Data integration for 3D structure determination.

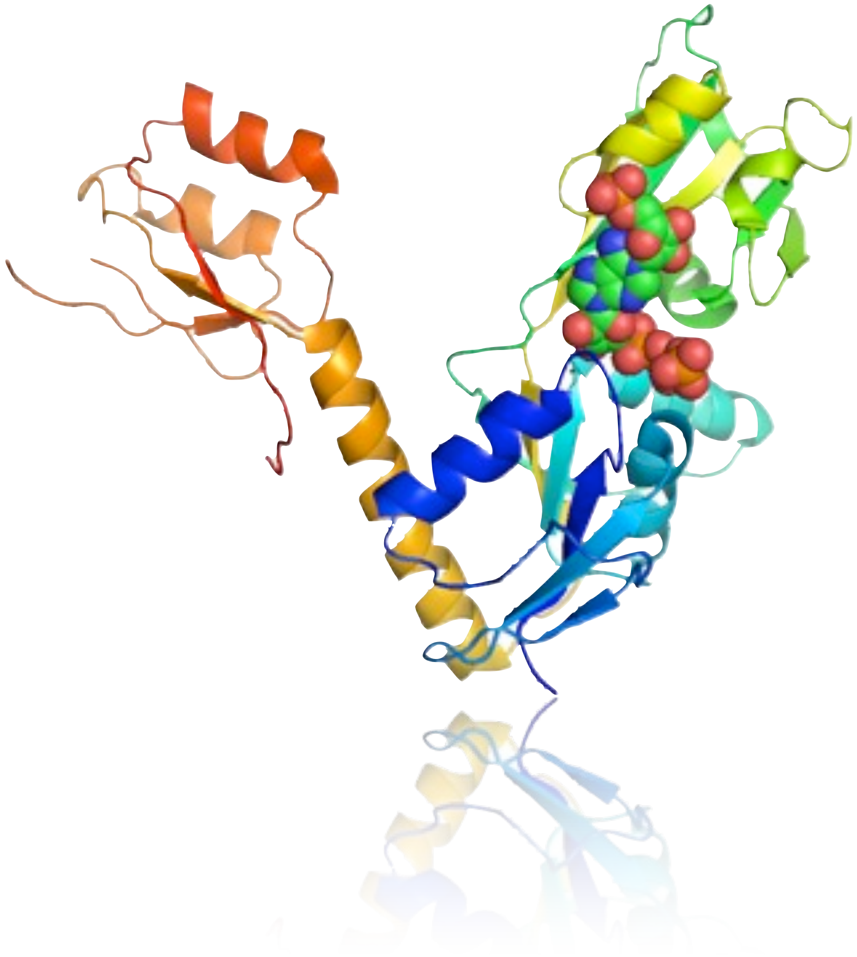
Marc A. Marti-Renom

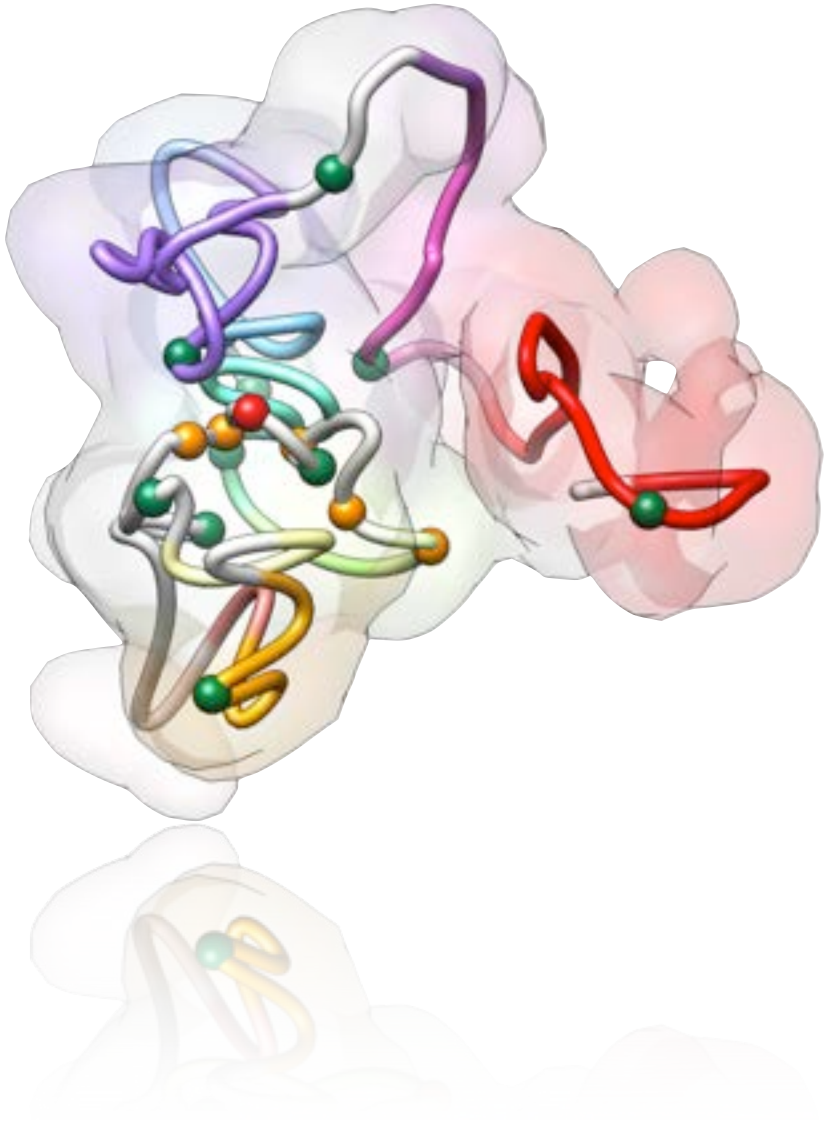
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

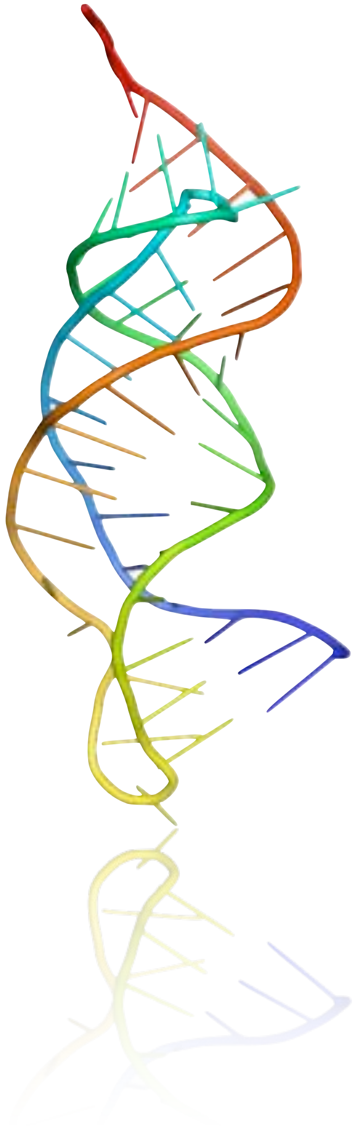
cnag CRG Centre for Genomic Regulation

Structural Genomics Group

http://www.marciuslab.org

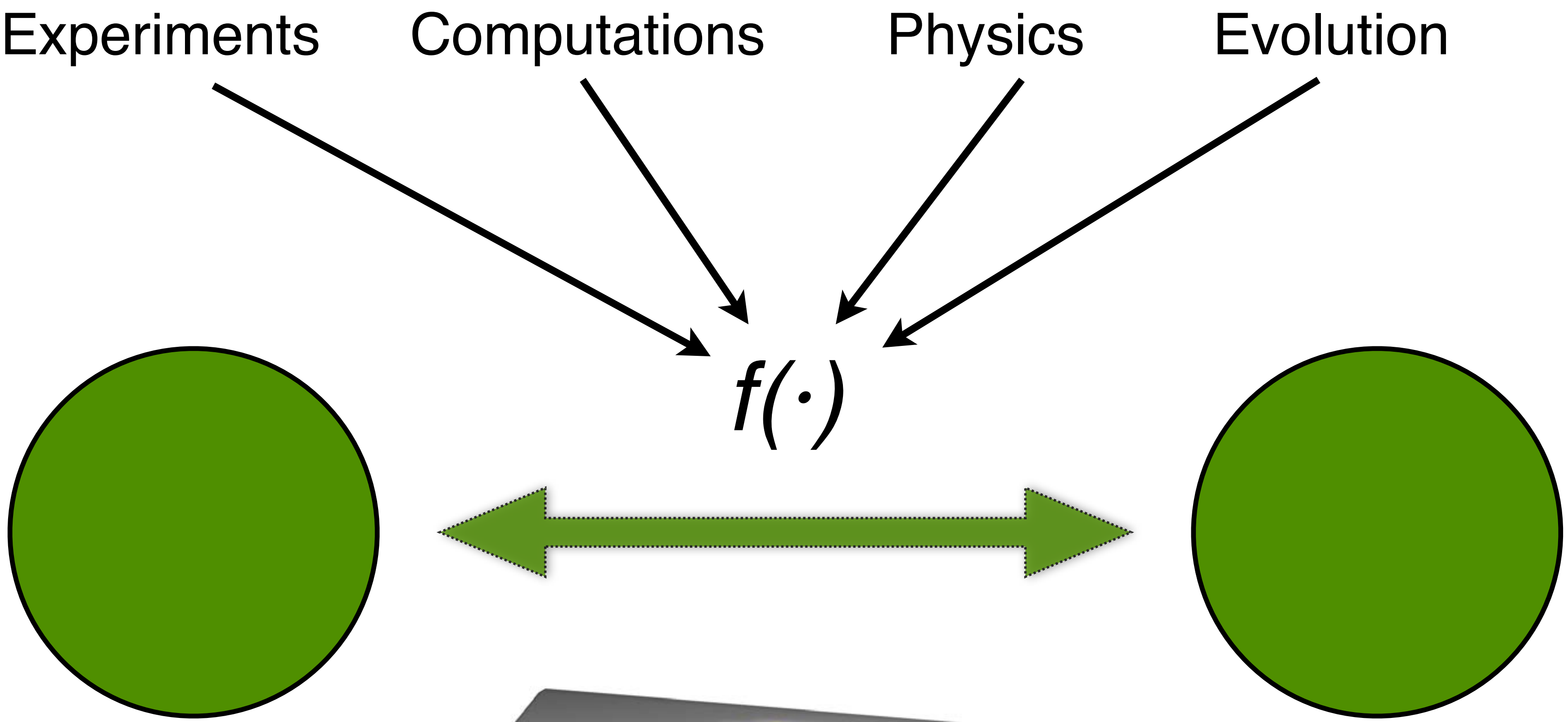






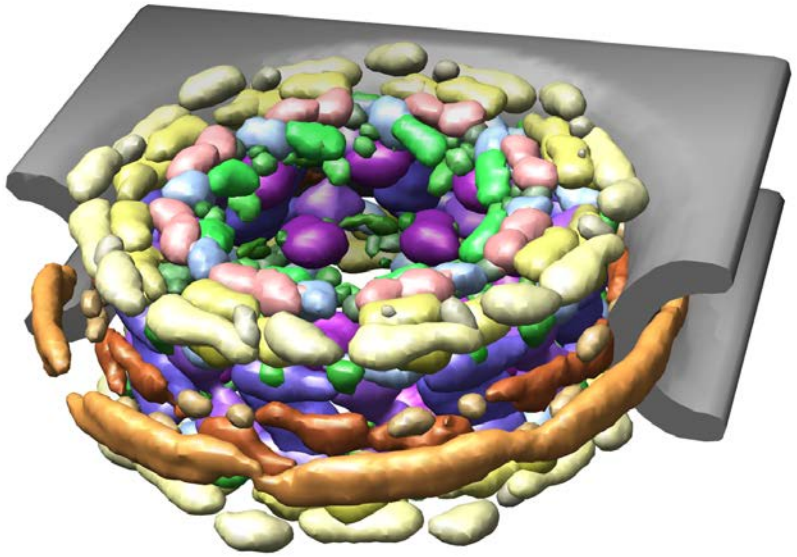
Integrative Modeling Platform

http://www.integrativemodeling.org



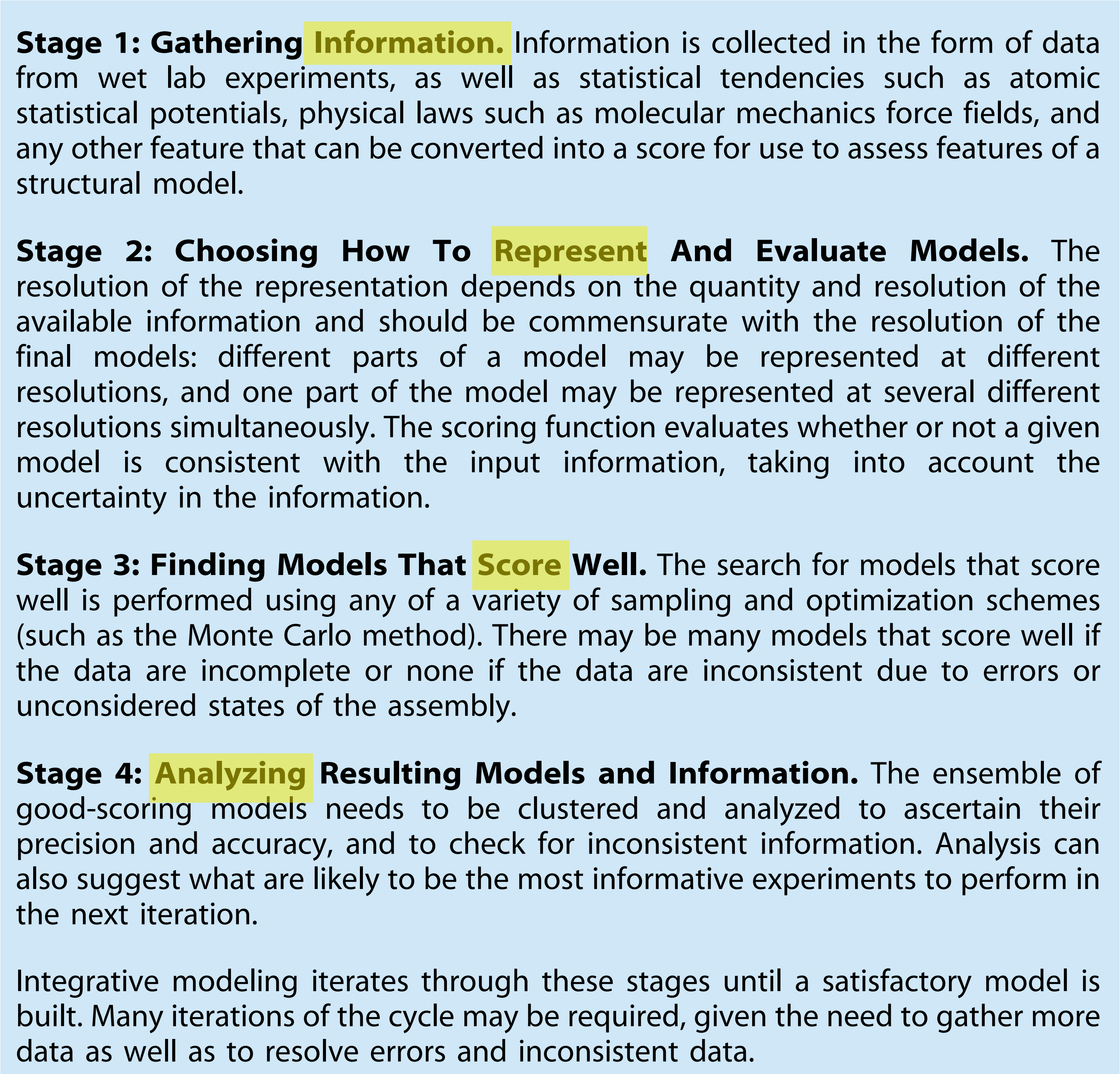
Experiments Computations Physics Evolution

$f(\cdot)$



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

Stages



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

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

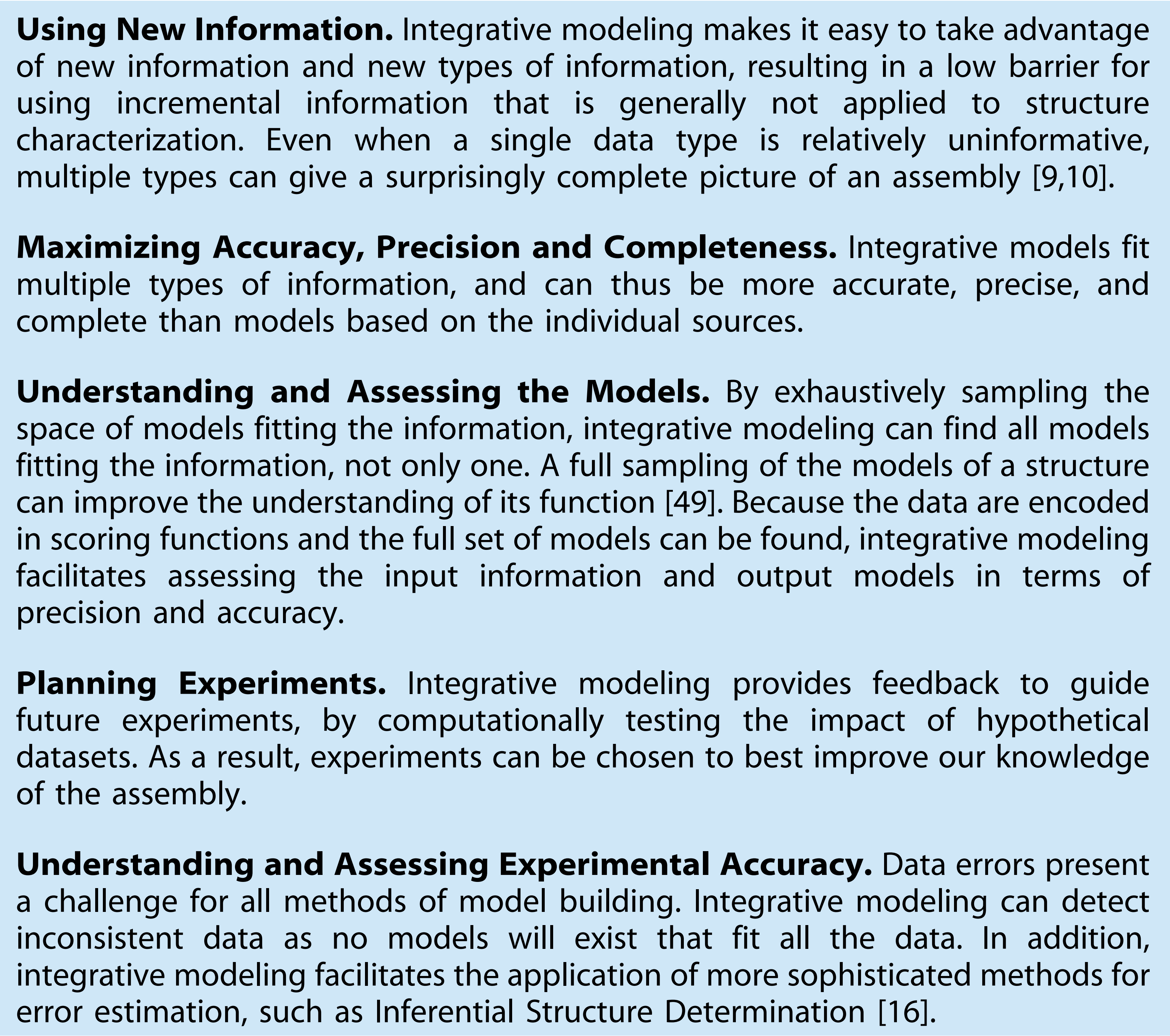
Stage 3: Finding Models That Score Well. The search for models that score well is performed using any of a variety of sampling and optimization schemes (such as the Monte Carlo method). There may be many models that score well if the data are incomplete or none if the data are inconsistent due to errors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

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

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

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

Advantages



Using New Information. Integrative modeling makes it easy to take advantage of new information and new types of information, resulting in a low barrier for using incremental information that is generally not applied to structure characterization. Even when a single data type is relatively uninformative, multiple types can give a surprisingly complete picture of an assembly [9,10].

Maximizing Accuracy, Precision and Completeness. Integrative models fit multiple types of information, and can thus be more accurate, precise, and complete than models based on the individual sources.

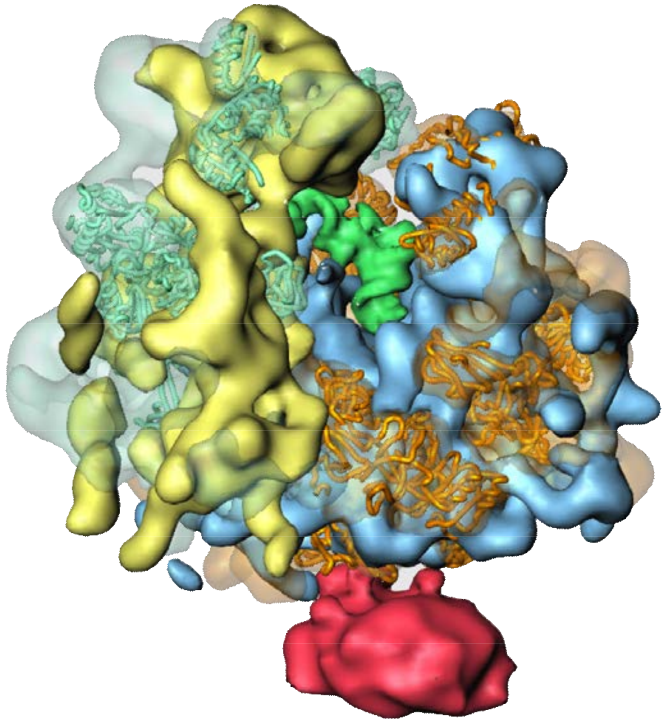
Understanding and Assessing the Models. By exhaustively sampling the space of models fitting the information, integrative modeling can find all models fitting the information, not only one. A full sampling of the models of a structure can improve the understanding of its function [49]. Because the data are encoded in scoring functions and the full set of models can be found, integrative modeling facilitates assessing the input information and output models in terms of precision and accuracy.

