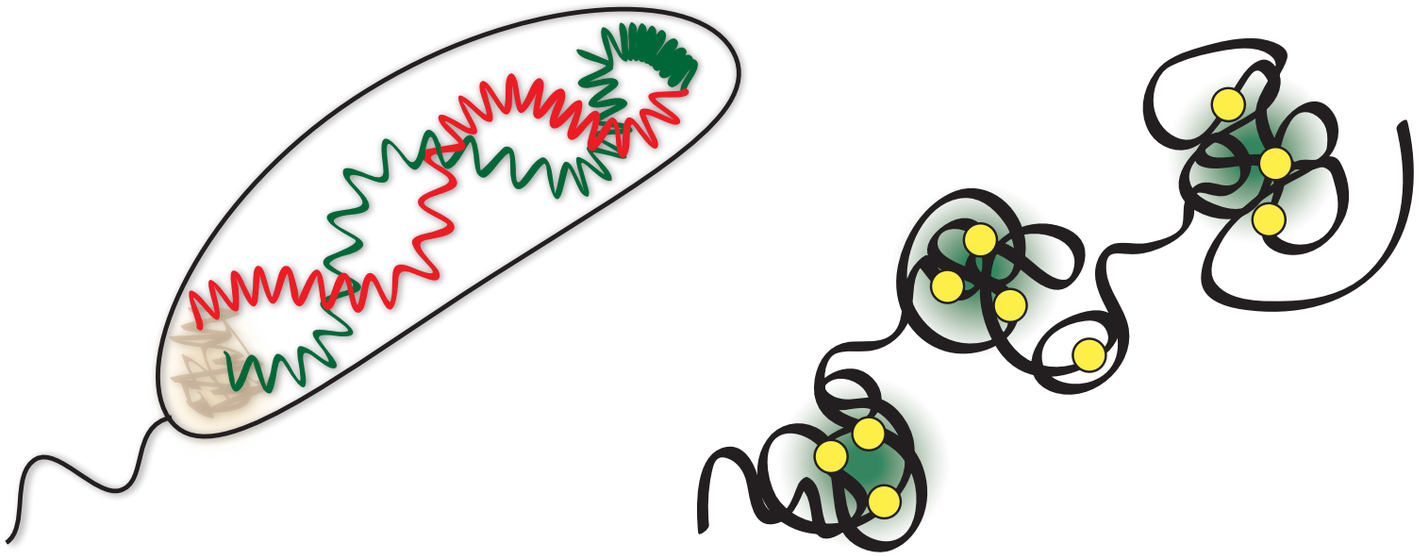
Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., & Darling, A. E. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. doi:10.7287/peerj.preprints.260v1

Modeling 3D Genomes

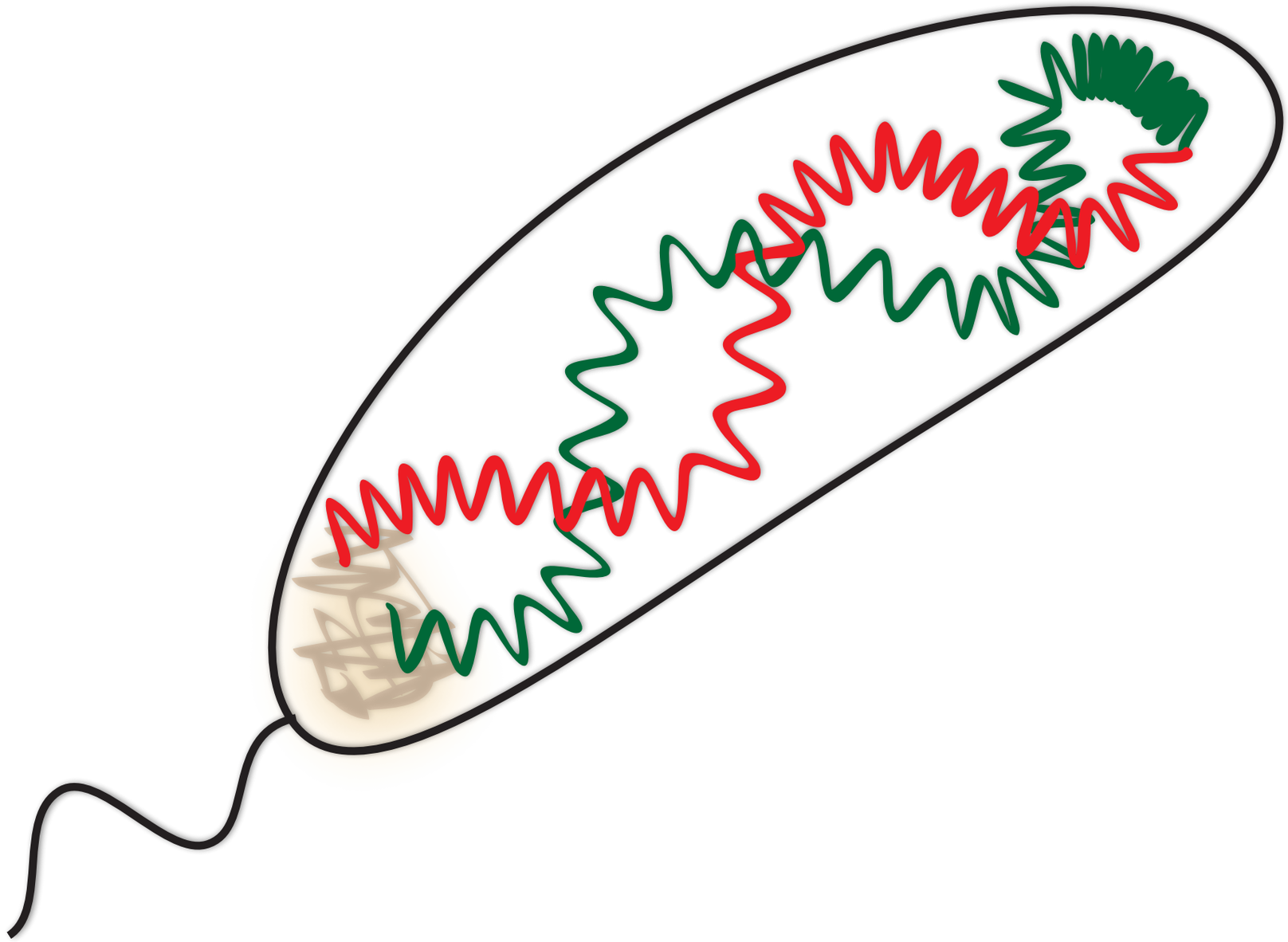
Baù, D. & Marti-Renom, M. A. Methods 58, 300–306 (2012).



Examples...

