# 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

cnag CRG ICREA
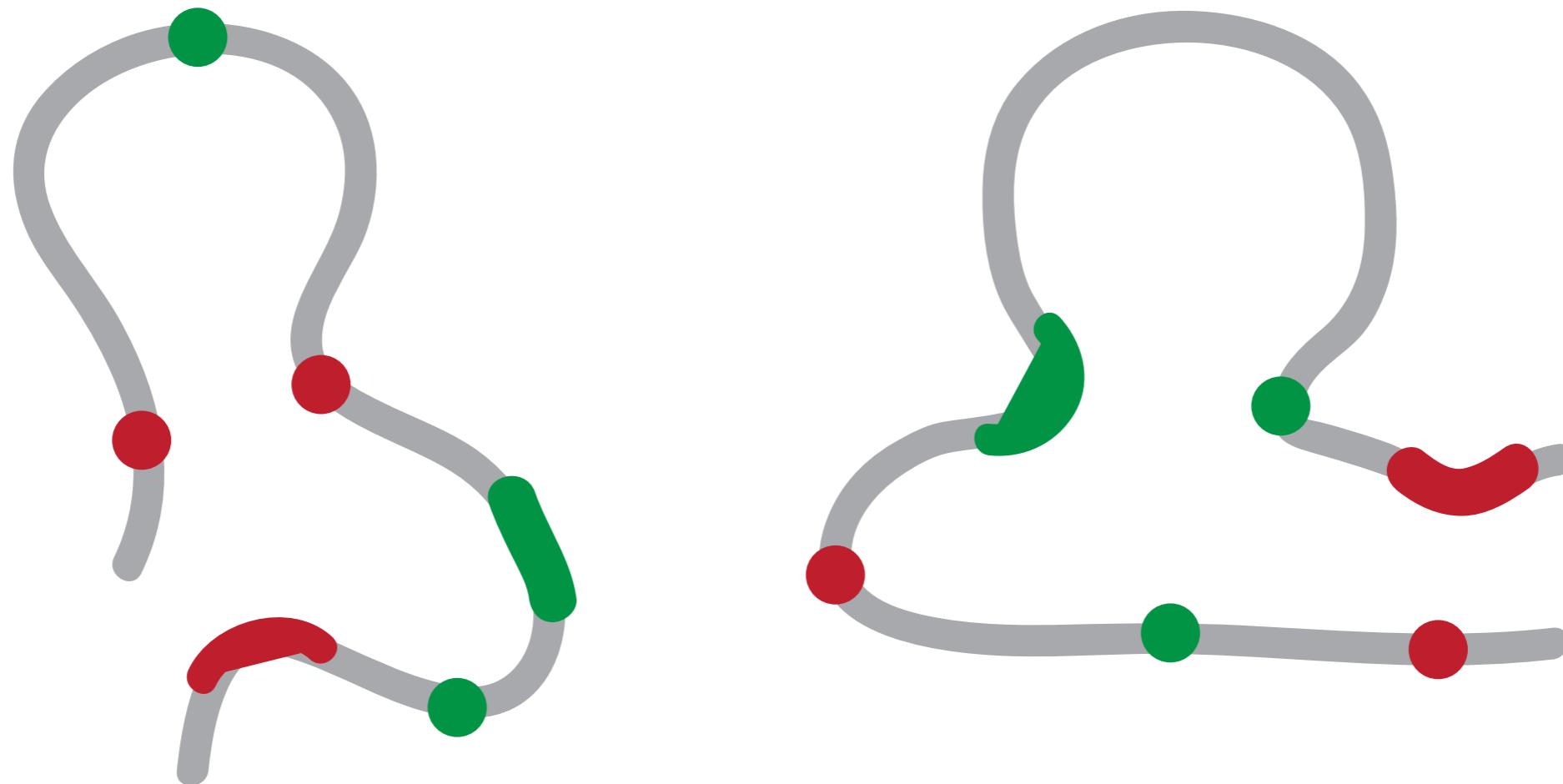
Complex genome organization

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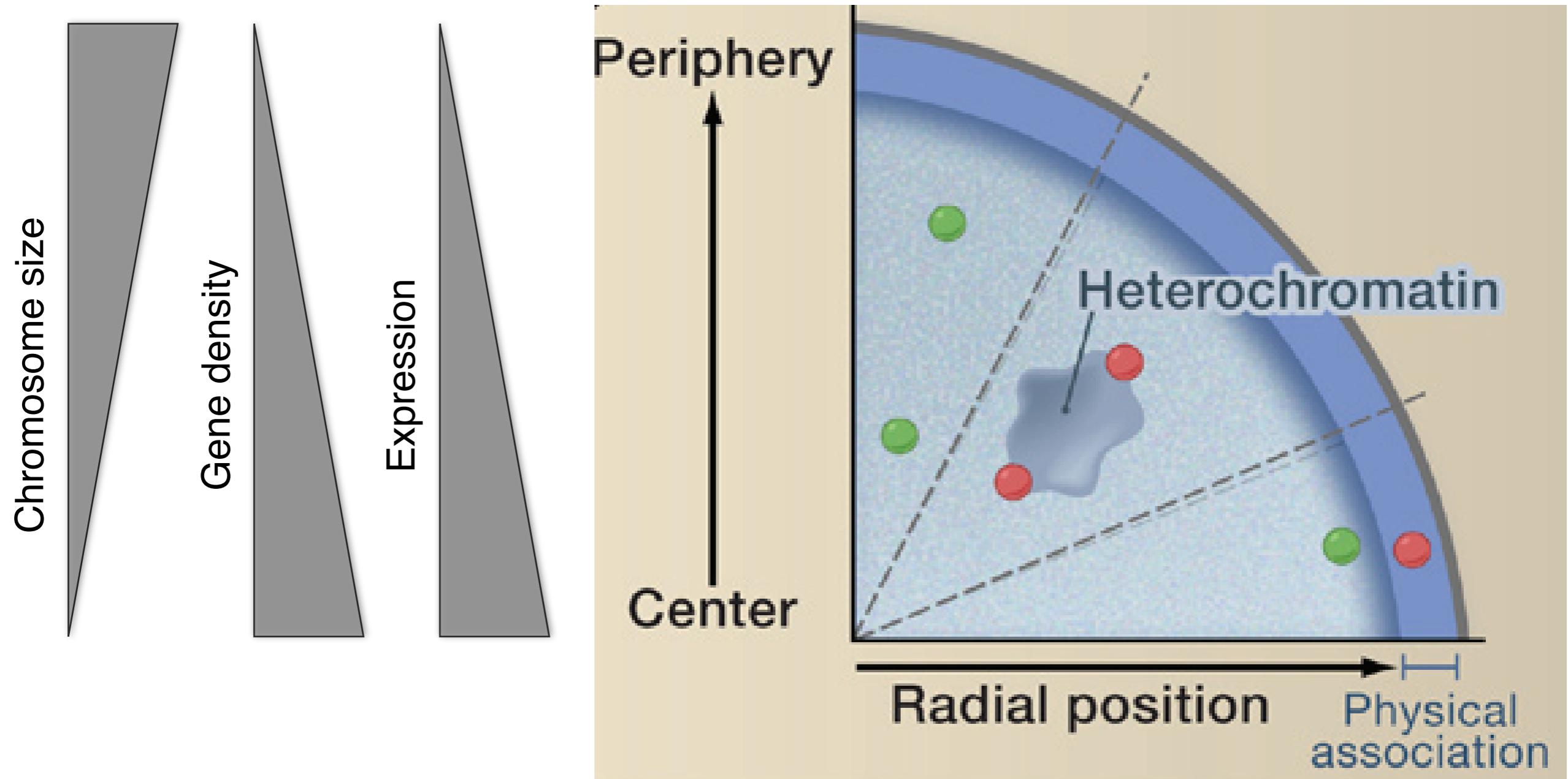
Complex genome organization

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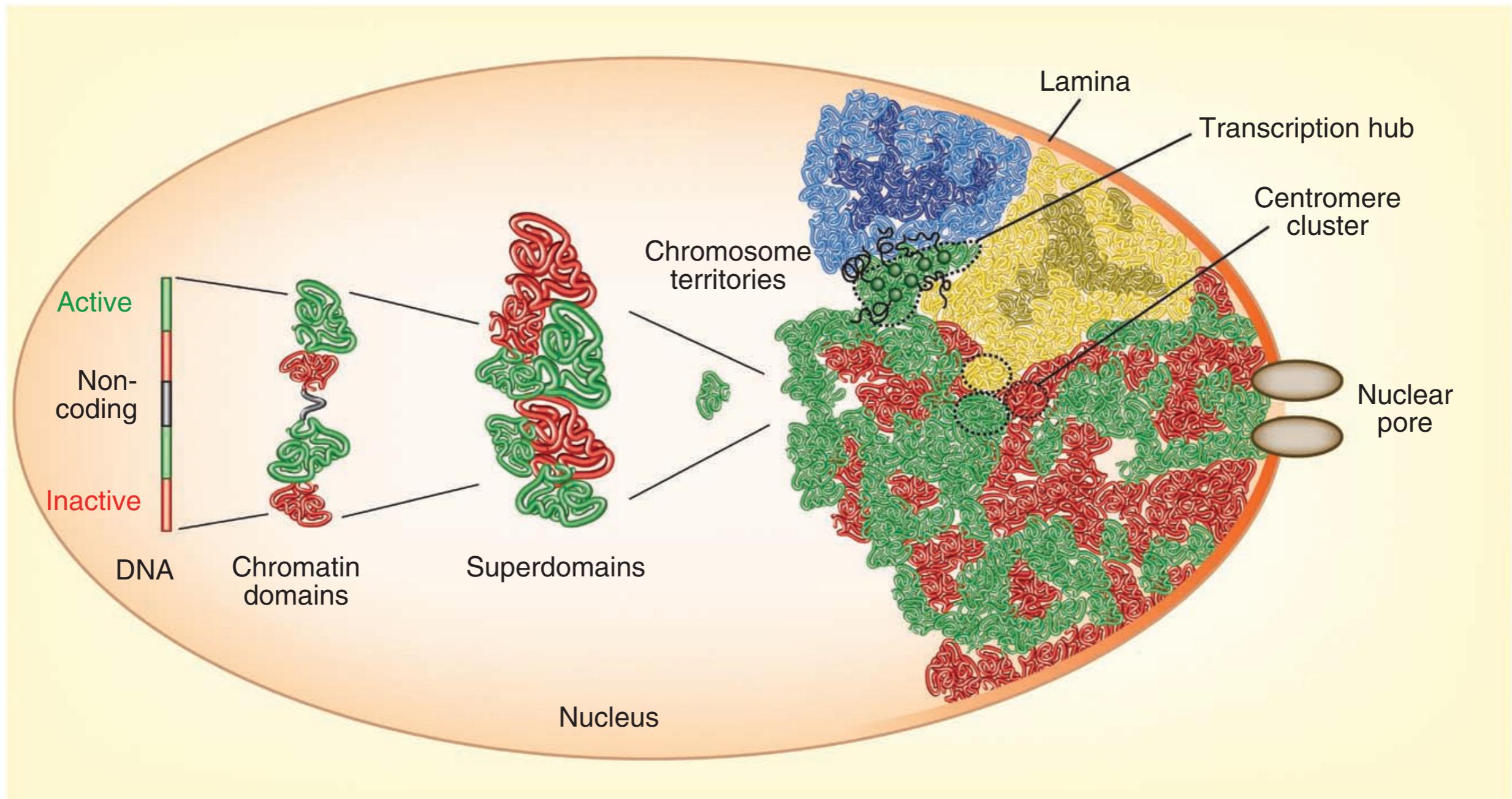
Marina Corral

Resolution Gap

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 µm³

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
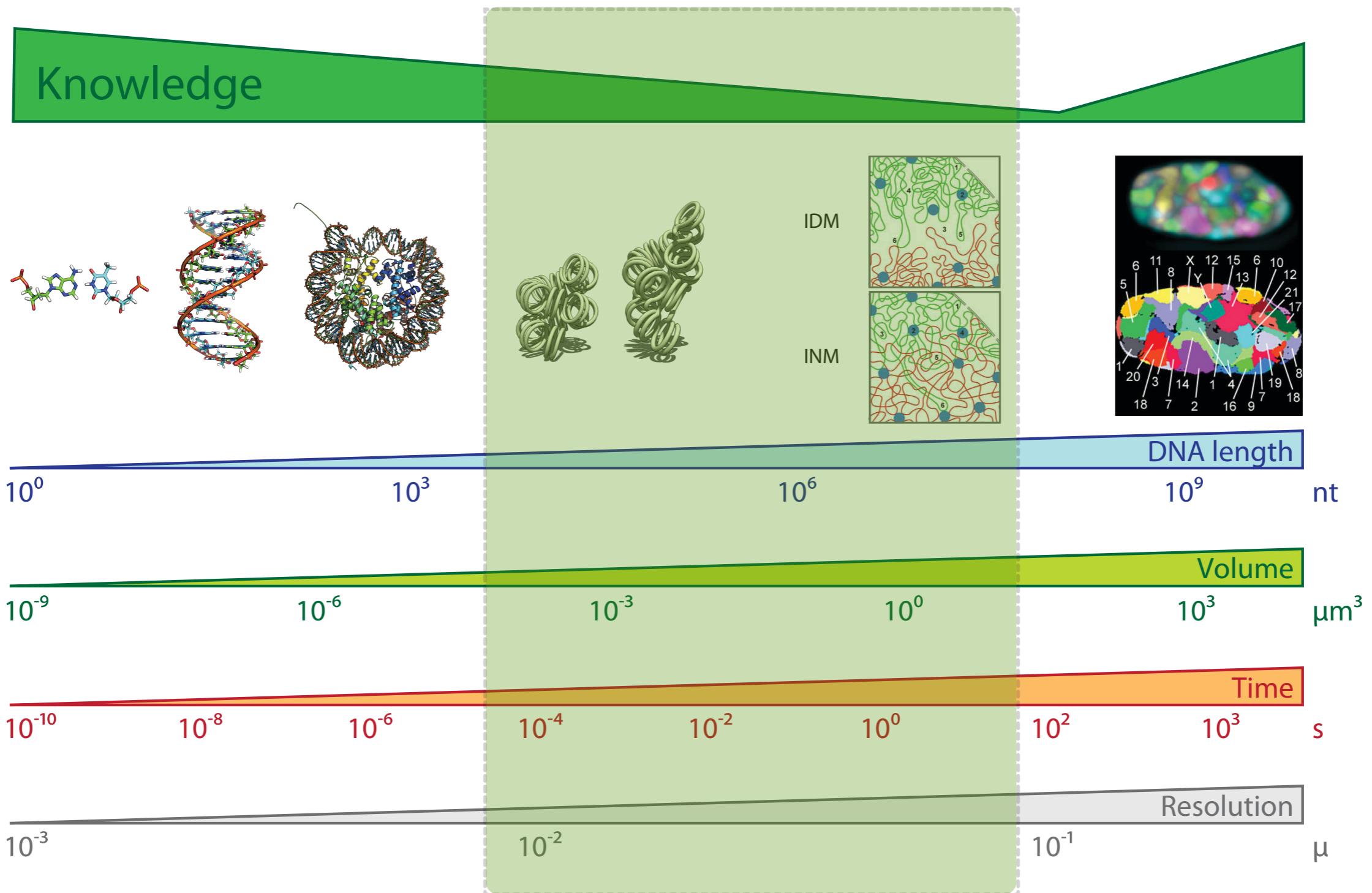
Hybrid Method

