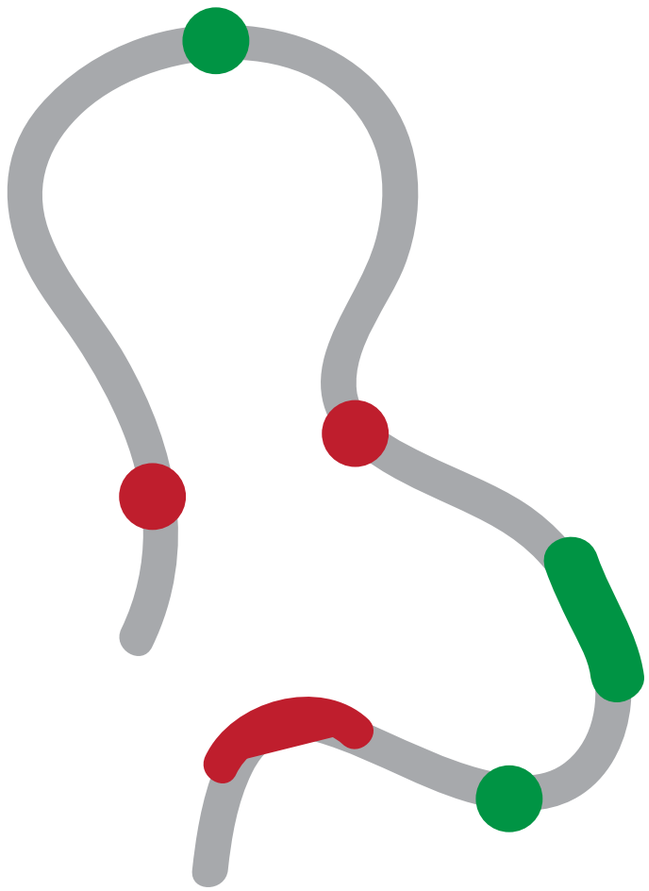
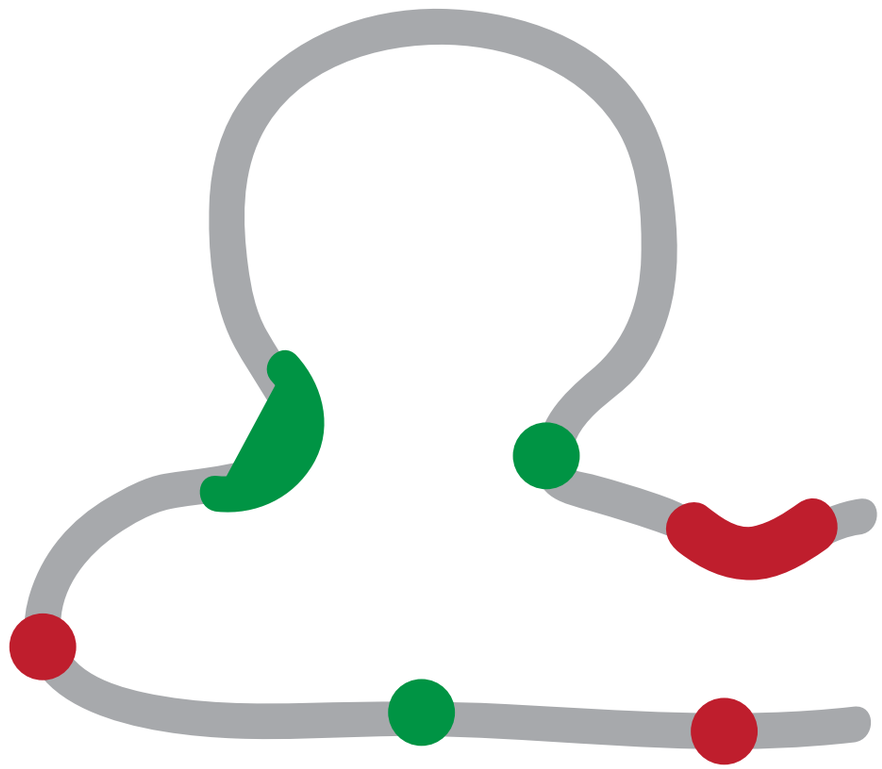
Structure determination of genomes and genomic domains by satisfaction of spatial restraints

Marc A. Marti-Renom
Structural Genomics Group (ICREA, CNAG-CRG)

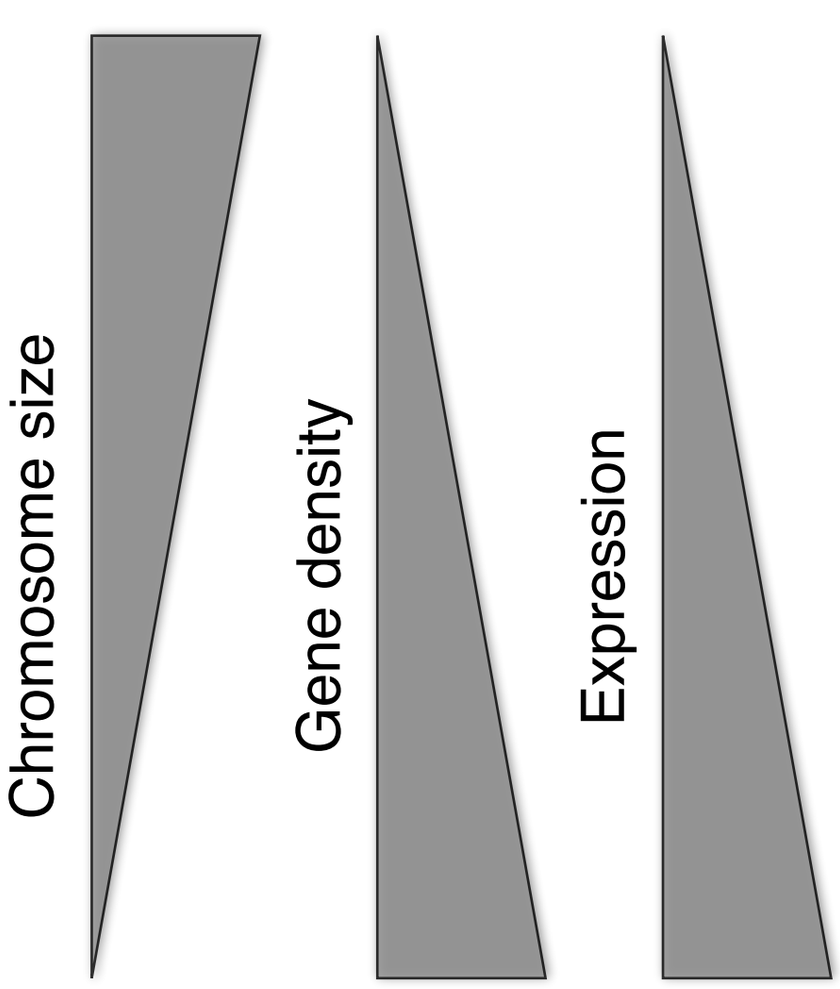
http://marciuslab.org
http://3DGenomes.org
http://cnag.crg.eu

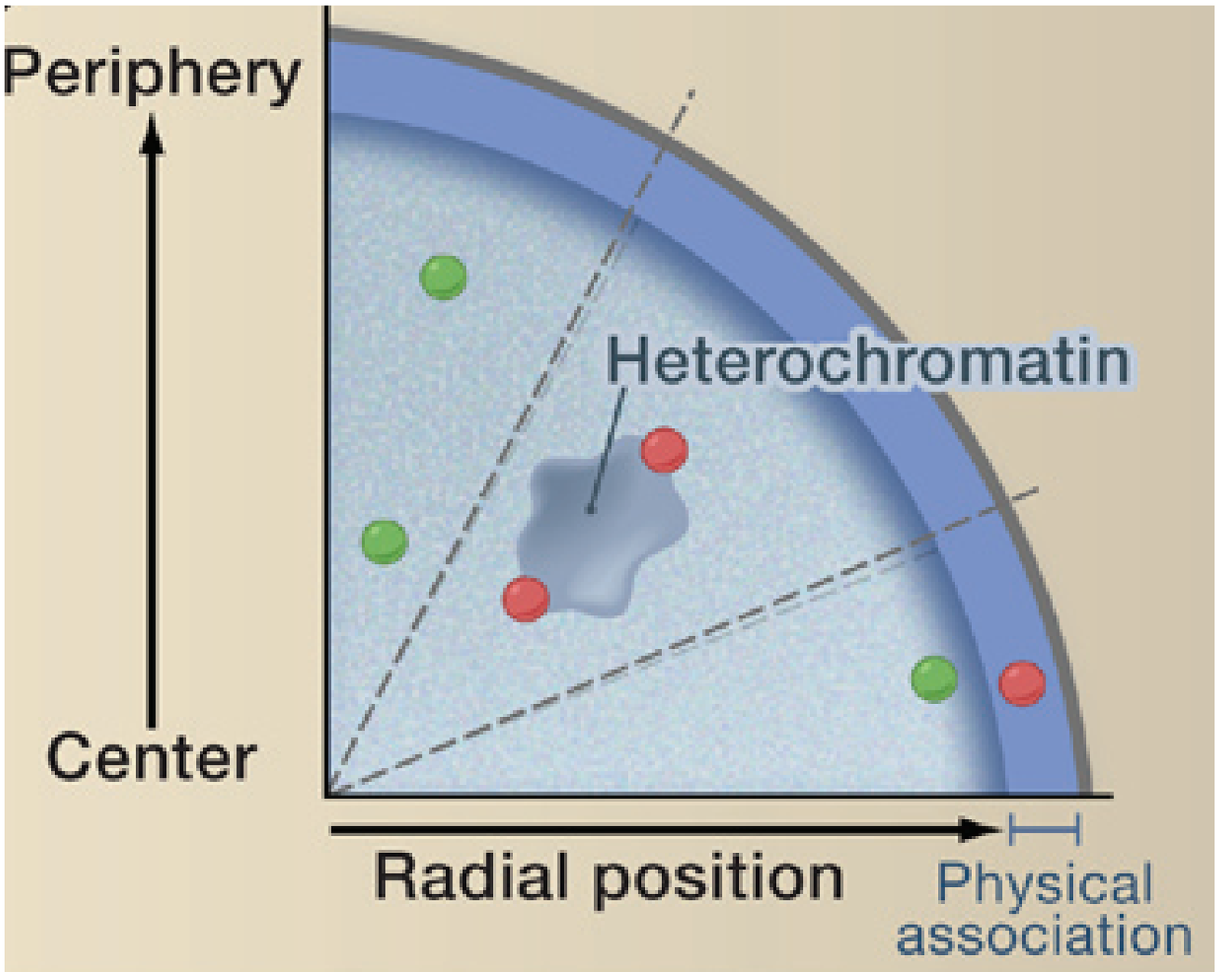




Complex genome organization

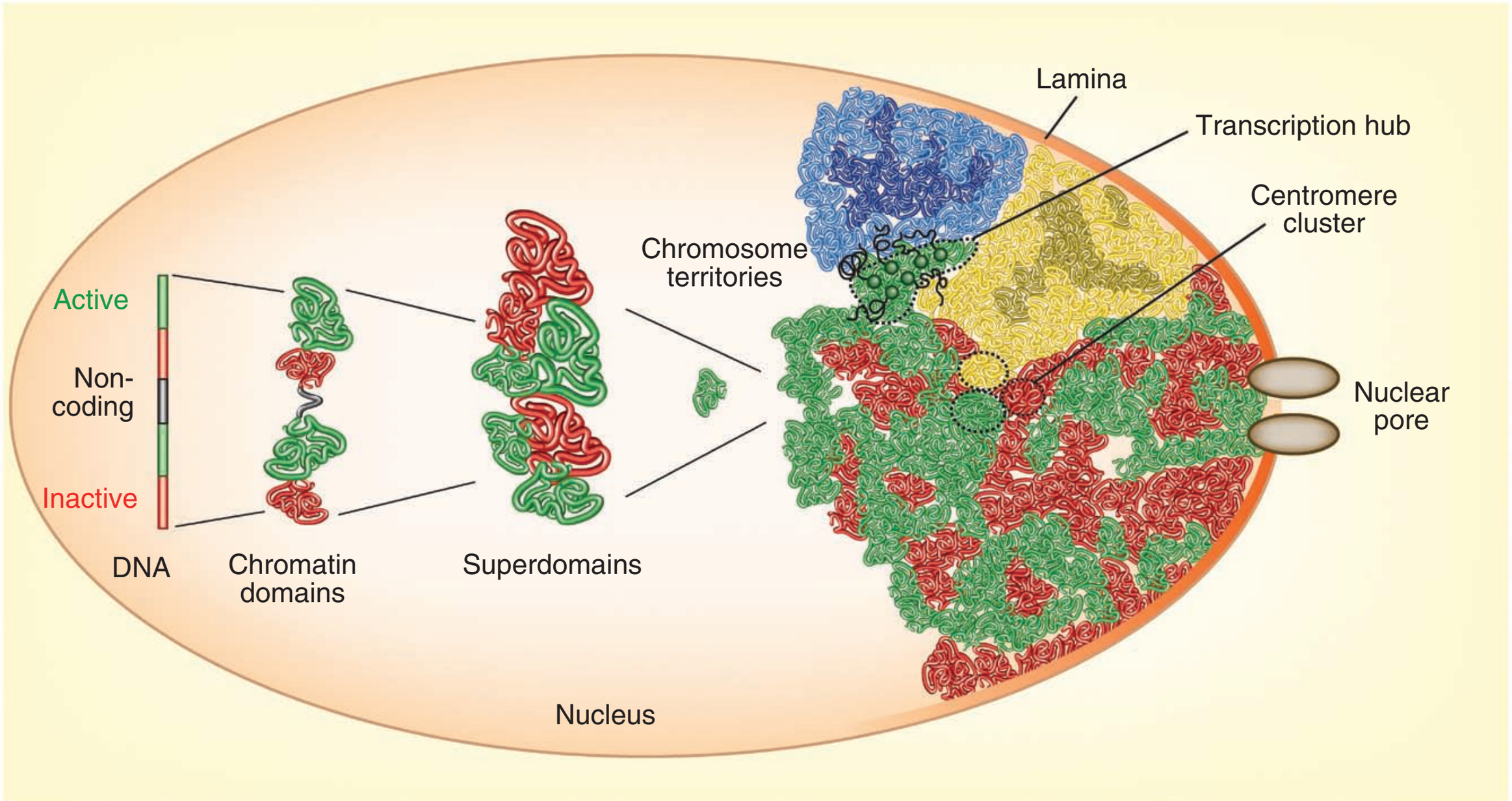
Takizawa, T., Meaburn, K. J. & Misteli, T. The meaning of gene positioning. Cell 135, 9–13 (2008).





Complex genome organization

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

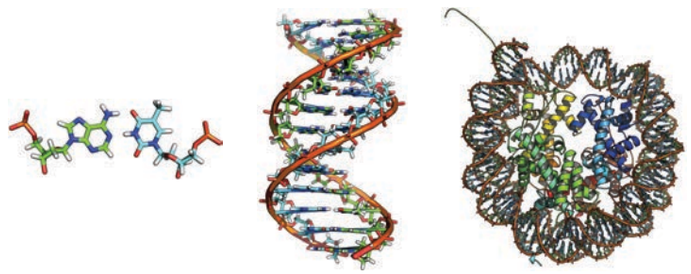


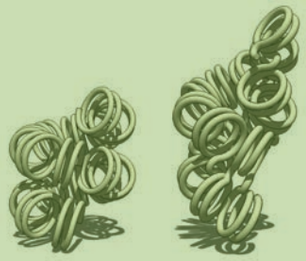
Marina Corral

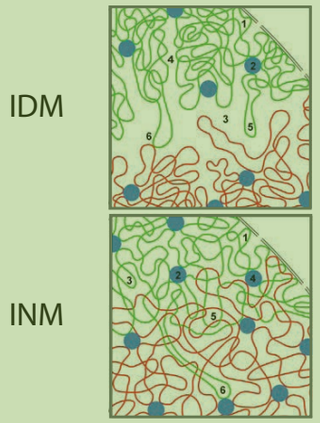
Resolution Gap

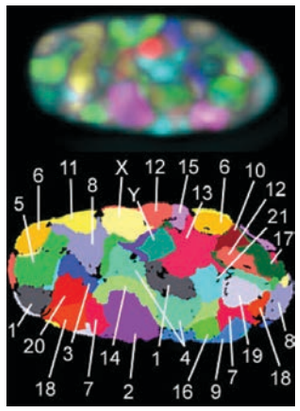
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)

Knowledge









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

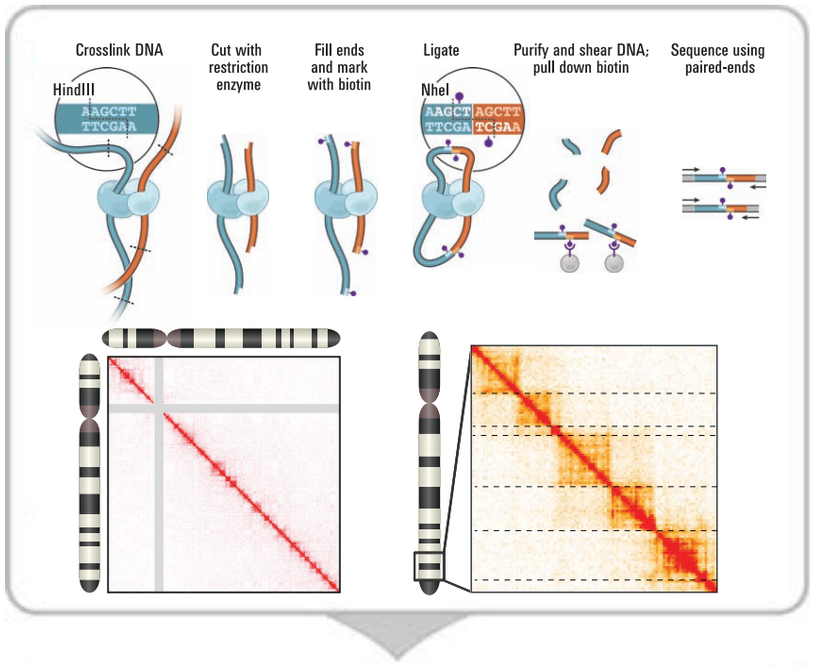
Resolution

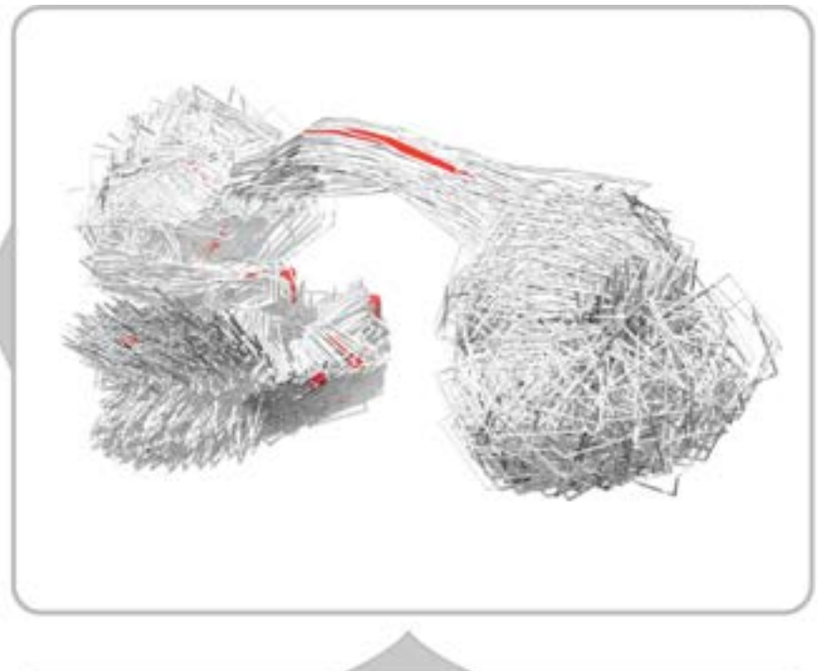
10^{-3} 10^{-2} 10^{-1} μ

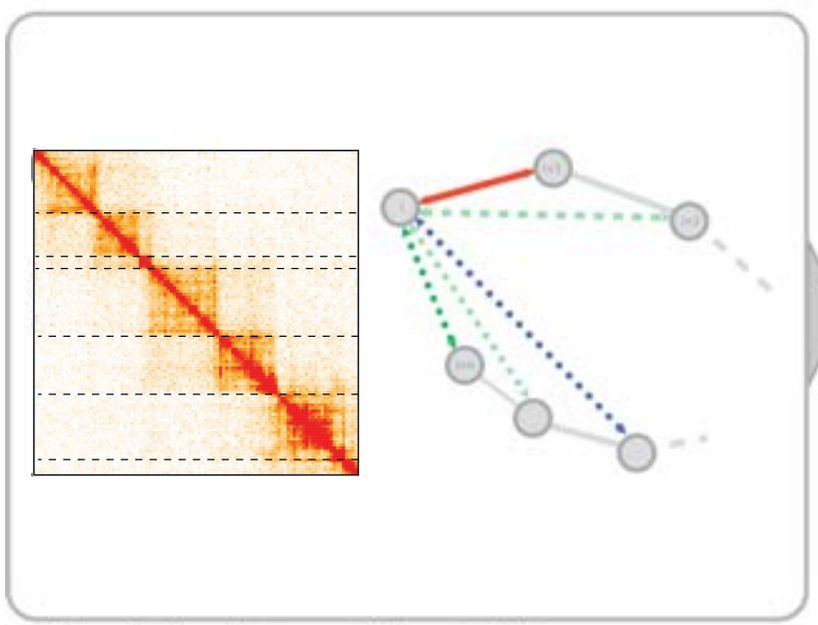
Hybrid Method

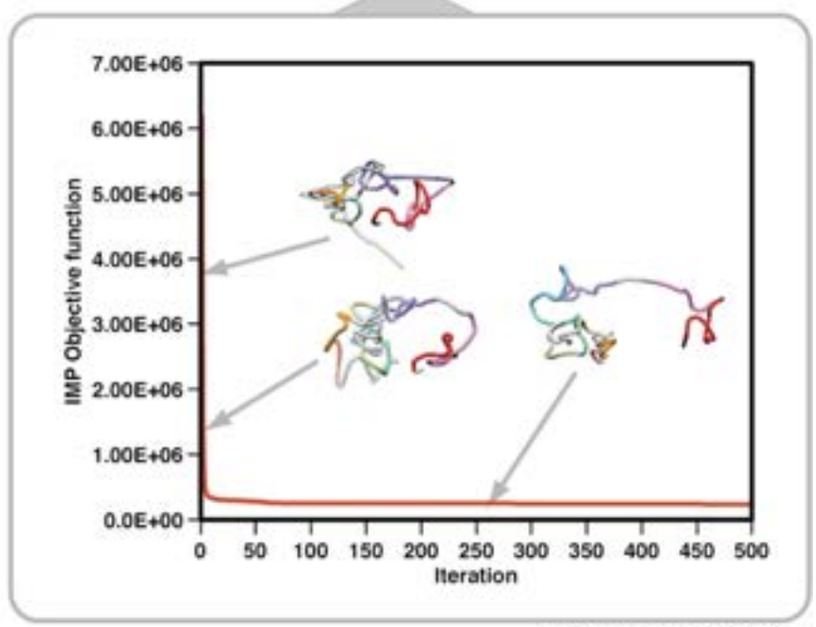
Baù, D. & Marti-Renom, M. A. Methods 58, 300–306 (2012).