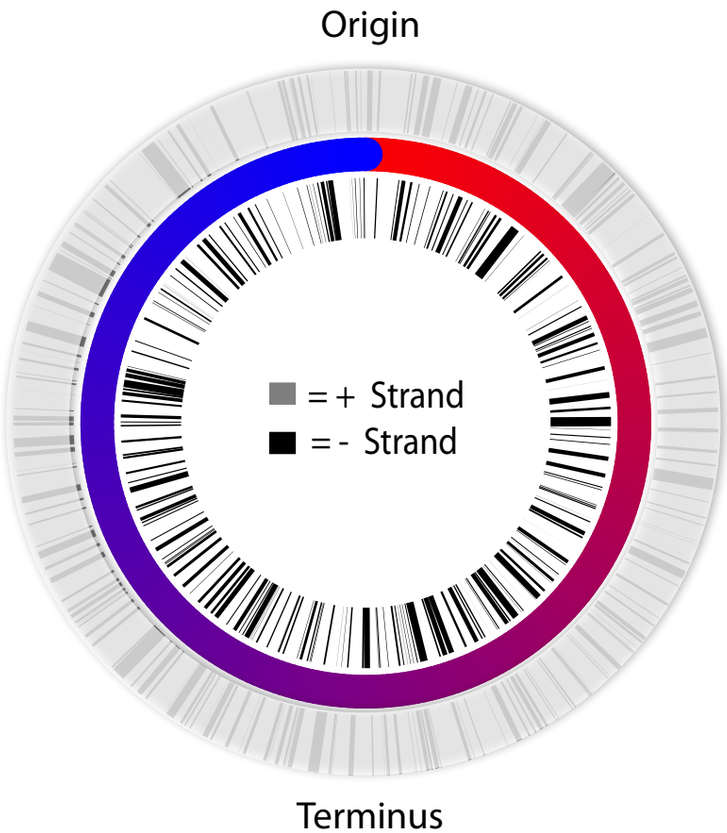


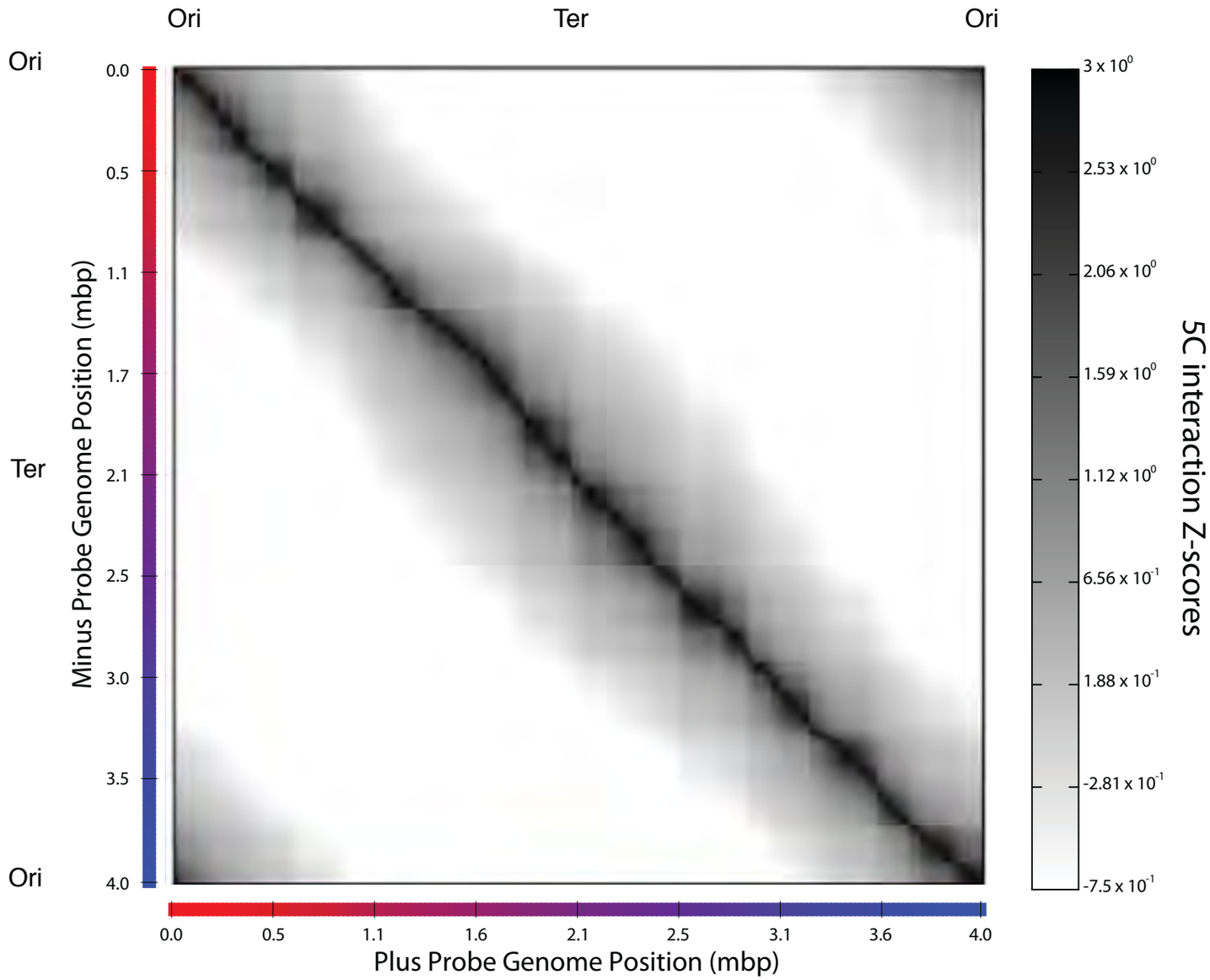
Caulobacter crescentus genome



The 3D architecture of Caulobacter Crescentus

4,016,942 bp & 3,767 genes





169 5C primers on + strand

170 5C primers on – strand

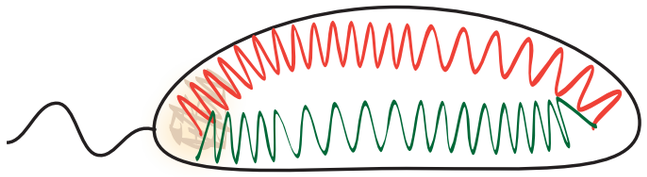
28,730 chromatin interactions

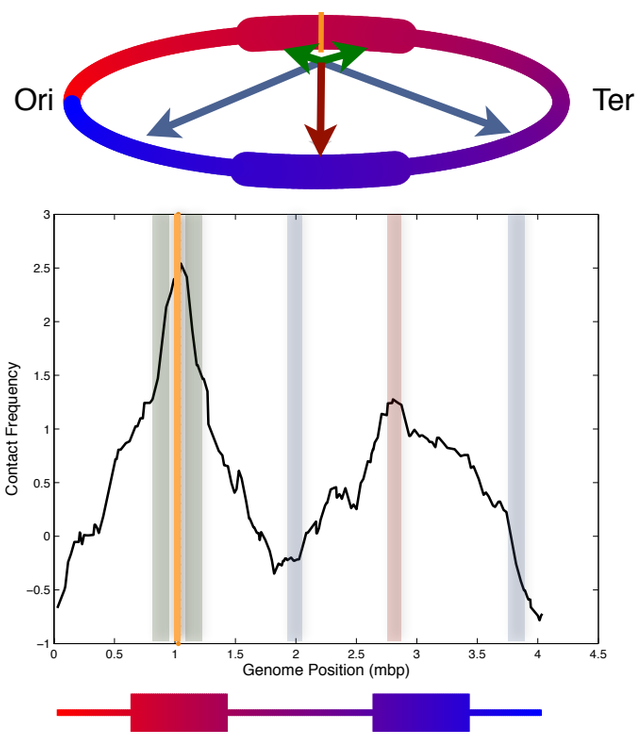
~13Kb

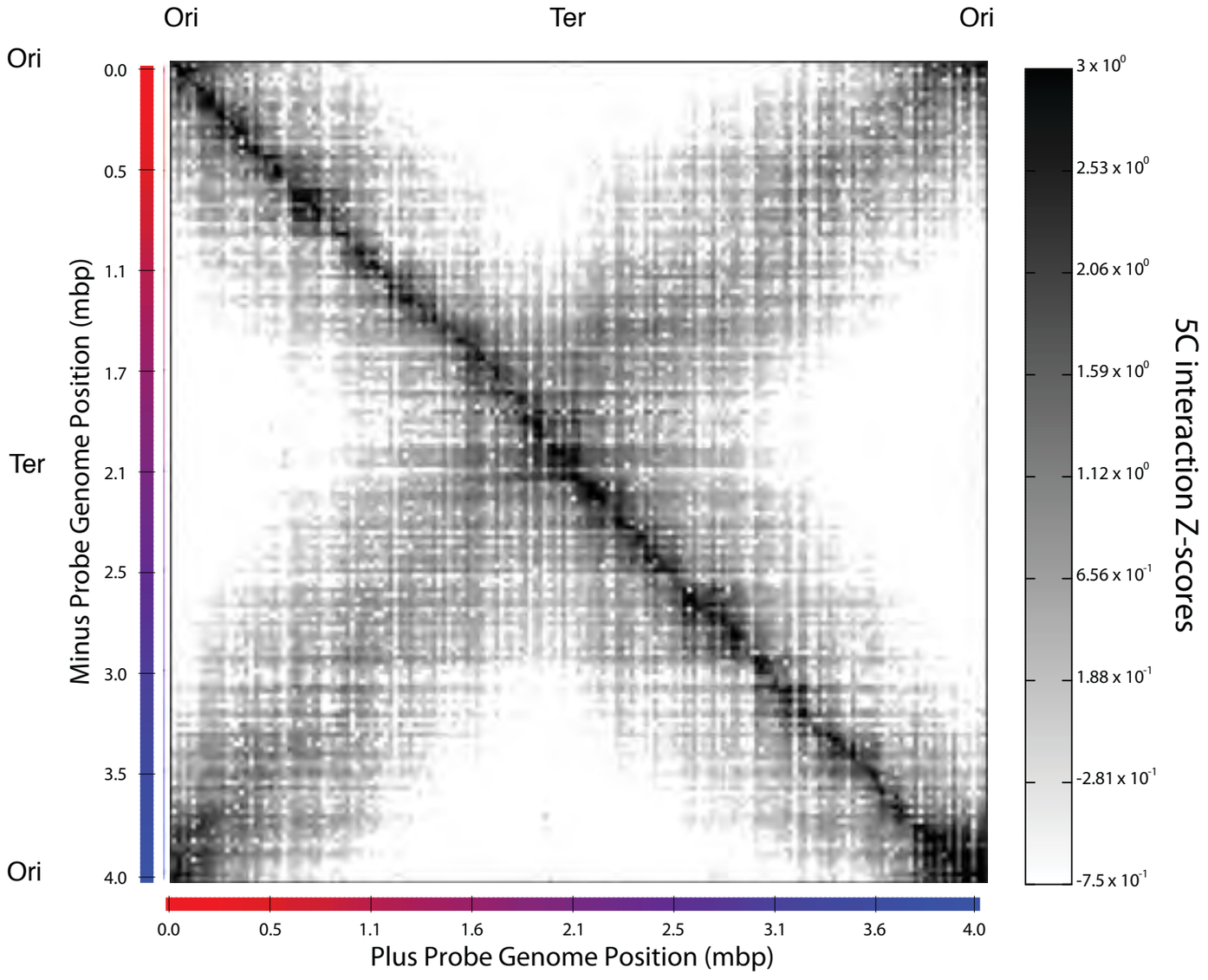
cnag CRG Centre for Genomic Regulation

5C interaction matrix

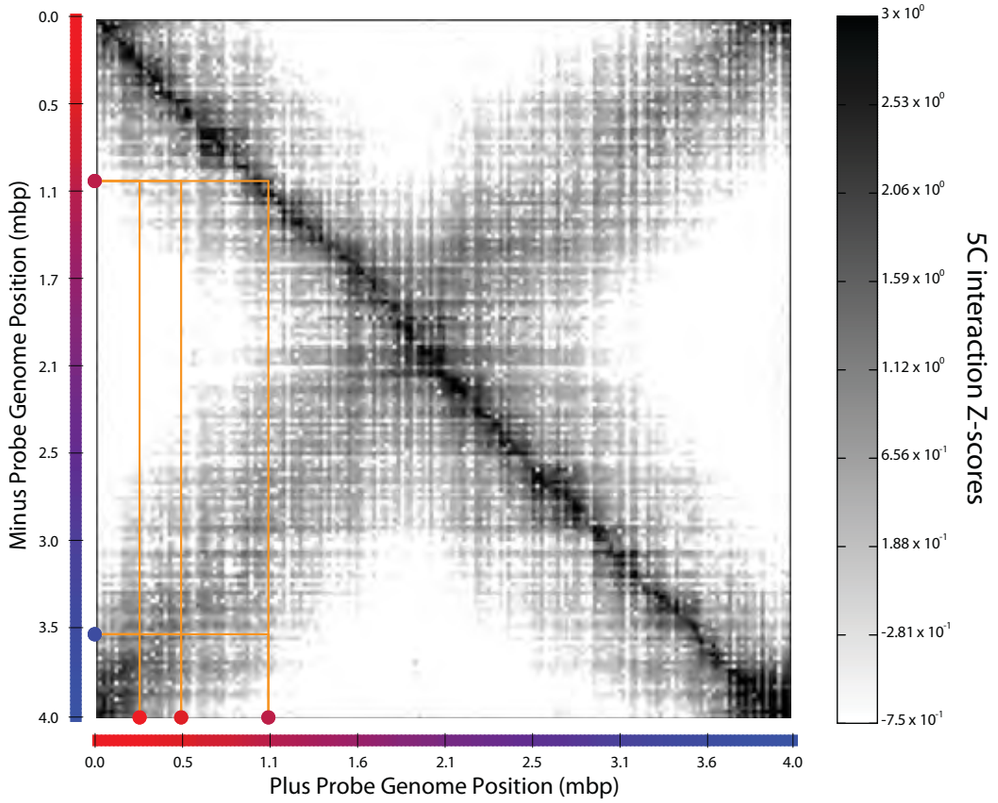
ELLIPSOID for Caulobacter cresentus

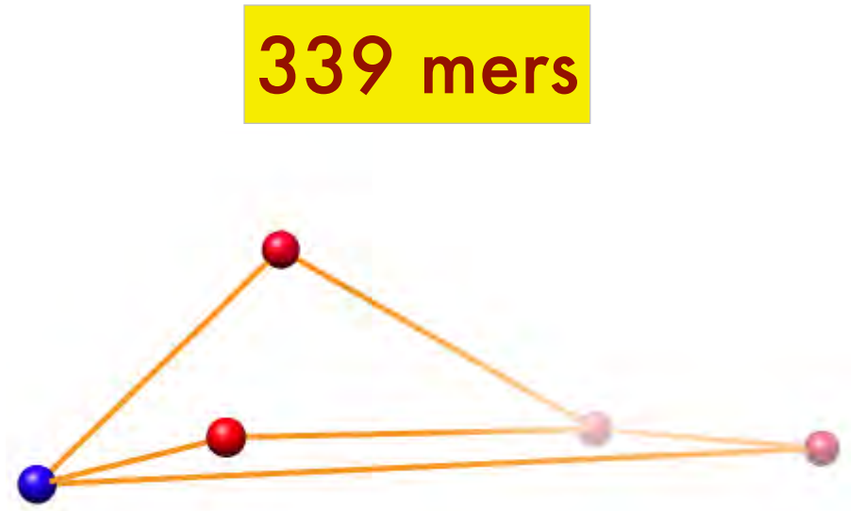


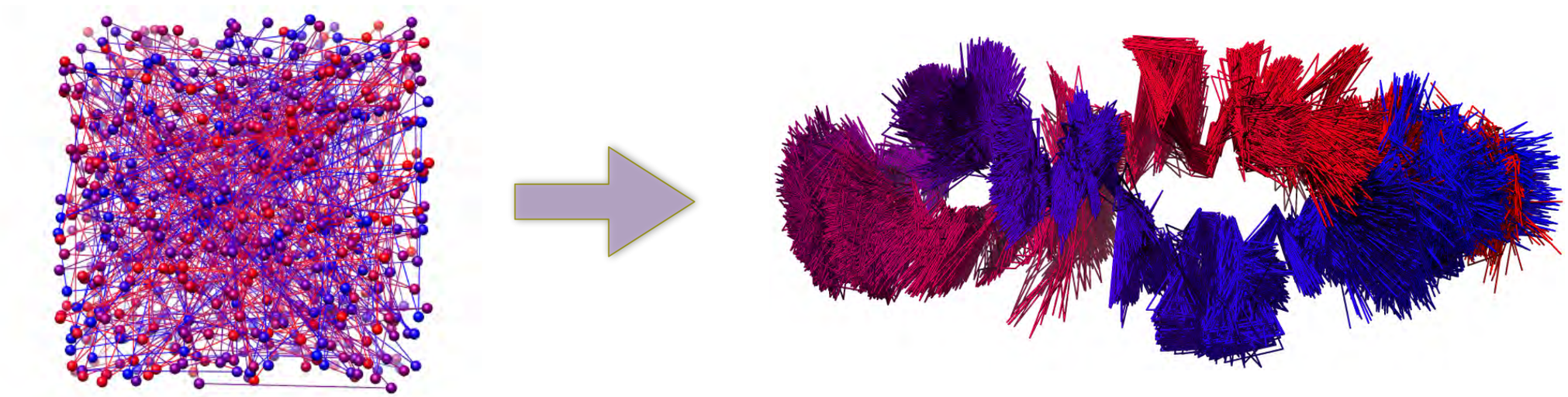