Experiments
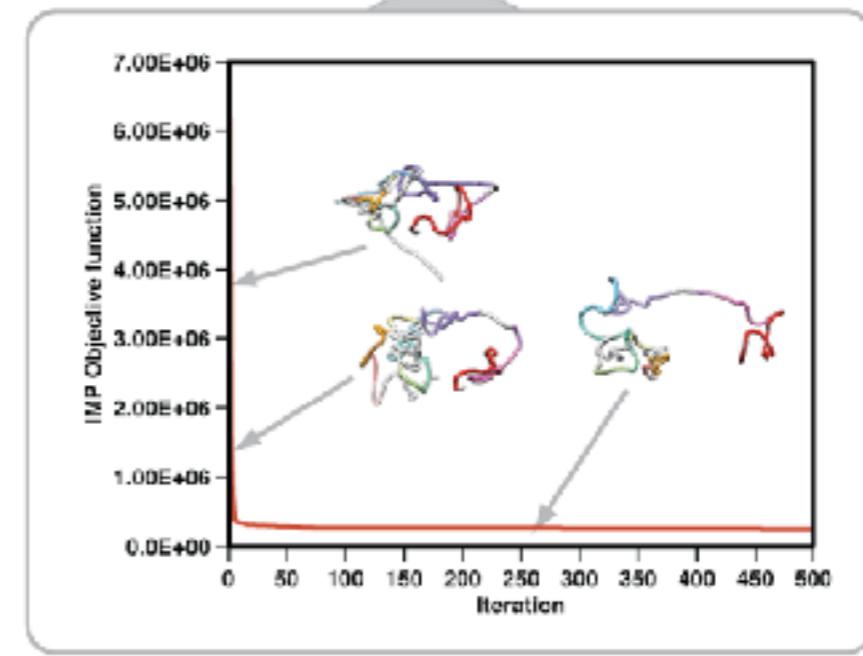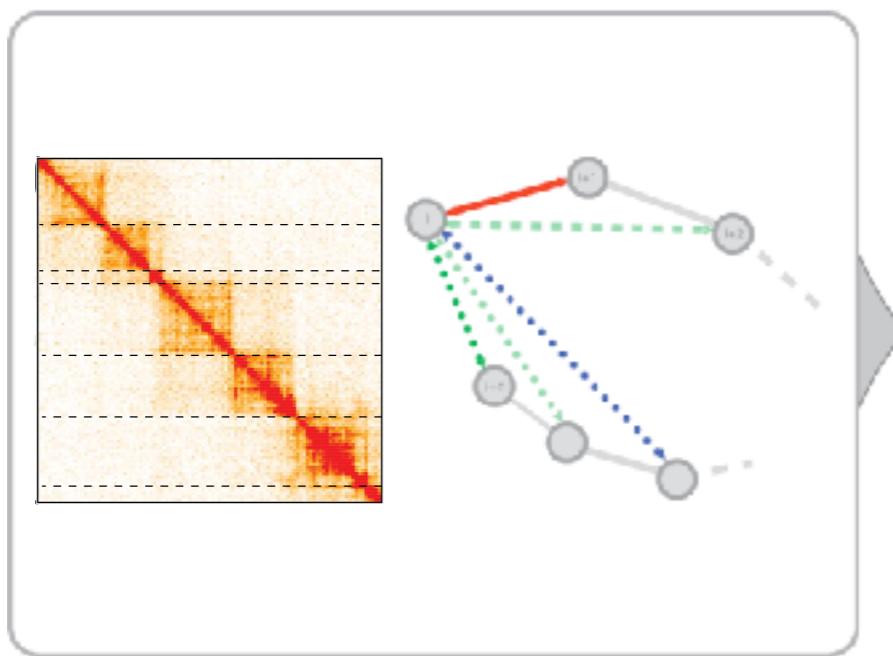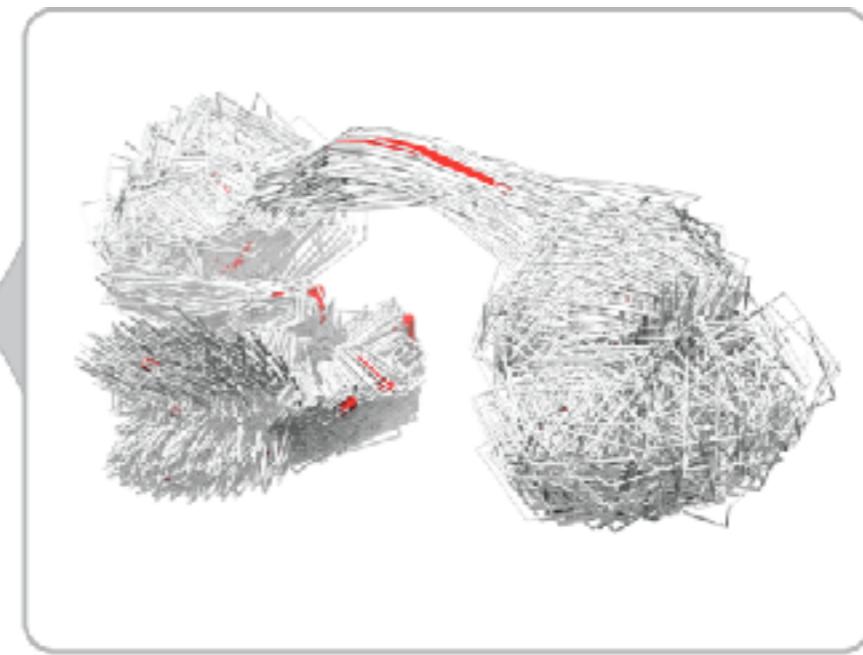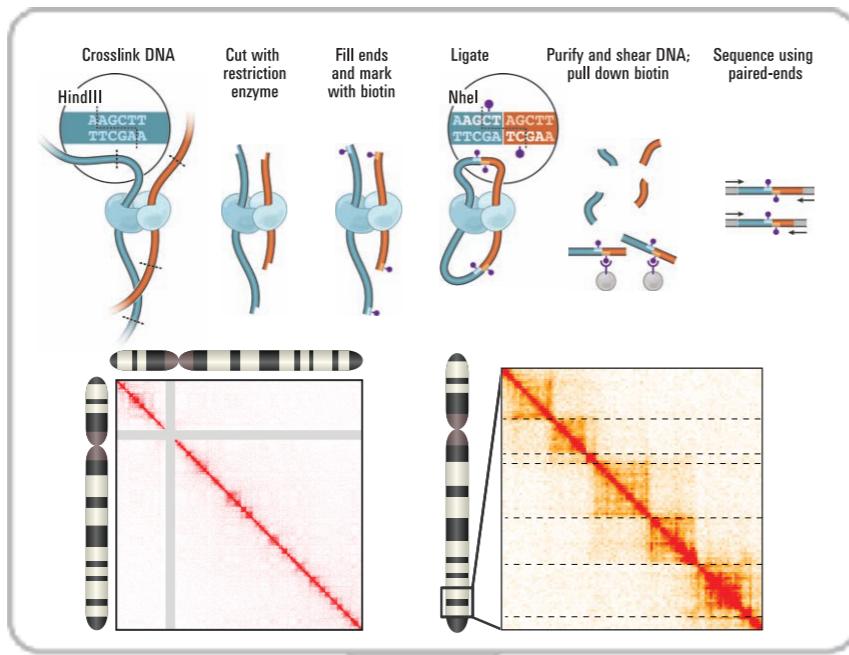
Computation

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

TADbit

http://3DGenomes.org
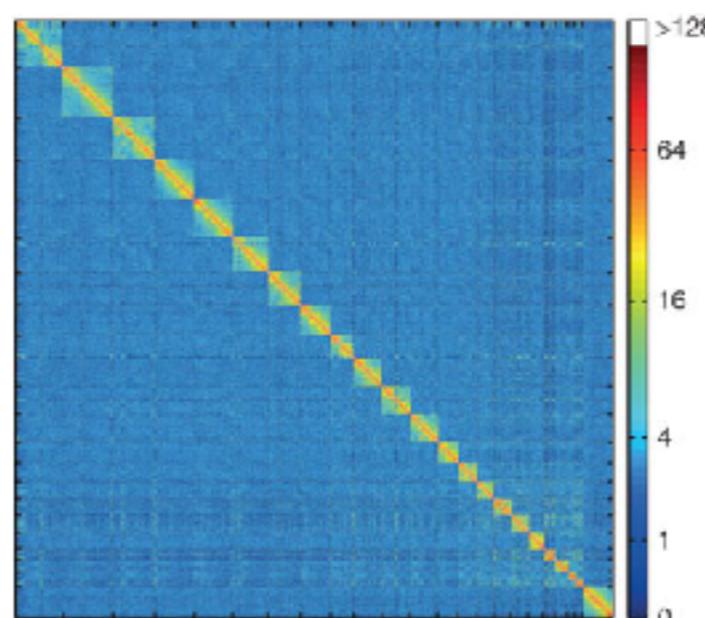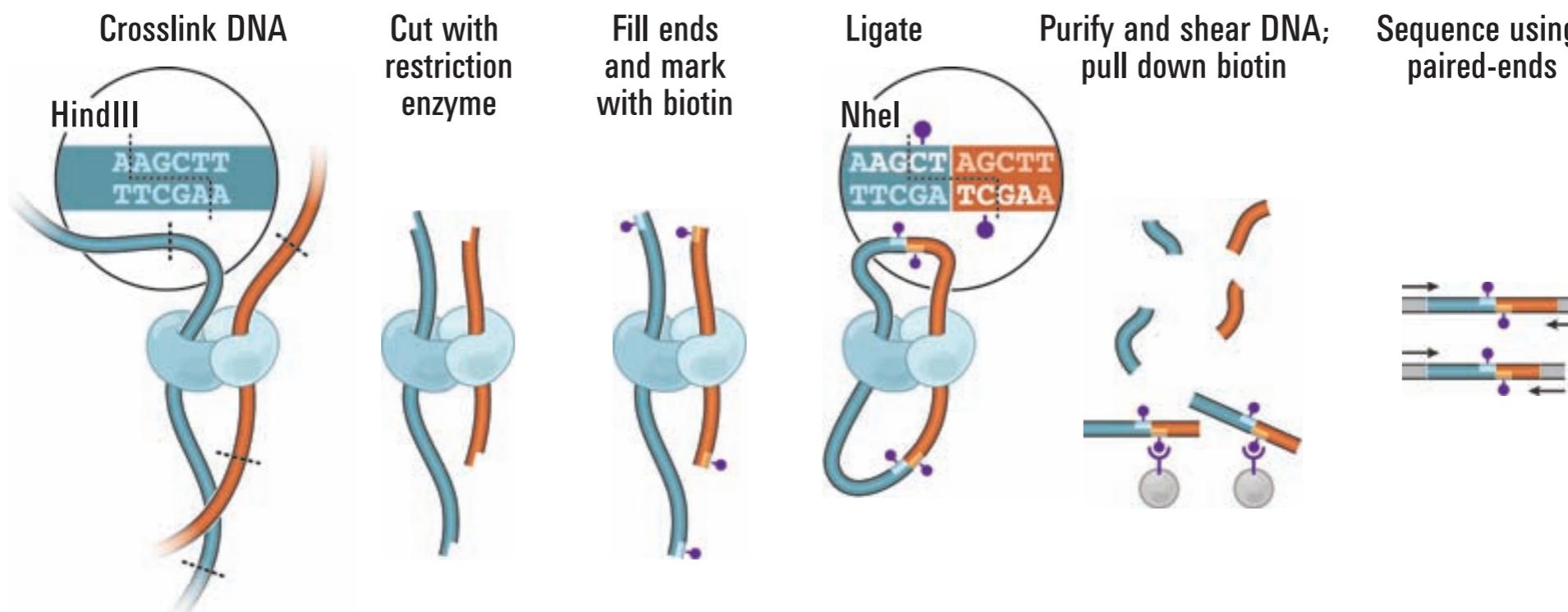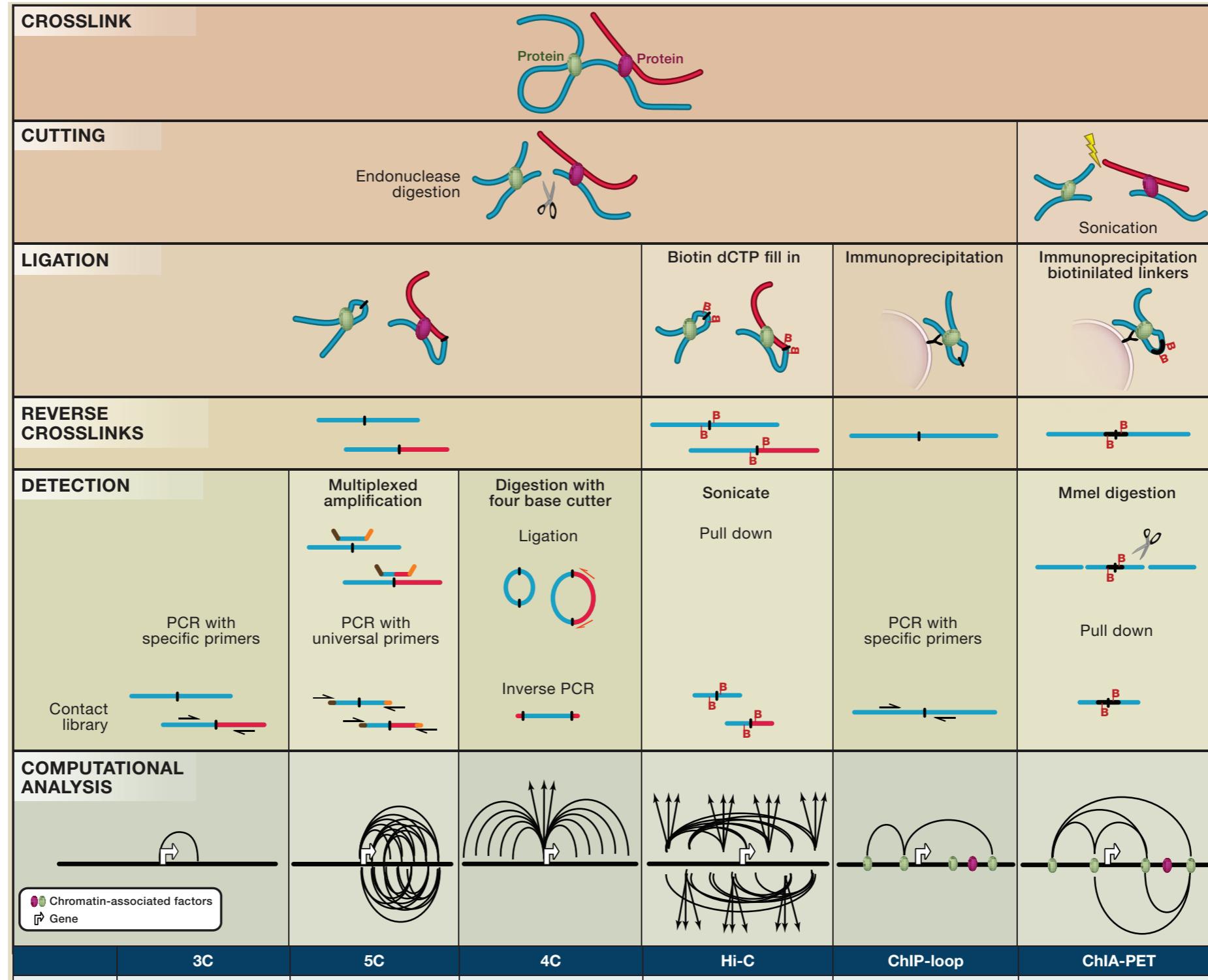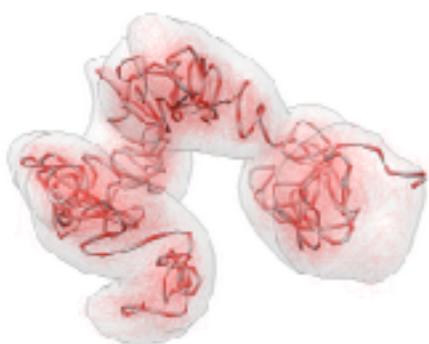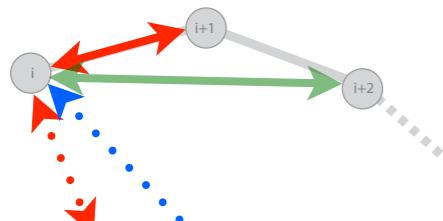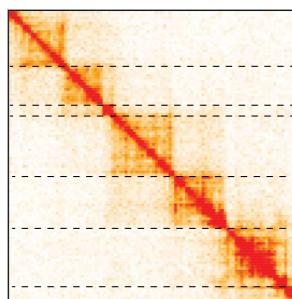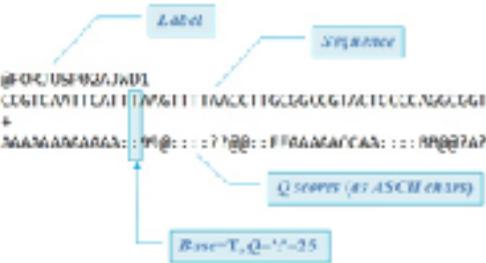
FastQ files to Maps