Planning Experiments. Integrative modeling provides feedback to guide future experiments, by computationally testing the impact of hypothetical datasets. As a result, experiments can be chosen to best improve our knowledge of the assembly.

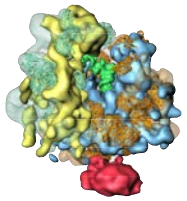
Understanding and Assessing Experimental Accuracy. Data errors present a challenge for all methods of model building. Integrative modeling can detect inconsistent data as no models will exist that fit all the data. In addition, integrative modeling facilitates the application of more sophisticated methods for error estimation, such as Inferential Structure Determination [16].

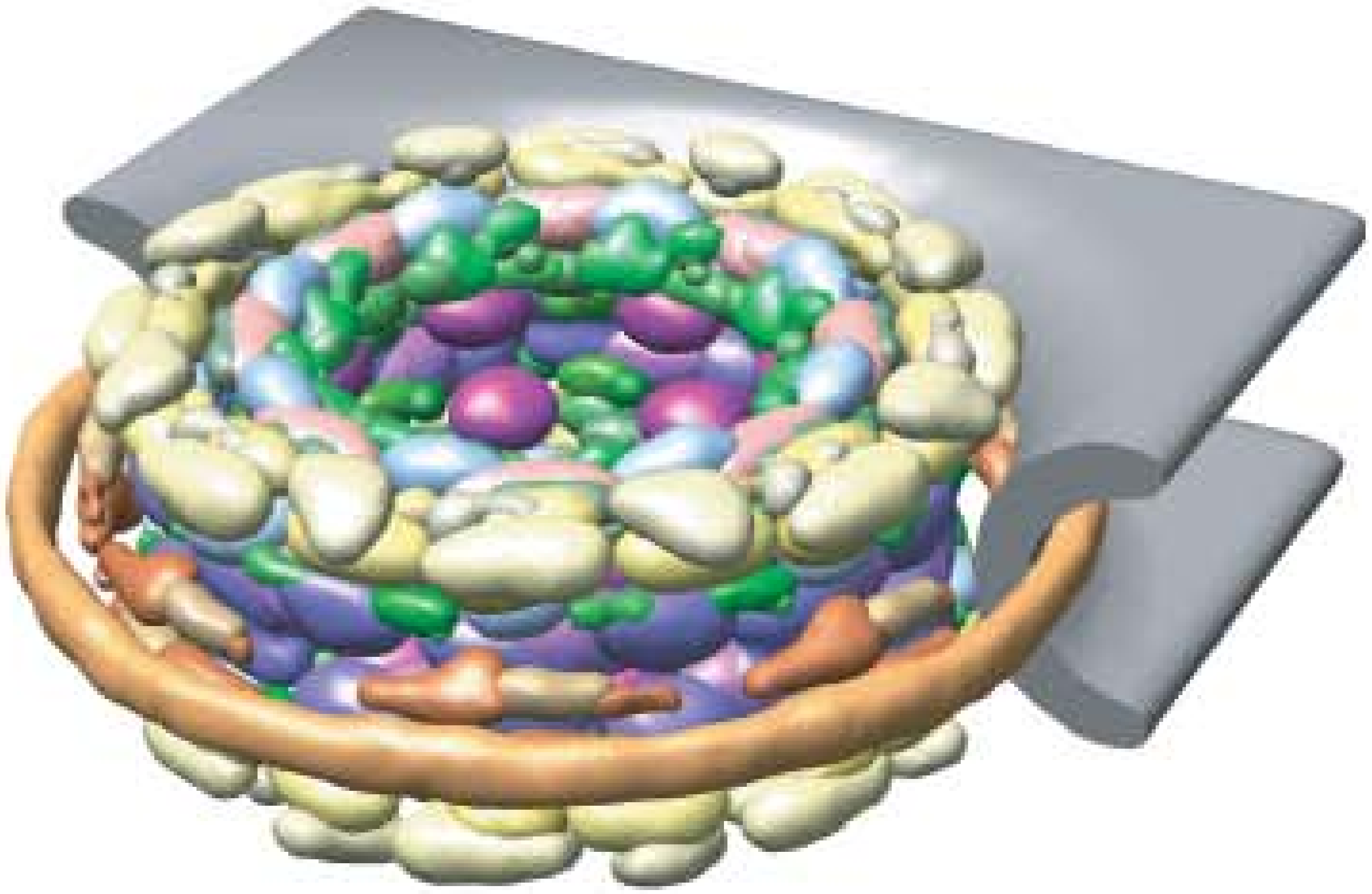
Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., et al. (2012). *PLoS Biology*, *10*(1), e1001244

Data Integration

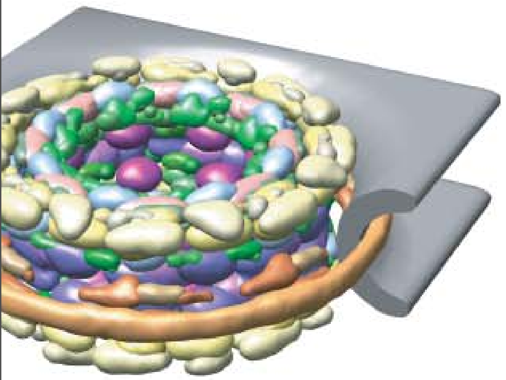


Data Integration



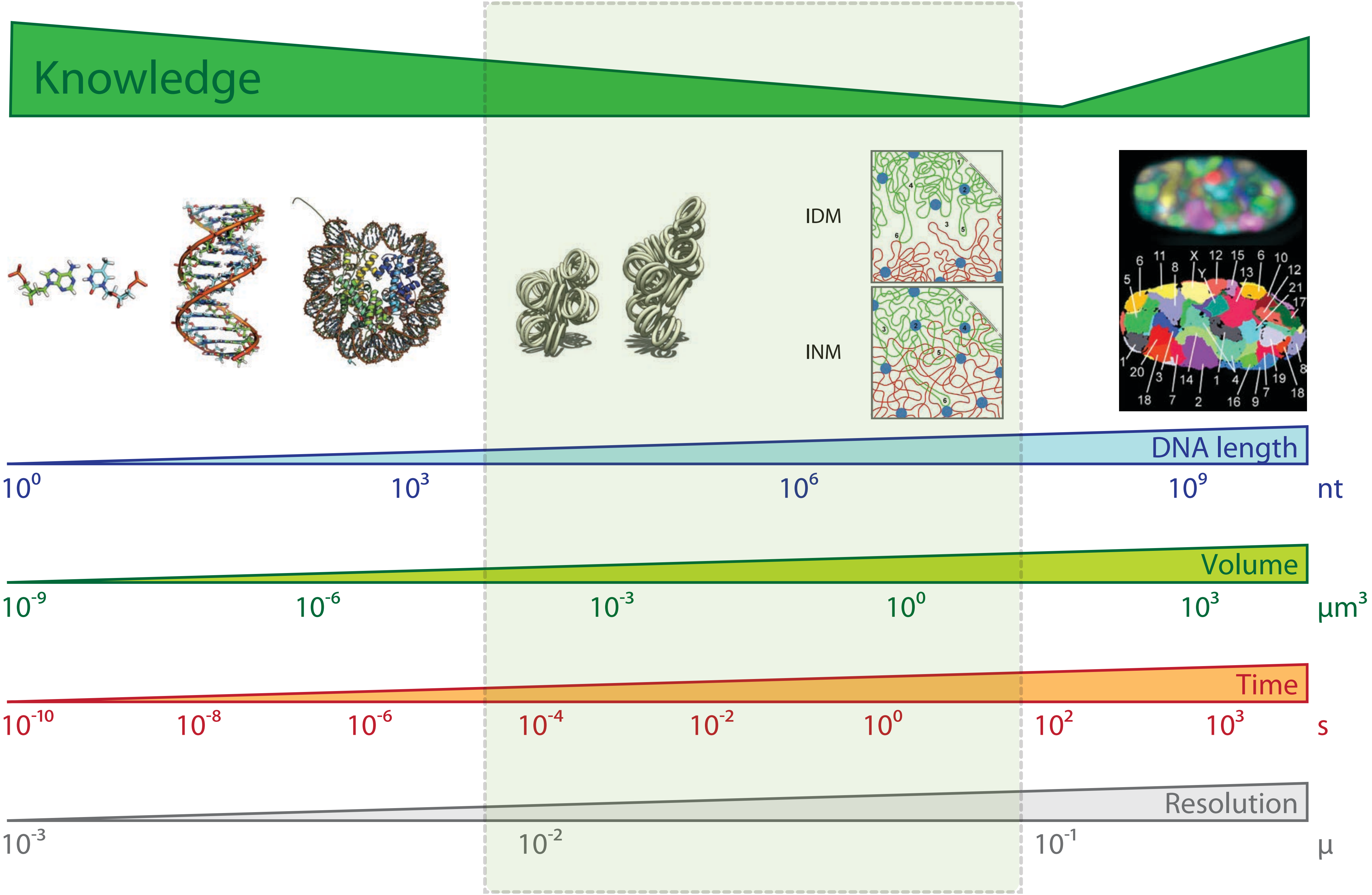


Data Integration



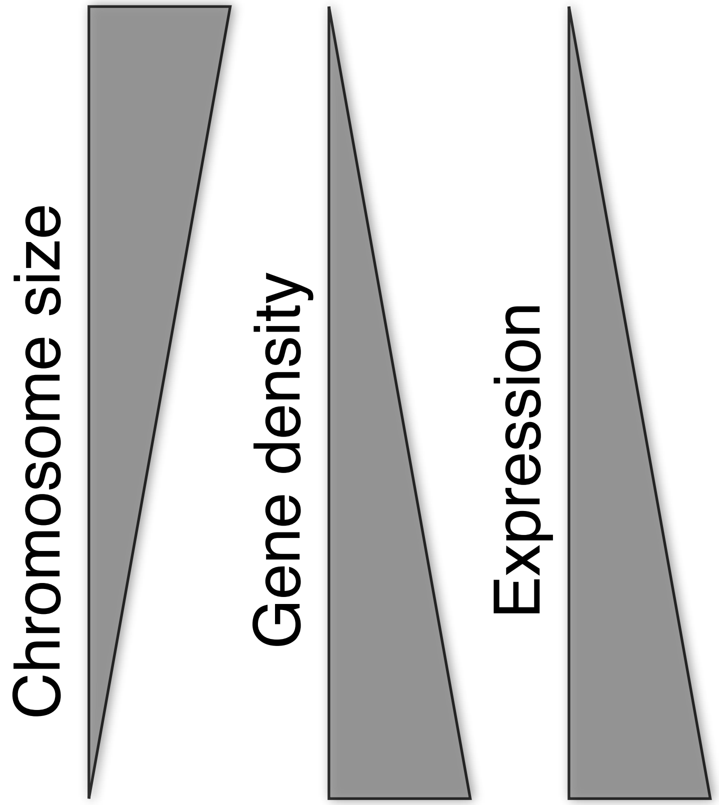
Resolution Gap

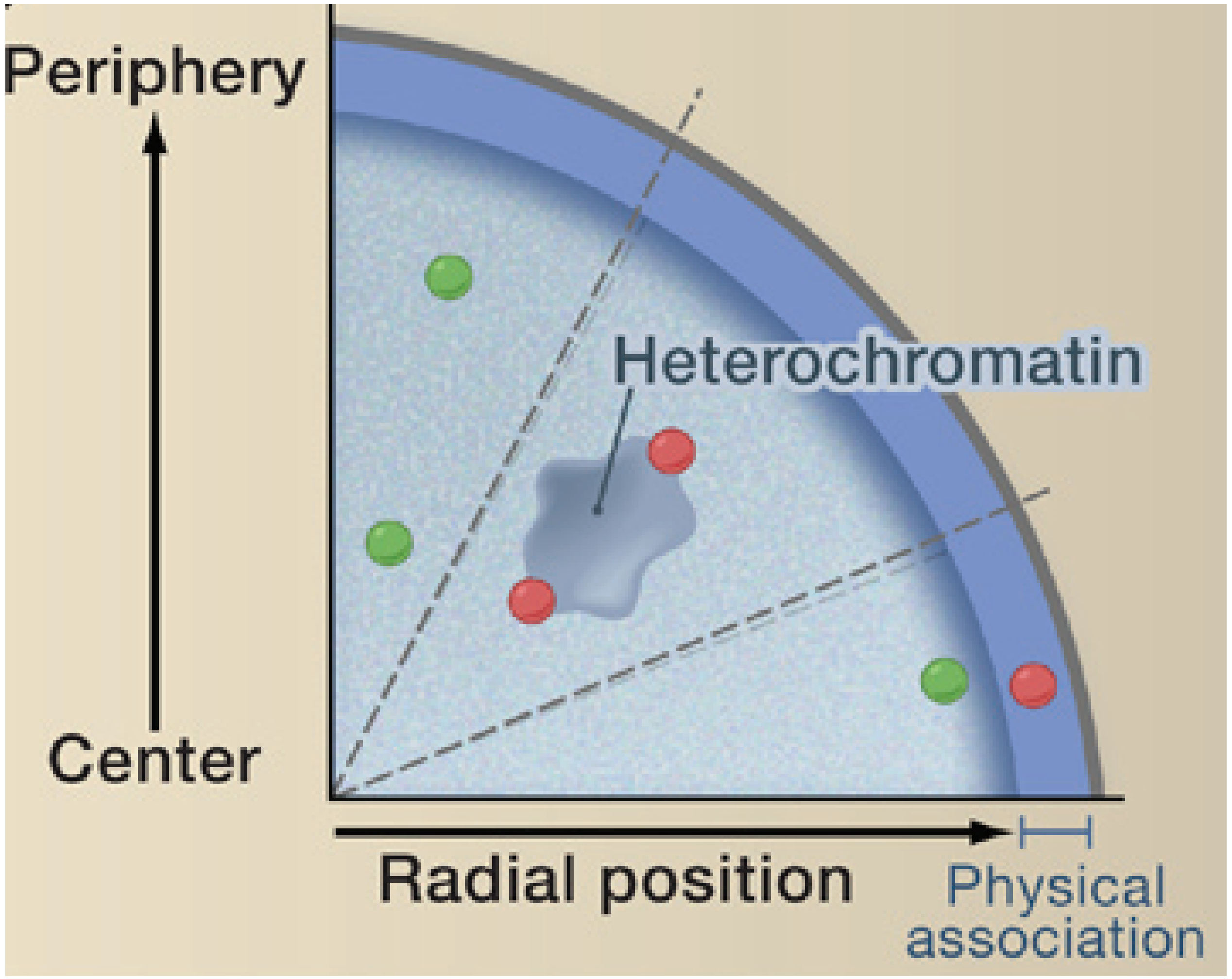
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)



Complex genome organization

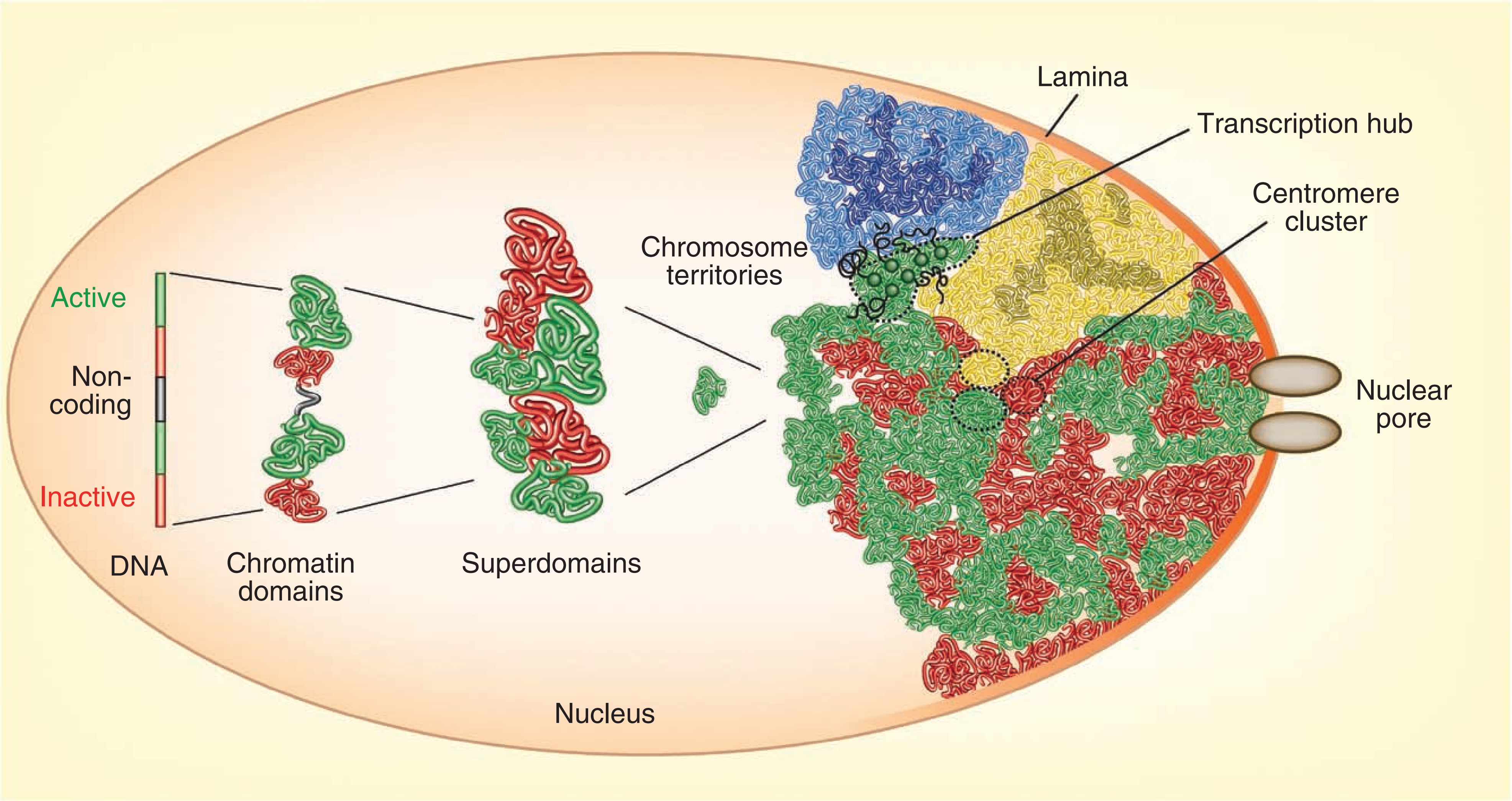
Takizawa, T., Meaburn, K. J. & Misteli, T. The meaning of gene positioning. Cell 135, 9–13 (2008).





Complex genome organization

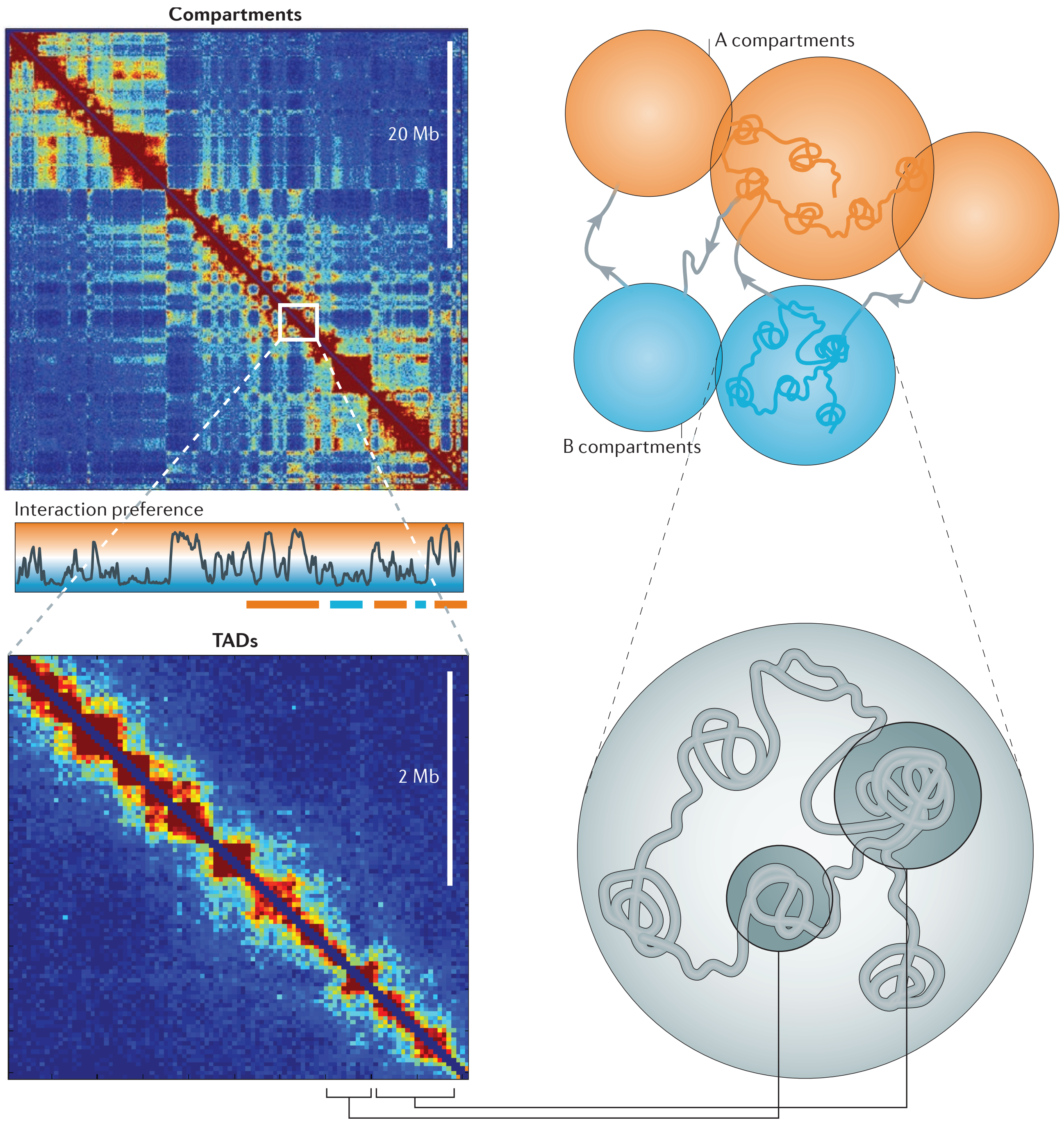
Cavalli, G. & Misteli, T. Functional implications of genome topology. Nat Struct Mol Biol 20, 290–299 (2013).



