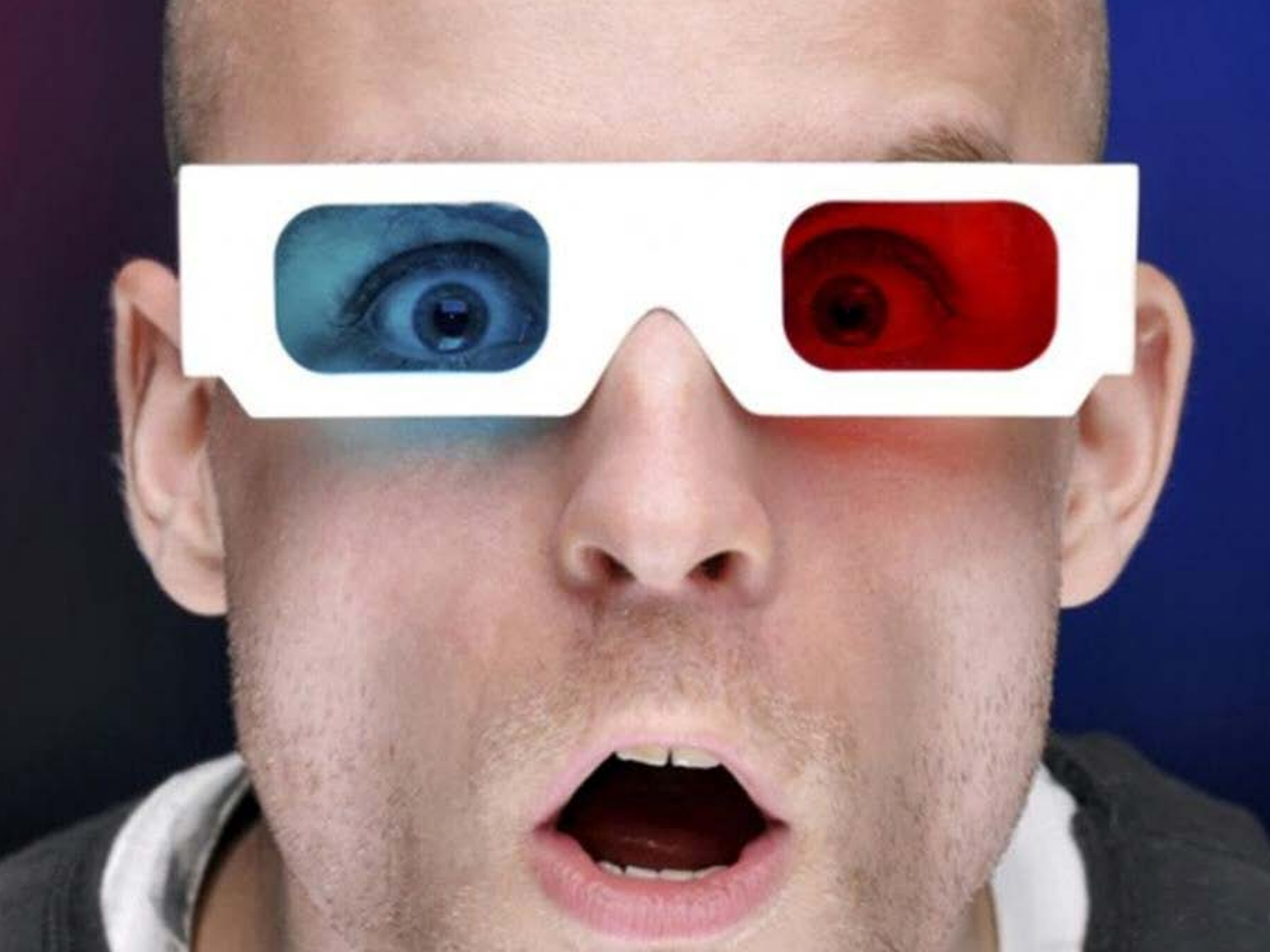
3DGenomics

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

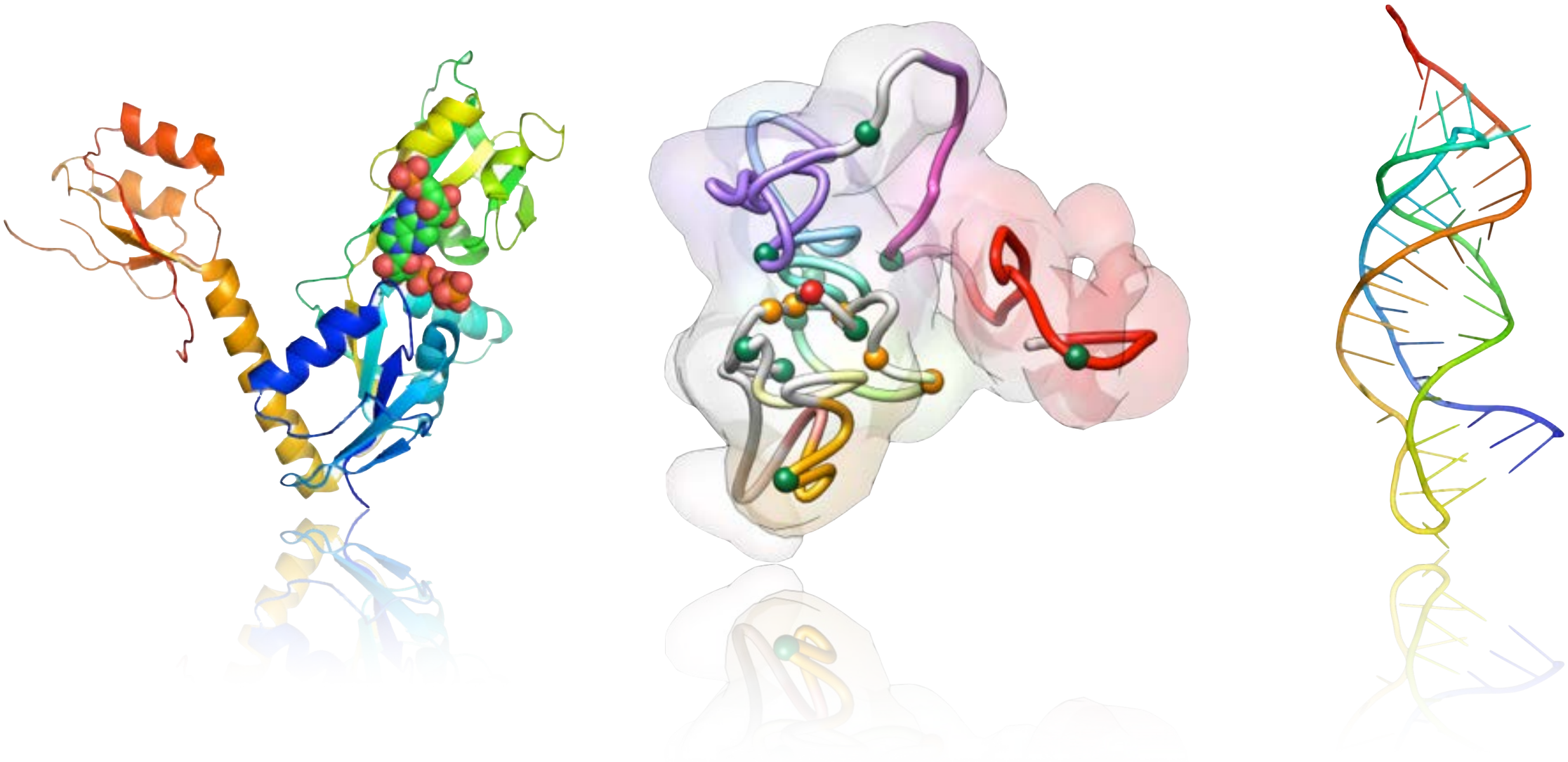
ICREA

cnag CRG Centre for Genomic Regulation

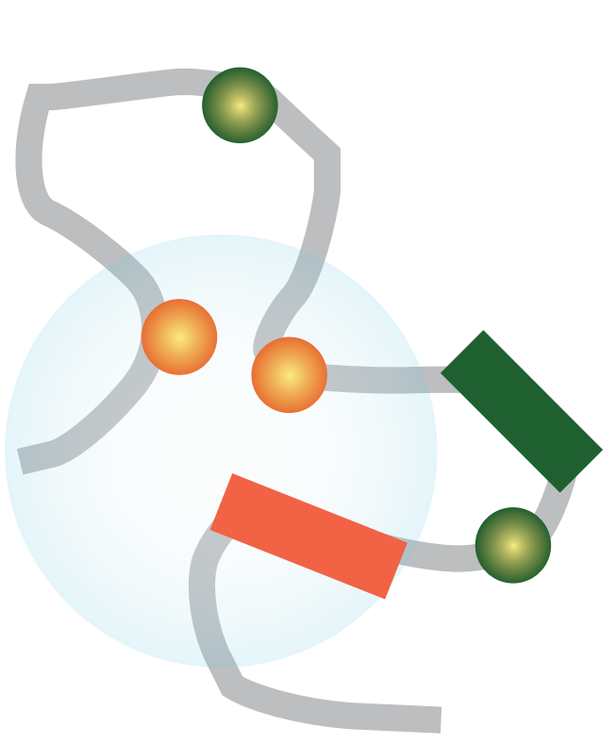


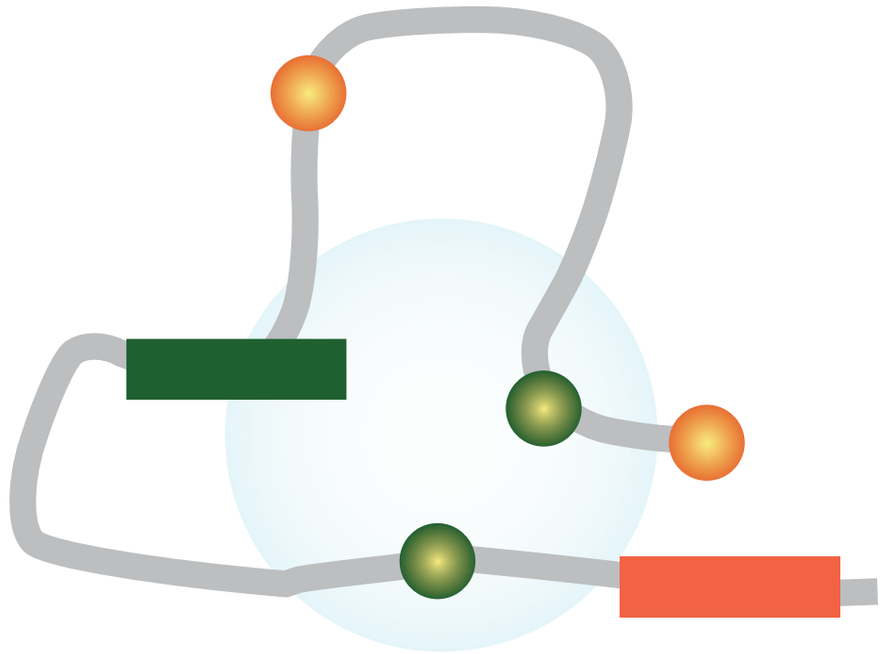
Structural Genomics Group

http://www.marciuslab.org



Complex genome organization





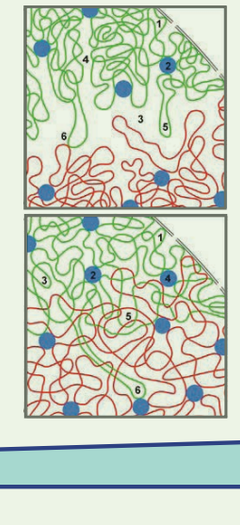
Resolution Gap

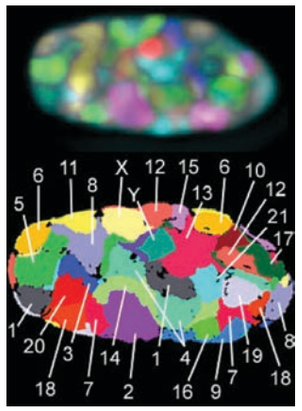
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)

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

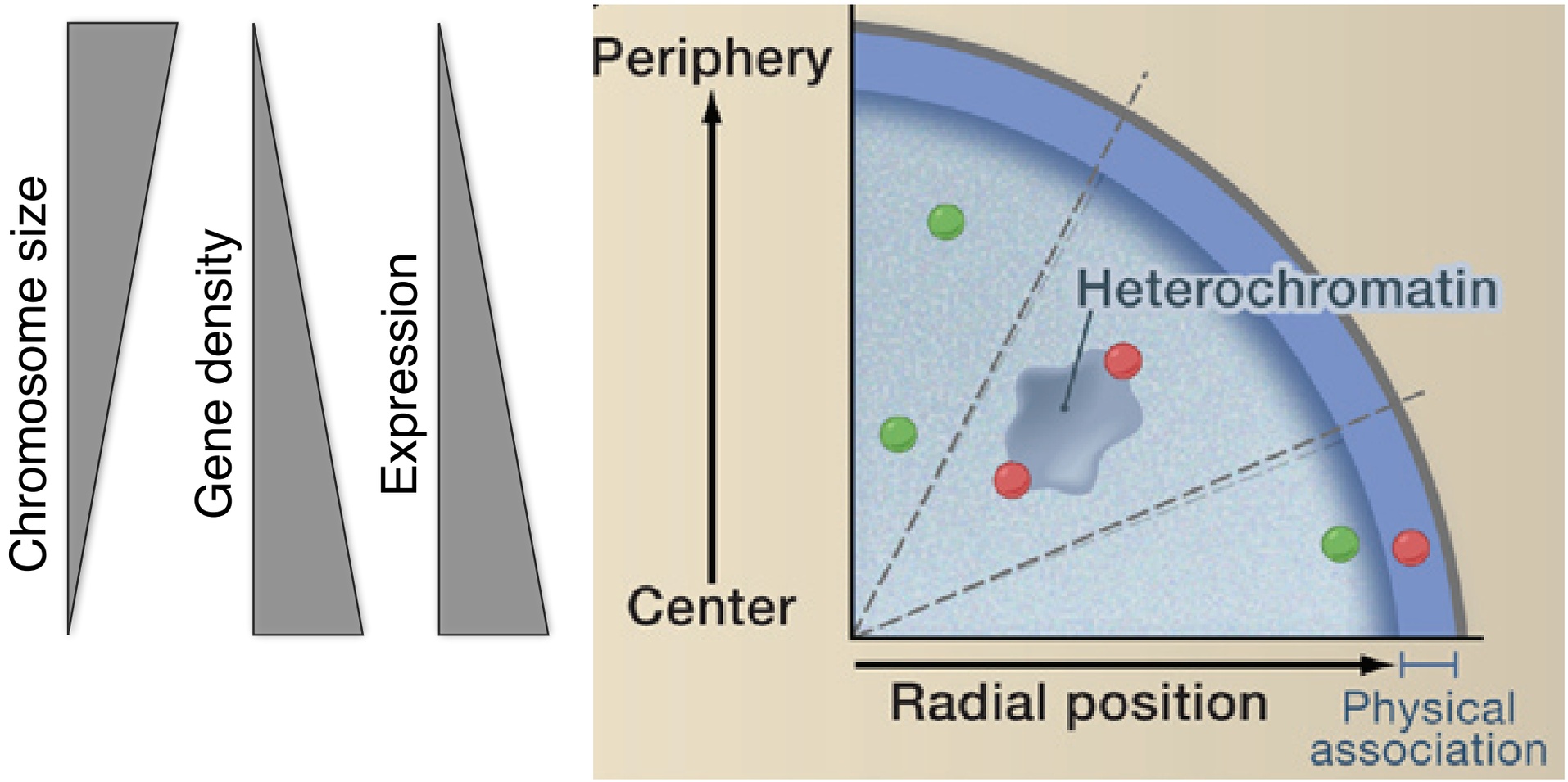
10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^0 10^2 10^3 s

Resolution

10^{-3} 10^{-2} 10^{-1} μ

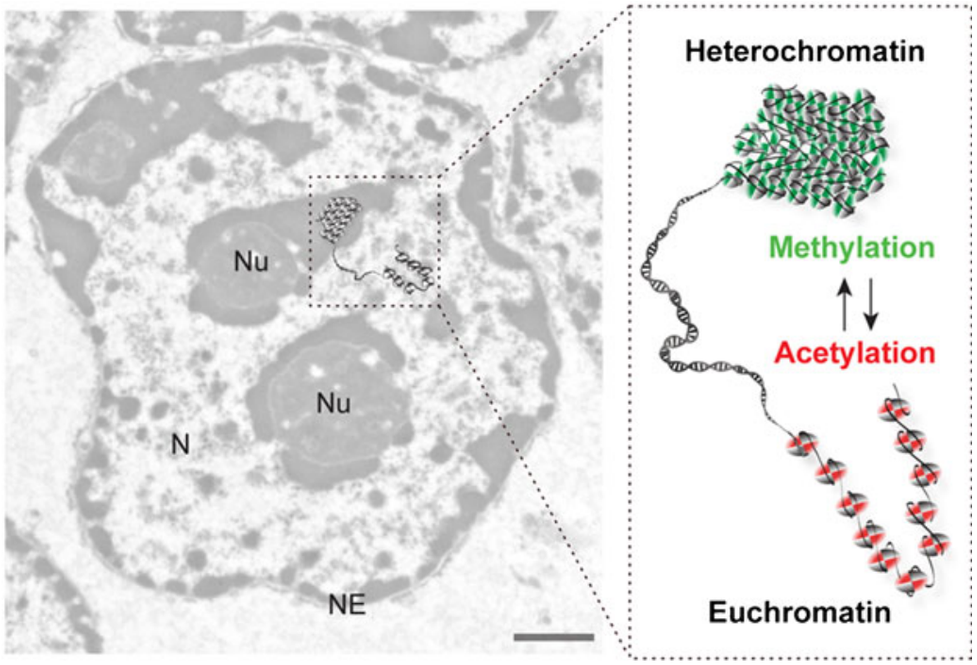
Level I: Radial genome organization

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

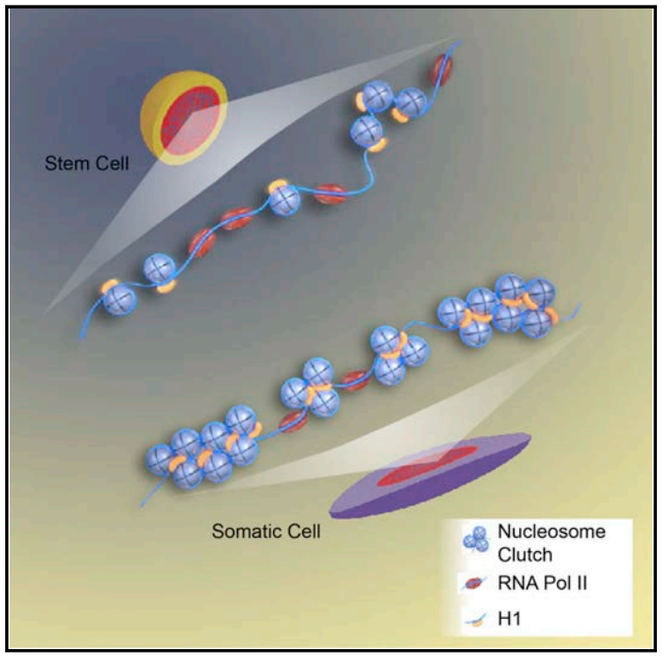


Level II: Euchromatin vs heterochromatin

Electron microscopy



Nanoscopy

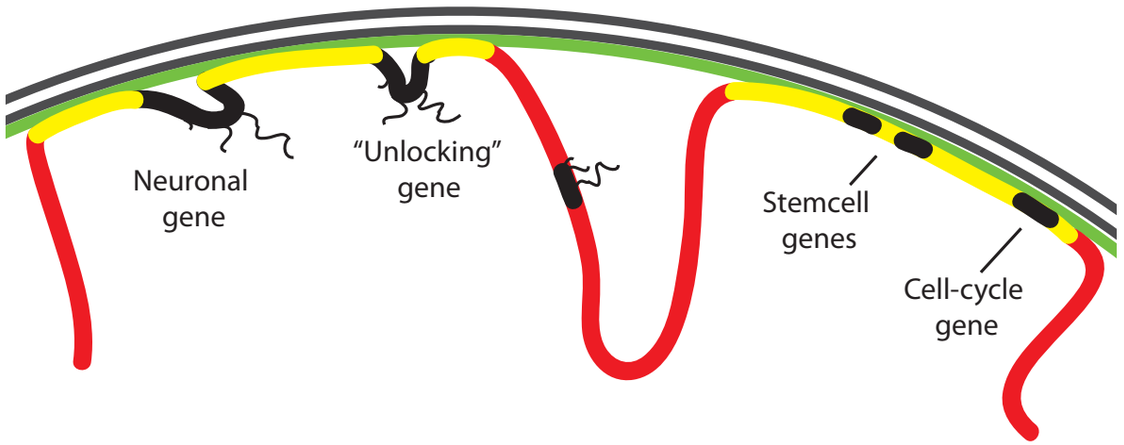


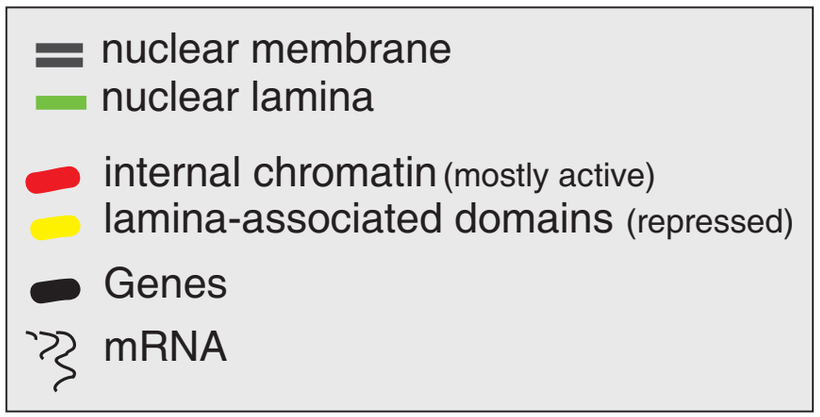
Euchromatin:
chromatin that is located away from the nuclear lamina, is generally less densely packed, and contains actively transcribed genes

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

Adapted from Cell, 160(6), 1145-1158. 2015

Level III: Lamina-genome interactions



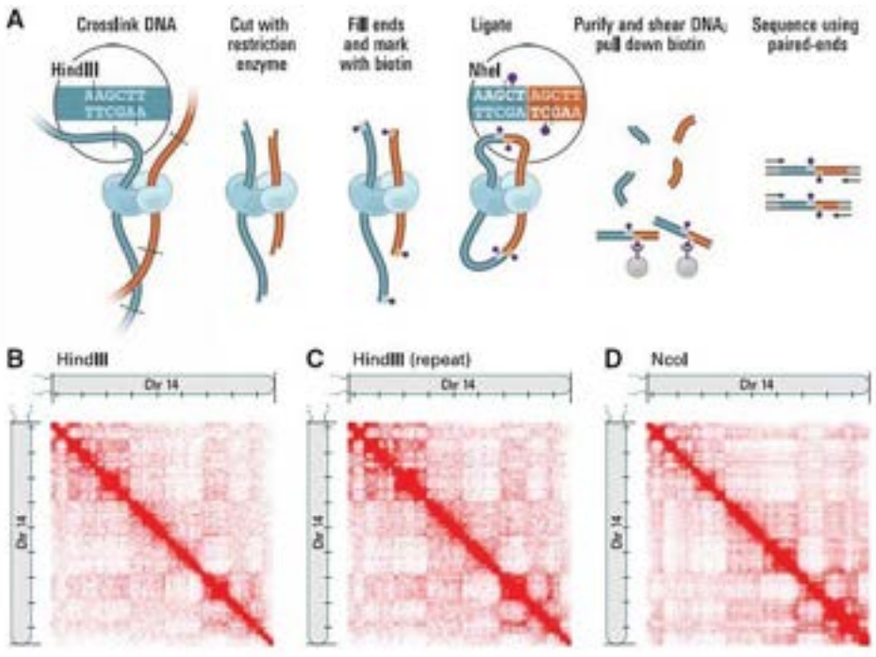


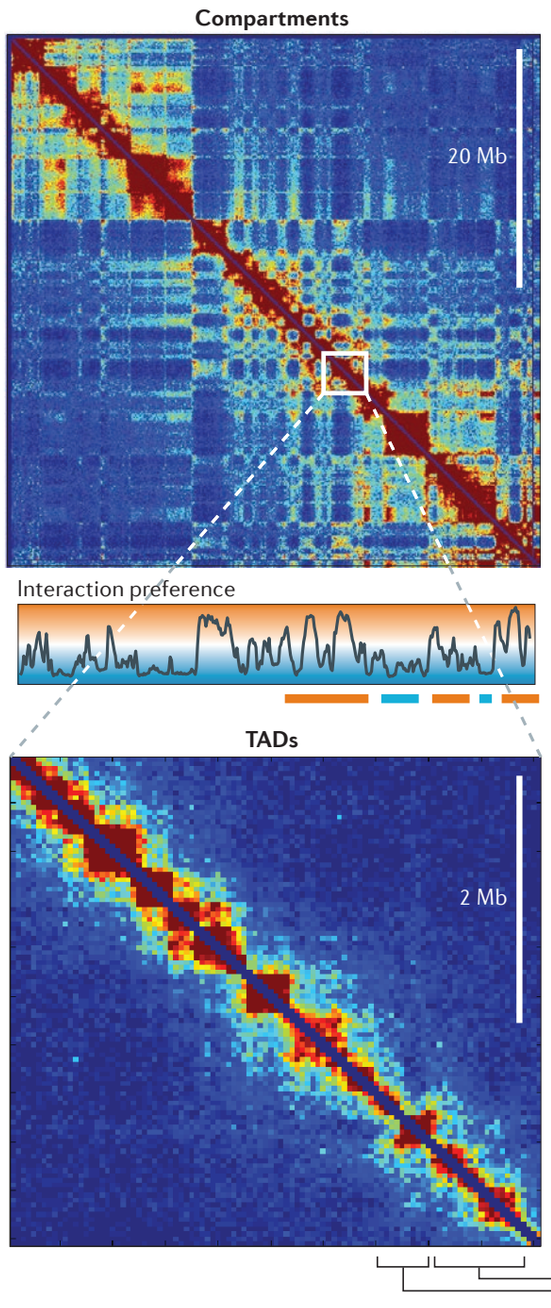
Most genes in Lamina Associated Domains are transcriptionally silent, suggesting that **lamina-genome interactions** are widely involved in the control of **gene expression**

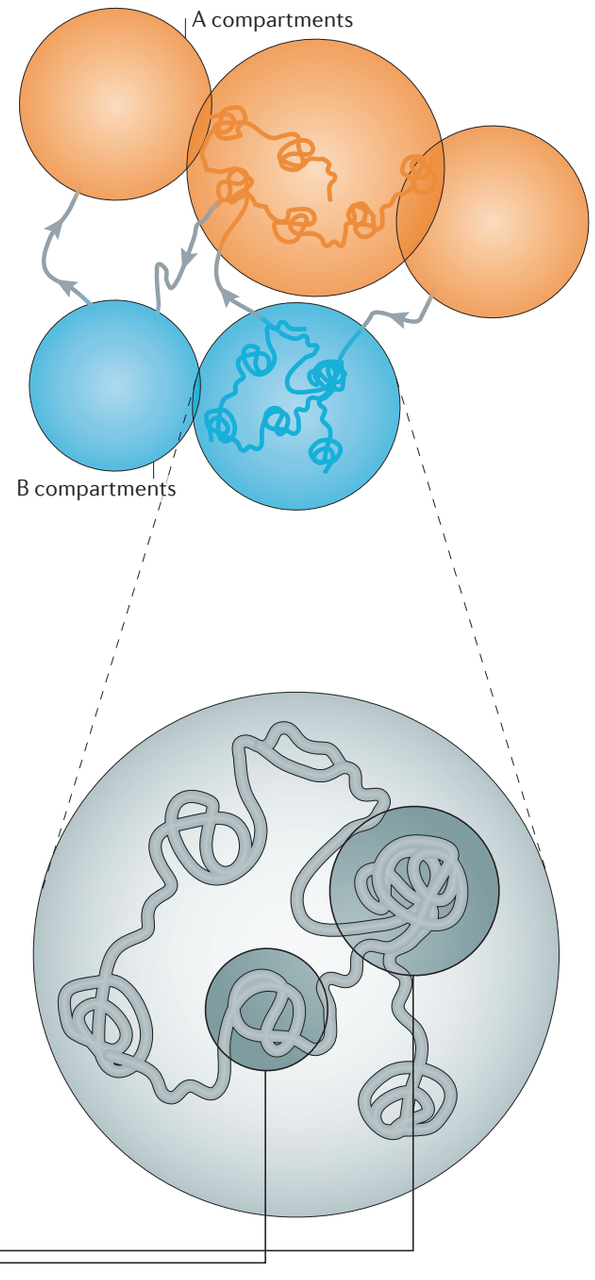
Adapted from Molecular Cell 38, 603-613, 2010

Level IV: Higher-order organization

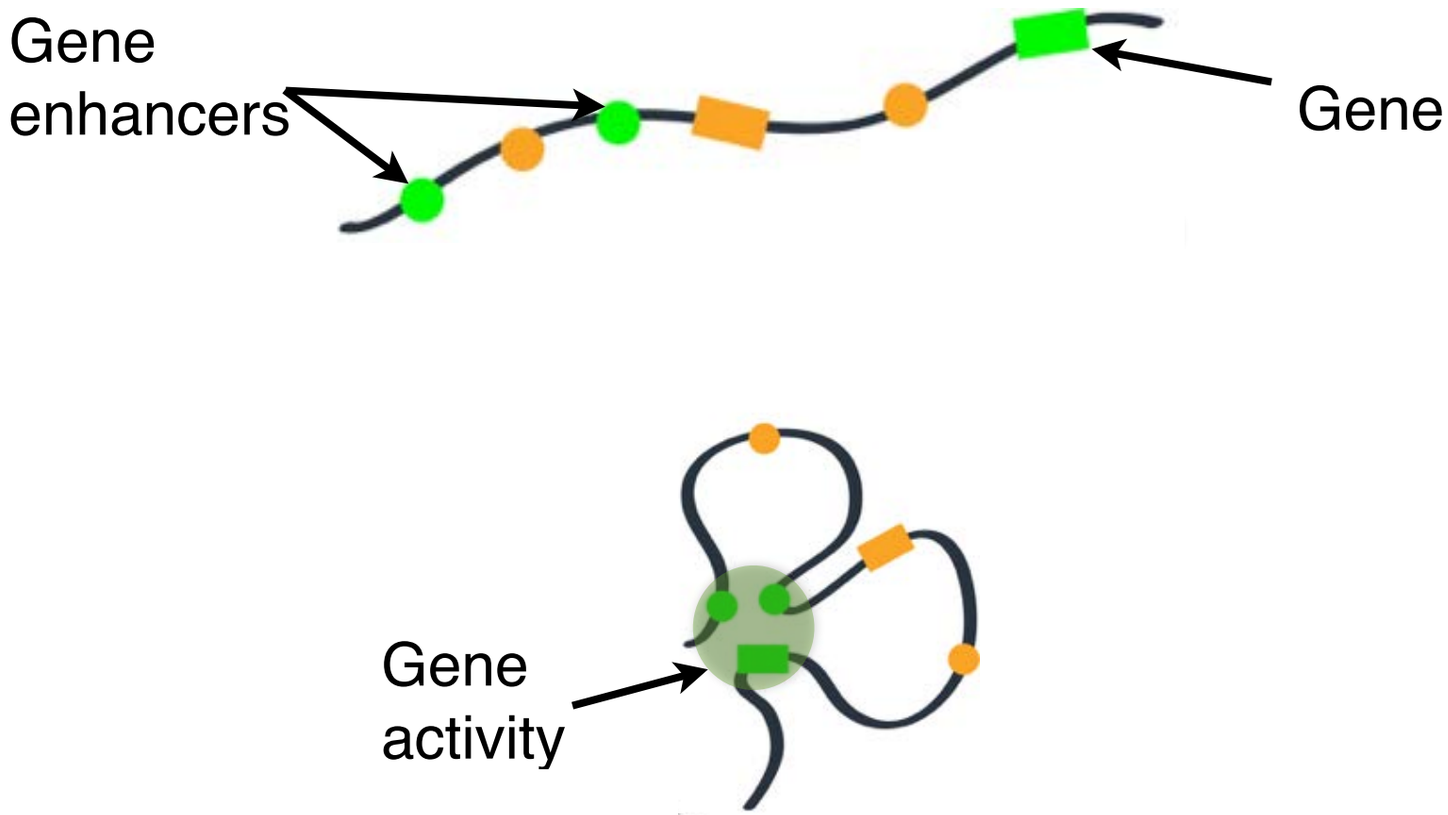
Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390–403 (2013).







Level V: Chromatin loops



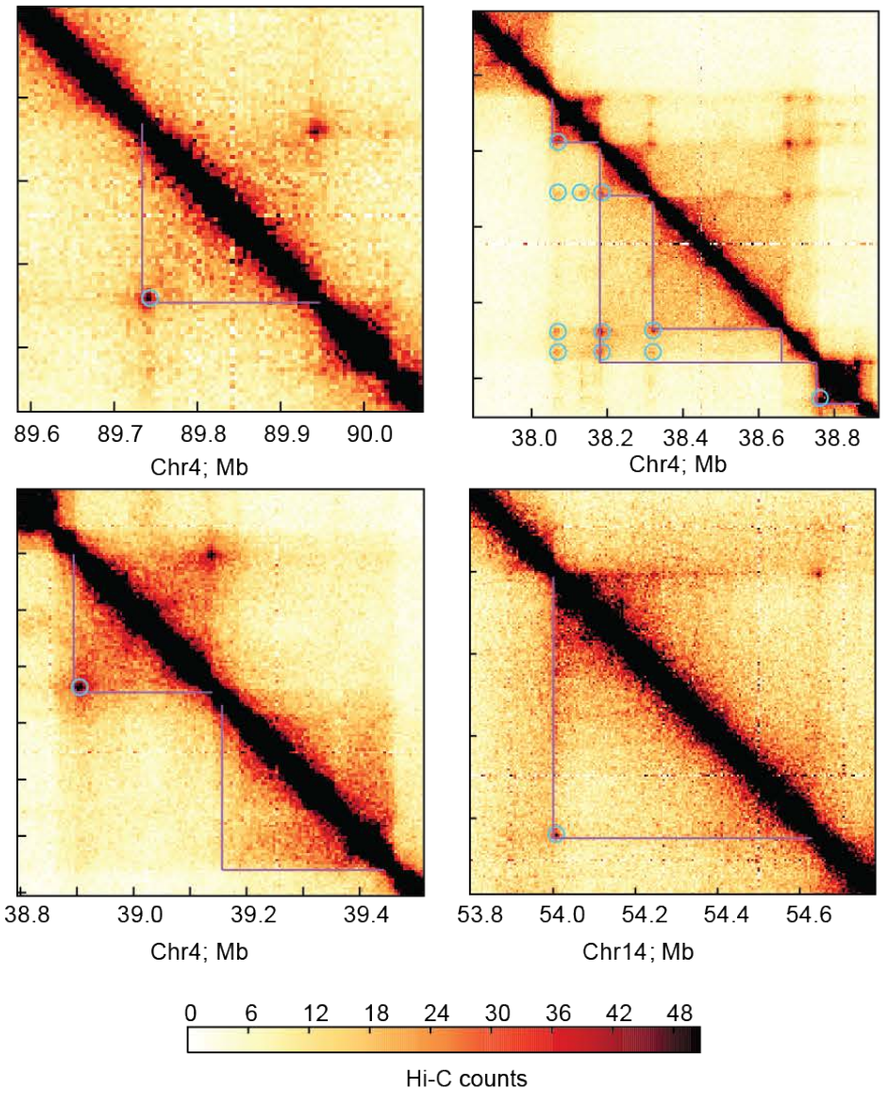
Loops bring distal genomic regions in close proximity to one another

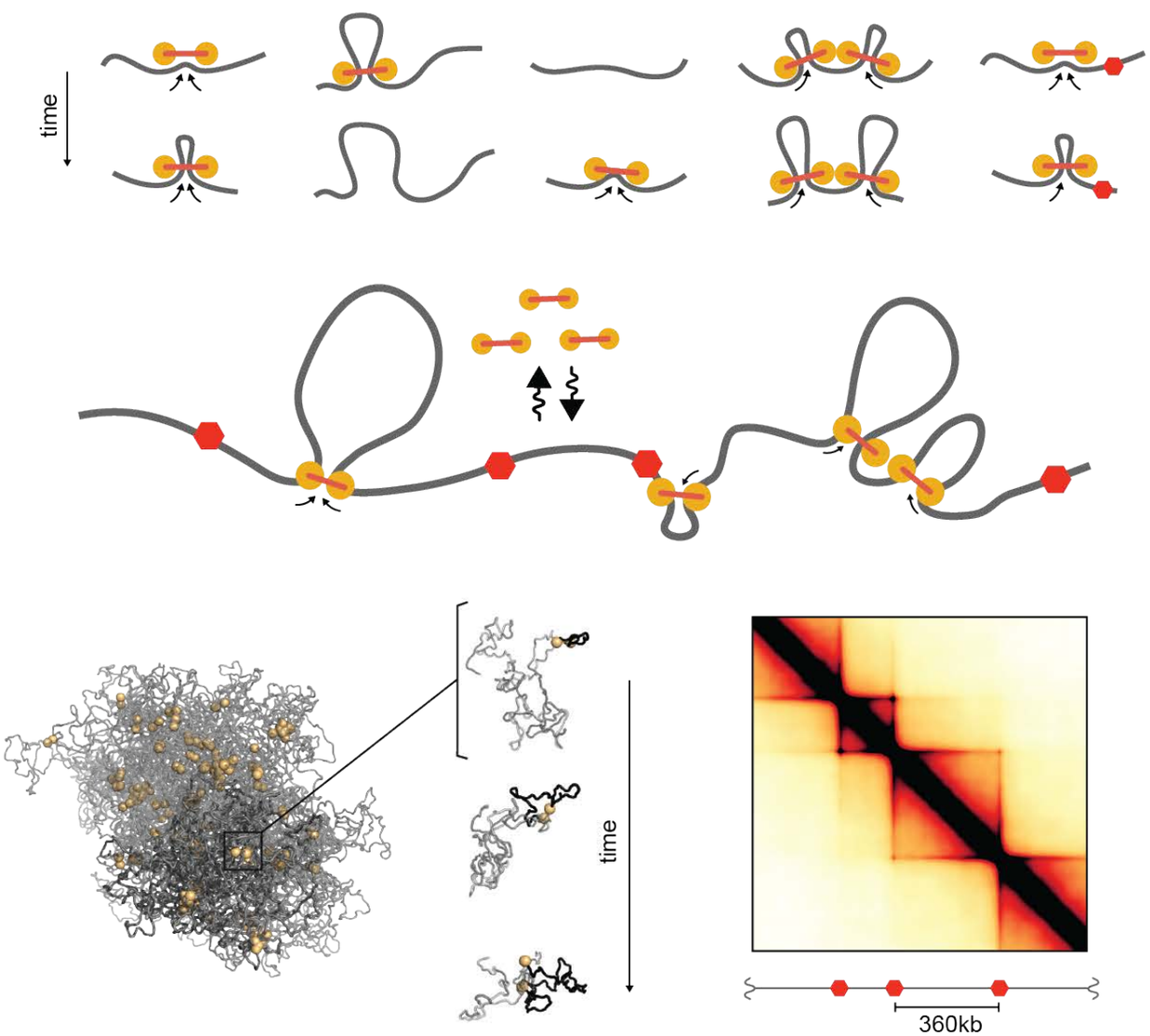
This in turn can have profound effects on gene transcription

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

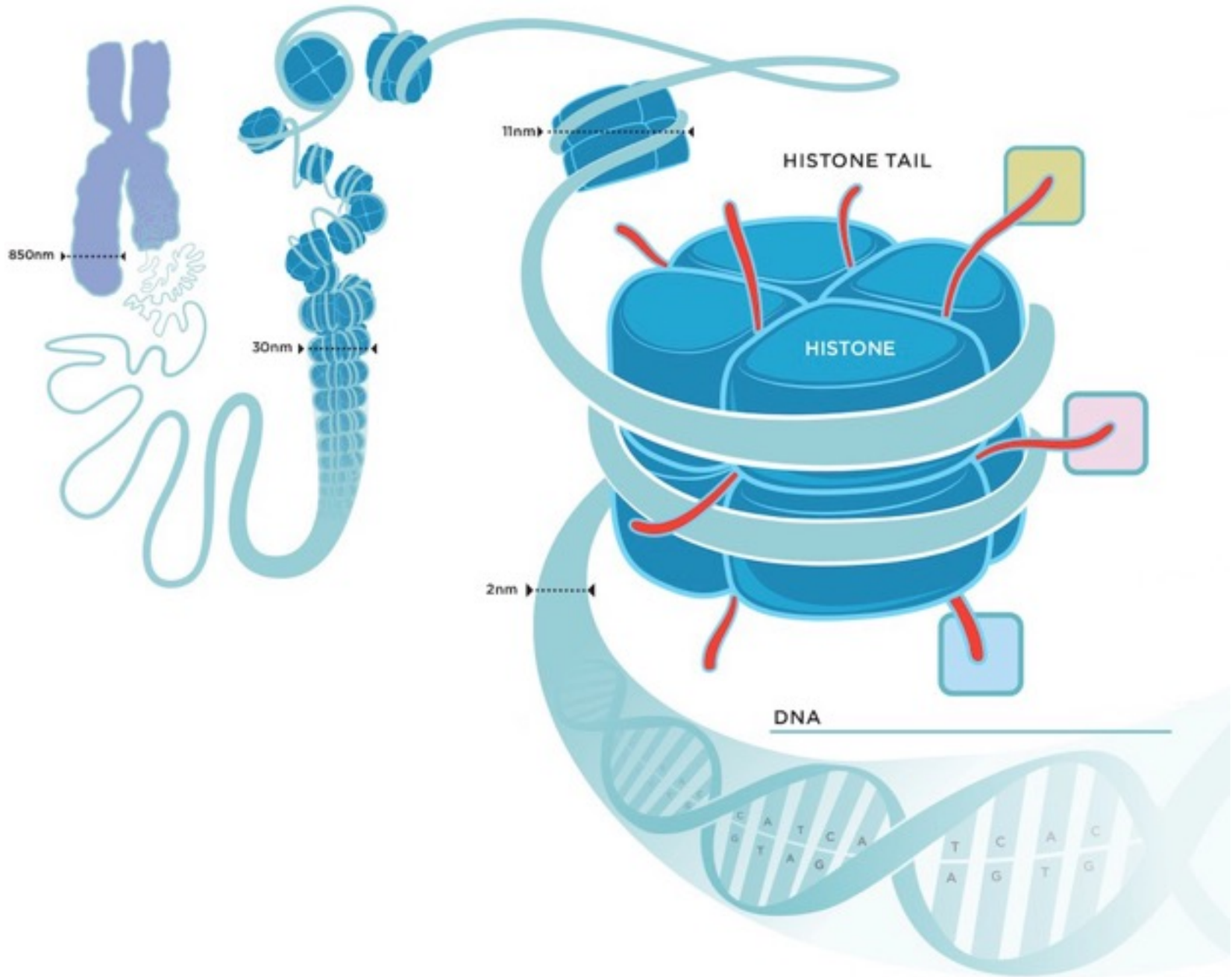
Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2015).
Formation of Chromosomal Domains by Loop Extrusion. bioRxiv.