Map analysis

Model building

Model analysis
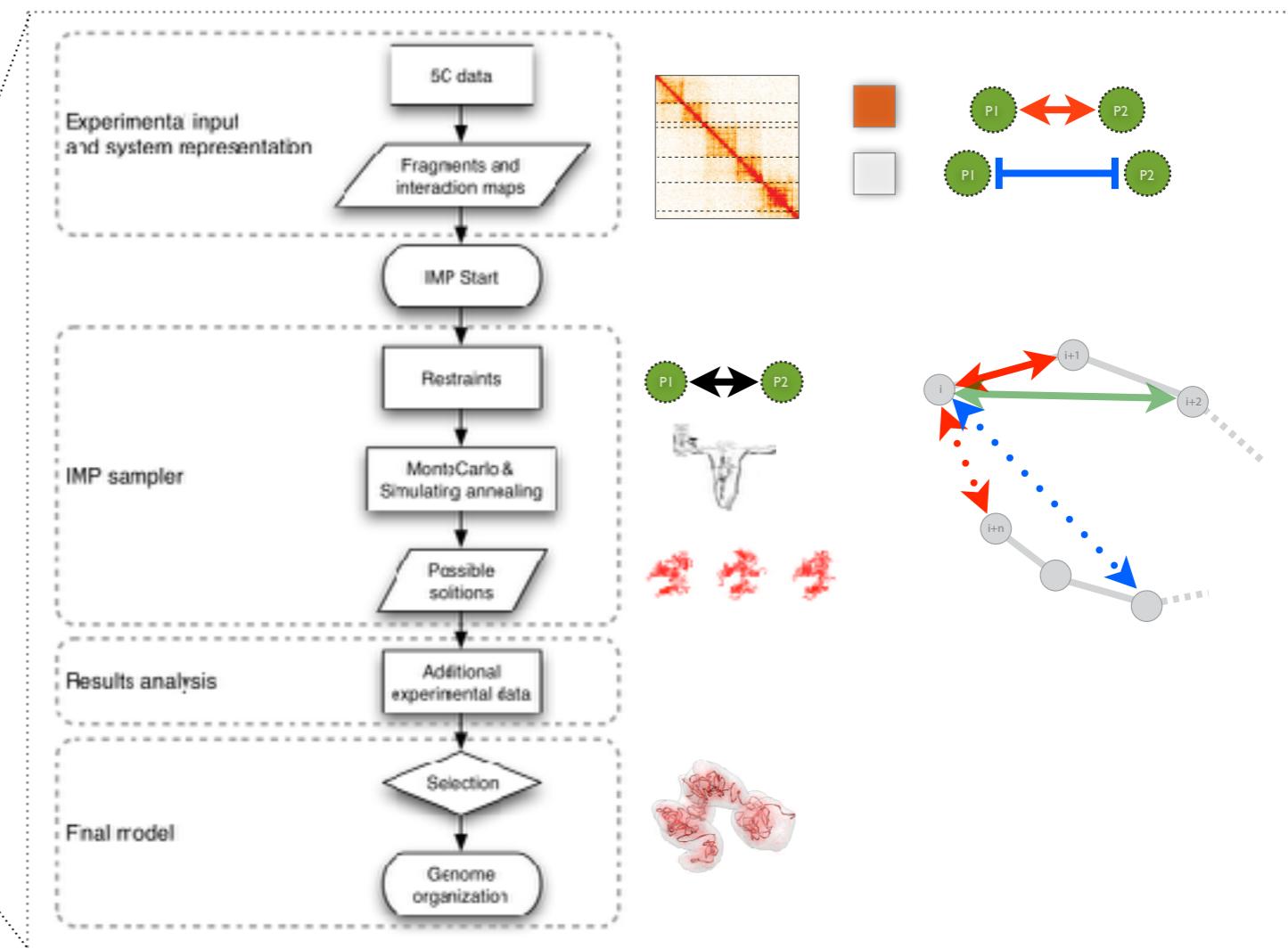
TADbit previous applications...

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)

Repressed TAD   Activated TAD

open   closed

closed   +Pg   closed

Structural transition

closed   open

Nucleosome   Histones H2A/H2B   Histone H1   PRBS
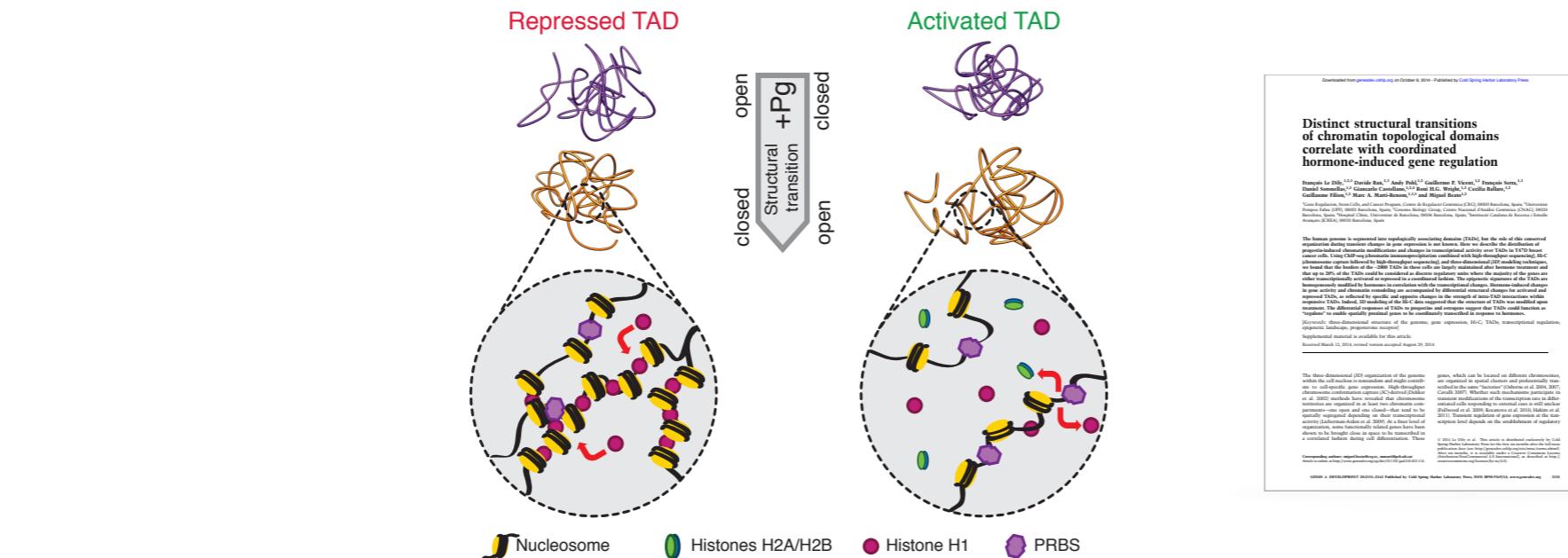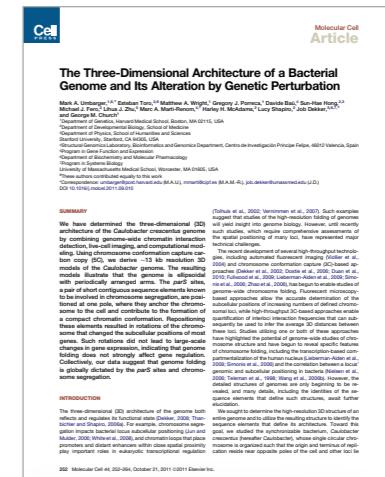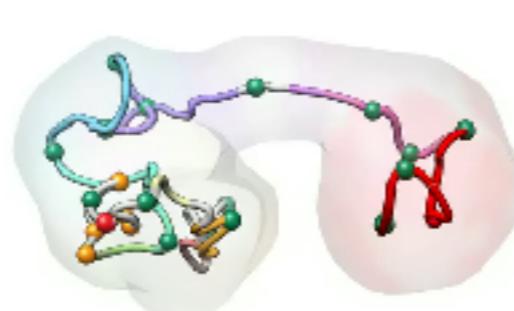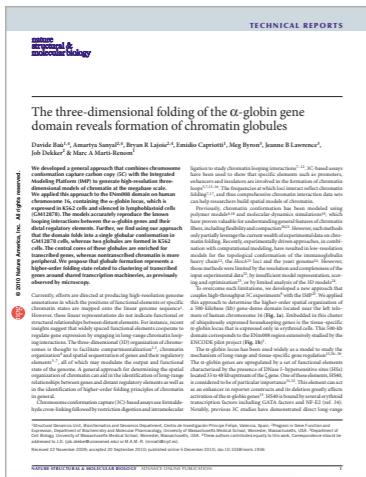
# Structuring the COLORs of chromatin
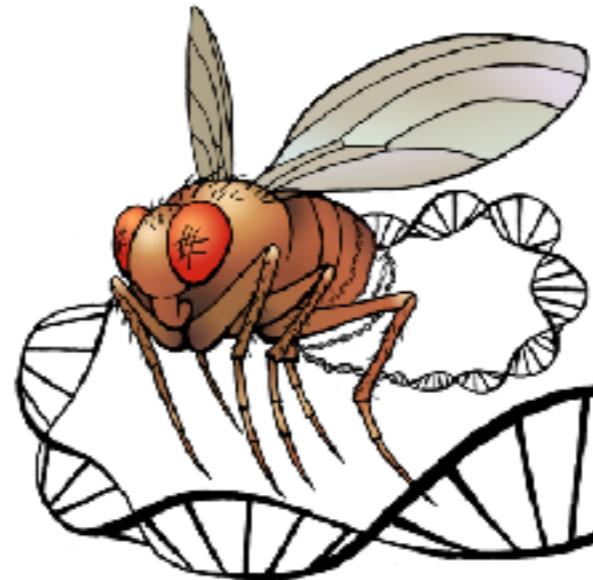
Fly Chromatin COLORs

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

drawing by Guillaume Filion

Position on chr2L (kb)

53 chromatin proteins
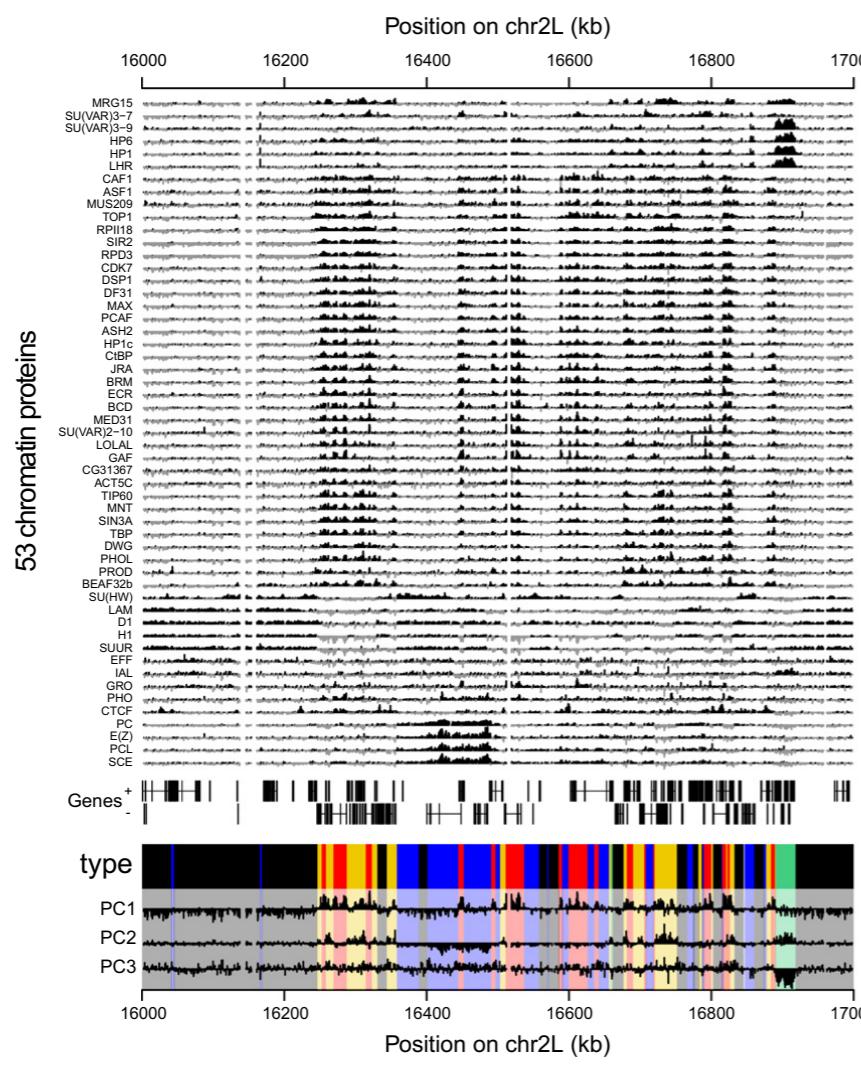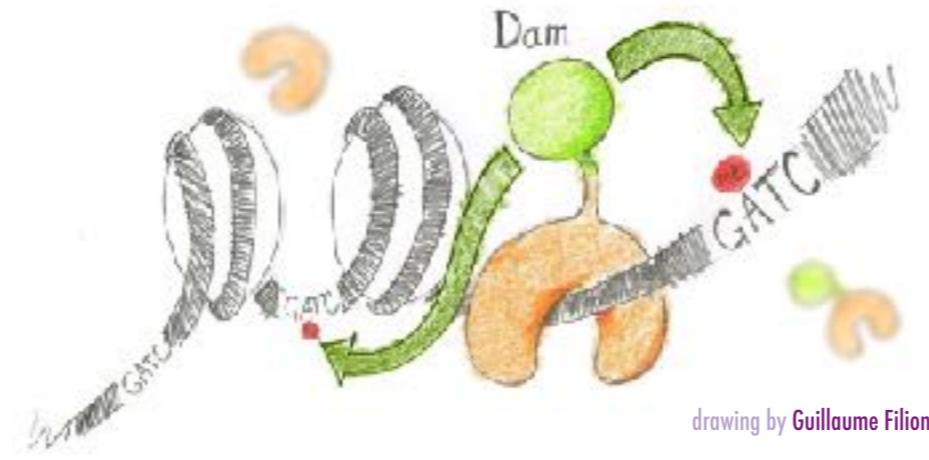
Principal component analysis

Hidden Markov model

GREEN
HP1

YELLOW
Active

RED
Active

BLUE
Polycomb

BLACK
Repressive

chromatin proteins

Nucleus
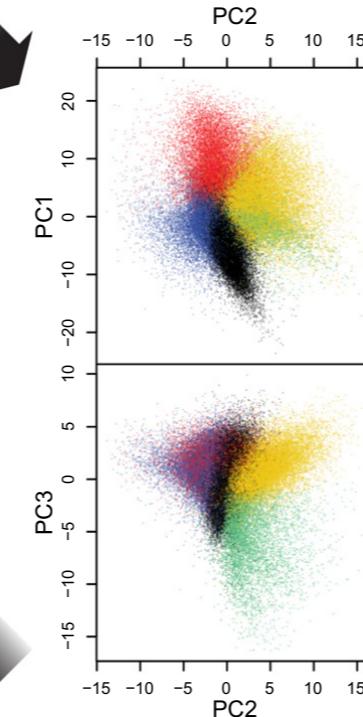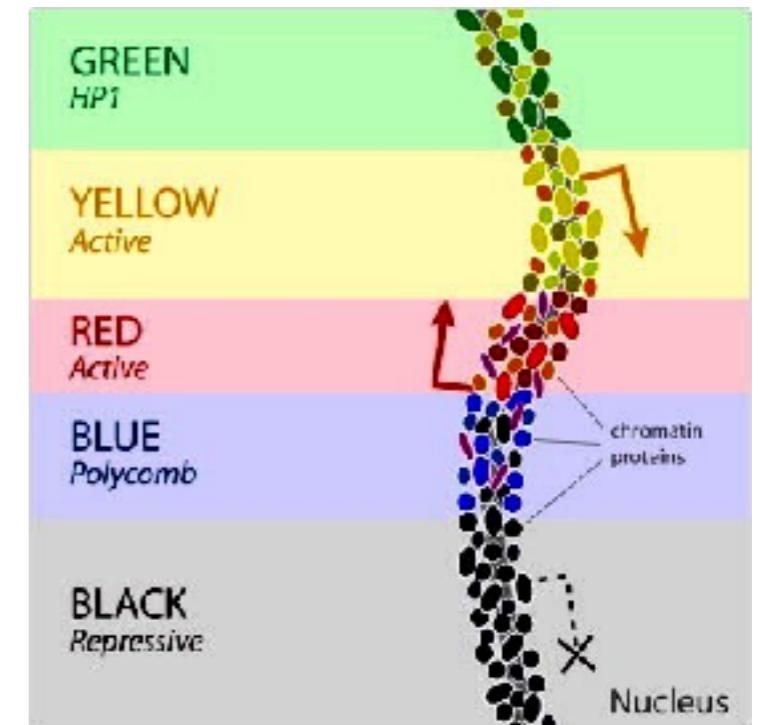
Fly Chromatin COLORs