Marina Corral

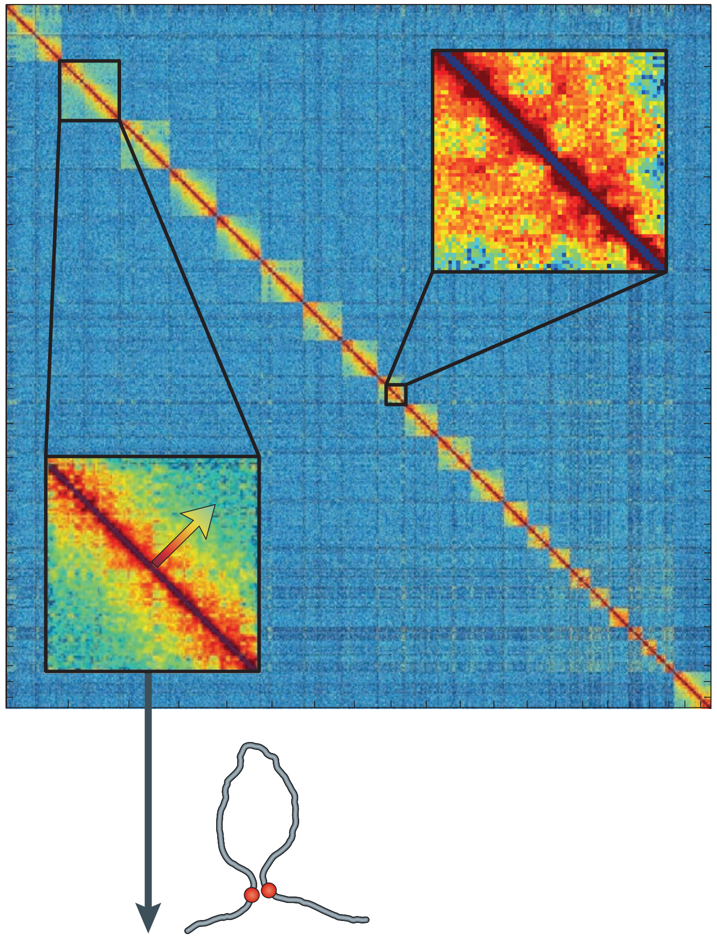
Complex genome organization

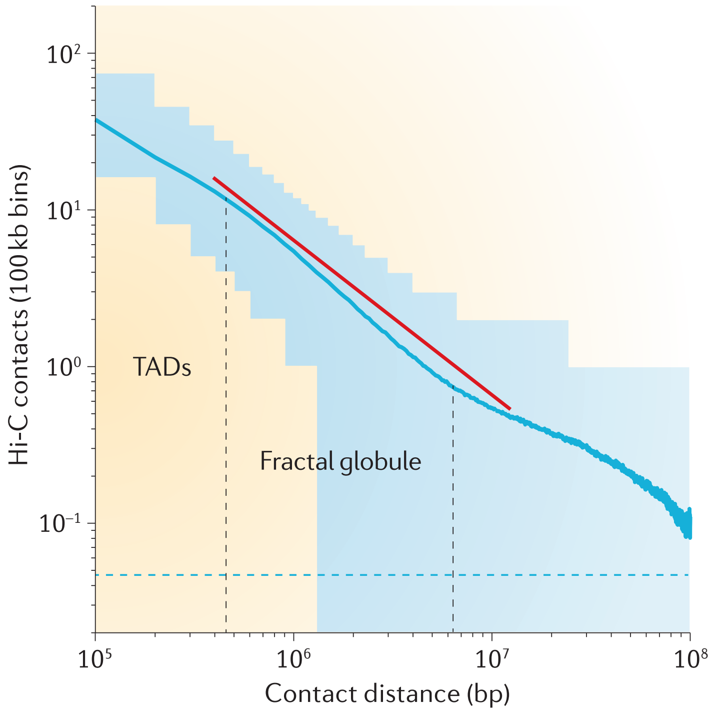
Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390–403 (2013).

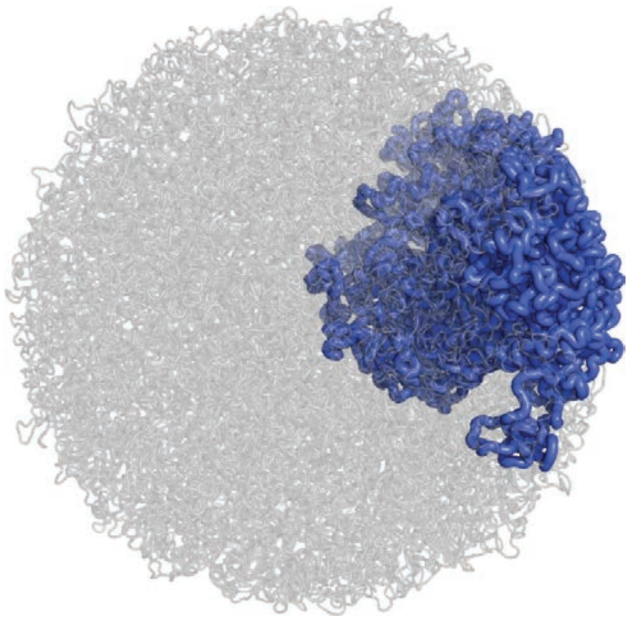


Complex genome organization

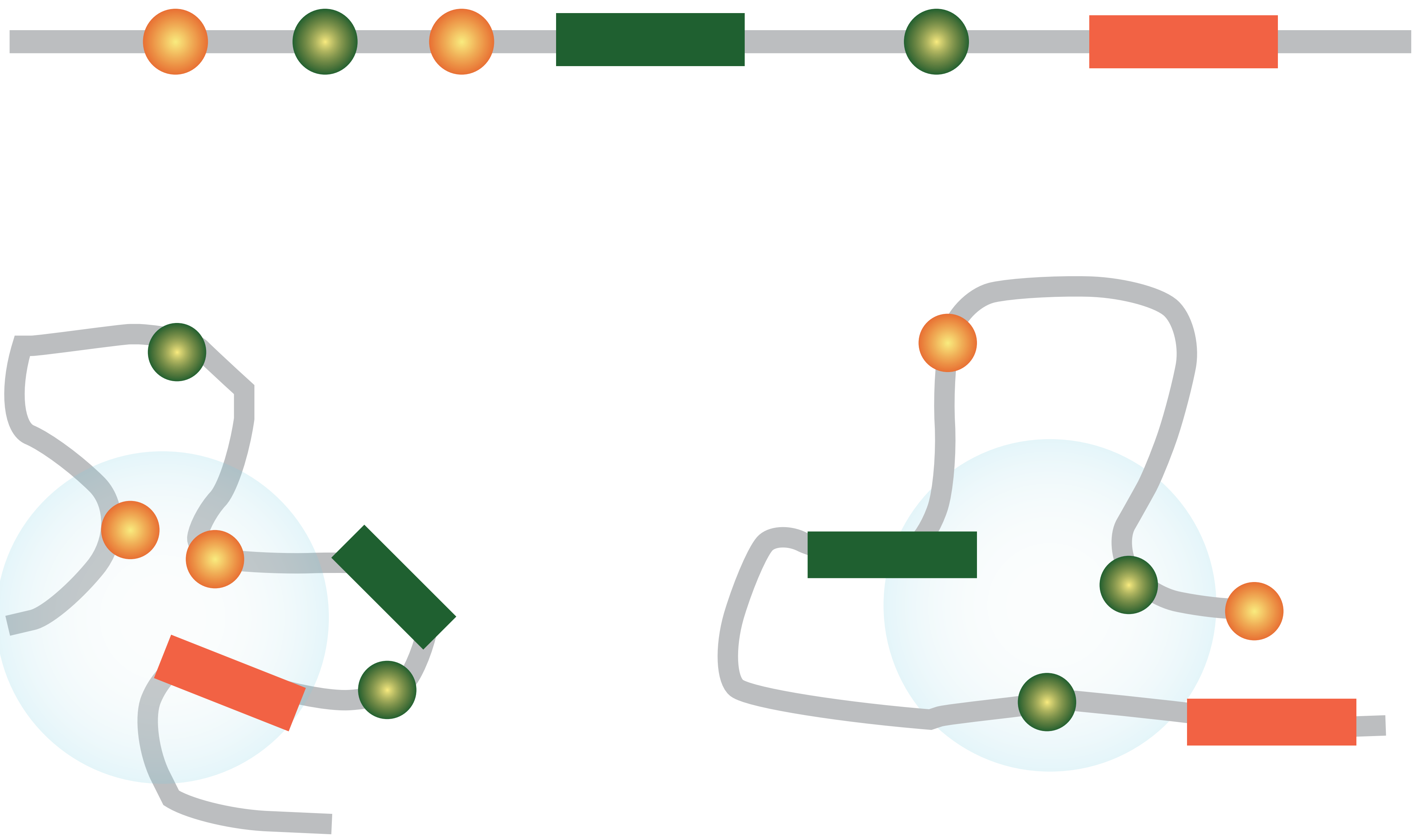
Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY) 326, 289–293 (2009).



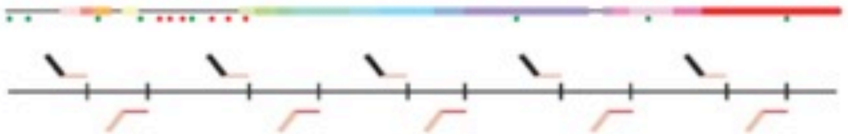




Complex genome organization



Experiments



Grow GM12878 and K562 cells

↓

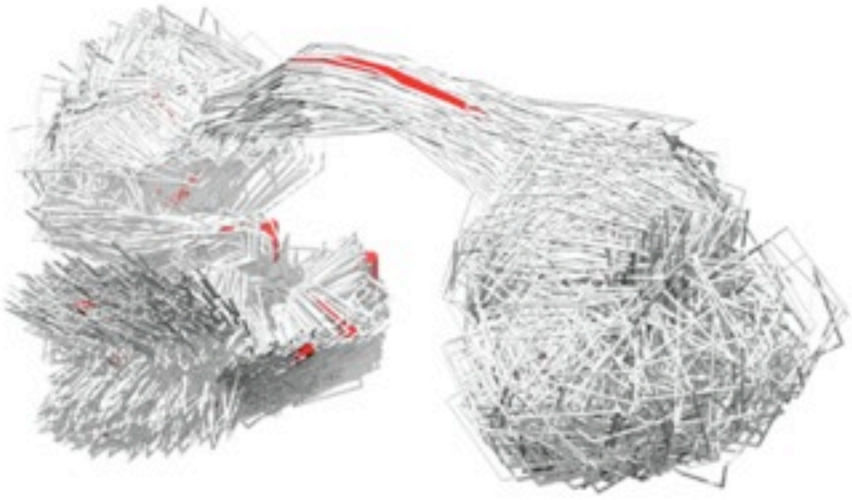
Perform 3C analysis

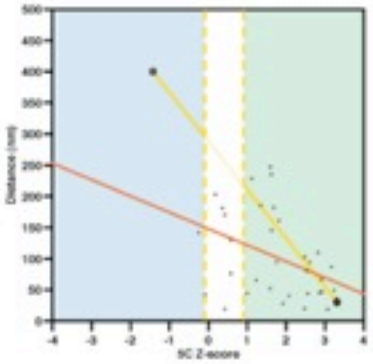
↓

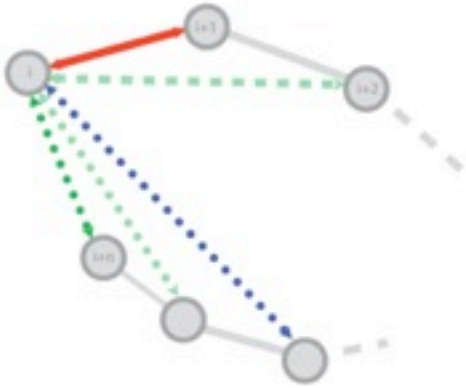
Perform 5C analysis with 30+25 primers

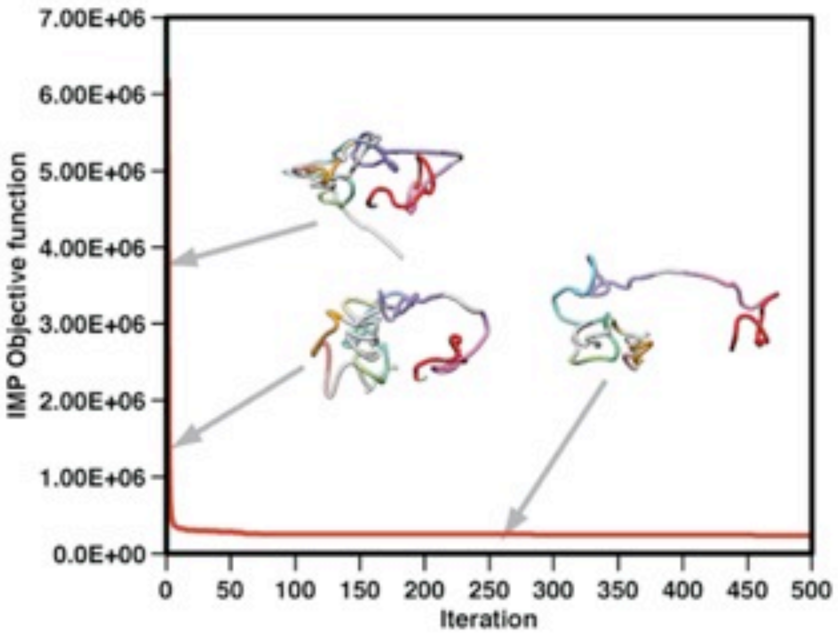
↓

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

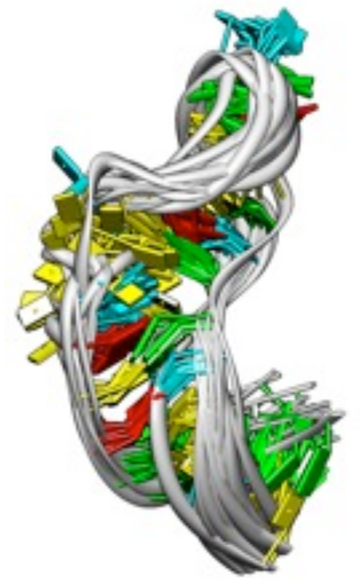


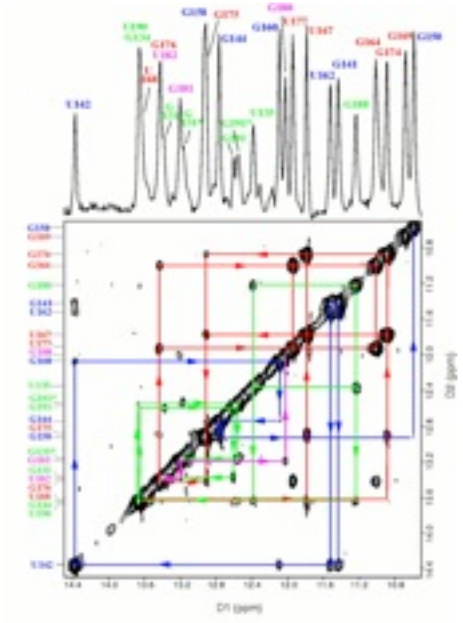




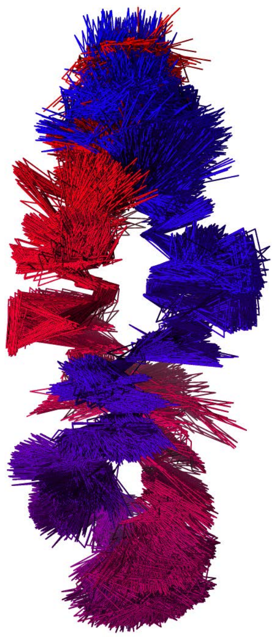


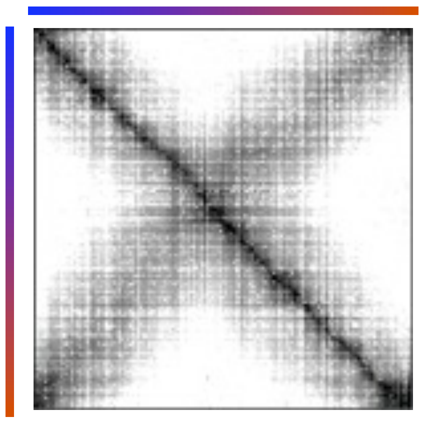
Computation





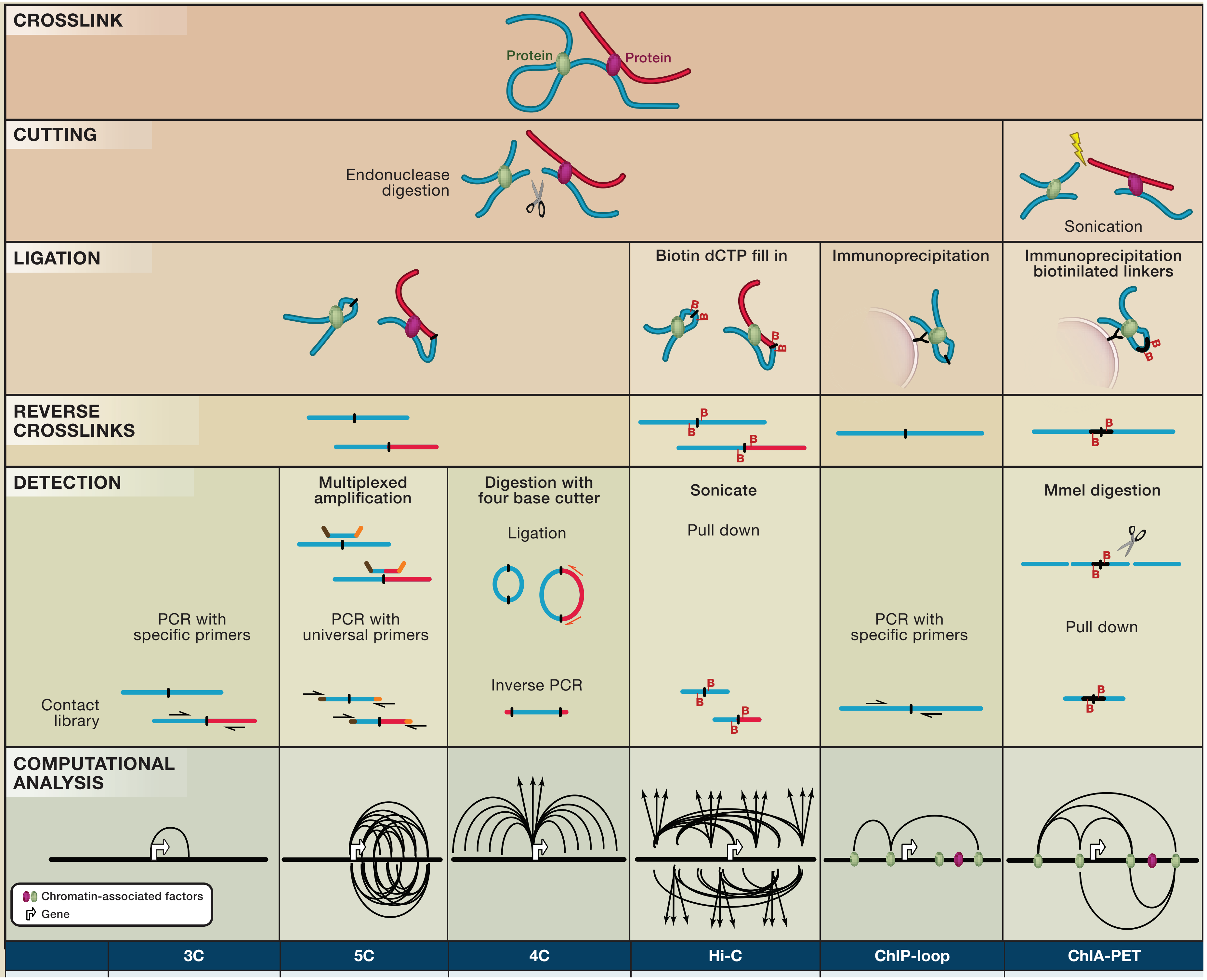
Biomolecular structure determination
2D-NOESY data