Experiments





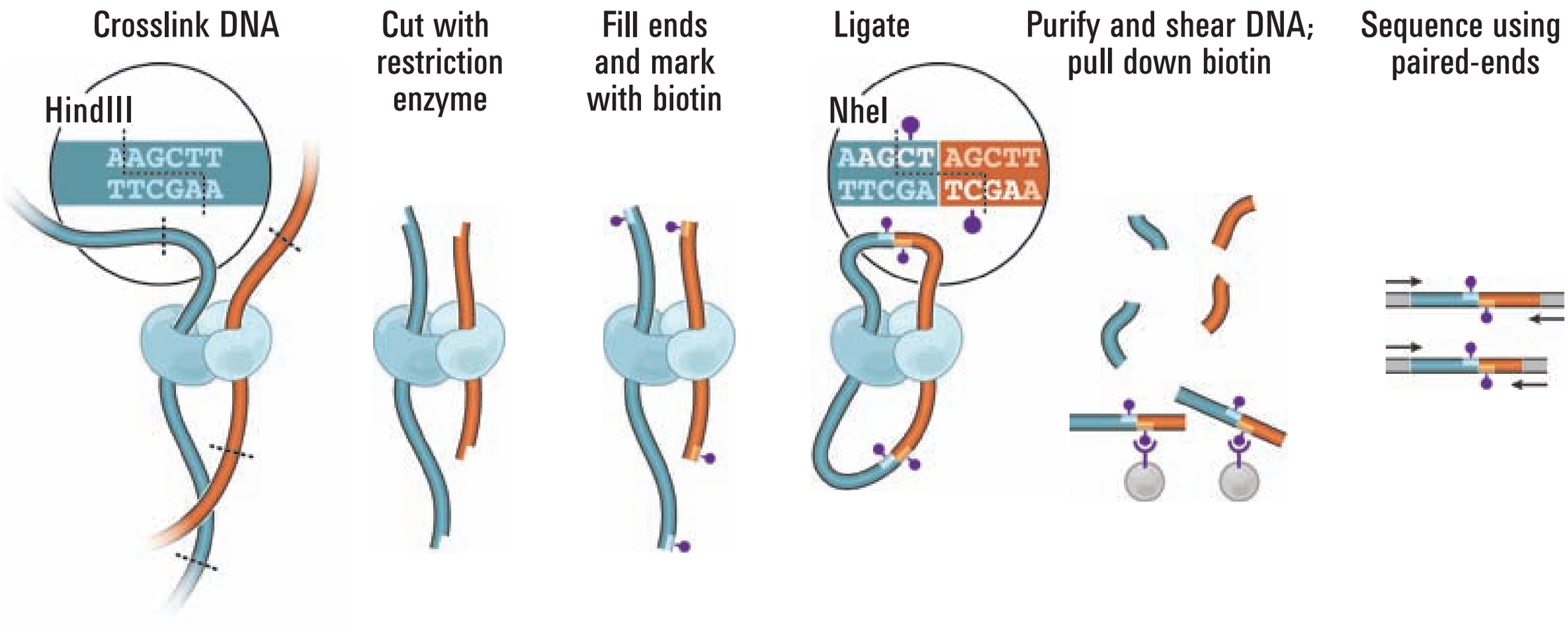




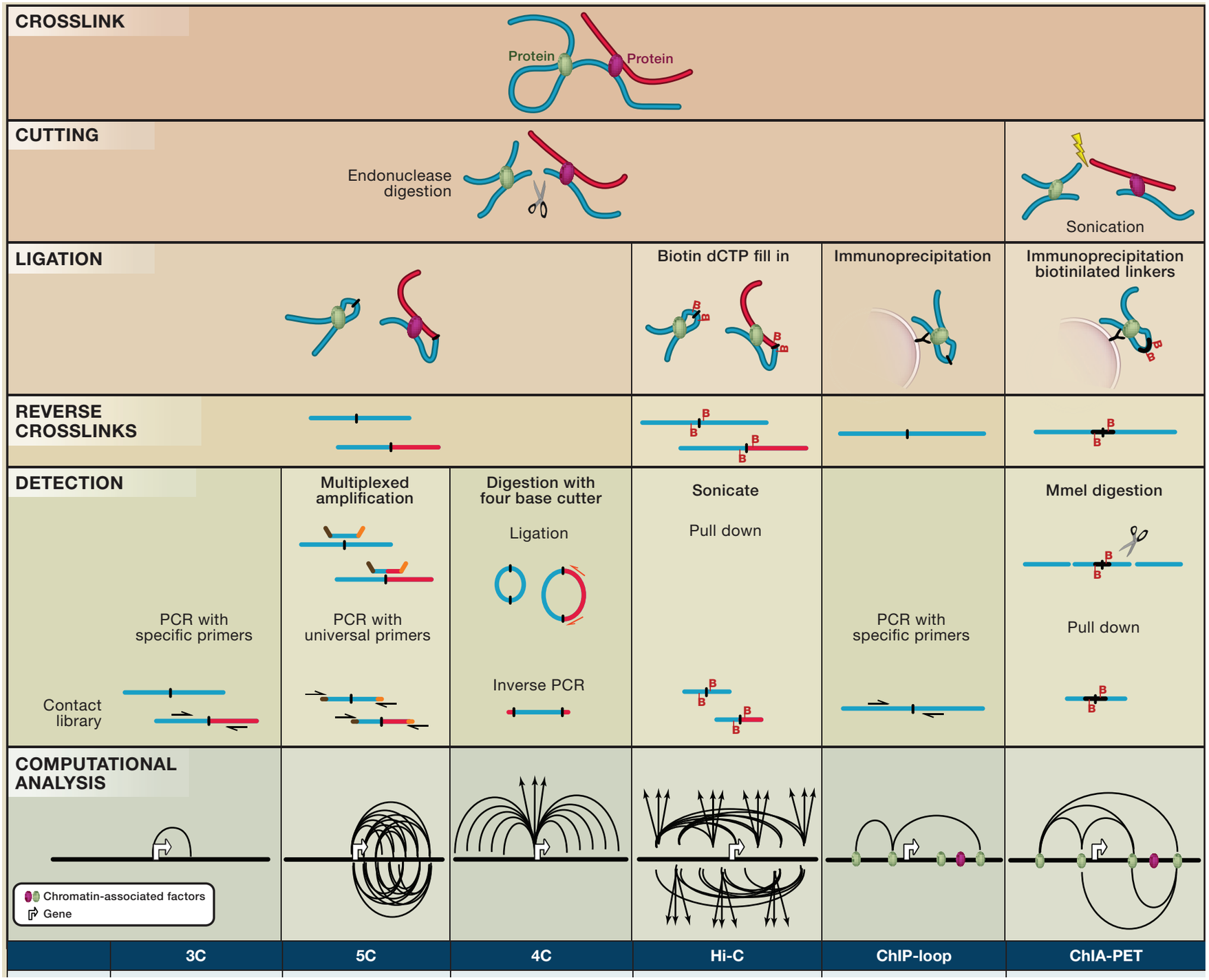
Computation

Chromosome Conformation Capture

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Science, 295(5558), 1306–1311.
Lieberman-Aiden, E., et al. (2009). Science, 326(5950), 289–293.

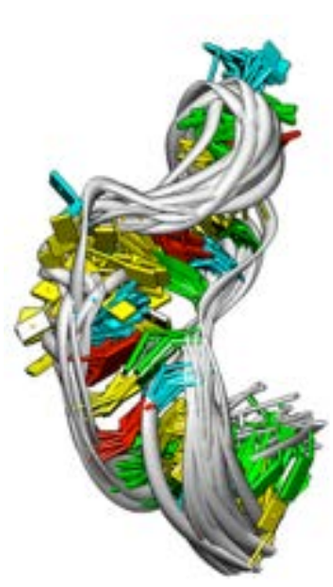


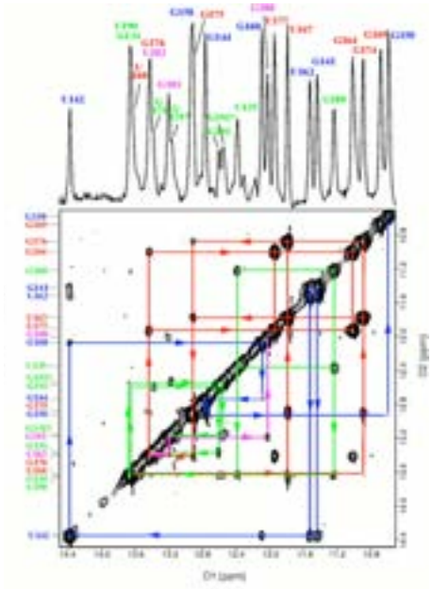
Chromosome Conformation Capture



Restraint-based Modeling

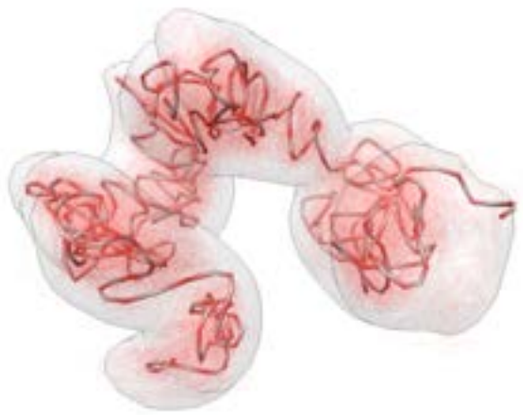
Baù, D. & Marti-Renom, M. A. Methods 58, 300–306 (2012).

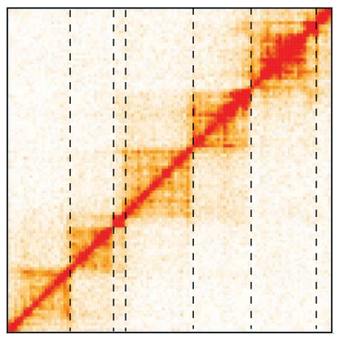




Biomolecular structure determination

2D-NOESY data



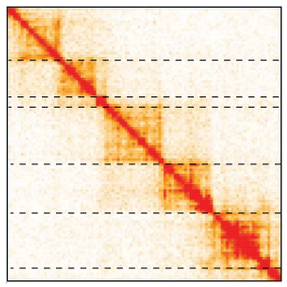


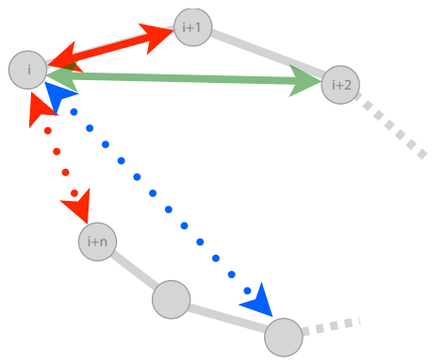
Chromosome structure determination

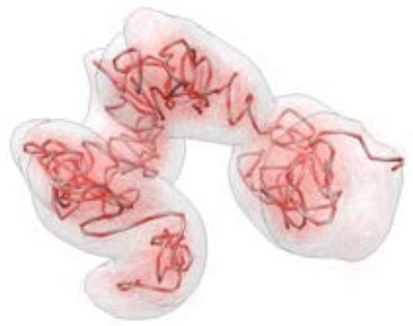
3C-based data

TADbit

http://3DGenomes.org







FastQ files to Maps

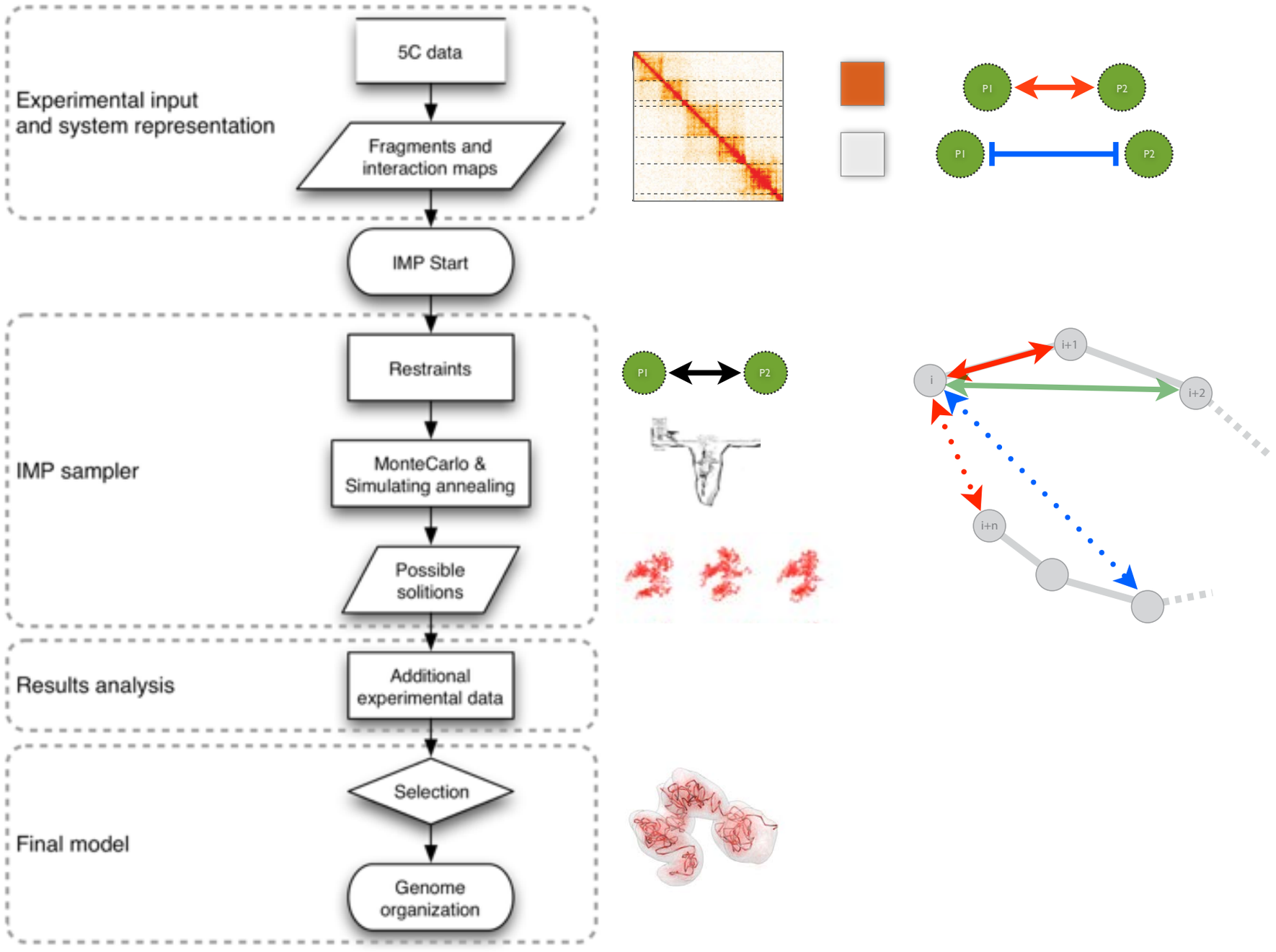
Map analysis

Model building

Model analysis

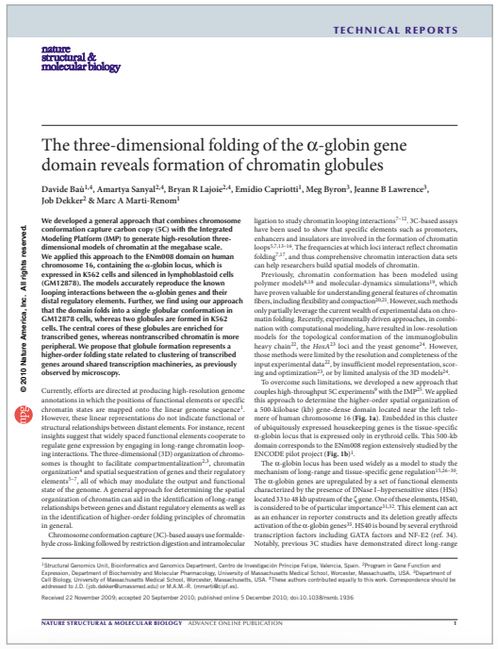
TADbit

http://3DGenomes.org



TADbit previous applications...

Baù, D. et al. Nat Struct Mol Biol (2011)
Umbarger, M. A. et al. Mol Cell (2011)
Le Dily, F. et al. Genes & Dev (2014)



nature structural & molecular biology

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

Davide Baù[1,4], Amartya Sanyal[2,4], Bryan R Lajoie[2,4], Emidio Capriotti[1], Meg Byron[3], Jeanne B Lawrence[3], Job Dekker[2] & Marc A Marti-Renom[1]

We developed a general approach that combines chromosome conformation capture carbon copy (5C) with the Integrated Modeling Platform (IMP) to generate high-resolution three-dimensional models of chromatin at the megabase scale. We applied this approach to the ENm008 domain on human chromosome 16, containing the α-globin locus, which is expressed in K562 cells and silenced in lymphoblastoid cells (GM12878). The models accurately reproduce the known looping interactions between the α-globin genes and their distal regulatory elements. Further, we find using our approach that the domain folds into a single globular conformation in GM12878 cells, whereas two globules are formed in K562 cells. The central cores of these globules are enriched for transcribed genes, whereas nontranscribed chromatin is more peripheral. We propose that globule formation represents a higher-order folding state related to clustering of transcribed genes around shared transcription machineries, as previously observed by microscopy.

Currently, efforts are directed at producing high-resolution genome annotations in which the positions of functional elements or specific chromatin states are mapped onto the linear genome sequence[1]. However, these linear representations do not indicate functional or structural relationships between distant elements. For instance, recent insights suggest that widely spaced functional elements cooperate to regulate gene expression by engaging in long-range chromatin looping interactions. The three-dimensional (3D) organization of chromosomes is thought to facilitate compartmentalization[2,3], chromatin organization[4] and spatial sequestration of genes and their regulatory elements[5–7], all of which may modulate the output and functional state of the genome. A general approach for determining the spatial organization of chromatin can aid in the identification of long-range relationships between genes and distant regulatory elements as well as in the identification of higher-order folding principles of chromatin in general.

Chromosome conformation capture (3C)-based assays use formaldehyde cross-linking followed by restriction digestion and intramolecular ligation to study chromatin looping interactions[7–12]. 3C-based assays have been used to show that specific elements such as promoters, enhancers and insulators are involved in the formation of chromatin loops[5,7,13–16]. The frequencies at which loci interact reflect chromatin folding[7,17], and thus comprehensive chromatin interaction data sets can help researchers build spatial models of chromatin.