Hou et al. (2012). Molecular Cell, 48(3), 471–484.
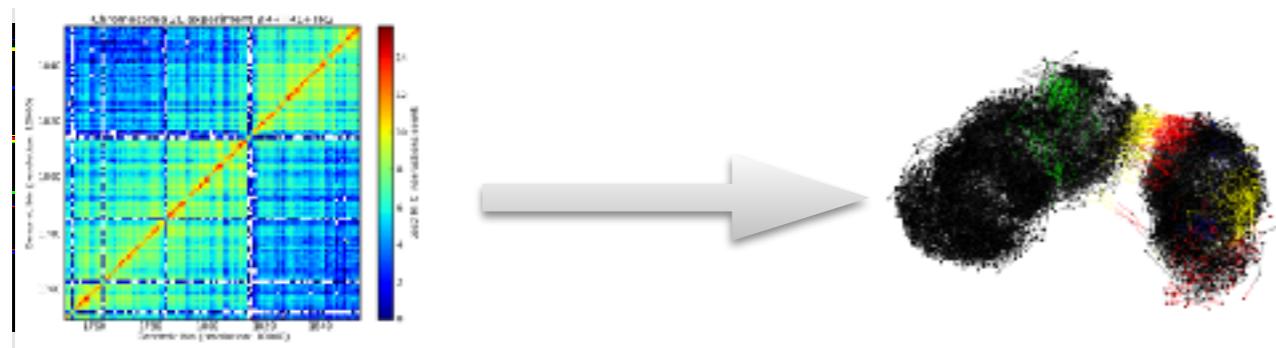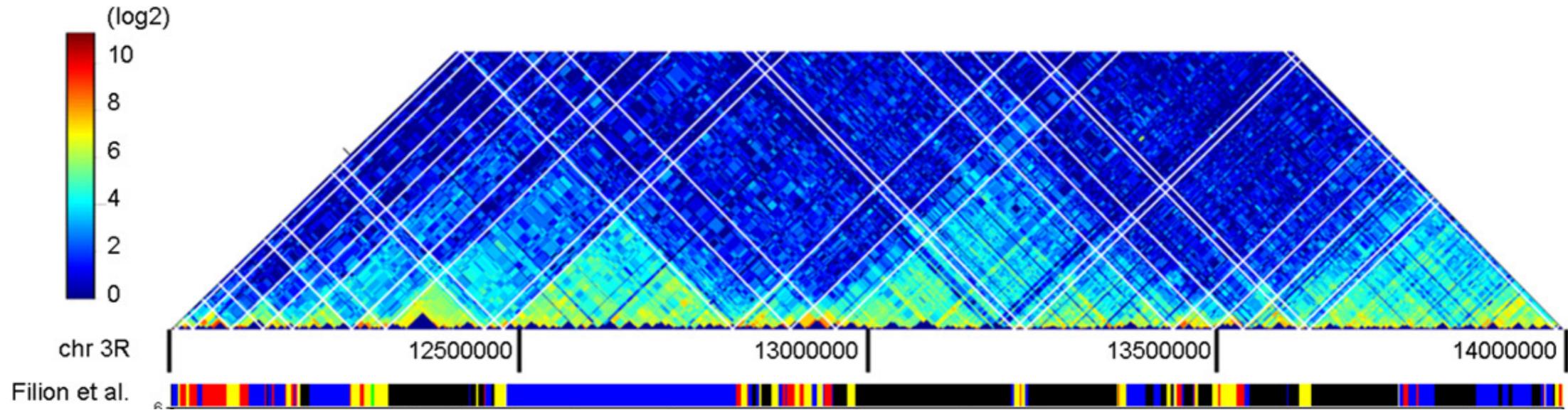
~200 regions of ~5Mb each
2Kb resolution

Model accuracy

Model accuracy

Boettiger, A. N., et al. (2016). Nature, 1–15.
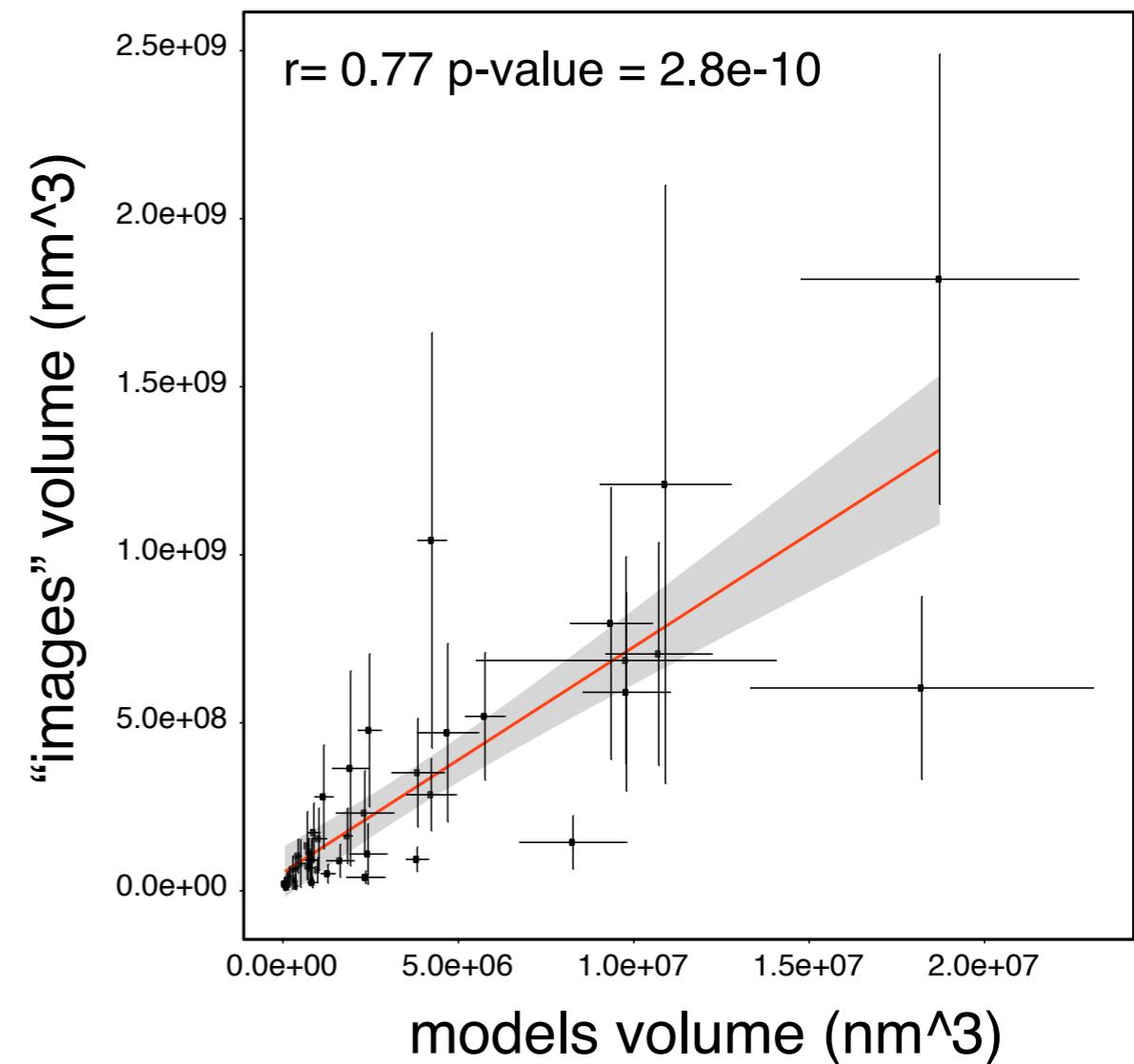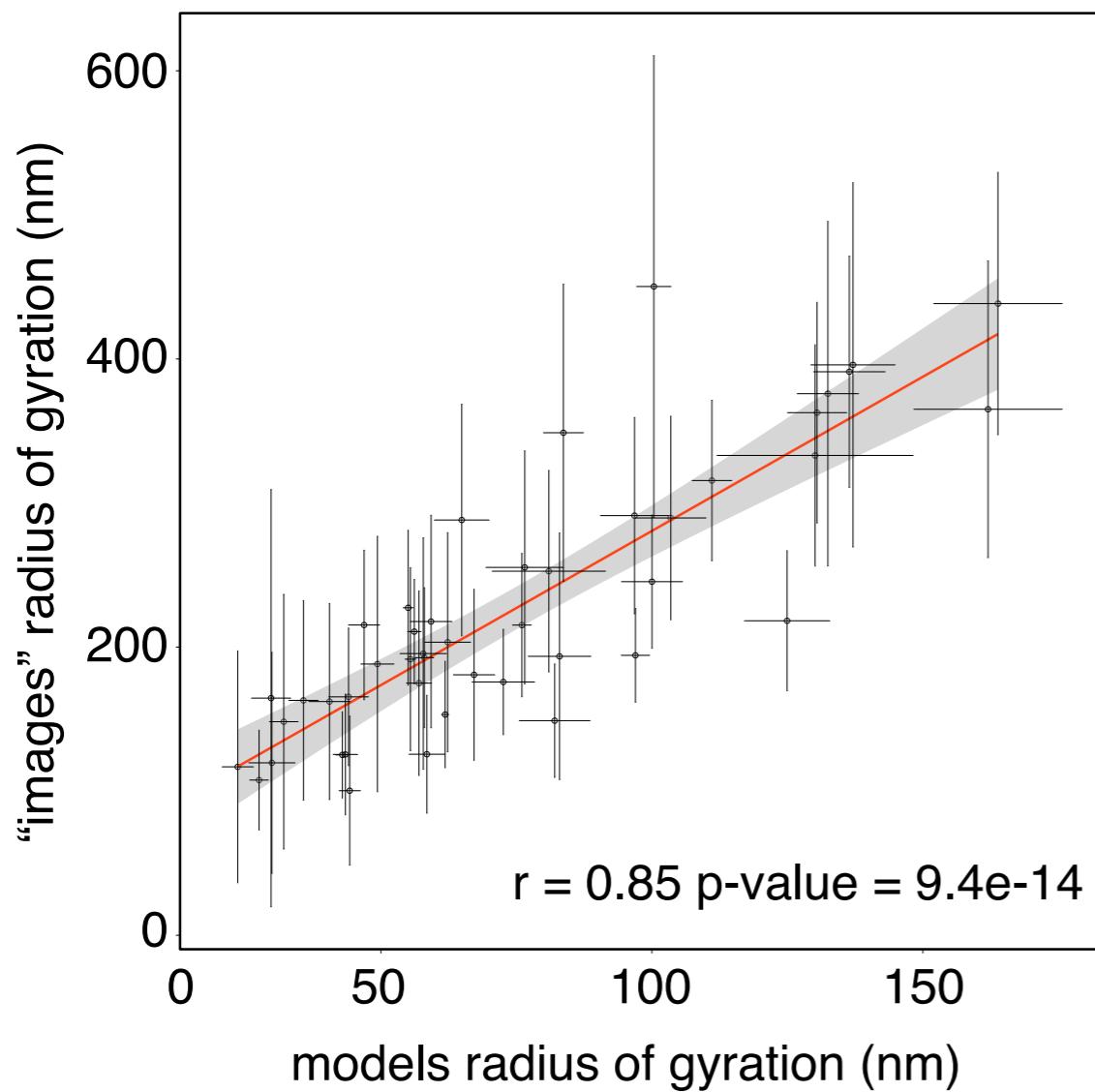
Structural properties

50 1Mb regions. 10 enriched for each color.