Chromosome structure determination
5C data

Chromosome Conformation Capture



CROSSLINK

Protein

Protein

CUTTING

Endonuclease digestion

Sonication

LIGATION

Biotin dCTP fill in

Immunoprecipitation

Immunoprecipitation biotinilated linkers

REVERSE CROSSLINKS

DETECTION

Multiplexed amplification

Digestion with four base cutter

Ligation

Inverse PCR

Sonicate

Pull down

Mmel digestion

Pull down

PCR with specific primers

Contact library

PCR with universal primers

PCR with specific primers

COMPUTATIONAL ANALYSIS

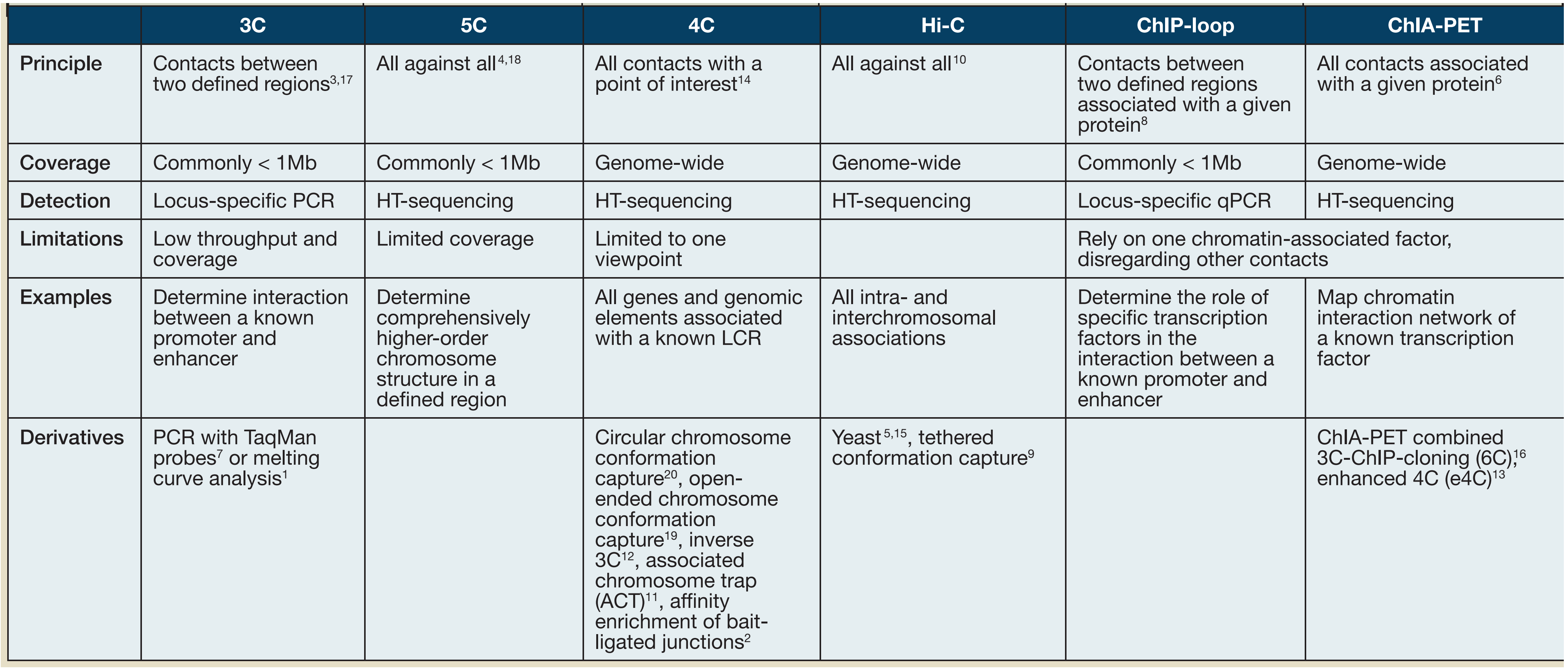
Chromatin-associated factors

Gene

3C | 5C | 4C | Hi-C | ChIP-loop | ChIA-PET

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

Chromosome Conformation Capture

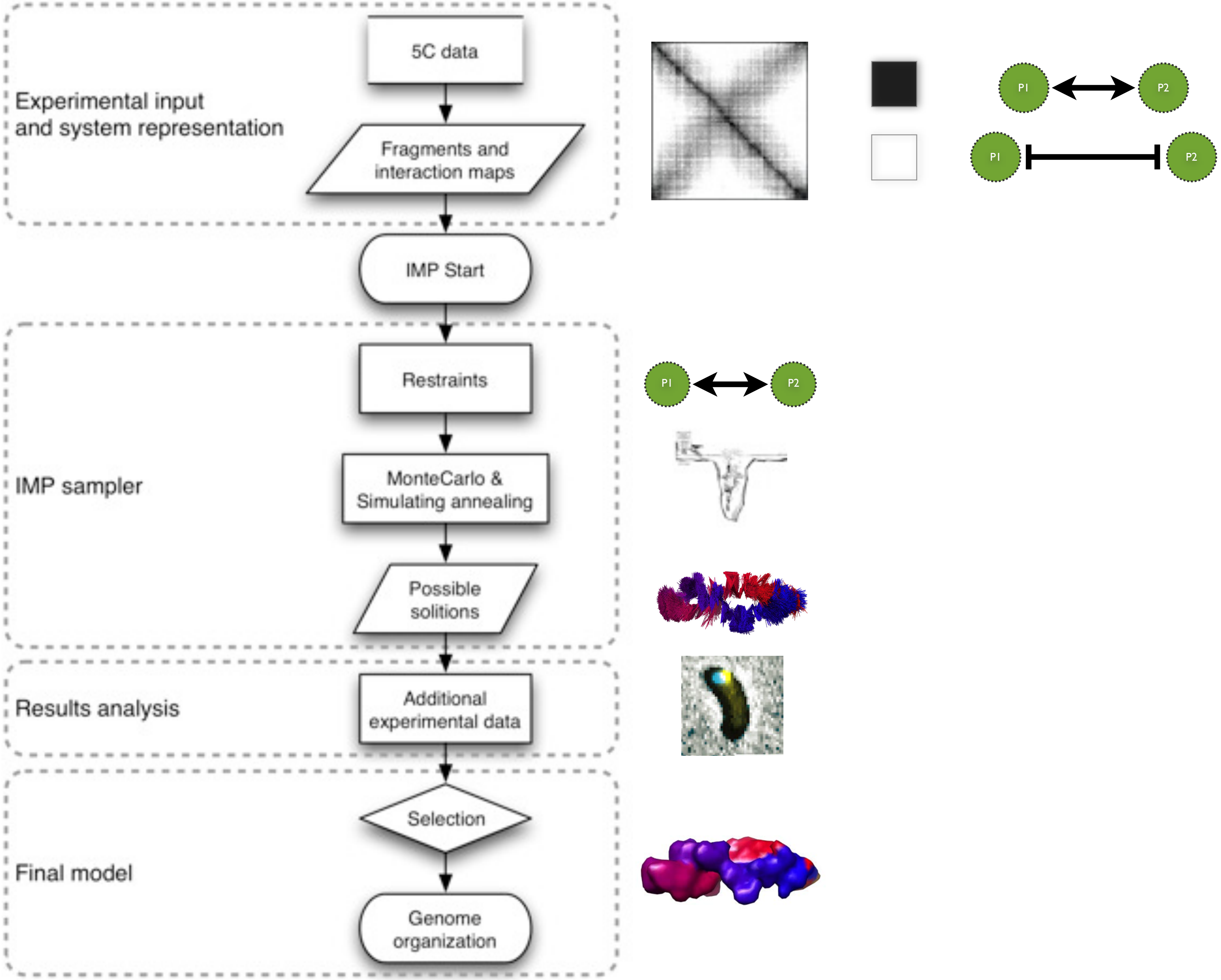


	3C	5C	4C	Hi-C	ChIP-loop	ChIA-PET
Principle	Contacts between two defined regions[3,17]	All against all[4,18]	All contacts with a point of interest[14]	All against all[10]	Contacts between two defined regions associated with a given protein[8]	All contacts associated with a given protein[6]
Coverage	Commonly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

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

Modeling 3D Genomes

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



Experimental input and system representation

5C data

Fragments and interaction maps

IMP Start

IMP sampler

Restraints

MonteCarlo & Simulating annealing

Possible solitions

Results analysis

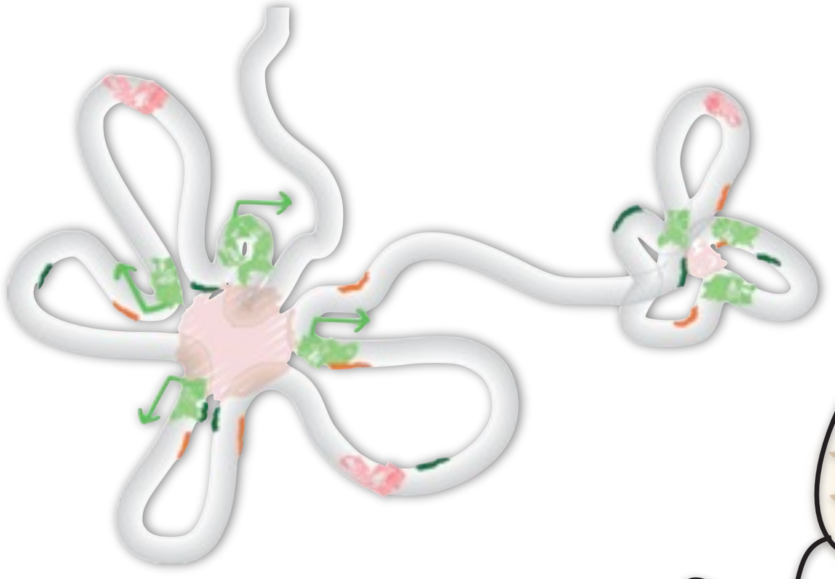
Additional experimental data

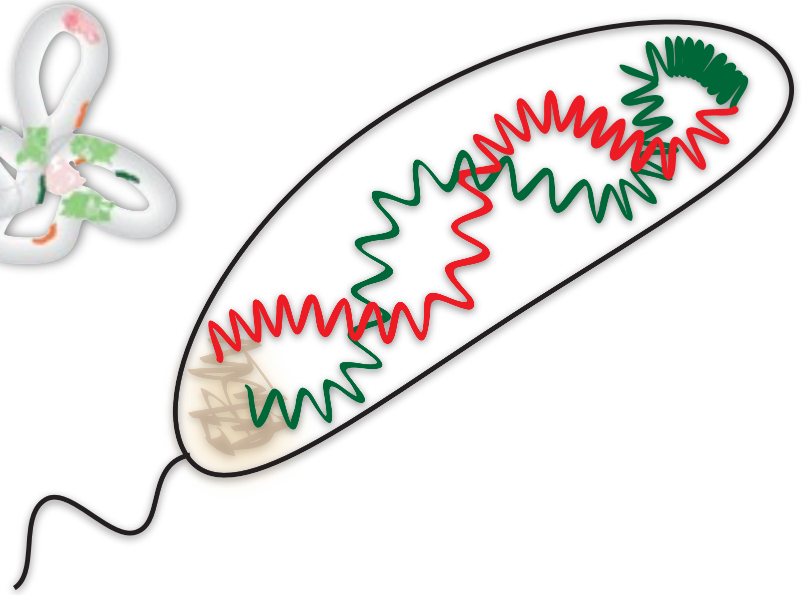
Final model

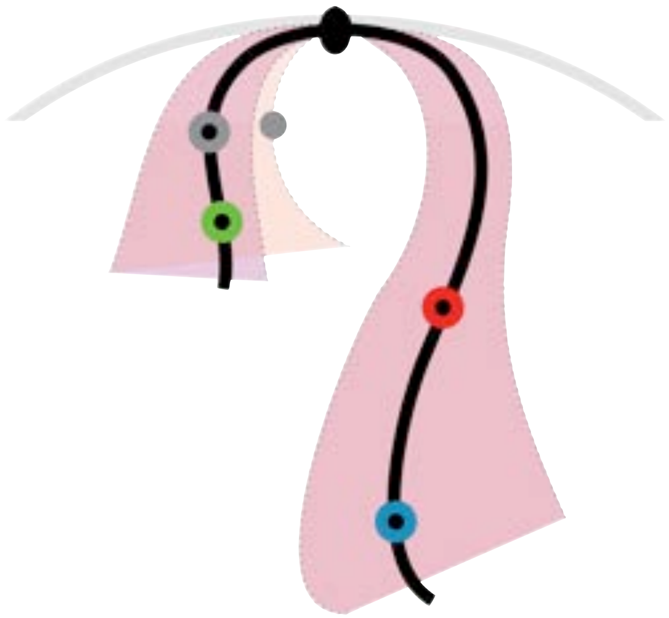
Selection

Genome organization

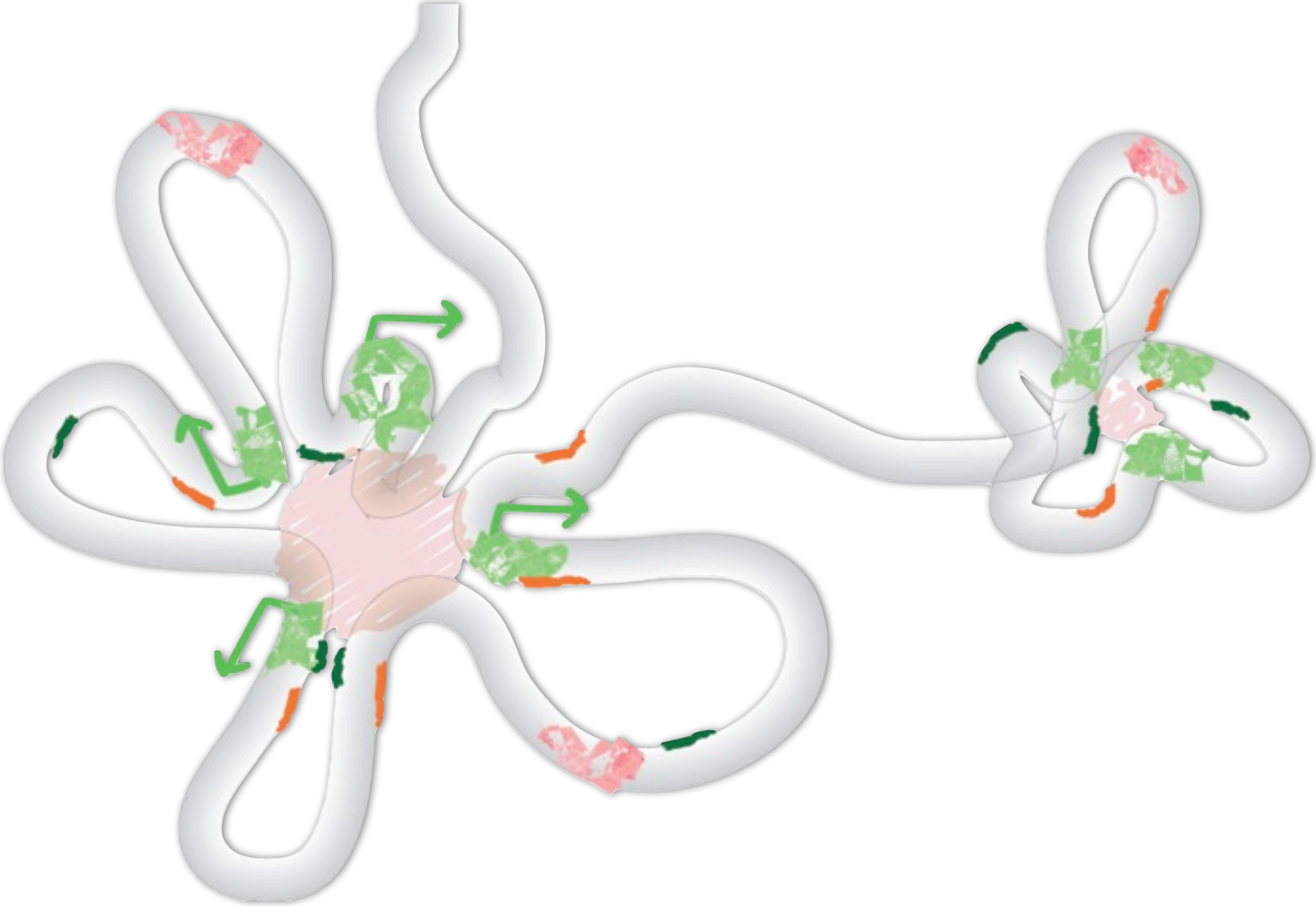
Examples...





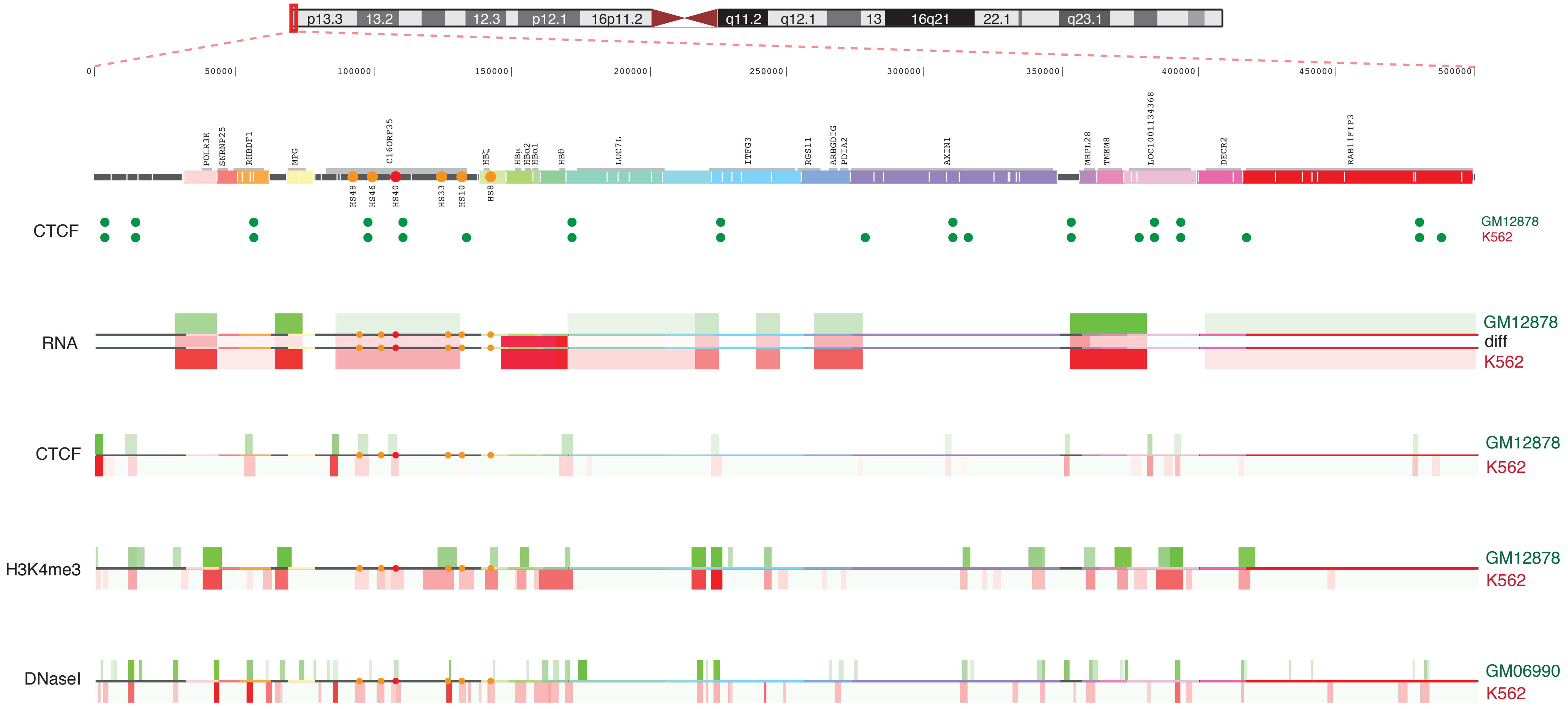


Human α-globin domain



Human α-globin domain

ENm008 genomic structure and environment

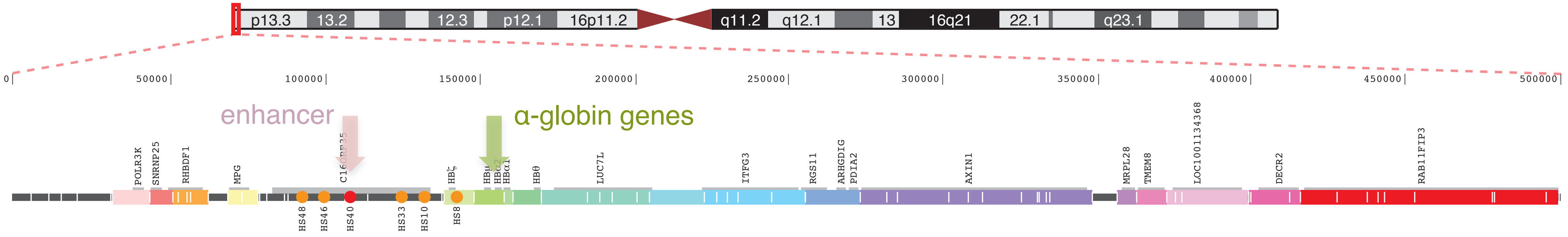


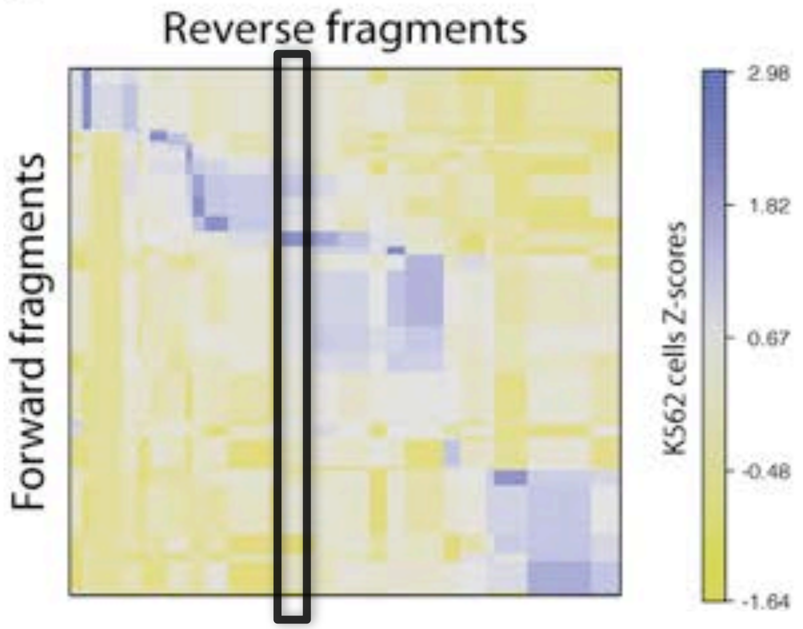
The ENCODE data for ENm008 region was obtained from the UCSC Genome Browser tracks for: RefSeq annotated genes, Affymetrix/CSHL expression data (Gingeras Group at Cold Spring Harbor), Duke/NHGRI DNaseI Hypersensitivity data (Crawford Group at Duke University), and Histone Modifications by Broad Institute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