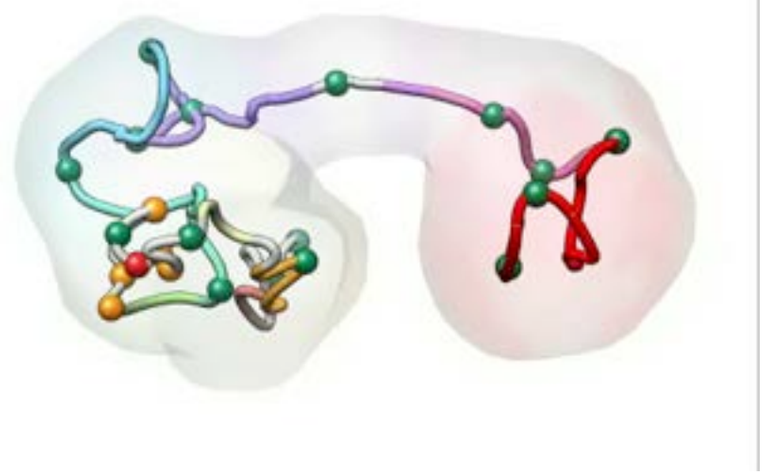
Previously, chromatin conformation has been modeled using polymer models[8,18] and molecular-dynamics simulations[19], which have proven valuable for understanding general features of chromatin fibers, including flexibility and compaction[20,21]. However, such methods only partially leverage the current wealth of experimental data on chromatin folding. Recently, experimentally driven approaches, in combination with computational modeling, have resulted in low-resolution models for the topological conformation of the immunoglobulin heavy chain[22], the *Hox*A[23] loci and the yeast genome[24]. However, those methods were limited by the resolution and completeness of the input experimental data[22], by insufficient model representation, scoring and optimization[23], or by limited analysis of the 3D models[24].

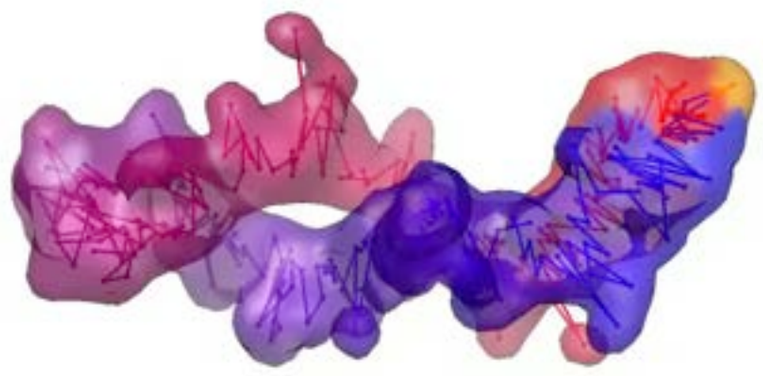
To overcome such limitations, we developed a new approach that couples high-throughput 5C experiments[9] with the IMP[25]. We applied this approach to determine the higher-order spatial organization of a 500-kilobase (kb) gene-dense domain located near the left telomere of human chromosome 16 (**Fig. 1a**). Embedded in this cluster of ubiquitously expressed housekeeping genes is the tissue-specific α-globin locus that is expressed only in erythroid cells. This 500-kb domain corresponds to the ENm008 region extensively studied by the ENCODE pilot project (**Fig. 1b**)[1].

The α-globin locus has been used widely as a model to study the mechanism of long-range and tissue-specific gene regulation[15,26–30]. The α-globin genes are upregulated by a set of functional elements characterized by the presence of DNase I–hypersensitive sites (HSs) located 33 to 48 kb upstream of the ζ gene. One of these elements, HS40, is considered to be of particular importance[31,32]. This element can act as an enhancer in reporter constructs and its deletion greatly affects activation of the α-globin genes[33]. HS40 is bound by several erythroid transcription factors including GATA factors and NF-E2 (ref. 34). Notably, previous 3C studies have demonstrated direct long-range

[1]Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain. [2]Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [3]Department of Cell Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to J.D. (job.dekker@umassmed.edu) or M.A.M.-R. (mmarti@cipf.es).

Received 22 November 2009; accepted 20 September 2010; published online 5 December 2010; doi:10.1038/nsmb.1936





Cell PRESS

Molecular Cell
Article

The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation

Mark A. Umbarger,[1,6,*] Esteban Toro,[2,6] Matthew A. Wright,[1] Gregory J. Porreca,[1] Davide Baù,[4] Sun-Hae Hong,[2,3] Michael J. Fero,[2] Lihua J. Zhu,[5] Marc A. Marti-Renom,[4,*] Harley H. McAdams,[2] Lucy Shapiro,[2] Job Dekker,[5,6,7,*] and George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2]Department of Developmental Biology, School of Medicine
[3]Department of Physics, School of Humanities and Sciences
Stanford University, Stanford, CA 94305, USA
[4]Structural Genomics Laboratory, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain
[5]Program in Gene Function and Expression
[6]Department of Biochemistry and Molecular Pharmacology
[7]Program in Systems Biology
University of Massachusetts Medical School, Worcester, MA 01605, USA
[6]These authors contributed equally to this work
*Correspondence: umbarger@post.harvard.edu (M.A.U.), mmarti@cipf.es (M.A.M.-R.), job.dekker@umassmed.edu (J.D.)
DOI 10.1016/j.molcel.2011.09.010

SUMMARY

We have determined the three-dimensional (3D) architecture of the *Caulobacter crescentus* genome by combining genome-wide chromatin interaction detection, live-cell imaging, and computational modeling. Using chromosome conformation capture carbon copy (5C), we derive ~13 kb resolution 3D models of the *Caulobacter* genome. The resulting models illustrate that the genome is ellipsoidal with periodically arranged arms. The *parS* sites, a pair of short contiguous sequence elements known to be involved in chromosome segregation, are positioned at one pole, where they anchor the chromosome to the cell and contribute to the formation of a compact chromatin conformation. Repositioning these elements resulted in rotations of the chromosome that changed the subcellular positions of most genes. Such rotations did not lead to large-scale changes in gene expression, indicating that genome folding does not strongly affect gene regulation. Collectively, our data suggest that genome folding is globally dictated by the *parS* sites and chromosome segregation.

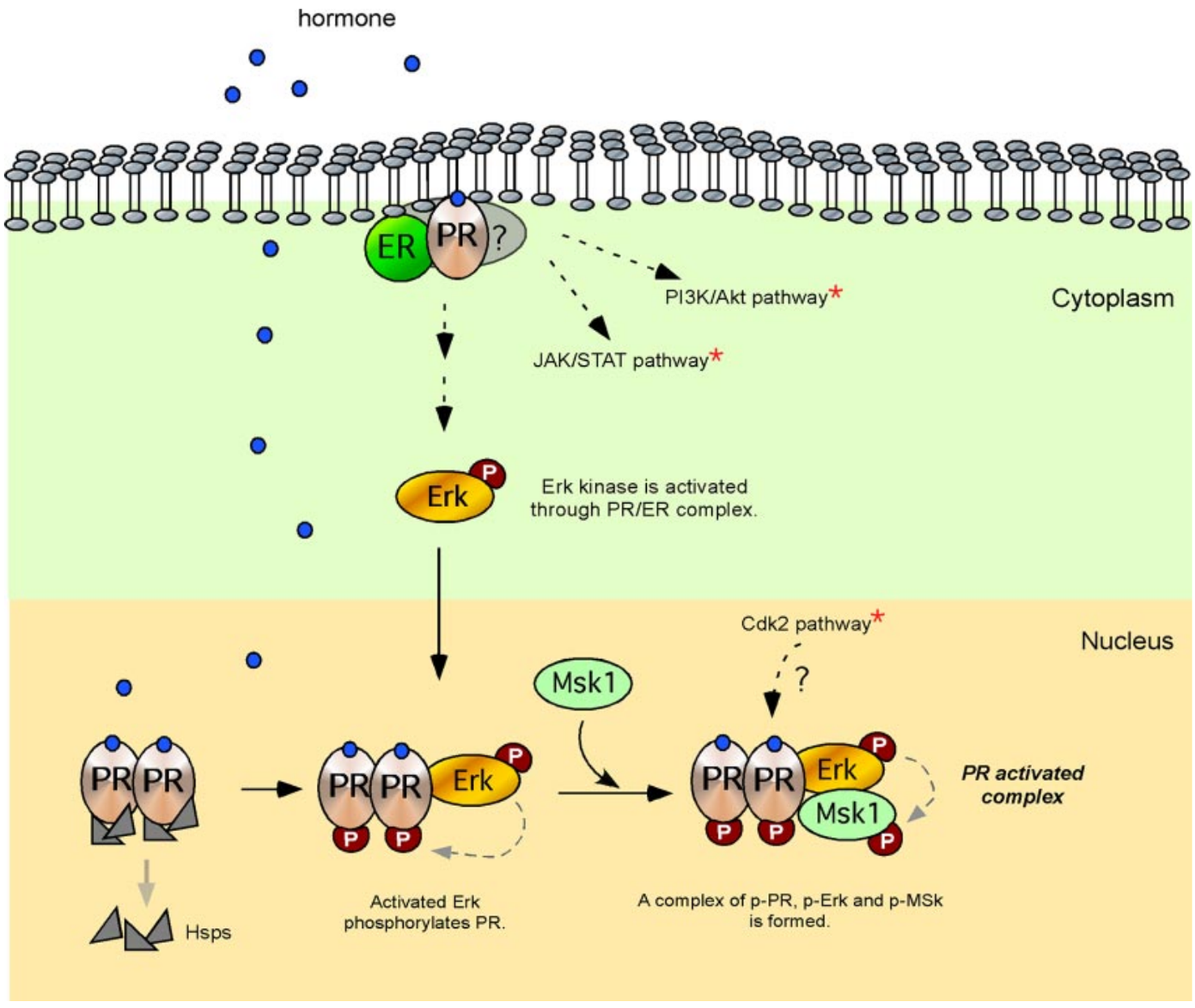
INTRODUCTION

The three-dimensional (3D) architecture of the genome both reflects and regulates its functional state (Dekker, 2008; Thanbichler and Shapiro, 2006a). For example, chromosome segregation impacts bacterial locus subcellular positioning (Jun and Mulder, 2006; White et al., 2008), and chromatin loops that place promoters and distant enhancers within close spatial proximity play important roles in eukaryotic transcriptional regulation (Tolhuis et al., 2002; Vernimmen et al., 2007). Such examples suggest that studies of the high-resolution folding of genomes will yield insight into genome biology. However, until recently such studies, which require comprehensive assessments of the spatial positioning of many loci, have represented major technical challenges.

The recent development of several high-throughput technologies, including automated fluorescent imaging (Viollier et al., 2004) and chromosome conformation capture (3C)-based approaches (Dekker et al., 2002; Dostie et al., 2006; Duan et al., 2010; Fullwood et al., 2009; Lieberman-Aiden et al., 2009; Simonis et al., 2006; Zhao et al., 2006), has begun to enable studies of genome-wide chromosome folding. Fluorescent microscopy-based approaches allow the accurate determination of the subcellular positions of increasing numbers of defined chromosomal loci, while high-throughput 3C-based approaches enable quantification of interloci interaction frequencies that can subsequently be used to infer the average 3D distances between these loci. Studies utilizing one or both of these approaches have highlighted the potential of genome-wide studies of chromosome structure and have begun to reveal specific features of chromosome folding, including the transcription-based compartmentalization of the human nucleus (Lieberman-Aiden et al., 2009; Simonis et al., 2006) and the correlation between a locus' genomic and subcellular positioning in bacteria (Nielsen et al., 2006; Teleman et al., 1998; Wang et al., 2006b). However, the detailed structures of genomes are only beginning to be revealed, and many details, including the identities of the sequence elements that define such structures, await further elucidation.

We sought to determine the high-resolution 3D structure of an entire genome and to utilize the resulting structure to identify the sequence elements that define its architecture. Toward this goal, we studied the synchronizable bacterium, *Caulobacter crescentus* (hereafter Caulobacter), whose single circular chromosome is organized such that the origin and terminus of replication reside near opposite poles of the cell and other loci lie

Progesterone-regulated transcription in breast cancer



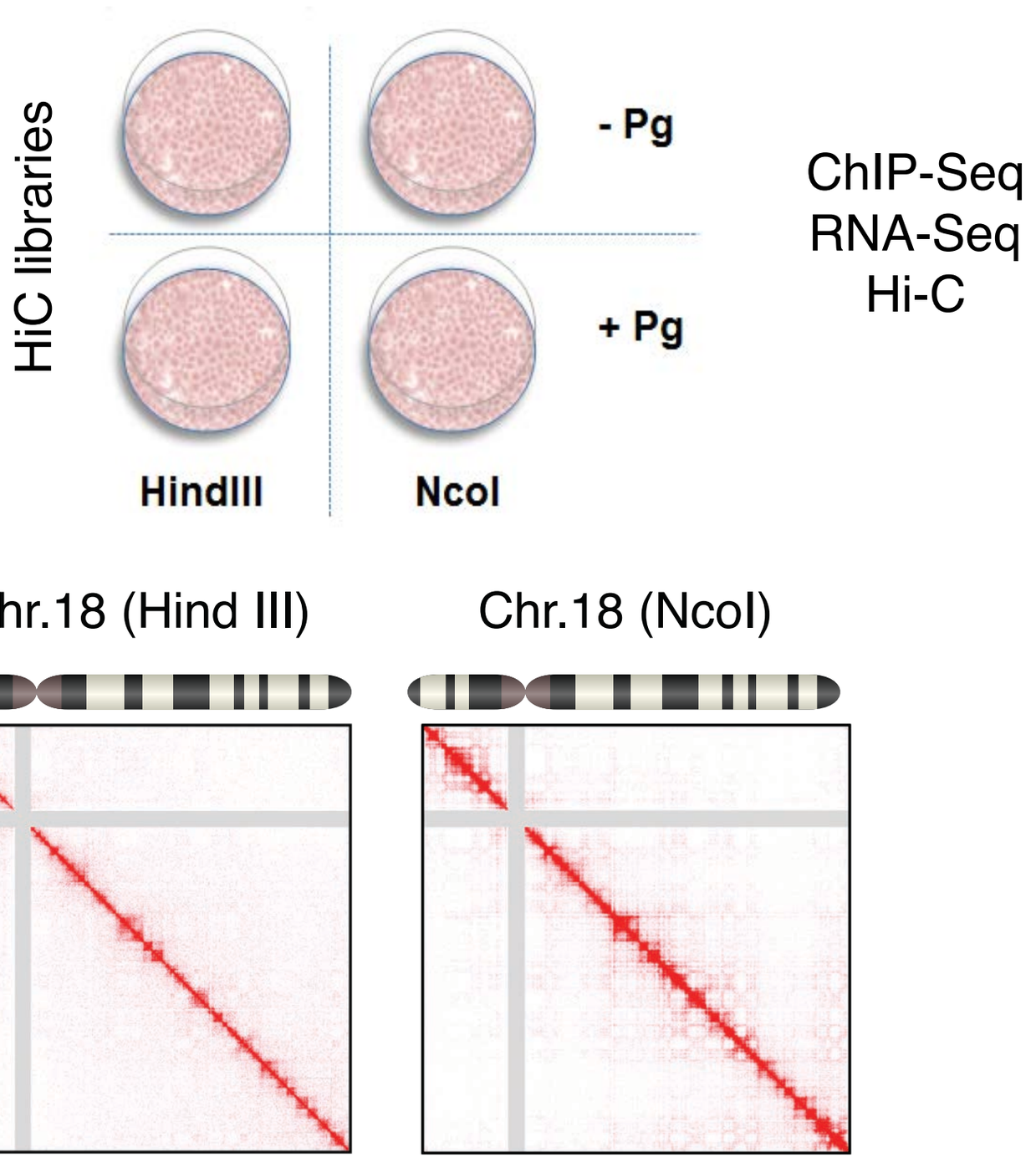
Vicent *et al* 2011, Wright *et al* 2012, Ballare *et al* 2012

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

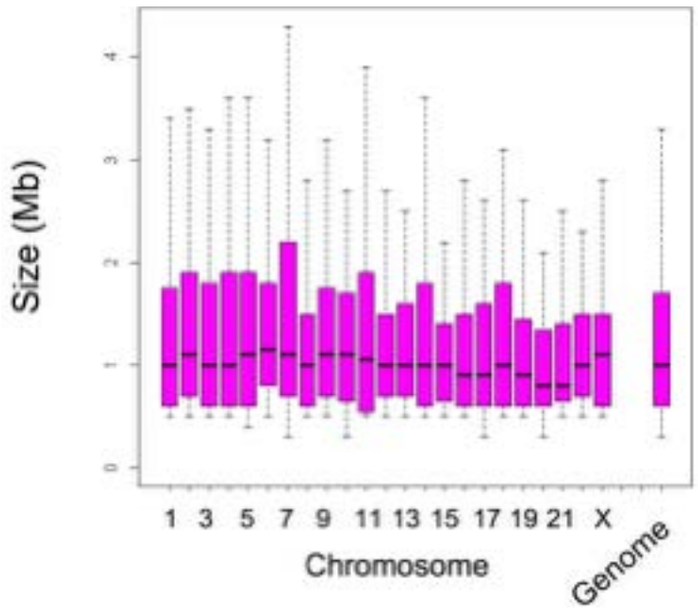
> 2,000 genes **Down**-regulated

Regulation in 3D?

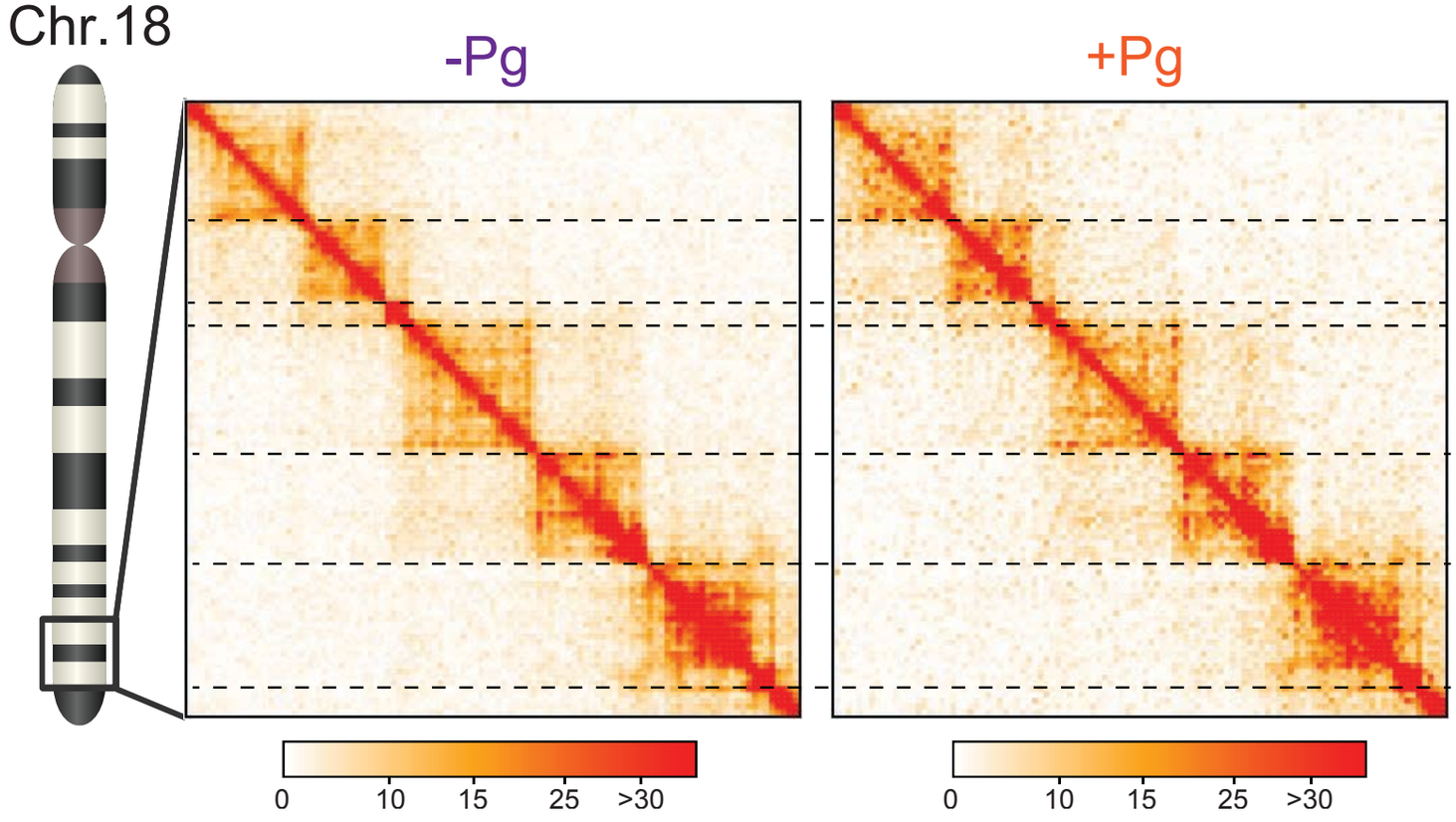
Experimental design

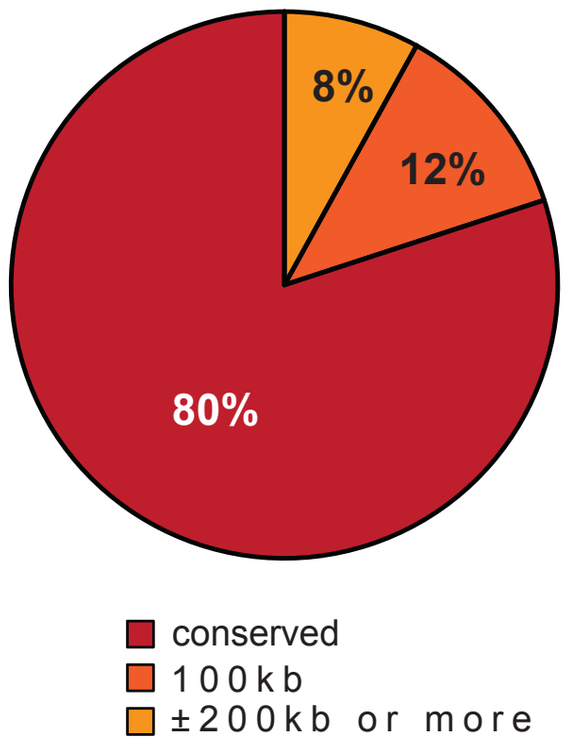


Are there TADs? how robust?