Level VI: Nucleosome

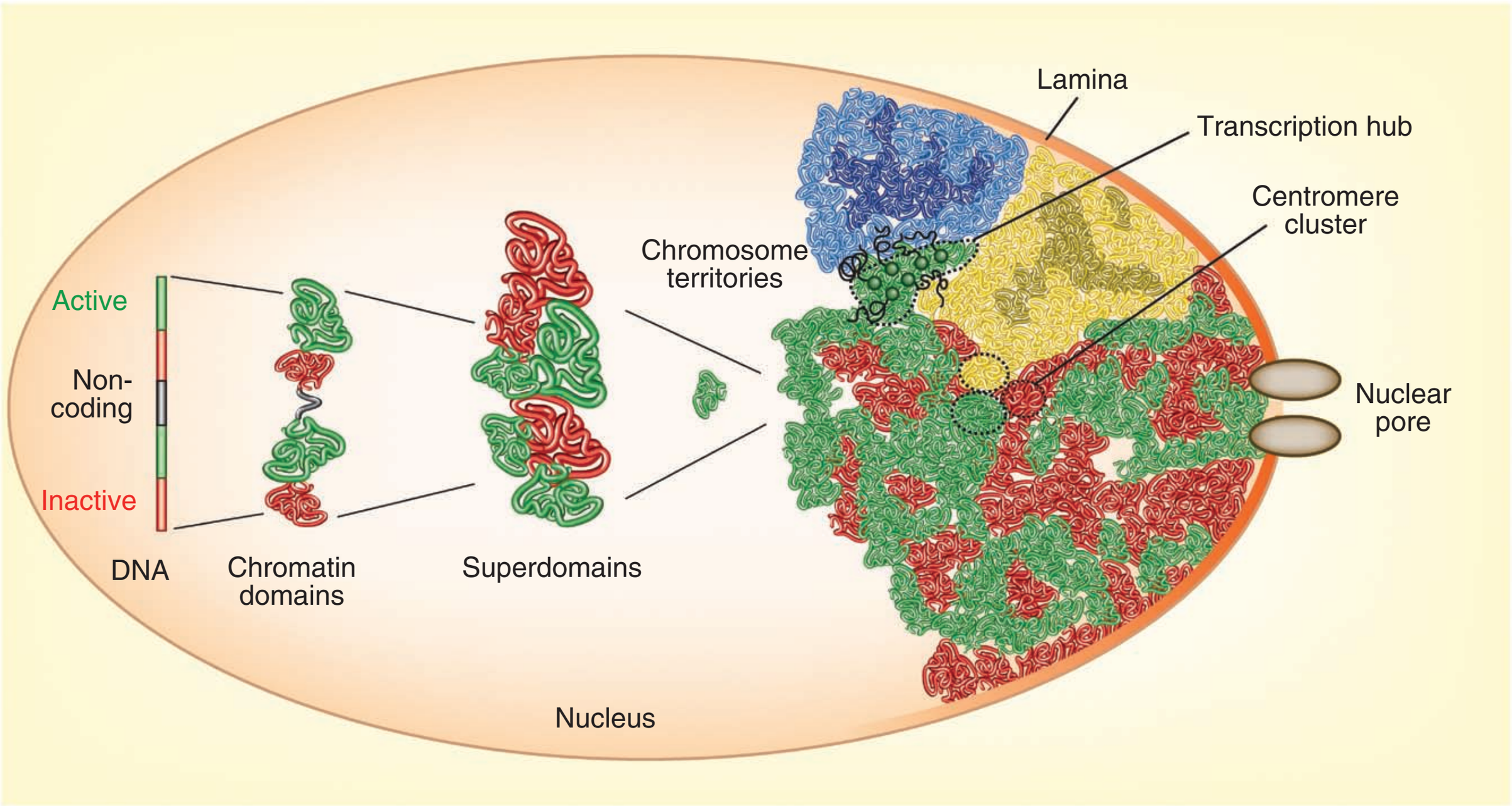
Chromosome **Chromatin fibre** **Nucleosome**



Adapted from Richard E. Ballermann, 2012

Complex genome organization

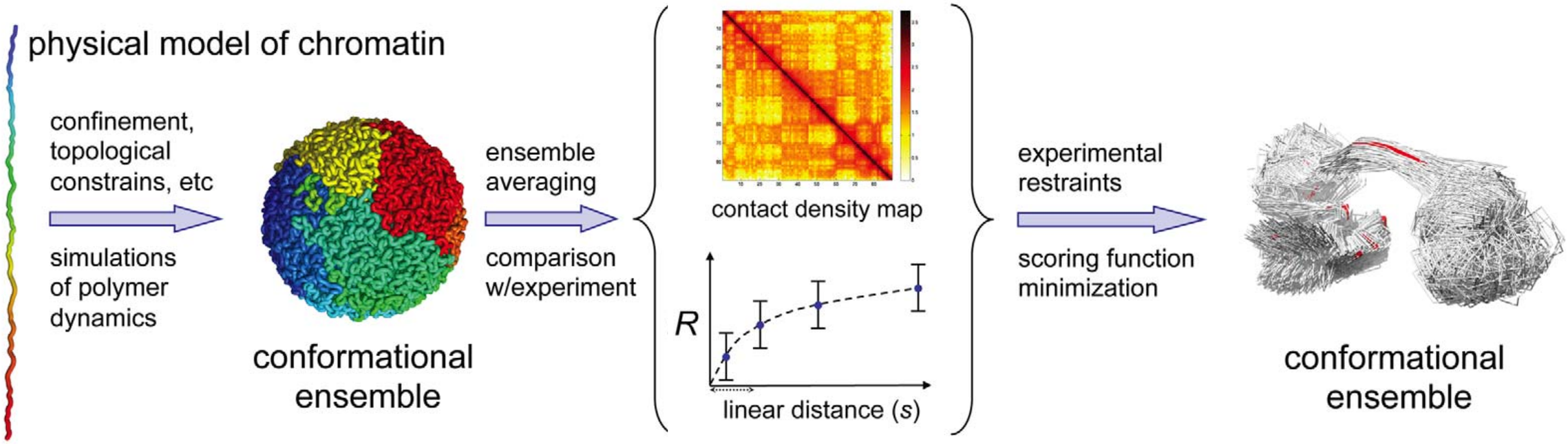
Cavalli, G. & Misteli, T. Functional implications of genome topology. Nat Struct Mol Biol 20, 290–299 (2013).



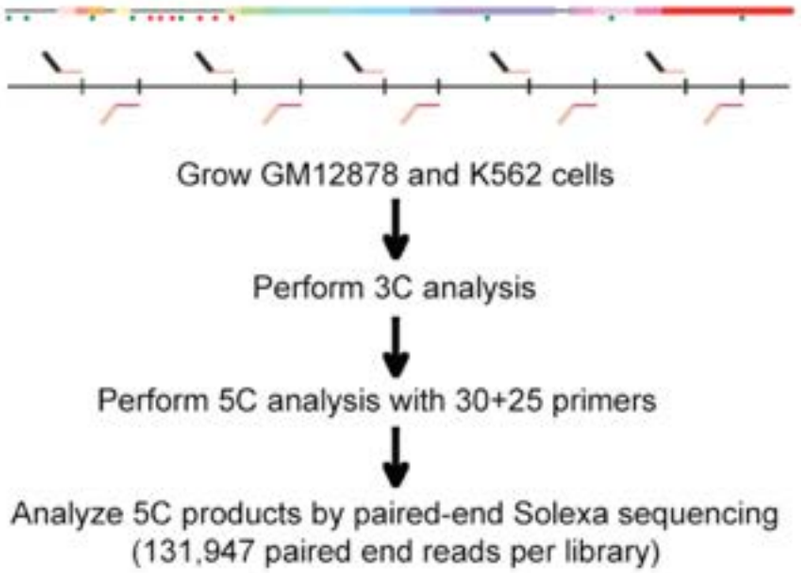
Marina Corral

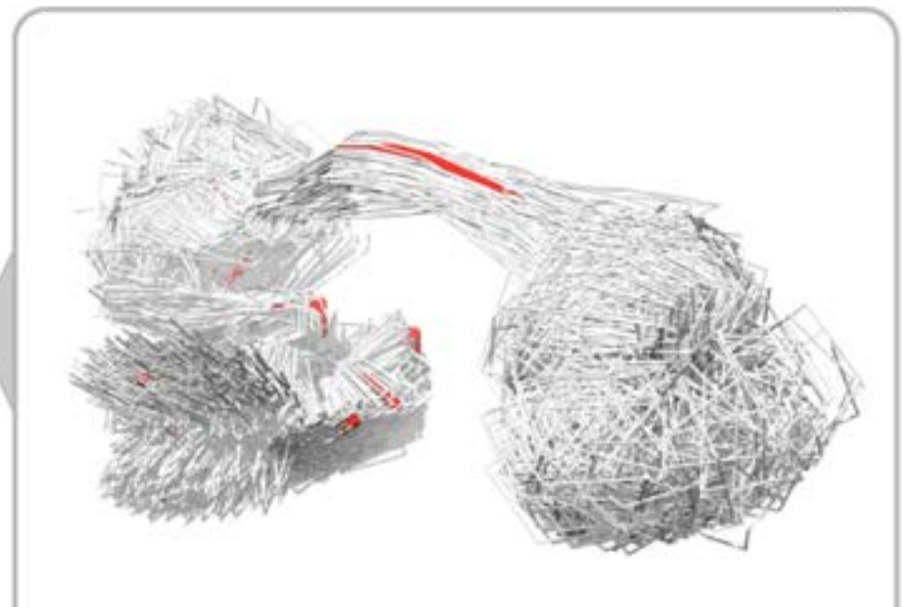
Modeling Genomes

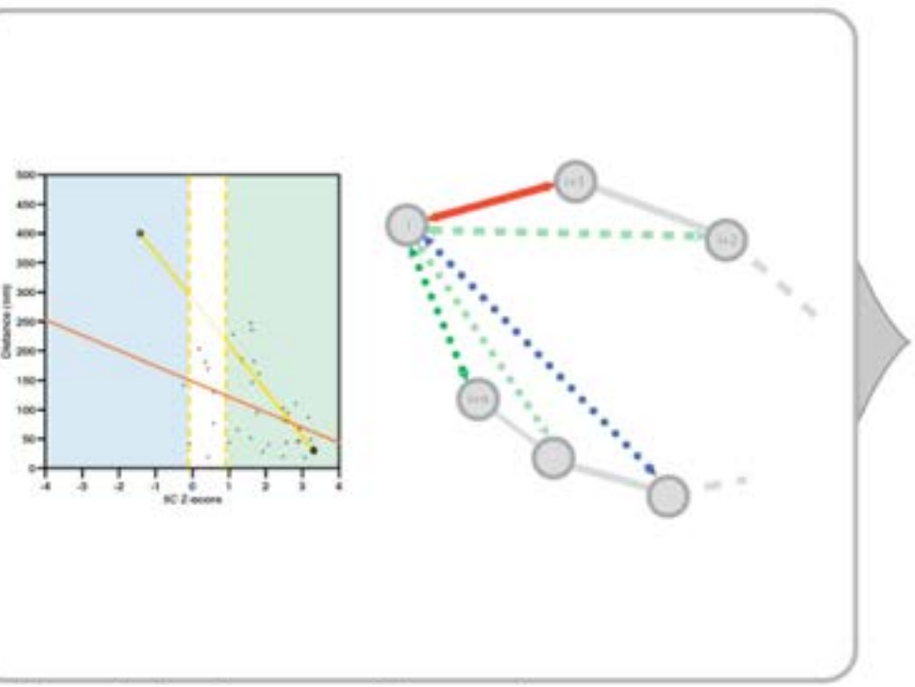
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)

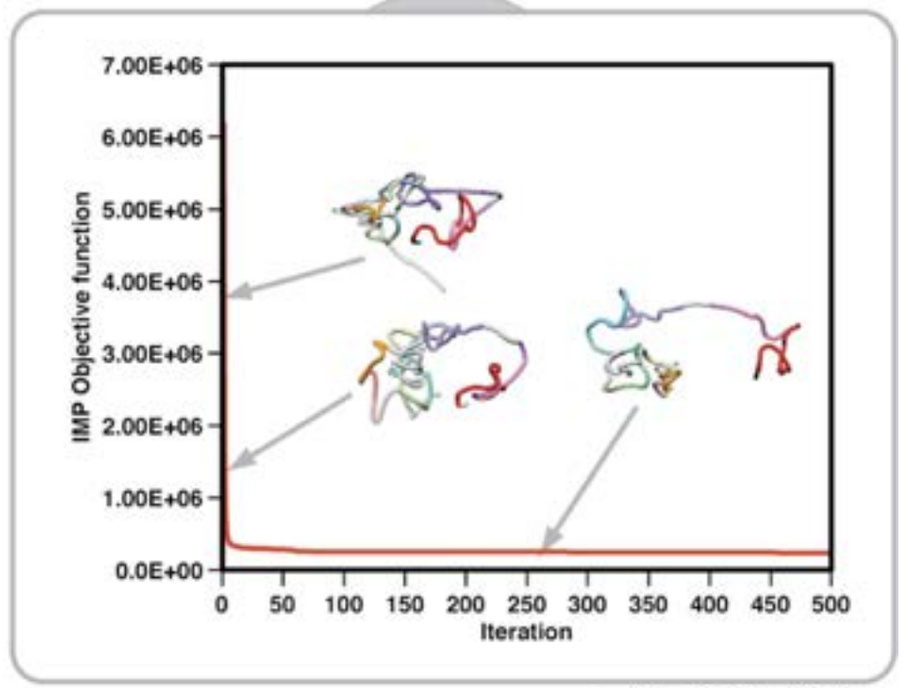


Experiments

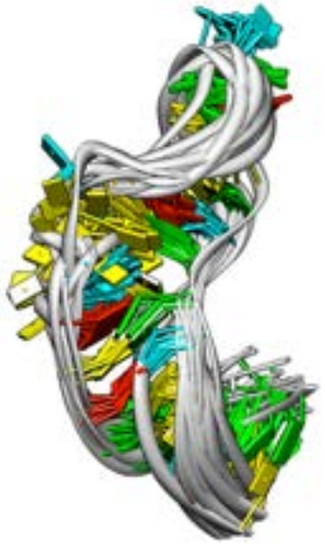


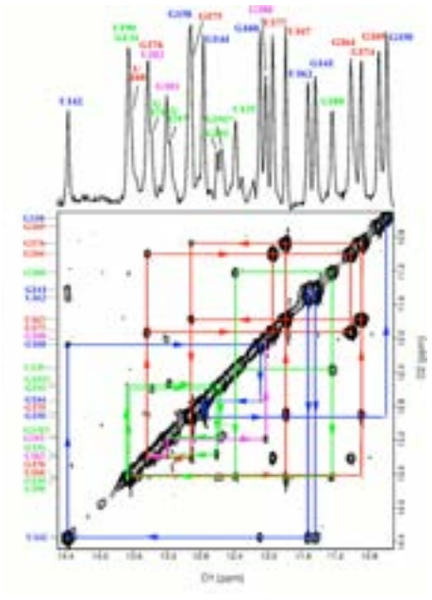






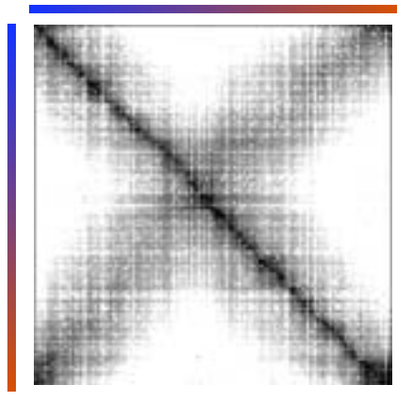
Computation





Biomolecular structure determination

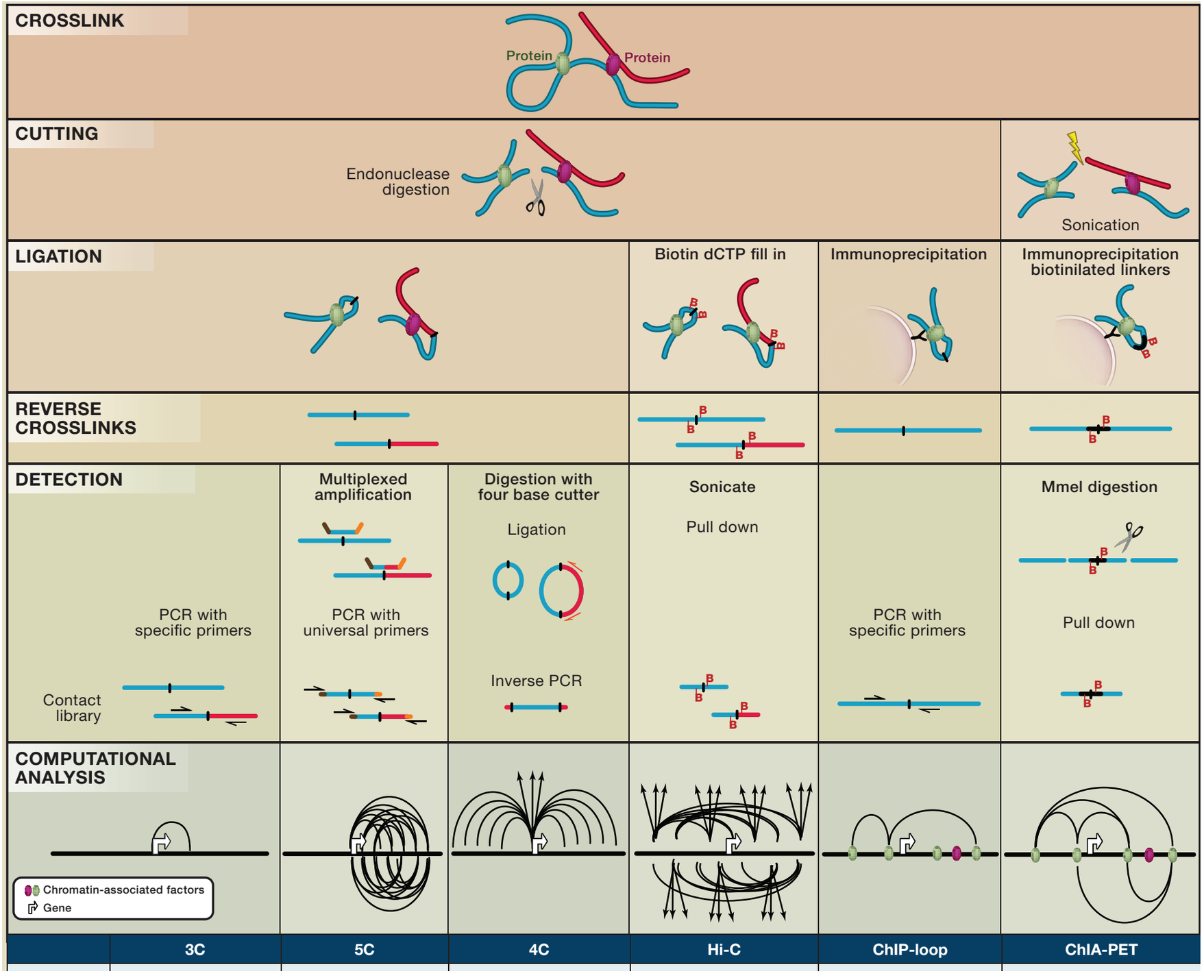
2D-NOESY data



Chromosome structure determination

5C data

Chromosome Conformation Capture



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

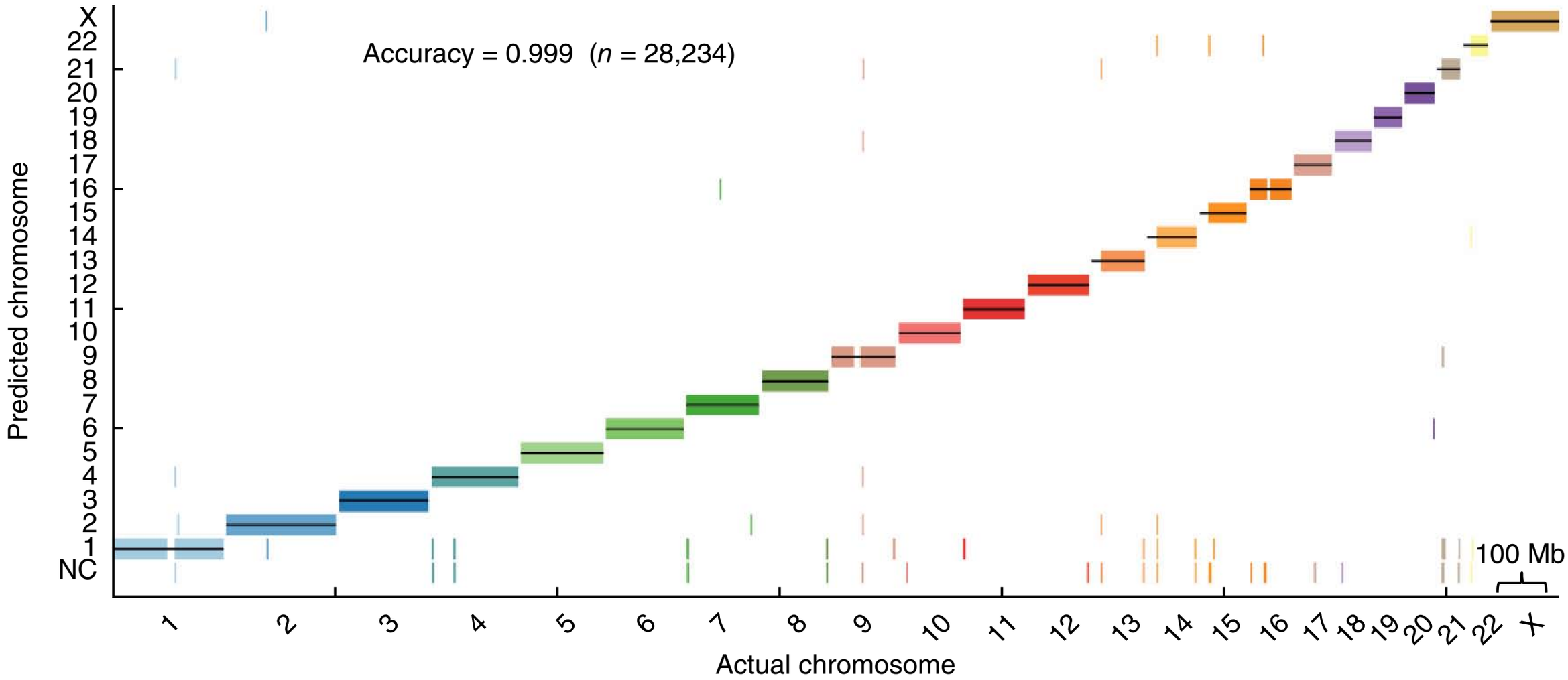
Chromosome Conformation Capture

	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.

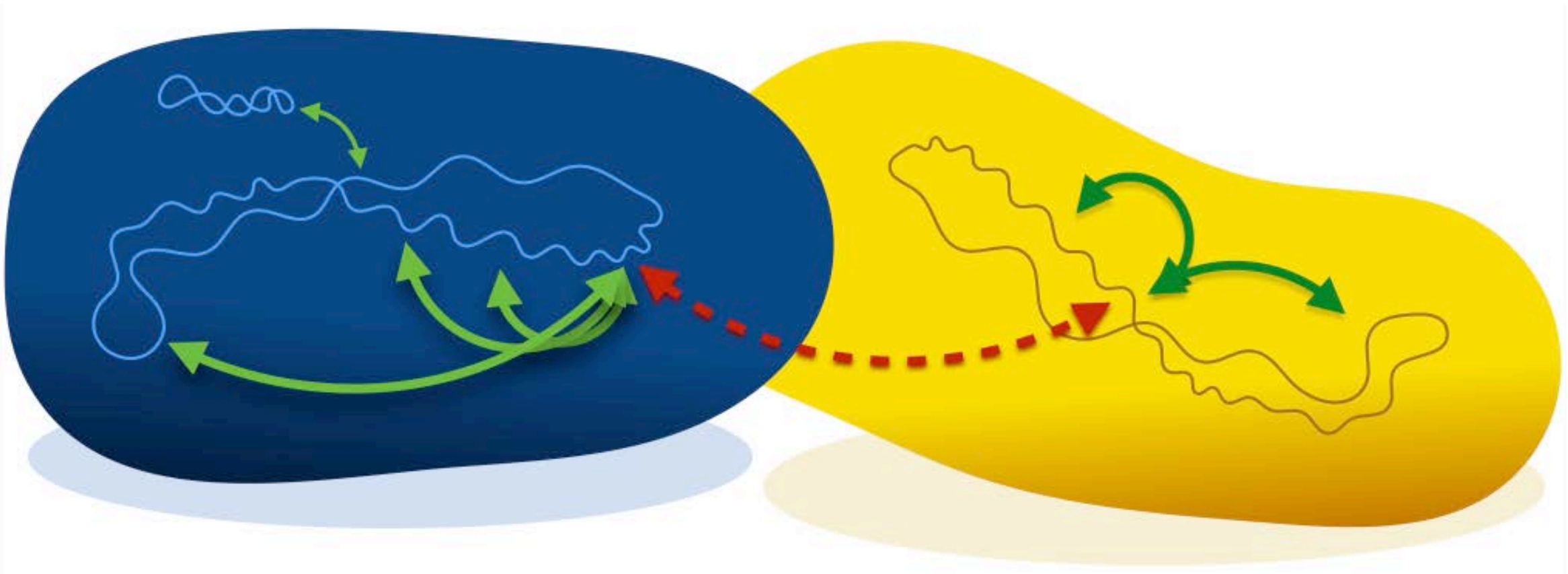
... and one more thing

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.

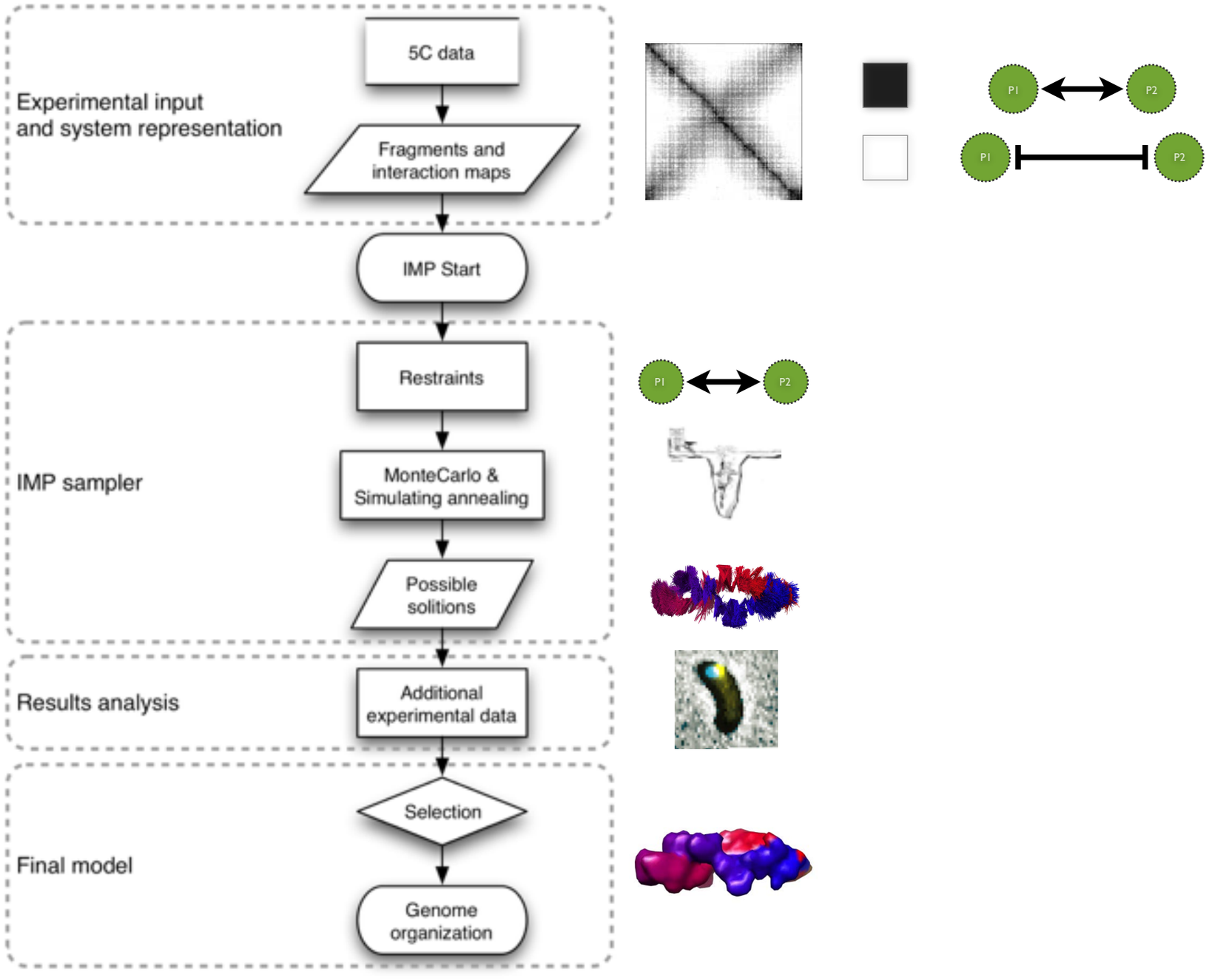
Chromosome Conformation Capture for meta genomics



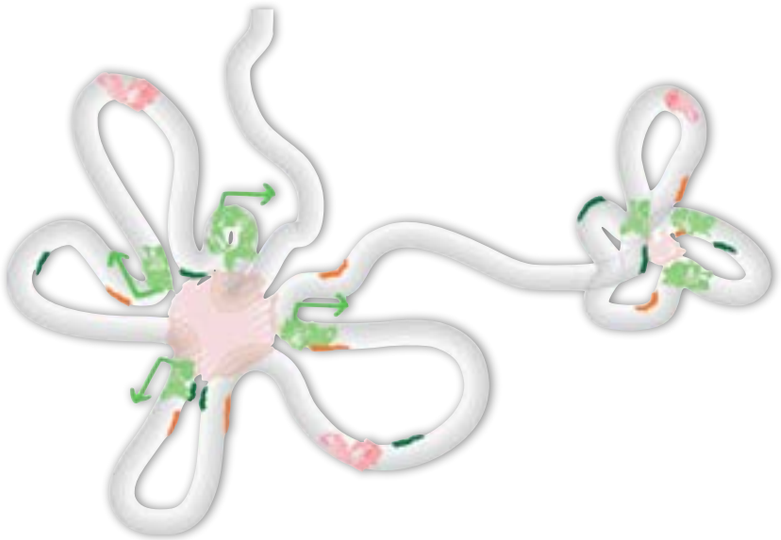
Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., & Darling, A. E. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. doi:10.7287/peerj.preprints.260v1

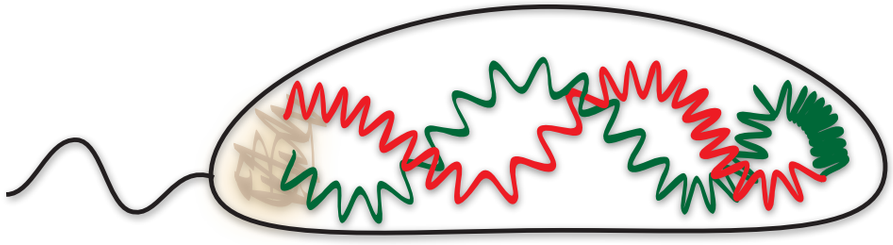
Modeling 3D Genomes

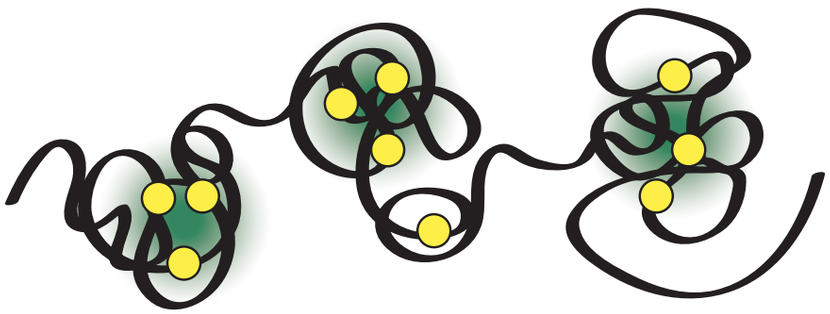
Baù, D. & Marti-Renom, M. A. Methods 58, 300–306 (2012).