3D model building with the 5C + IMP approach





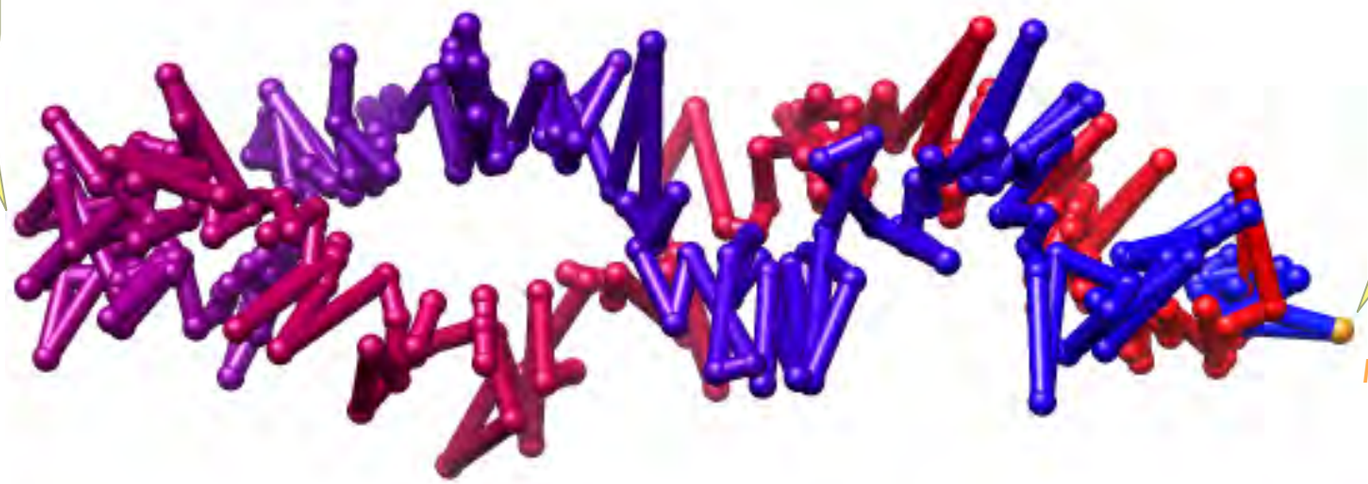


Genome organization in Caulobacter crescentus

Arms are helical

Resolution

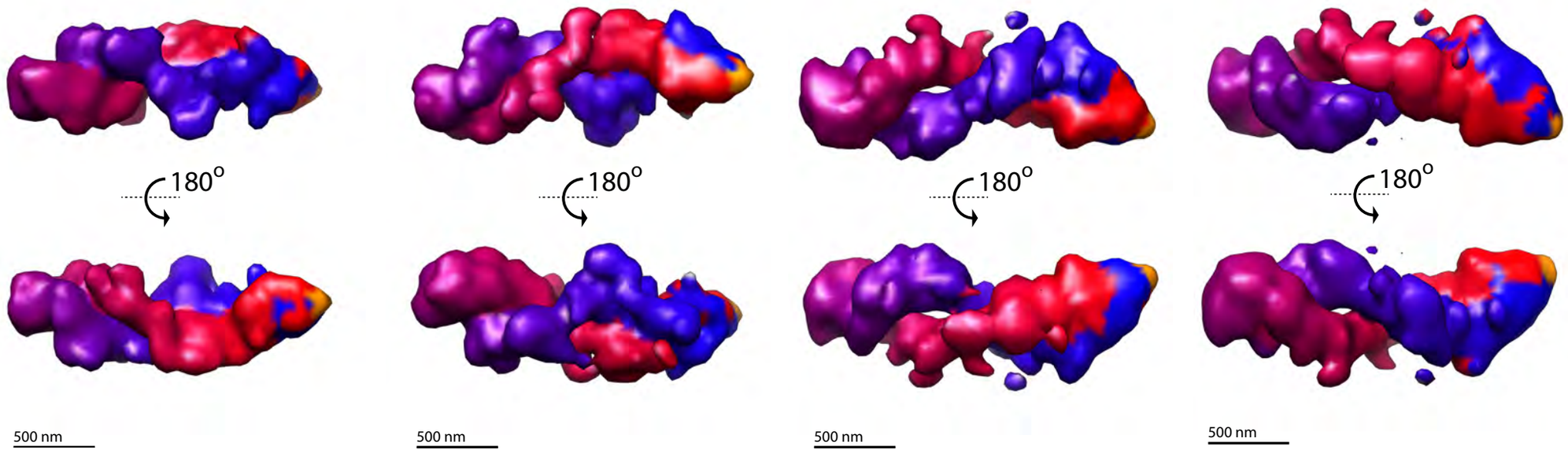
dif site 47±17Kb from Ter



Centromer-like

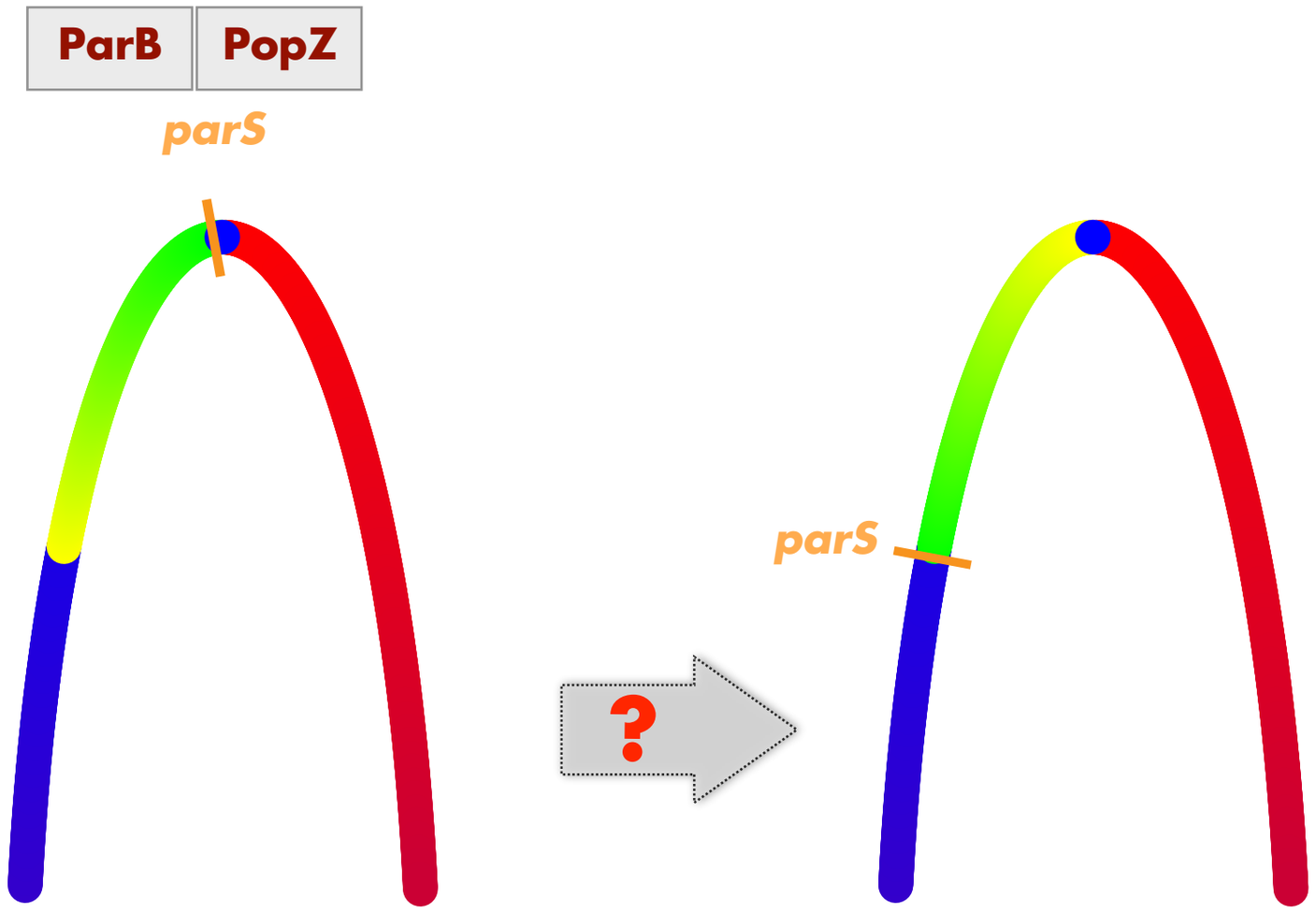
parS sites 25±17Kb from Ori

Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4



MIRRORS!

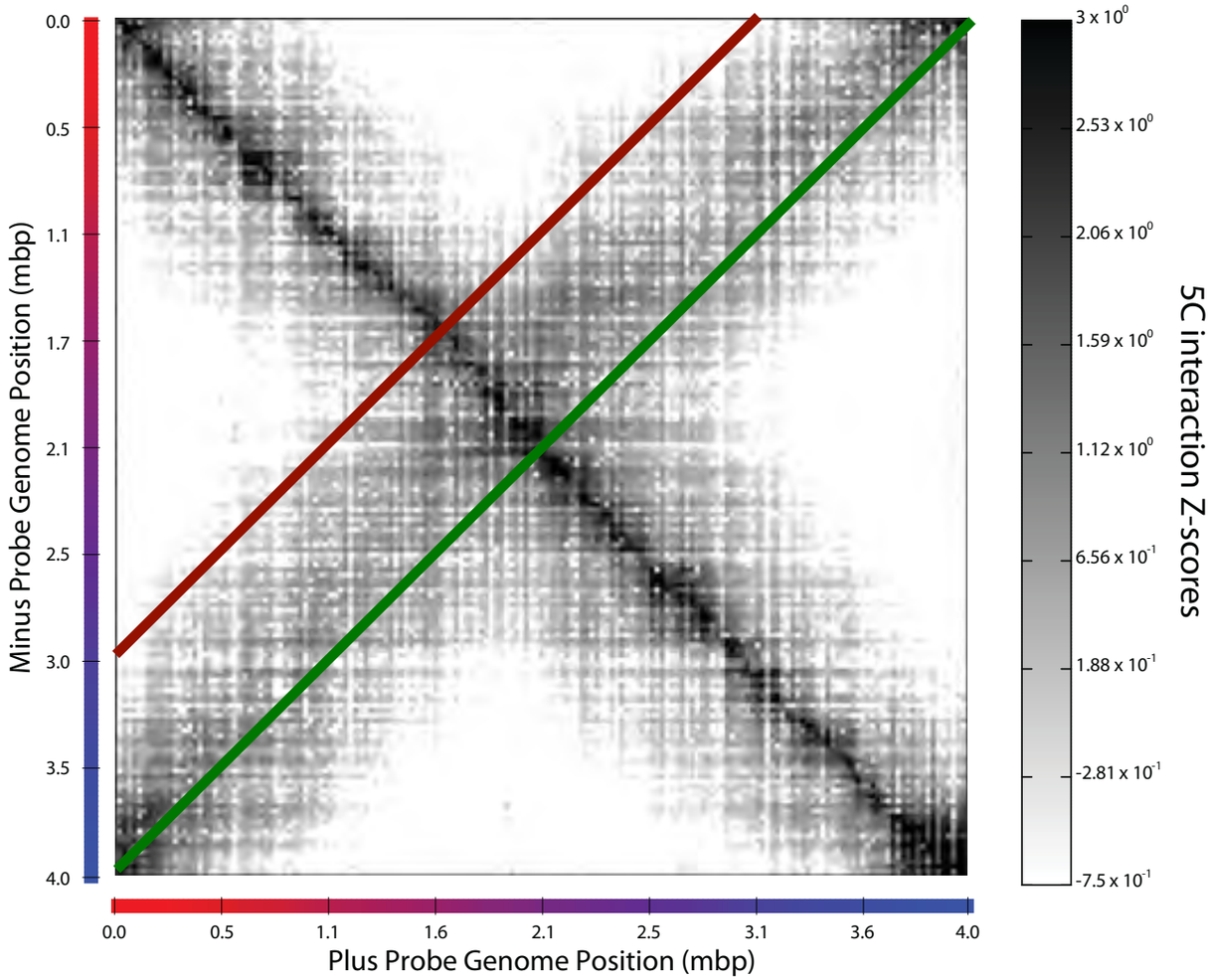
Moving the parS sites 400 Kb away from Ori

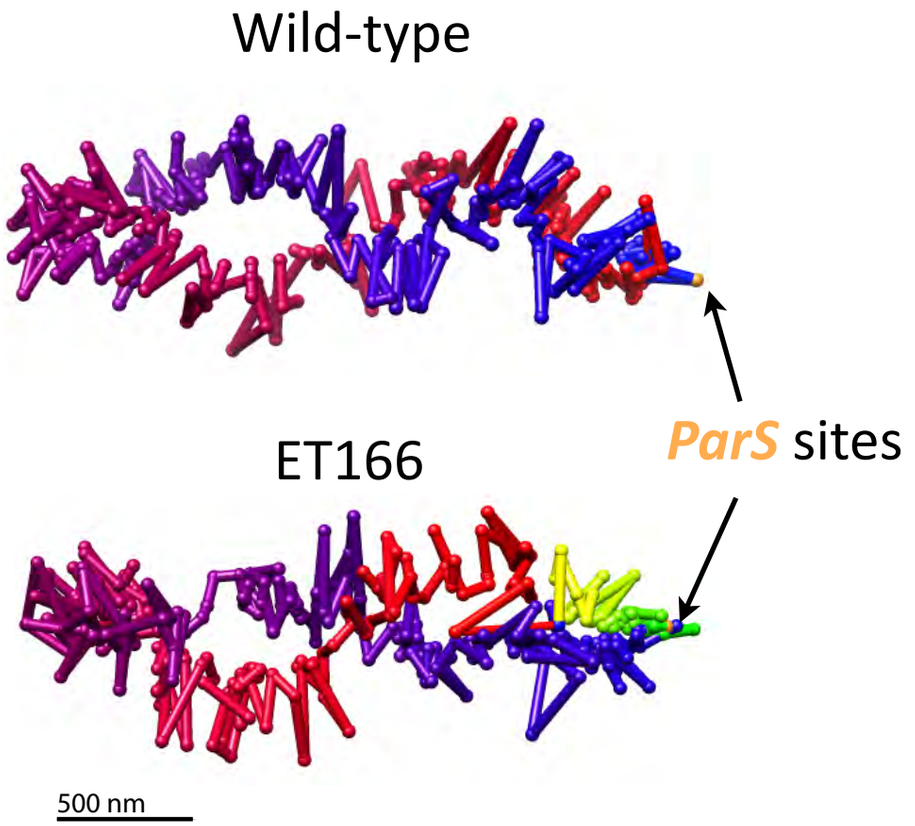


Wild-type

ET166

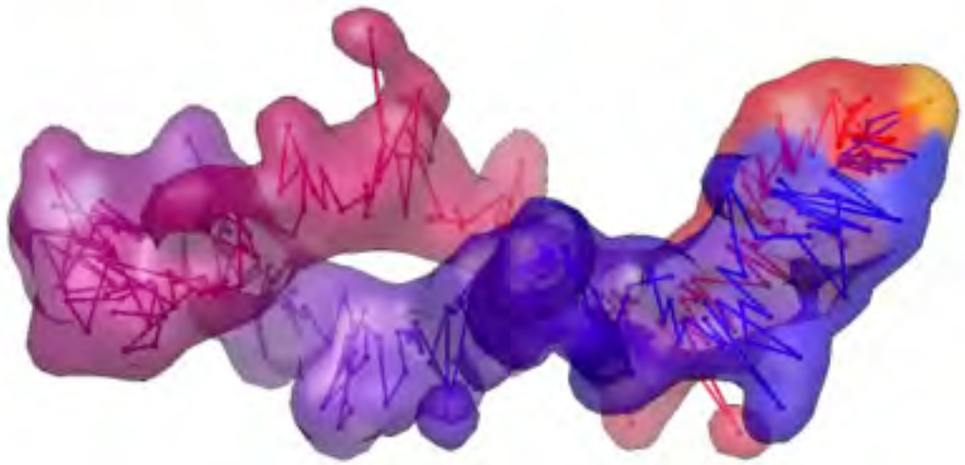
Moving the parS sites results in whole genome rotation!