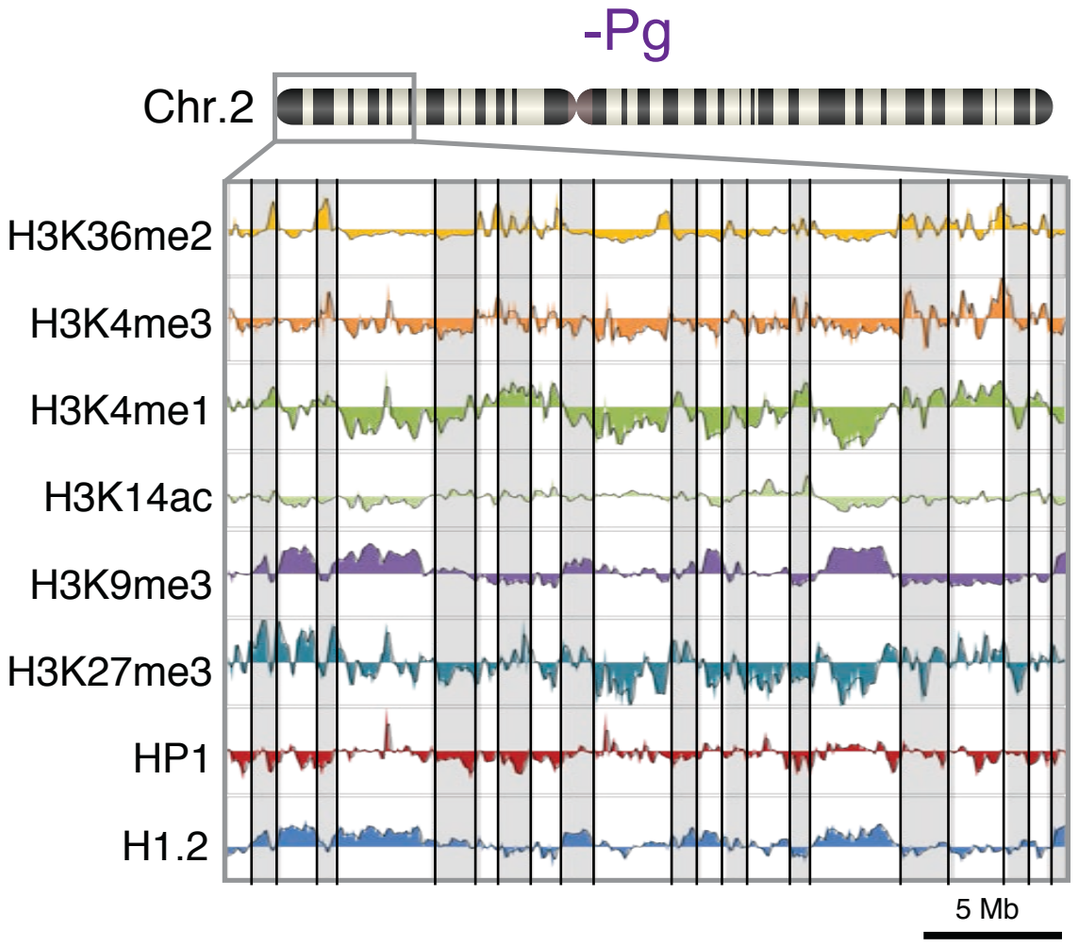


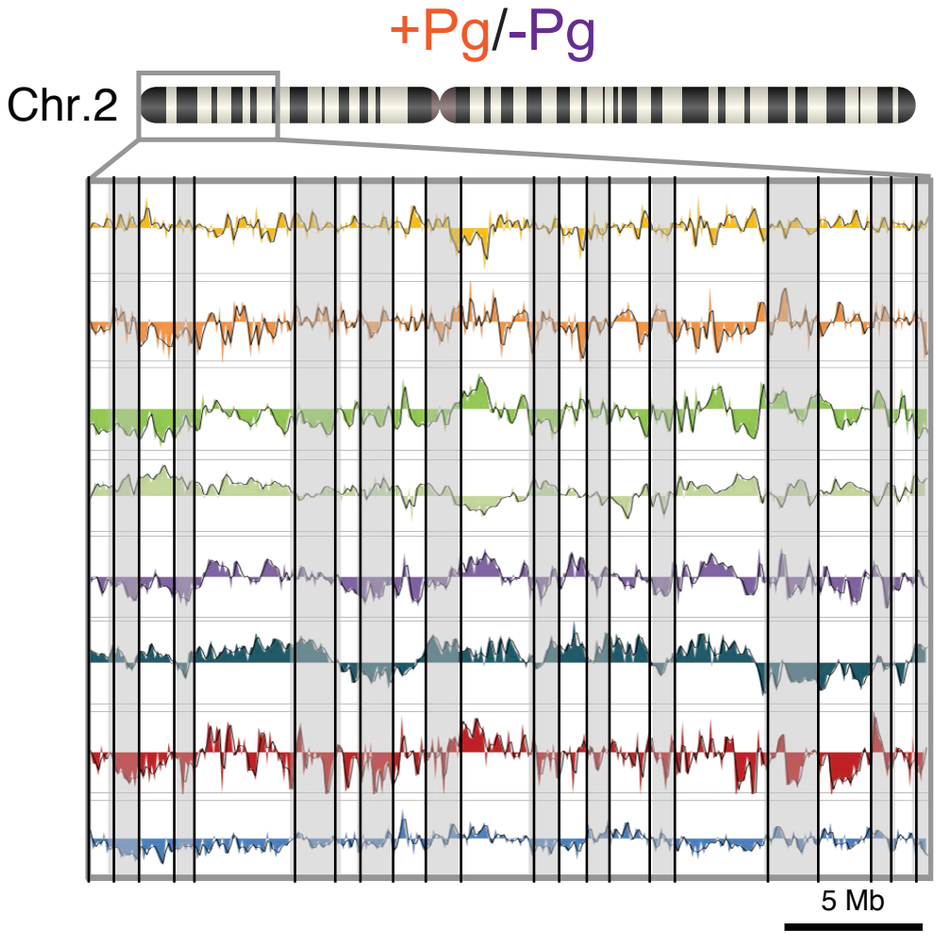
>2,000 detected TADs

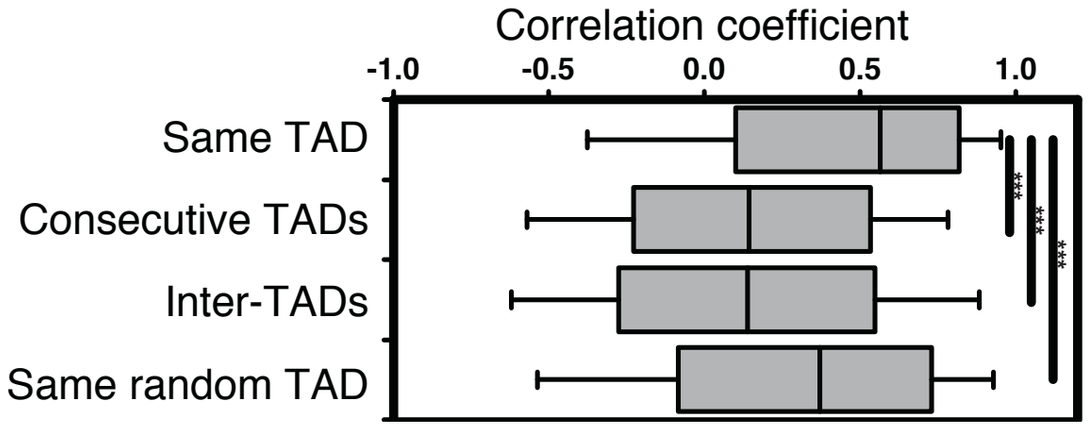


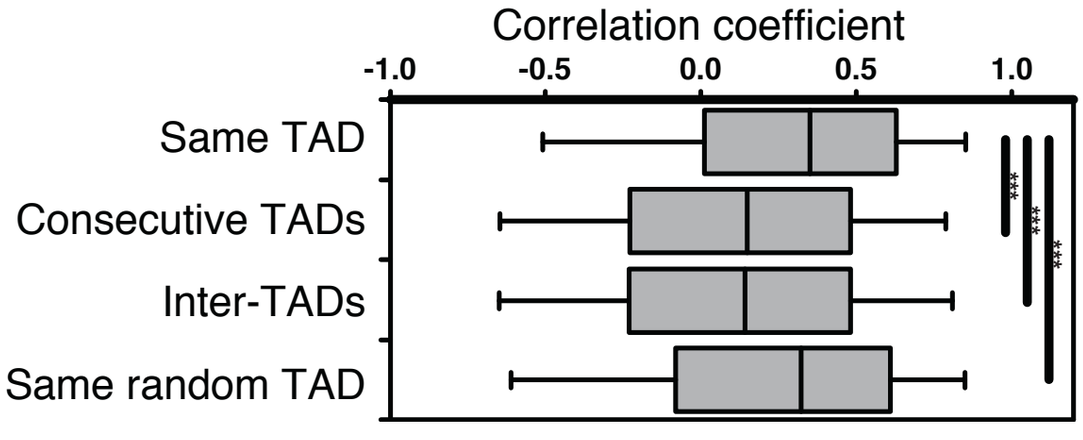


Are TADs homogeneous?

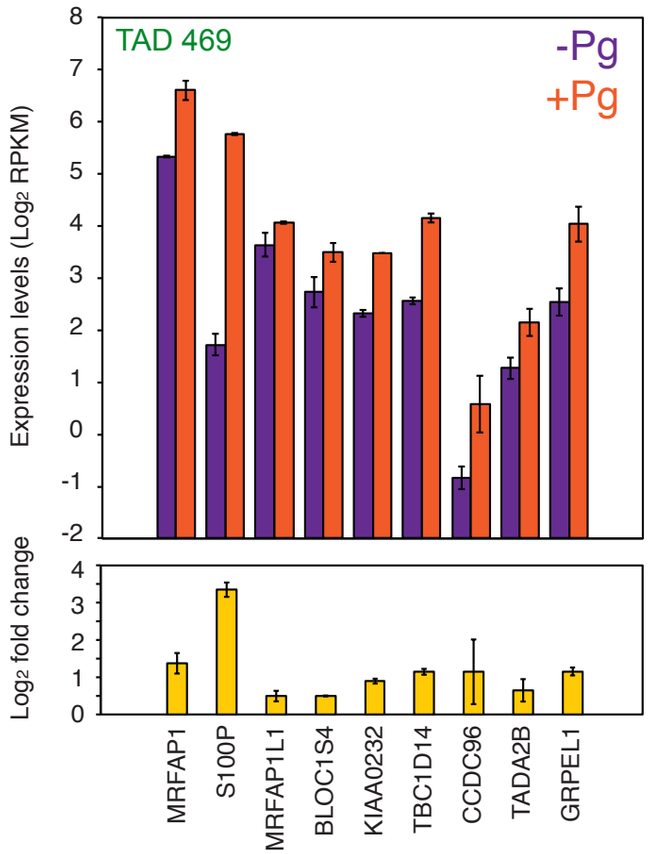


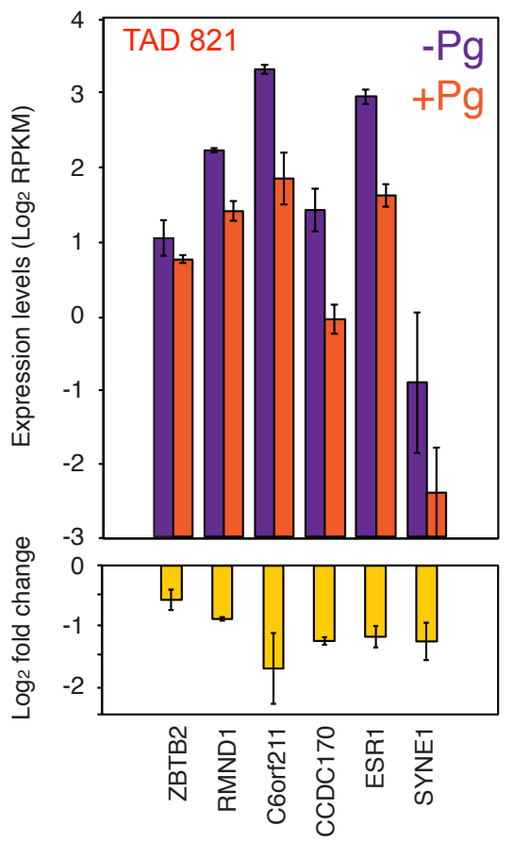


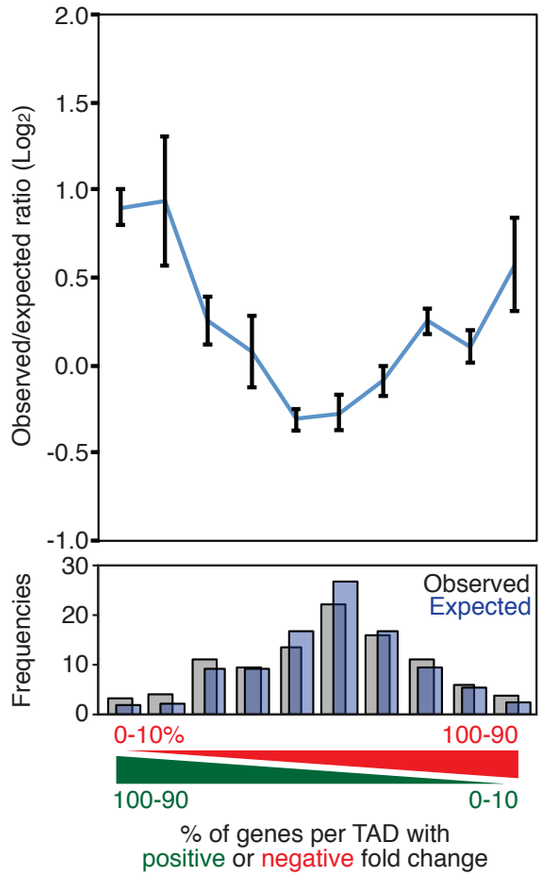




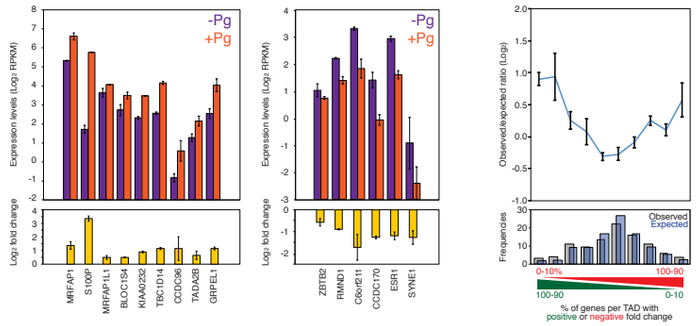
Do TADs respond differently to Pg treatment?





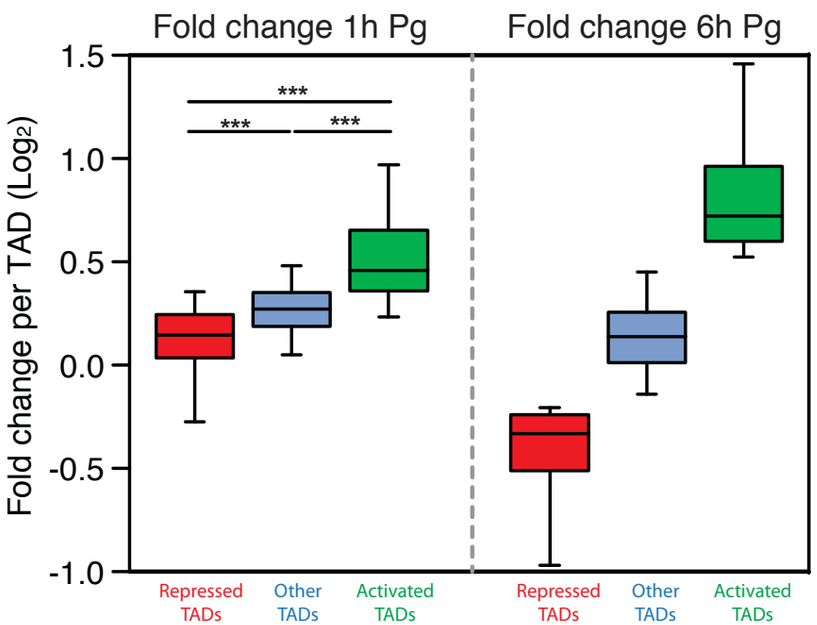


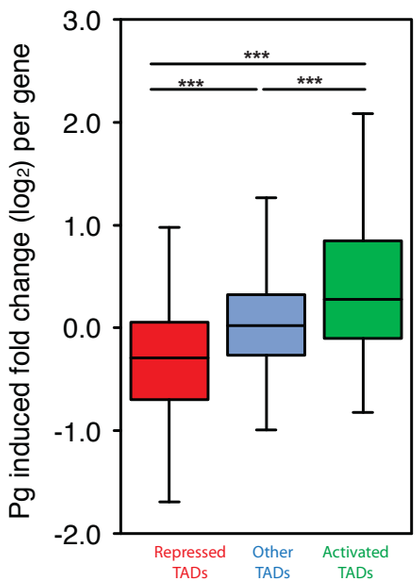
Do TADs respond differently to Pg treatment?