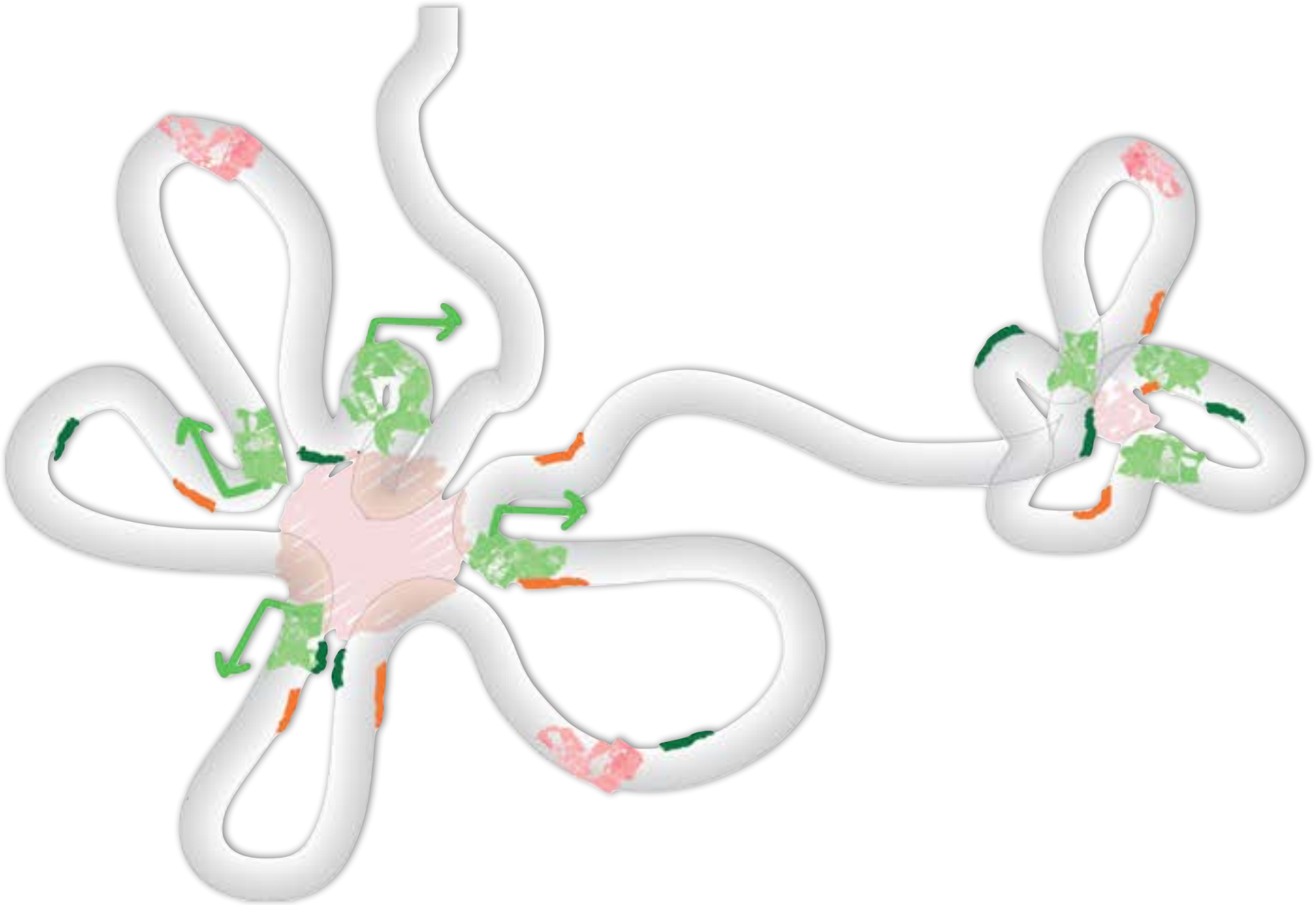
Examples...







Human α-globin domain



Human α-globin domain

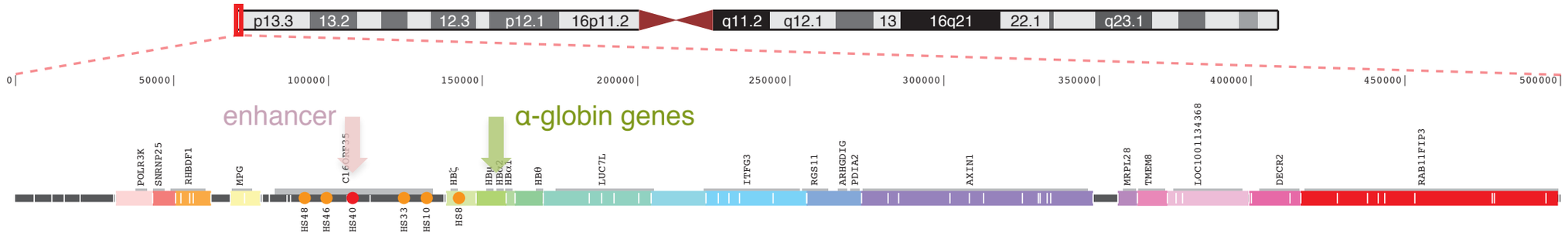
ENm008 genomic structure and environment

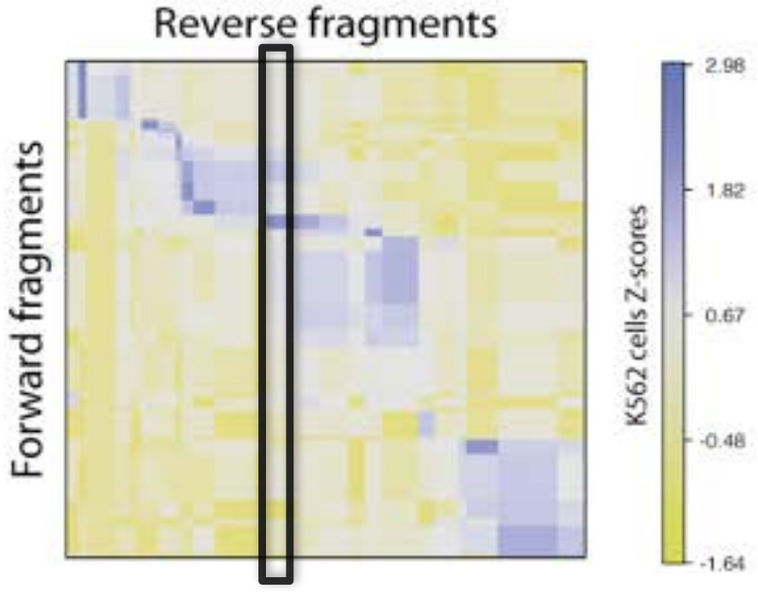
The ENCODE data for ENm008 region was obtained from the UCSC Genome Browser tracks for: RefSeq annotated genes, Affymetrix/CSHL expression data (Gingeras Group at Cold Spring Harbor), Duke/NHGRI DNaseI Hypersensitivity data (Crawford Group at Duke University), and Histone Modifications by Broad Institute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

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

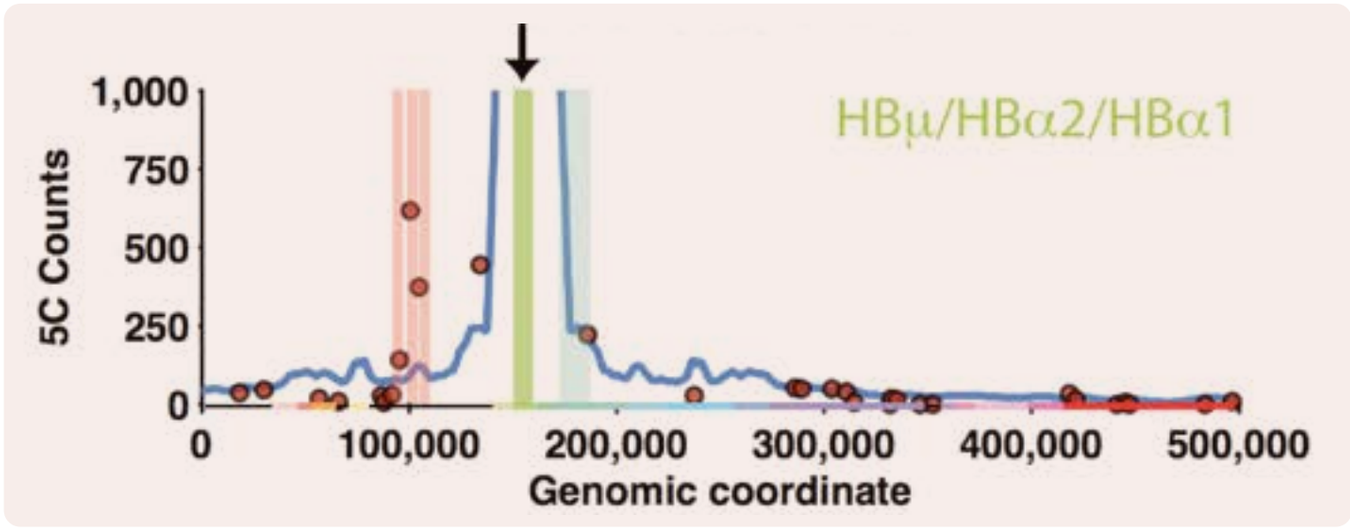
Human α-globin domain

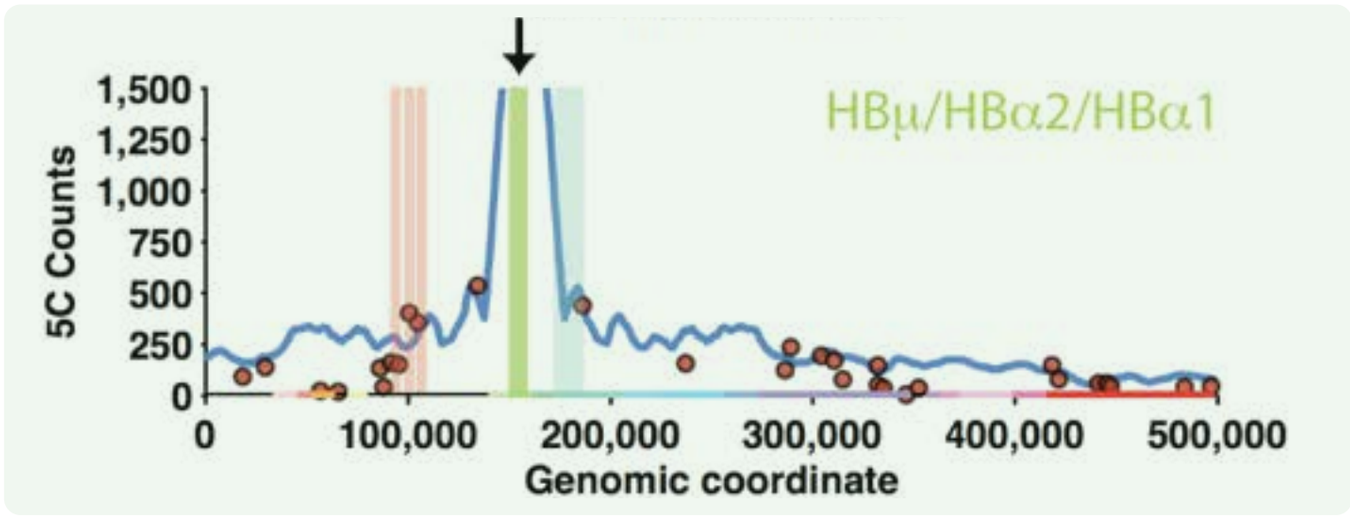
ENm008 genomic structure and environment





K562 cells:
α-globin genes active

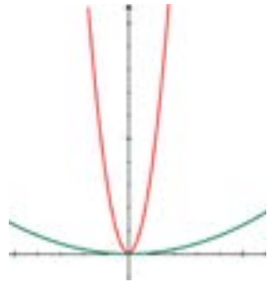




Representation

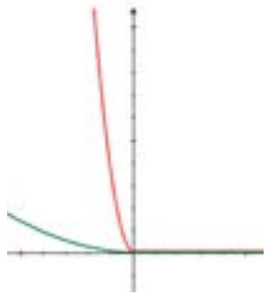
Harmonic

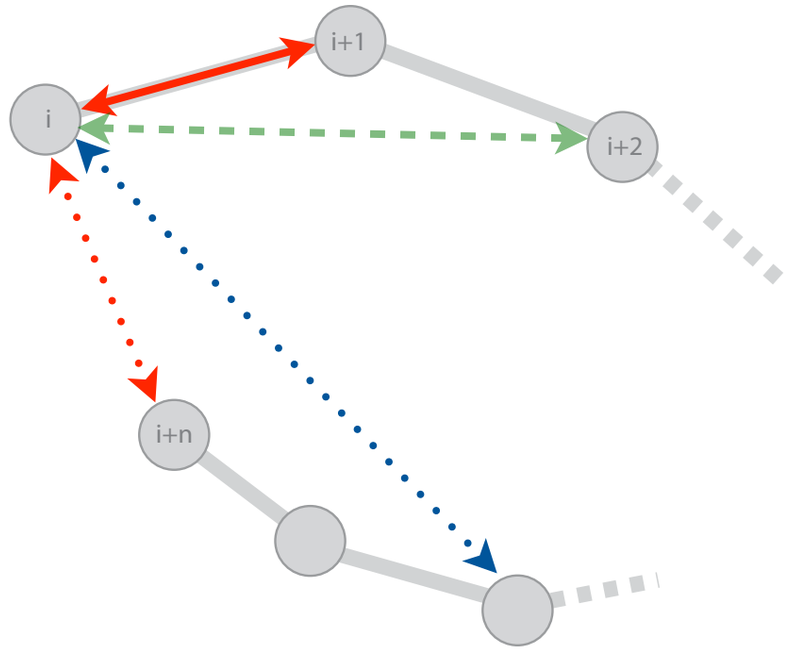
$$H_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2$$



Harmonic Lower Bound

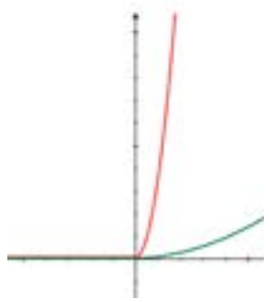
$$\begin{cases} if\ \ d_{i,j} \le d^0_{i,j}; \quad lbH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} > d^0_{i,j}; \quad lbH_{i,j} = 0 \end{cases}$$





Harmonic Upper Bound

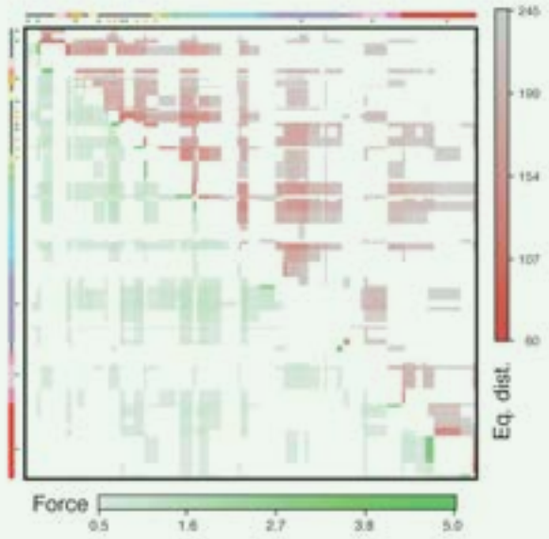
$$\begin{cases} if\ \ d_{i,j} \ge d^0_{i,j}; \quad ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; \quad ubH_{i,j} = 0 \end{cases}$$



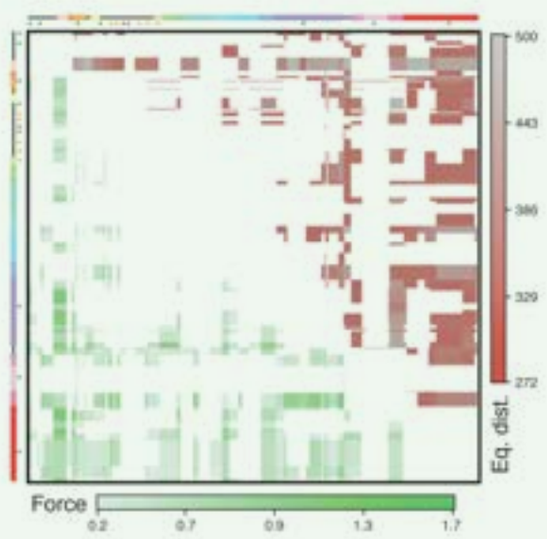
Scoring

GM12878

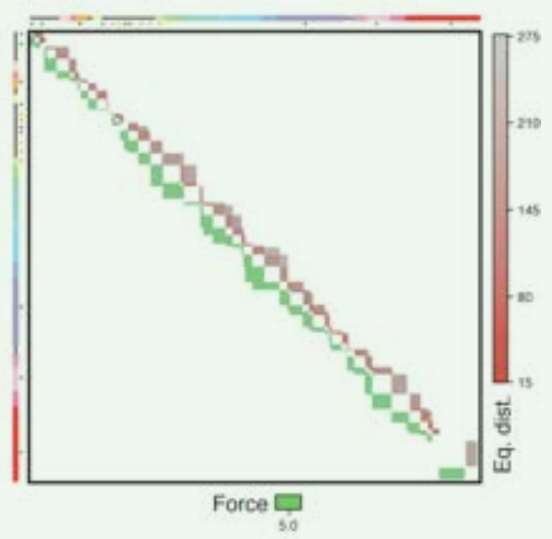
70 fragments
1,520 restraints



Harmonic



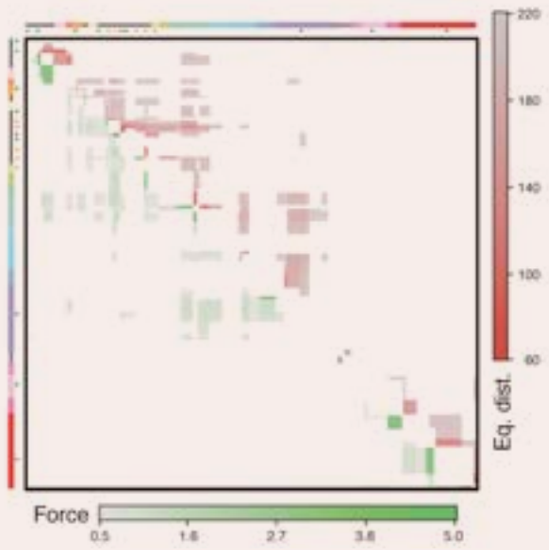
Harmonic Lower Bound

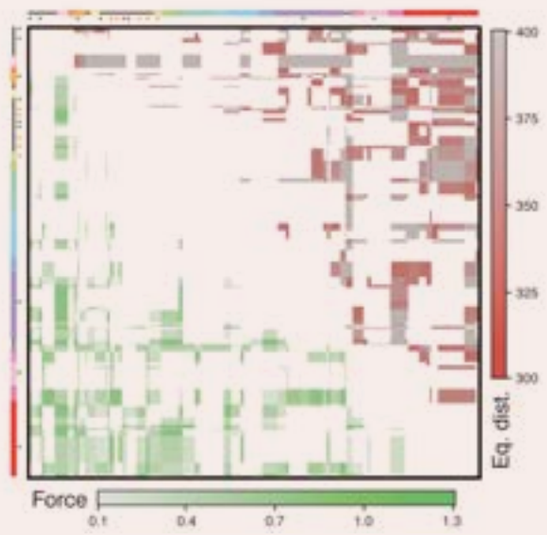


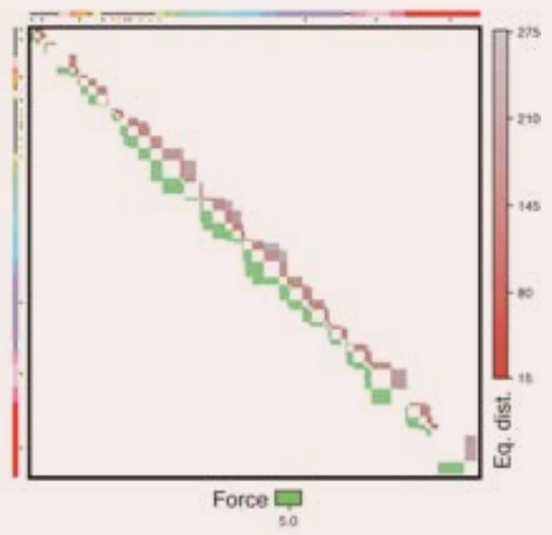
Harmonic Upper Bound

K562

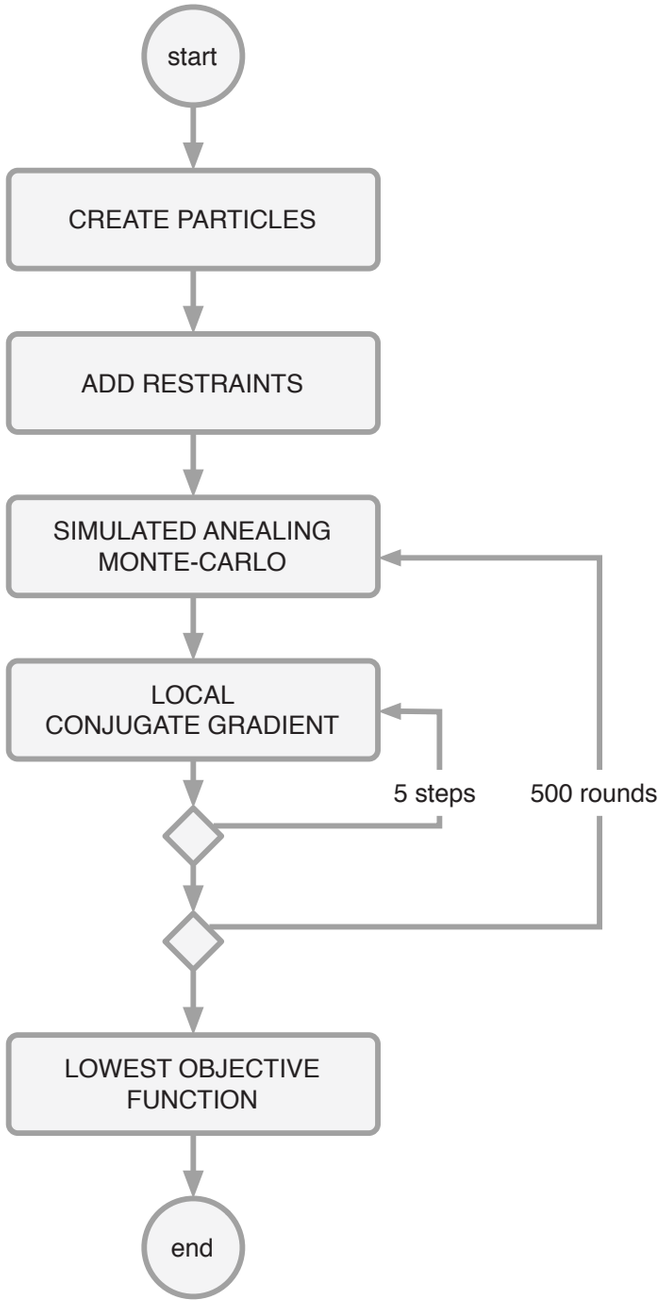
70 fragments
1,049 restraints

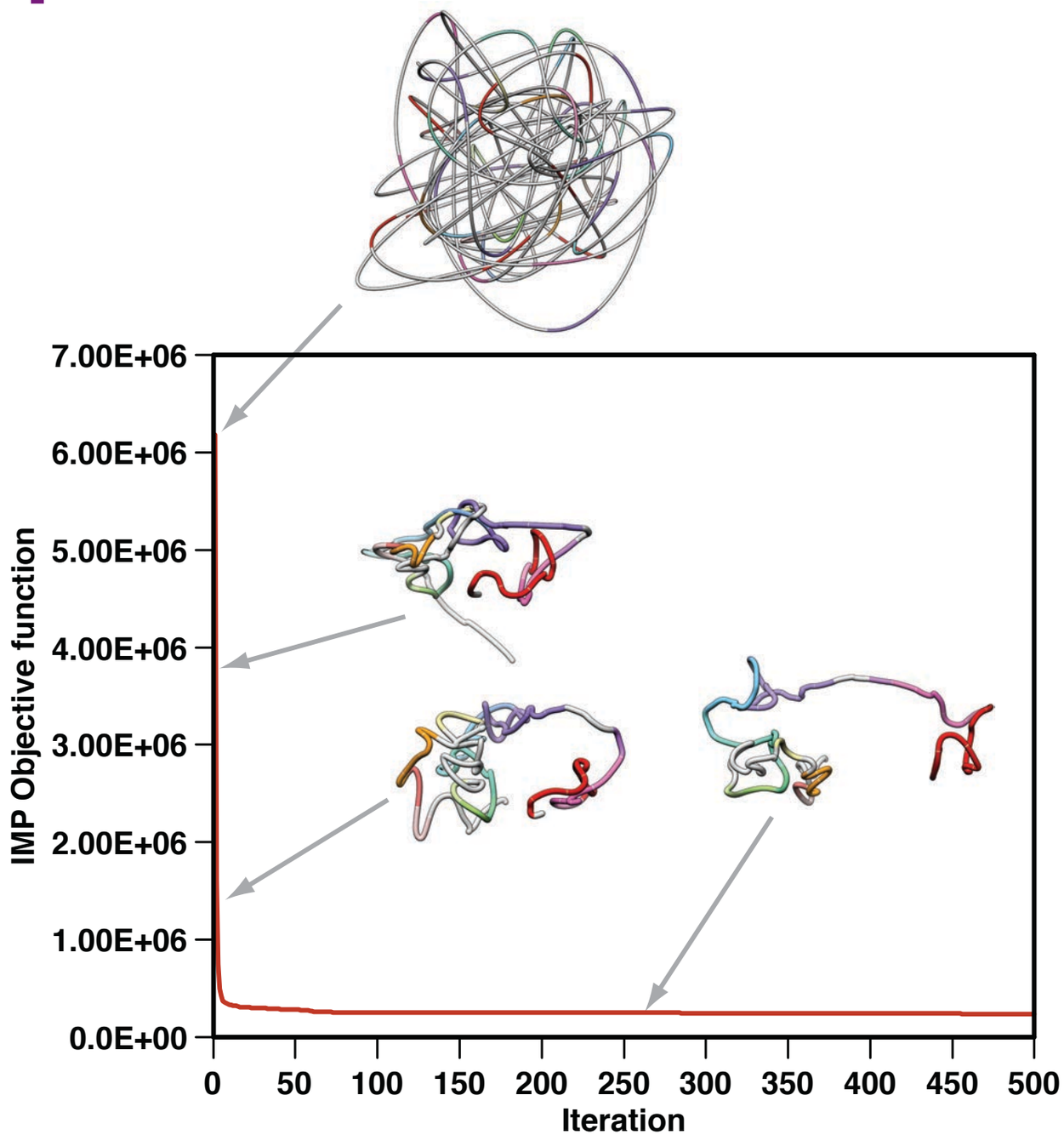




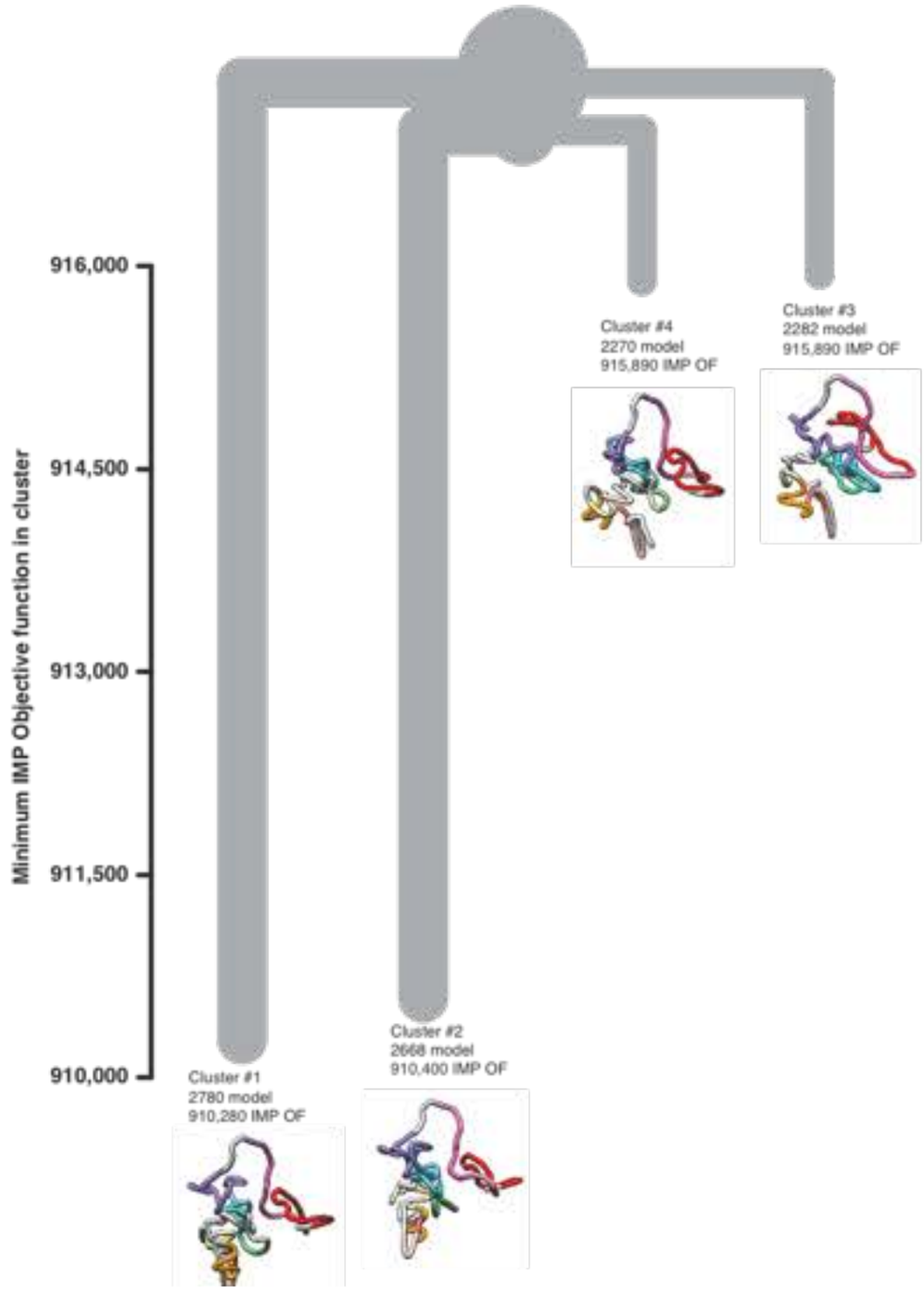


Optimization



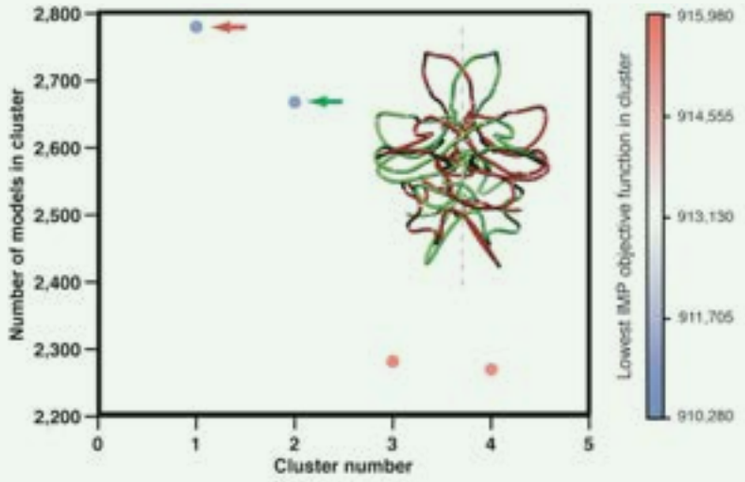


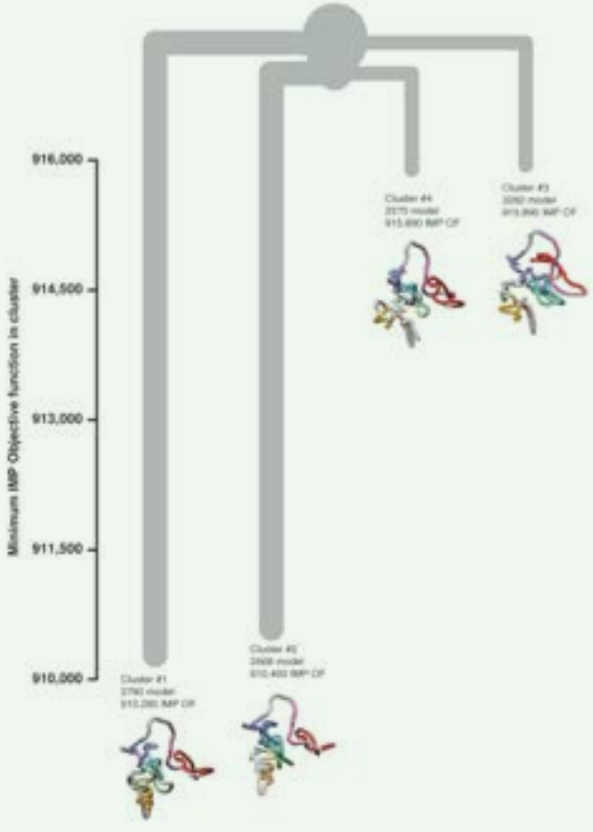
Clustering


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

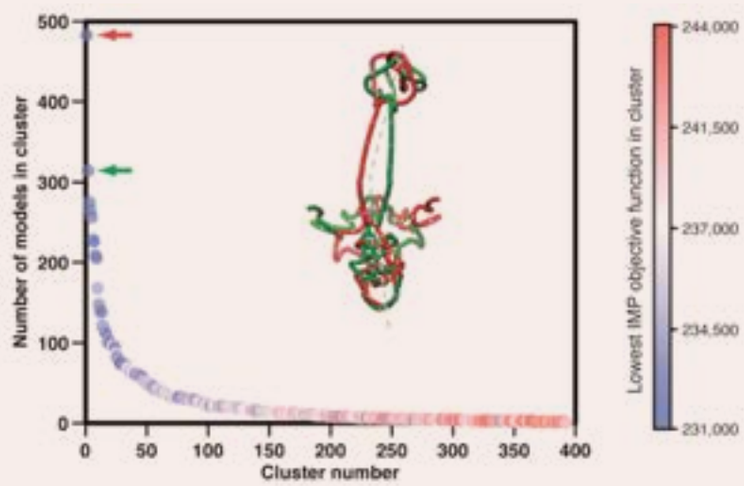
Not just one solution

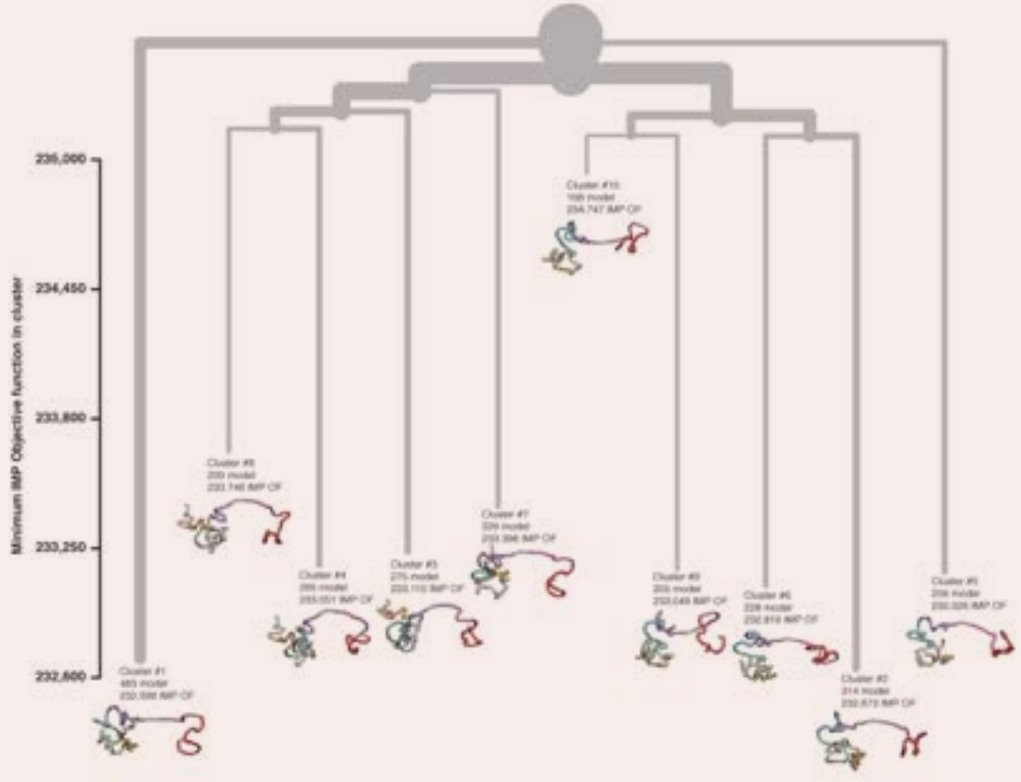
GM12878





K562

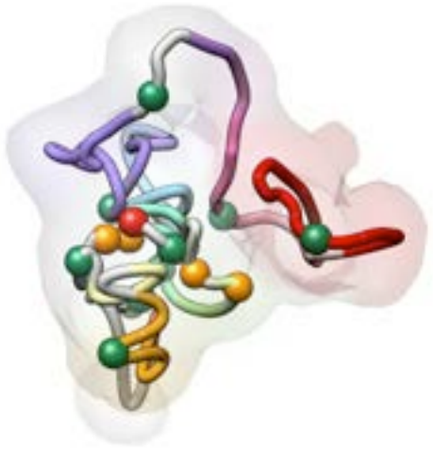


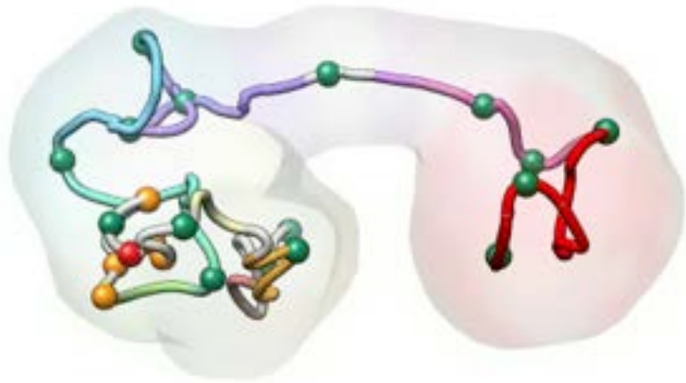


Reg[illegible]ents

GM12878

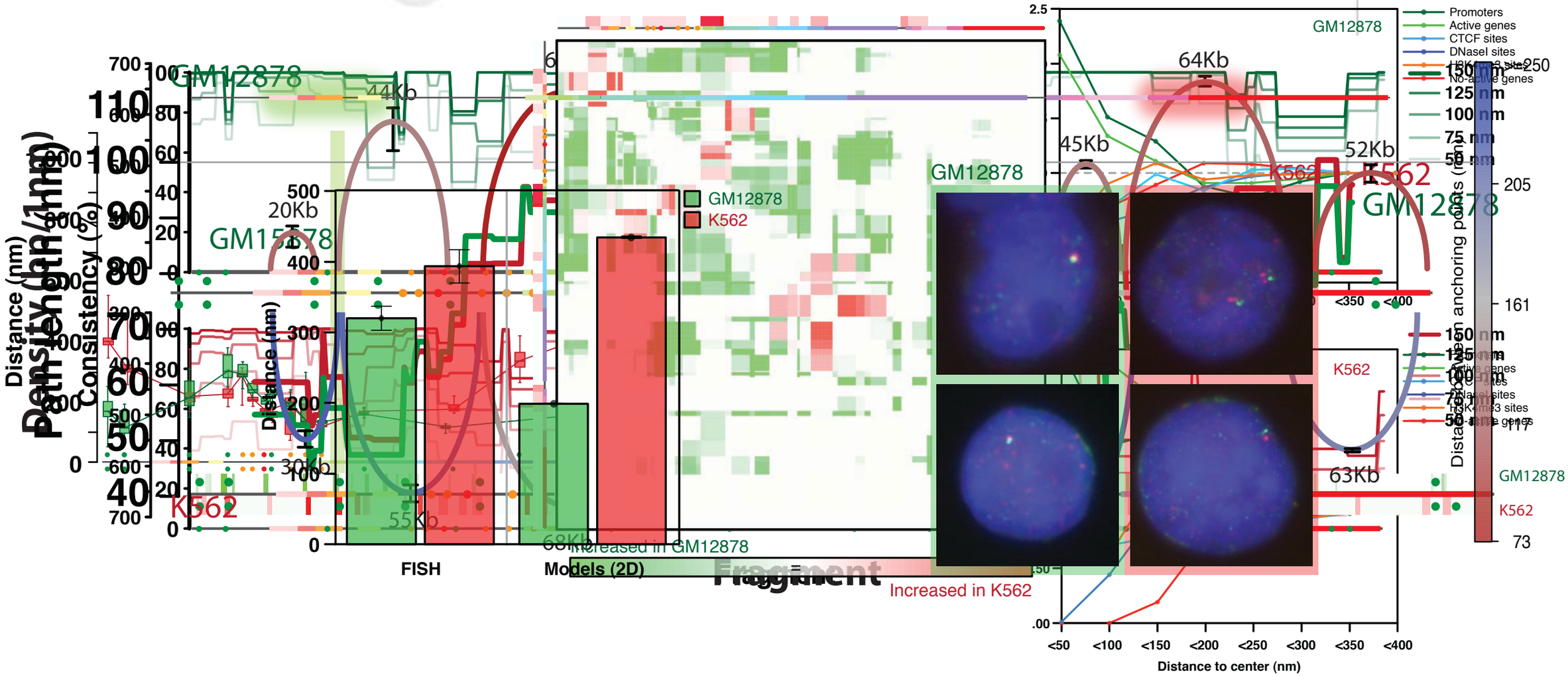
Cluster #1
2780 model



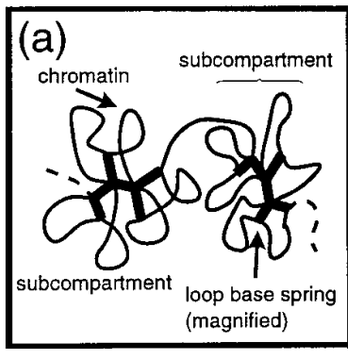


K562

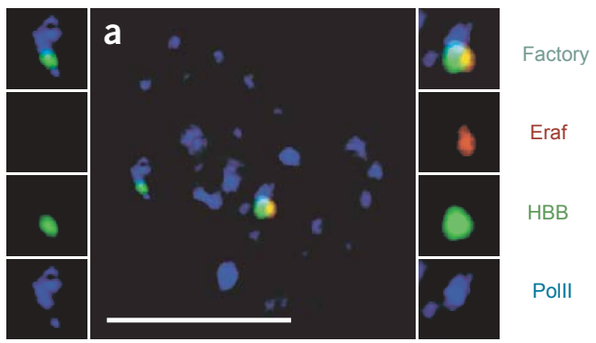
Cluster #2
314 model



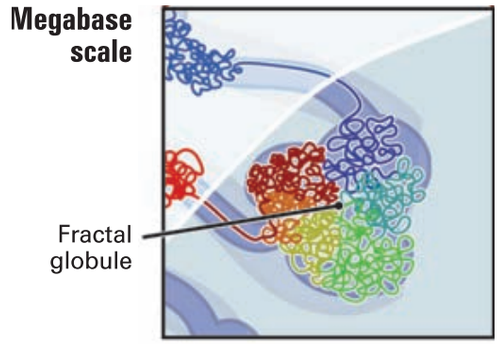
The "Chromatin Globule" model



Münkel et al. JMB (1999)