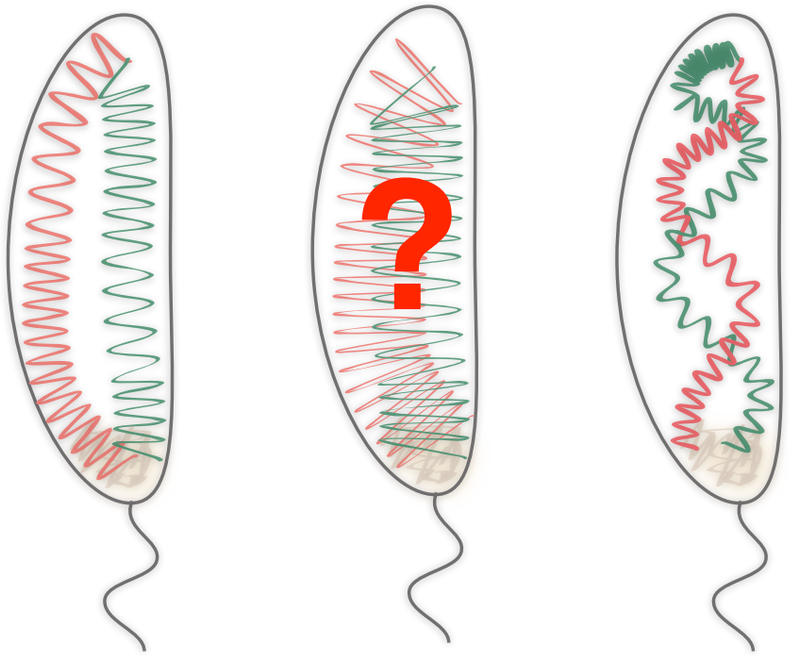


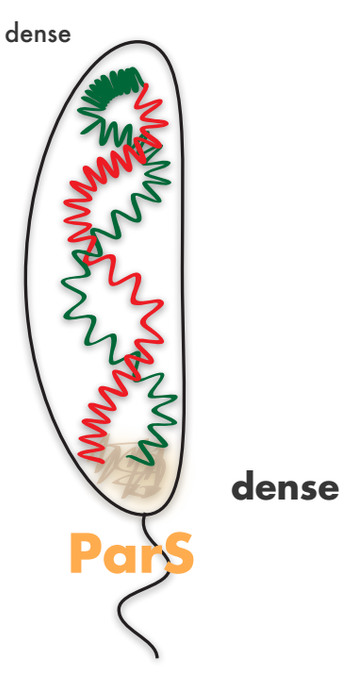


Arms are **STILL** helical

Genome architecture in Caulobacter





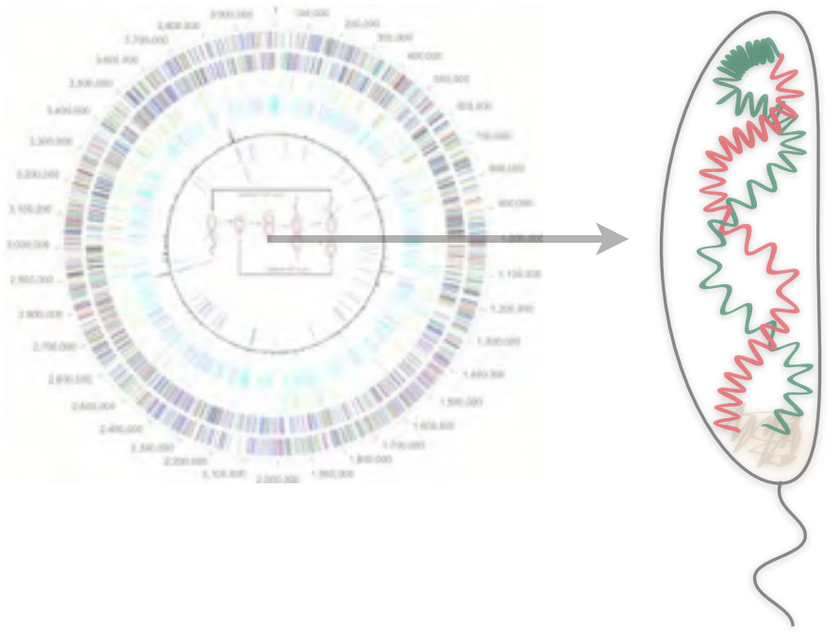


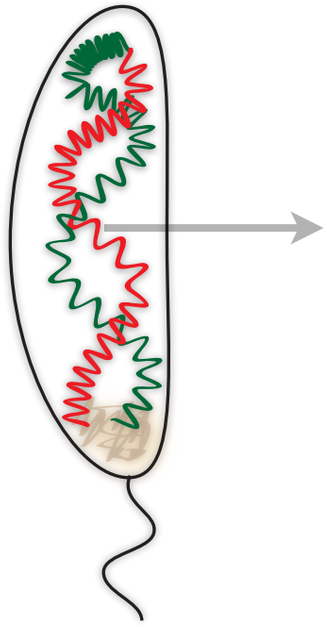
M.A. Umbarger, *et al.* **Molecular Cell** (2011) **44**:252–264

From Sequence to Function

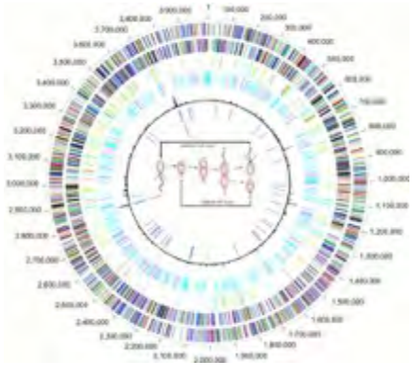
5C + IMP

Technology





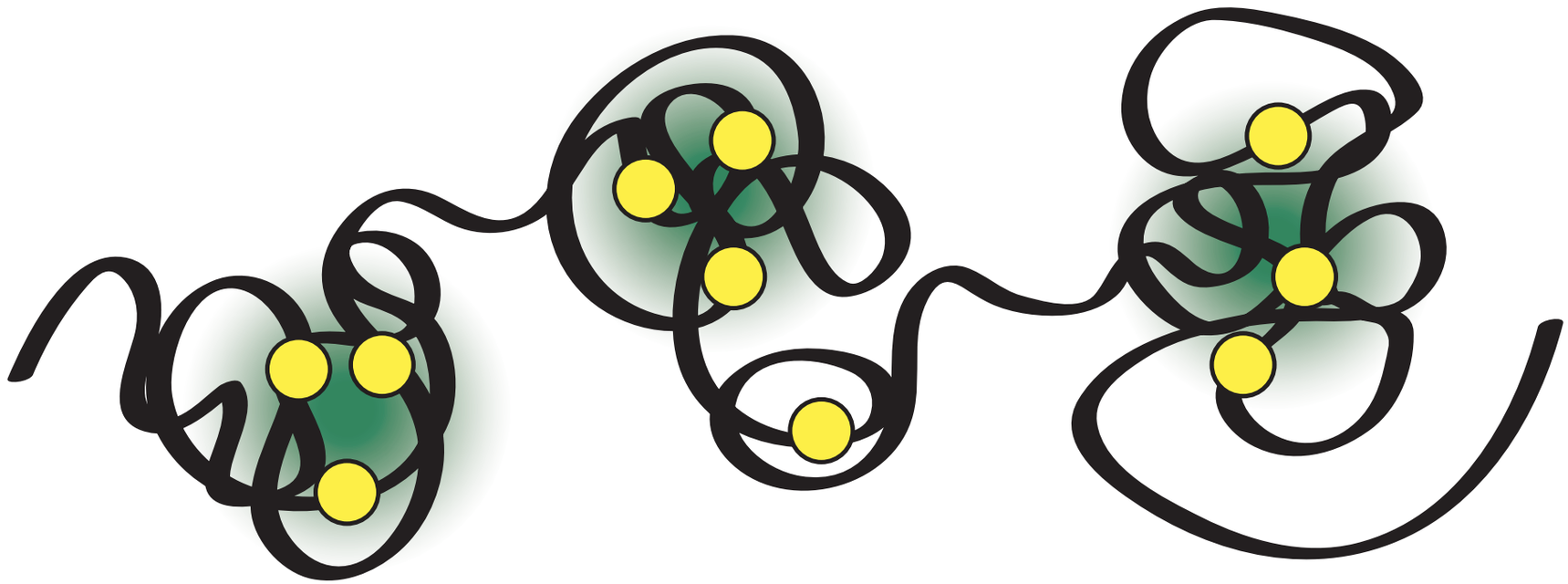
Hypothesis



Function!

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

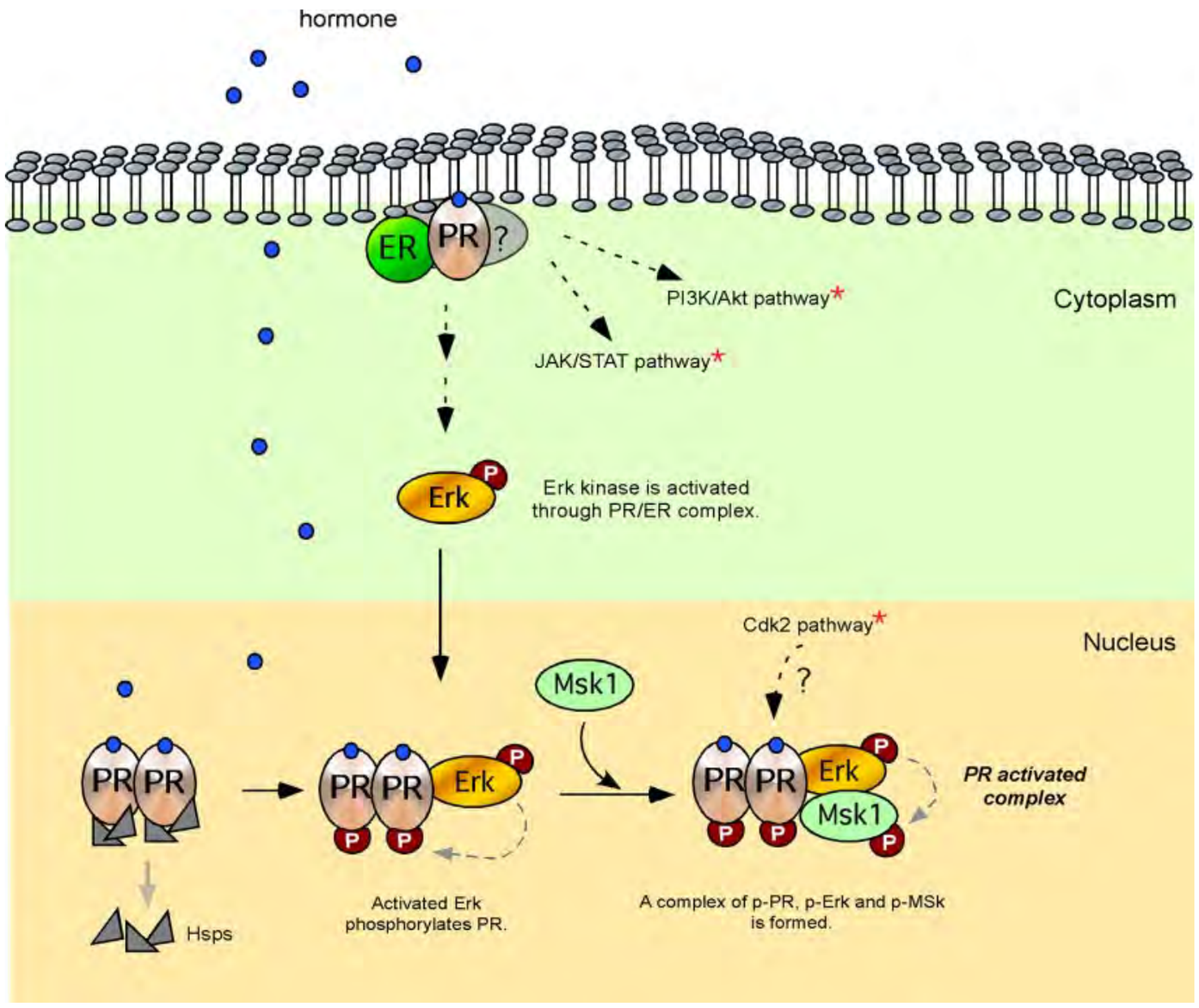
On TADs and hormones



Davide Baù

François le Dily

Progesterone-regulated transcription in breast cancer



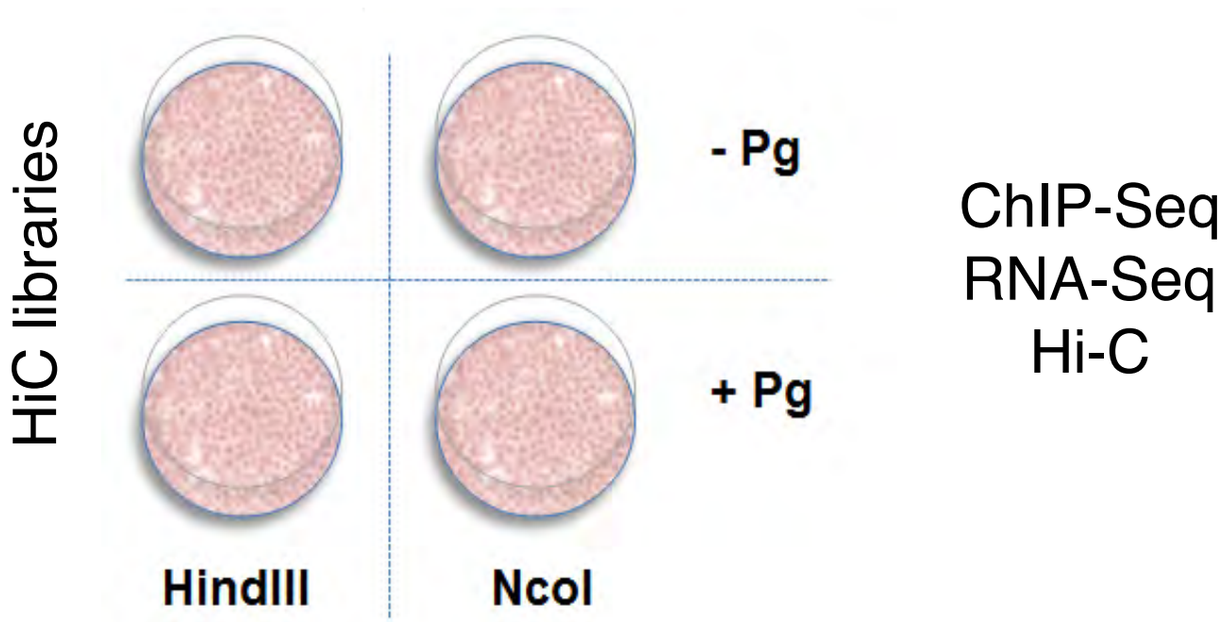
> 2,000 genes **Up**-regulated

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

Regulation in 3D?