| RED dense region	YELLOW dense region	GREEN dense region	BLUE dense region	BLACK dense region
3R:18920000-19920000	X:15590000-16600000	2R:510000-1530000	3L:210000-1230000	2L:17500000-18530000
22% 17% 0% 11% 45% 6%	0% 48% 4% 20% 26% 3%	0% 0% 100% 0% 0% 0%	11% 17% 0% 52% 13% 0%	1% 0% 0% 0% 98% 1%
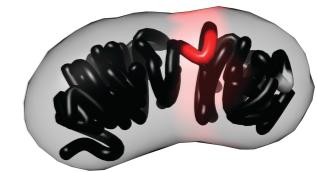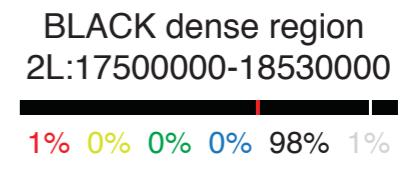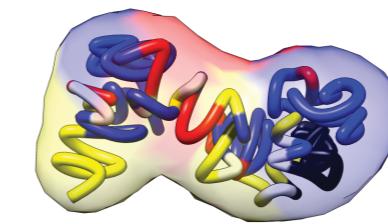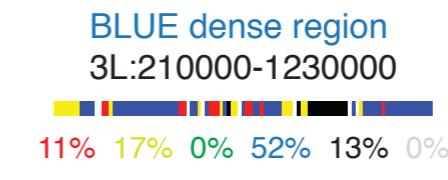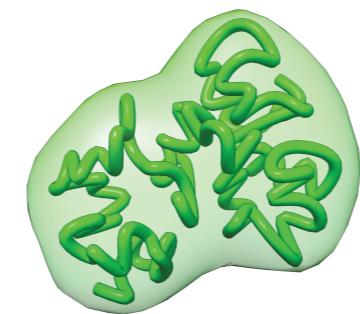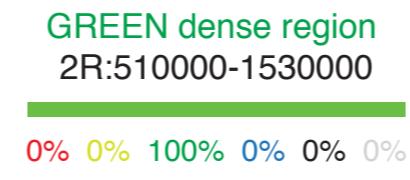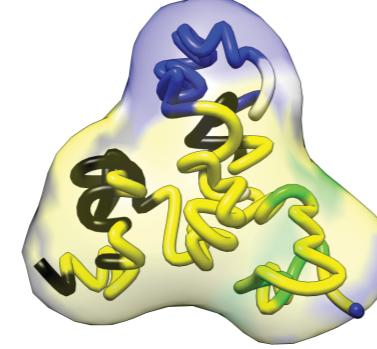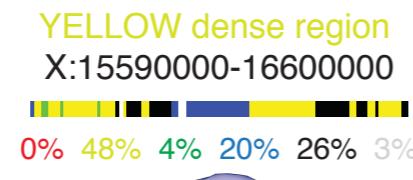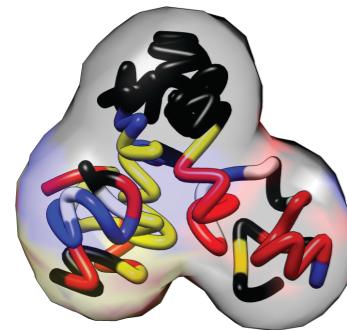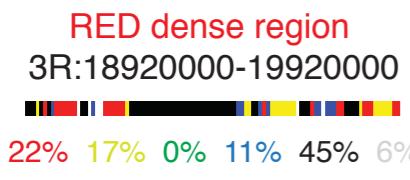
Accessibility (%)
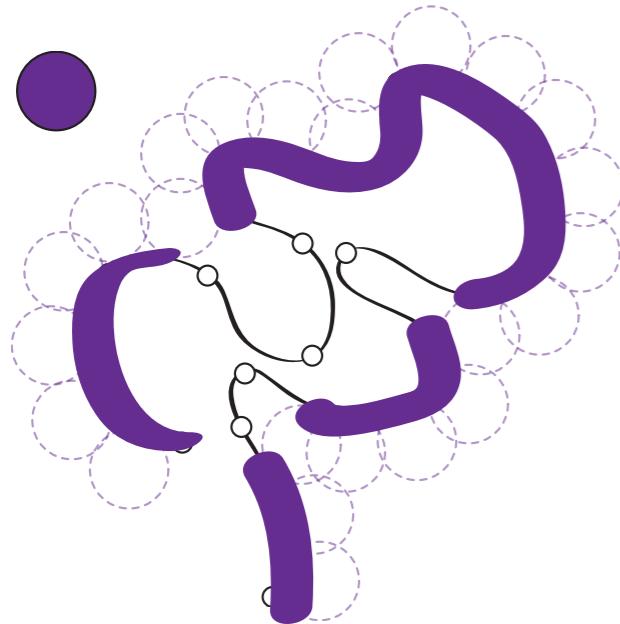
Density (bp/nm)
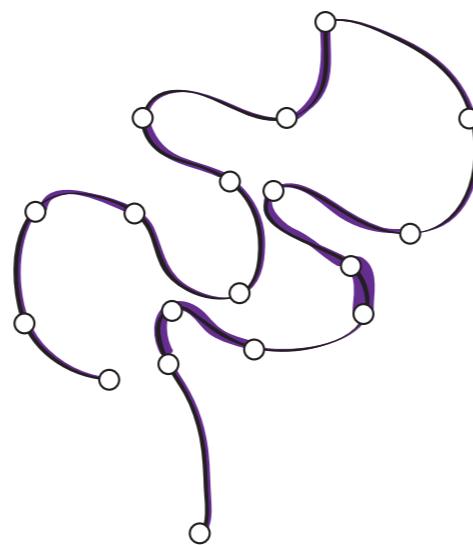
Interactions
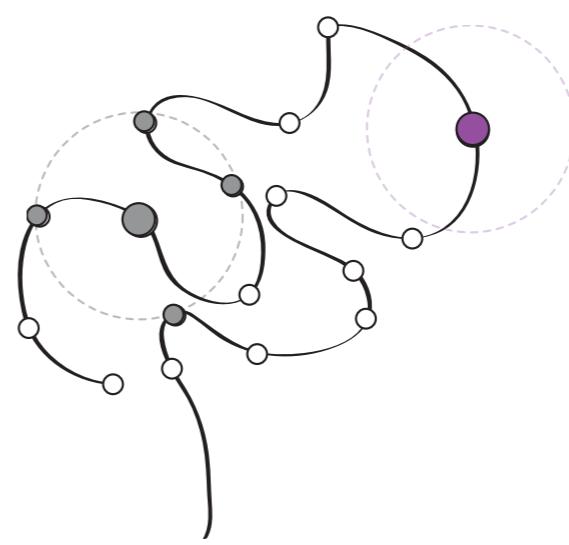
Angle
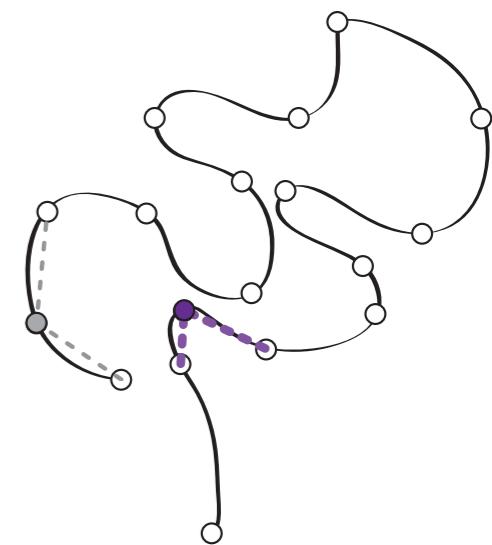
Structural COLORs
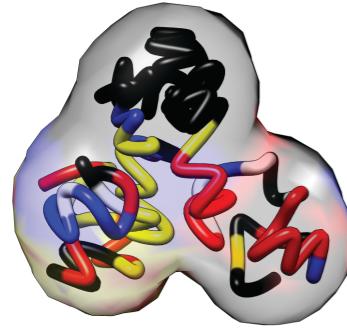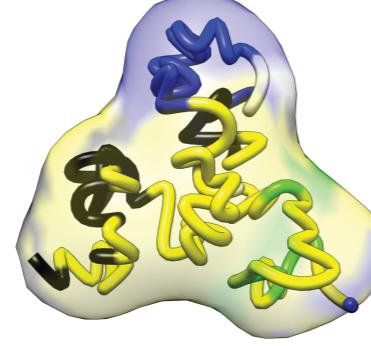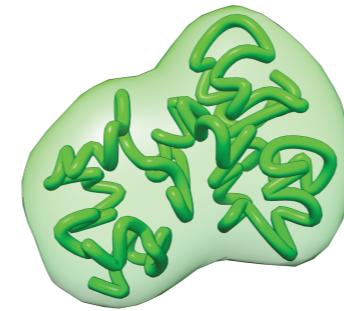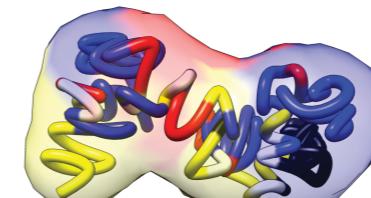
RED dense region	YELLOW dense region	GREEN dense region	BLUE dense region	BLACK dense region
3R:18920000-19920000	X:15590000-16600000	2R:510000-1530000	3L:210000-1230000	2L:17500000-18530000
22% 17% 0% 11% 45% 6%	0% 48% 4% 20% 26% 3%	0% 0% 100% 0% 0% 0%	11% 17% 0% 52% 13% 0%	1% 0% 0% 0% 98% 1%
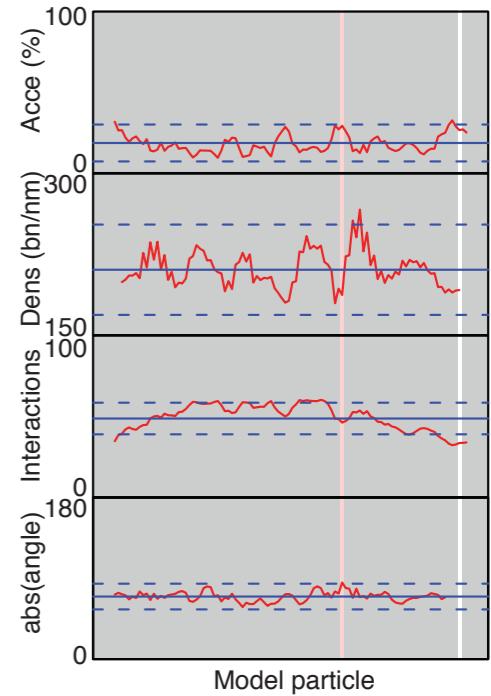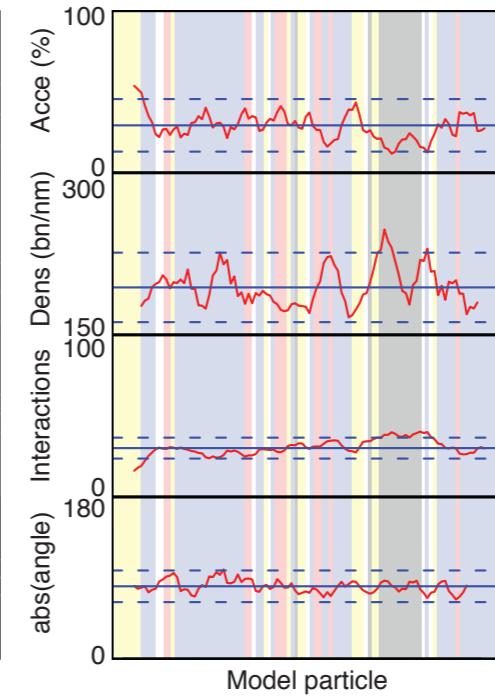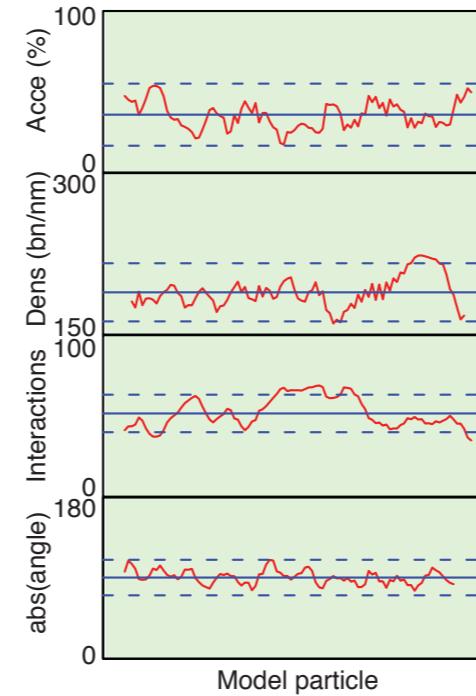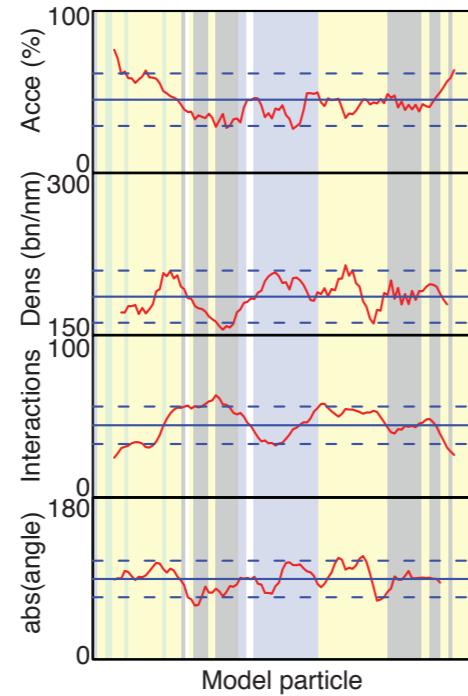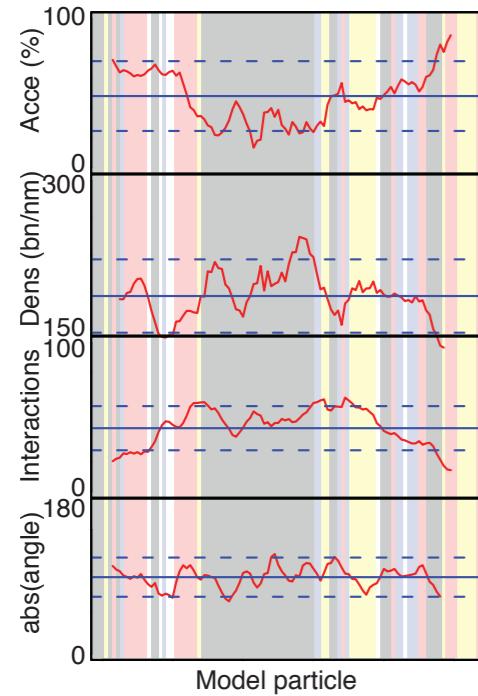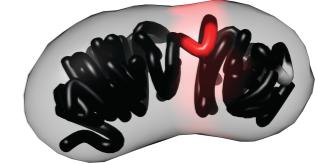
Structural **COLORs**
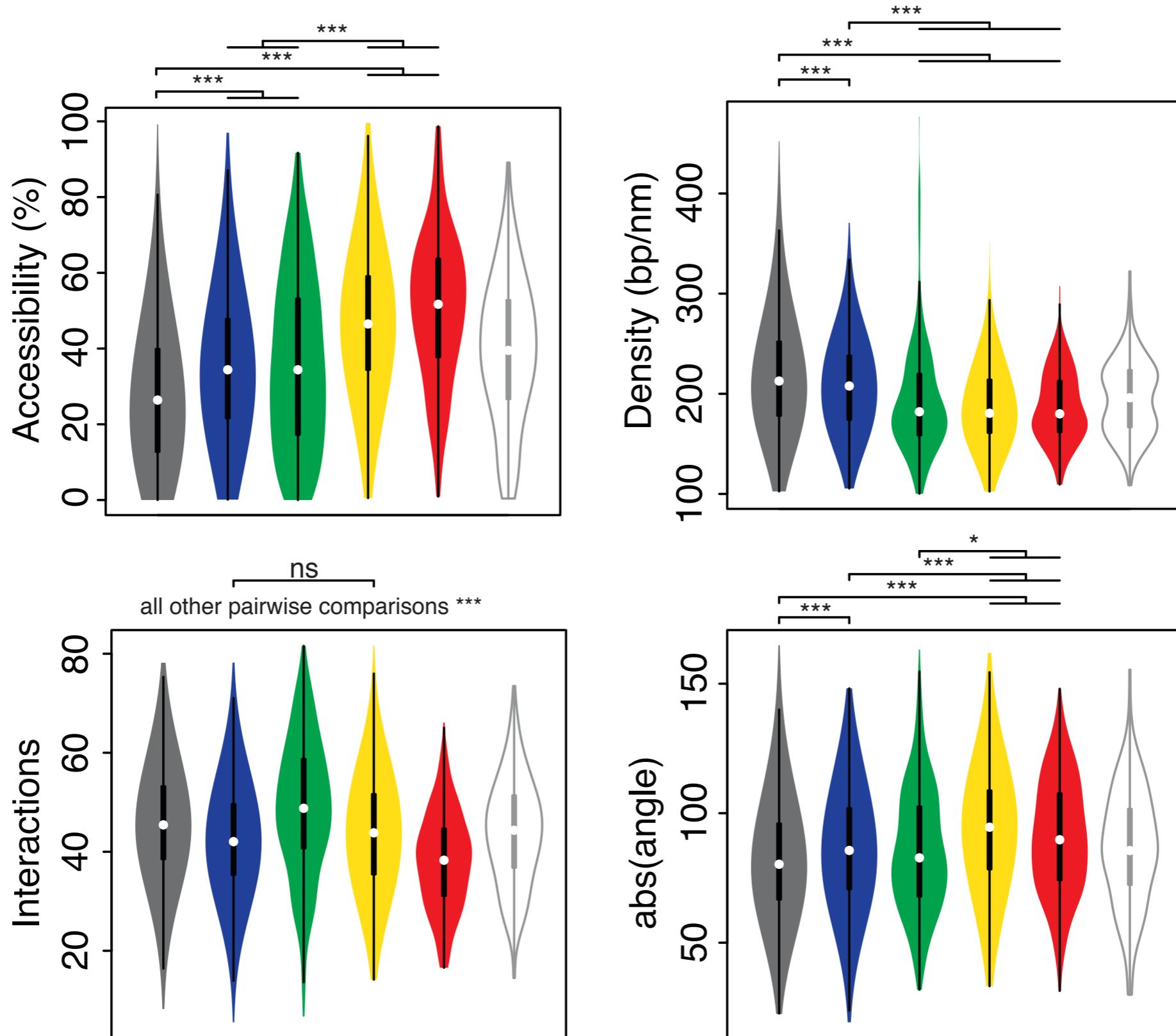
Structural COLORs
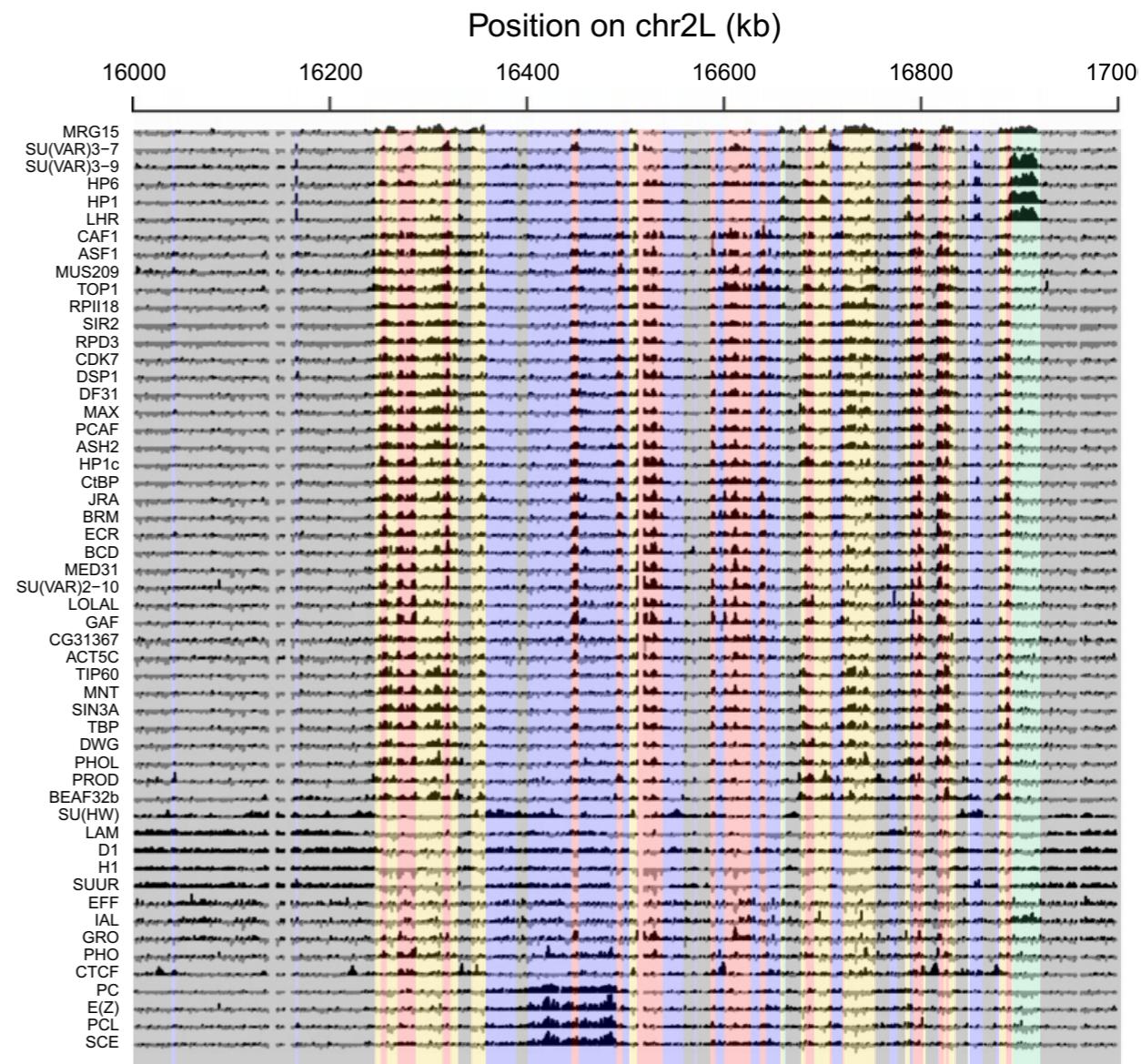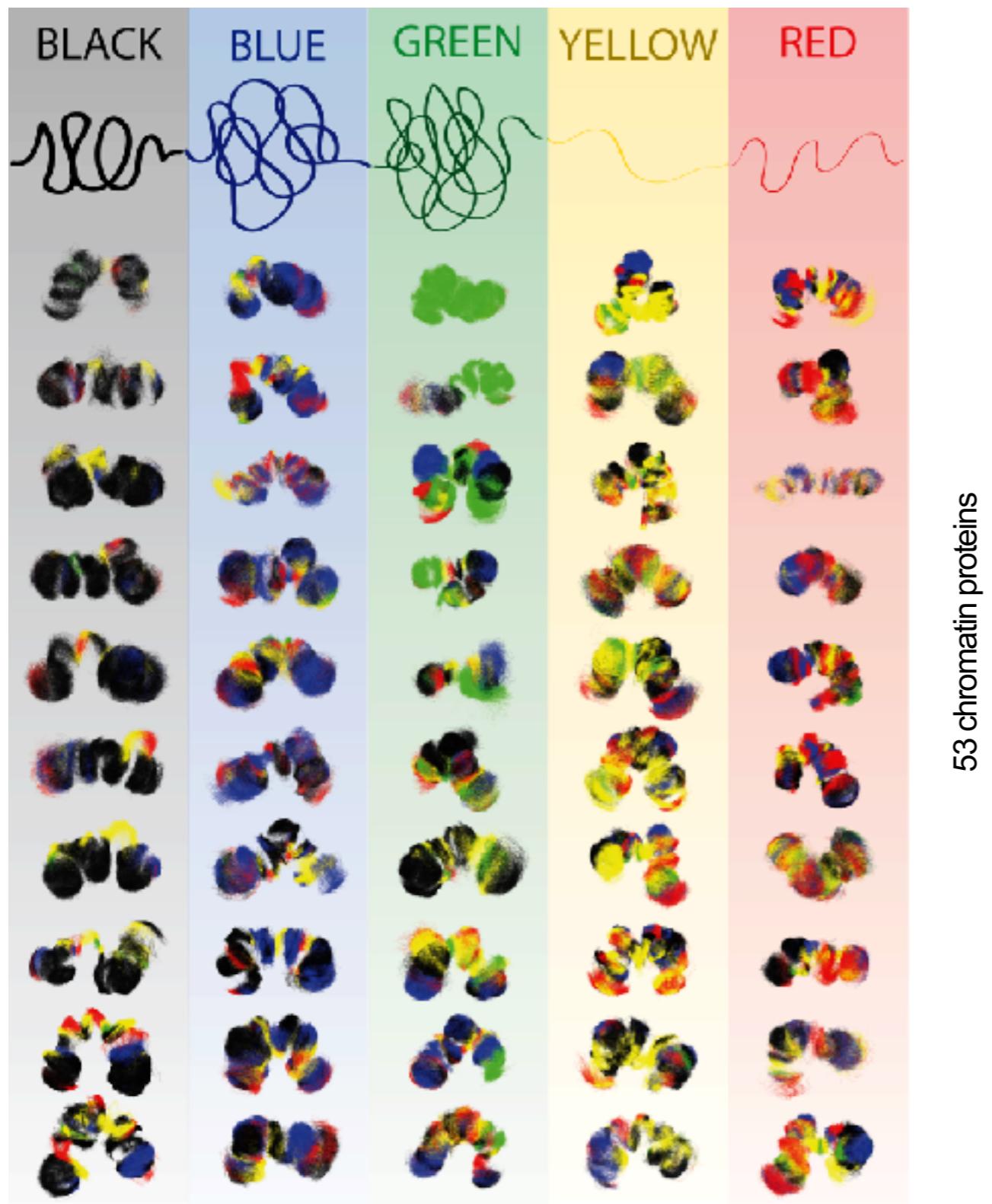
Are the models correct?
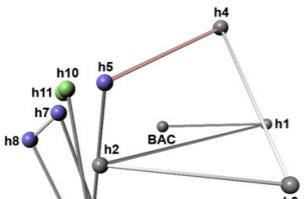
Jhunjhunwala (2008) Cell
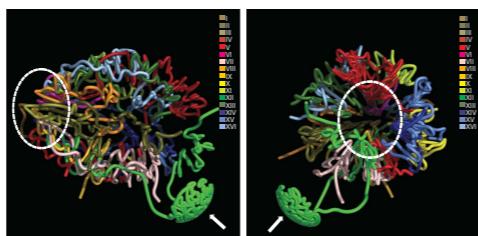
Duan (2010) Nature
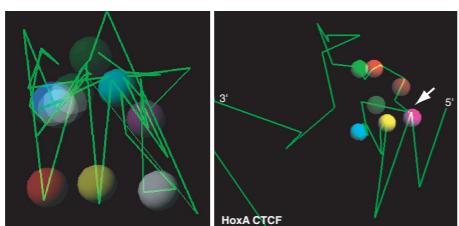
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research
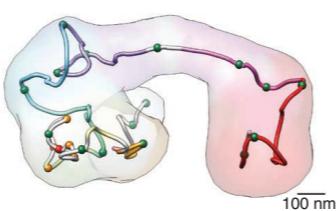
Baù (2011) Nature Structural & Molecular Biology
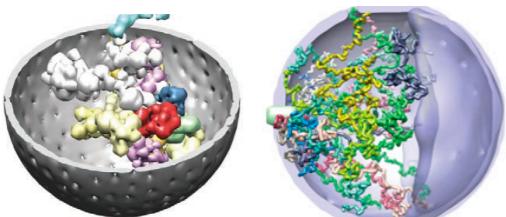
Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research
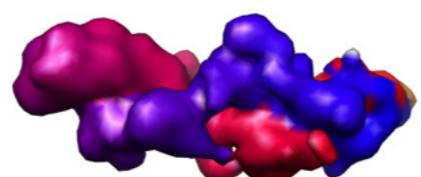
Umbarger (2011) Molecular Cell
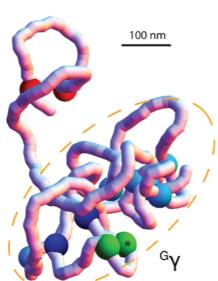
Junier (2012) Nucleic Acids Research
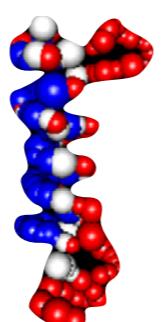
Hu (2013) PLoS Computational Biology