Osborne et al. Nat Genet (2004)

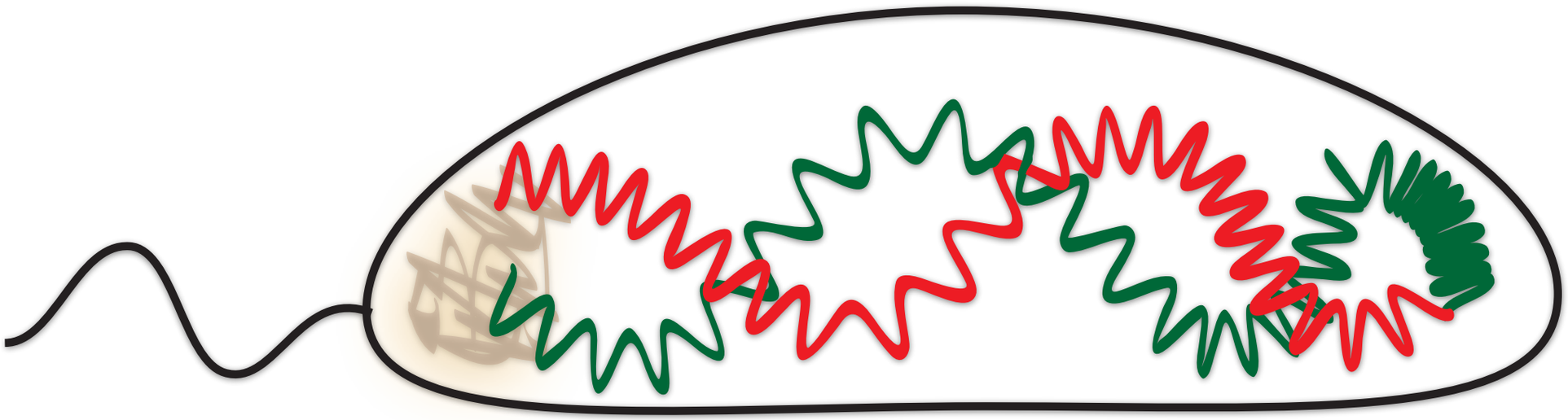


Lieberman-Aiden et al. Science (2009)

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

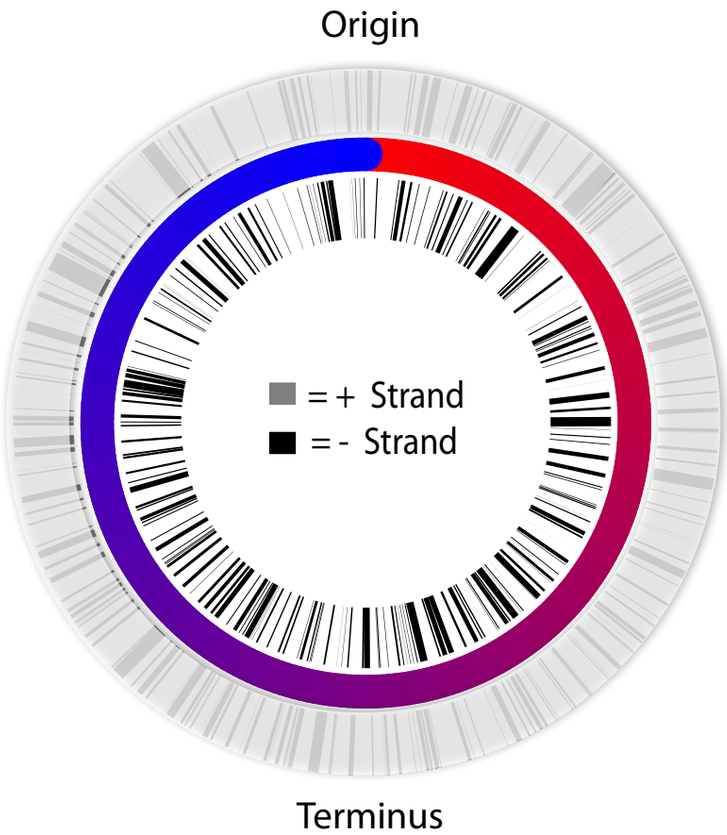
A. Sanyal *et al.* Current Opinion in Cell Biology (2011) 23:325–33.

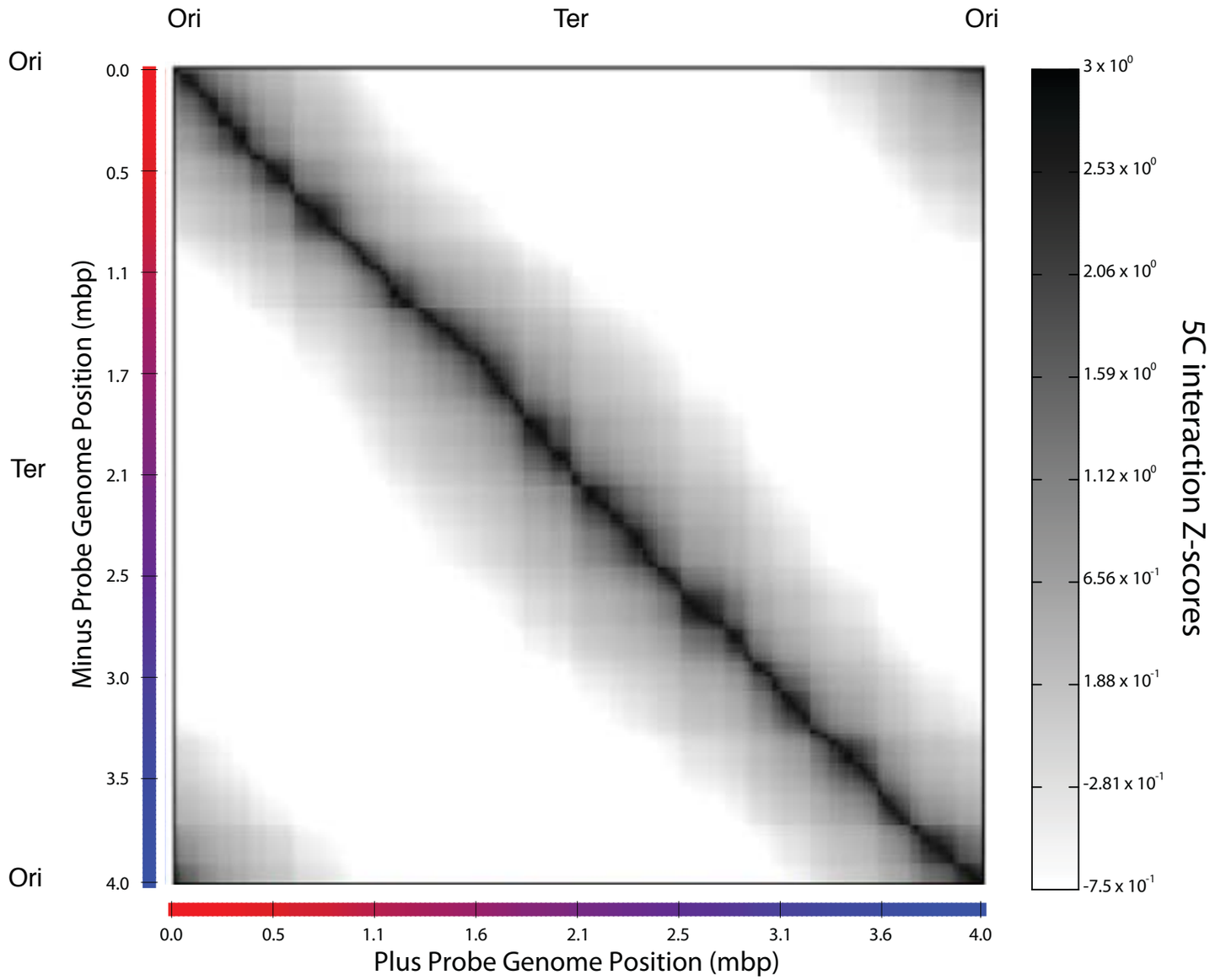
Caulobacter crescentus genome



The 3D architecture of Caulobacter Crescentus

4,016,942 bp & 3,767 genes





169 5C primers on + strand

170 5C primers on – strand

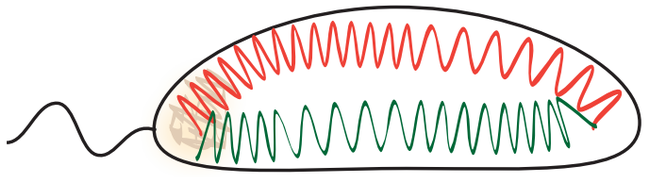
28,730 chromatin interactions

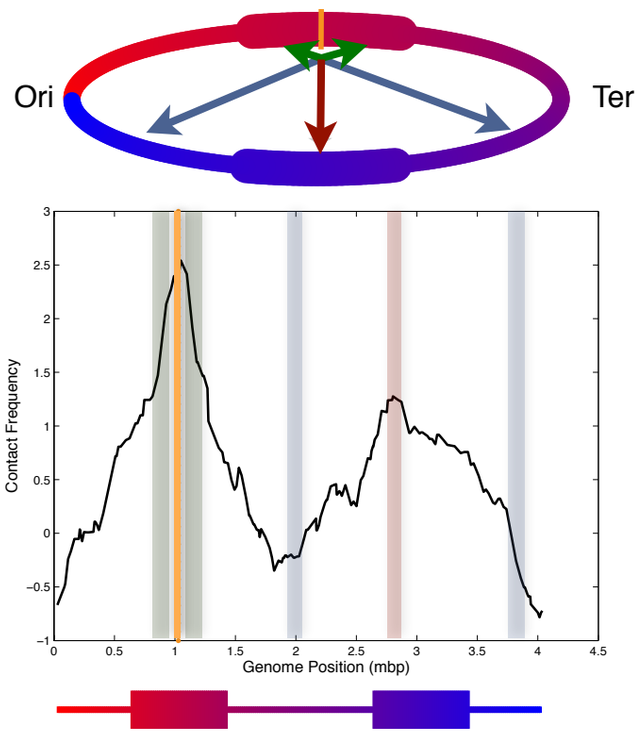
~13Kb

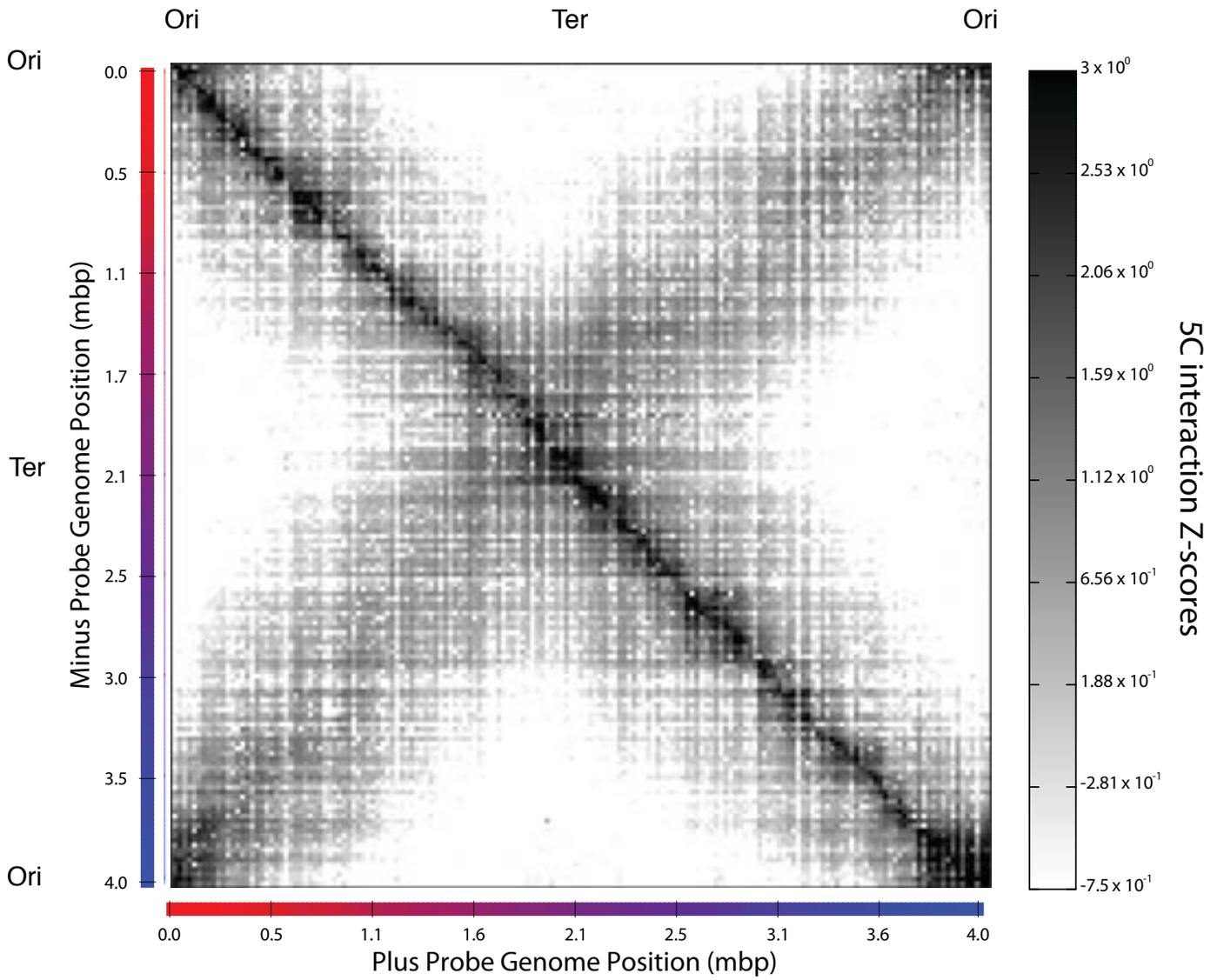
cnag CRG Centre for Genomic Regulation

5C interaction matrix

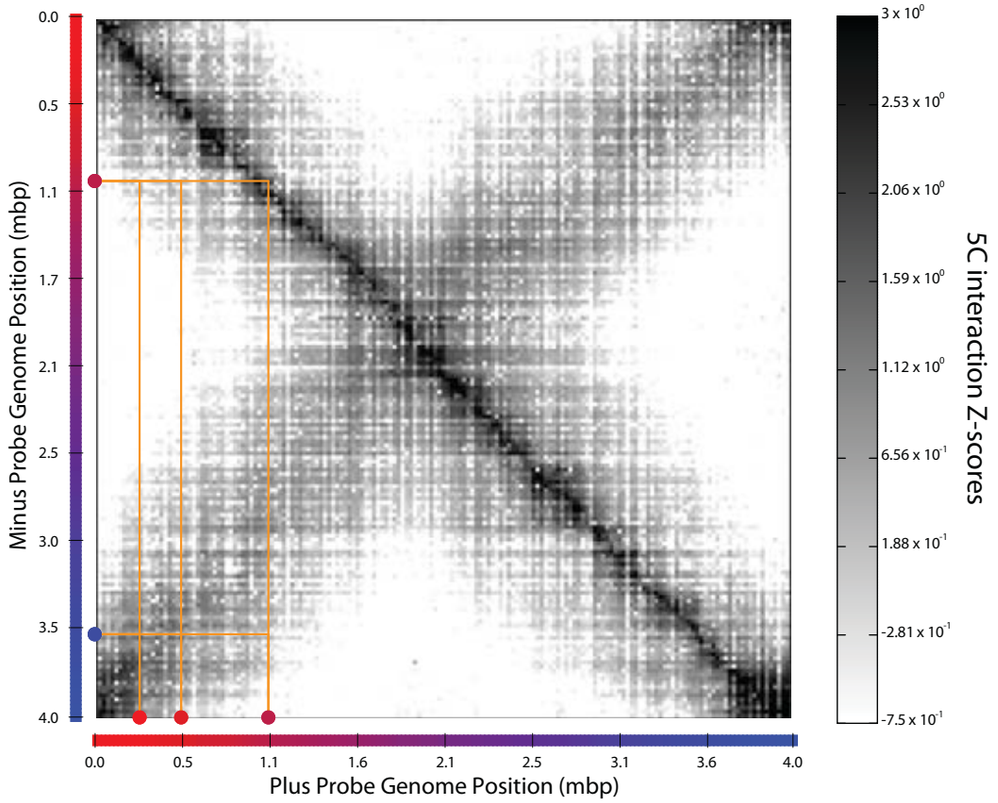
ELLIPSOID for Caulobacter cresentus



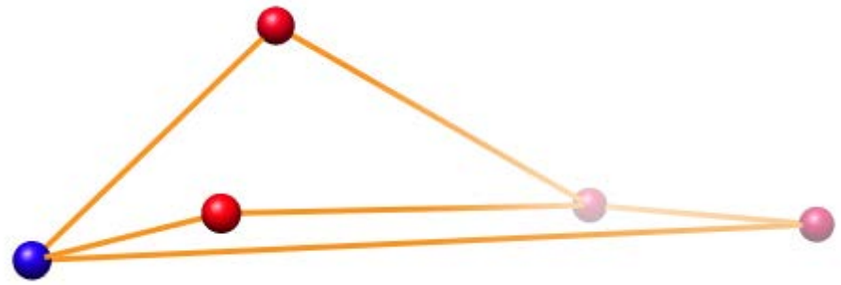


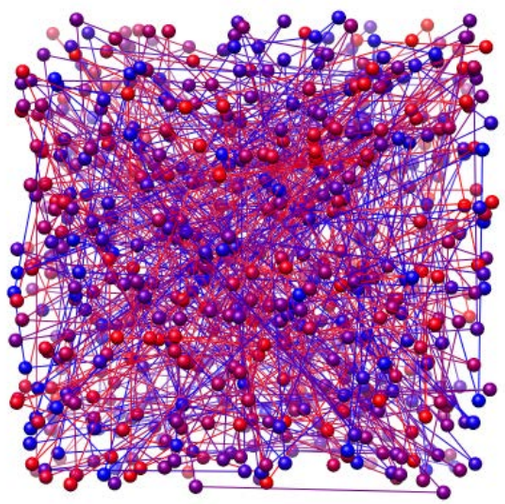


3D model building with the 5C + IMP approach



339 mers



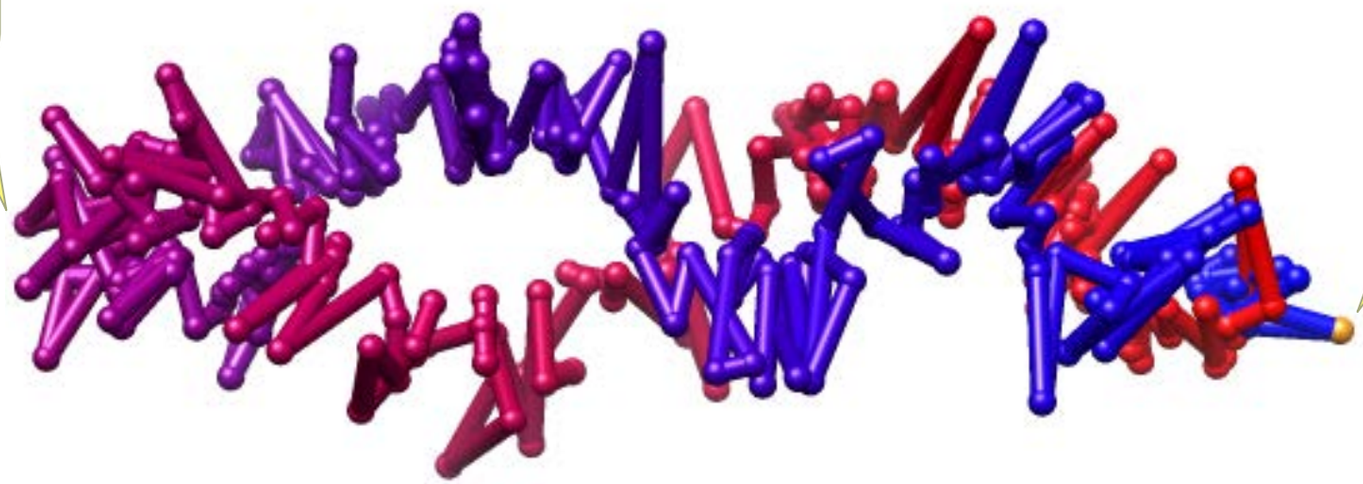


Genome organization in Caulobacter crescentus

Arms are helical

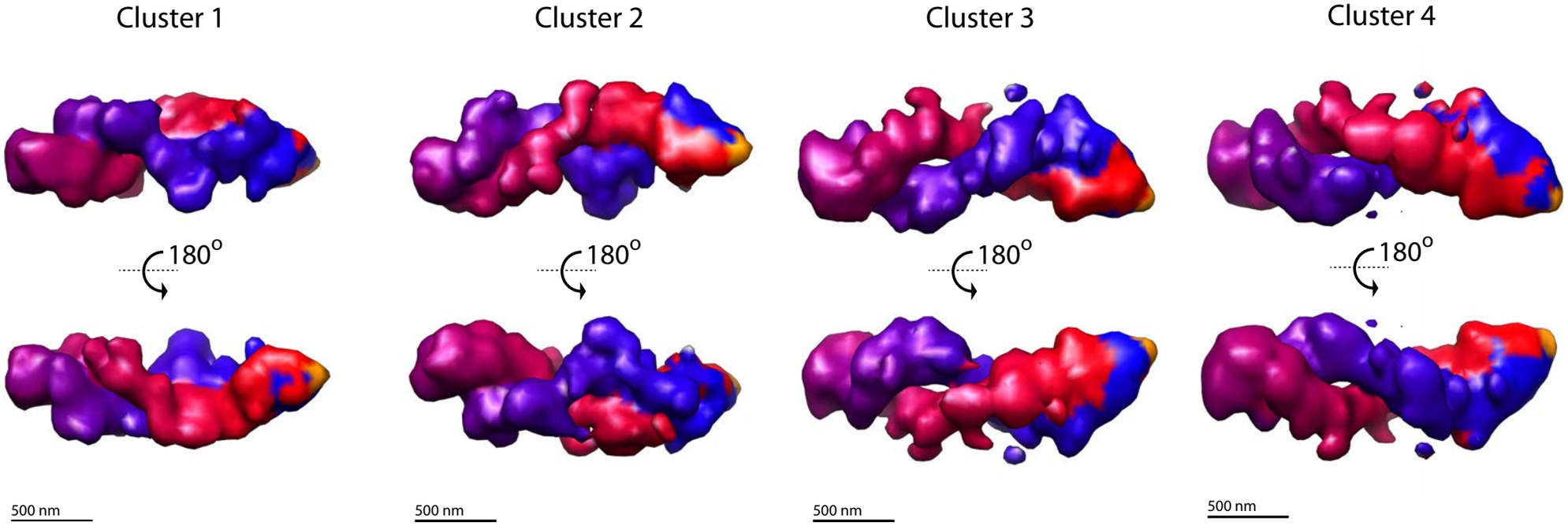
Resolution

dif site 47±17Kb from Ter



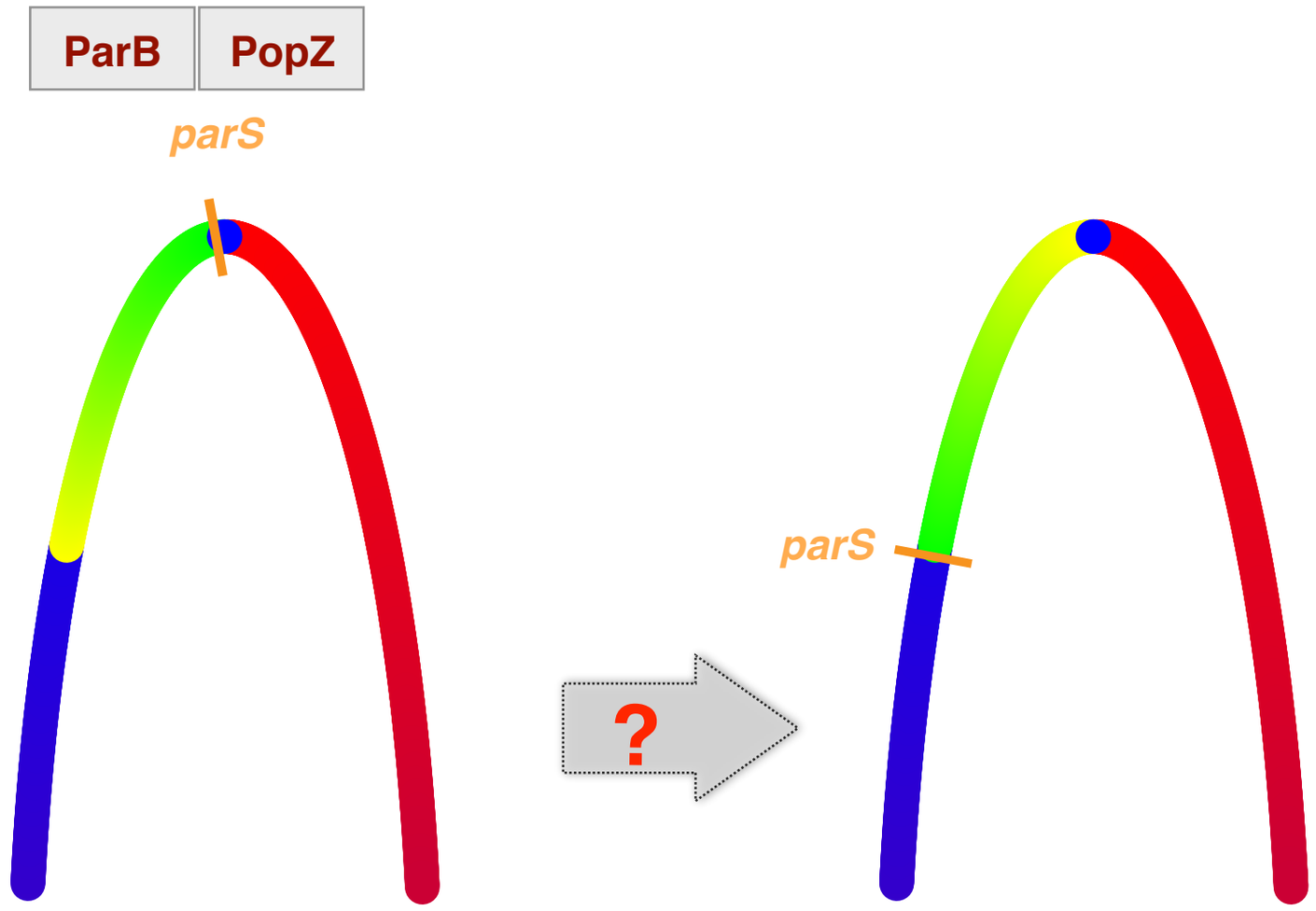
Centromer-like

parS sites 25±17Kb from Ori



MIRRORS!

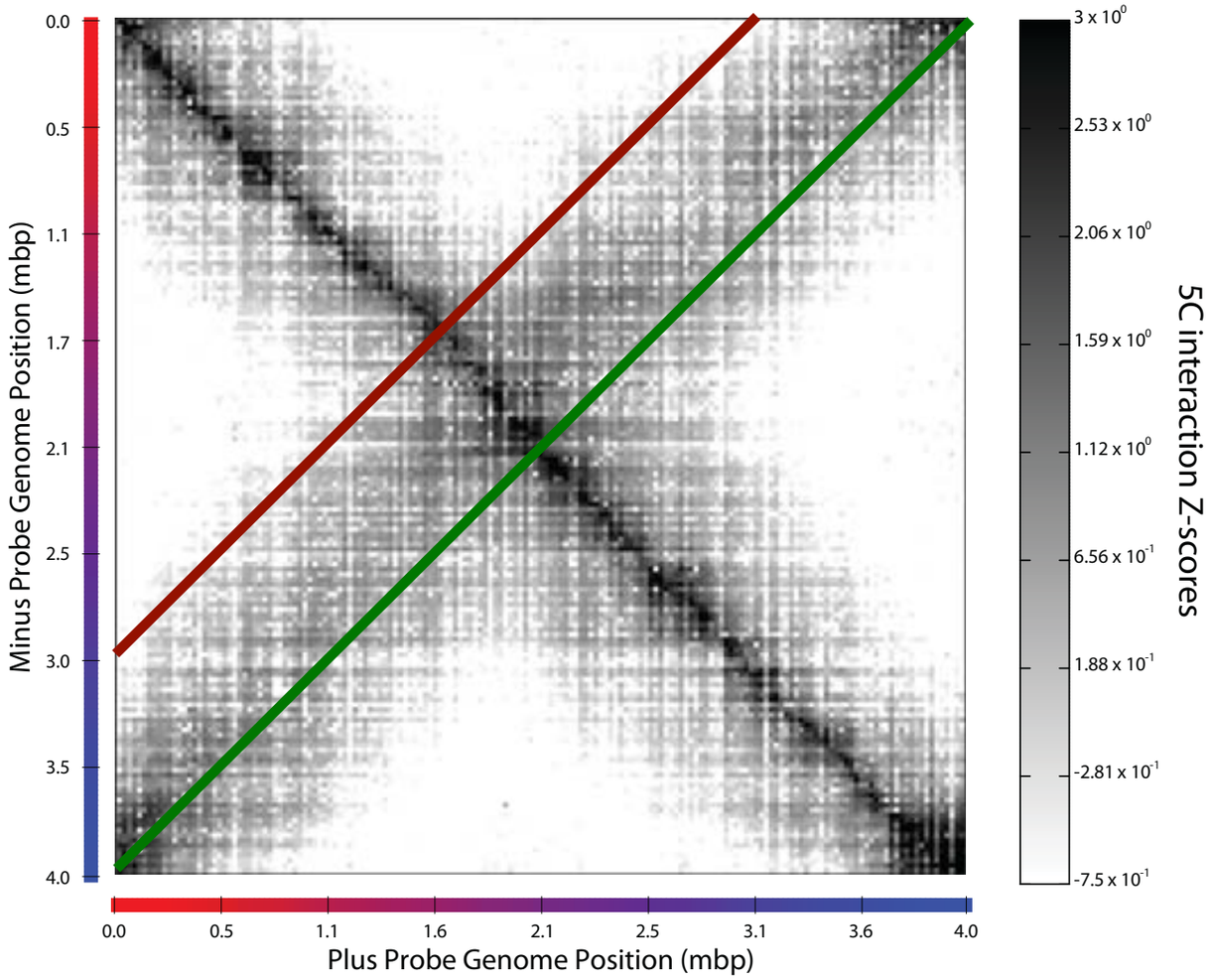
Moving the parS sites 400 Kb away from Ori

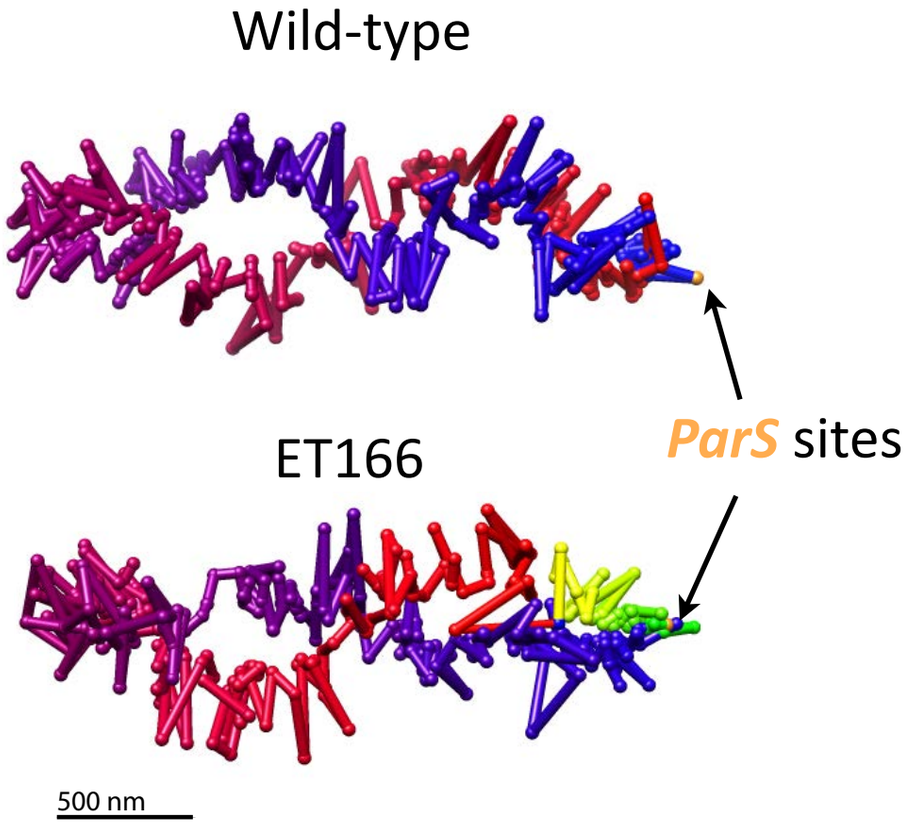


Wild-type

ET166

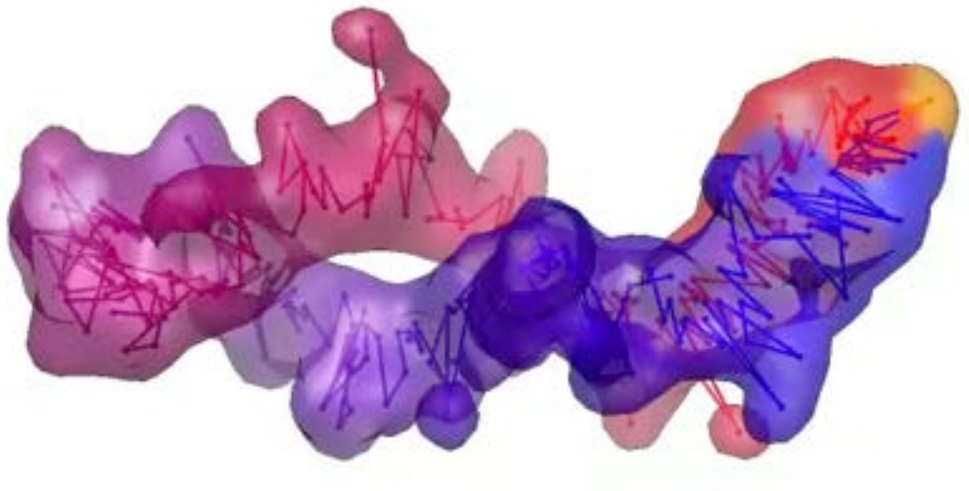
Moving the parS sites results in whole genome rotation!