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

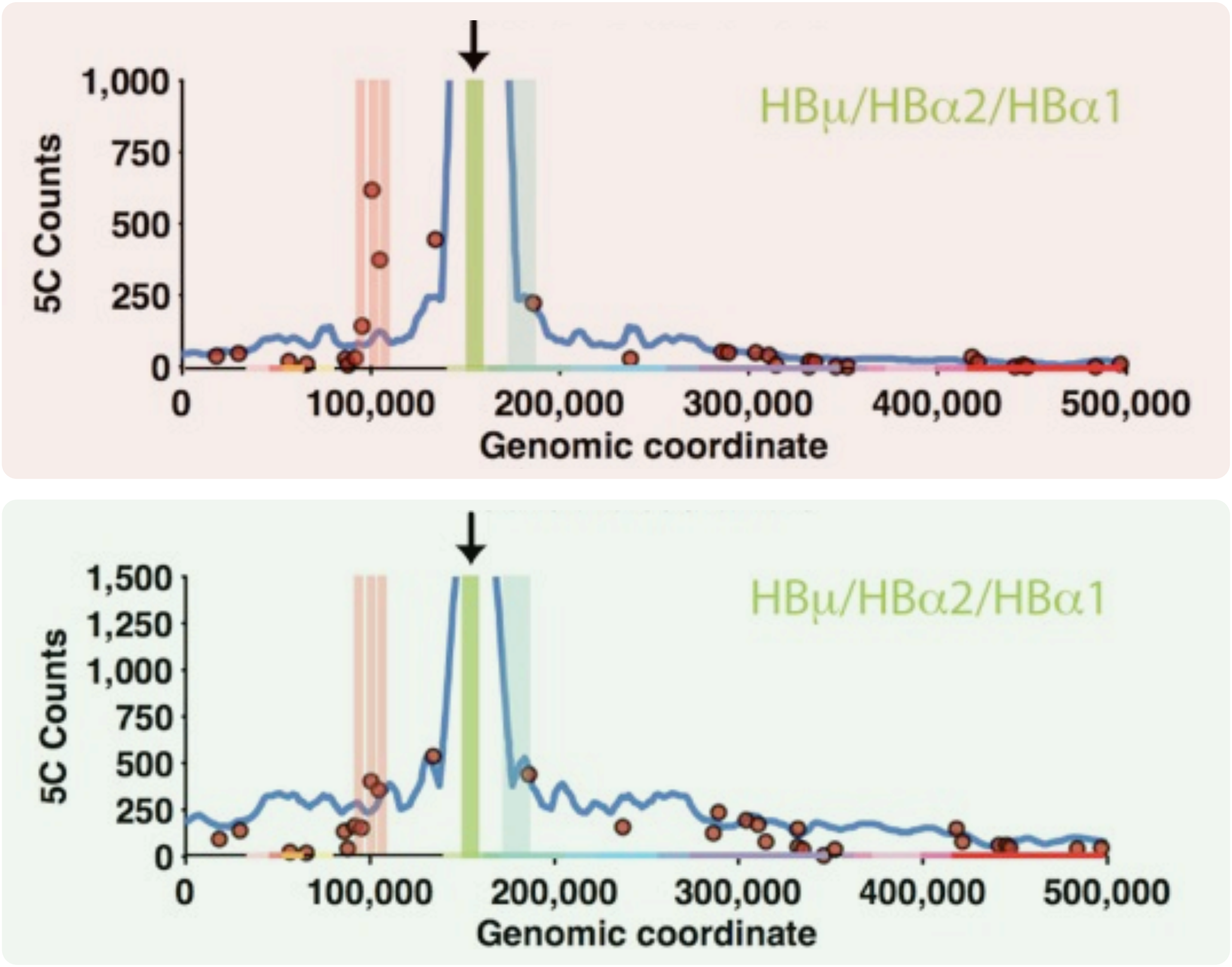
Human α-globin domain

ENm008 genomic structure and environment





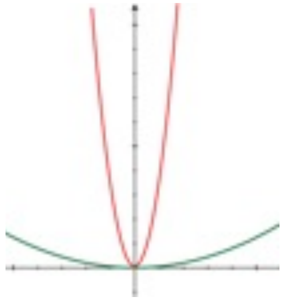
K562 cells:
α-globin genes active



Representation

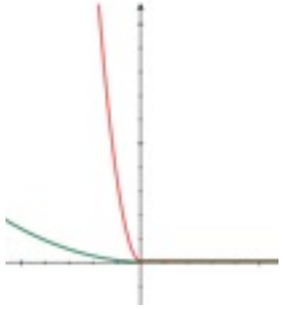
Harmonic

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



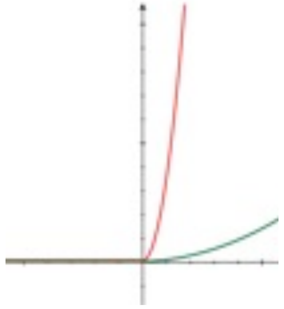
Harmonic Lower Bound

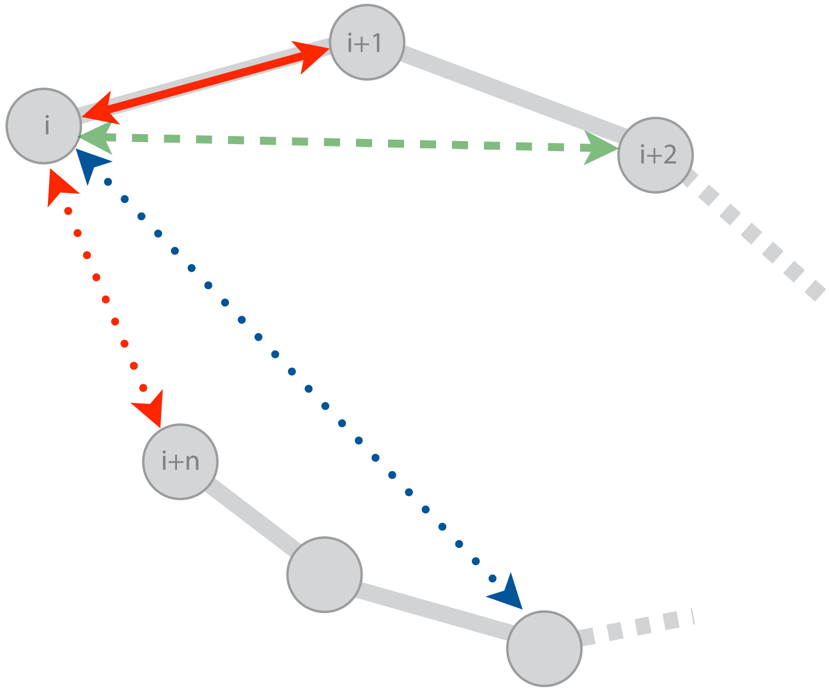
$$\begin{cases} if \;\; d_{i,j} \le d^{0}_{i,j}; & lbH_{i,j} = k\left(d_{i,j} - d^{0}_{i,j}\right)^{2} \\ if \;\; d_{i,j} > d^{0}_{i,j}; & lbH_{i,j} = 0 \end{cases}$$