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

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

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

Toy models
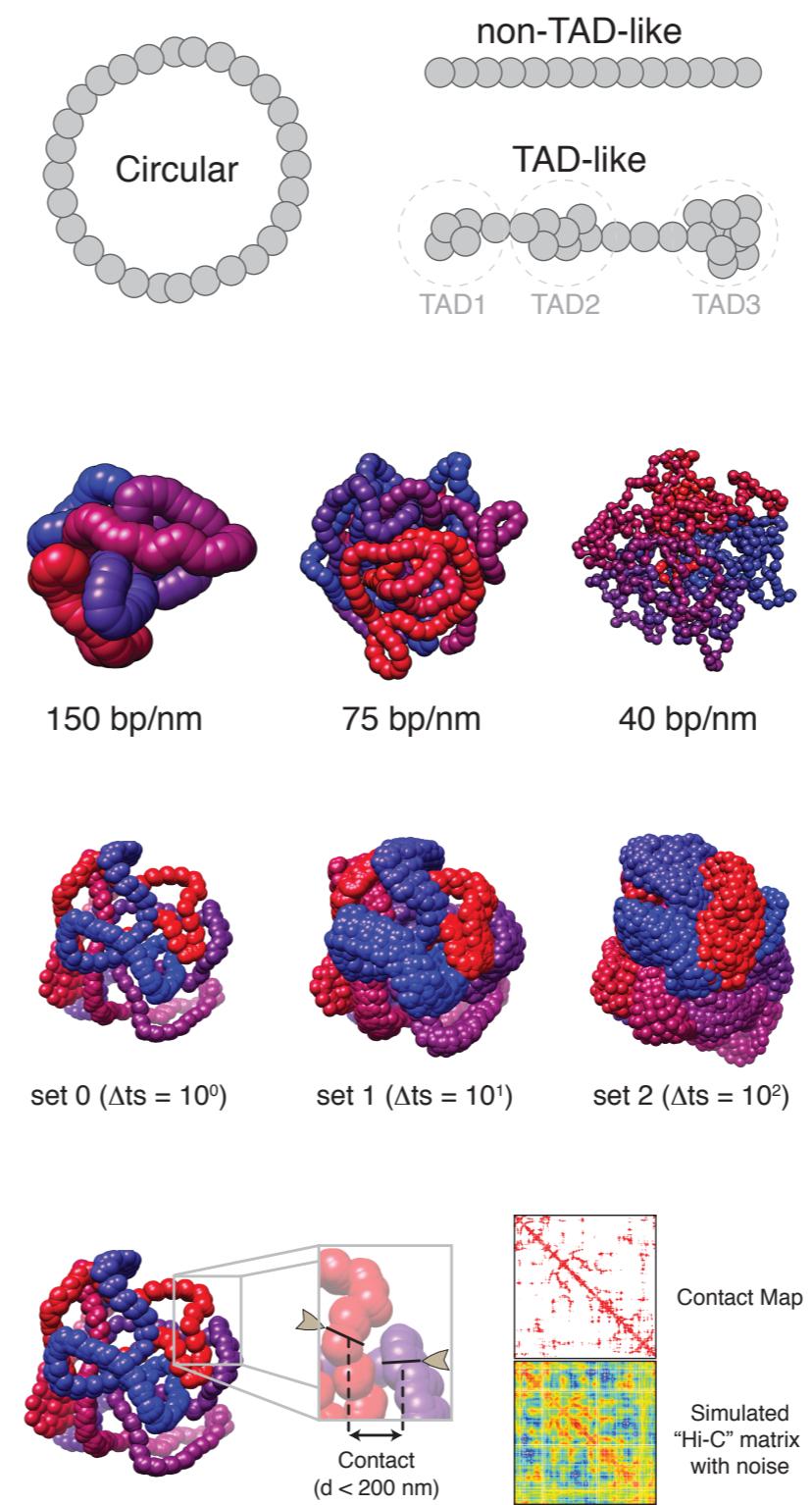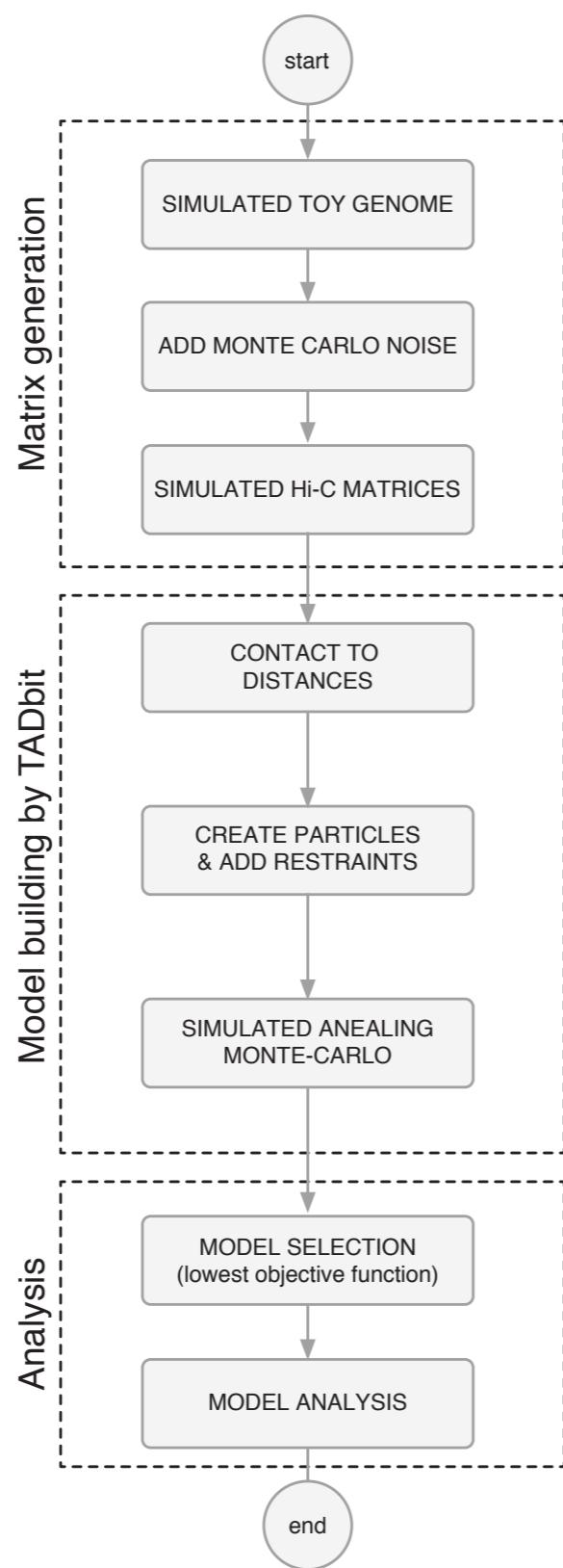
by Ivan Junier

Toy interaction matrices

set 0 (Δts=10^0)

0   1Mb

0

1Mb

set 4 (Δts=10^4)

0   1Mb

0

1Mb

set 6 (Δts=10^6)