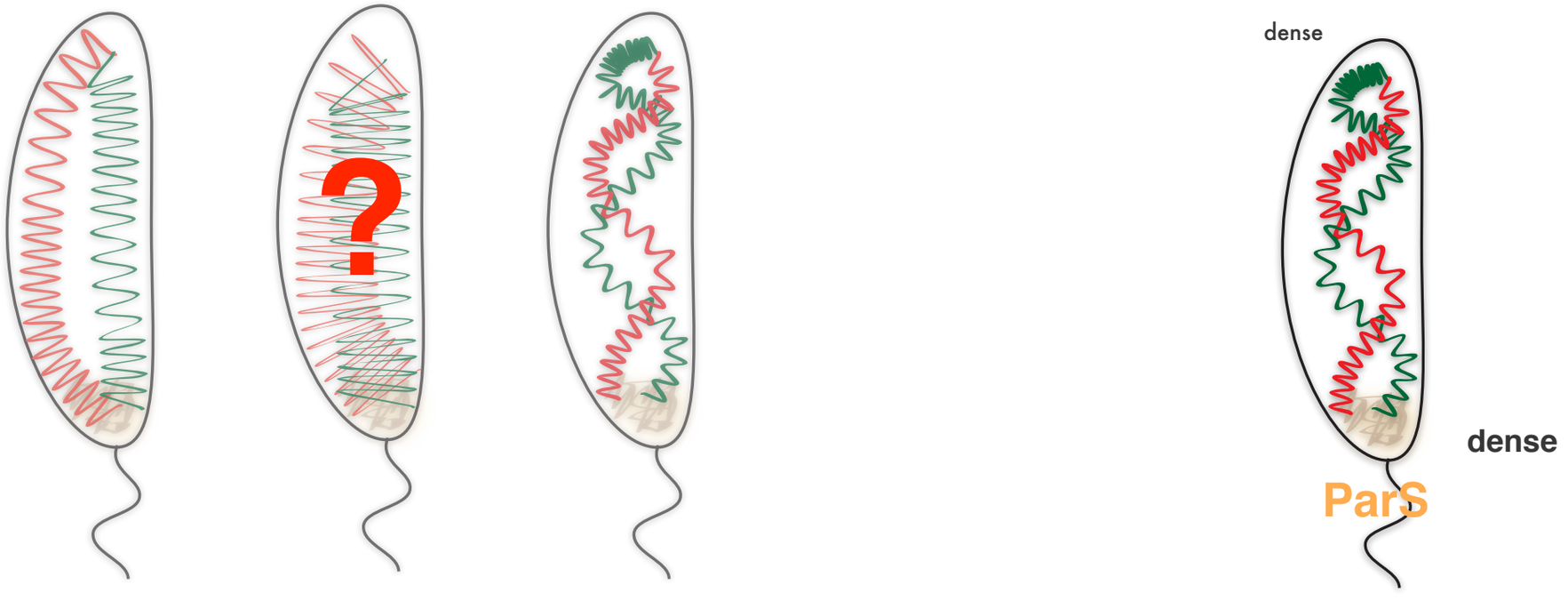




Arms are **STILL** helical

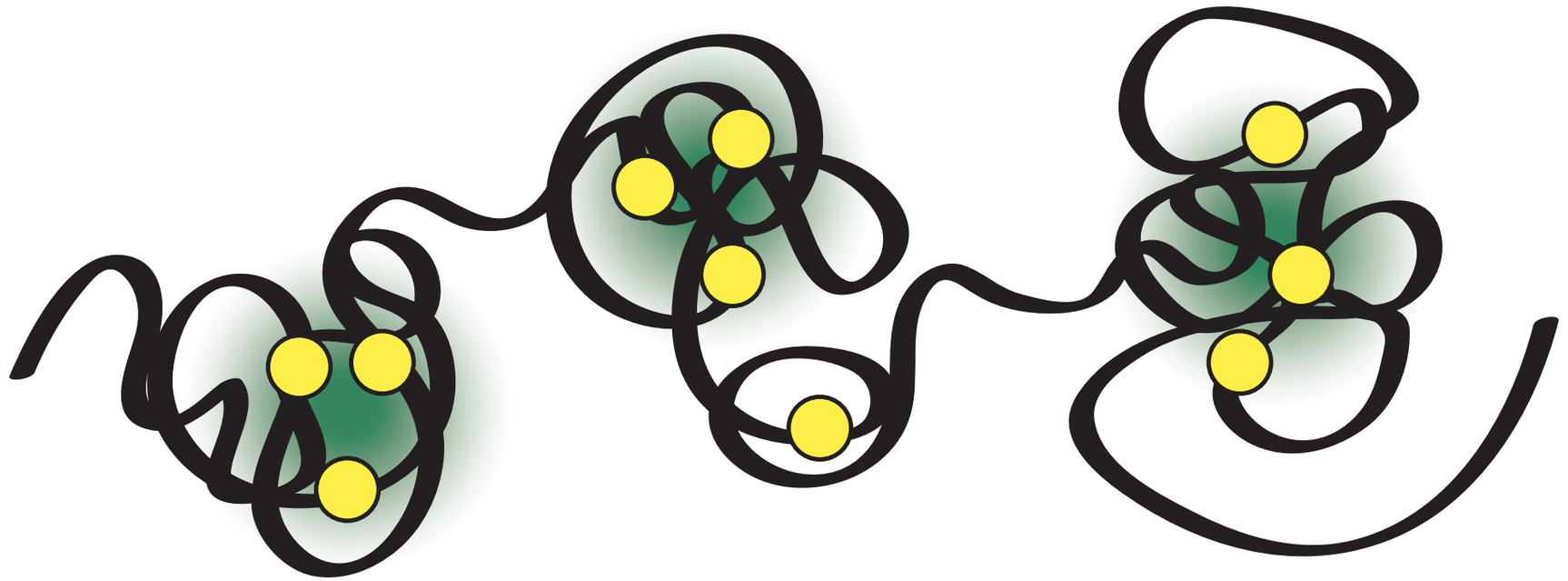
Genome architecture in Caulobacter



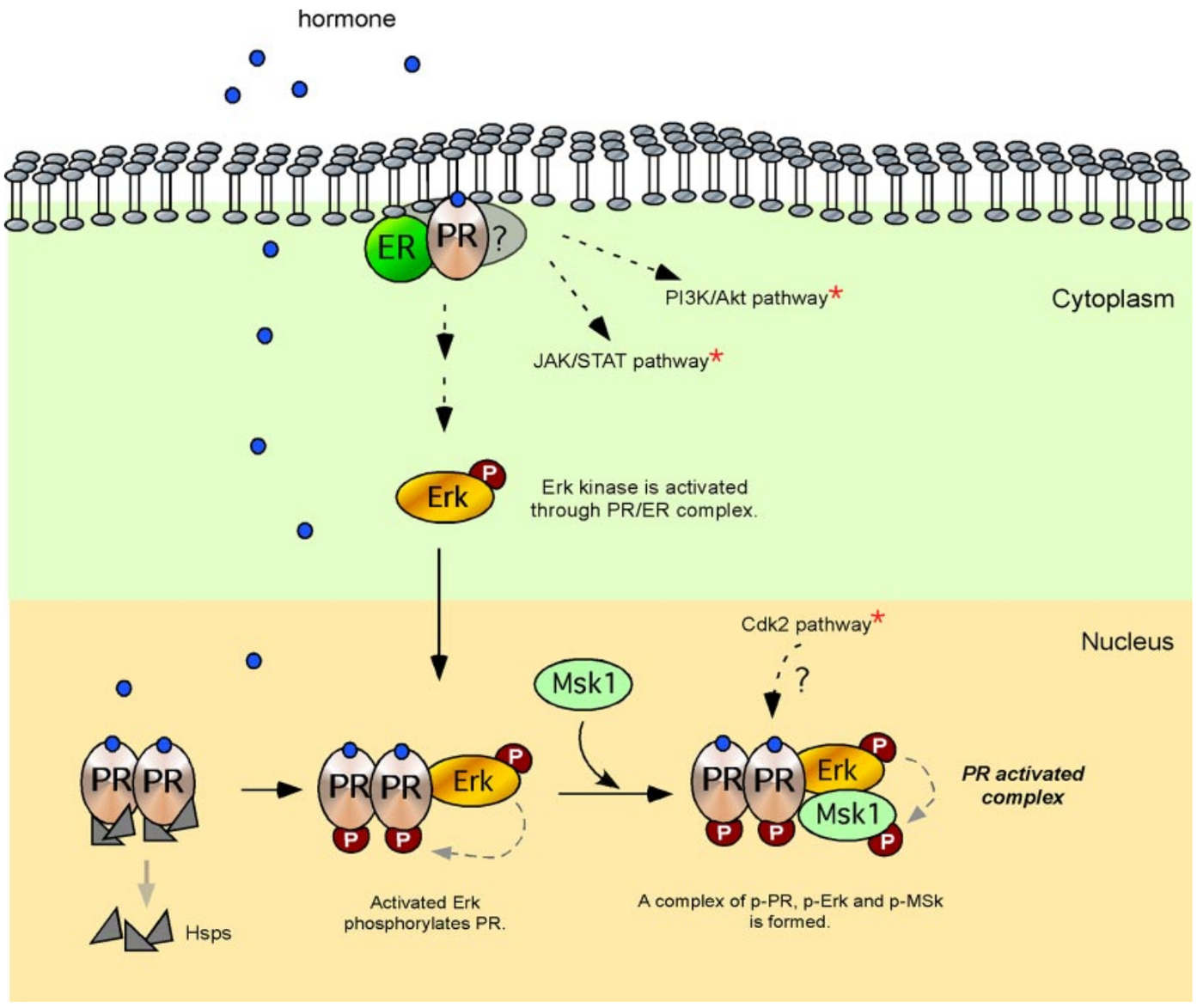


M.A. Umbarger, *et al.* Molecular Cell (2011) 44:252–264

On TADs and hormones



Progesterone-regulated transcription in breast cancer



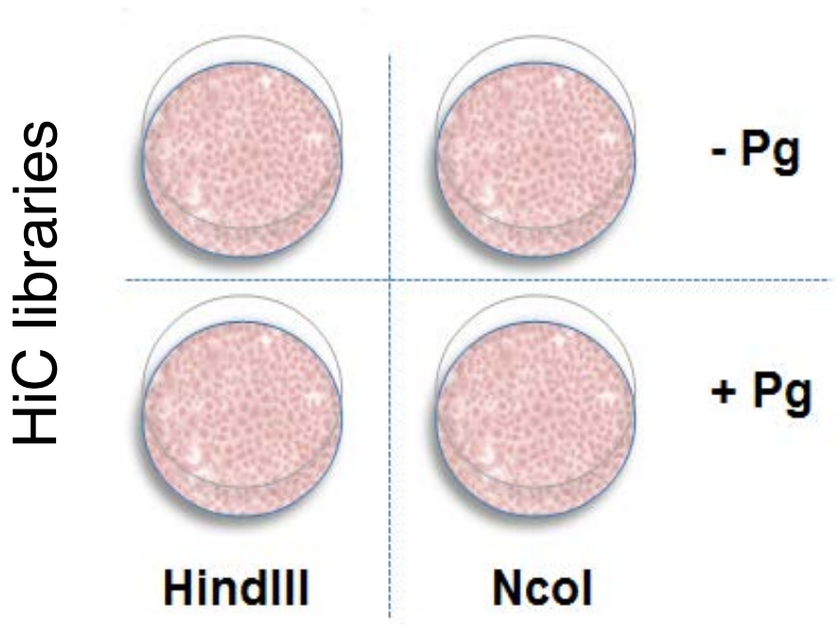
> 2,000 genes **Up**-regulated

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

Regulation in 3D?

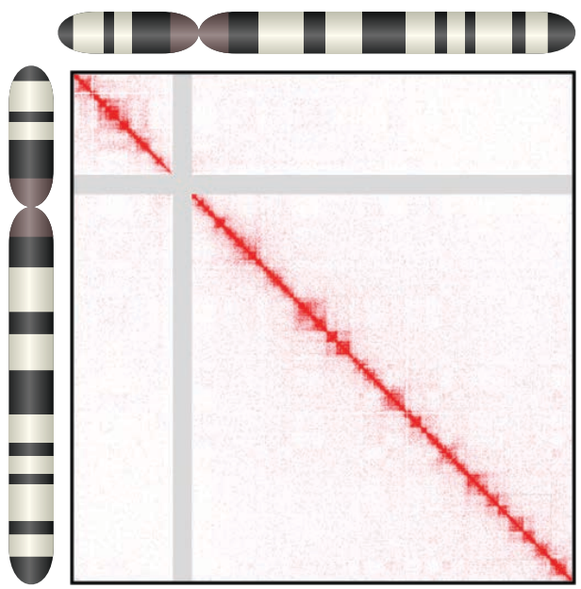
Vicent *et al* 2011, Wright *et al* 2012, Ballare *et al* 2012

Experimental design

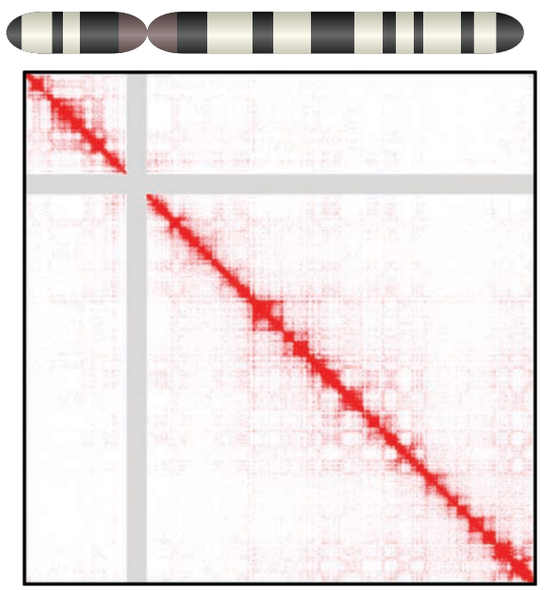


ChIP-Seq
RNA-Seq
Hi-C

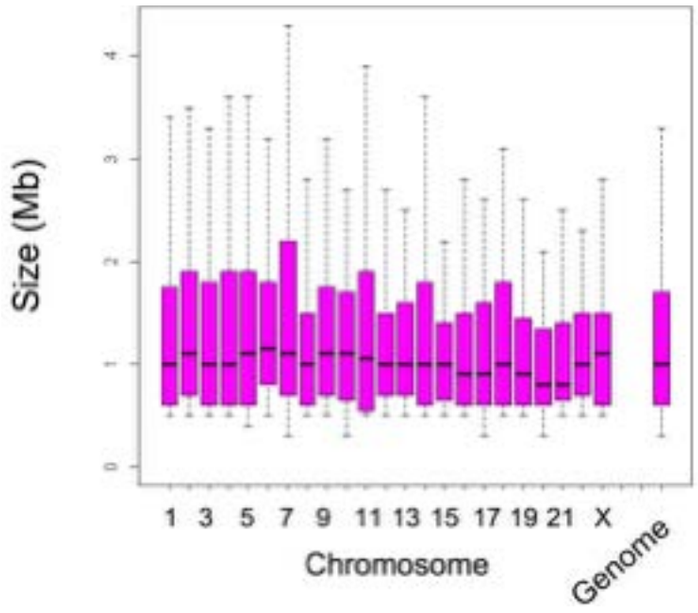
Chr.18 (Hind III)



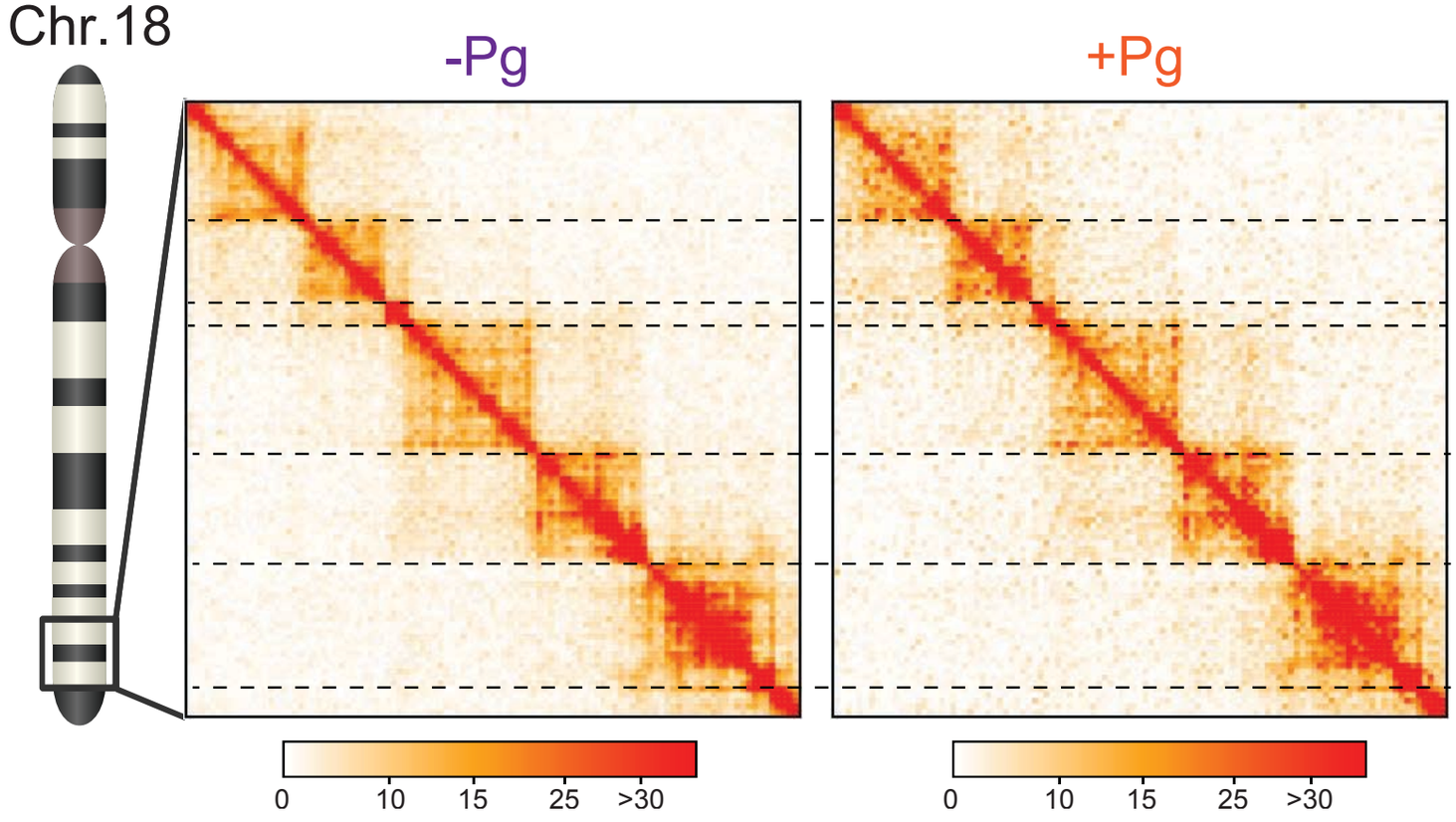
Chr.18 (NcoI)

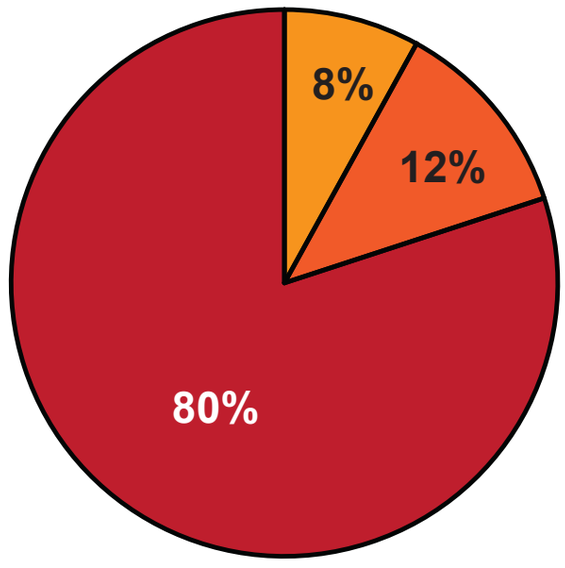


Are there TADs? how robust?

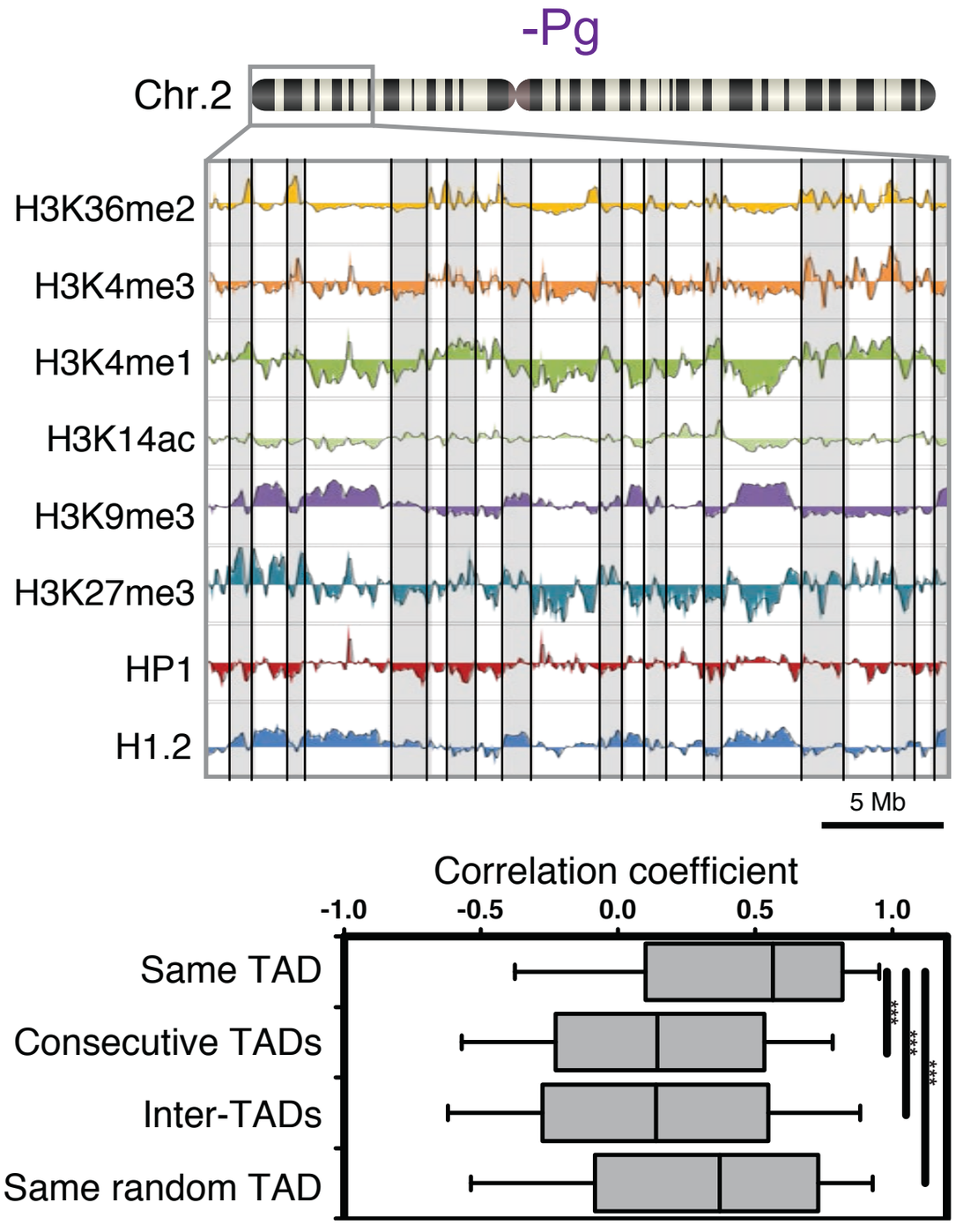


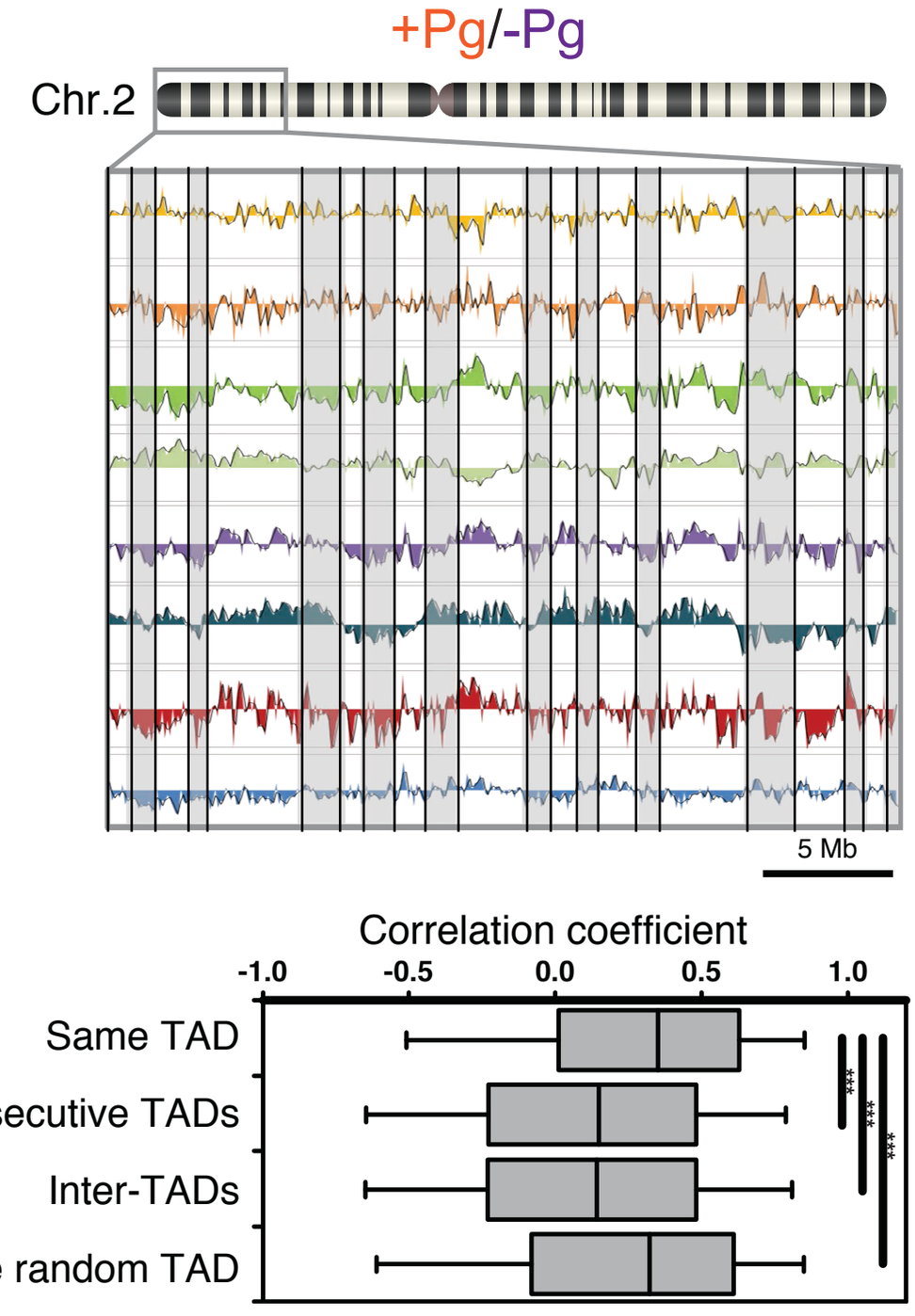
>2,000 detected TADs



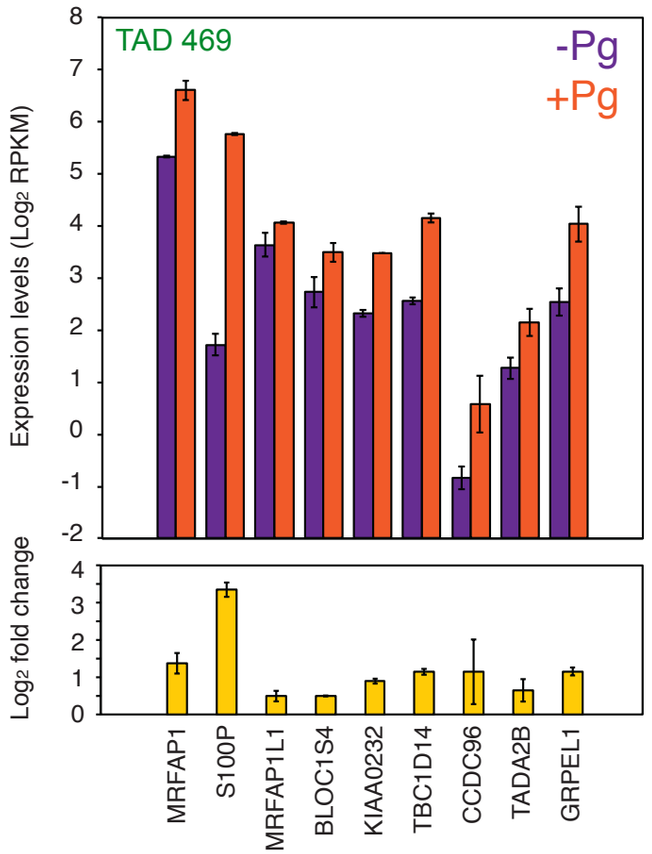


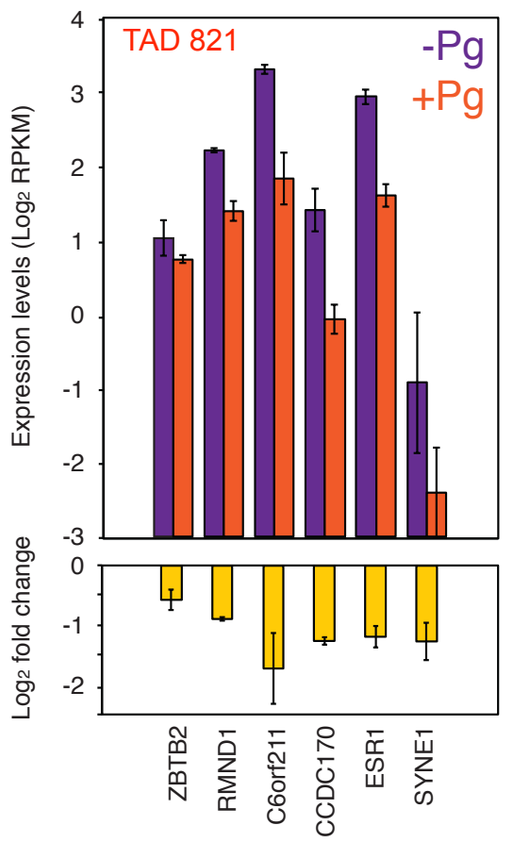
Are TADs homogeneous?

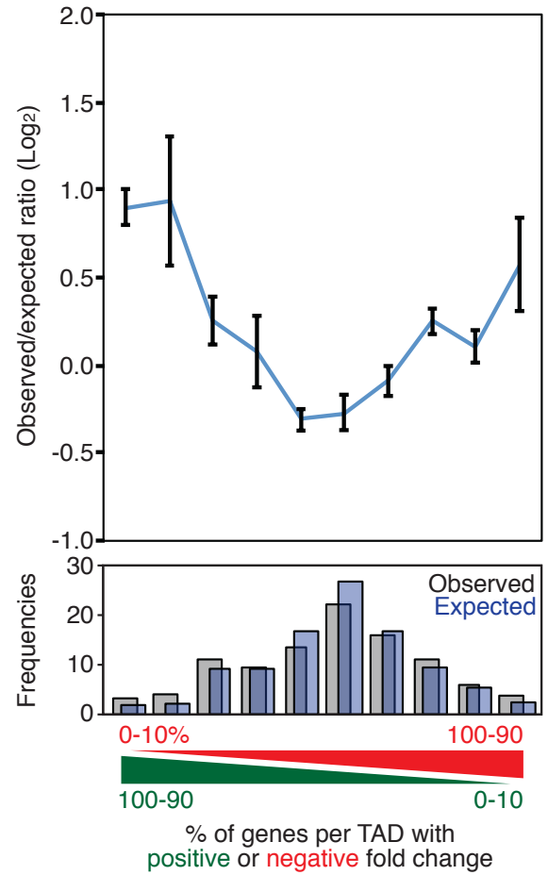




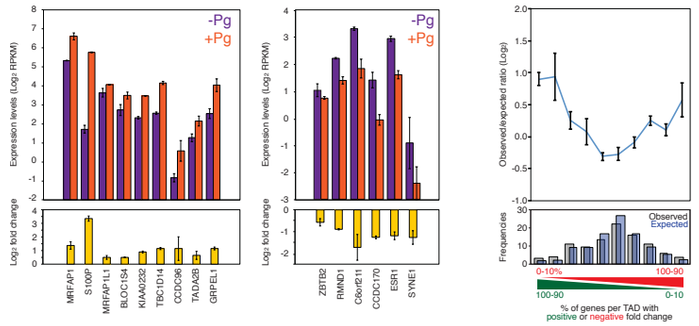
Do TADs respond differently to Pg treatment?



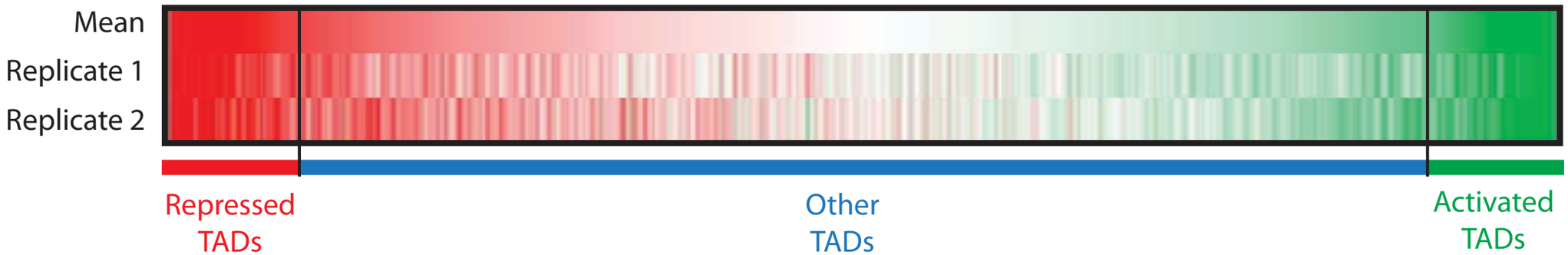


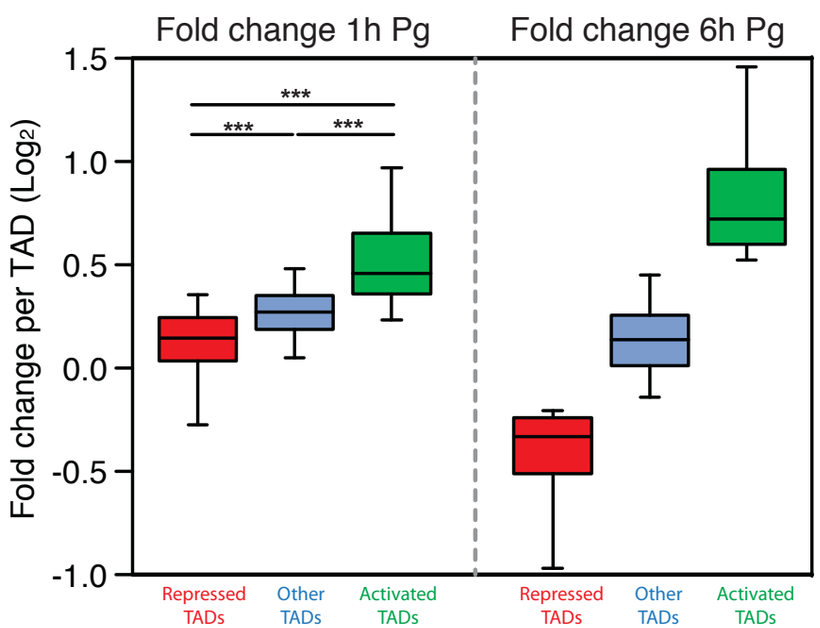


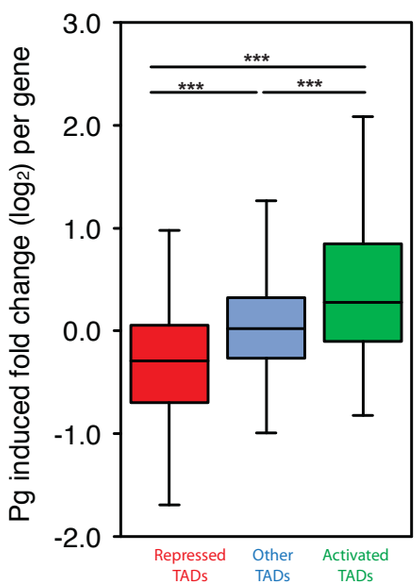
Do TADs respond differently to Pg treatment?