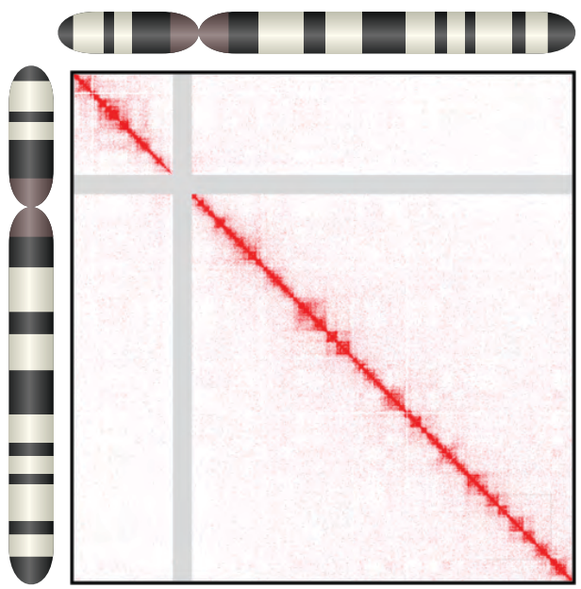
Vicent *et al* 2011, Wright *et al* 2012, Ballare *et al* 2012

Experimental design

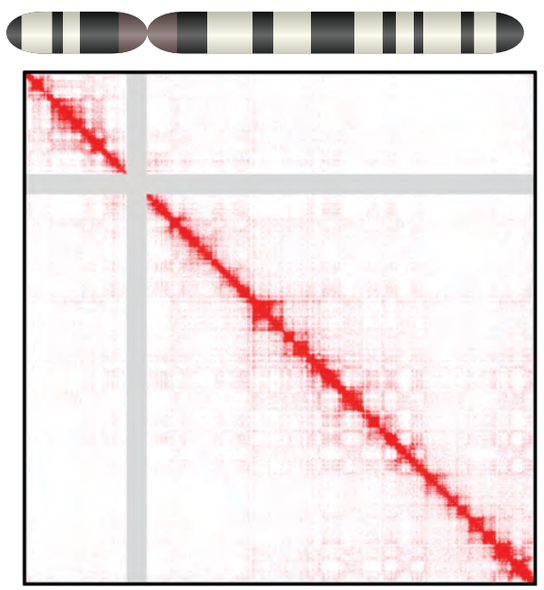


ChIP-Seq
RNA-Seq
Hi-C

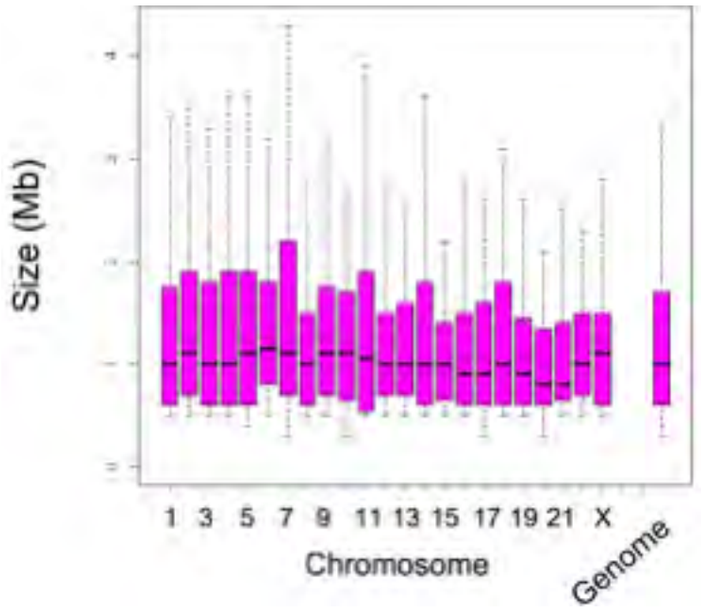
Chr.18 (Hind III)



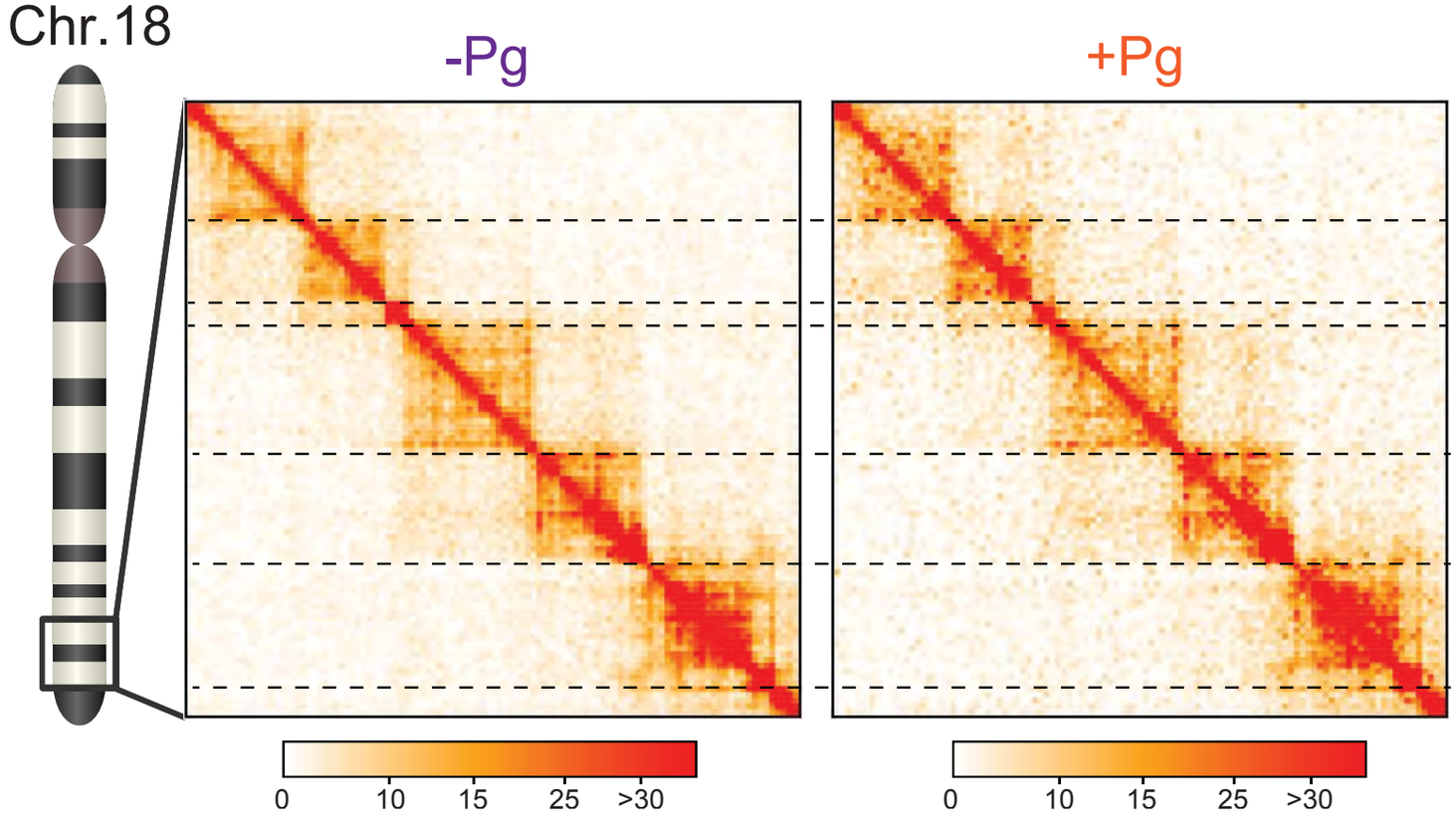
Chr.18 (NcoI)

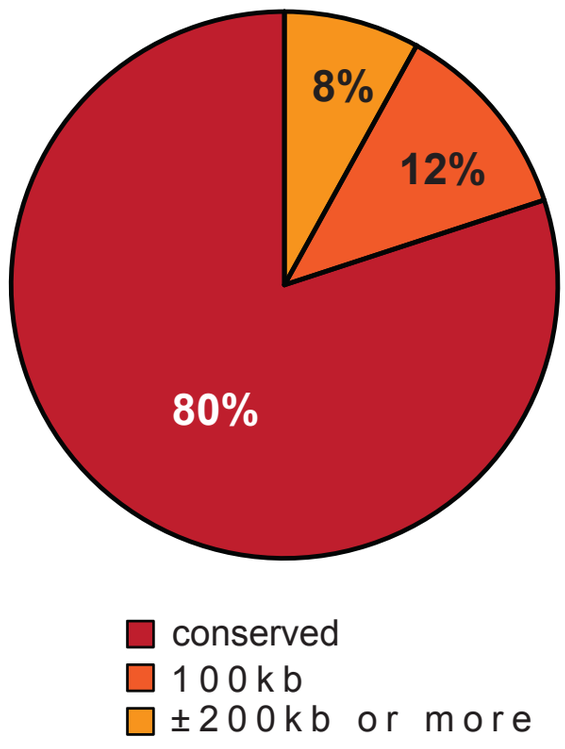


Are there TADs? how robust?