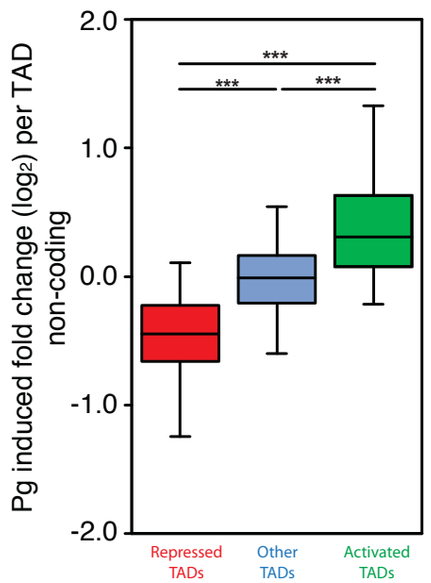


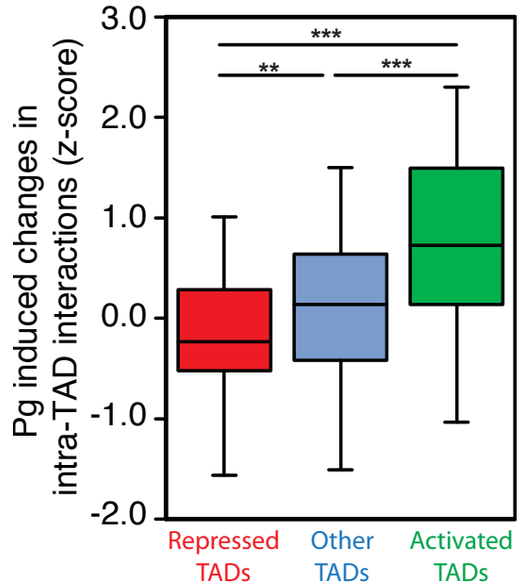
Pg induced fold change per TAD (6h)

Repressed TADs | Other TADs | Activated TADs

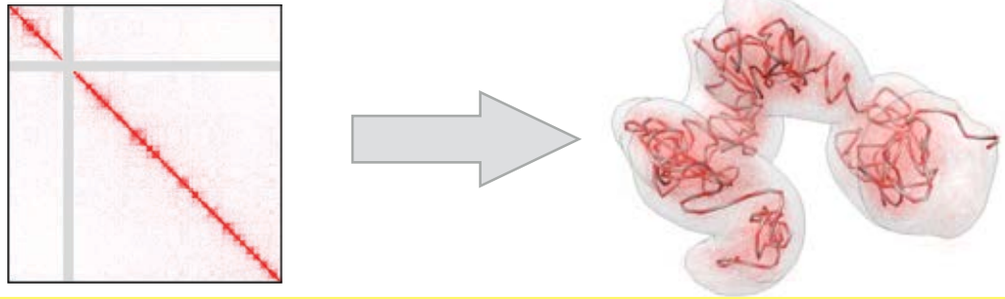




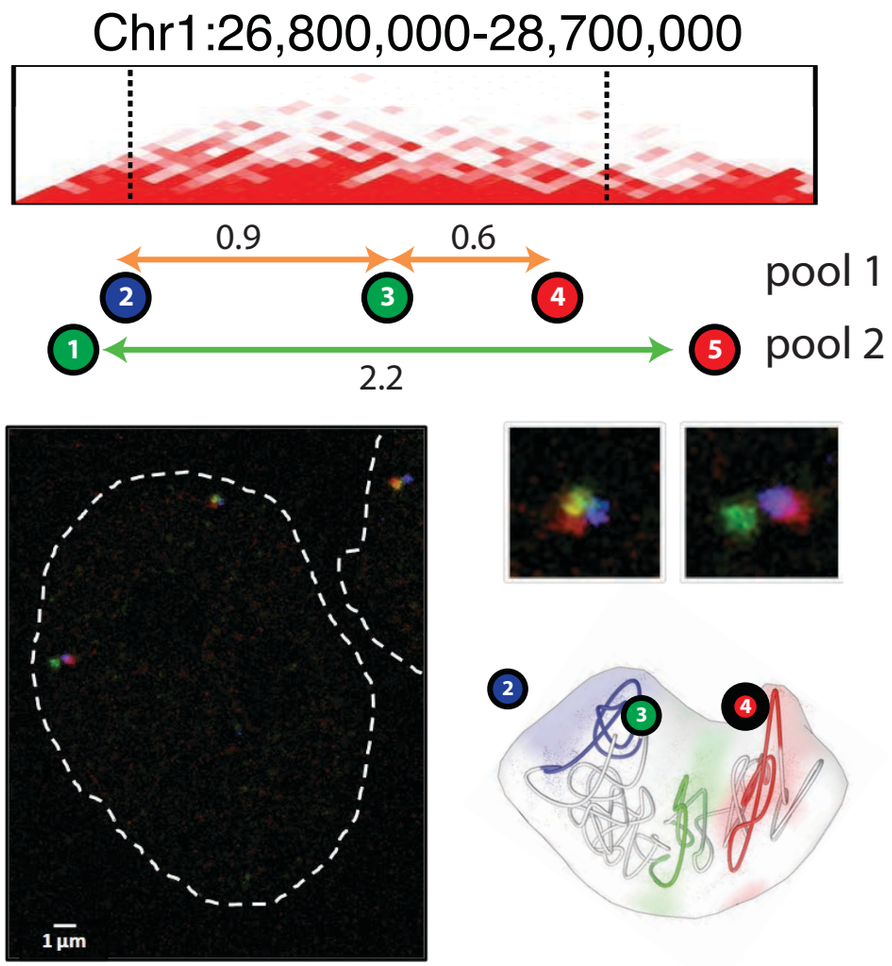


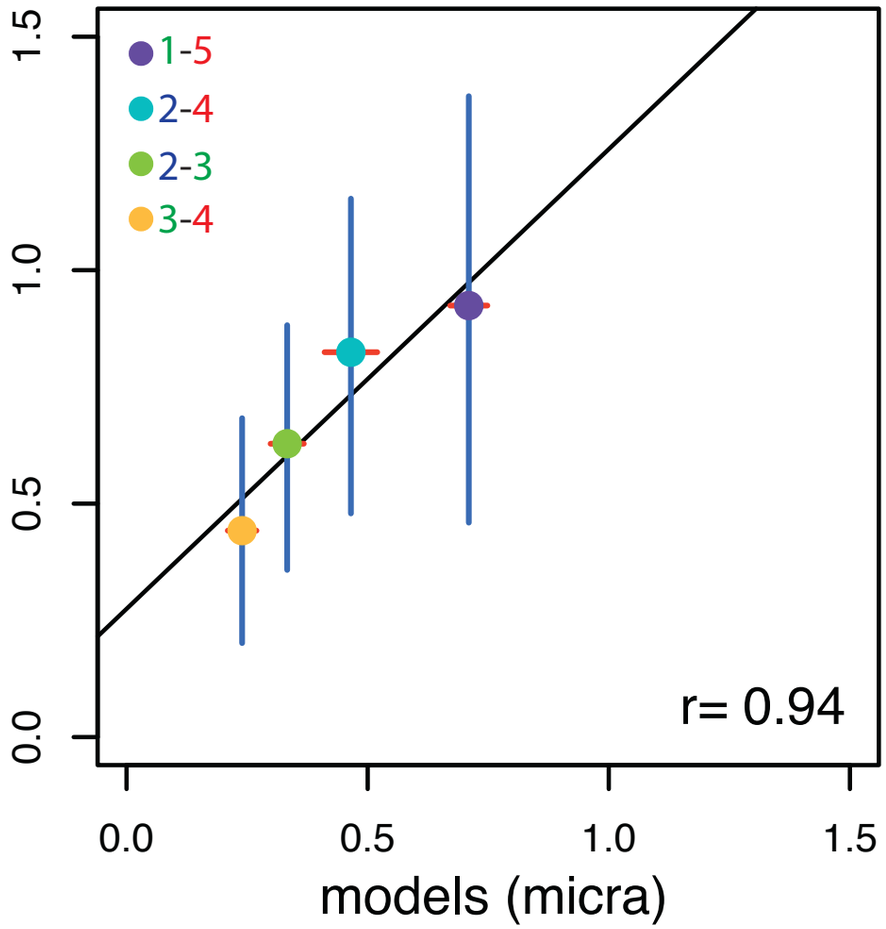


Modeling 3D TADs

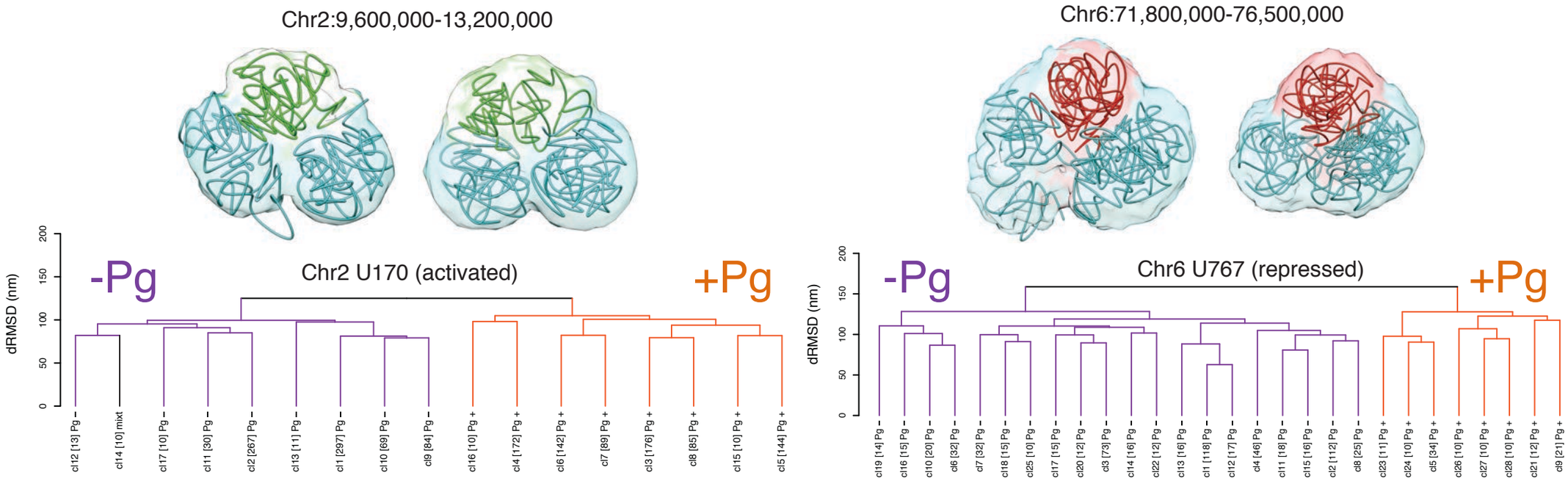


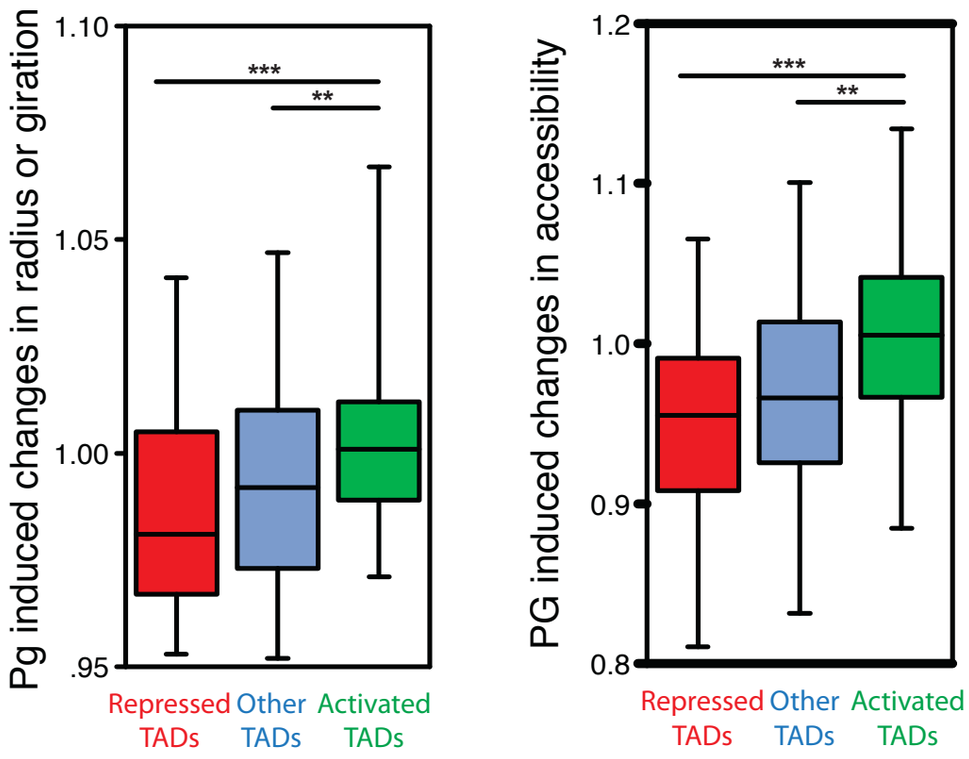
61 genomic regions containing 209 TADs covering 267Mb



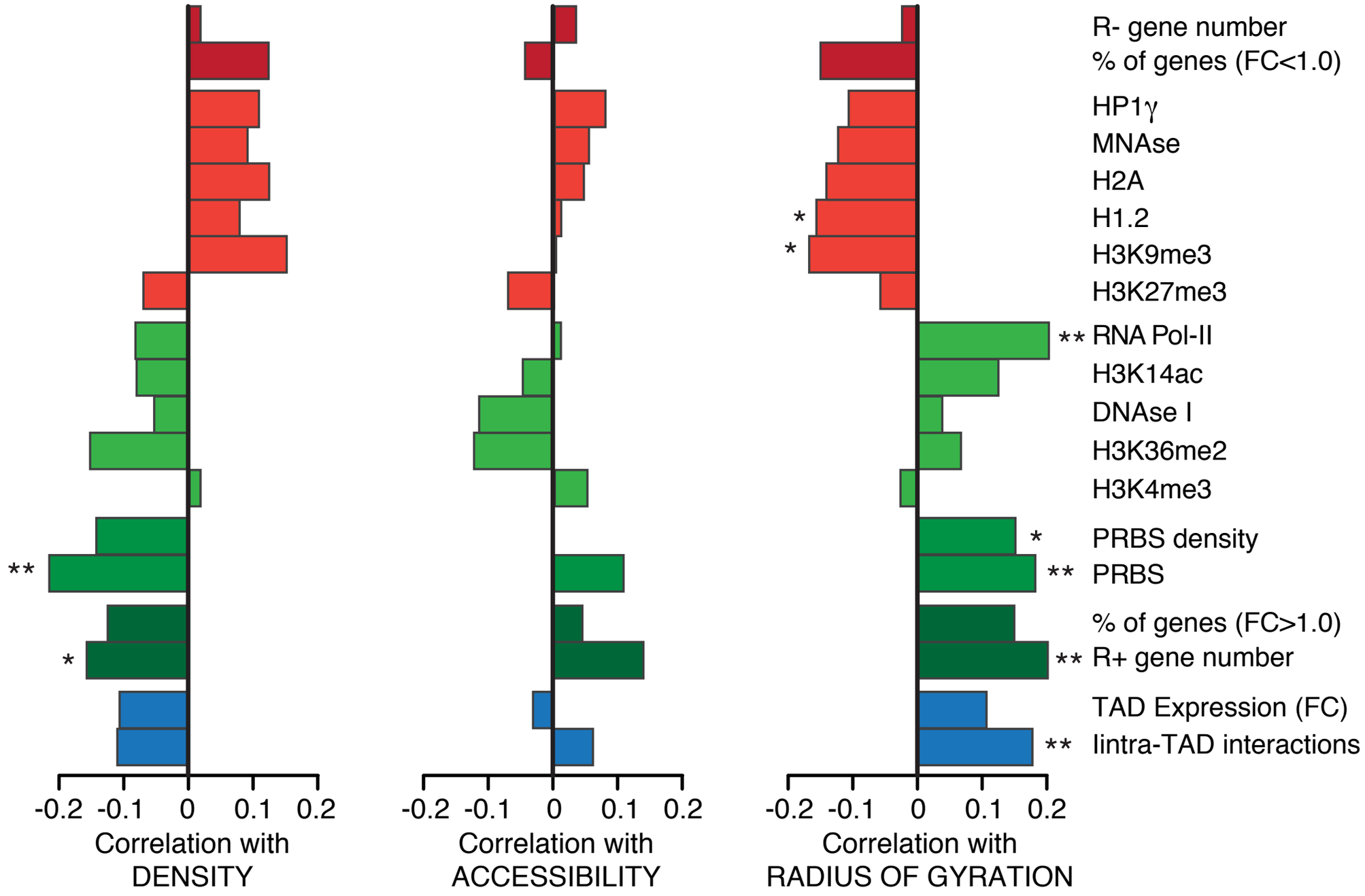


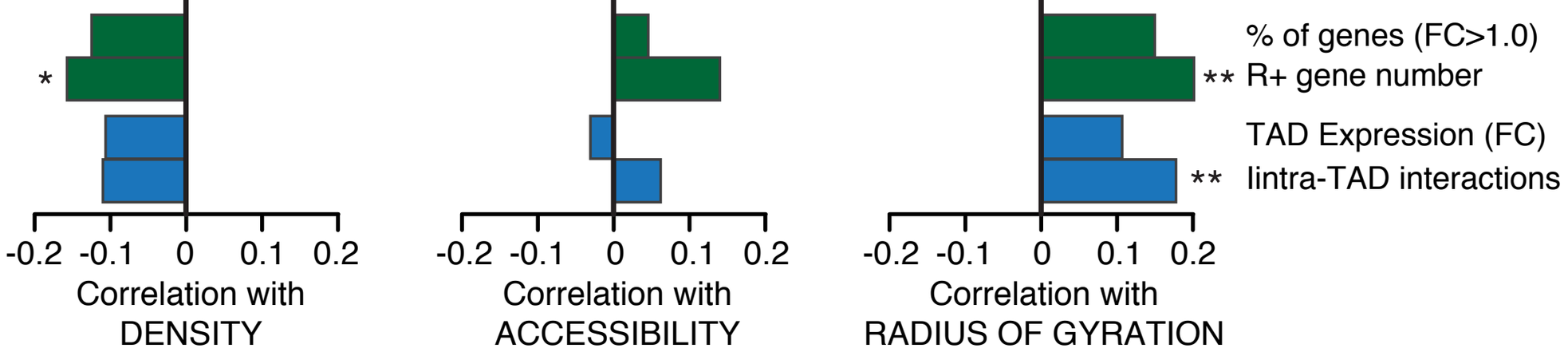
How TADs respond structurally to Pg?

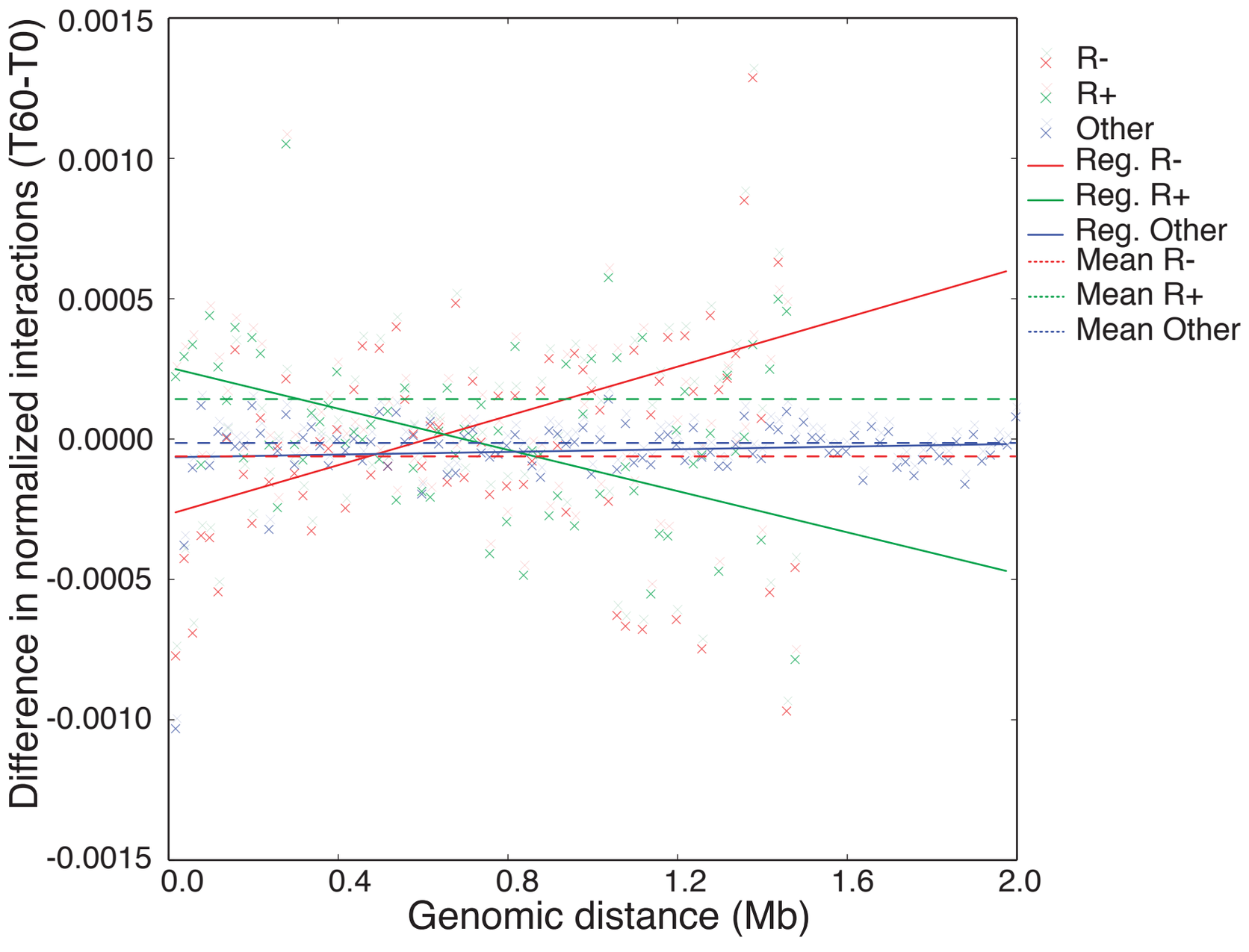




How TADs respond structurally to Pg?