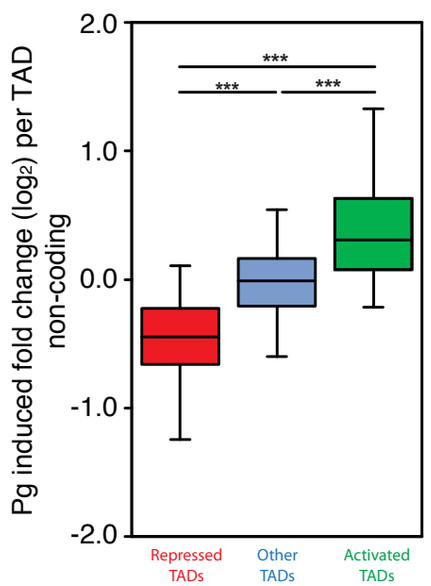


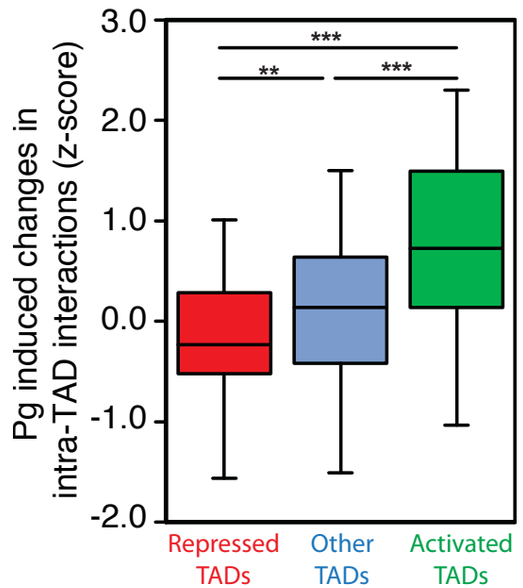
Pg induced fold change per TAD (6h)





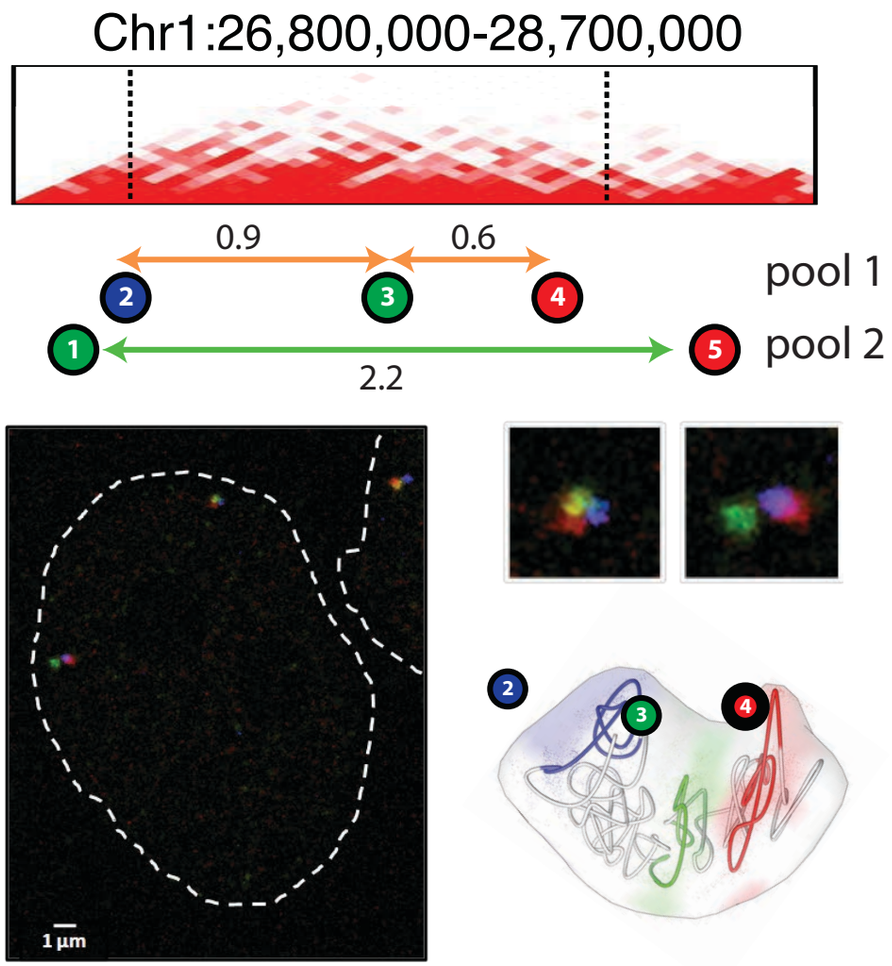


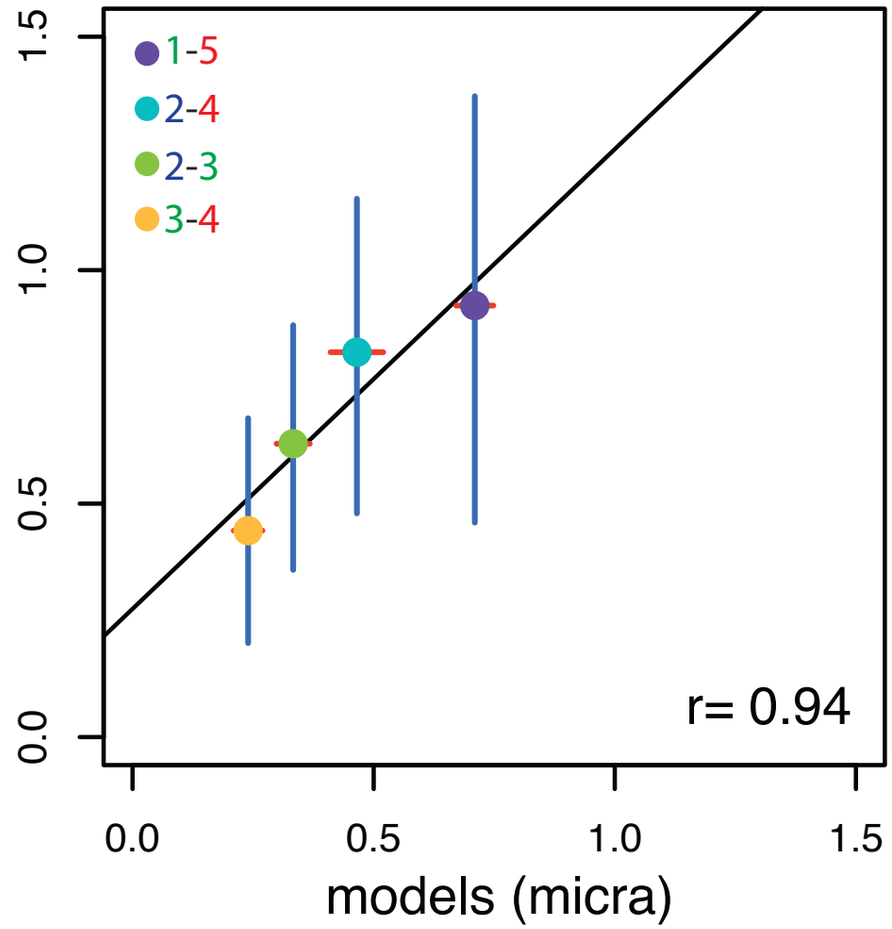




Modeling 3D TADs

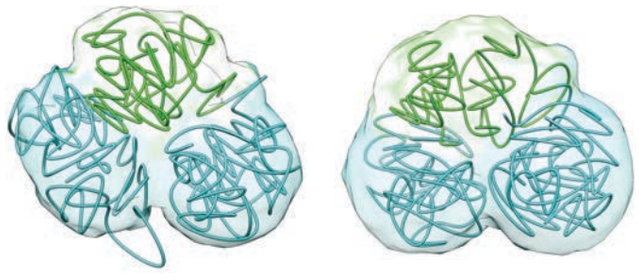
61 genomic regions containing 209 TADs covering 267Mb

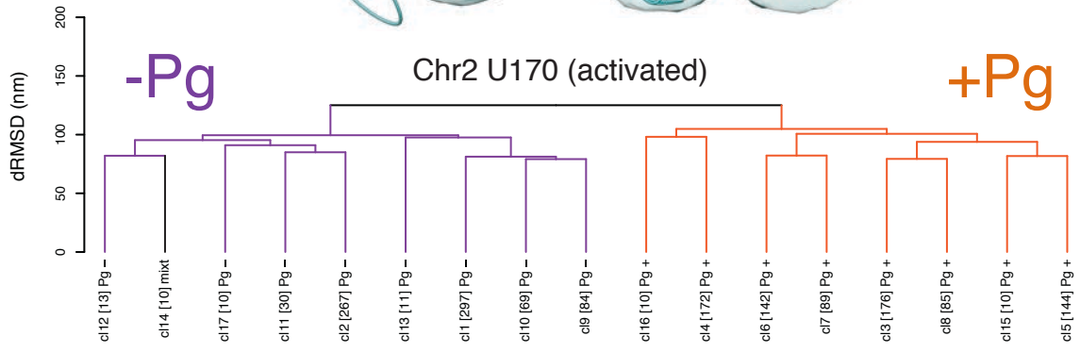




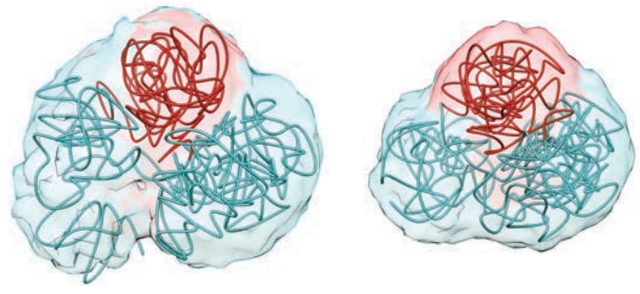
How TADs respond structurally to Pg?

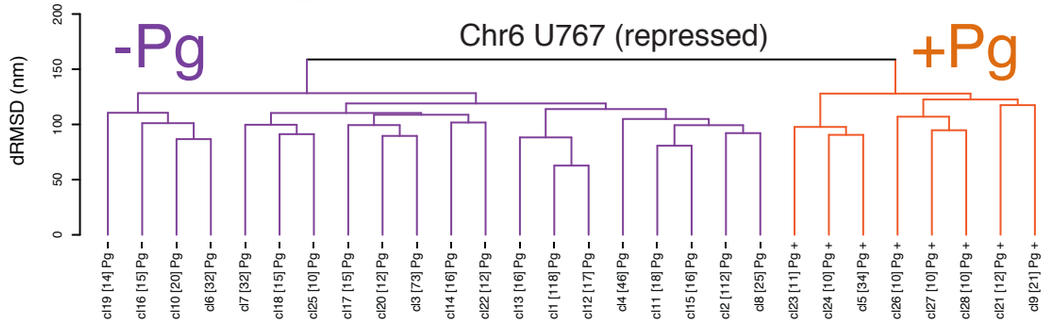
Chr2:9,600,000-13,200,000

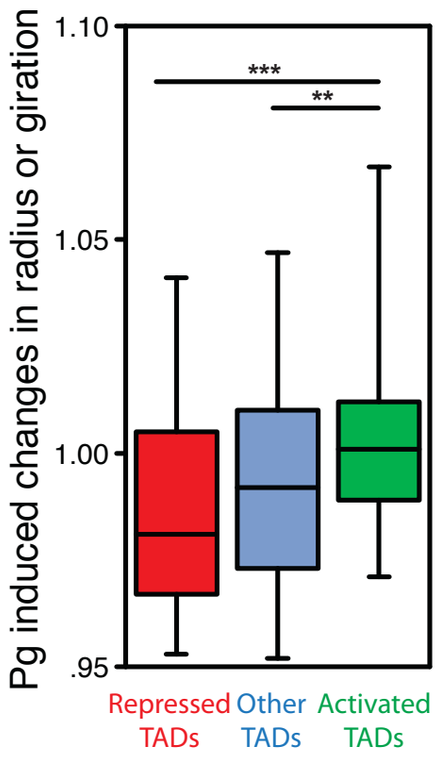


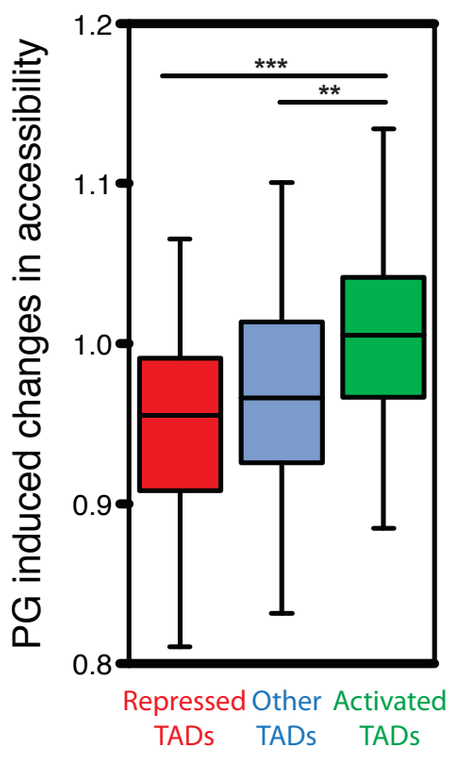


Chr6:71,800,000-76,500,000

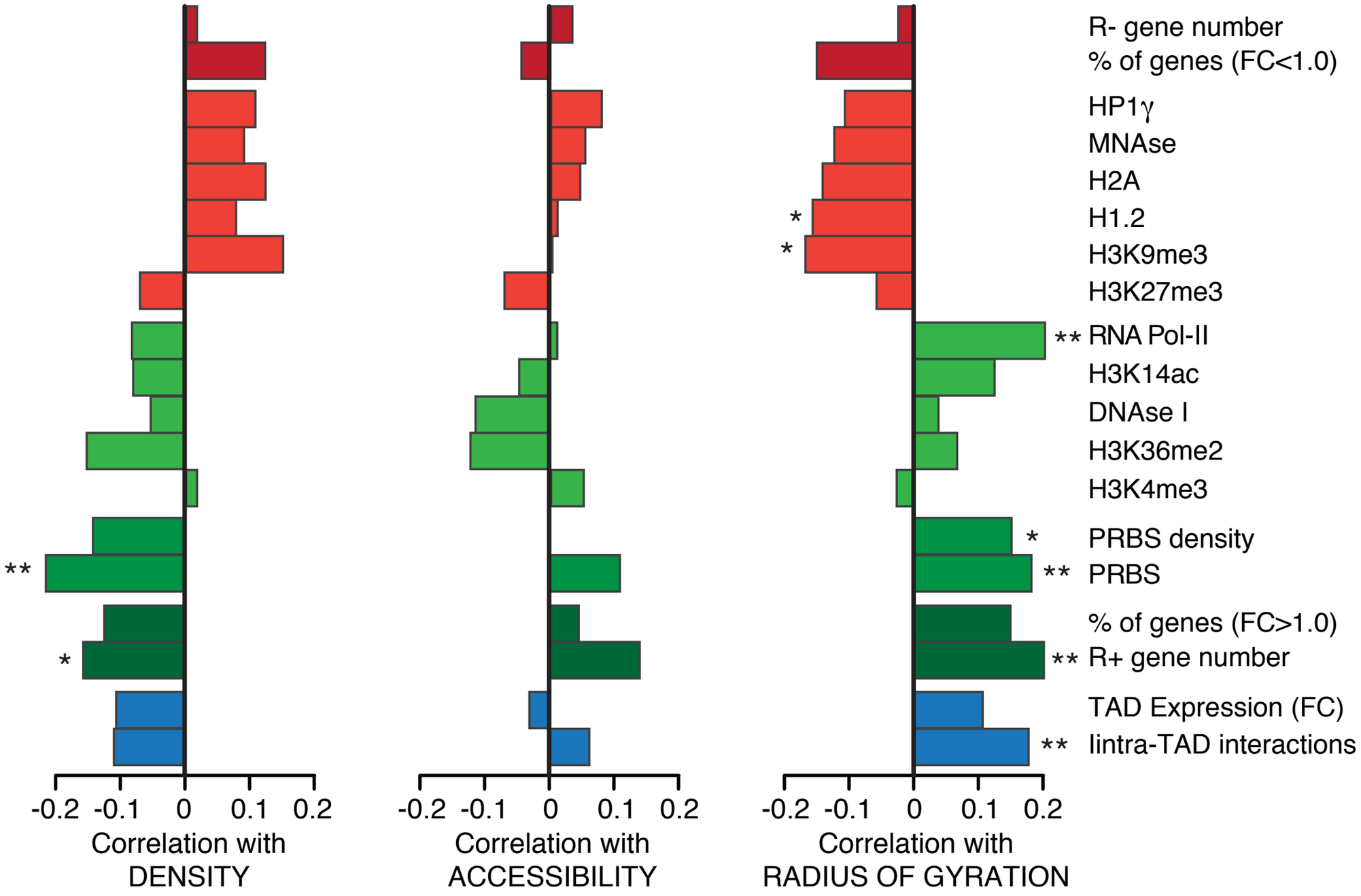


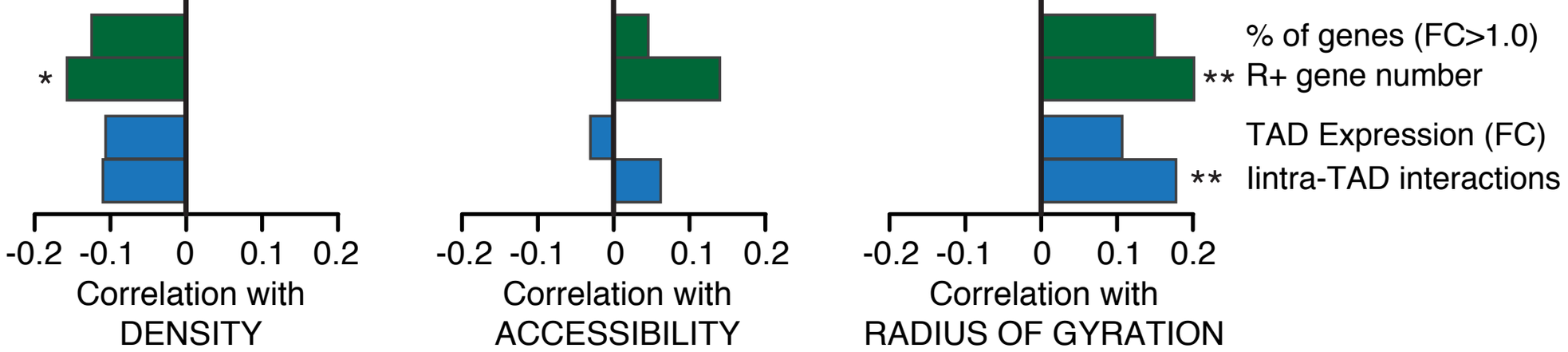


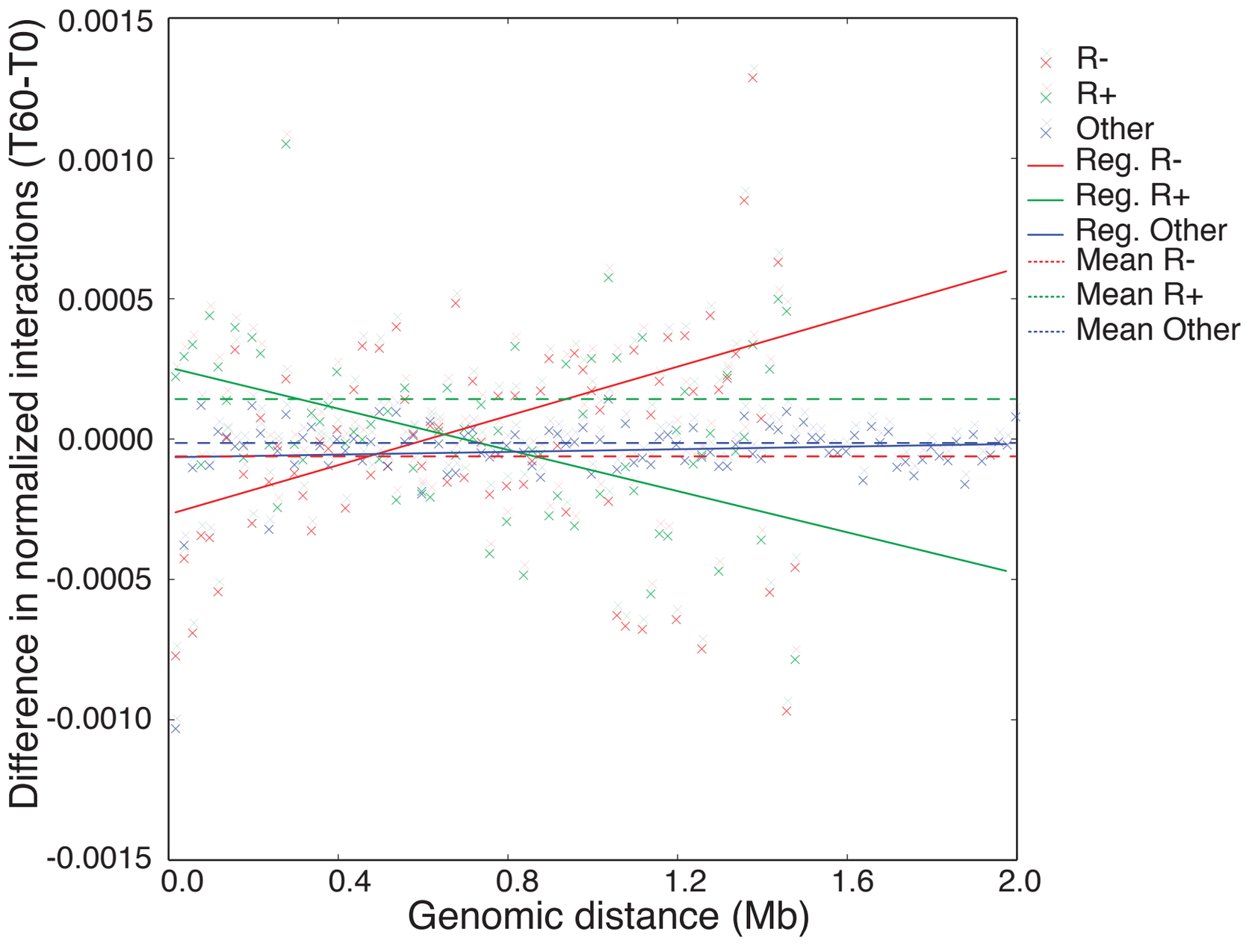




How TADs respond structurally to Pg?



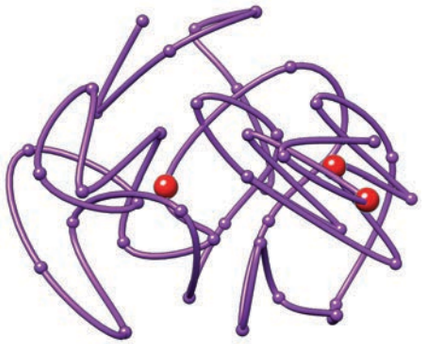

*
-0.2 -0.1 0 0.1 0.2
Correlation with
DENSITY
-0.2 -0.1 0 0.1 0.2
Correlation with
ACCESSIBILITY
-0.2 -0.1 0 0.1 0.2
Correlation with
RADIUS OF GYRATION
% of genes (FC>1.0)
** R+ gene number
TAD Expression (FC)
** Iintra-TAD interactions

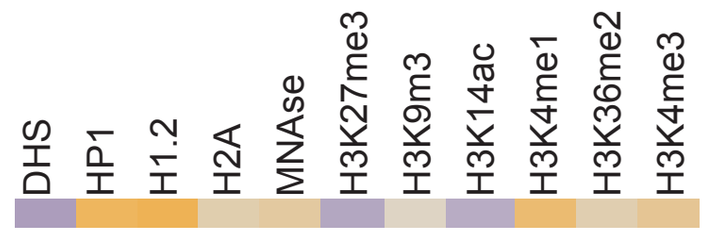

Difference in normalized interactions (T60-T0)
0.0015
0.0010
0.0005
0.0000
-0.0005
-0.0010
-0.0015
0.0
0.4
0.8
1.2
1.6
2.0
Genomic distance (Mb)
R-
R+
Other
Reg. R-
Reg. R+
Reg. Other
Mean R-
Mean R+
Mean Other

Model for TAD regulation

Repressed TAD

chr1 U41

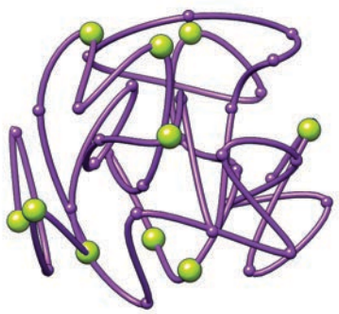


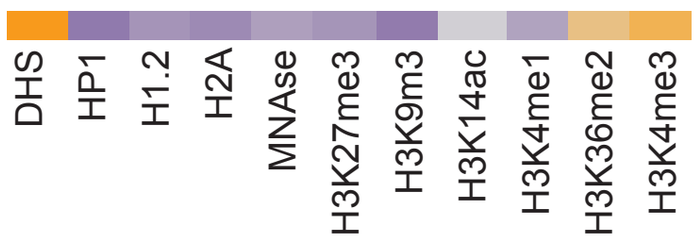


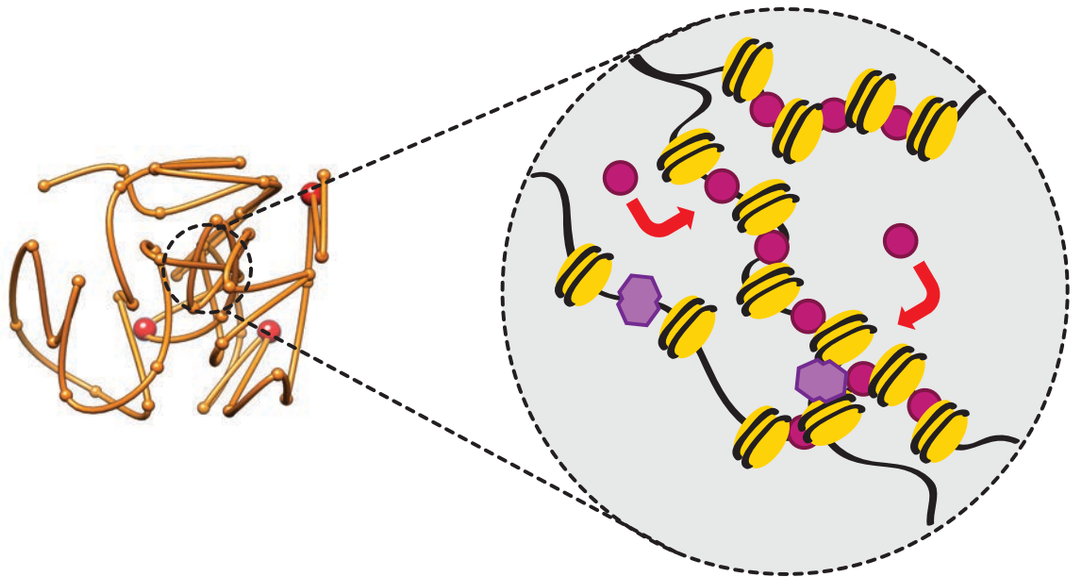
Structural transition **+Pg**

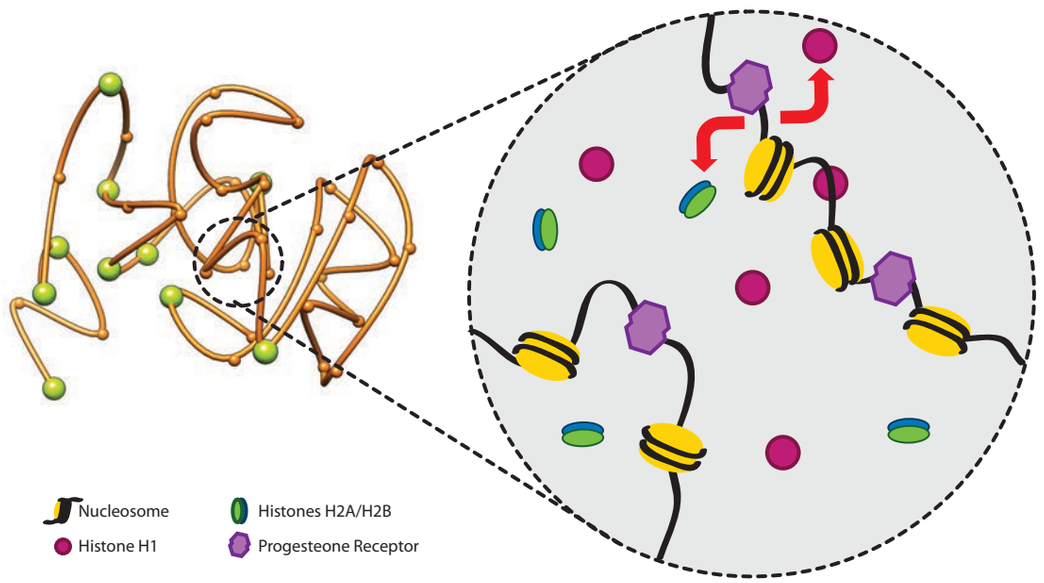
Activated TAD

chr2 U207





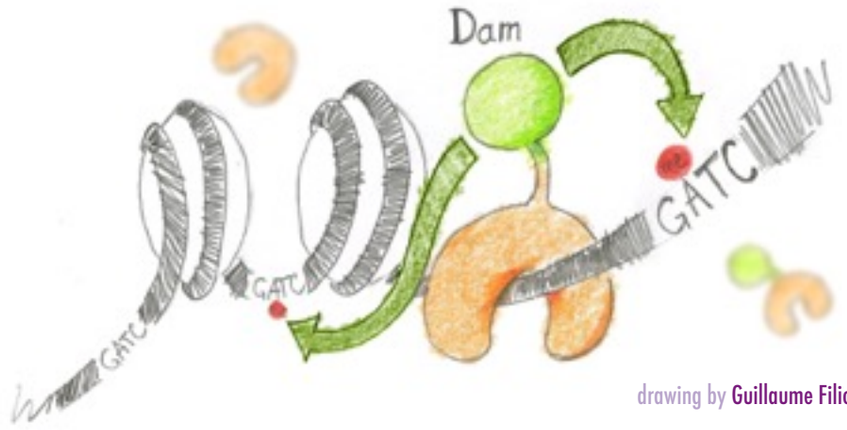




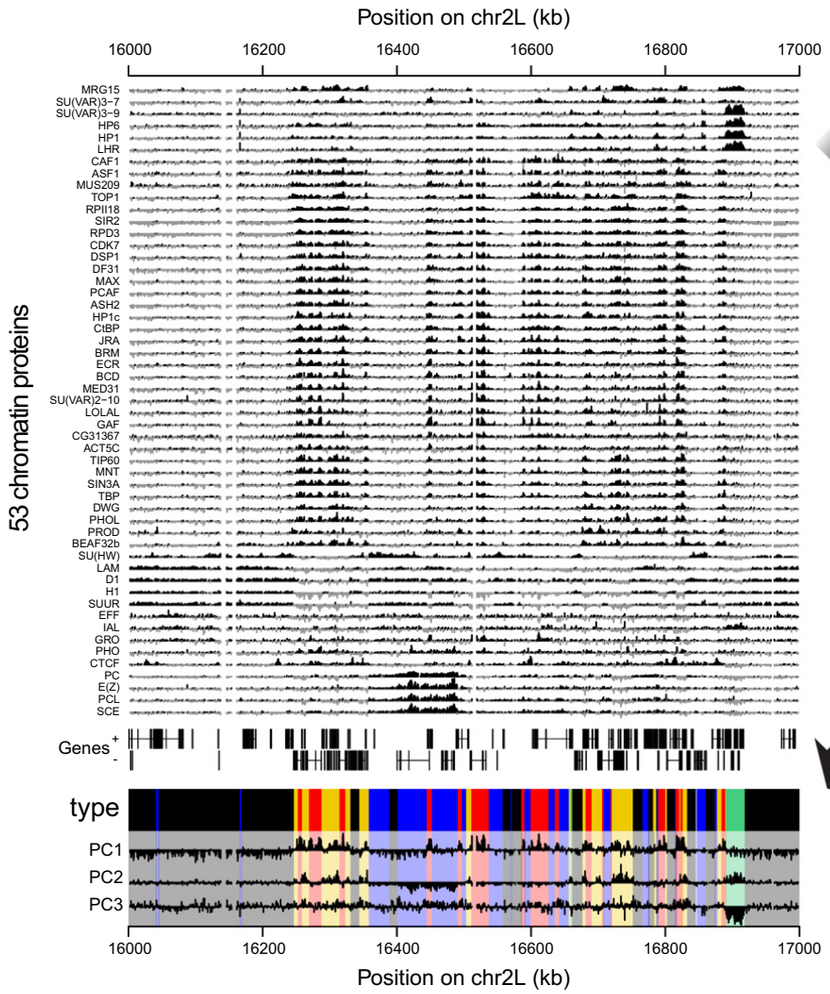
Structuring the COLORs of chromatin

Fly Chromatin COLORs

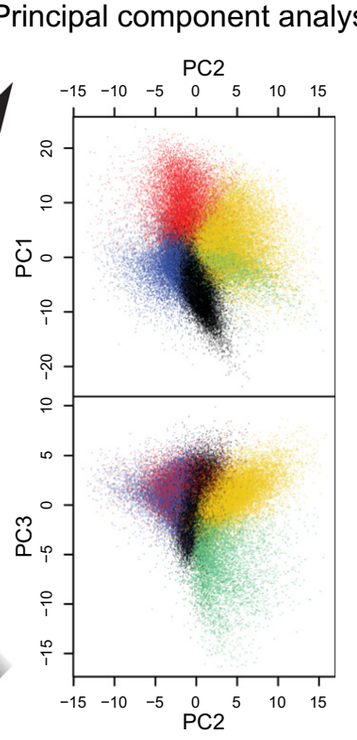
Filion et al. (2010). Cell, 143(2), 212–224.



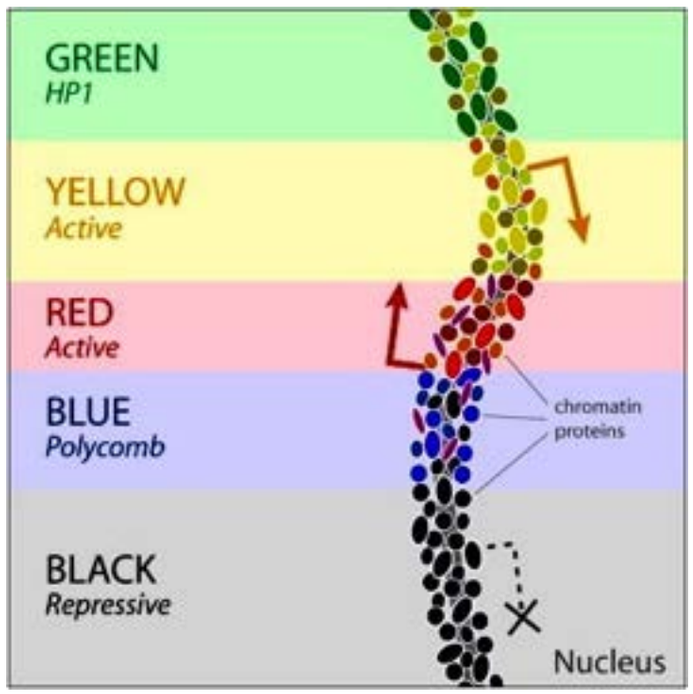
drawing by Guillaume Filion



Principal component analysis

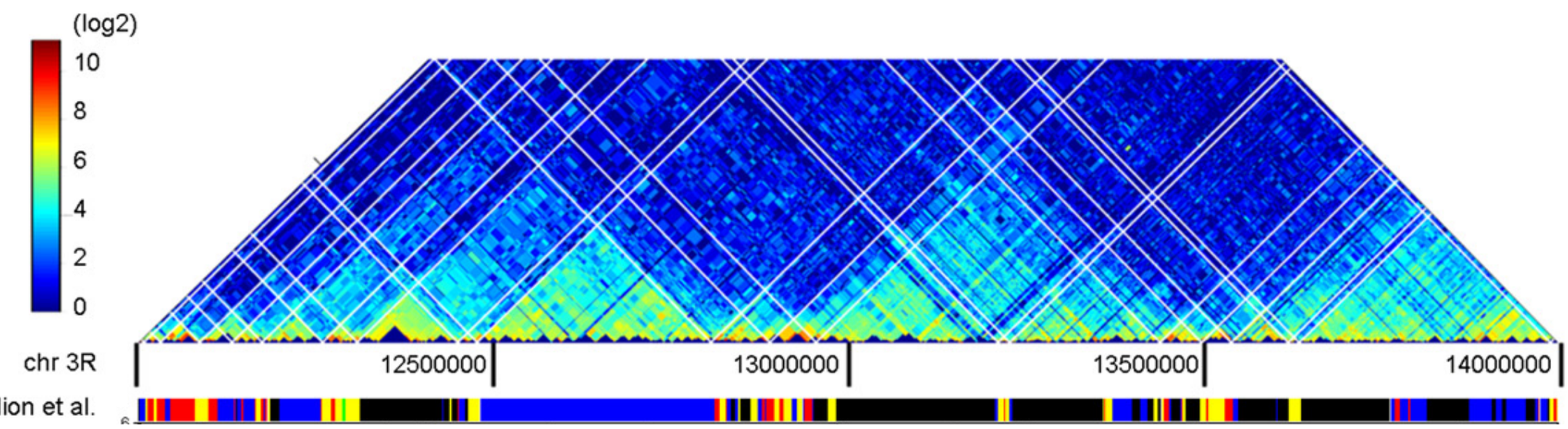


Hidden Markov model

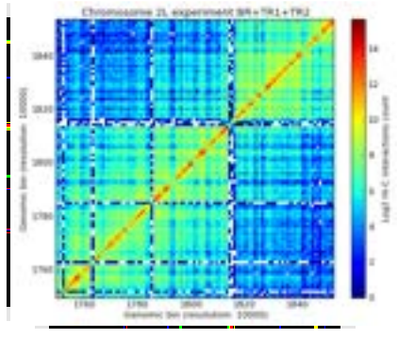


Fly Chromatin COLORs

Hou et al. (2012). Molecular Cell, 48(3), 471–484.



Filion et al.



50 ~1Mb regions
10 for each color

Structural properties

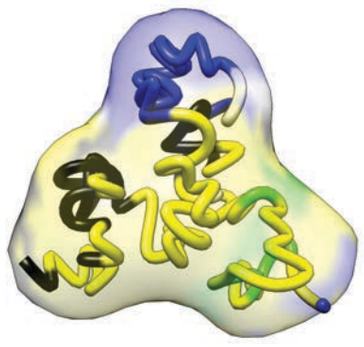
50 1Mb regions. 10 enriched for each color.