0   1Mb

0

1Mb
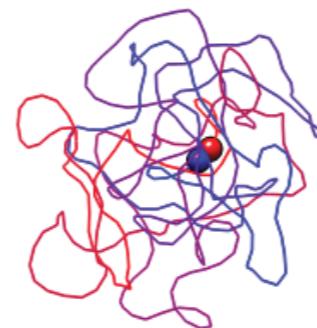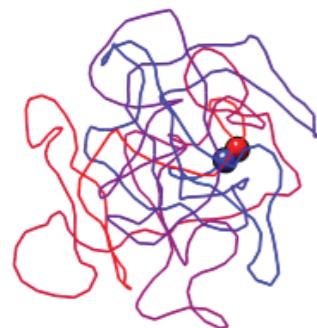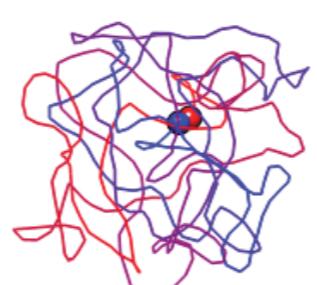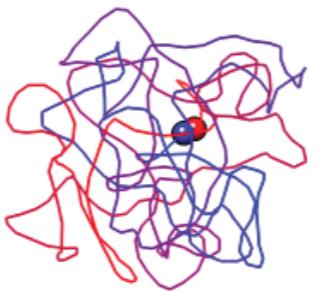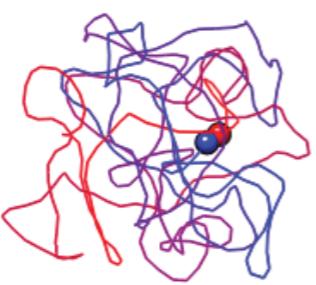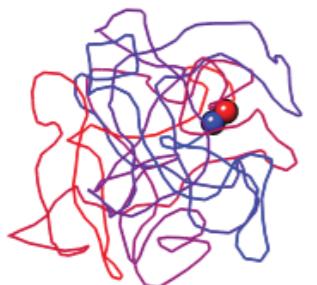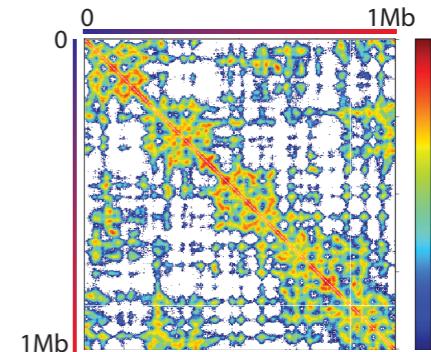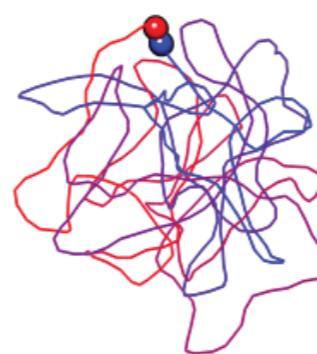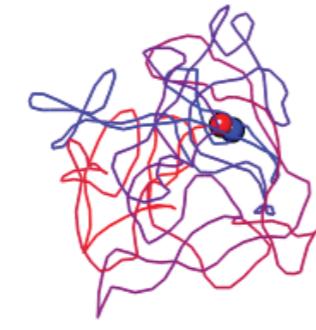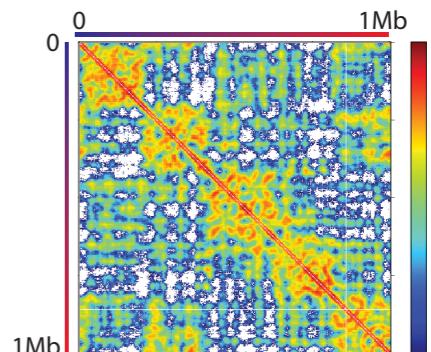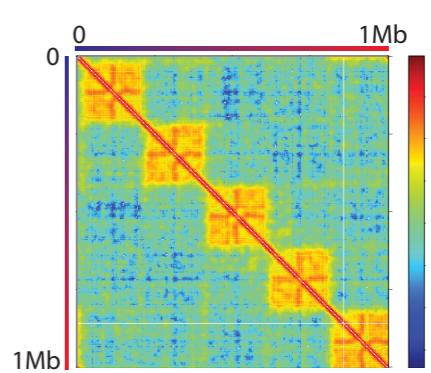
Reconstructing toy models
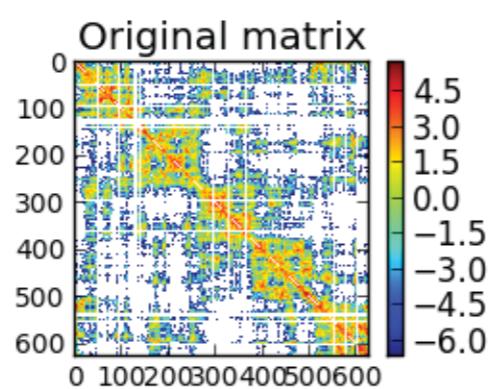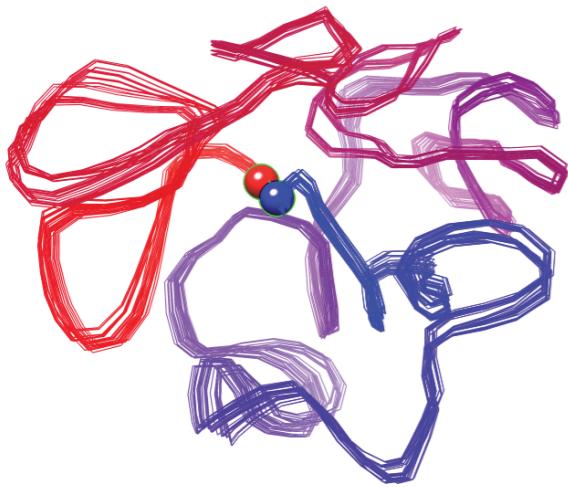
chr40_TAD
$\alpha=100$
$\Delta ts=10$
TADbit-SCC: 0.91
<dRMSD>: 32.7 nm
<dSCC>: 0.94
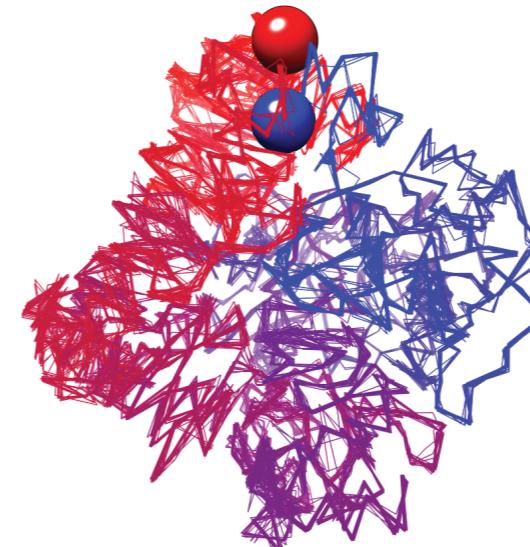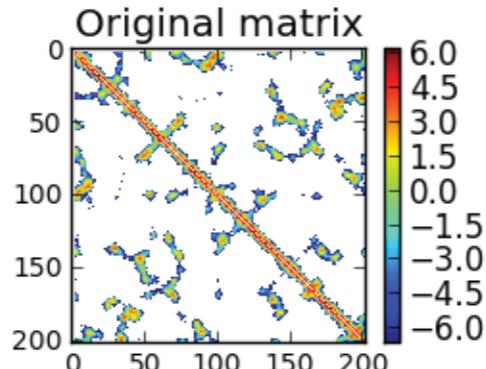
chr150_TAD
$\alpha=50$
$\Delta ts=1$
TADbit-SCC: 0.82
**<dRMSD>: 45.4 nm**
<dSCC>: 0.86
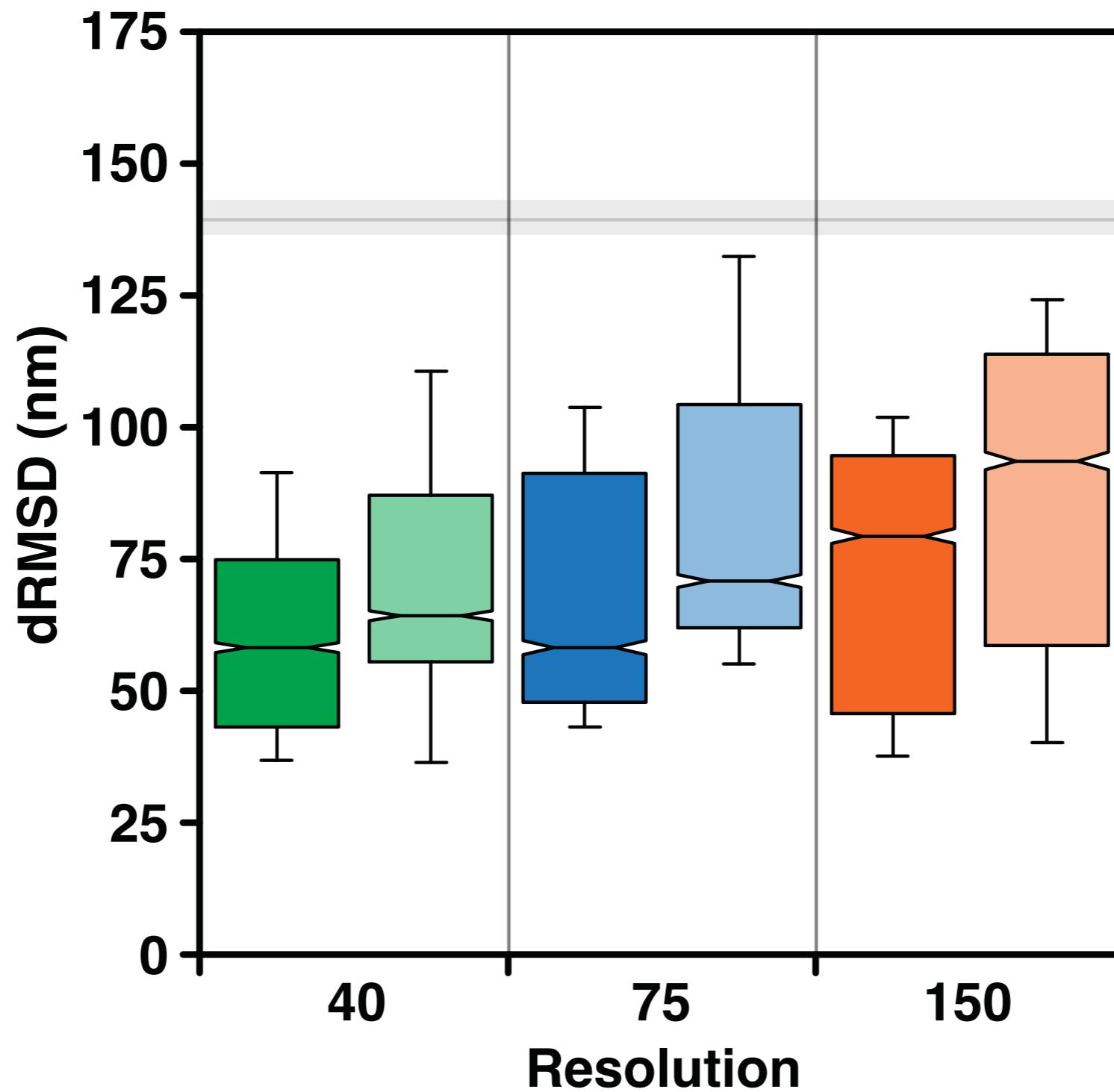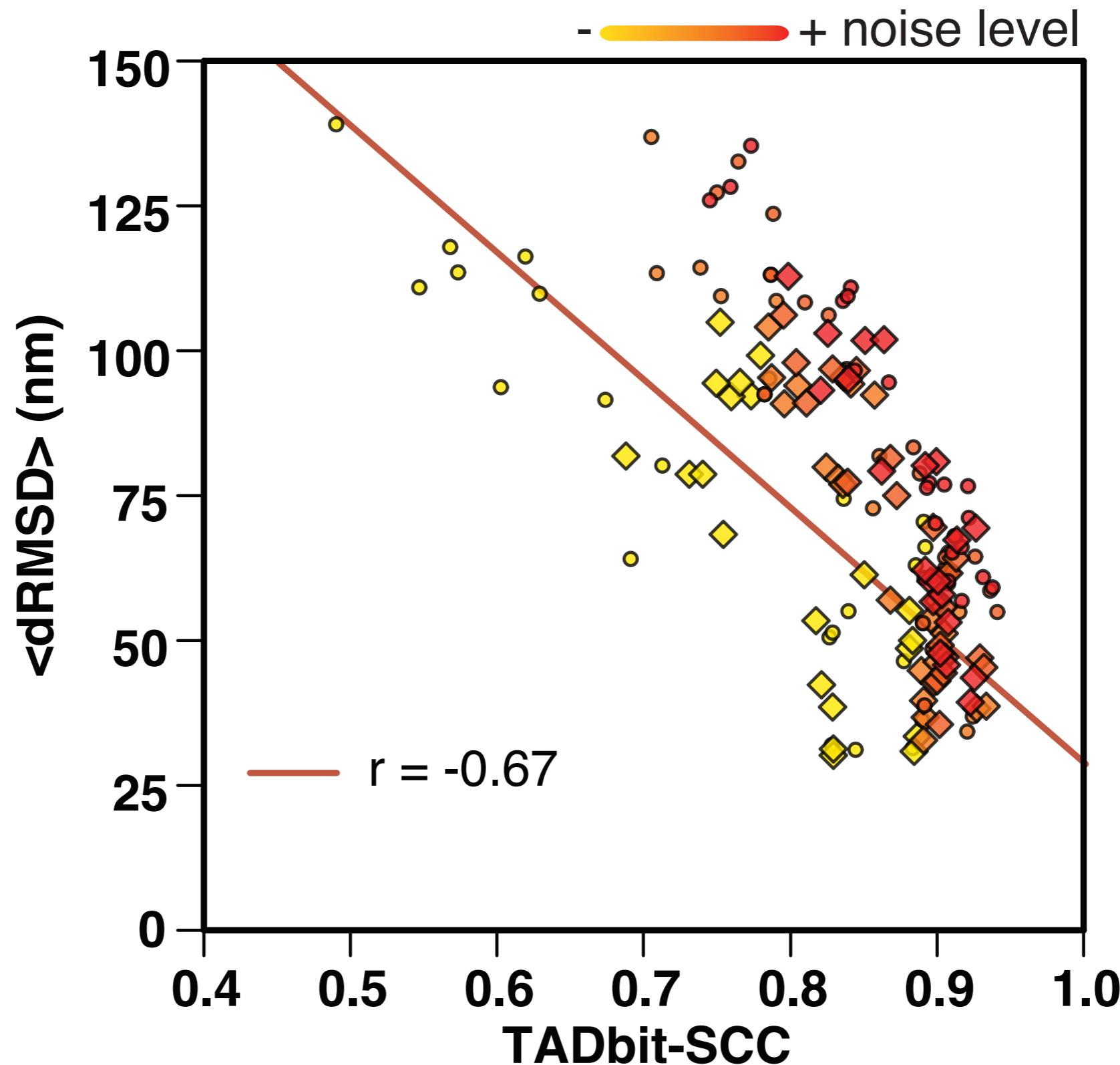
Structural variability is "NOT OK"

- ⬤⬤⬤⬤⬤ + structural variability
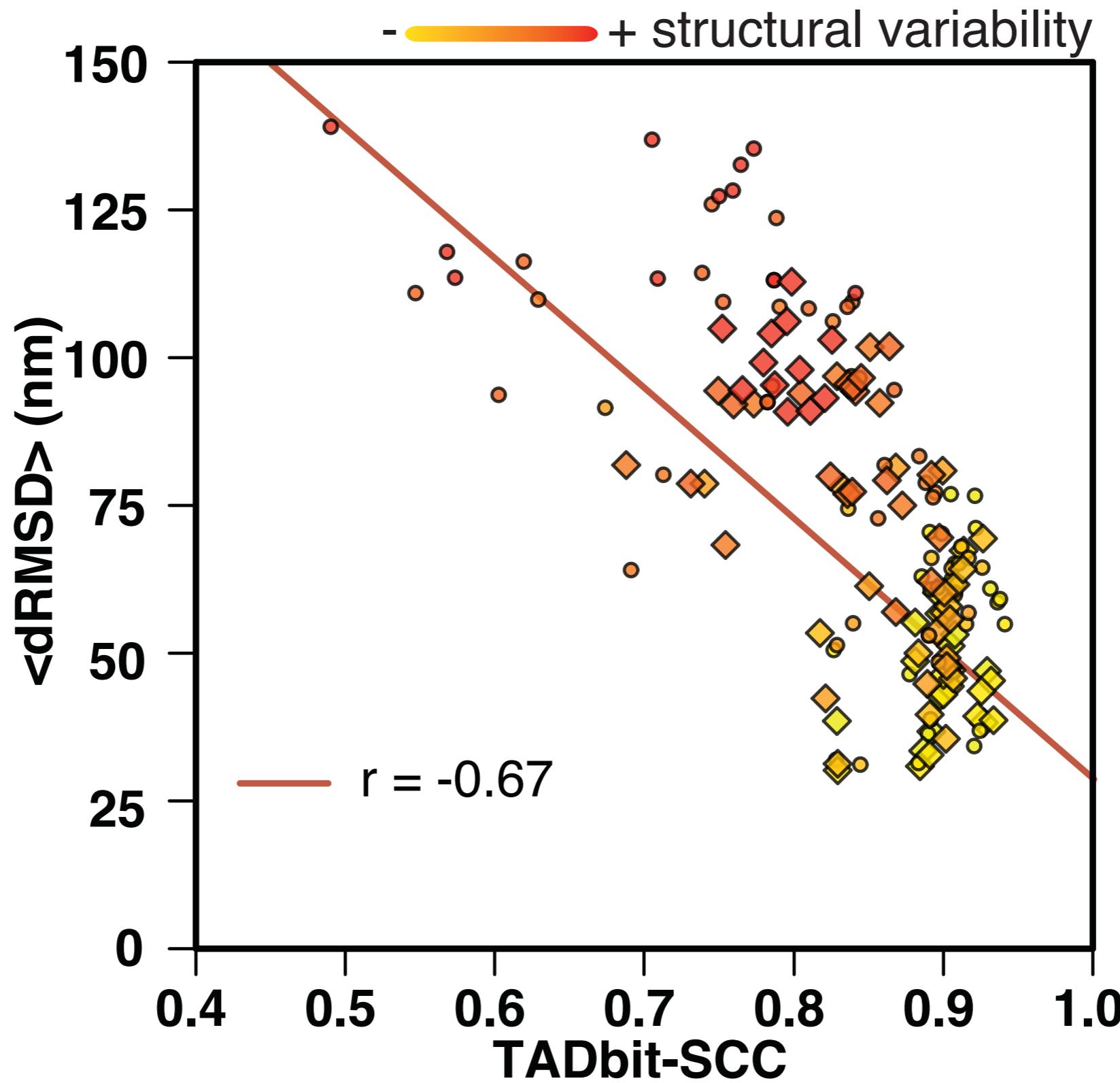
$\langle dRMSD \rangle$ (nm)