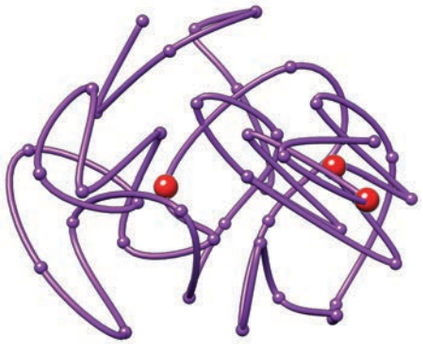

*
-0.2 -0.1 0 0.1 0.2
Correlation with
DENSITY
-0.2 -0.1 0 0.1 0.2
Correlation with
ACCESSIBILITY
-0.2 -0.1 0 0.1 0.2
Correlation with
RADIUS OF GYRATION
% of genes (FC>1.0)
** R+ gene number
TAD Expression (FC)
** Iintra-TAD interactions

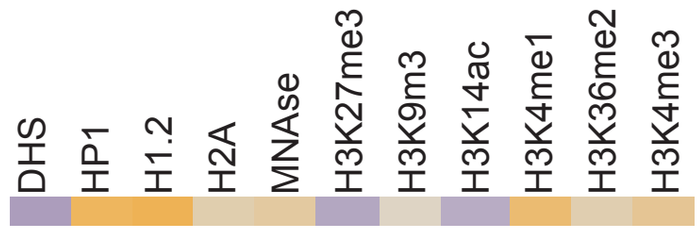

Difference in normalized interactions (T60-T0)
0.0015
0.0010
0.0005
0.0000
-0.0005
-0.0010
-0.0015
0.0
0.4
0.8
1.2
1.6
2.0
Genomic distance (Mb)
R-
R+
Other
Reg. R-
Reg. R+
Reg. Other
Mean R-
Mean R+
Mean Other

Model for TAD regulation

Repressed TAD

chr1 U41

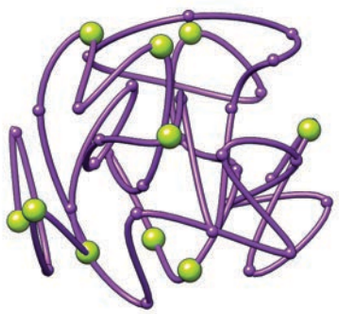


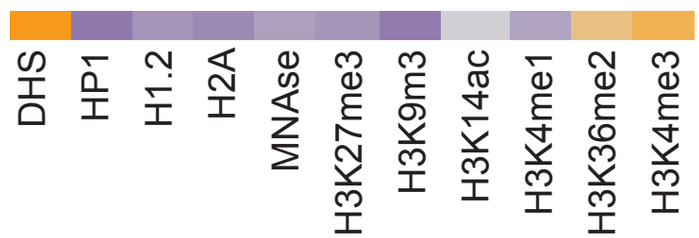


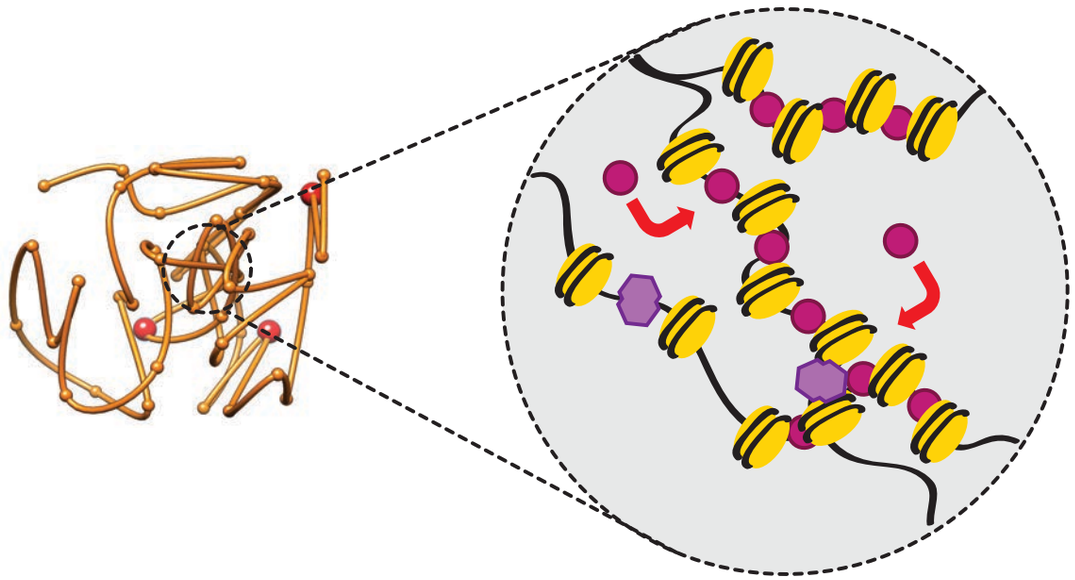
Structural transition **+Pg**

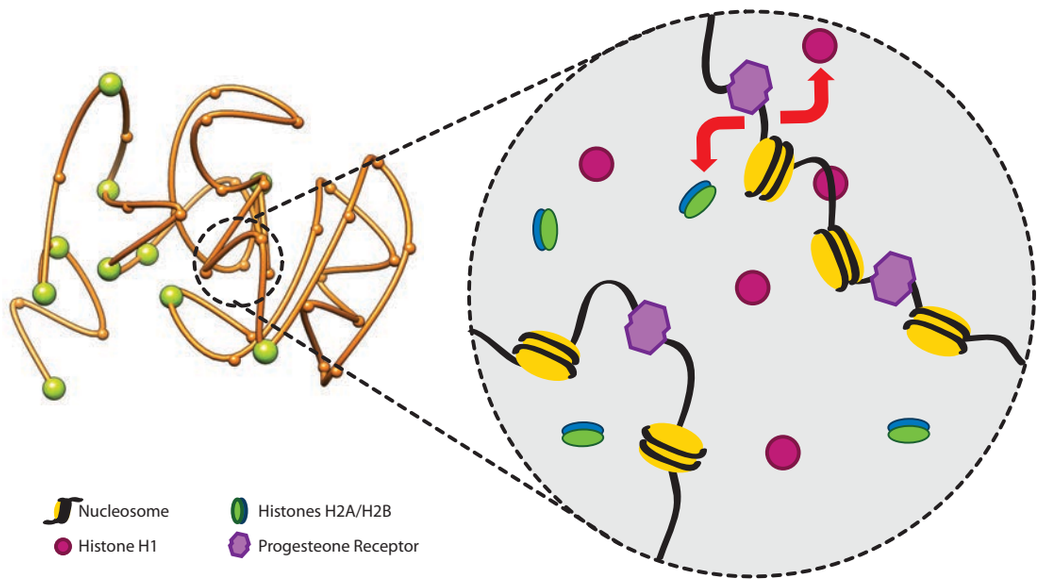
Activated TAD

chr2 U207



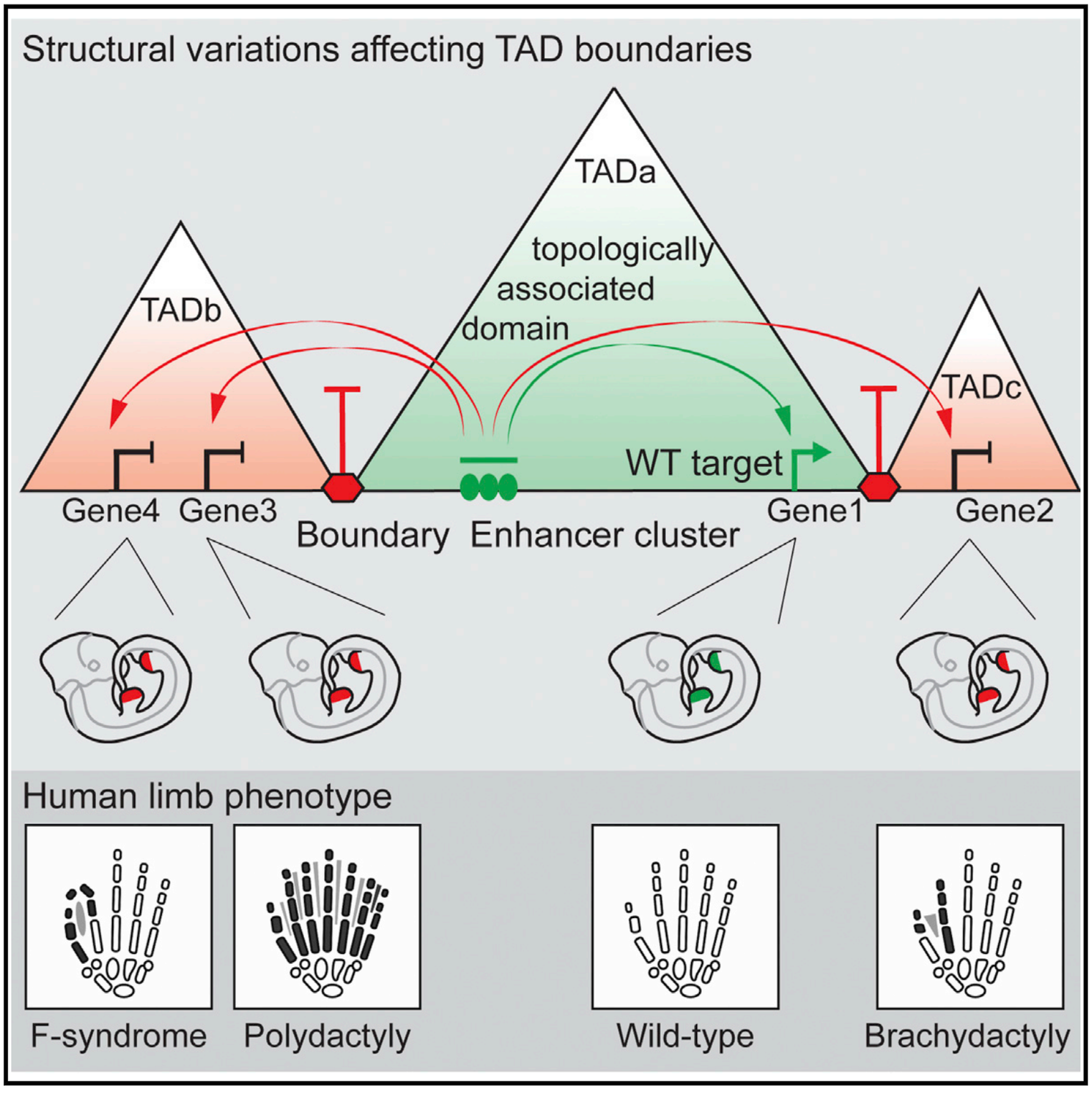






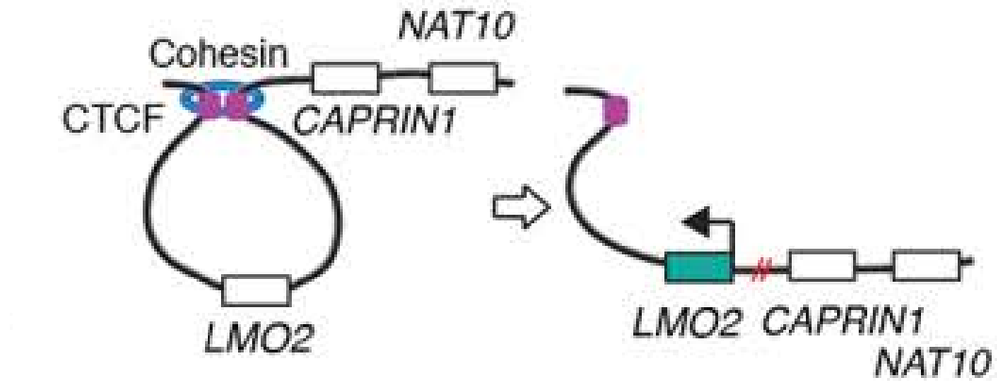
TADs are functional units

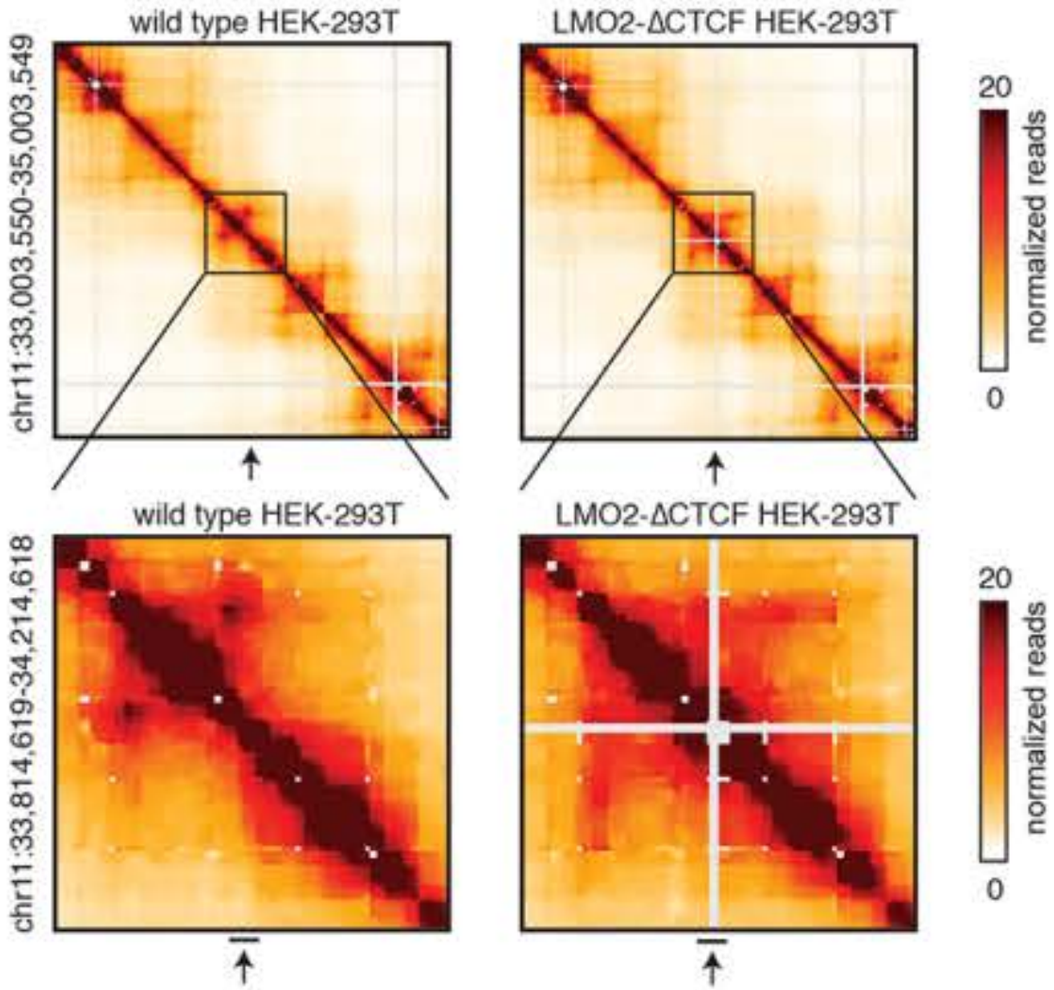
Lupiáñez, et al. (2015). Cell, 1–15.



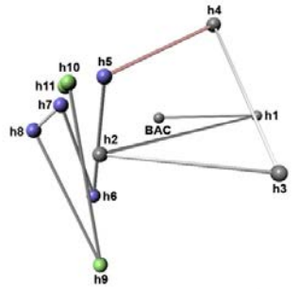
TADs are functional units

Hnisz, D., et al. (2016). Science, on line

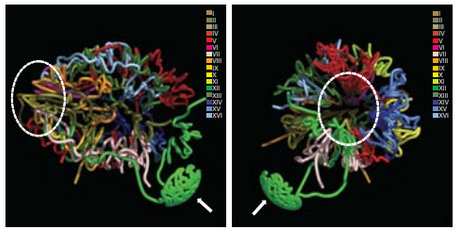




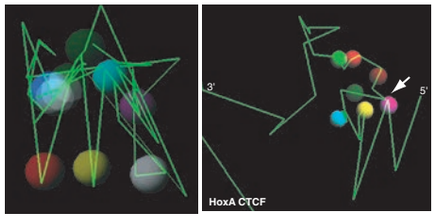
Are the models correct?



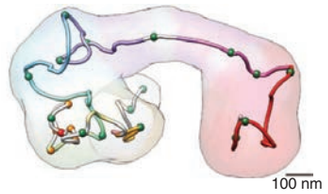
Jhunjhunwala (2008) Cell



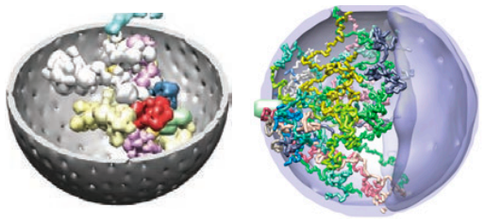
Duan (2010) Nature



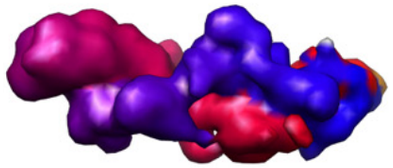
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research



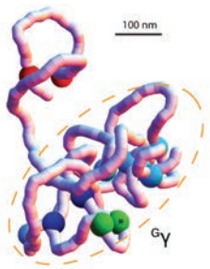
Baù (2011) Nature Structural & Molecular Biology



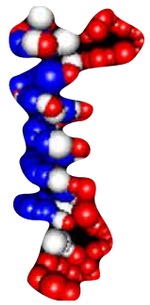
Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research



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

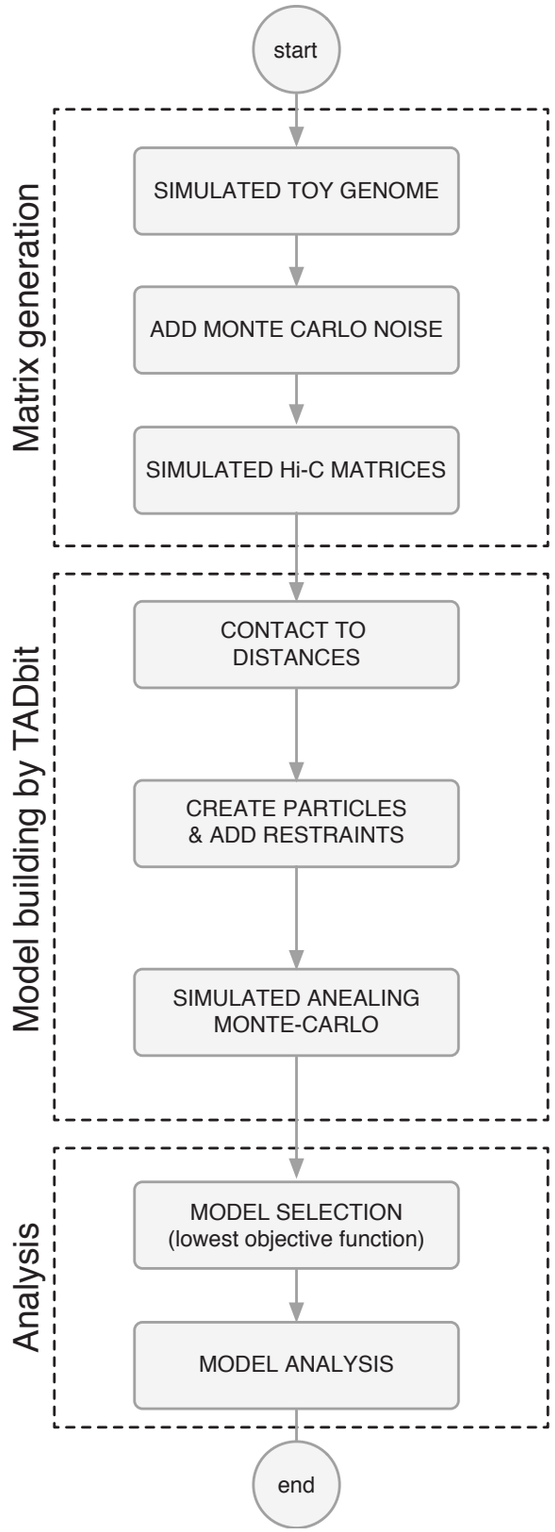
Recent studies of the three-dimensional (3D) conformation of genomes are revealing insights into the organization and the regulation of biological processes, such as gene expression regulation and replication (1–6). The advent of the so-called Chromosome Conformation Capture (3C) assays (7), which allowed 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).

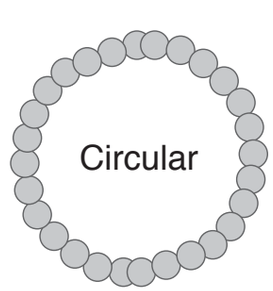
We recently developed a mean-field method for modeling 3D structures of genomes and genomic domains based on 3C interaction data (9). Our approach, called TADbit, was developed around the Integrative Modeling Platform (IMP, http://integrativemodeing.org), a general framework for restraint-based modeling 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

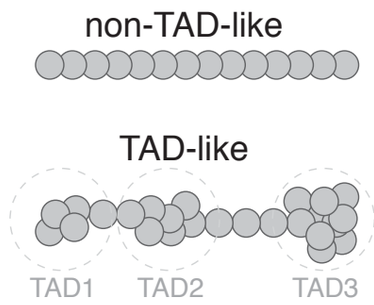
*To whom correspondence should be addressed. Tel: +34 934 020 542; Fax: +34 934 037 279; Email: mmarti@pcb.ub.cat

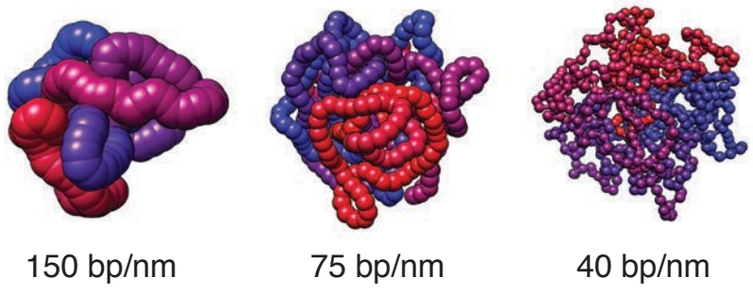
Trussart, et al. (2015). Nucleic Acids Research.