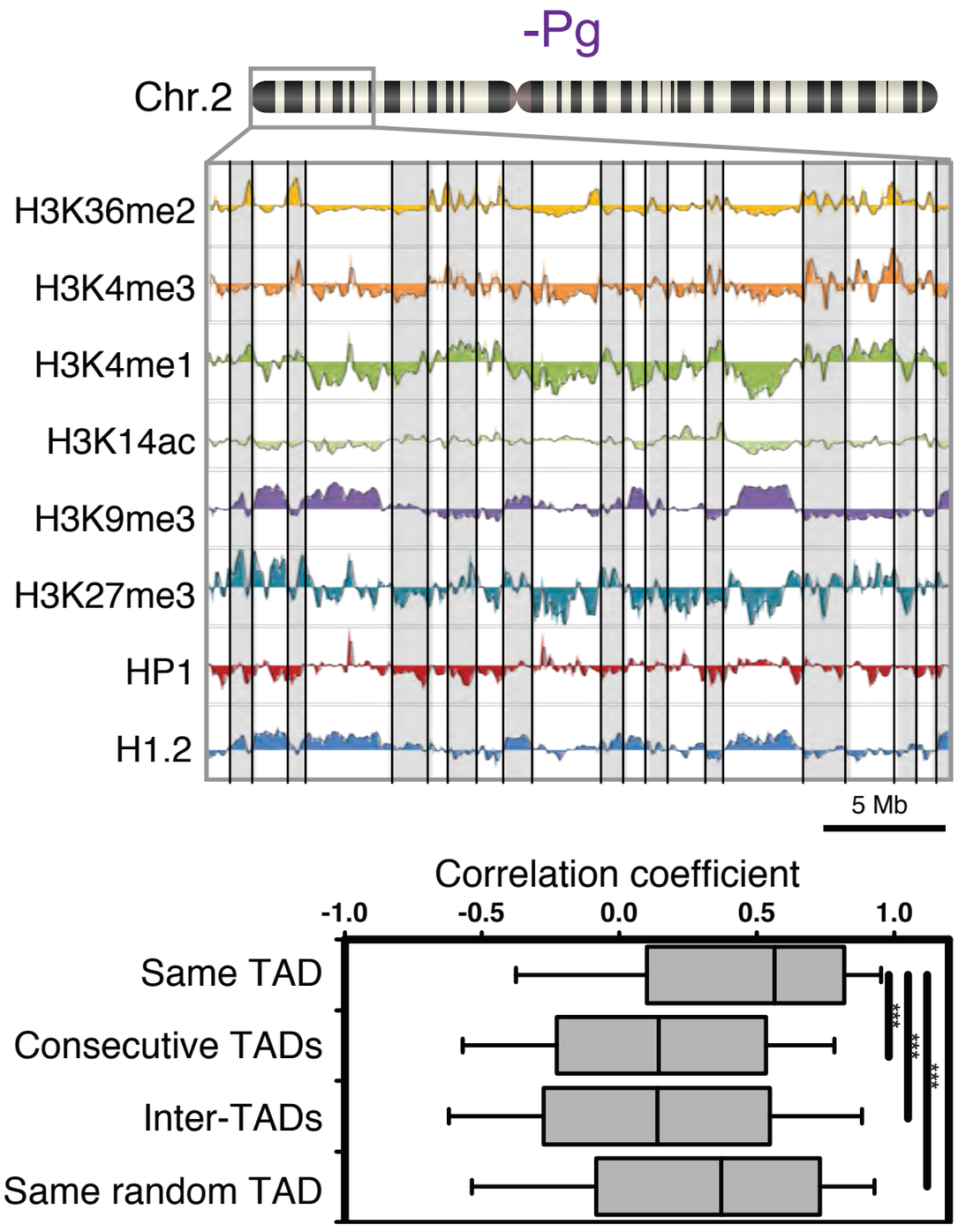


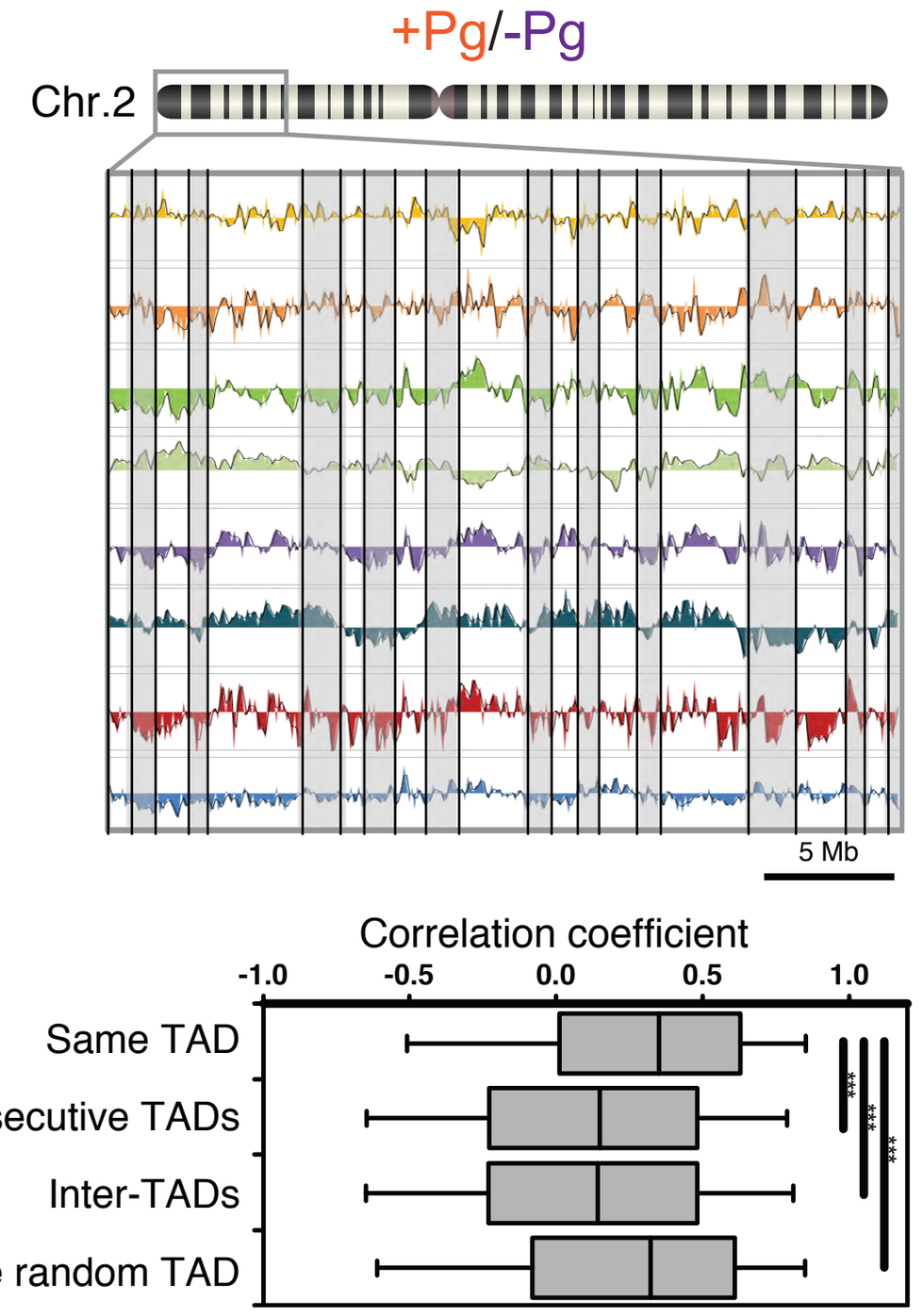
>2,000 detected TADs



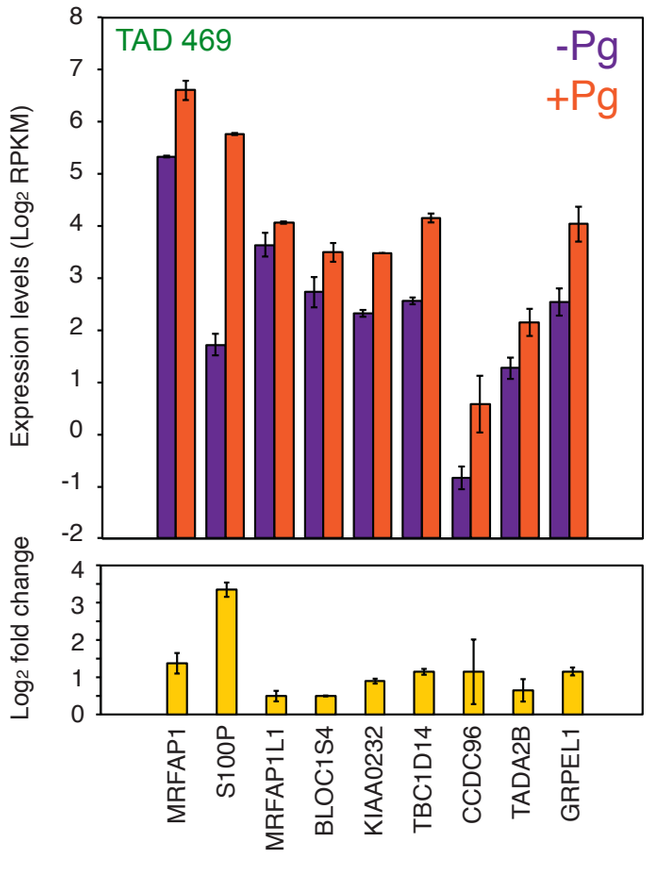


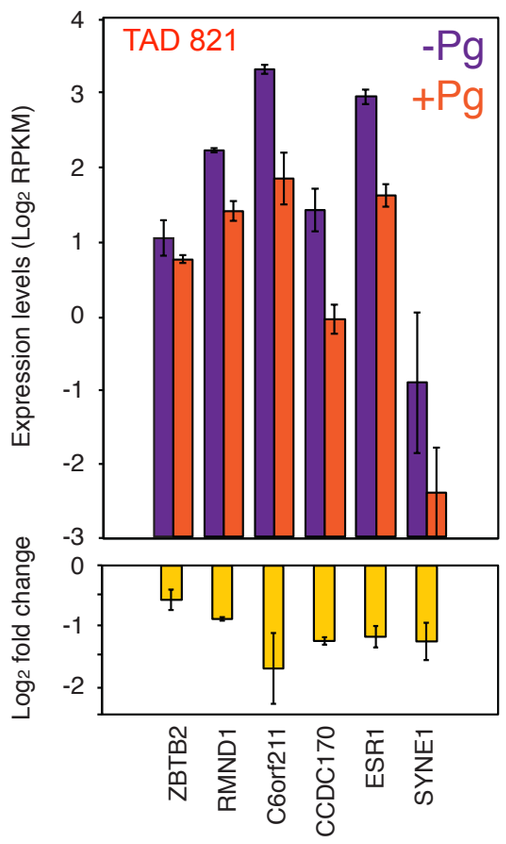
Are TADs homogeneous?

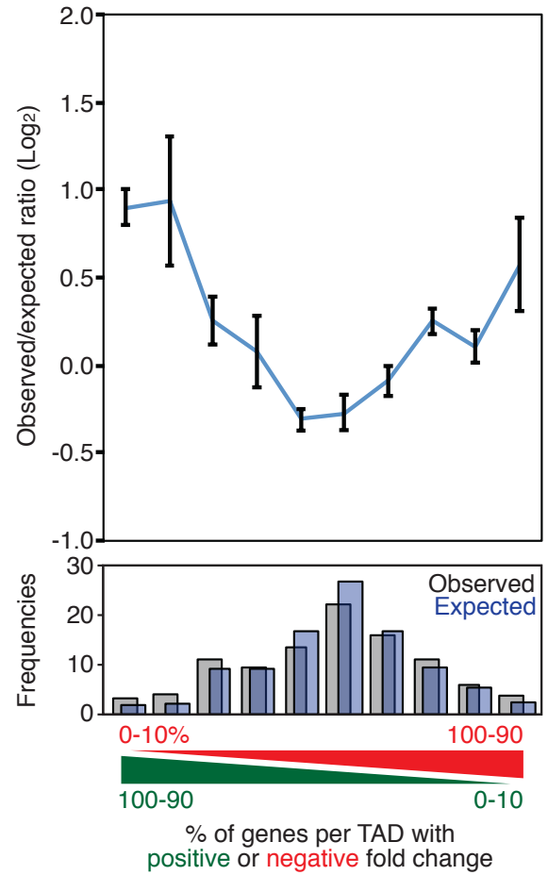




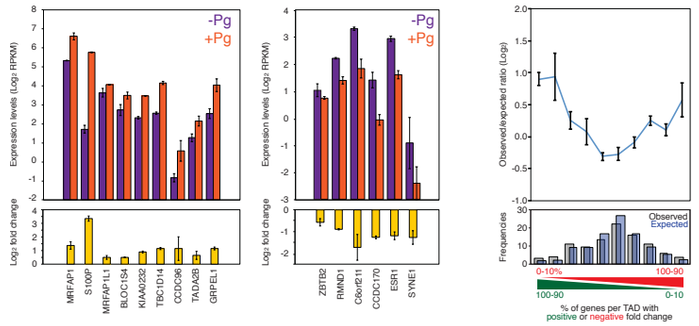
Do TADs respond differently to Pg treatment?



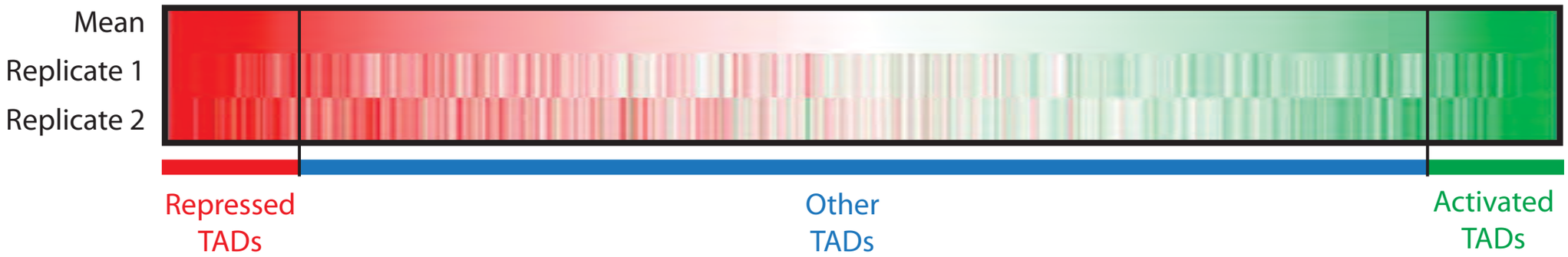




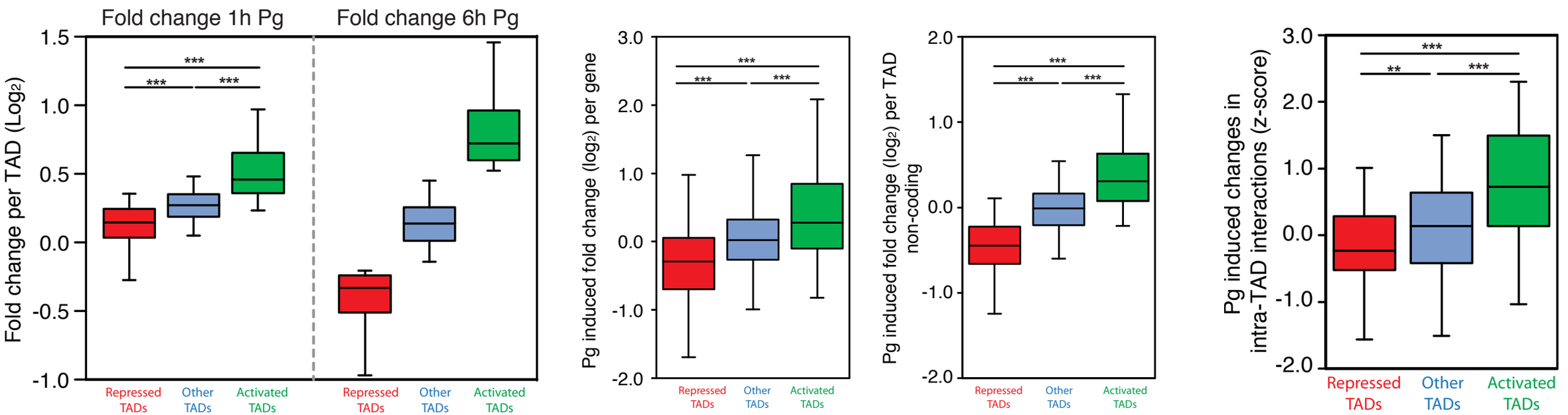
Do TADs respond differently to Pg treatment?