RED dense region
3R:18920000-19920000

22% 17% 0% 11% 45% 6%

YELLOW dense region
X:15590000-16600000

0% 48% 4% 20% 26% 3%

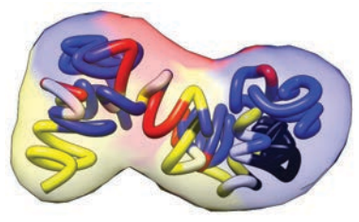


GREEN dense region
2R:510000-1530000

0% 0% 100% 0% 0% 0%

BLUE dense region
3L:210000-1230000

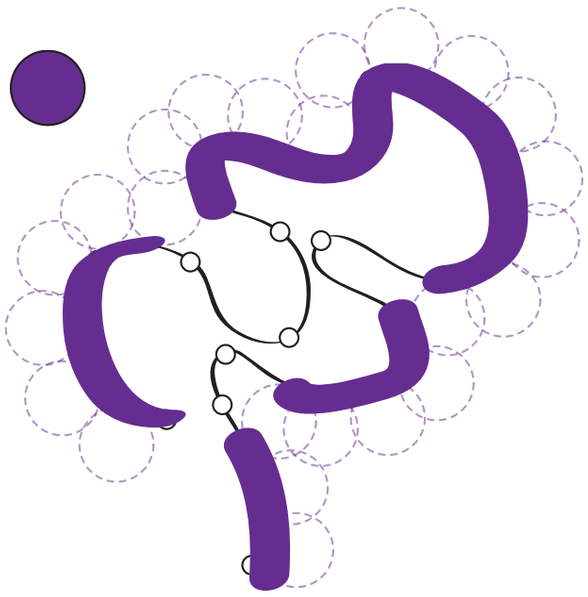
11% 17% 0% 52% 13% 0%



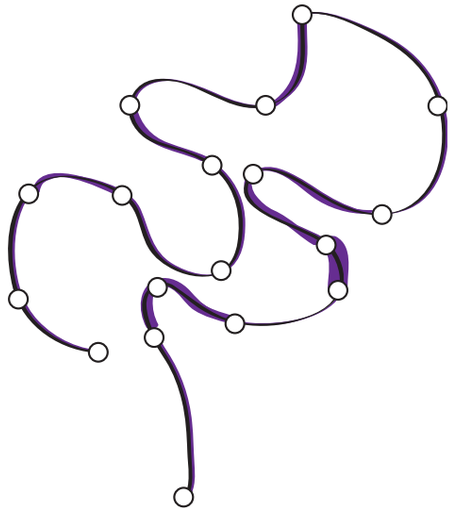
BLACK dense region
2L:17500000-18530000

1% 0% 0% 0% 98% 1%

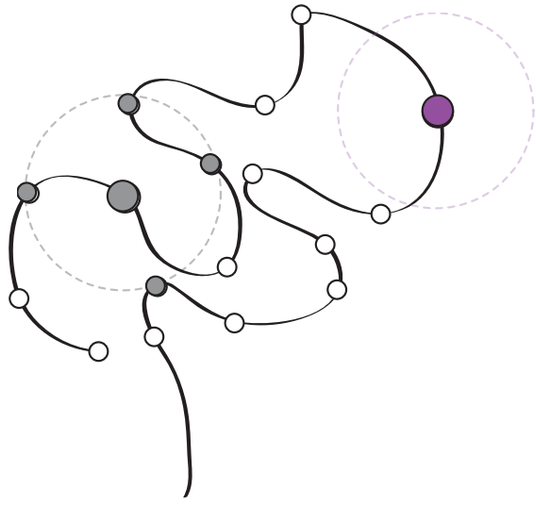
Accessibility (%)



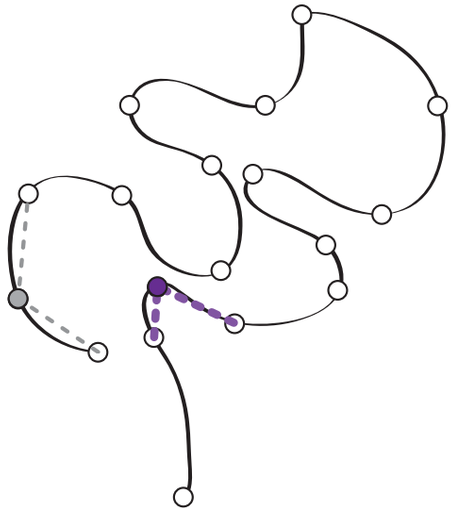
Density (bp/nm)



Interactions



Angle



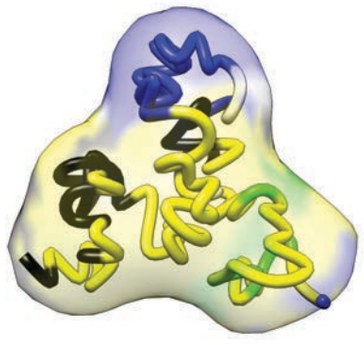
Structural COLORs

RED dense region
3R:18920000-19920000

22% 17% 0% 11% 45% 6%

YELLOW dense region
X:15590000-16600000

0% 48% 4% 20% 26% 3%



GREEN dense region
2R:510000-1530000

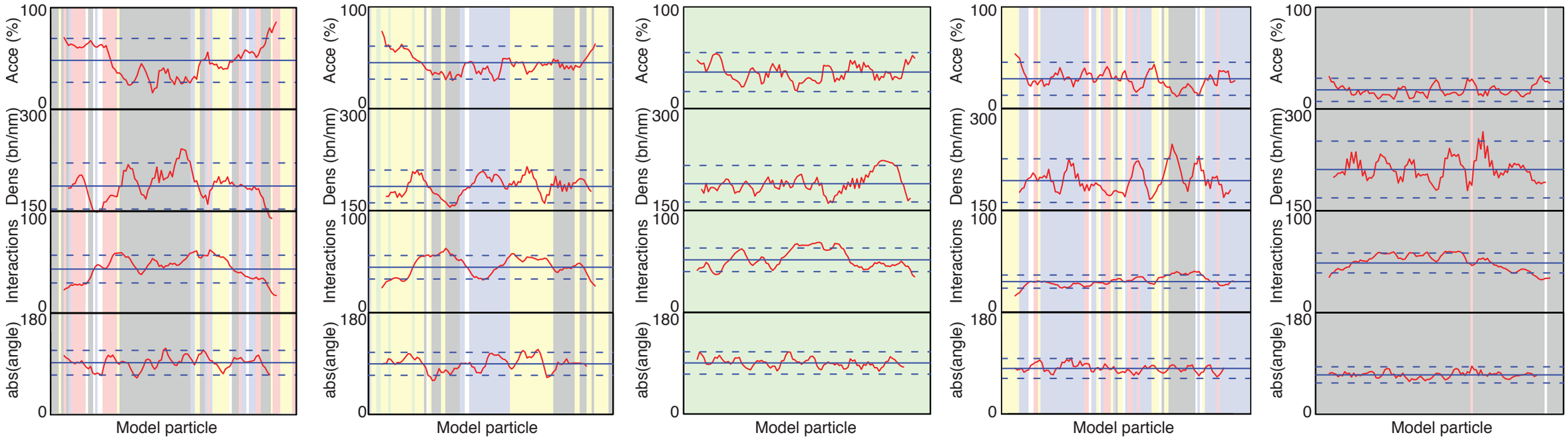
0% 0% 100% 0% 0% 0%

BLUE dense region
3L:210000-1230000

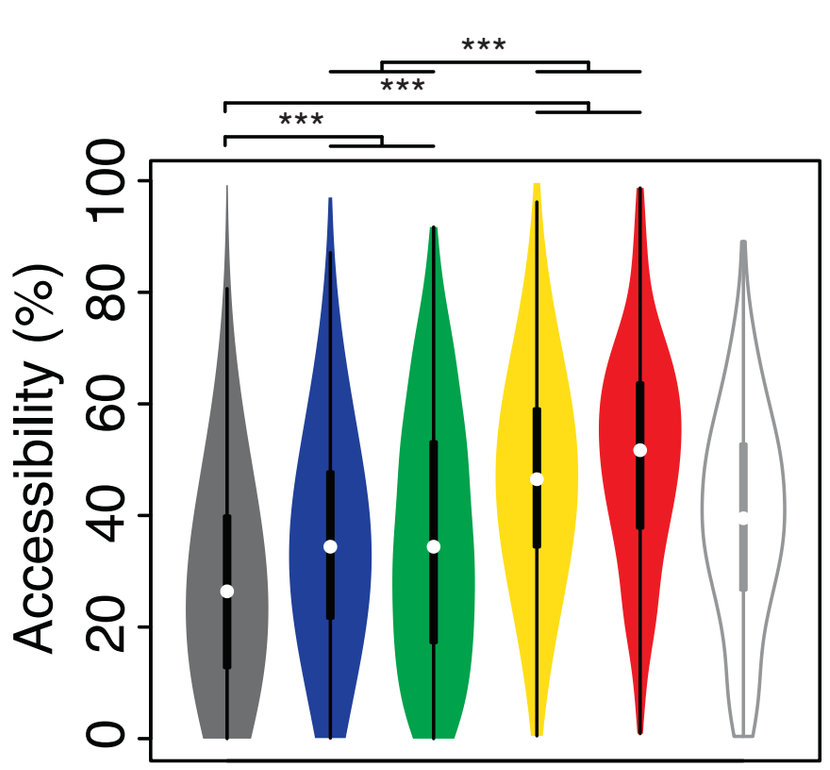
11% 17% 0% 52% 13% 0%

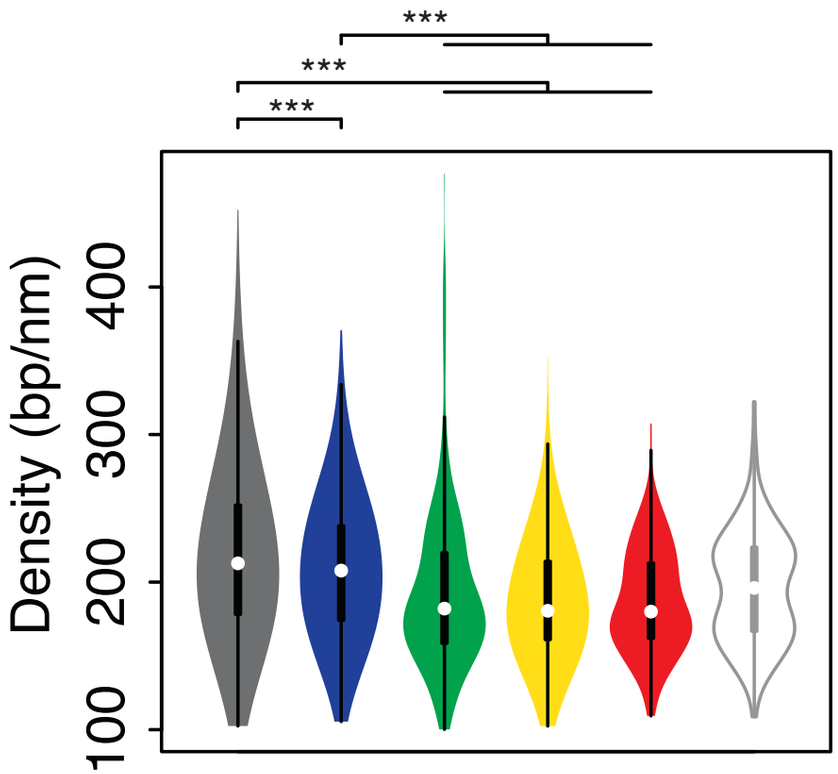
BLACK dense region
2L:17500000-18530000

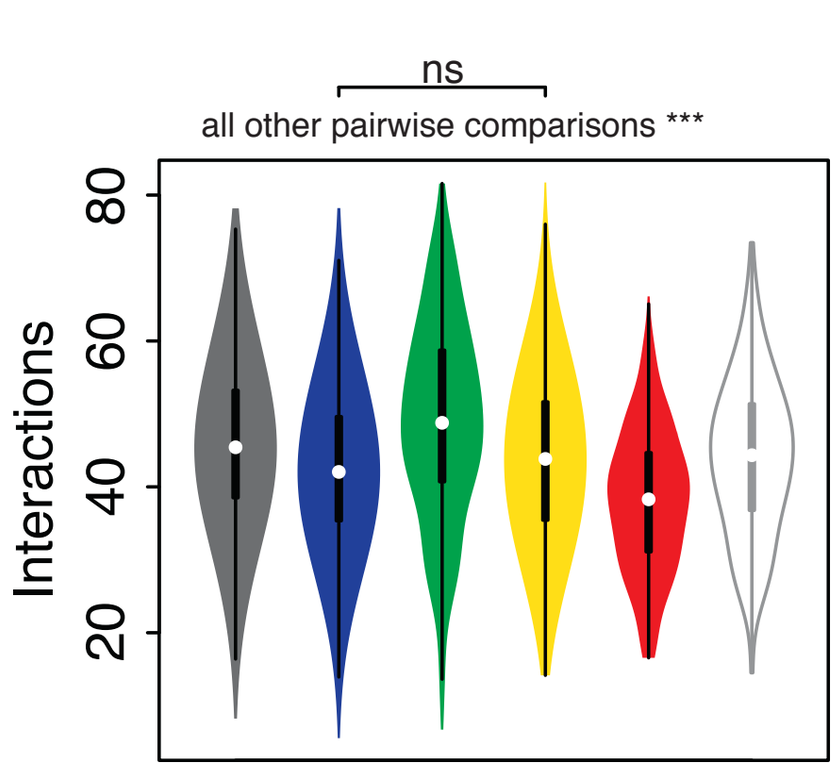
1% 0% 0% 0% 98% 1%

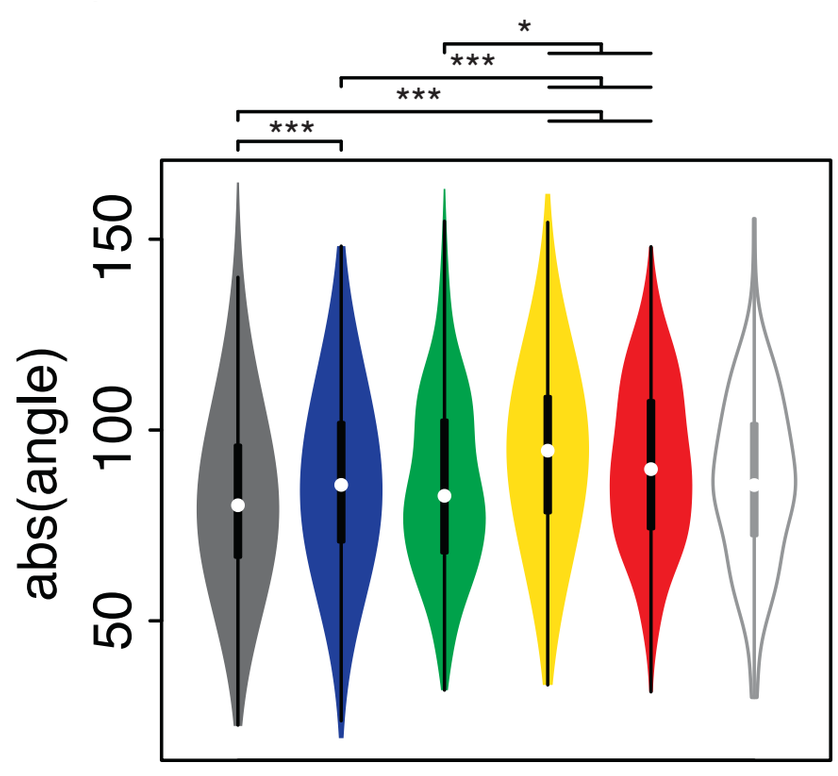


Structural COLORs

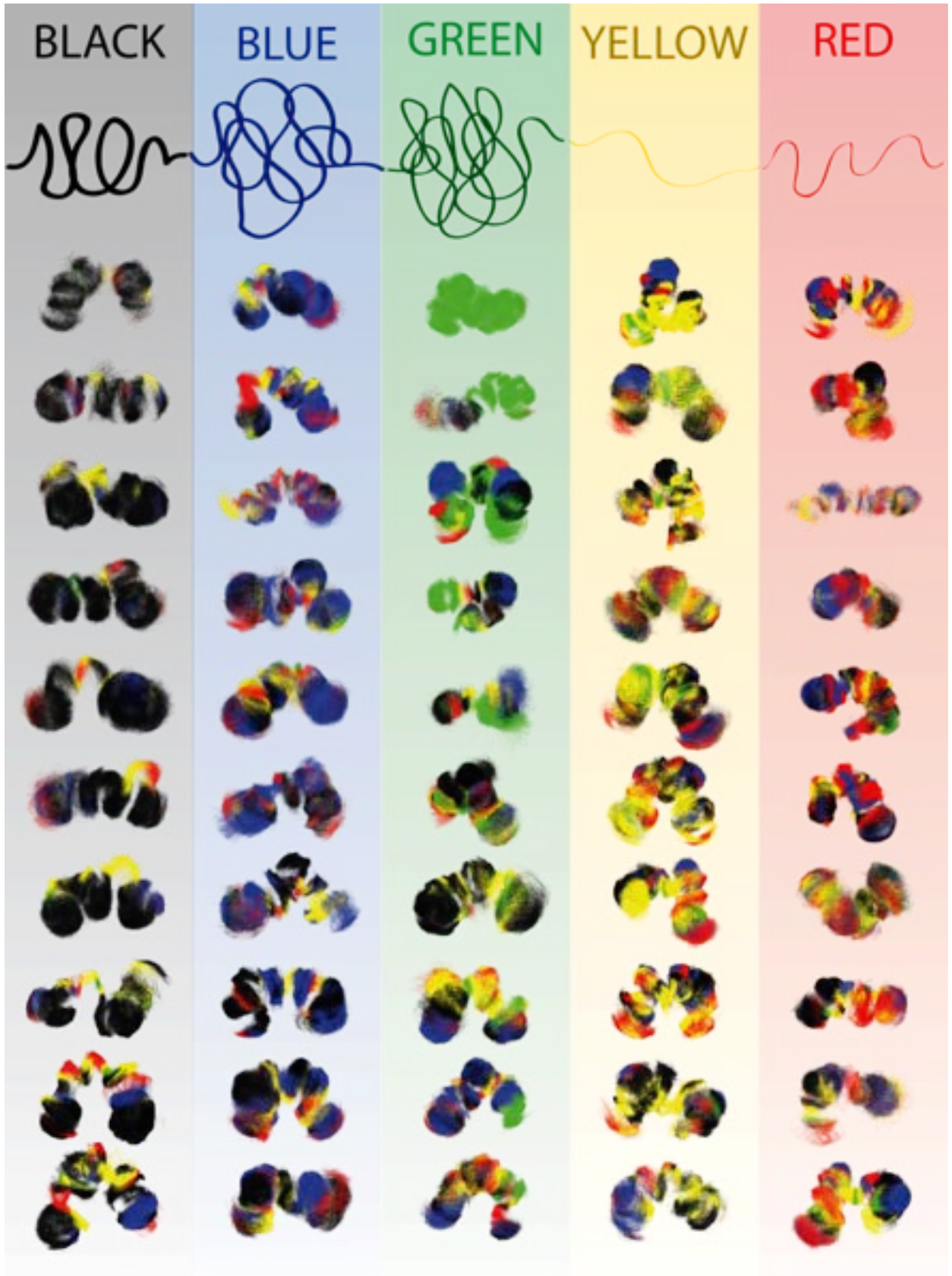


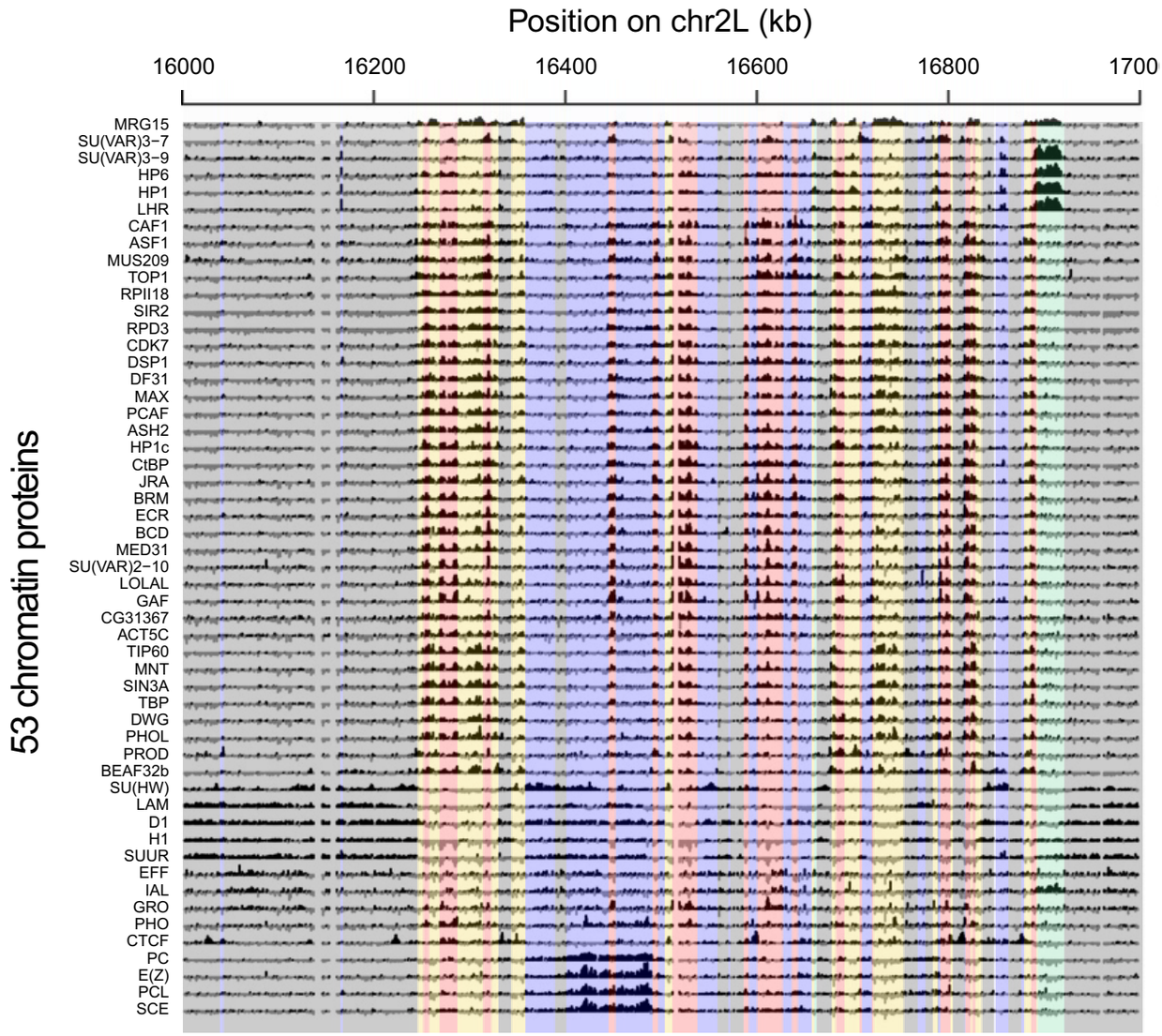




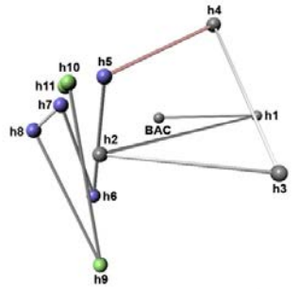


Structural COLORs

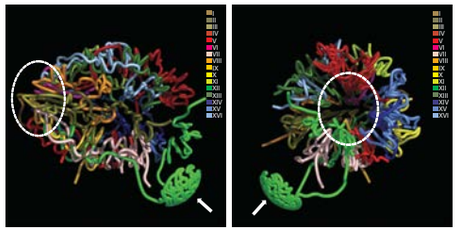




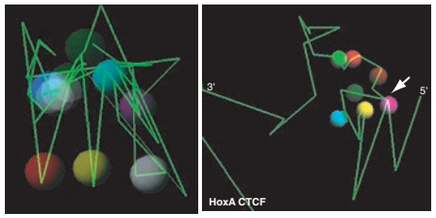
Are the models correct?



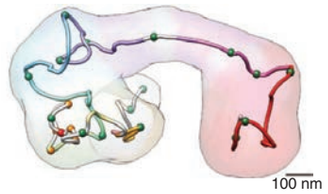
Jhunjhunwala (2008) Cell



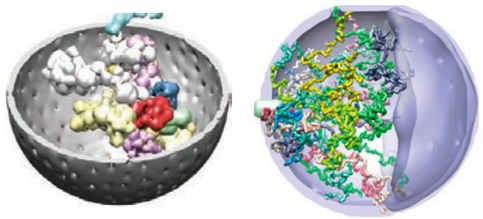
Duan (2010) Nature



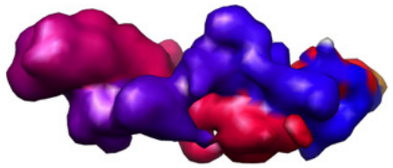
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research



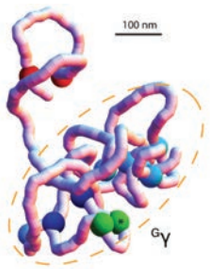
Baù (2011) Nature Structural & Molecular Biology



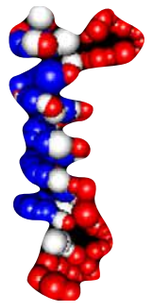
Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research



Umbarger (2011) Molecular Cell



Junier (2012) Nucleic Acids Research



Hu (2013) PLoS Computational Biology

Nucleic Acids Research Advance Access published March 23, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv221

Assessing the limits of restraint-based 3D modeling of genomes and genomic domains

Marie Trussart[1,2], François Serra[3,4], Davide Baù[3,4], Ivan Junier[2,3], Luís Serrano[1,2,5] and Marc A. Marti-Renom[3,4,5,*]

[1]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, [4]Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain and [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Received January 16, 2015; Revised February 16, 2015; Accepted February 22, 2015

ABSTRACT

Restraint-based modeling of genomes has been recently explored with the advent of Chromosome Conformation Capture (3C-based) experiments. We previously developed a reconstruction method to resolve the 3D architecture of both prokaryotic and eukaryotic genomes using 3C-based data. These models were congruent with fluorescent imaging validation. However, the limits of such methods have not systematically been assessed. Here we propose the first evaluation of a mean-field restraint-based reconstruction of genomes by considering diverse chromosome architectures and different levels of data noise and structural variability. The results show that: first, current scoring functions for 3D reconstruction correlate with the accuracy of the models; second, reconstructed models are robust to noise but sensitive to structural variability; third, the local structure organization of genomes, such as Topologically Associating Domains, results in more accurate models; fourth, to a certain extent, the models capture the intrinsic structural variability in the input matrices and fifth, the accuracy of the models can be *a priori* predicted by analyzing the properties of the interaction matrices. In summary, our work provides a systematic analysis of the limitations of a mean-field restrain-based method, which could be taken into consideration in further development of methods as well as their applications.

INTRODUCTION

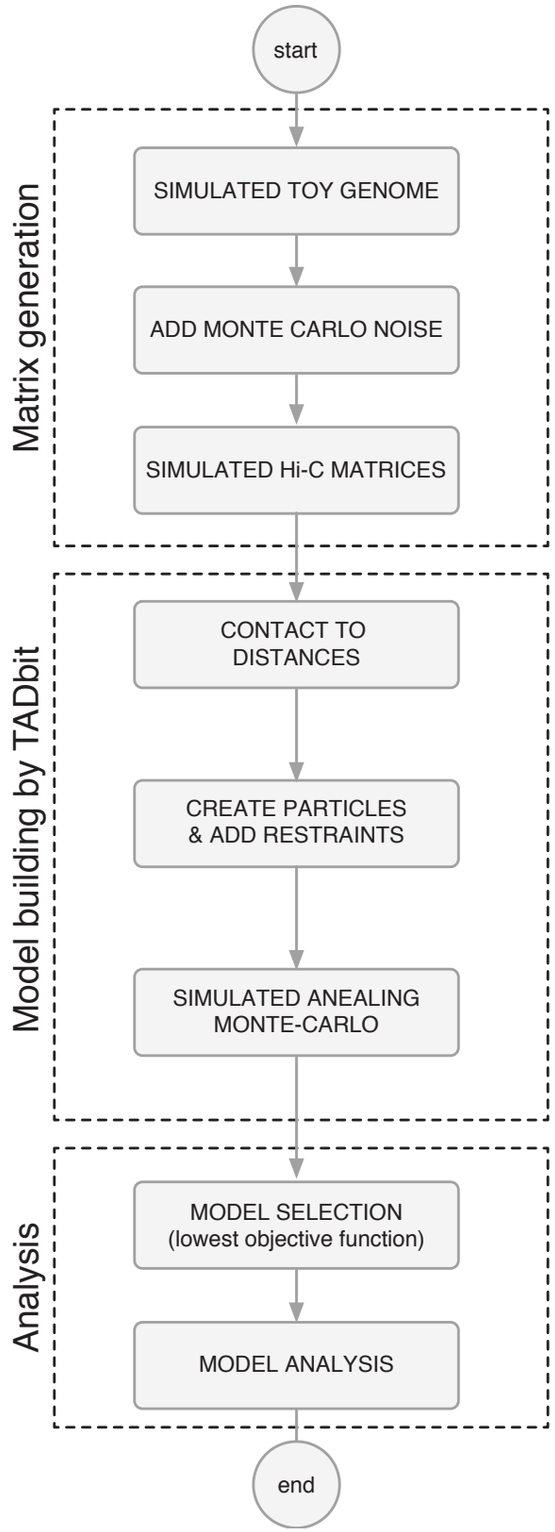
Recent studies of the three-dimensional (3D) conformation of genomes are revealing insights into the organization and the regulation of biological processes, such as gene expression regulation and replication (1–6). The advent of the so-called Chromosome Conformation Capture (3C) assays (7), which allowed identifying chromatin-looping interactions between pairs of loci, helped deciphering some of the key elements organizing the genomes. High-throughput derivations of genome-wide 3C-based assays were established with Hi-C technologies (8) for an unbiased identification of chromatin interactions. The resulting genome interaction matrices from Hi-C experiments have been extensively used for computationally analyzing the organization of genomes and genomic domains (5). In particular, a significant number of new approaches for modeling the 3D organization of genomes have recently flourished (9–14). The main goal of such approaches is to provide an accurate 3D representation of the bi-dimensional interaction matrices, which can then be more easily explored to extract biological insights. One type of methods for building 3D models from interaction matrices relies on the existence of a limited number of conformational states in the cell. Such methods are regarded as mean-field approaches and are able to capture, to a certain degree, the structural variability around these mean structures (15).

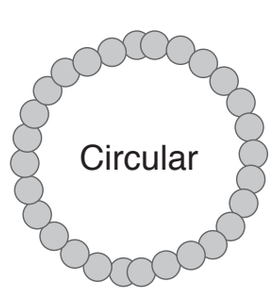
We recently developed a mean-field method for modeling 3D structures of genomes and genomic domains based on 3C interaction data (9). Our approach, called TADbit, was developed around the Integrative Modeling Platform (IMP, http://integrativemodeing.org), a general framework for restraint-based modeling of 3D bio-molecular structures (16). Briefly, our method uses chromatin interaction frequencies derived from experiments as a proxy of spatial proximity between the ligation products of the 3C libraries. Two fragments of DNA that interact with high frequency are dynamically placed close in space in our models while two fragments that do not interact as often will be kept apart. Our method has been successfully applied to model the structures of genomes and genomic domains in eukaryote and prokaryote organisms (17–19). In all of our studies, the final models were partially validated by assessing their

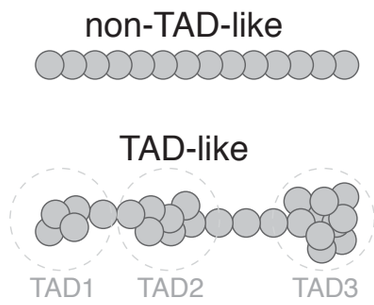
*To whom correspondence should be addressed. Tel: +34 934 020 542; Fax: +34 934 037 279; Email: mmarti@pcb.ub.cat

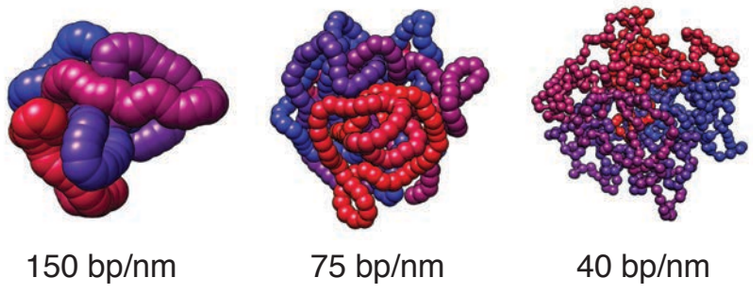
Trussart, et al. (2015). Nucleic Acids Research.