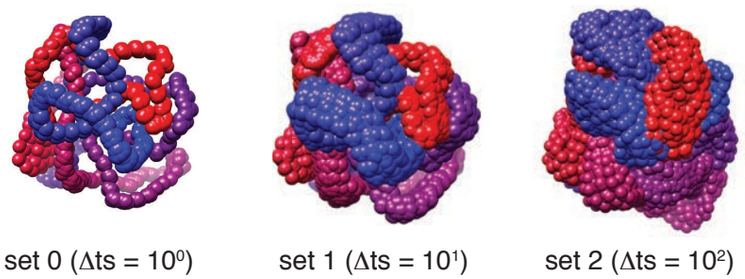
Toy models

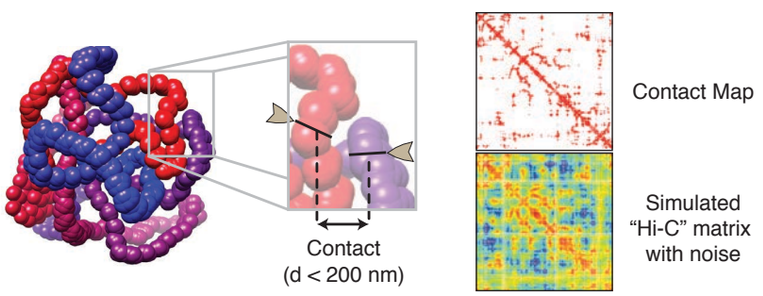








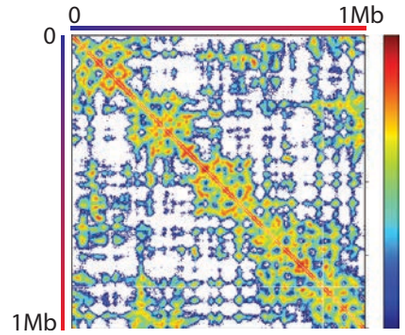


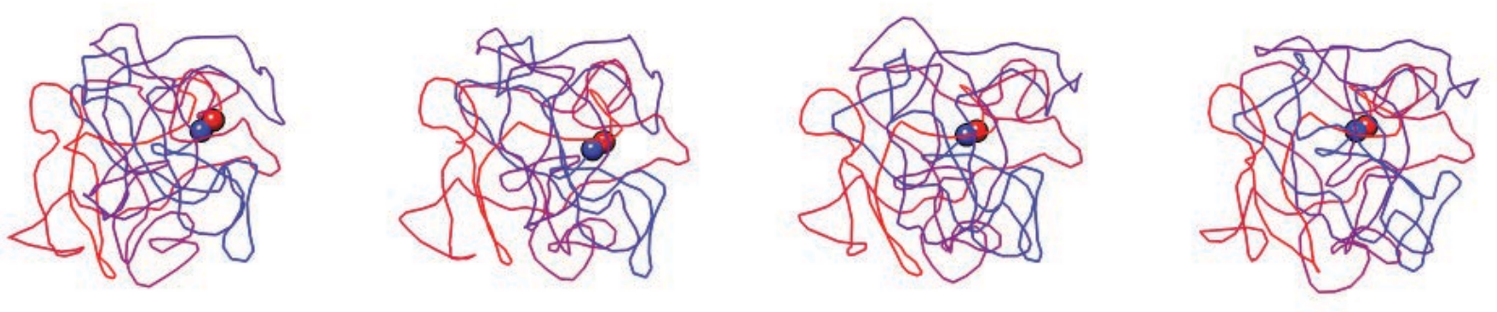


by Ivan Junier

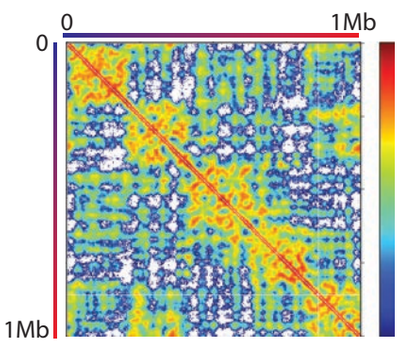
Toy interaction matrices

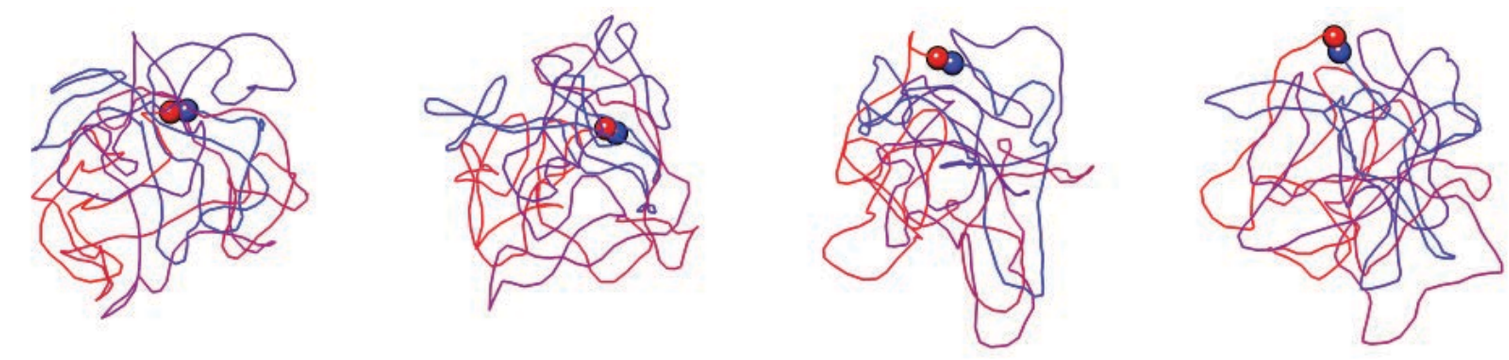
set 0 (Δts=10^0)



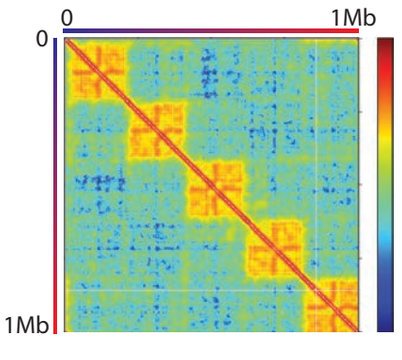


set 4 (Δts=10^4)

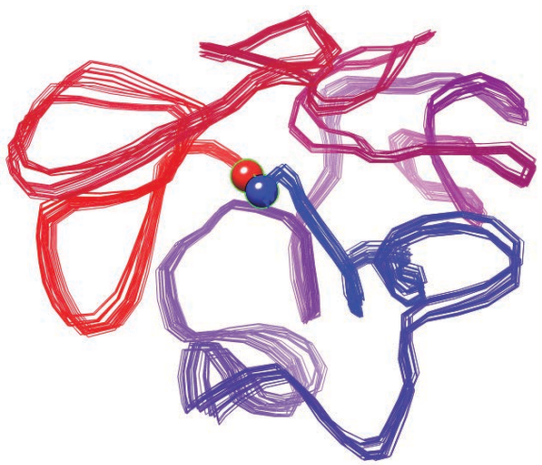


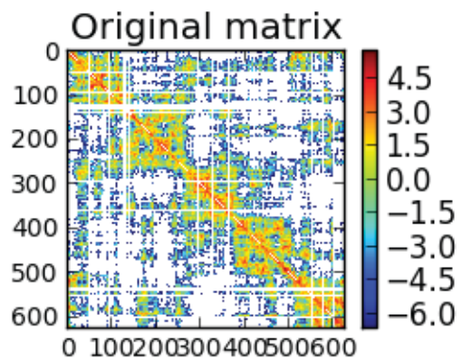


set 6 (Δts=10^6)

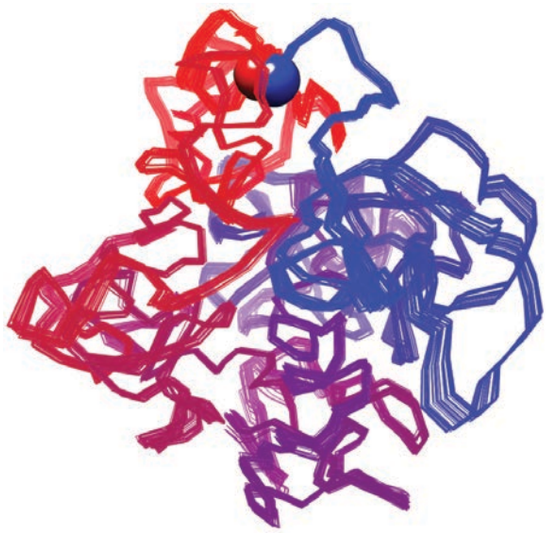


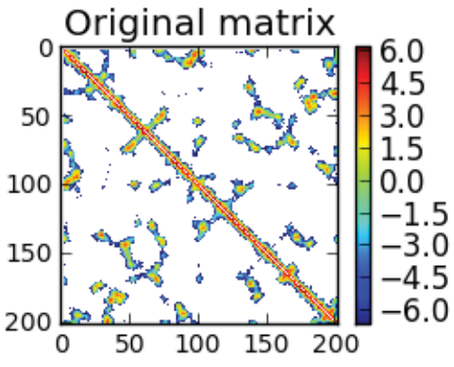
Reconstructing toy models

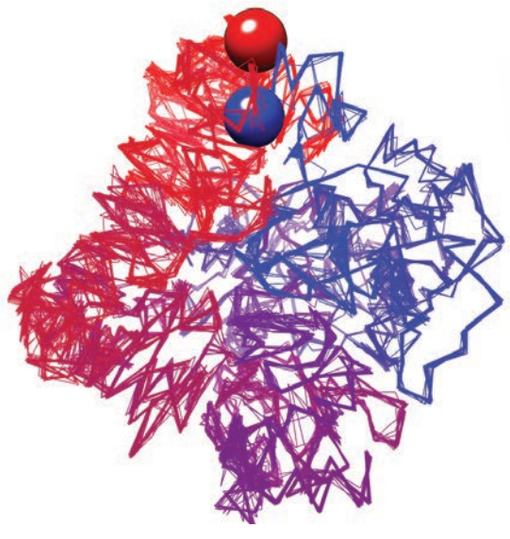




chr40_TAD
α=100
Δts=10
TADbit-SCC: 0.91
<dRMSD>: 32.7 nm
<dSCC>: 0.94

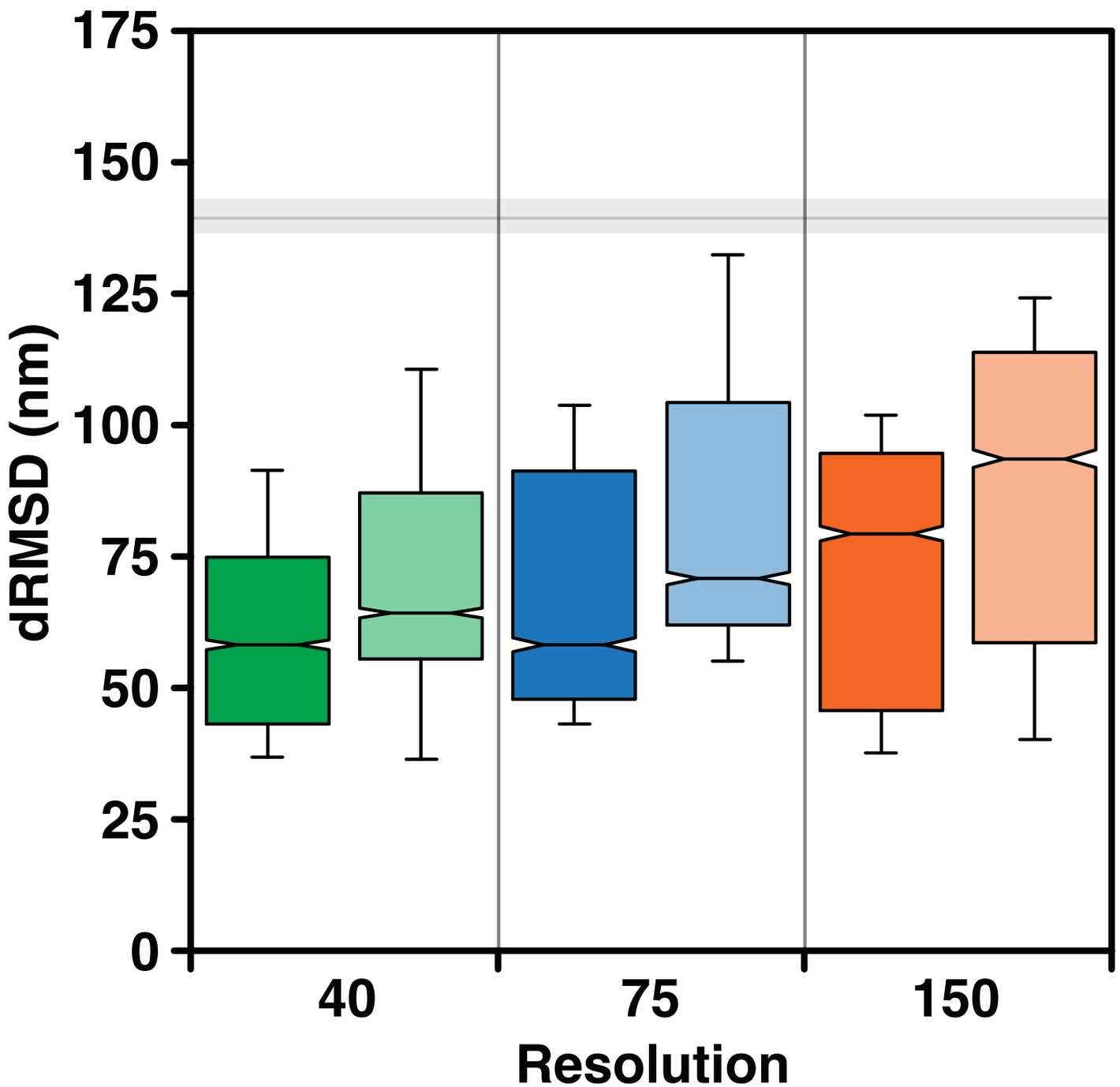




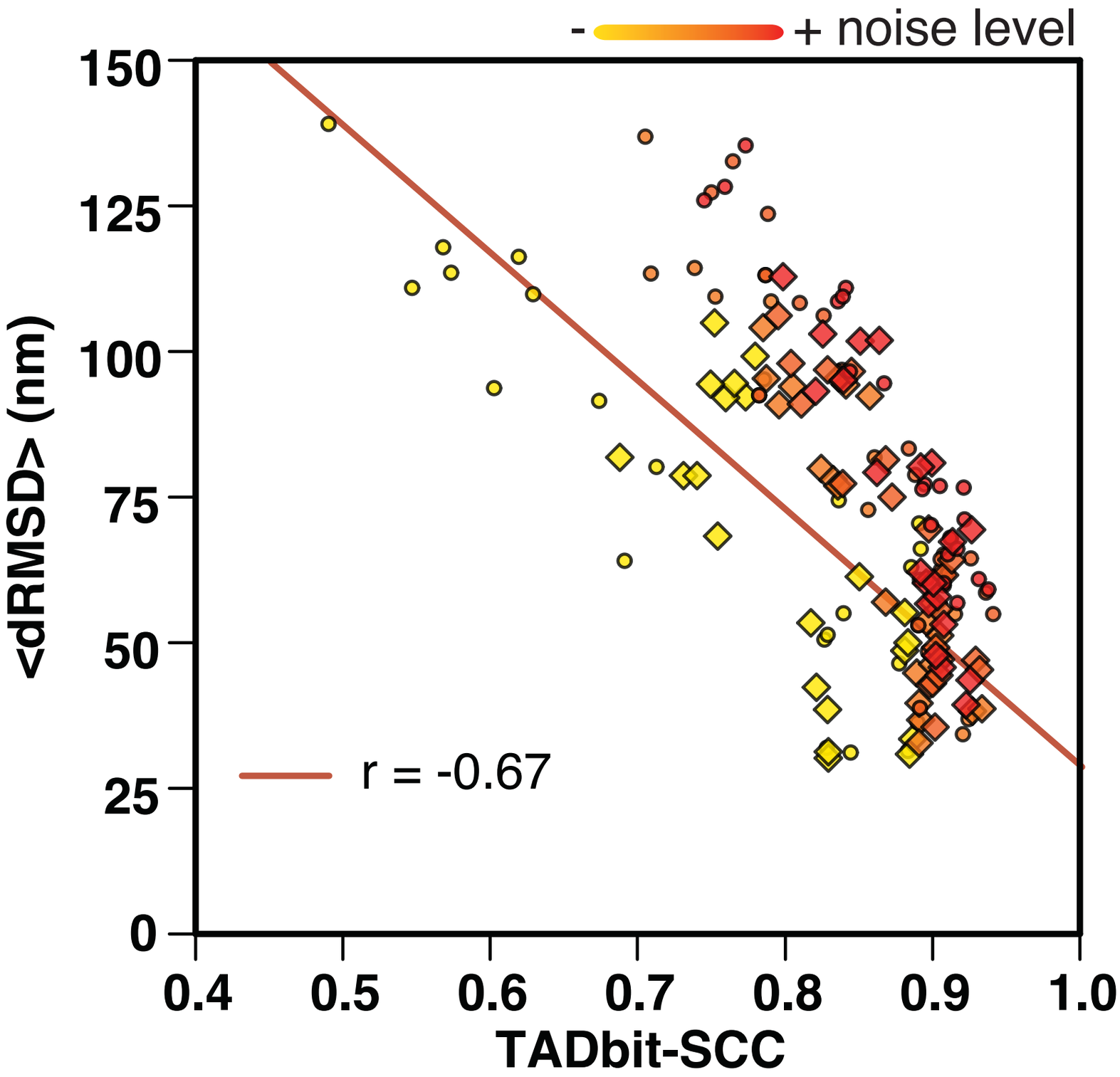


chr150_TAD
α=50
Δts=1
TADbit-SCC: 0.82
<dRMSD>: 45.4 nm
<dSCC>: 0.86

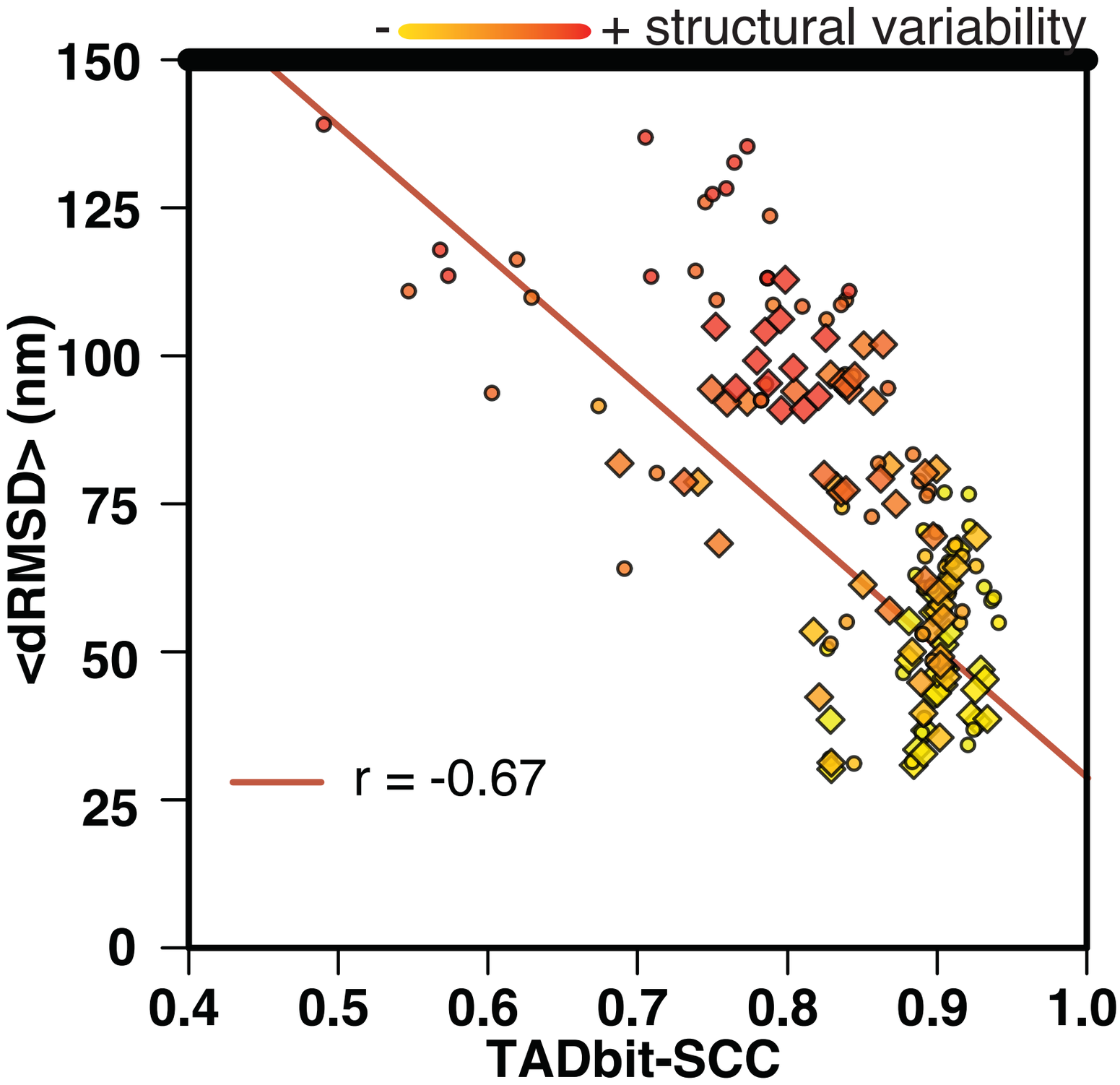
TADs & higher-res are "good"