Harmonic Upper Bound

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

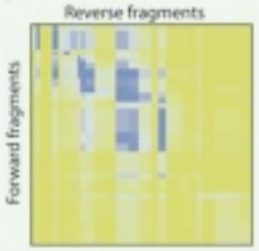


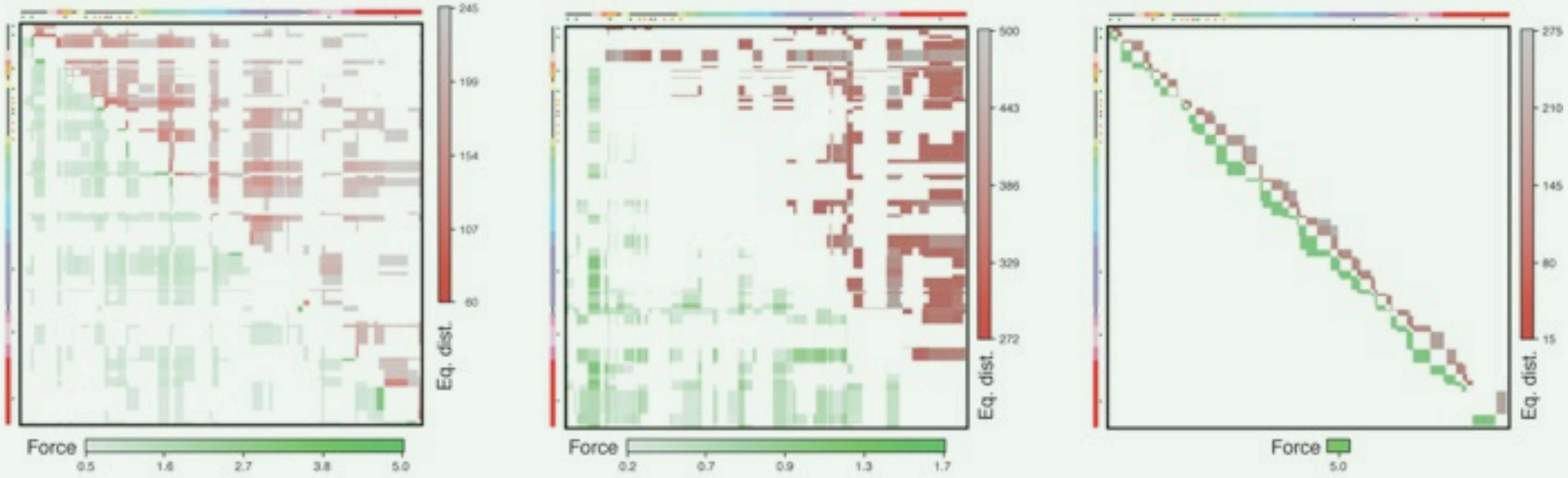


Scoring

GM12878

70 fragments
1,520 restraints





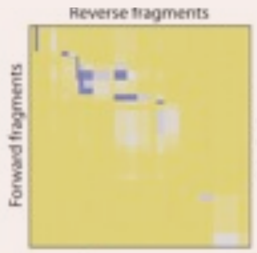
Harmonic

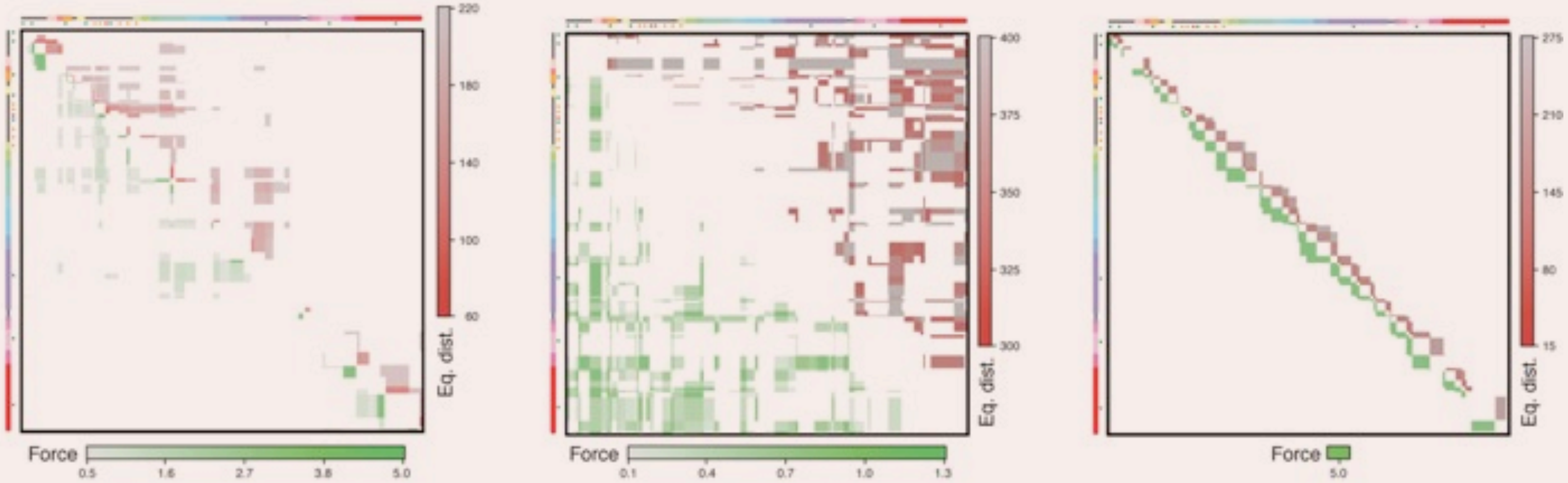
Harmonic Lower Bound

Harmonic Upper Bound

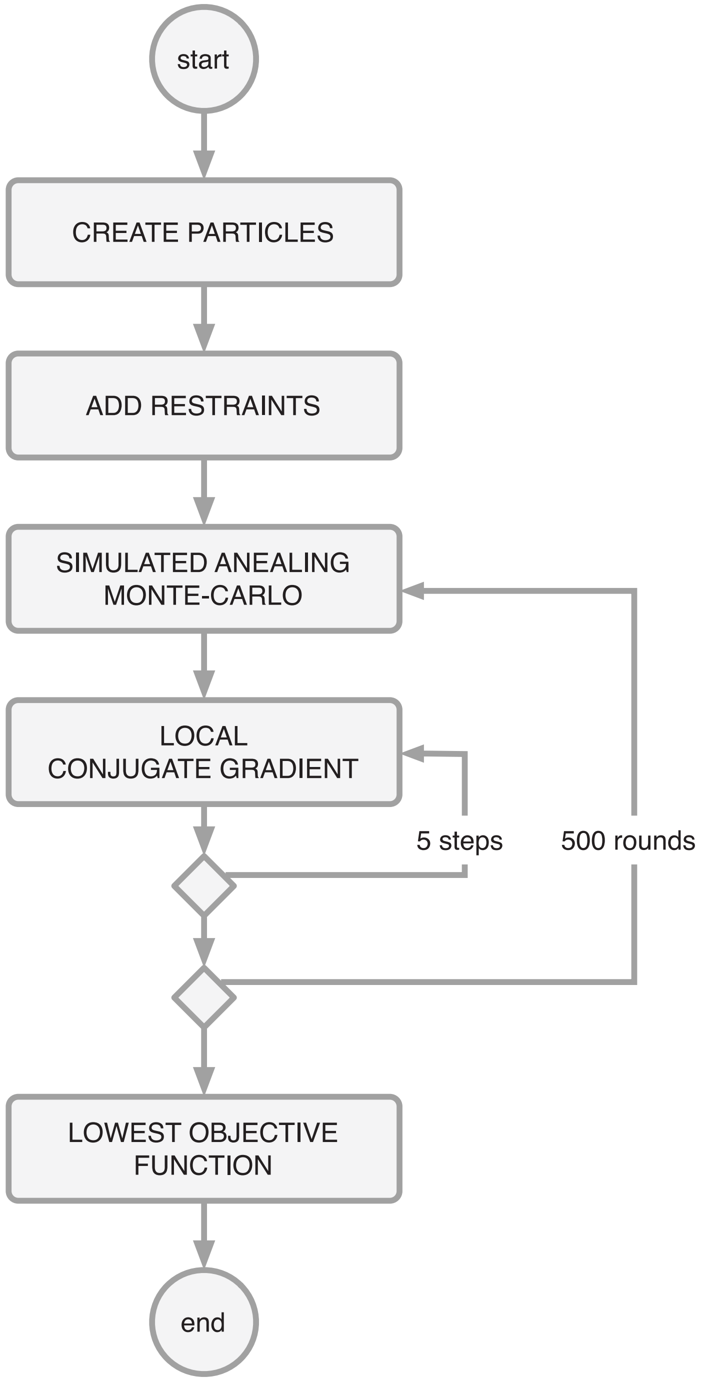
K562

70 fragments
1,049 restraints





Optimization



start

CREATE PARTICLES

ADD RESTRAINTS

SIMULATED ANEALING
MONTE-CARLO

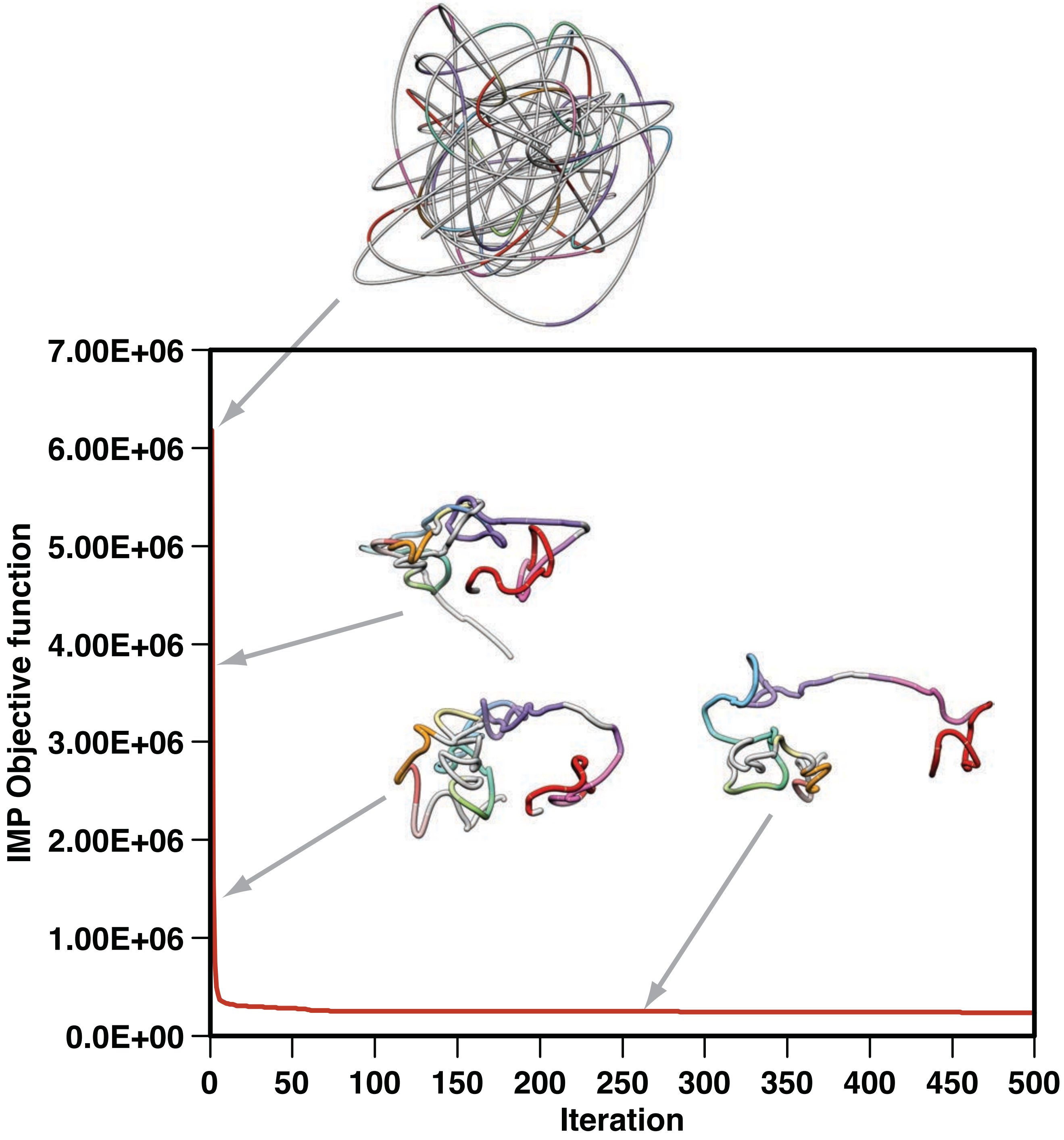
LOCAL
CONJUGATE GRADIENT

5 steps

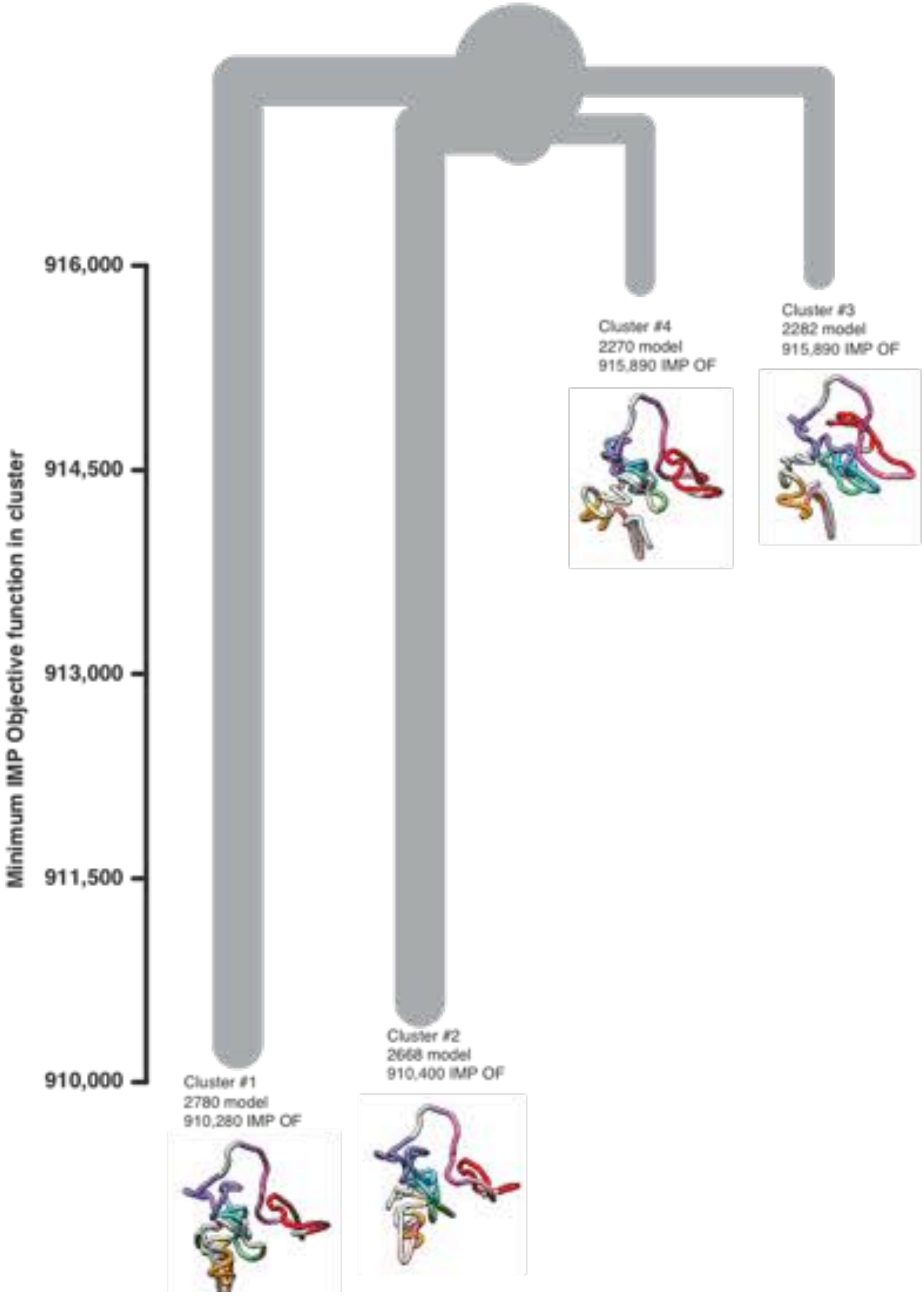
500 rounds

LOWEST OBJECTIVE
FUNCTION

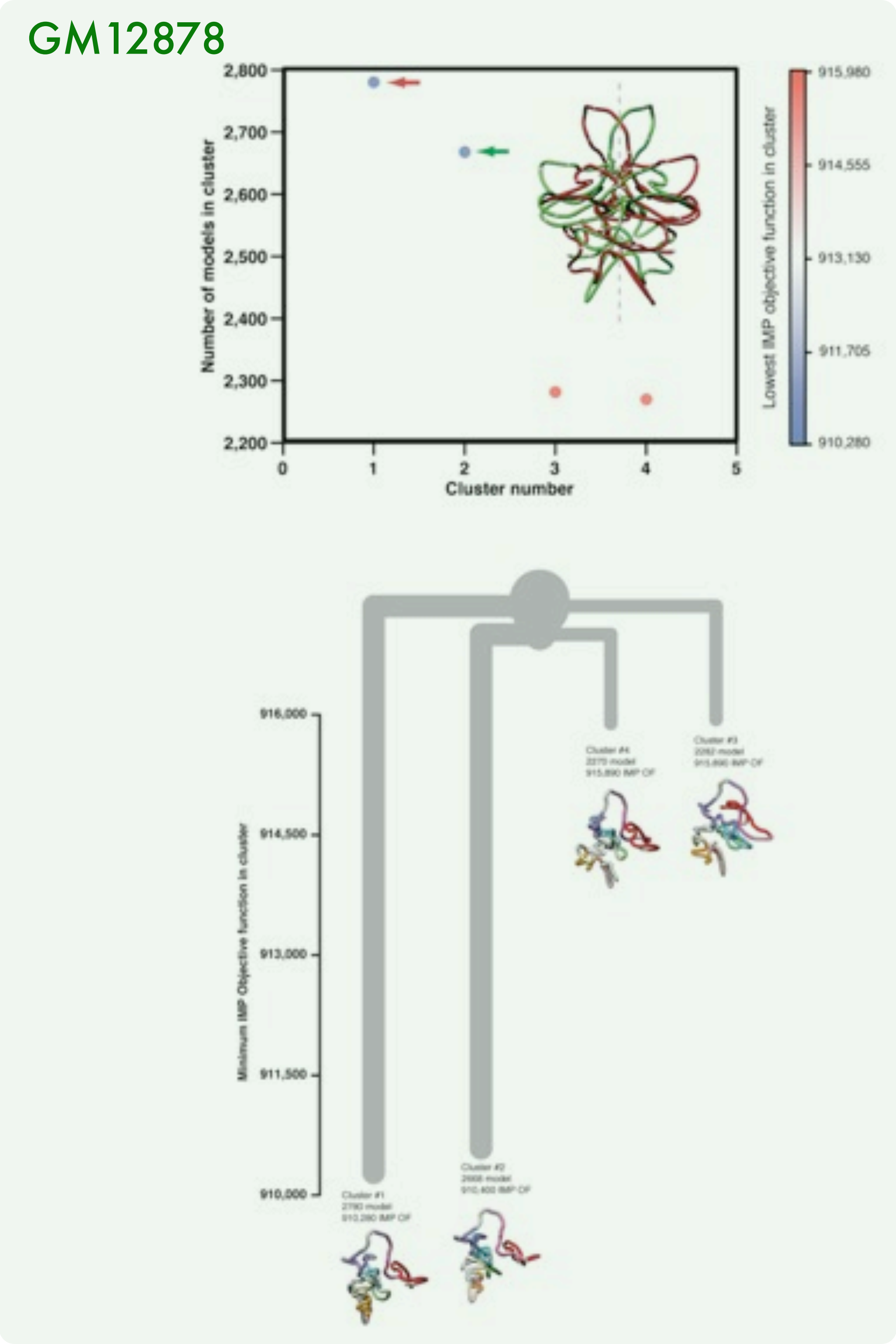
end

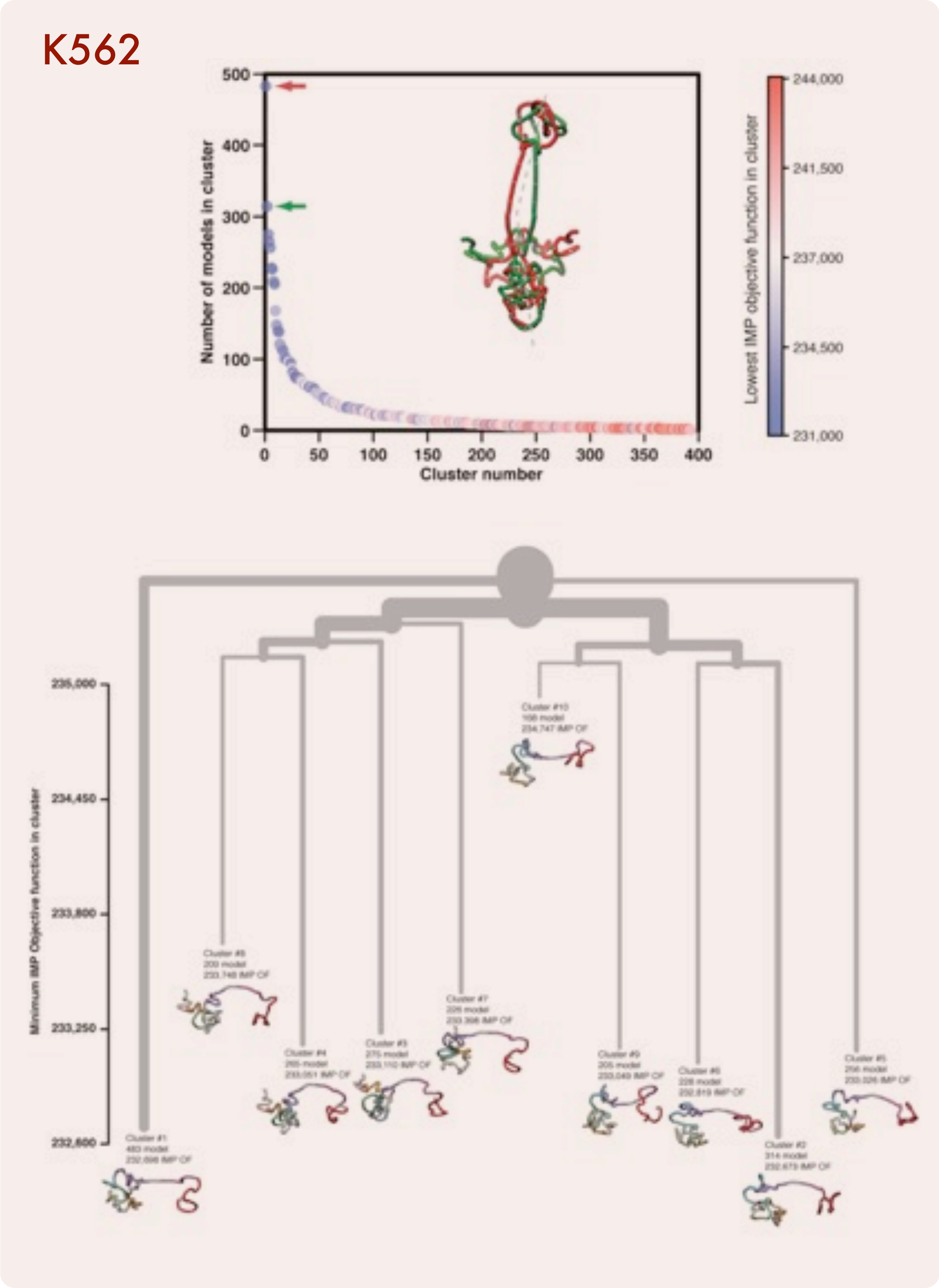


Clustering


Minimum IMP Objective function in cluster
916,000
914,500
913,000
911,500
910,000
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