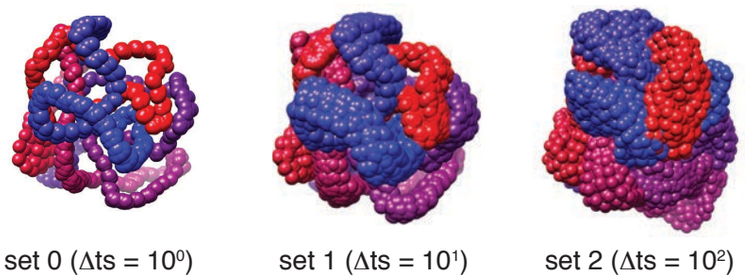
Toy models

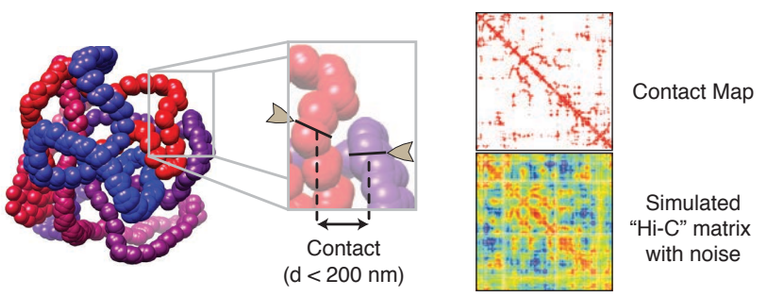








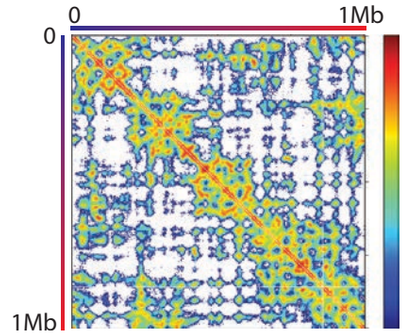


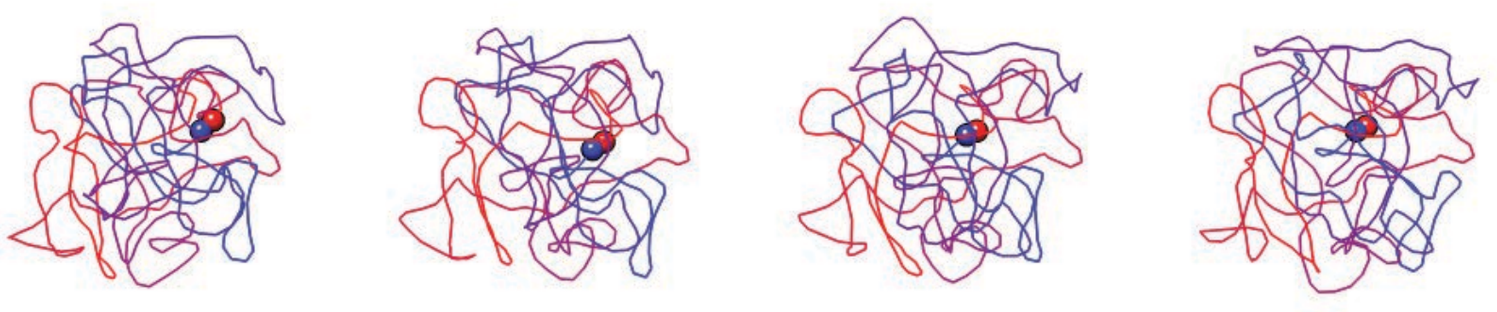


by Ivan Junier

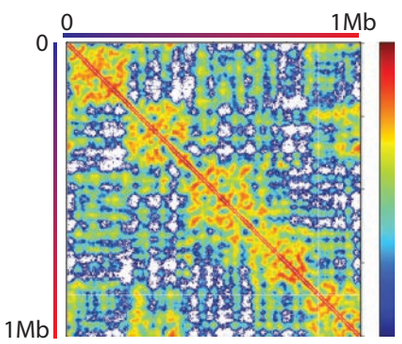
Toy interaction matrices

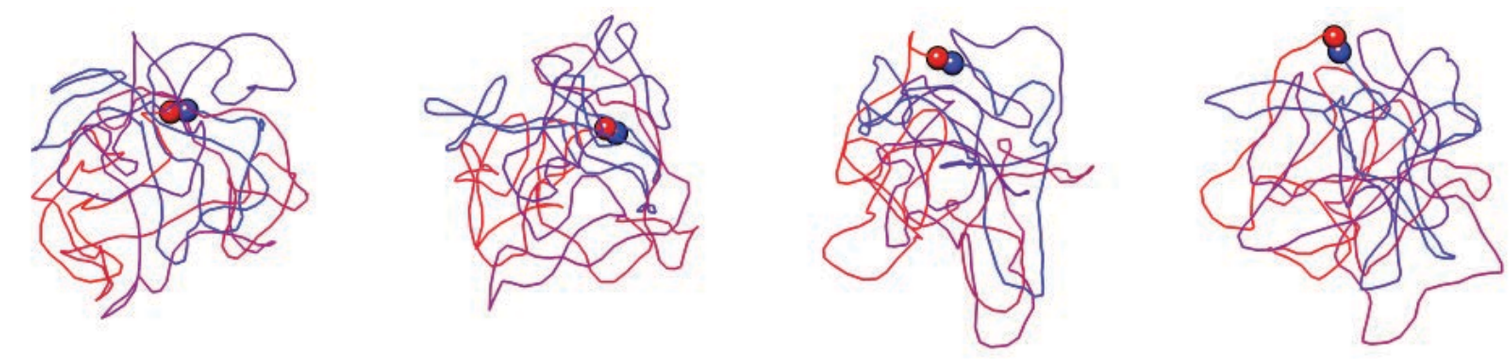
set 0 (Δts=10^0)



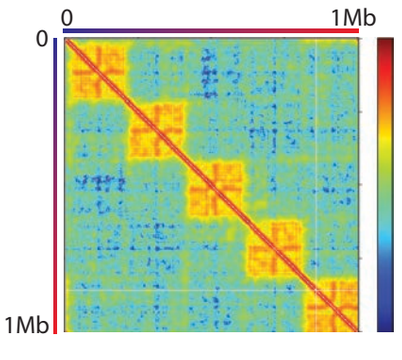


set 4 (Δts=10^4)

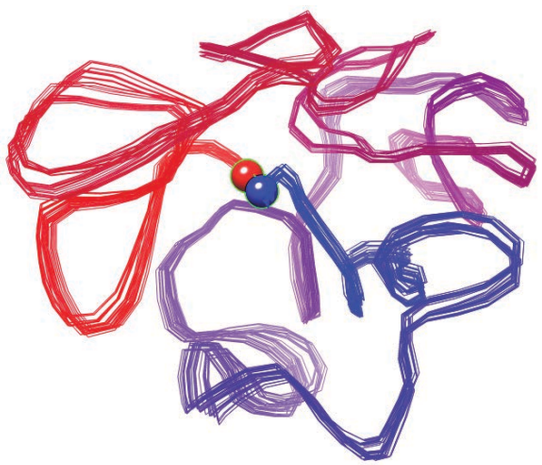


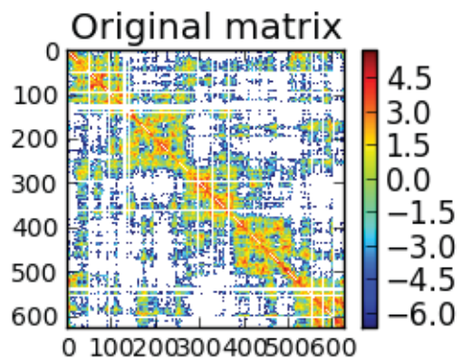


set 6 (Δts=10^6)

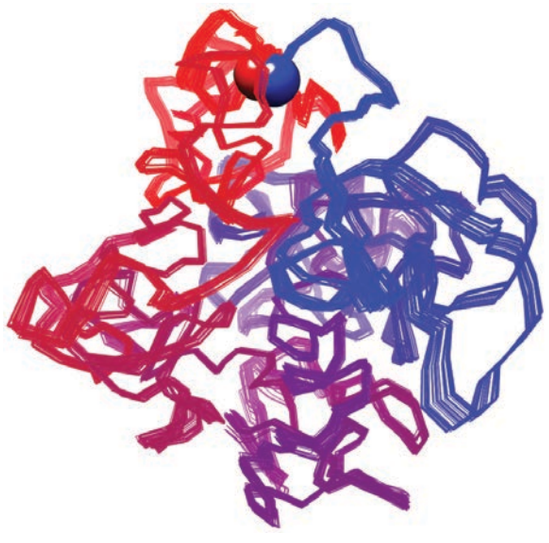


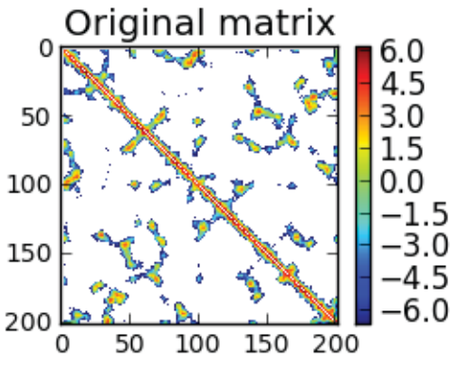
Reconstructing toy models

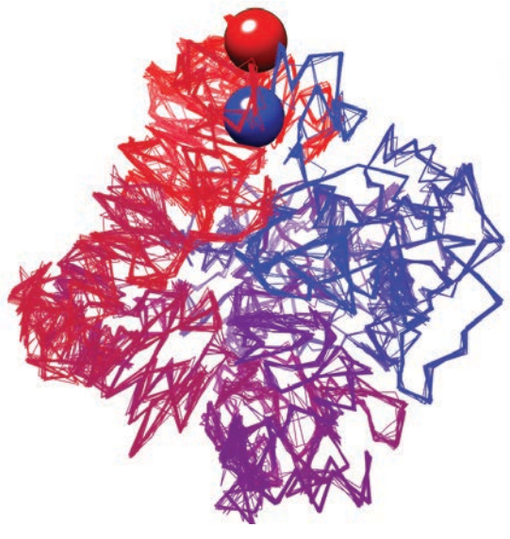




chr40_TAD
α=100
Δts=10
TADbit-SCC: 0.91
<dRMSD>: 32.7 nm
<dSCC>: 0.94

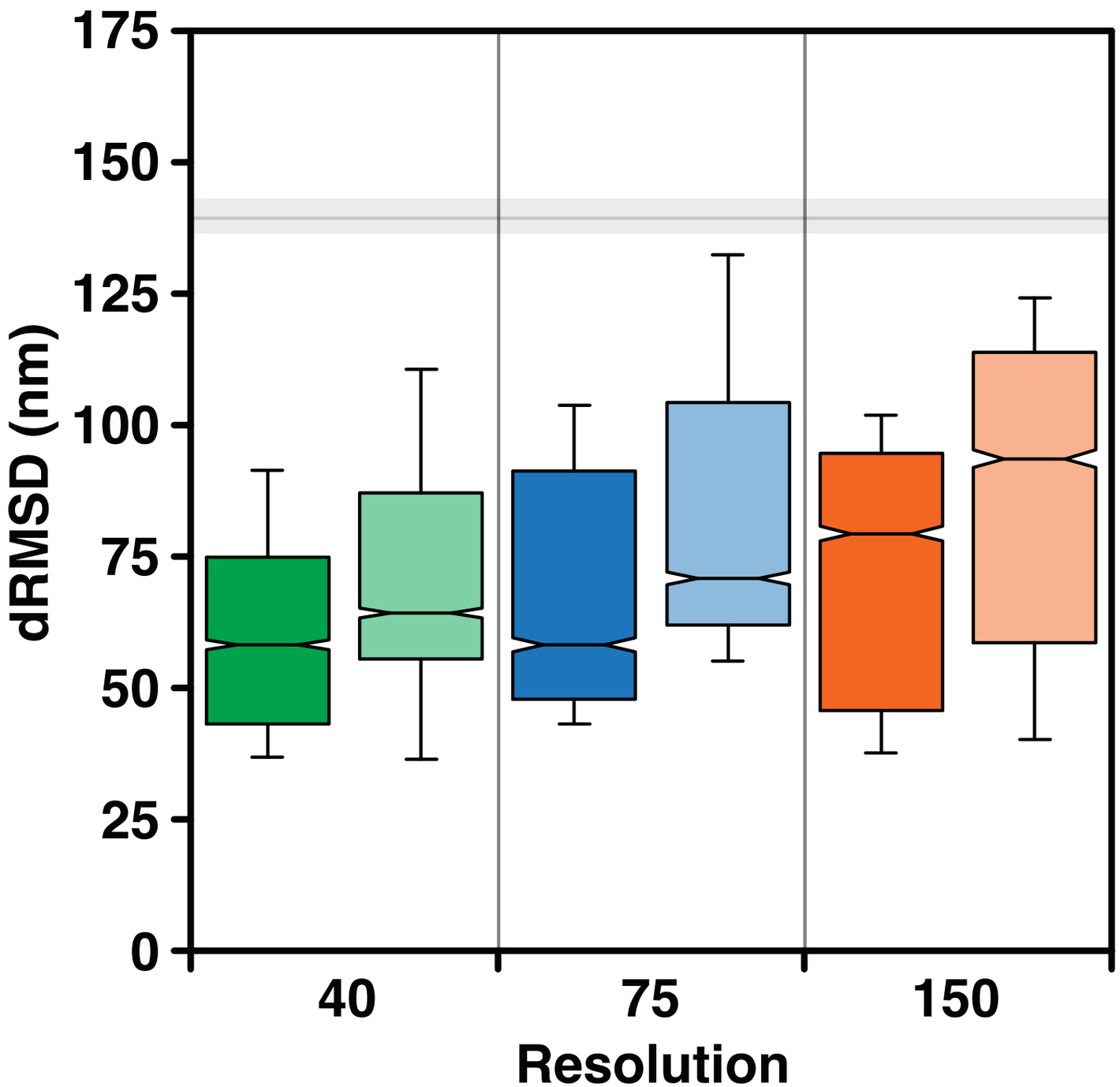




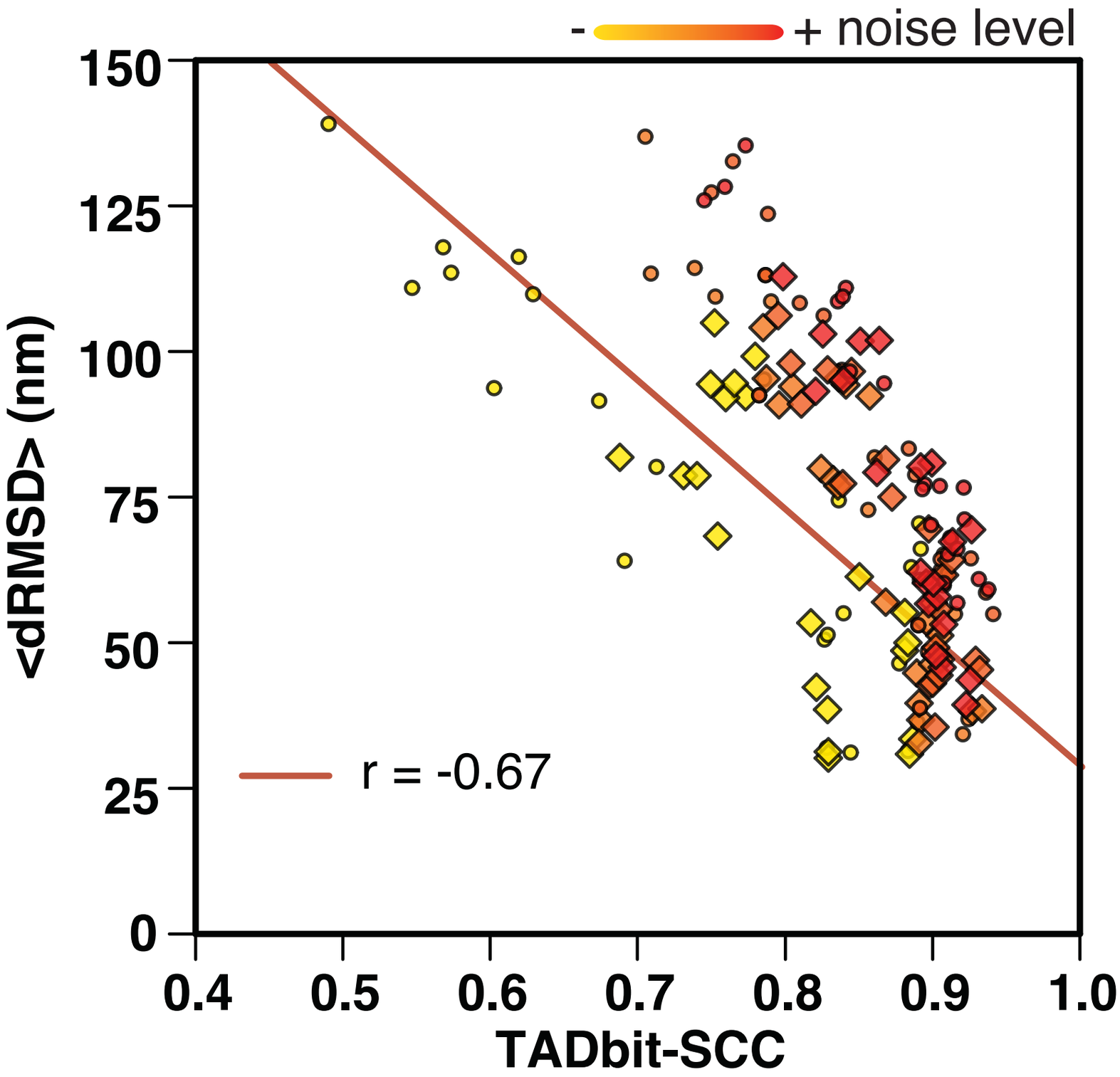


chr150_TAD
α=50
Δts=1
TADbit-SCC: 0.82
<dRMSD>: 45.4 nm
<dSCC>: 0.86

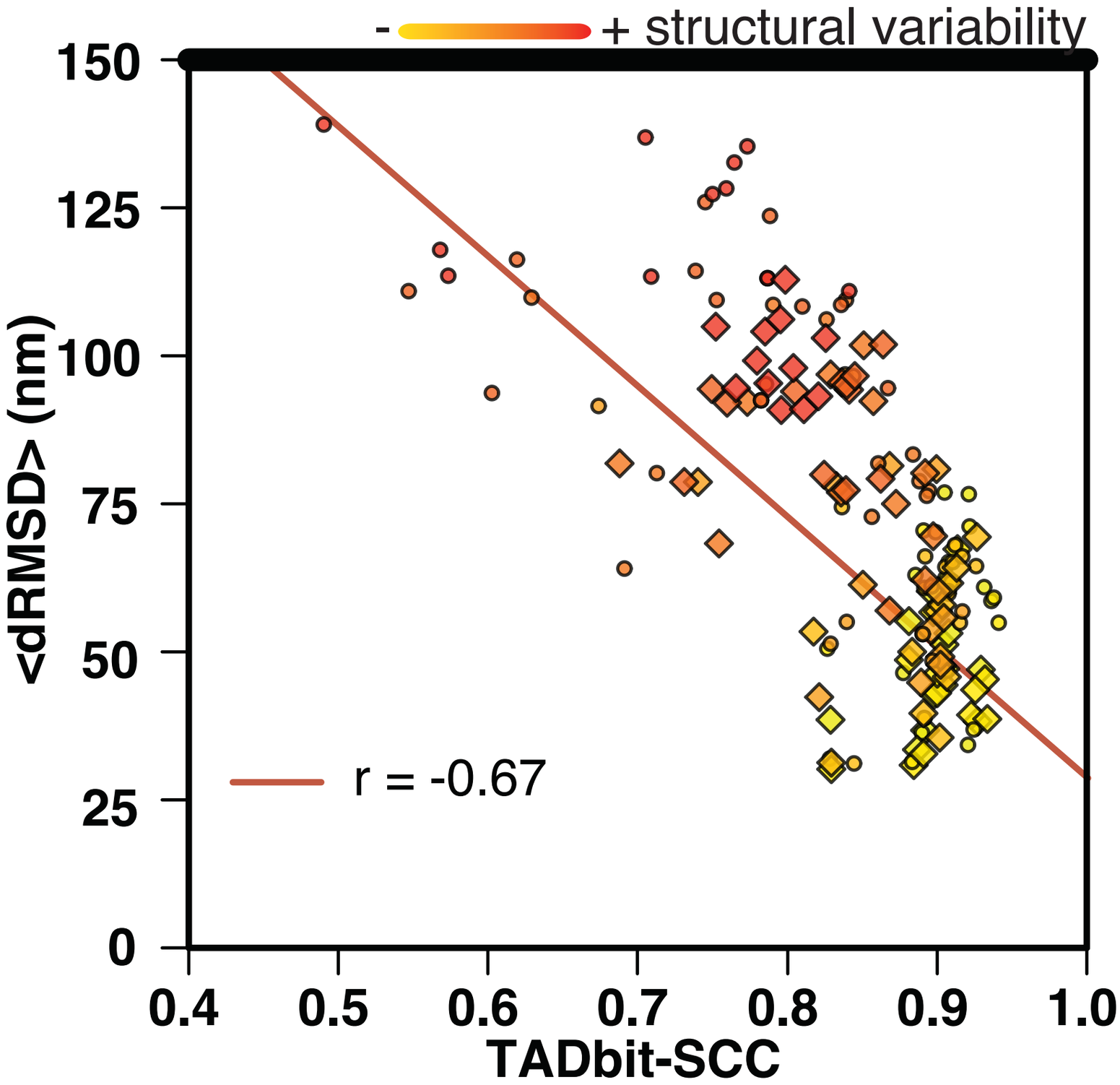
TADs & higher-res are "good"



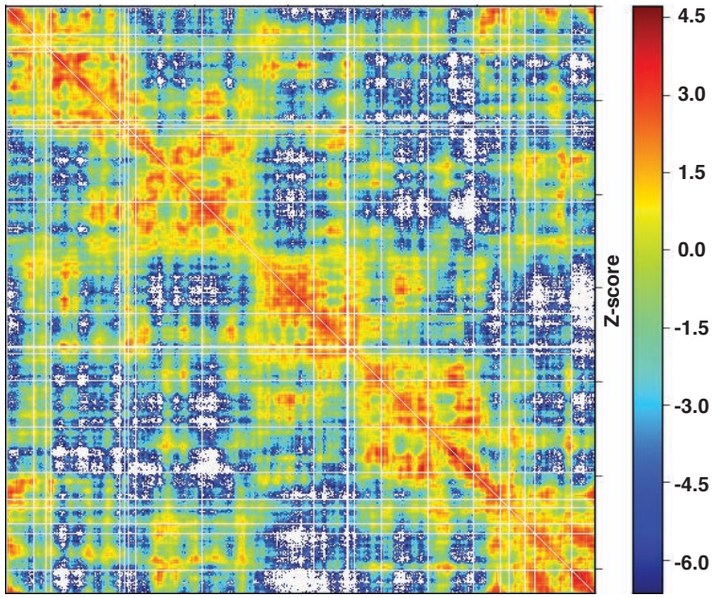
Noise is "OK"


- + noise level
150
125
100
75
50
25
0
<dRMSD> (nm)
r = -0.67
0.4
0.5
0.6
0.7
0.8
0.9
1.0
TADbit-SCC

Structural variability is "NOT OK"

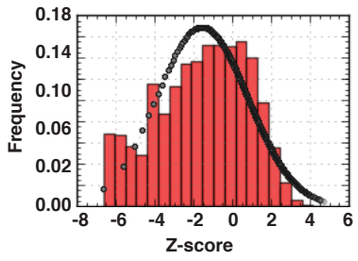


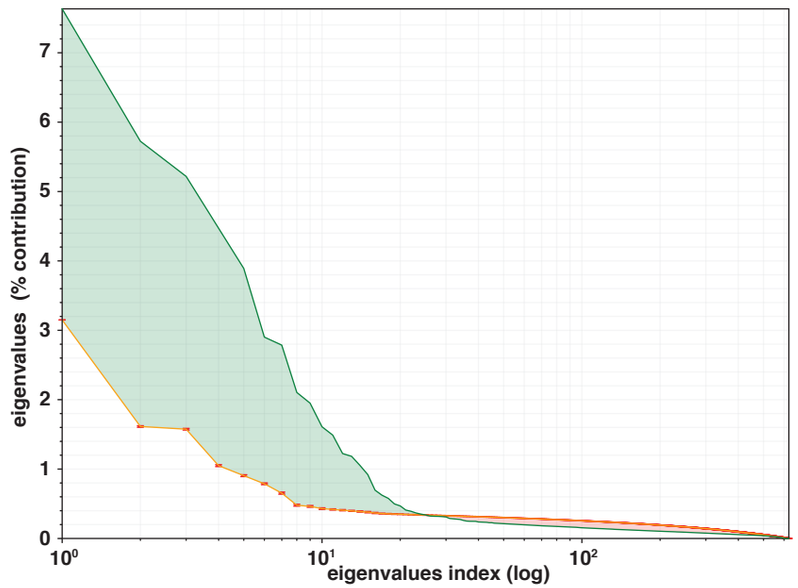
Can we predict the accuracy of the models?

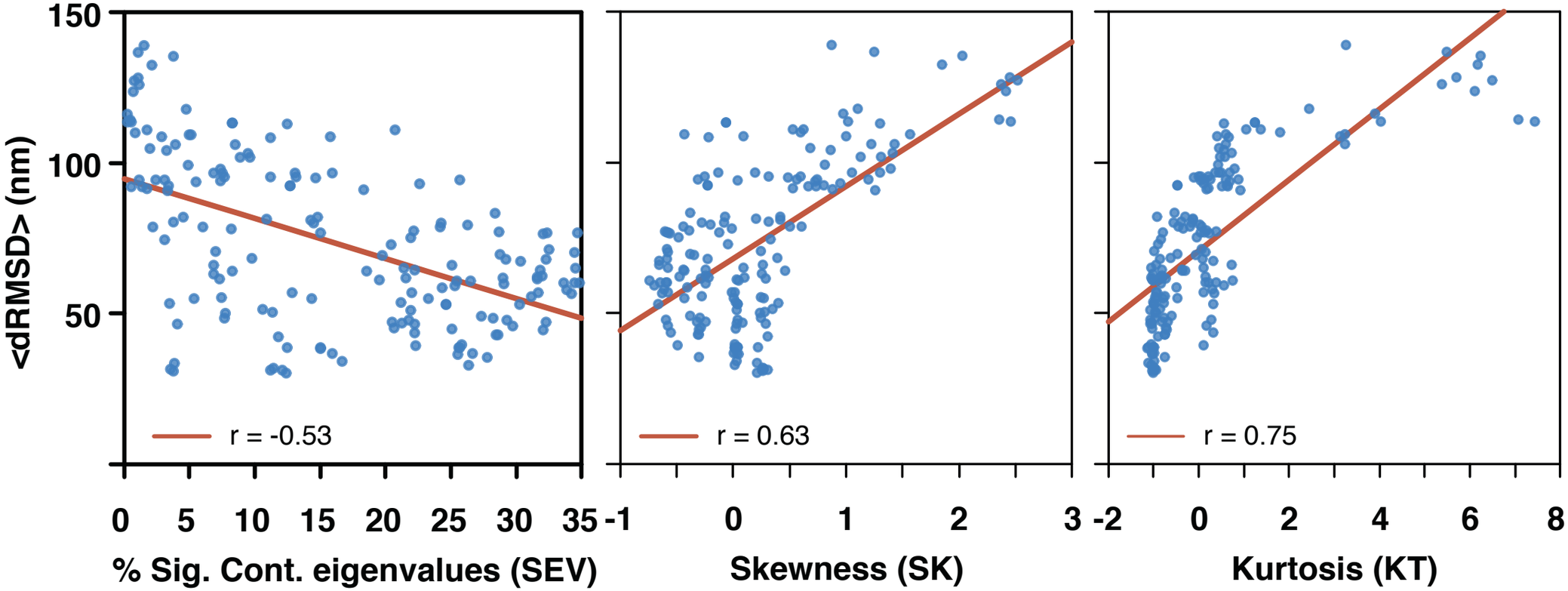


Toy genome: chr40_TAD
Density: 40 bp/nm
TADs: Yes
Noise: 150
Δts: 10^0

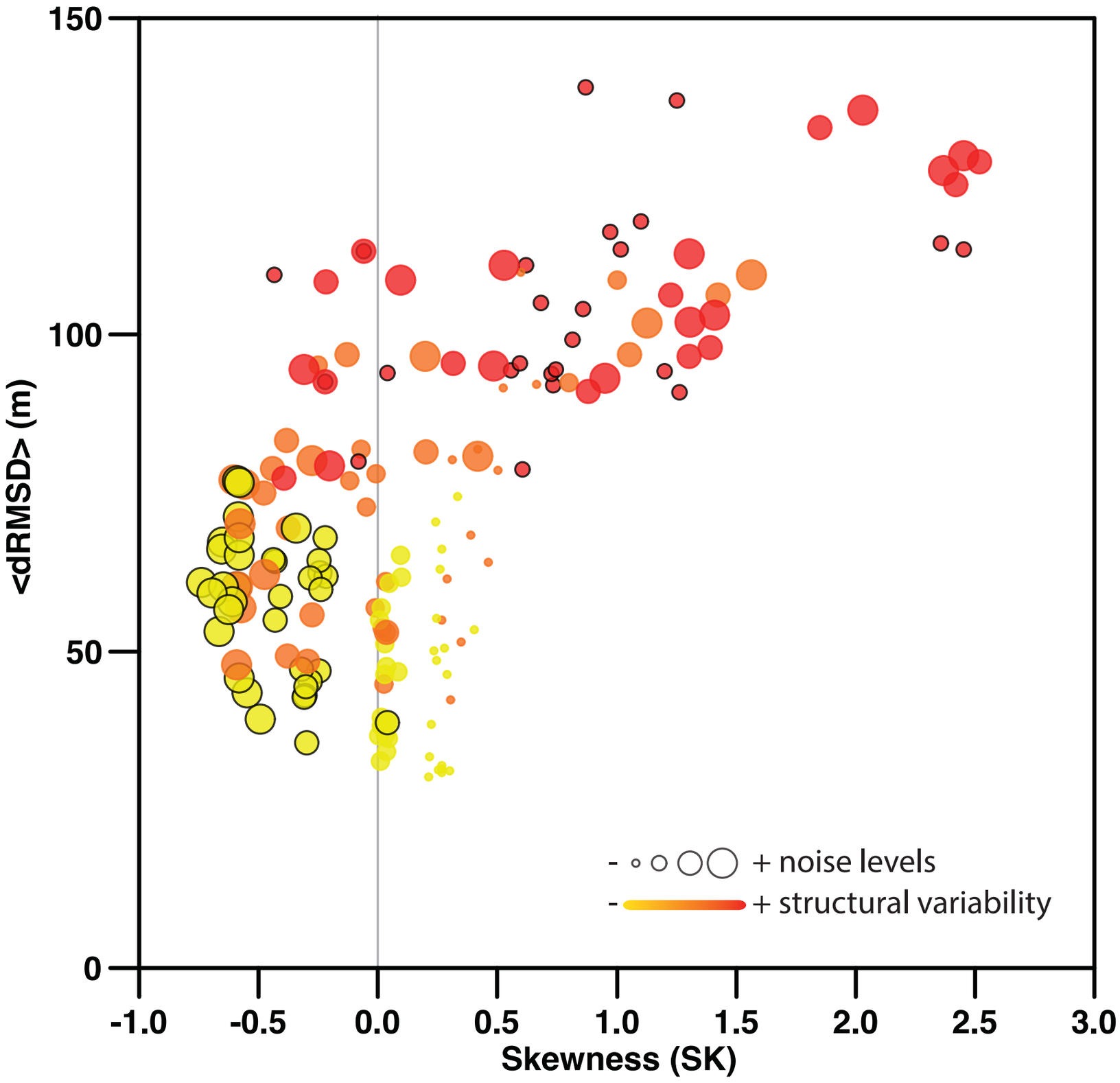
% Sig. Cont. EV: 32.3
Skewness: -0.32
Kurtosis: -0.69





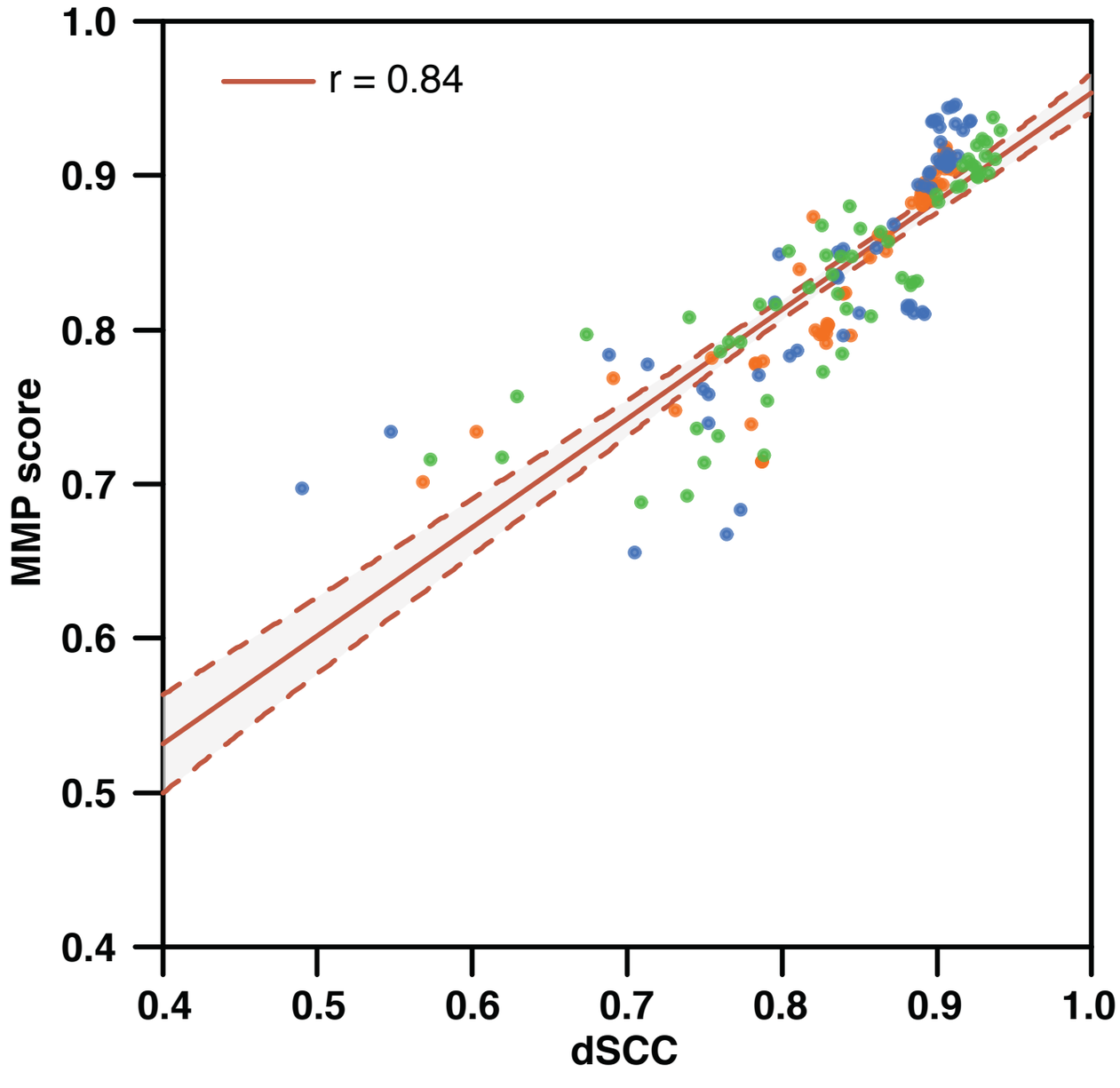


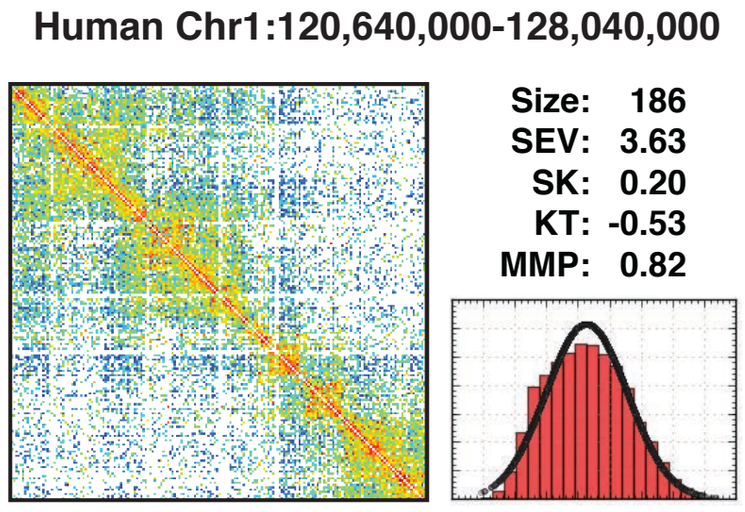
Skewness "side effect"

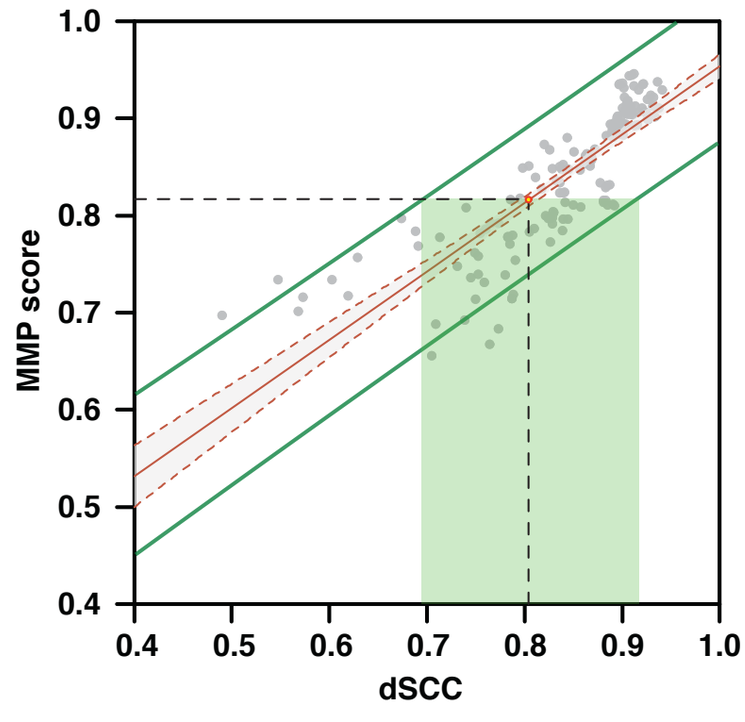


Can we predict the accuracy of the models?

$$\mathrm{MMP} = -0.0002 * \mathrm{Size} + 0.0335 * \mathrm{SK} - 0.0229 * \mathrm{KU} + 0.0069 * \mathrm{SEV} + 0.8126$$







Higher-res is "good"

put your $$ in sequencing

Noise is "OK"

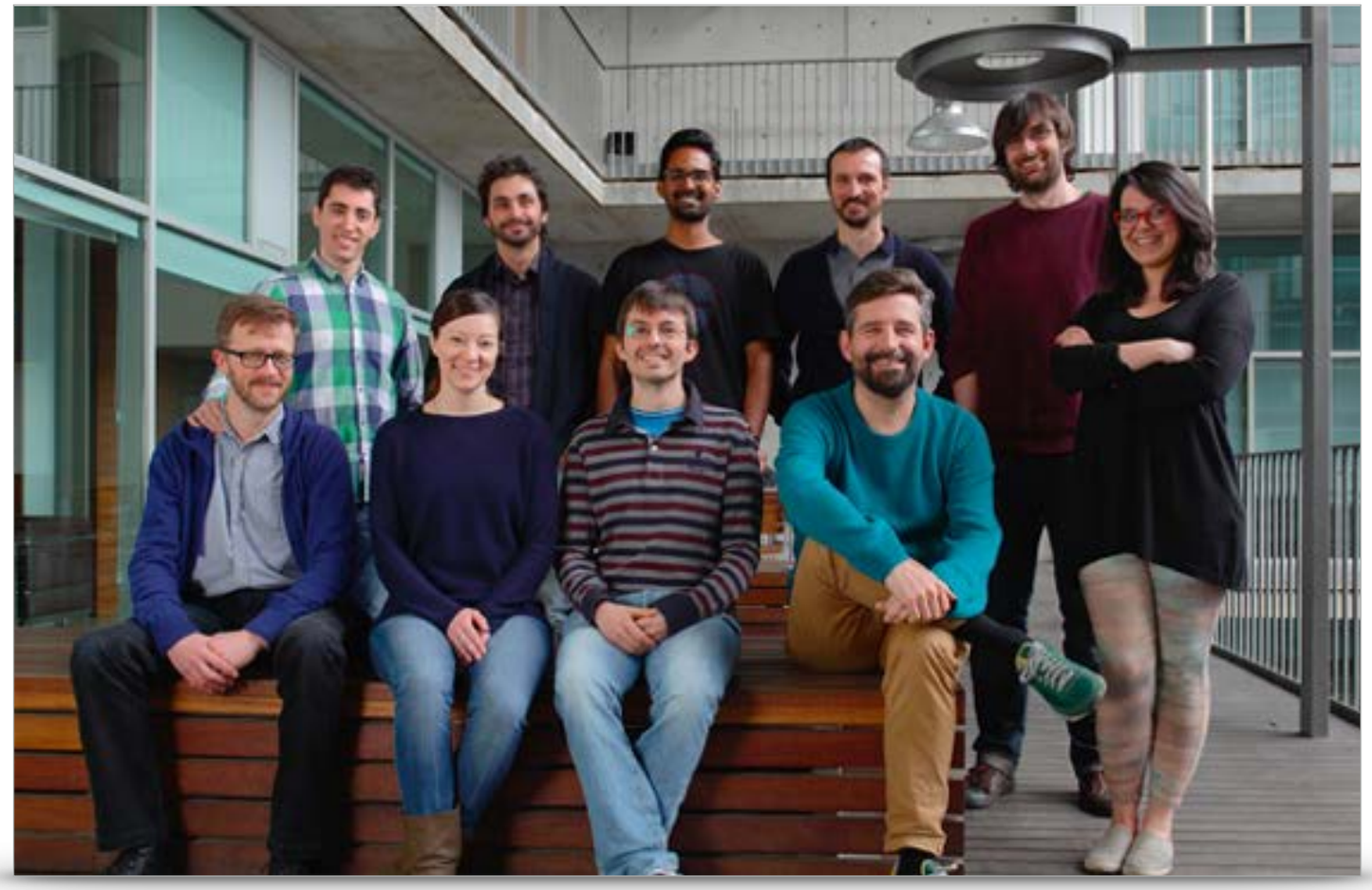
no need to worry much

Structural variability is "NOT OK"

homogenize your cell population!

...but we can differentiate between noise and structural variability

and we can a priori predict the accuracy of the models



Marie Trussart
François Serra
Davide Baù

Gireesh K. Bogu
Yasmina Cuartero
François le Dily
David Dufour
Irene Farabella
Silvia Galan
Mike Goodstadt
Francisco Martínez-Jiménez
Paula Soler
Yannick Spill
Marco di Stefano

in collaboration with Ivan Junier (Université Joseph Fourier) & Luís Serrano (CRG)

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