r = -0.67

TADbit-SCC

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
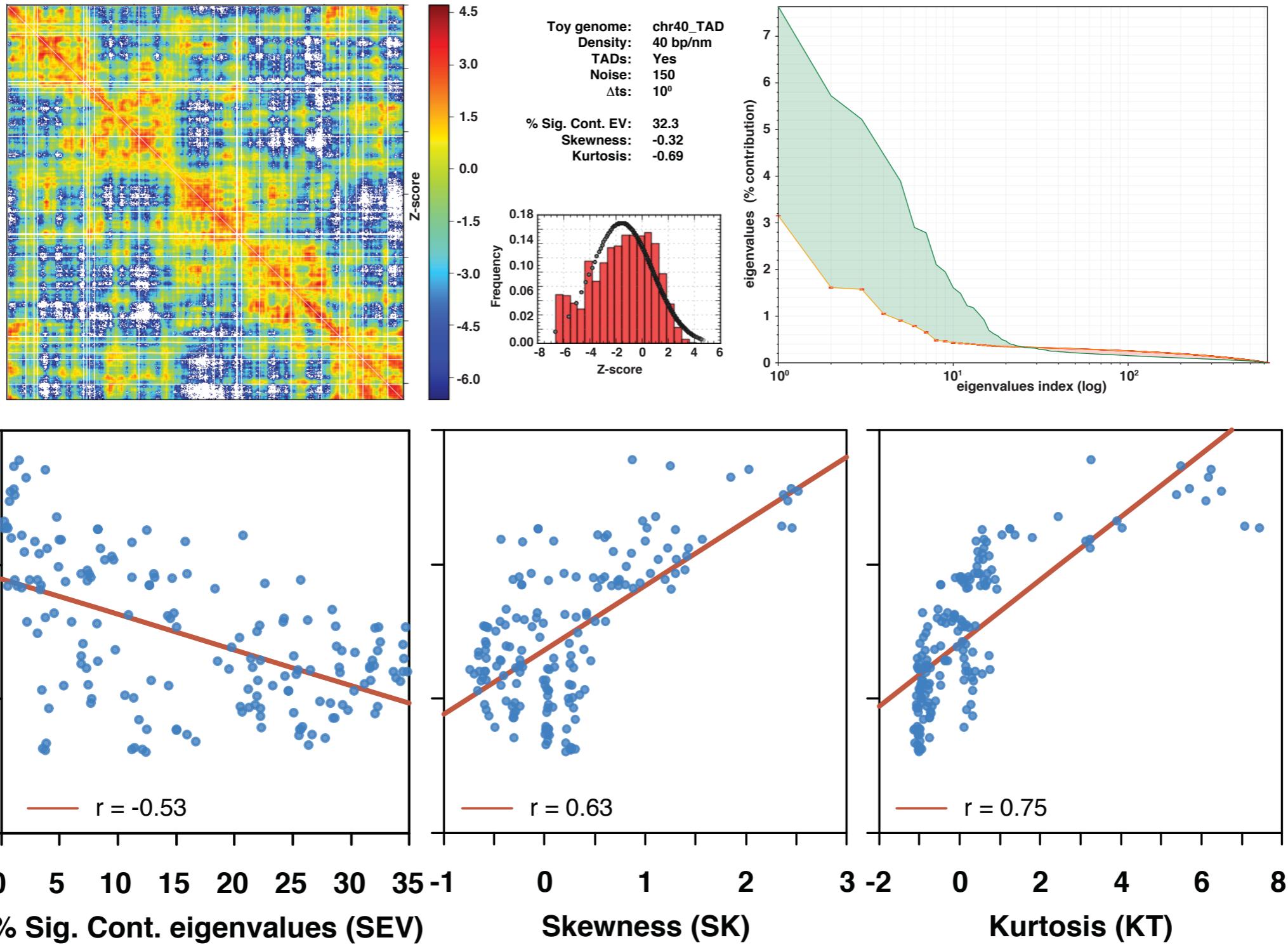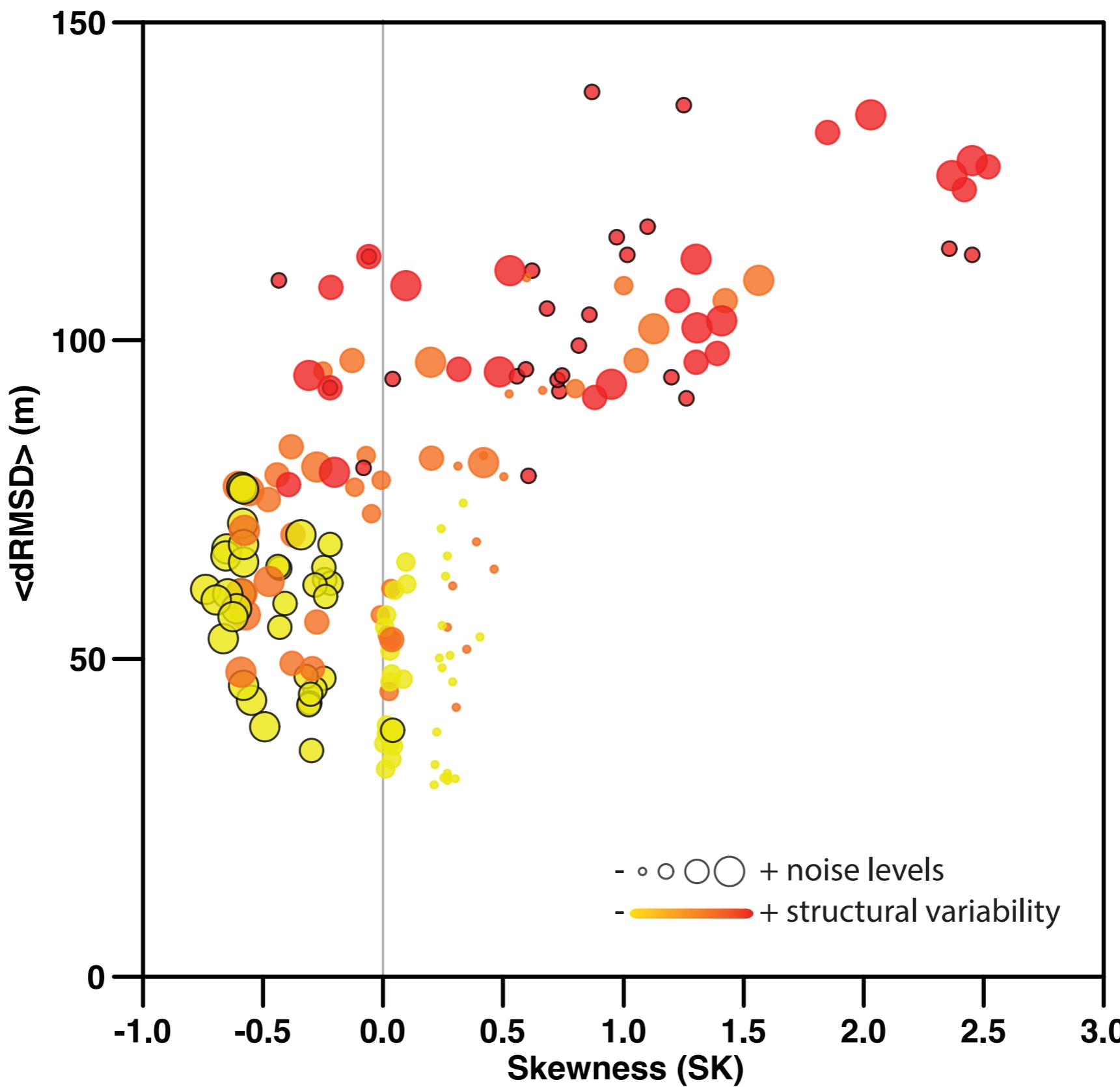
Can we predict the accuracy of the models?

$$\text{MMP} = -0.0002 * \text{Size} + 0.0335 * \text{SK} - 0.0229 * \text{KU} + 0.0069 * \text{SEV} + 0.8126$$
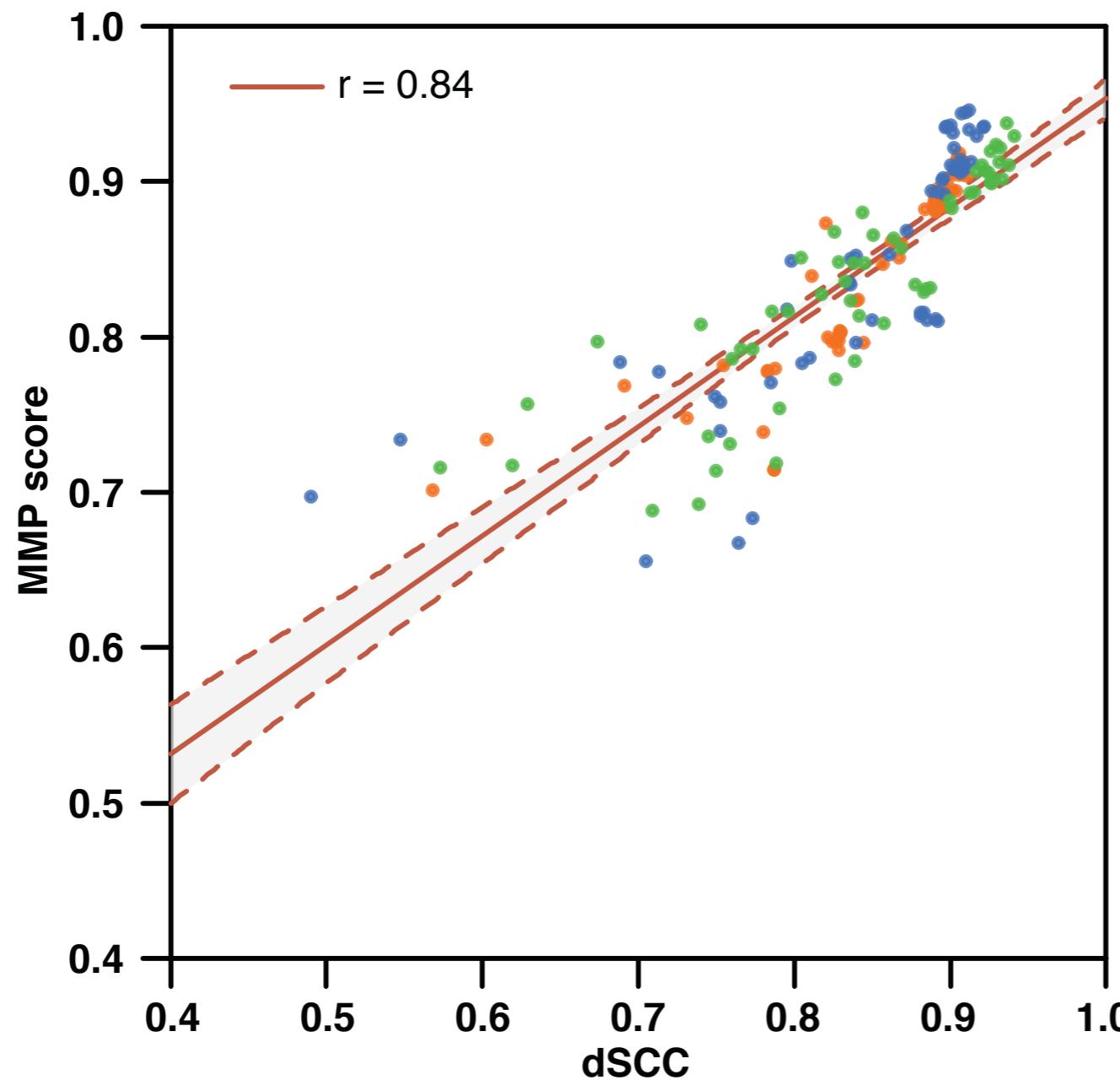
Human Chr1:120,640,000-128,040,000

Size: 186
SEV: 3.63
SK: 0.20
KT: -0.53
MMP: 0.82
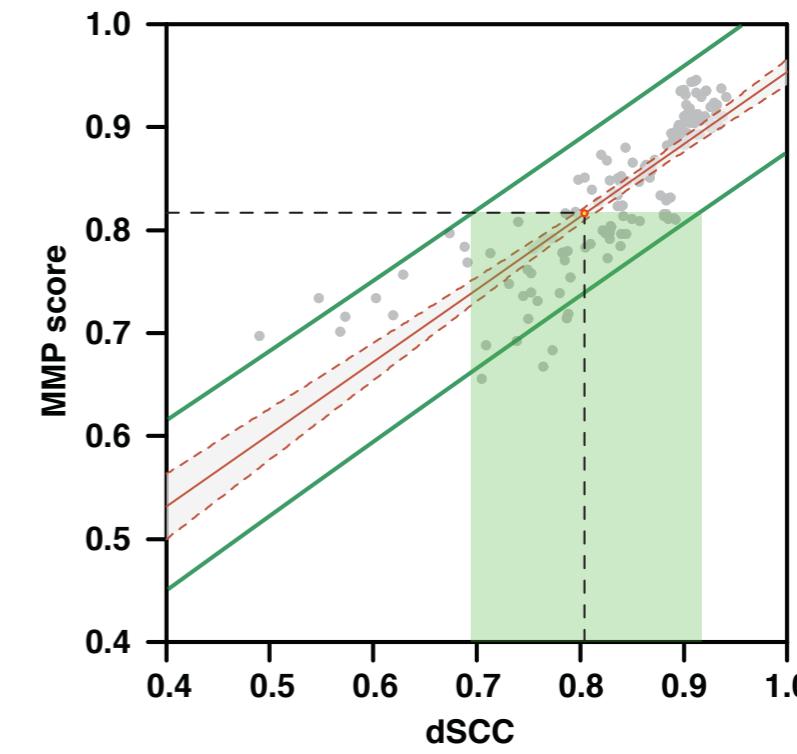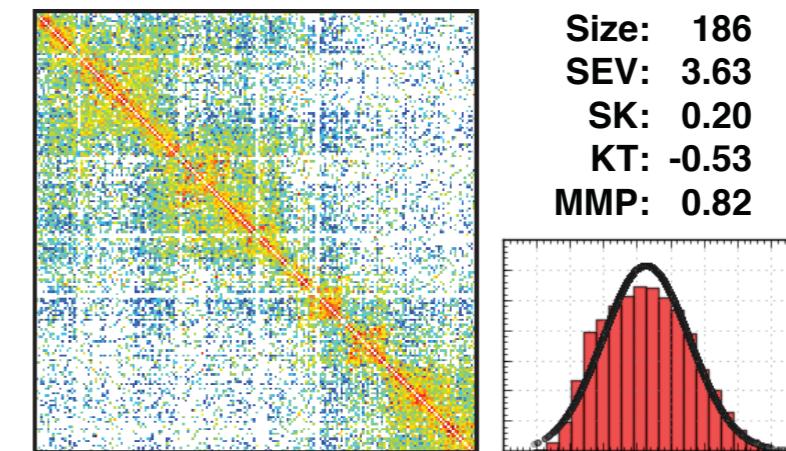
Higher-res is "good"
put your $$ in sequencing

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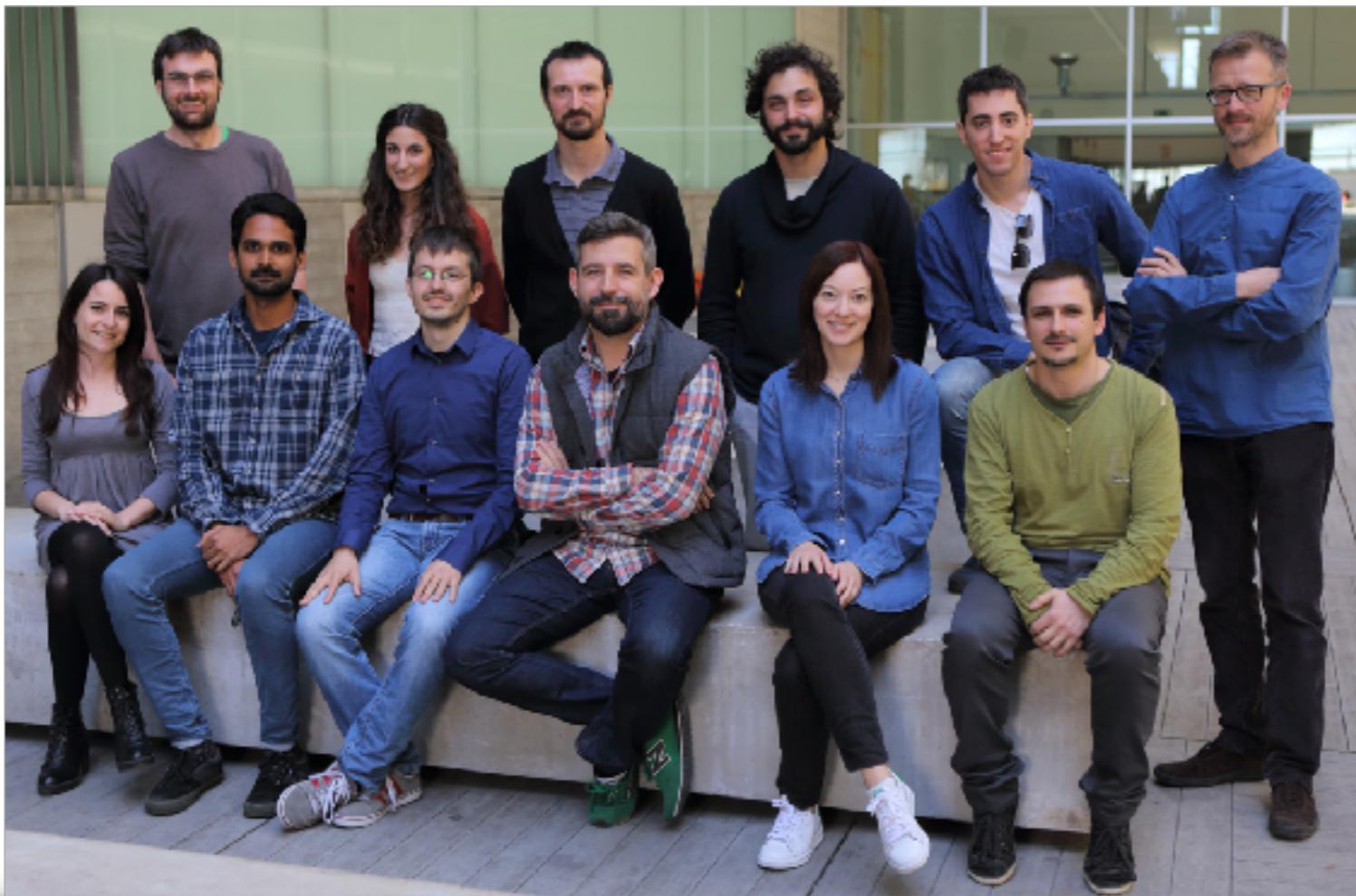
Davide Baù
Gireesh K. Bogu
Yasmina Cuartero
François le Dily
David Dufour
Irene Farabella
Silvia Galan
Francesca di Giovanni
Mike Goodstadt
Francisco Martínez-Jiménez
François Serra
Paula Soler
Yannick Spill
Marco di Stefano
Marie Trussart

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