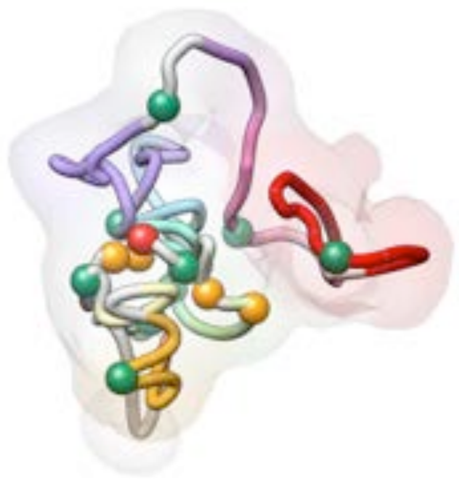


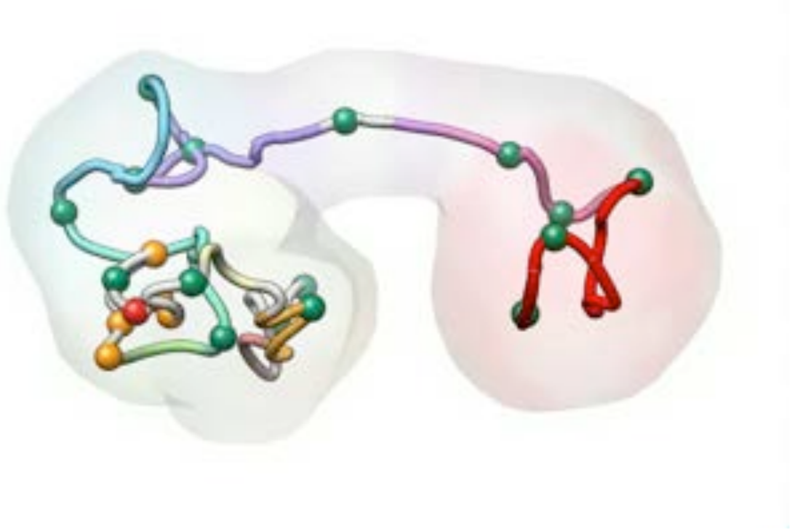


Consistency

GM12878

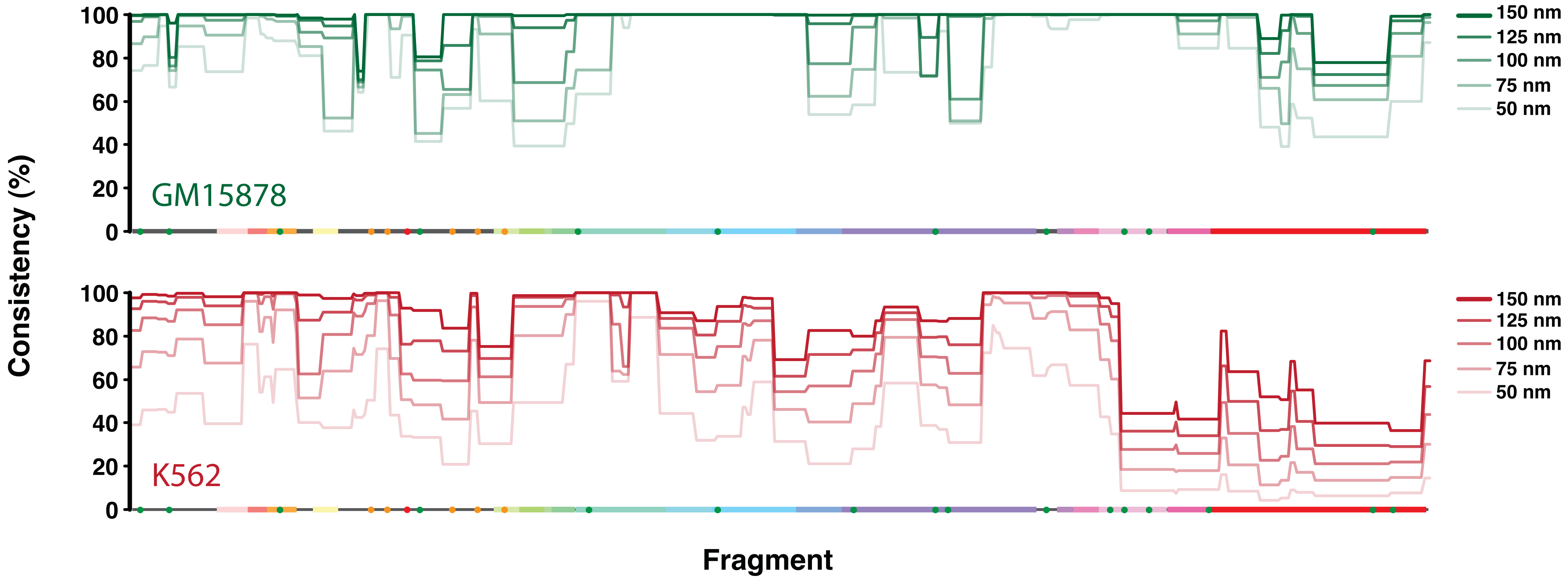
Cluster #1
2780 model





K562

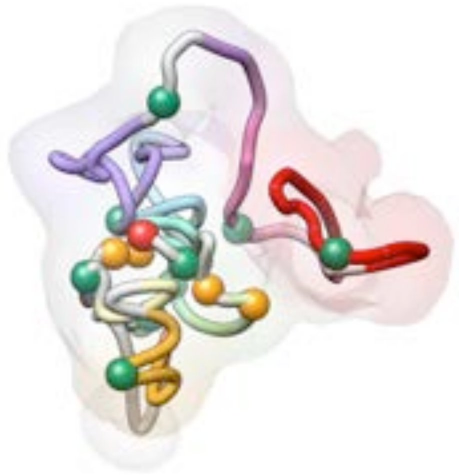
Cluster #2
314 model

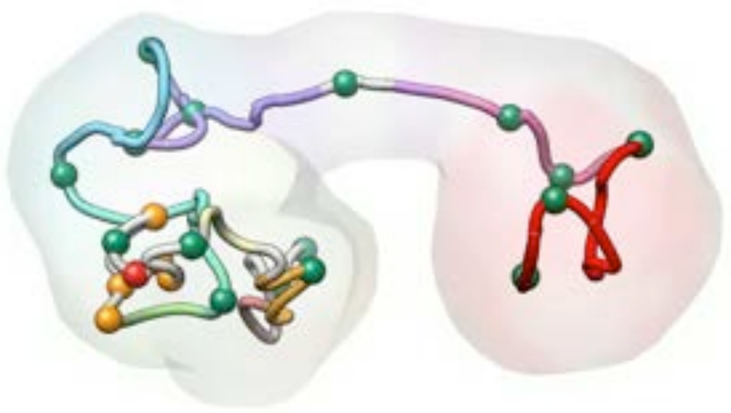


Regulatory elements

GM12878

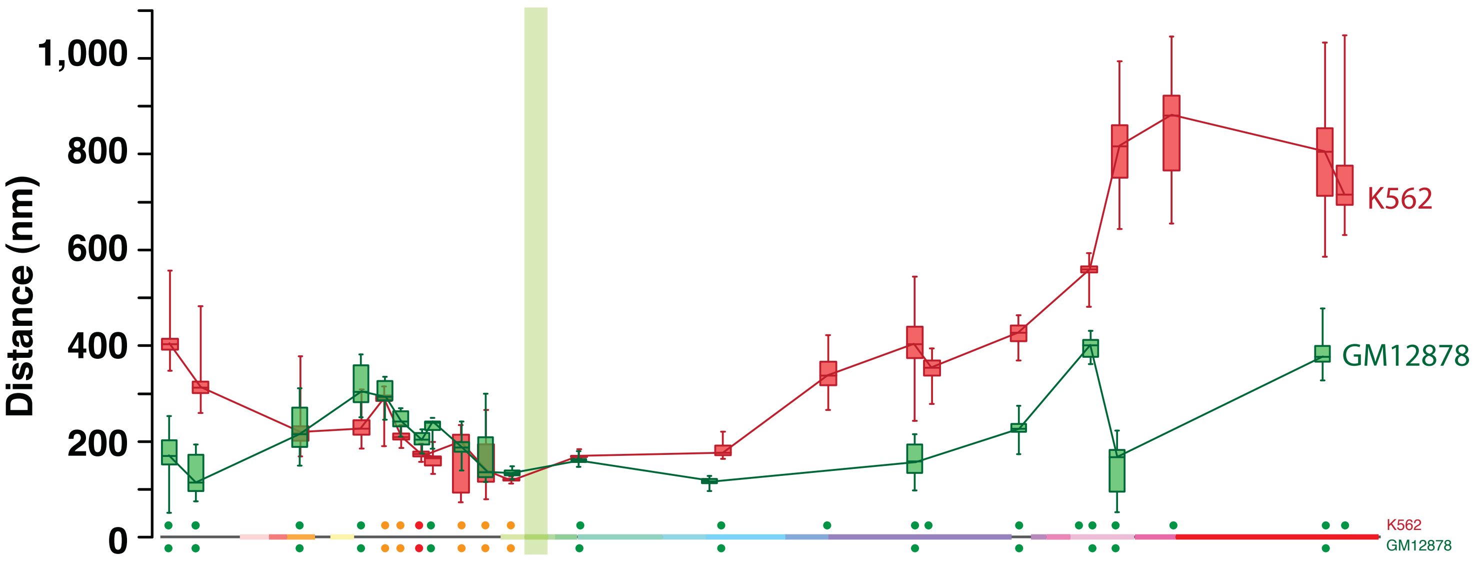
Cluster #1
2780 model

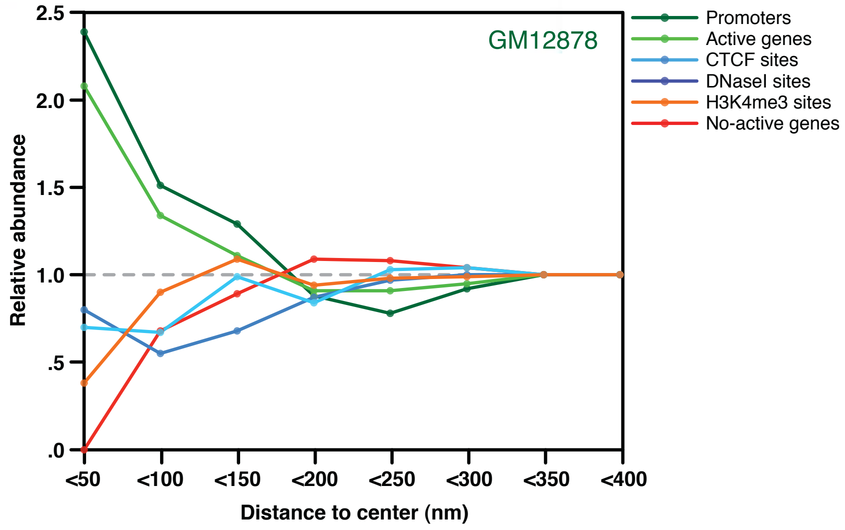


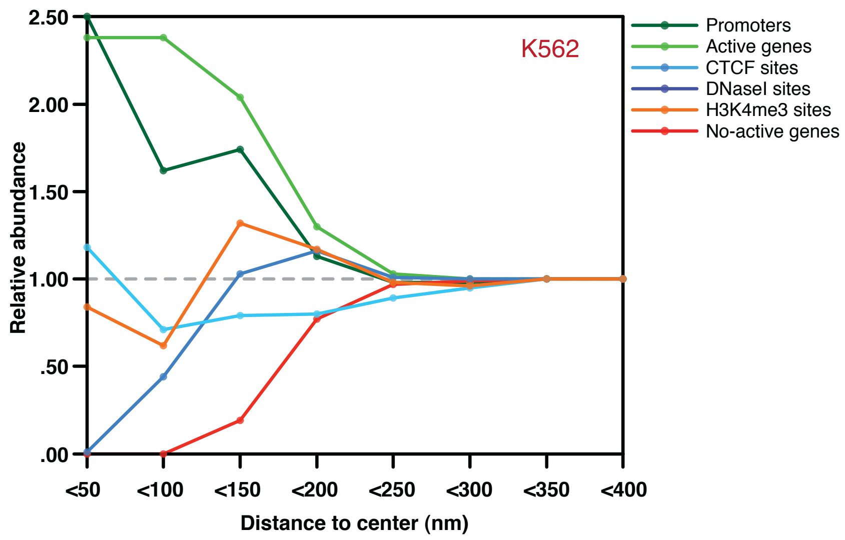


K562

Cluster #2
314 model

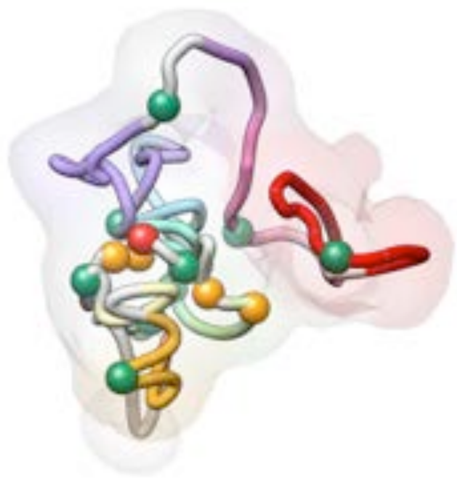


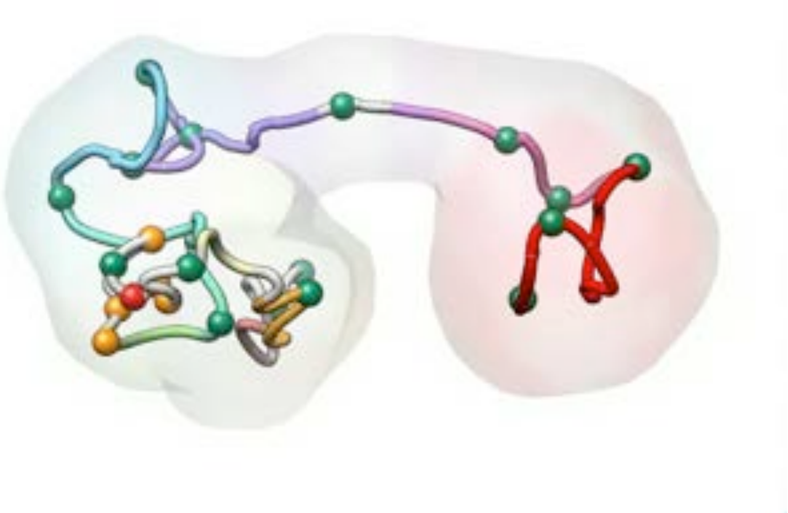




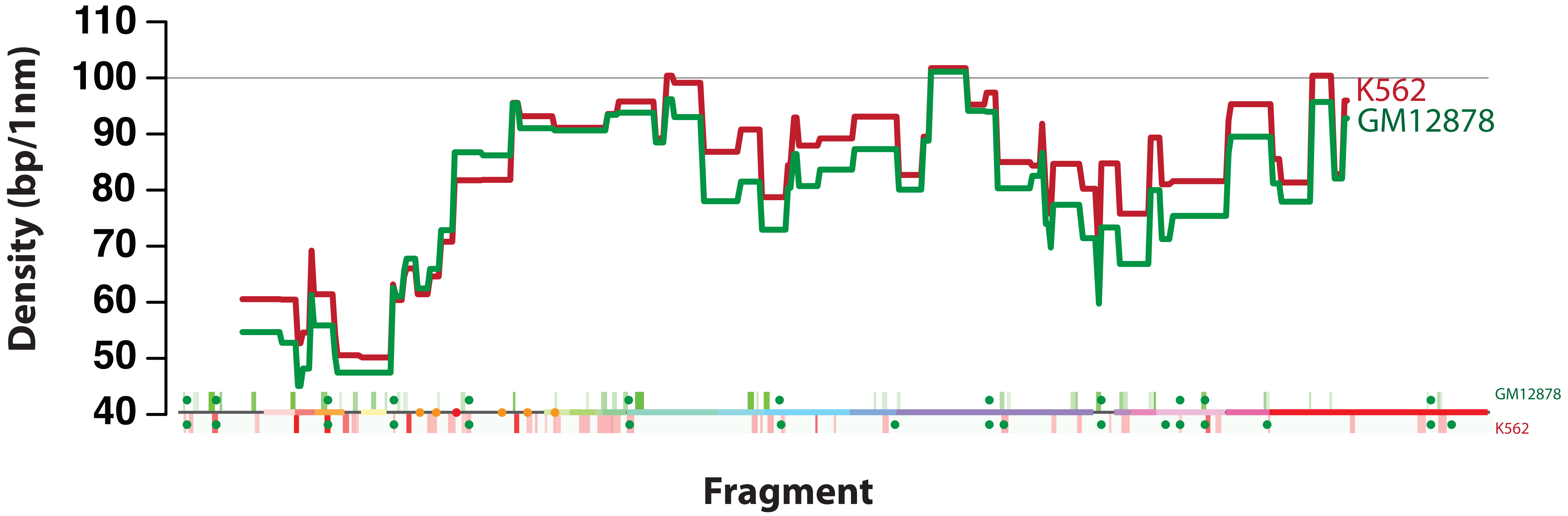
Compactness

GM12878
Cluster #1
2780 model





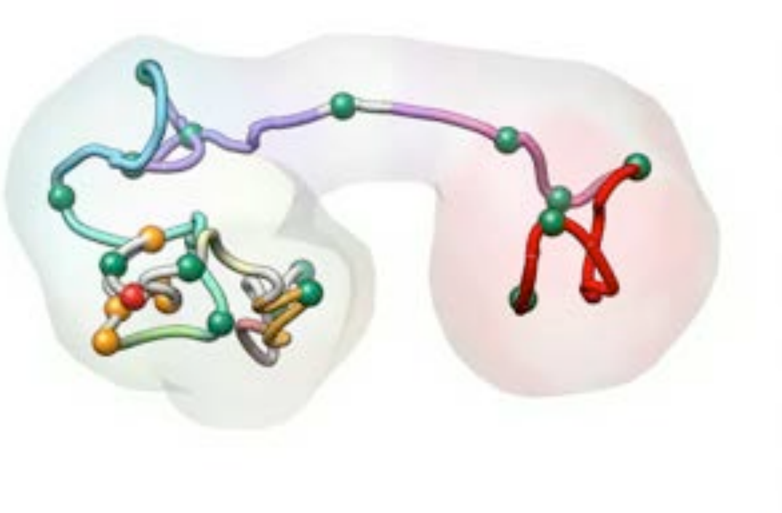
K562
Cluster #2
314 model



Multi-loops

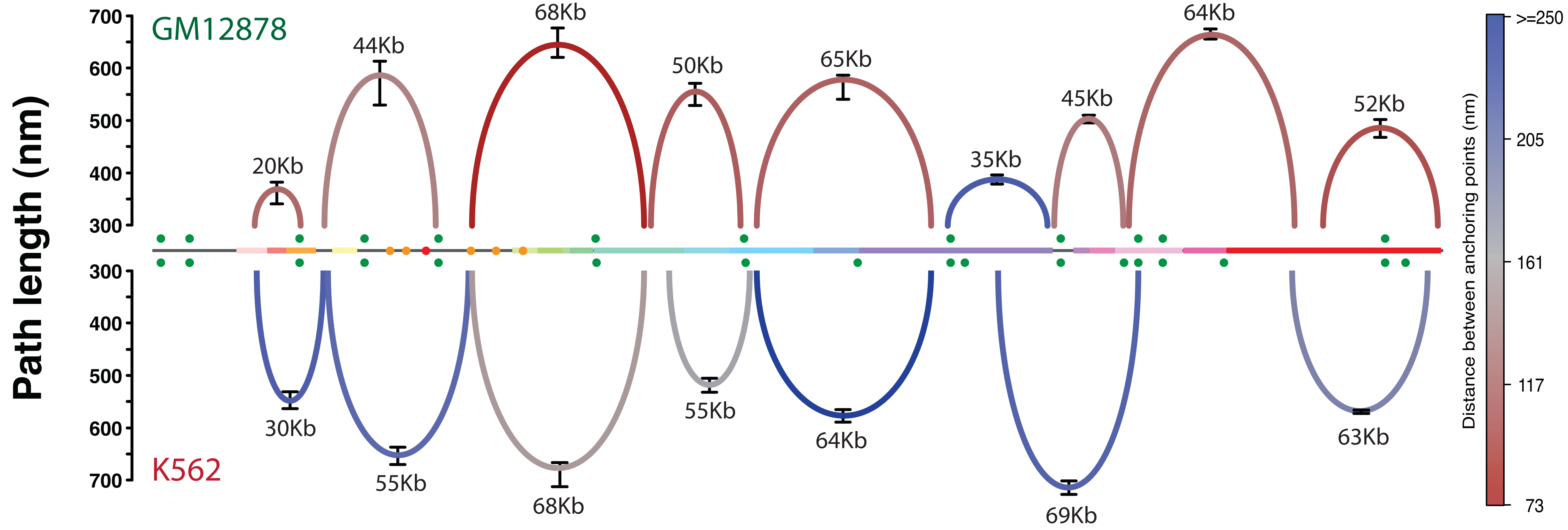
GM12878

Cluster #1
2780 model



K562

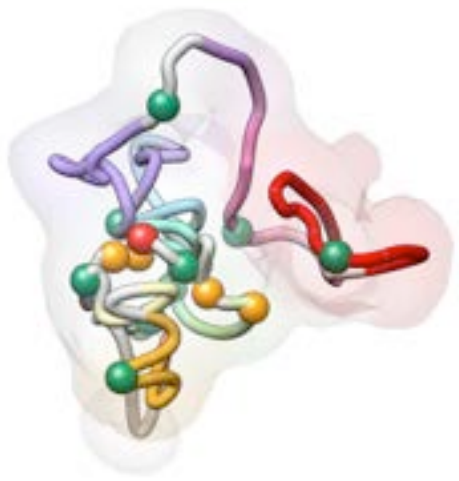
Cluster #2
314 model



Expression

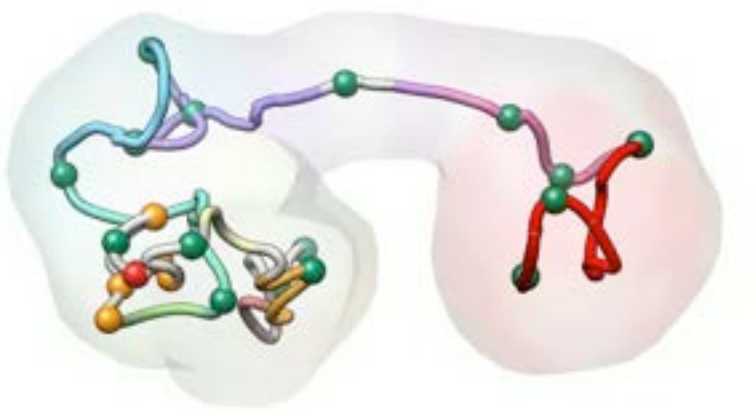
GM12878

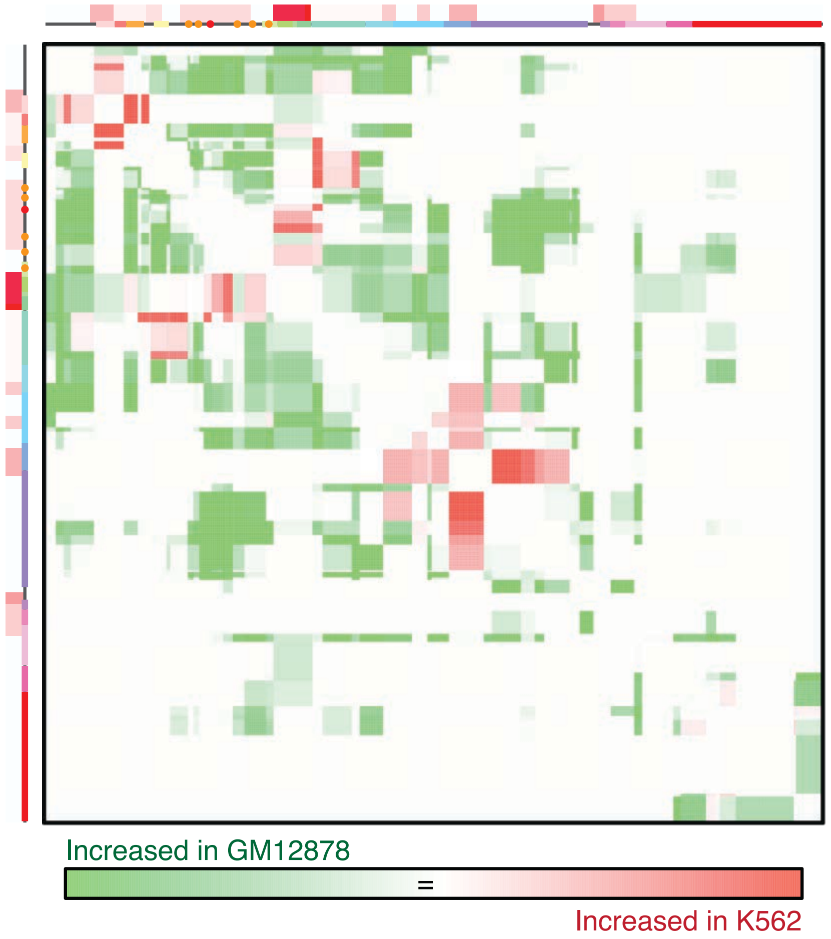
Cluster #1
2780 model



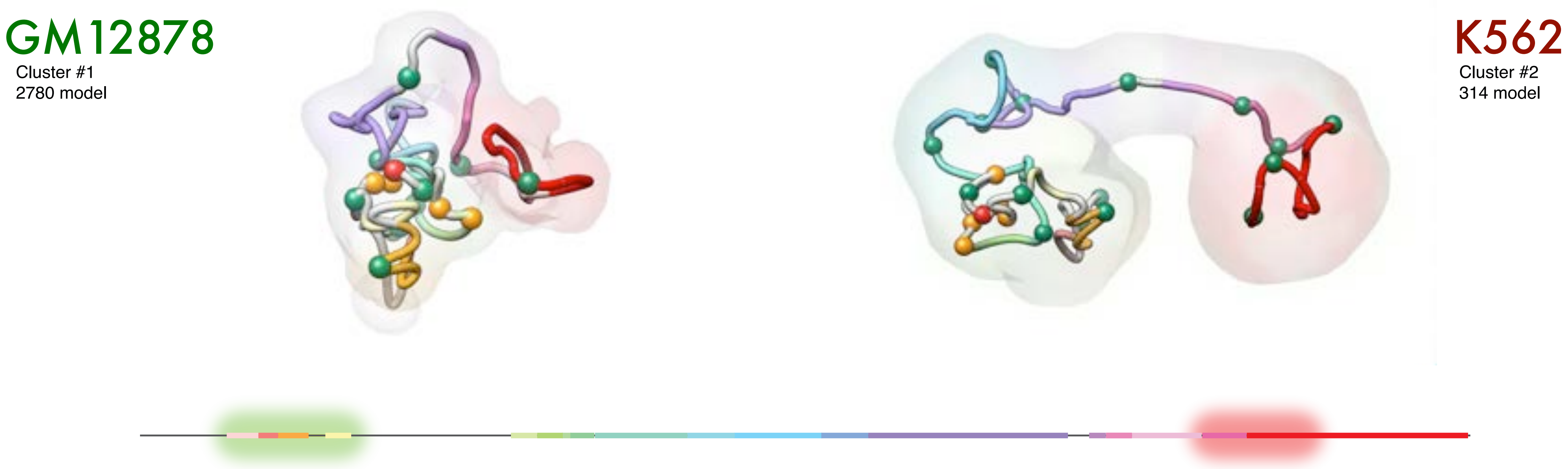
K562

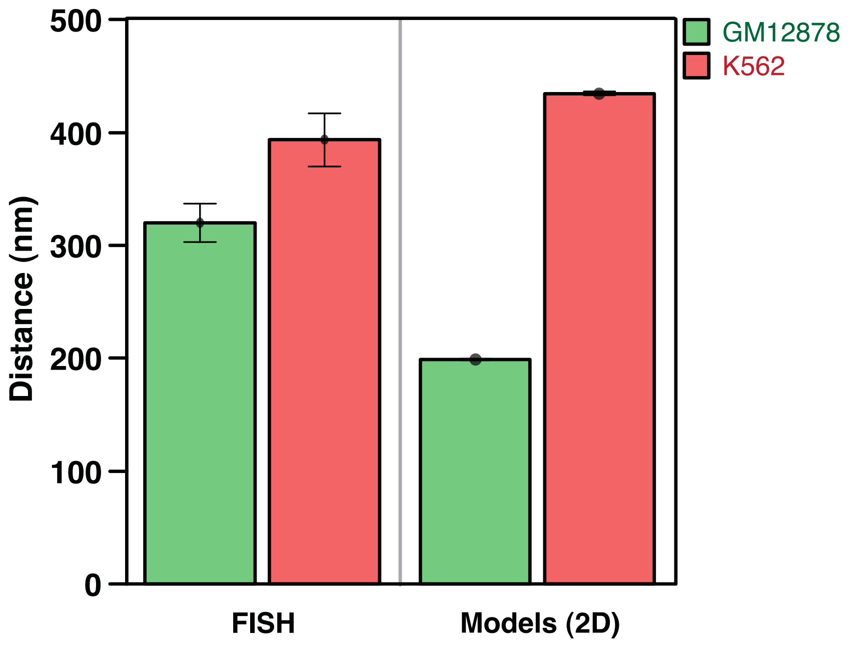
Cluster #2
314 model

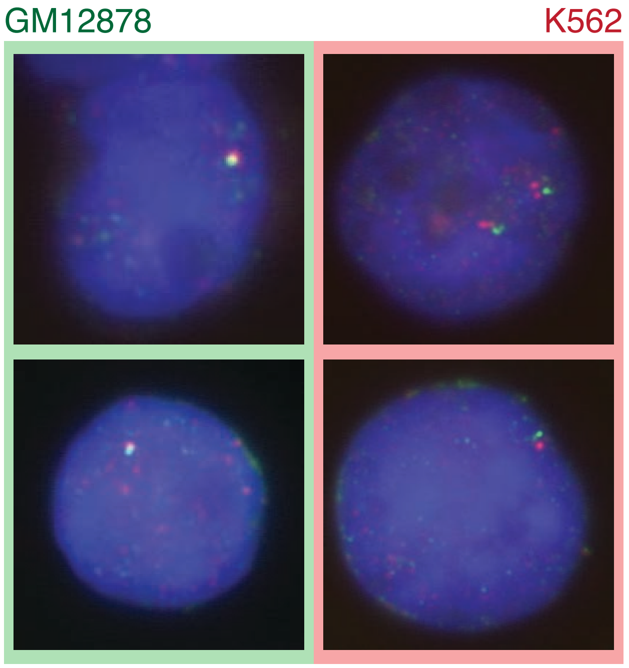




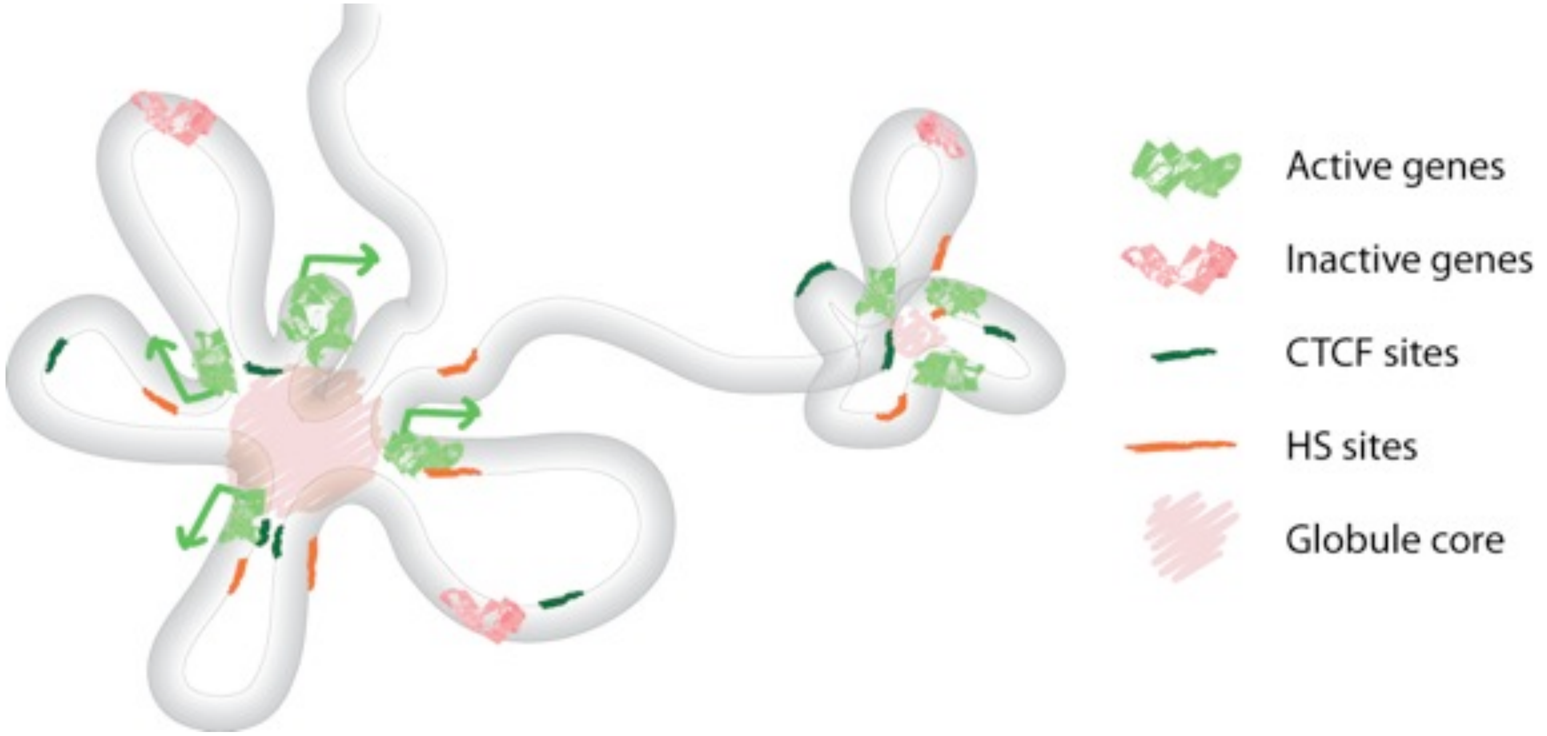
FISH validation

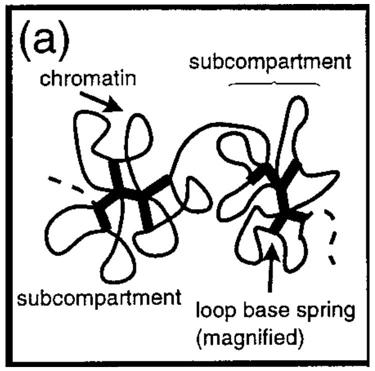




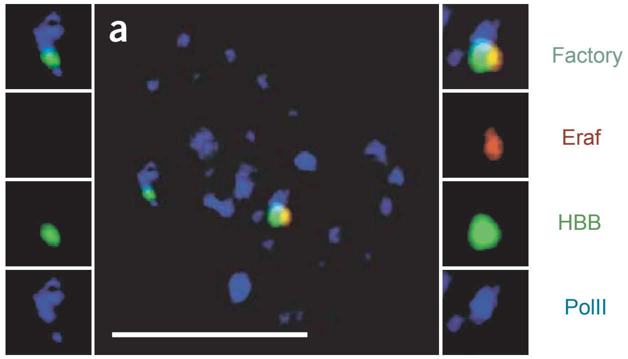


The “Chromatin Globule” model

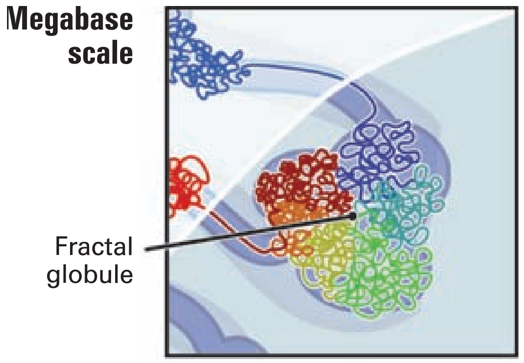




Münkel et al. JMB (1999)