Pg induced fold change per TAD (6h)

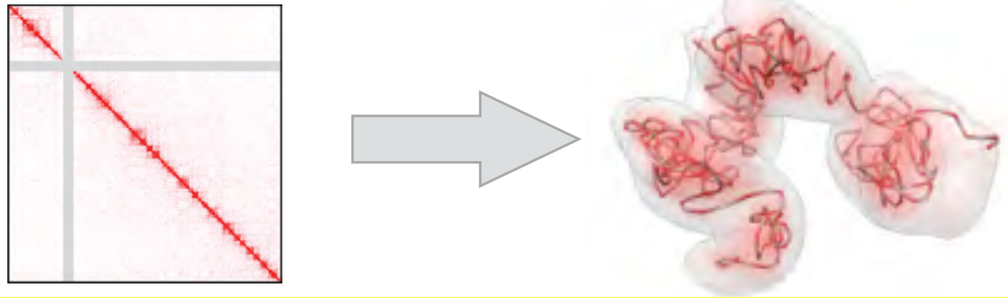


Repressed TADs | Other TADs | Activated TADs

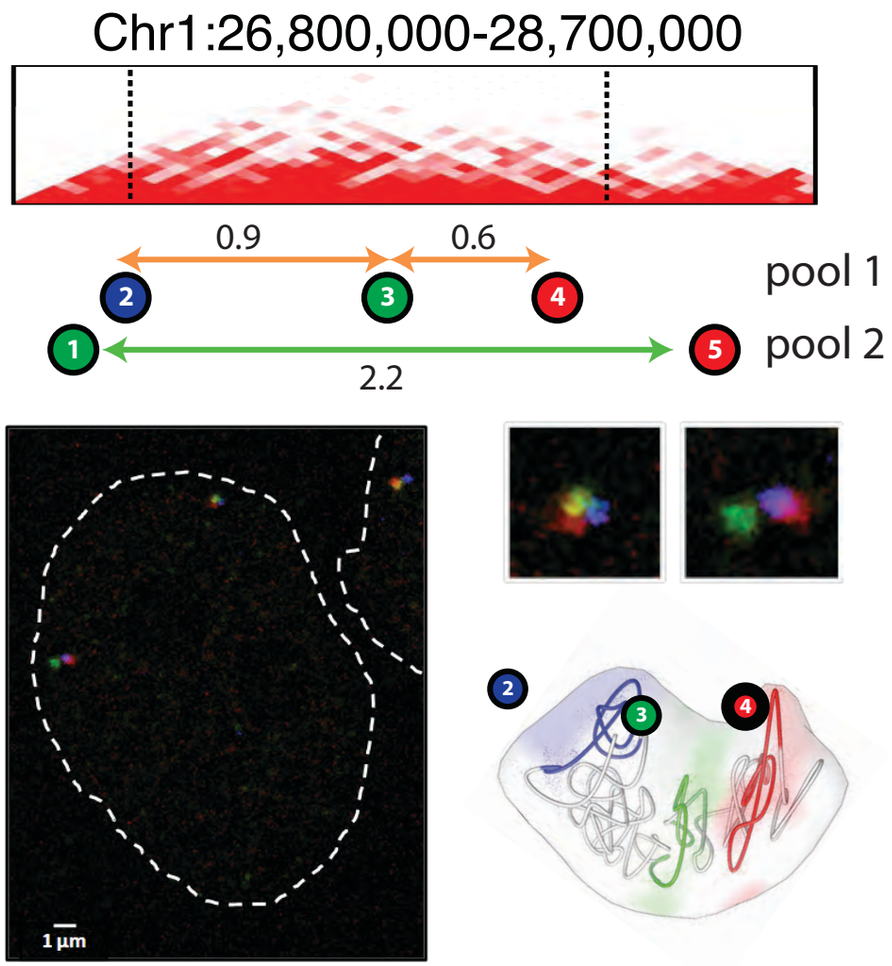


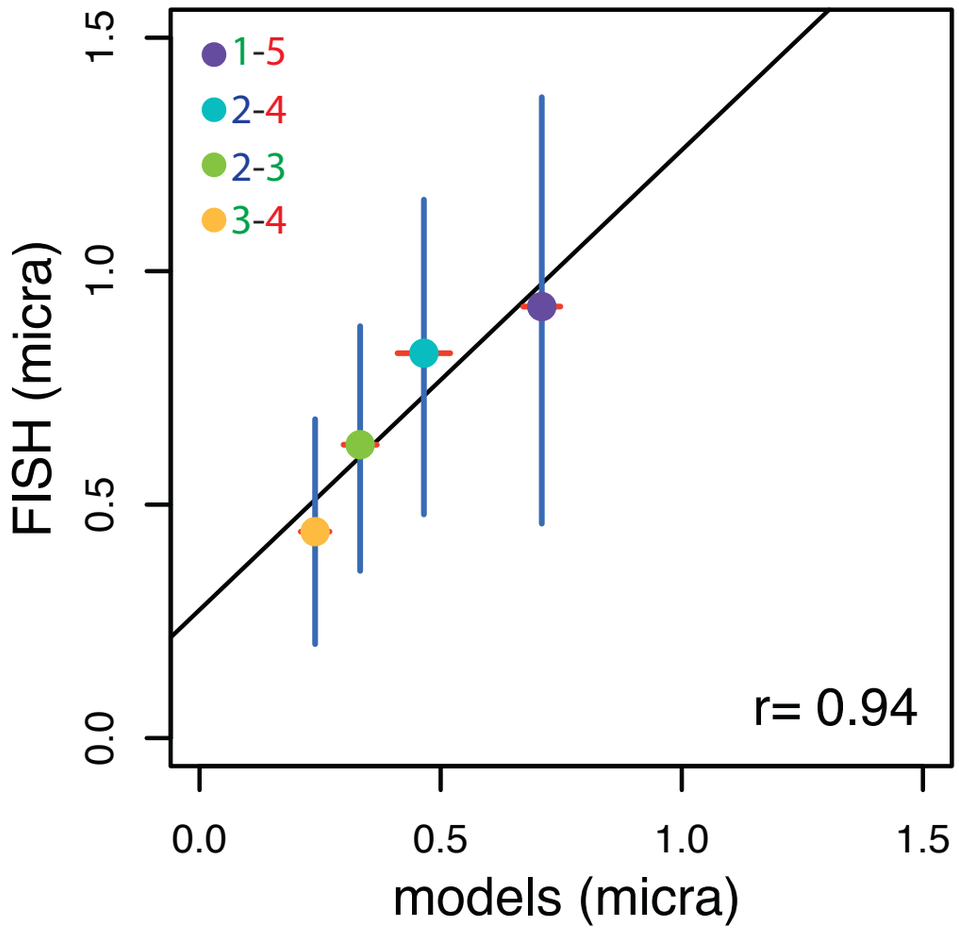
cnag CRG

Modeling 3D TADs



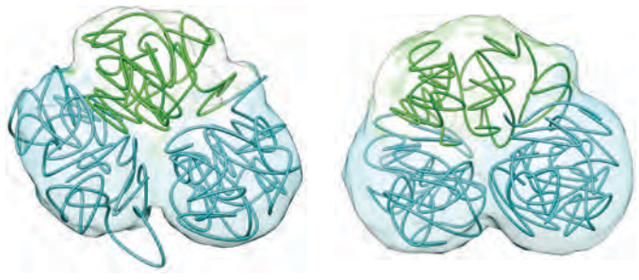
61 genomic regions containing 209 TADs covering 267Mb

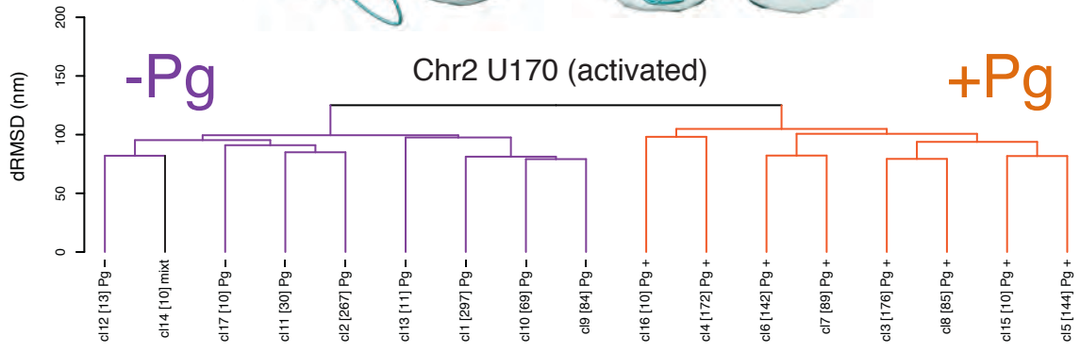




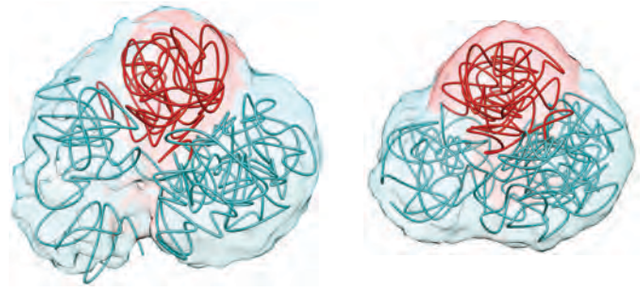
How TADs respond structurally to Pg?