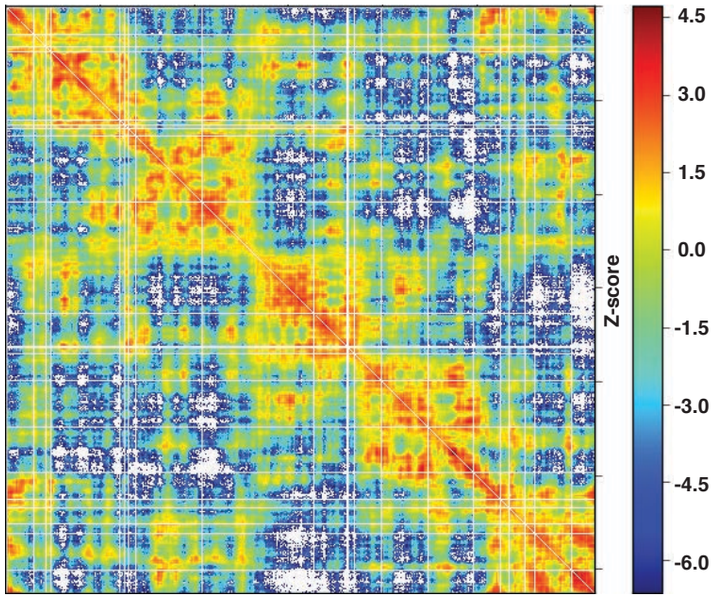
Noise is "OK"


- + 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"

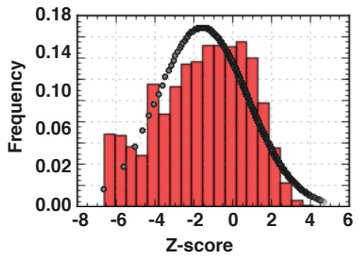


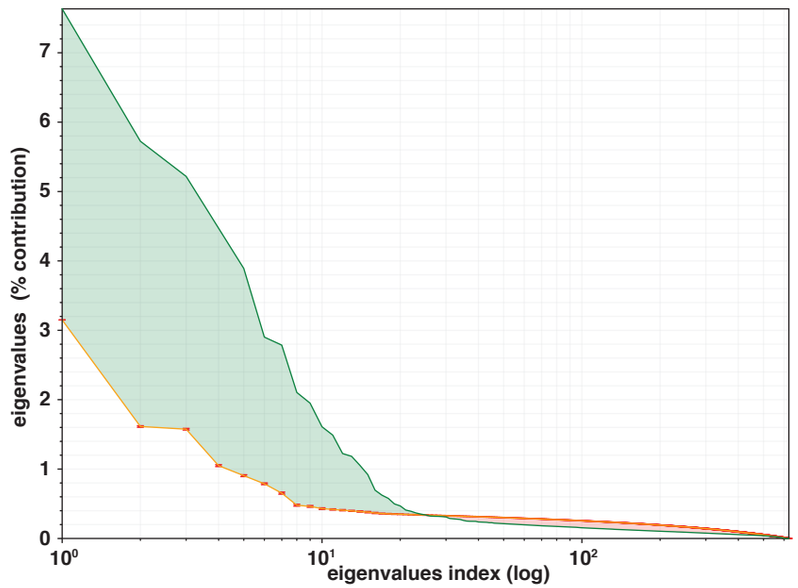
Can we predict the accuracy of the models?

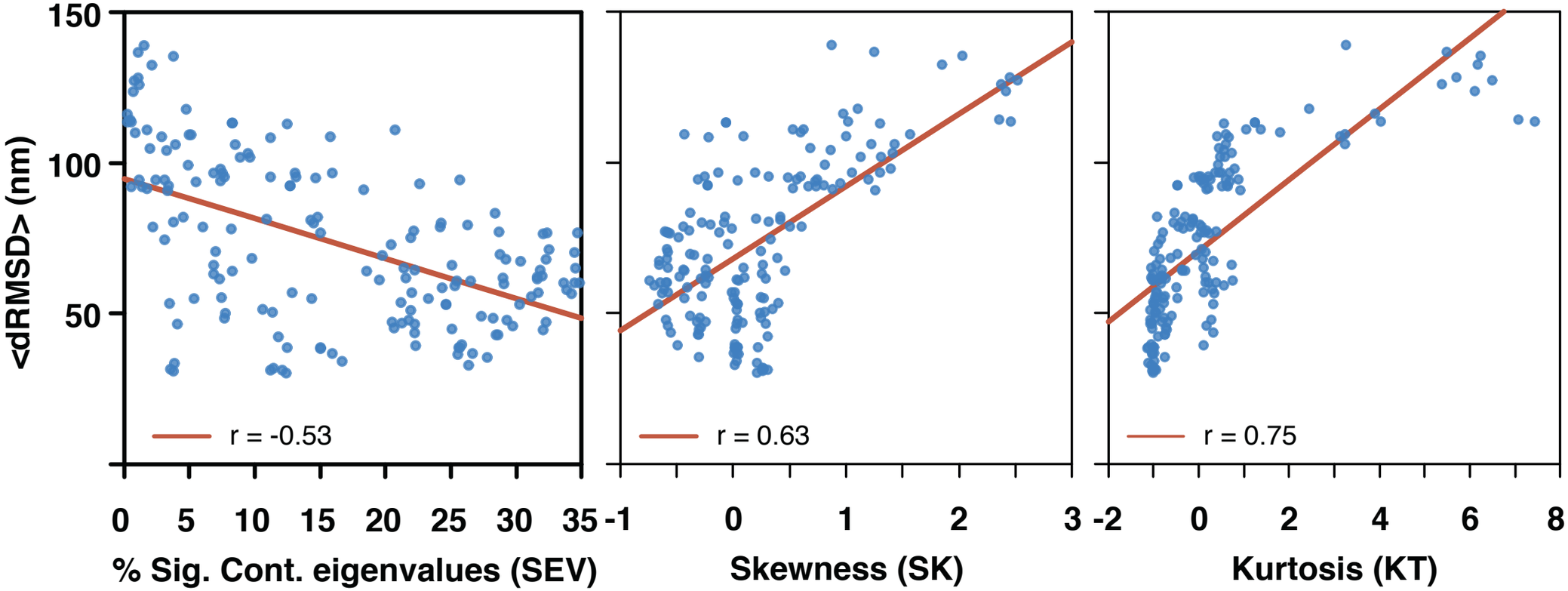


Toy genome: chr40_TAD
Density: 40 bp/nm
TADs: Yes
Noise: 150
Δts: 10^0

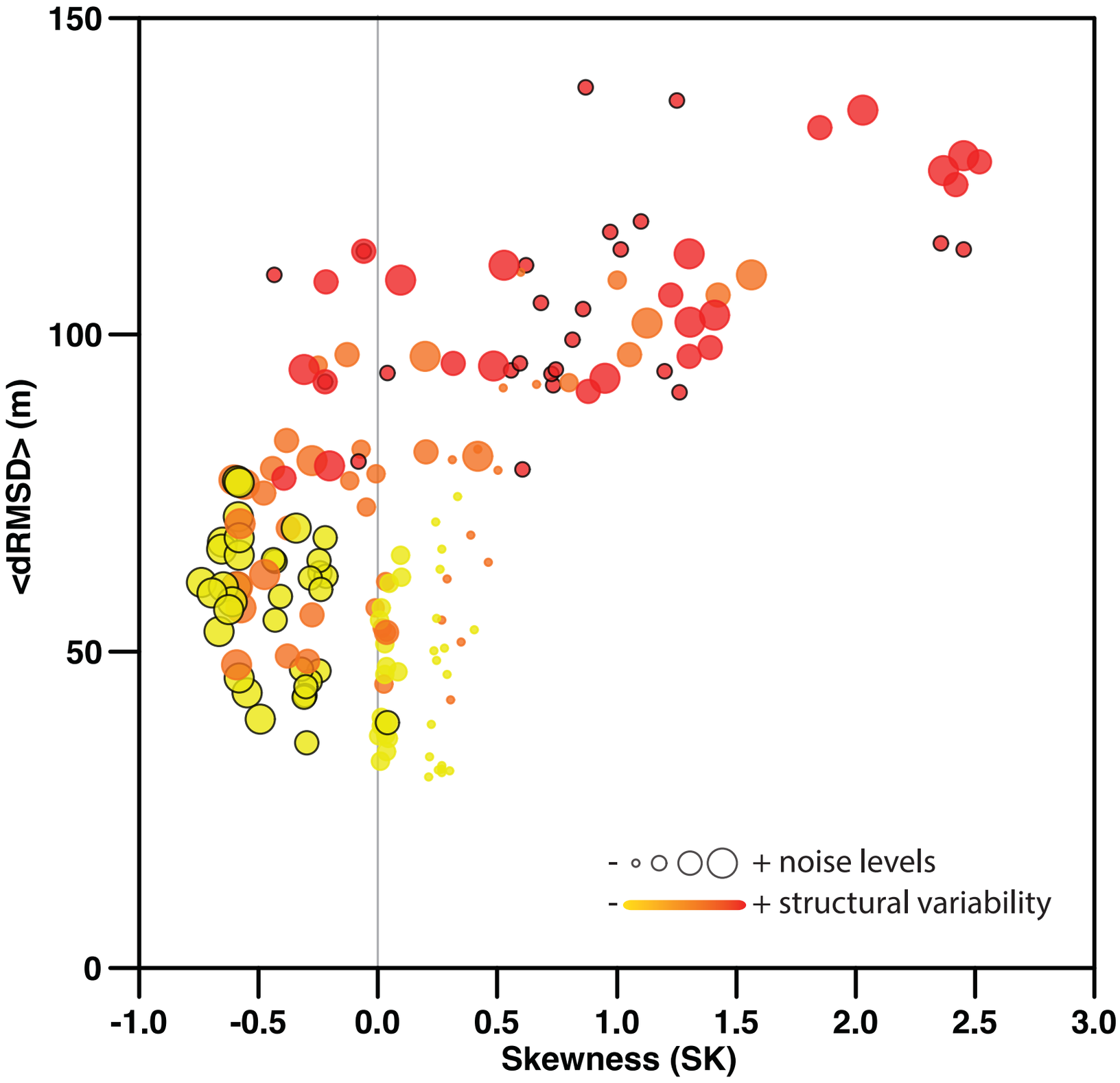
% Sig. Cont. EV: 32.3
Skewness: -0.32
Kurtosis: -0.69





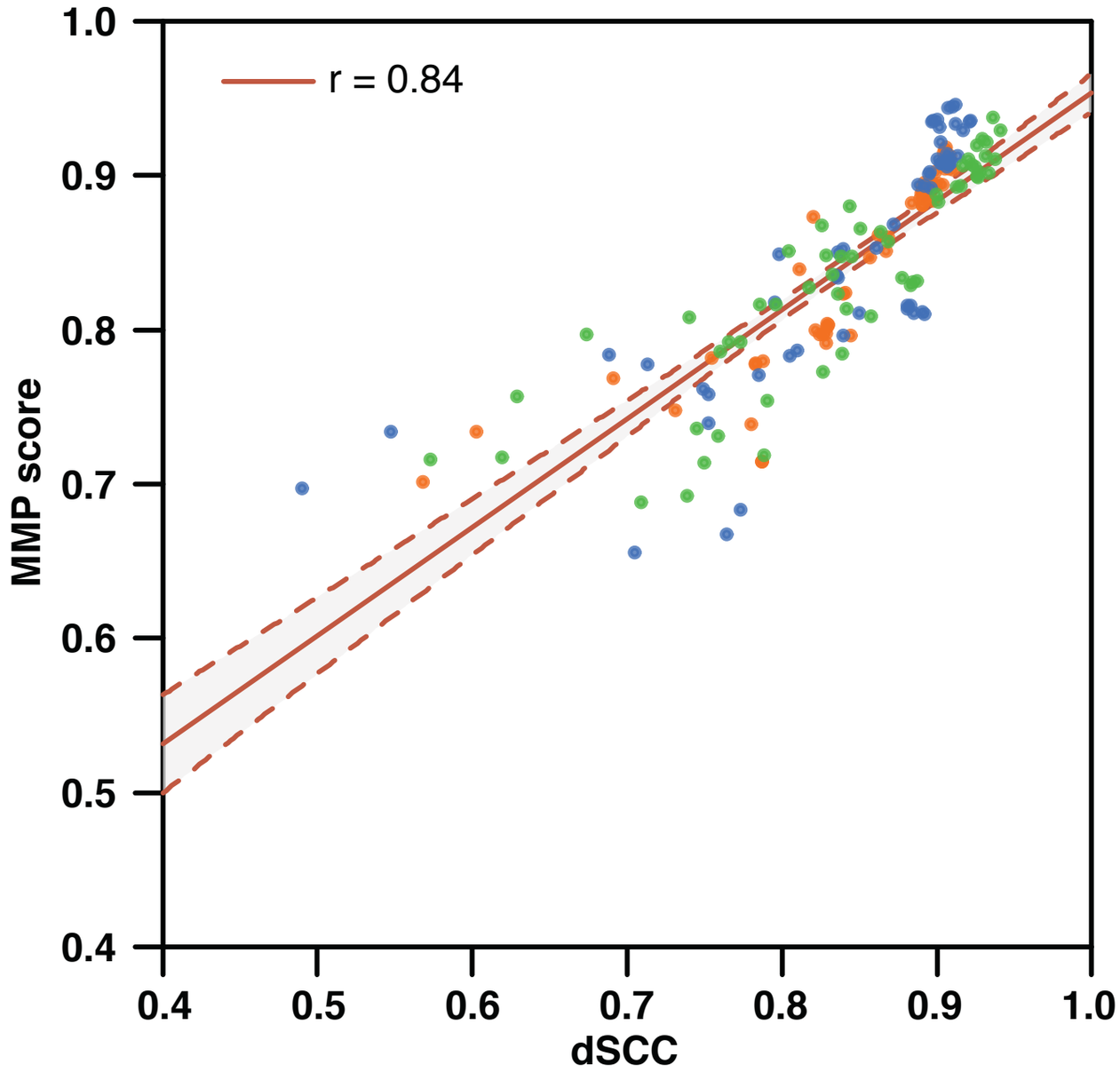


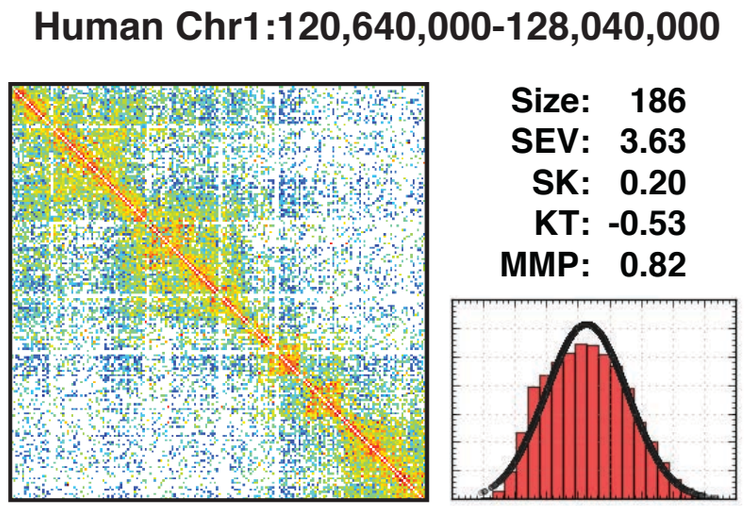
Skewness "side effect"

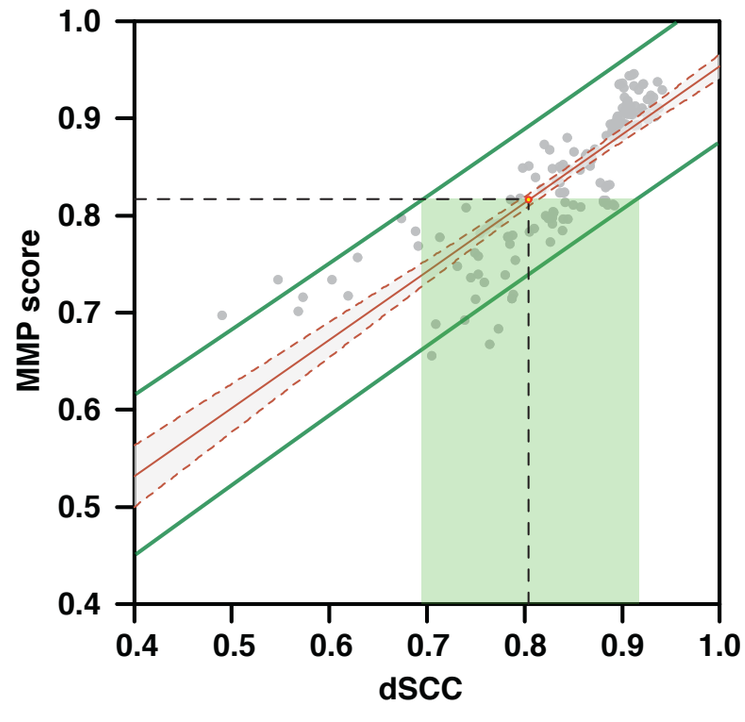


Can we predict the accuracy of the models?

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







Higher-res is "good"

put your $$ in sequencing

Noise is "OK"

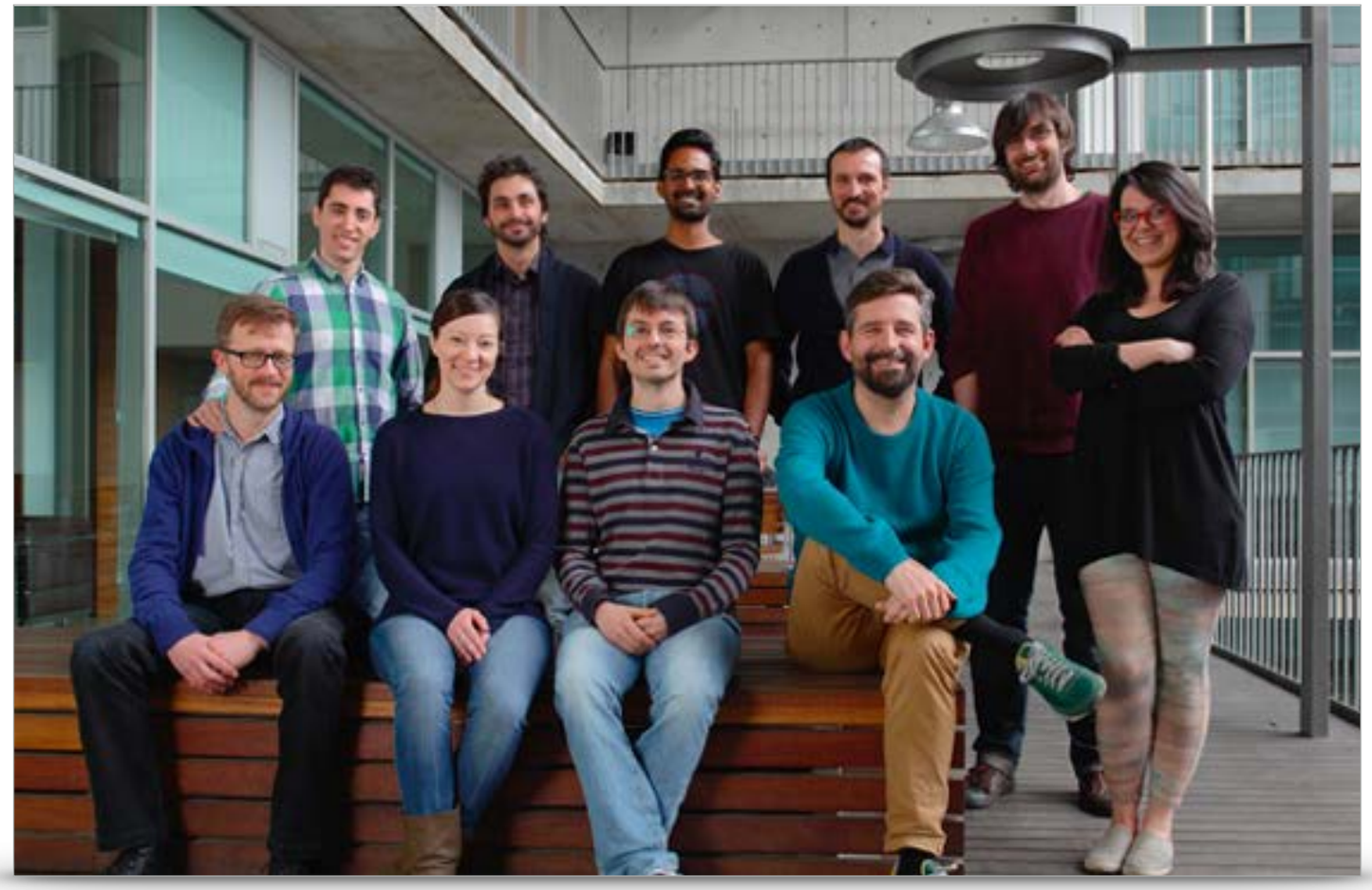
no need to worry much

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



François le Dily
Marie Trussart
Davide Baù

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

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

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