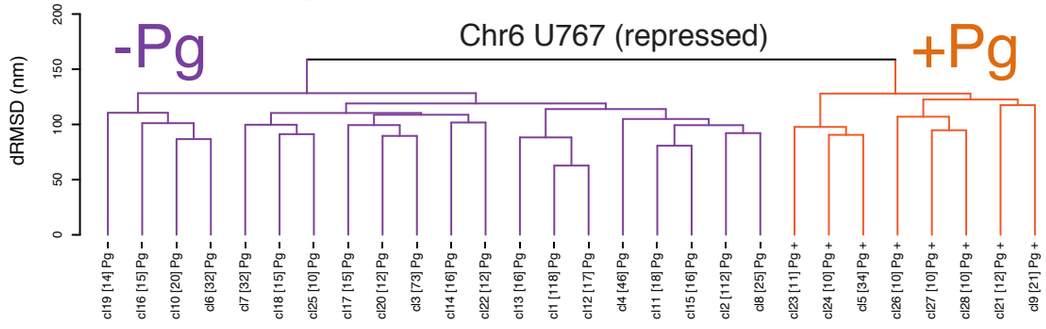
Chr2:9,600,000-13,200,000

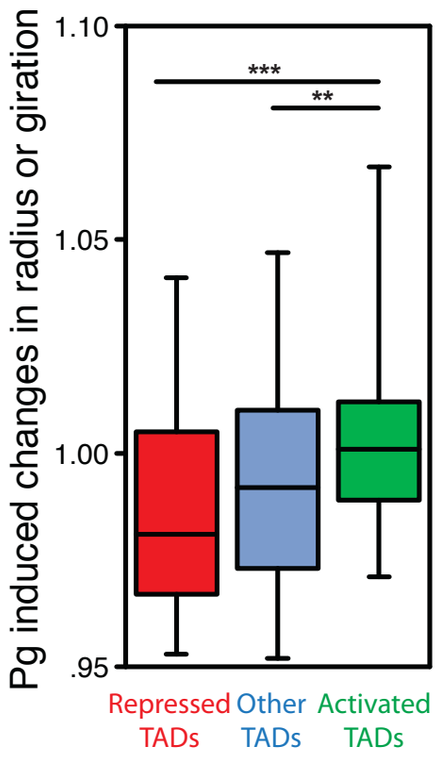


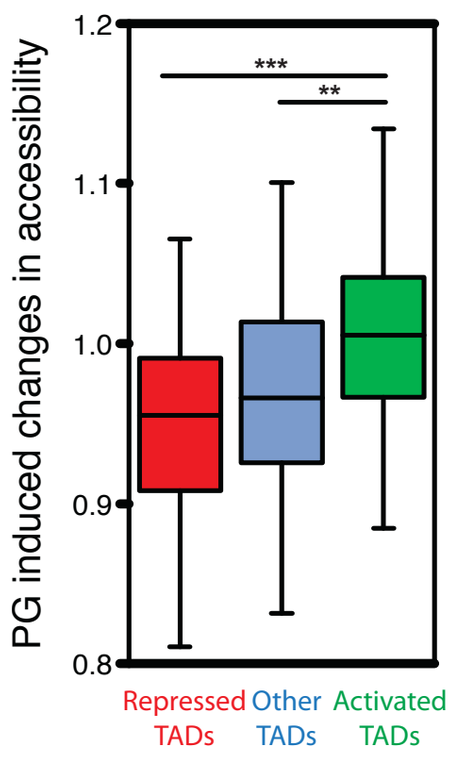


Chr6:71,800,000-76,500,000

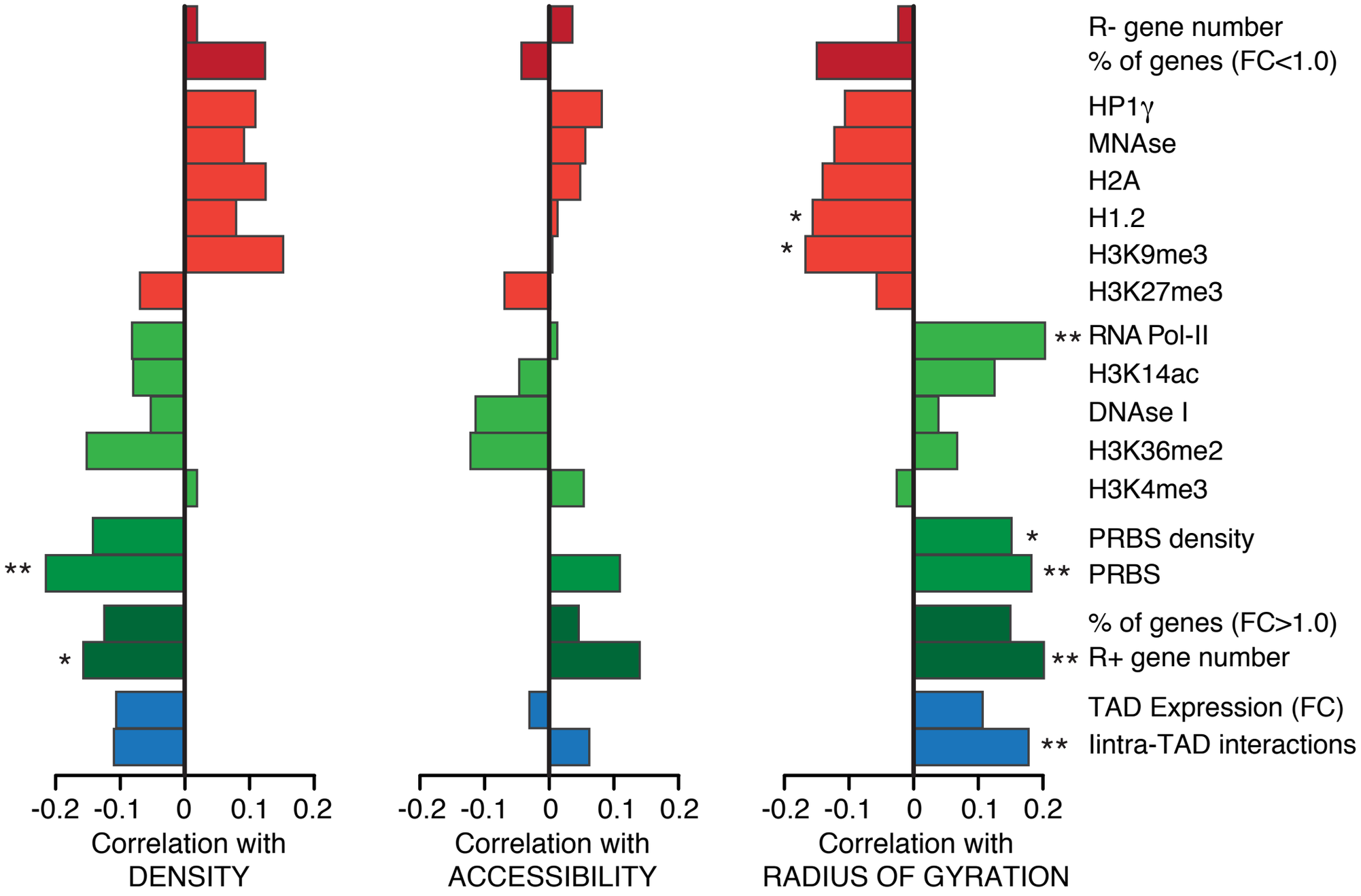


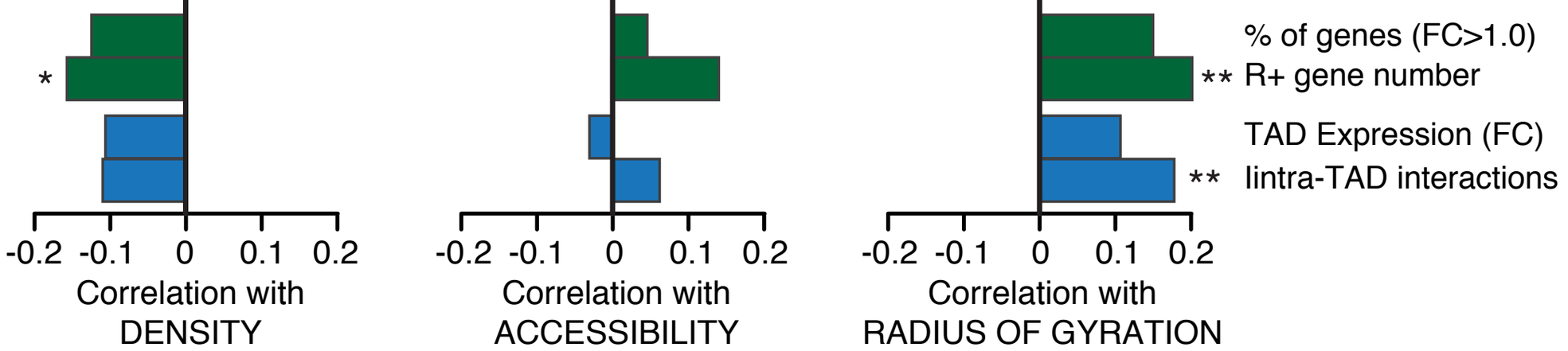


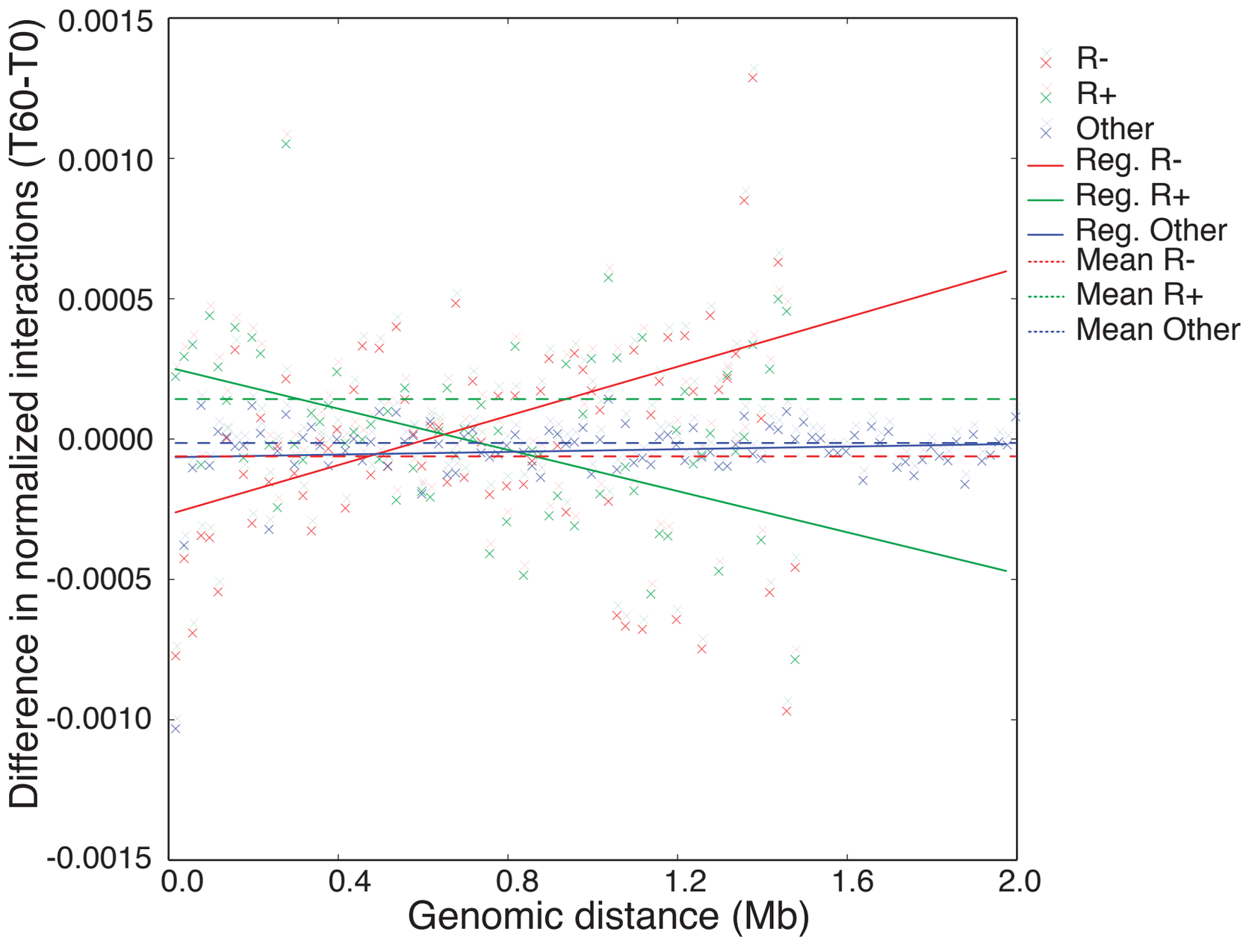




How TADs respond structurally to Pg?



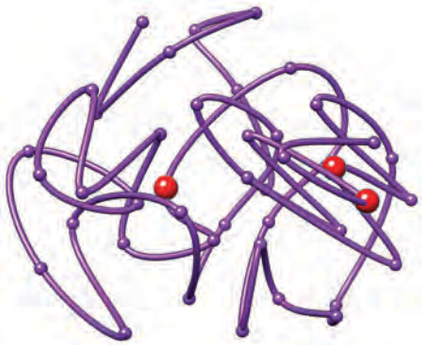

*
-0.2 -0.1 0 0.1 0.2
Correlation with
DENSITY
-0.2 -0.1 0 0.1 0.2
Correlation with
ACCESSIBILITY
** % of genes (FC>1.0)
** R+ gene number
TAD Expression (FC)
** lintra-TAD interactions
-0.2 -0.1 0 0.1 0.2
Correlation with
RADIUS OF GYRATION

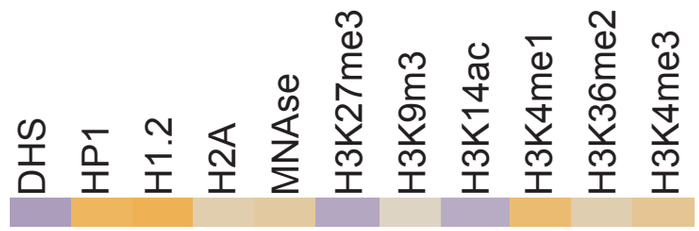

Difference in normalized interactions (T60-T0)
0.0015
0.0010
0.0005
0.0000
-0.0005
-0.0010
-0.0015
0.0
0.4
0.8
1.2
1.6
2.0
Genomic distance (Mb)
R-
R+
Other
Reg. R-
Reg. R+
Reg. Other
Mean R-
Mean R+
Mean Other

Model for TAD regulation

Repressed TAD

chr1 U41

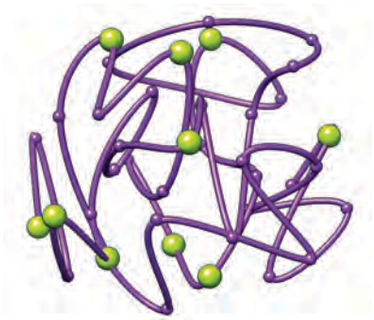


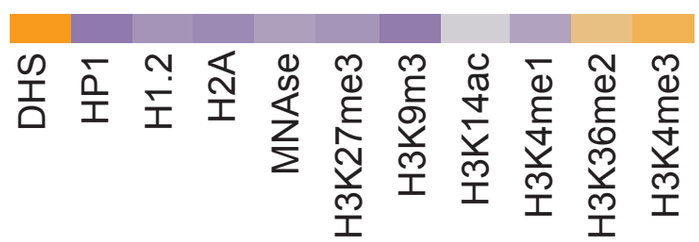


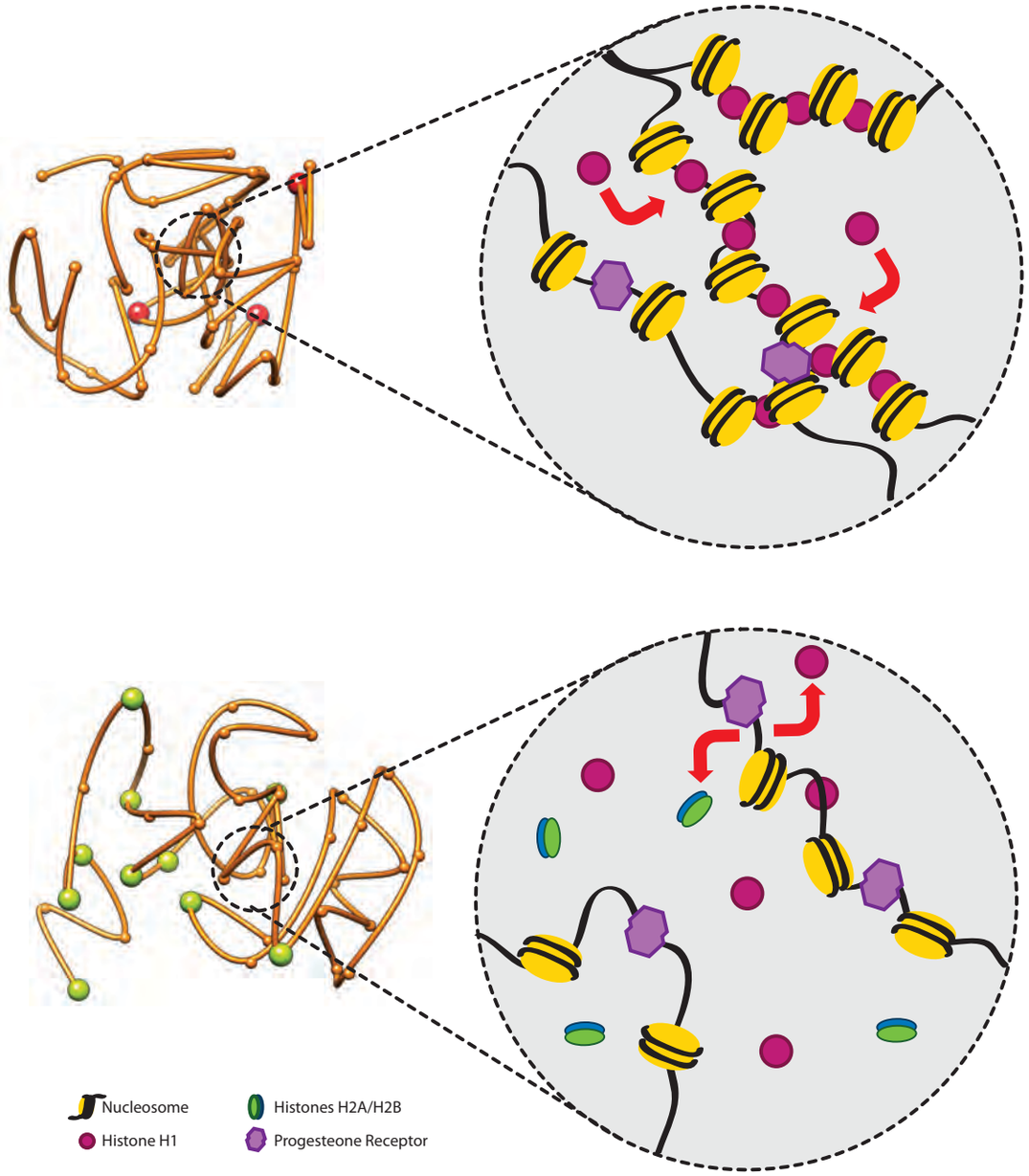
Structural transition **+Pg**

Activated TAD

chr2 U207







cnag CRG

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

Acknowledgments

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