Osborne et al. Nat Genet (2004)

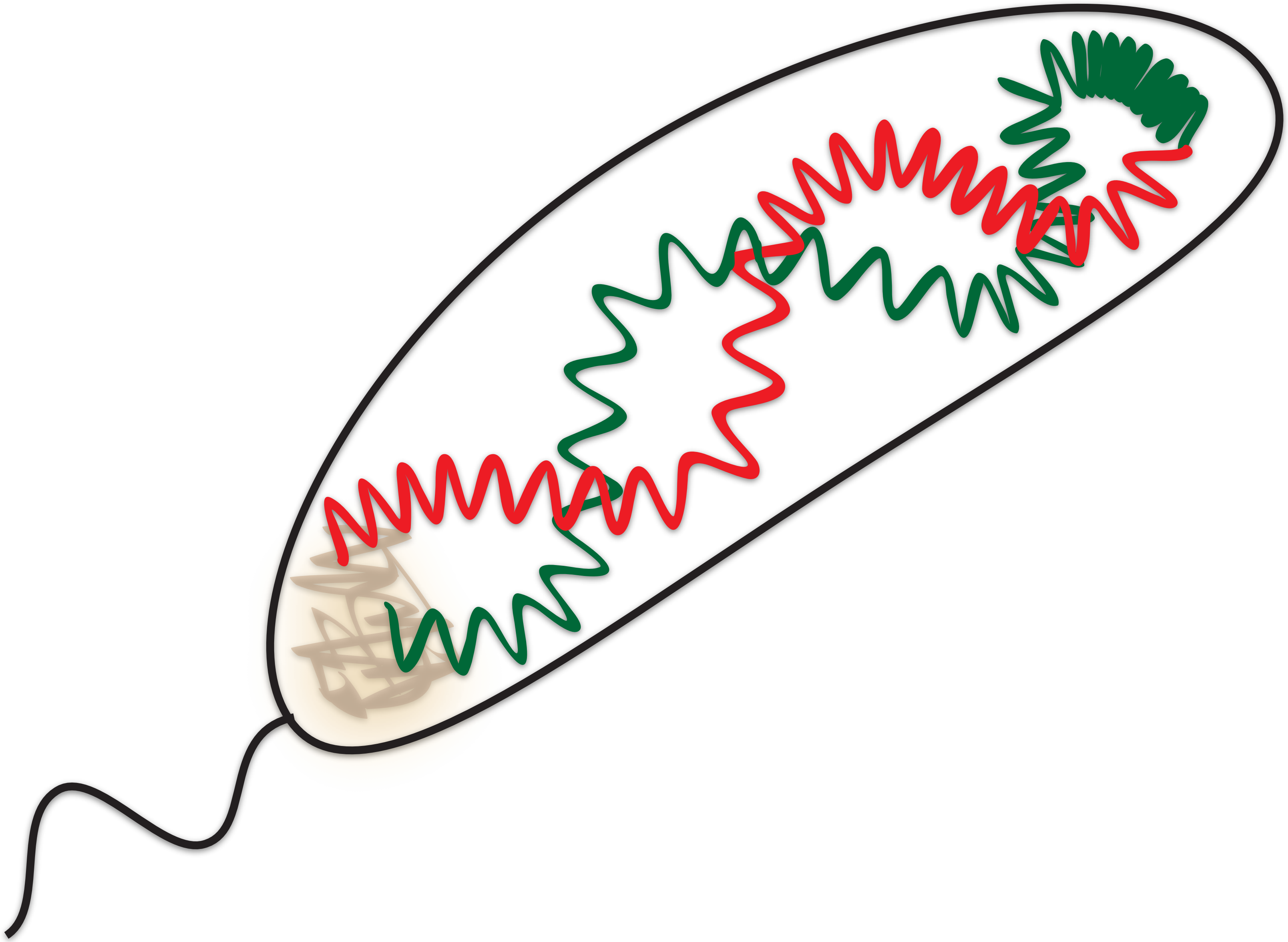


Lieberman-Aiden et al. Science (2009)

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

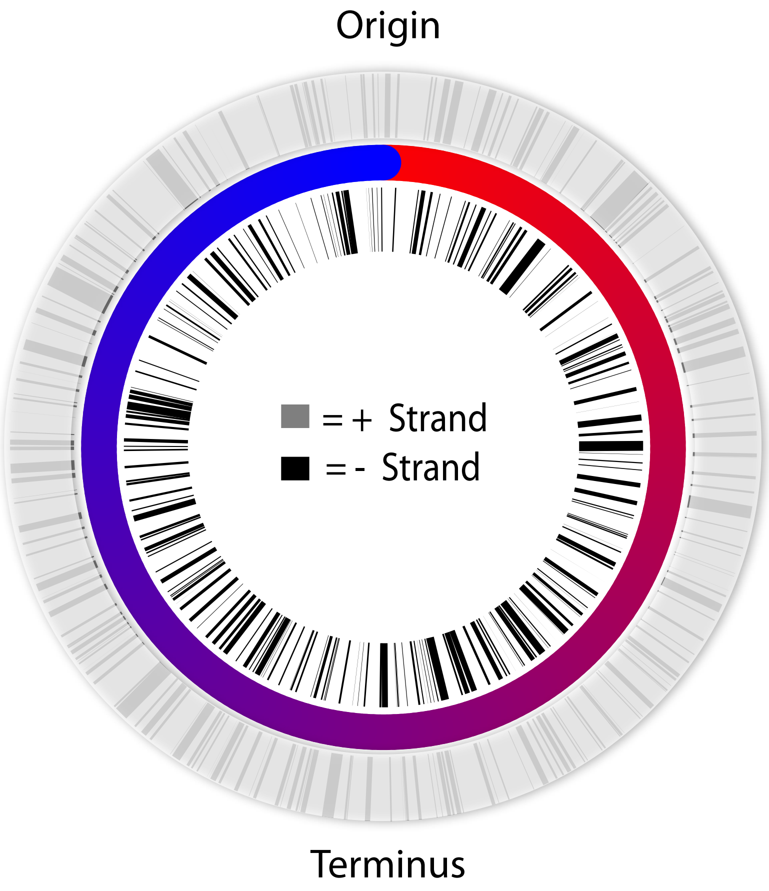
A. Sanyal *et al.* **Current Opinion in Cell Biology** (2011) 23:325–33.

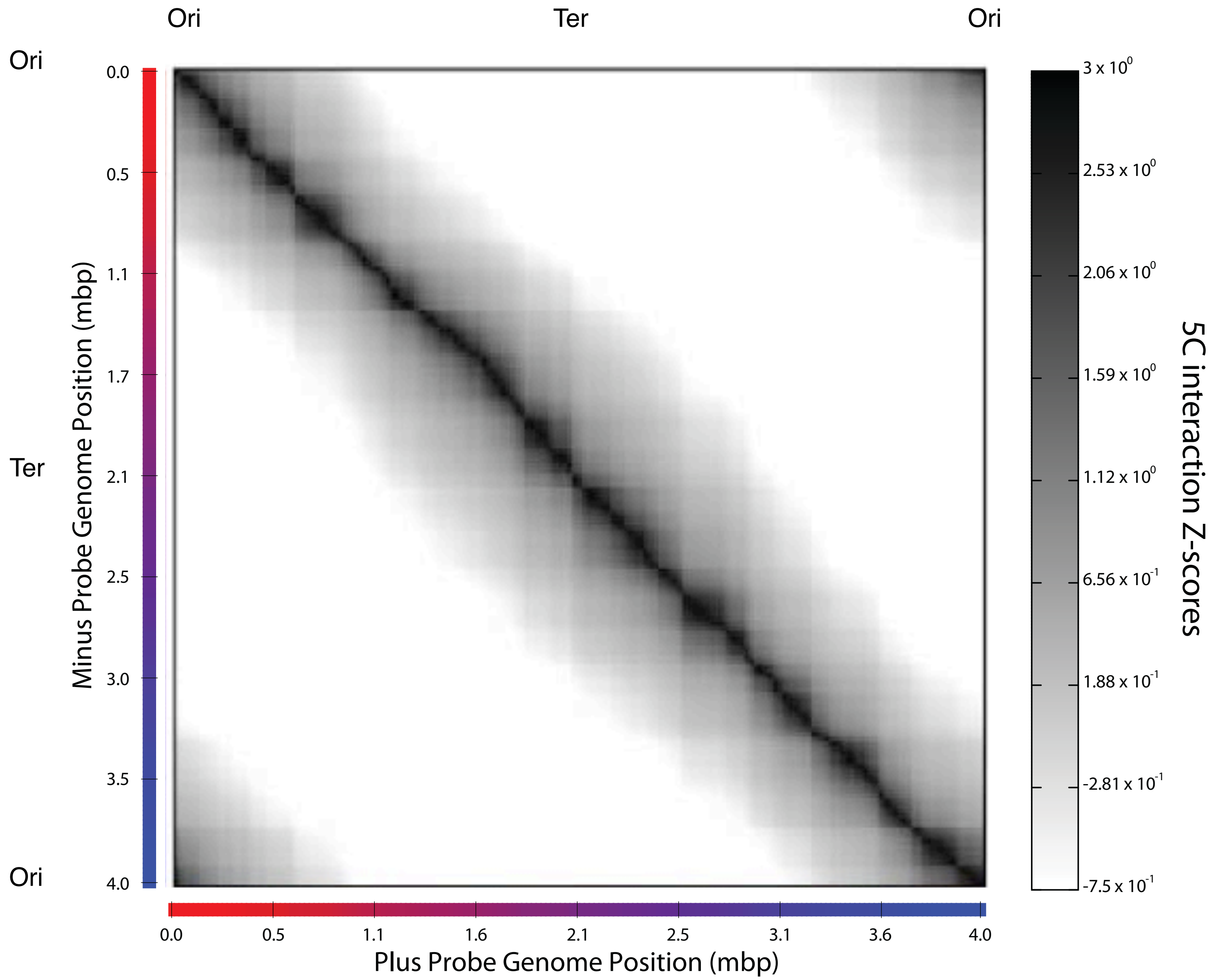
Caulobacter crescentus genome



The 3D architecture of Caulobacter Crescentus

4,016,942 bp & 3,767 genes





169 5C primers on + strand

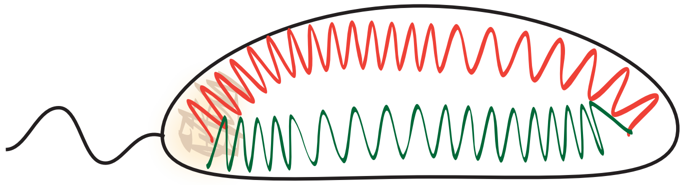
170 5C primers on – strand

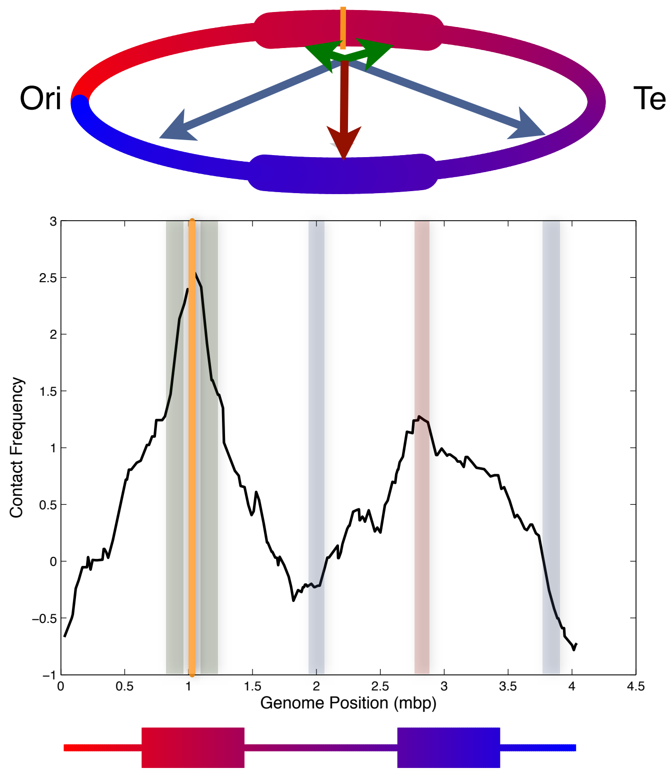
28,730 chromatin interactions

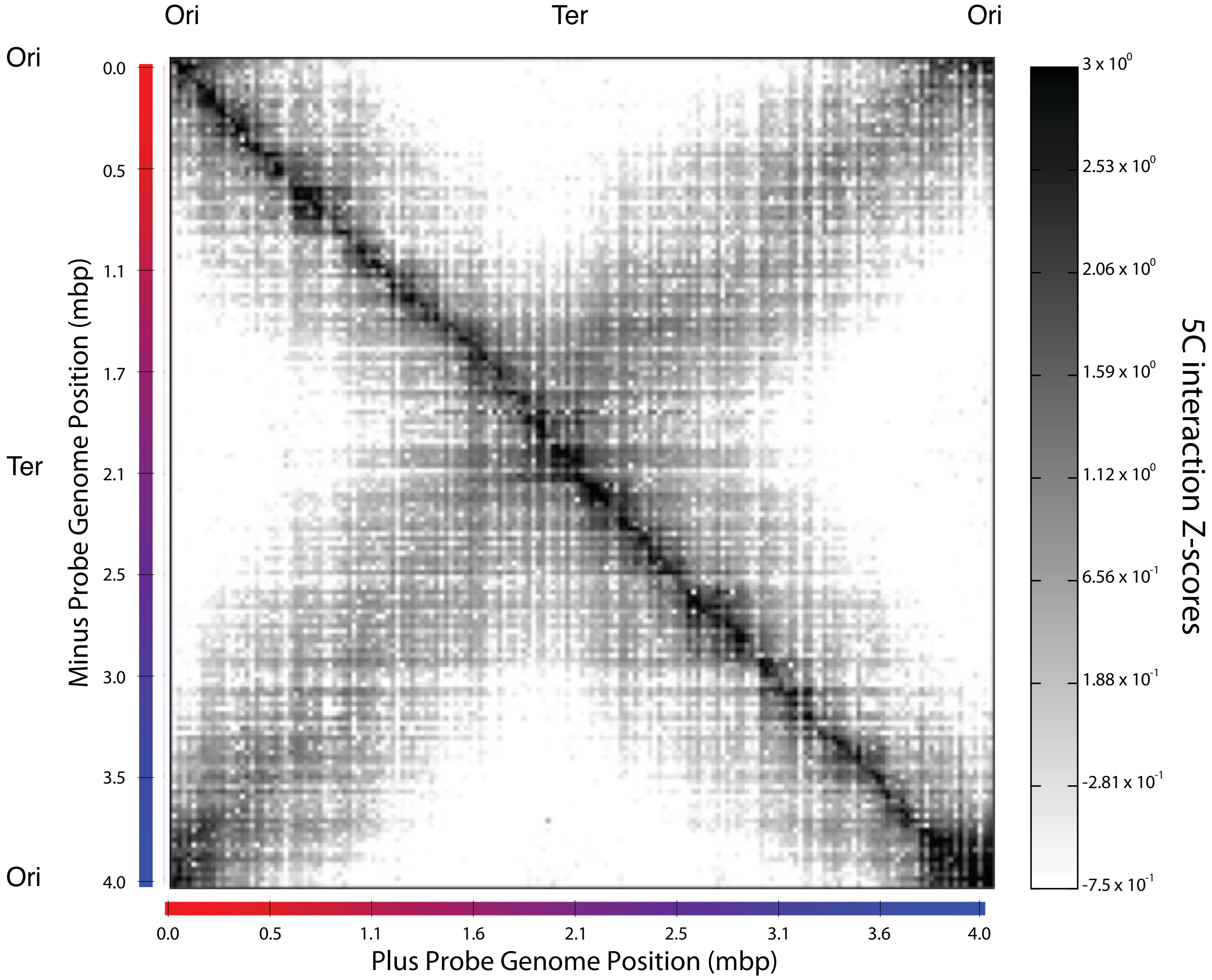
~13Kb

5C interaction matrix

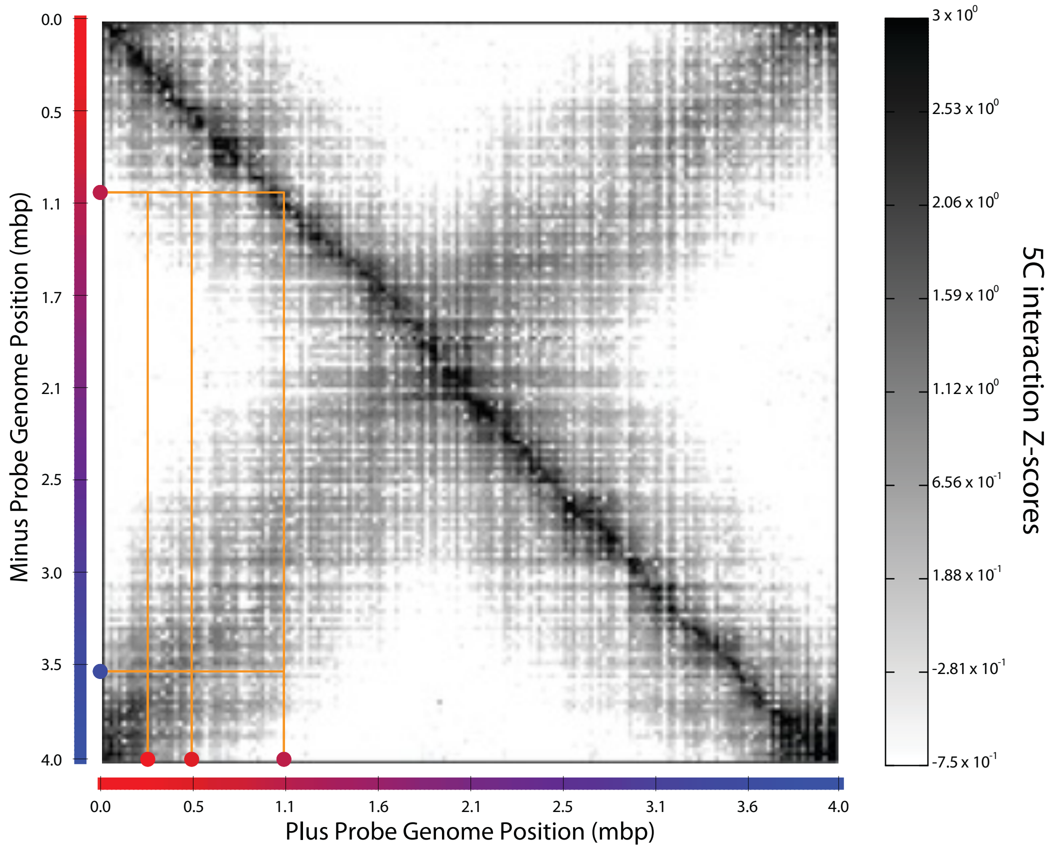
ELLIPSOID for Caulobacter cresentus



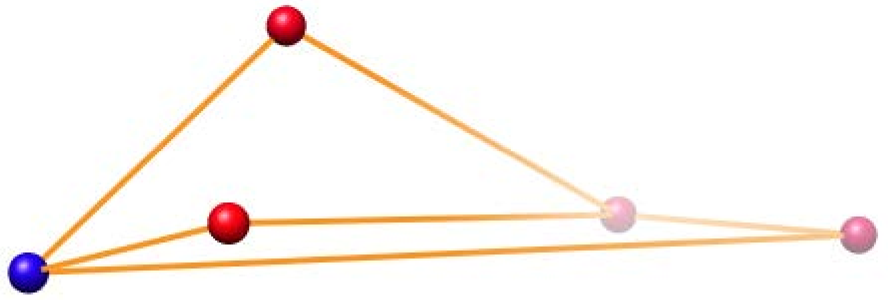


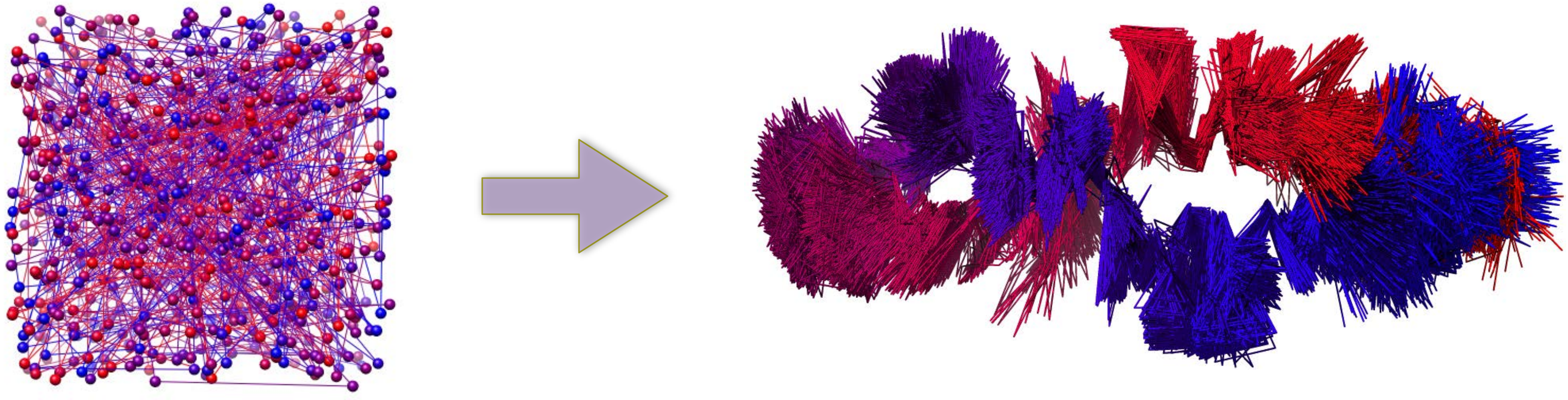


3D model building with the 5C + IMP approach



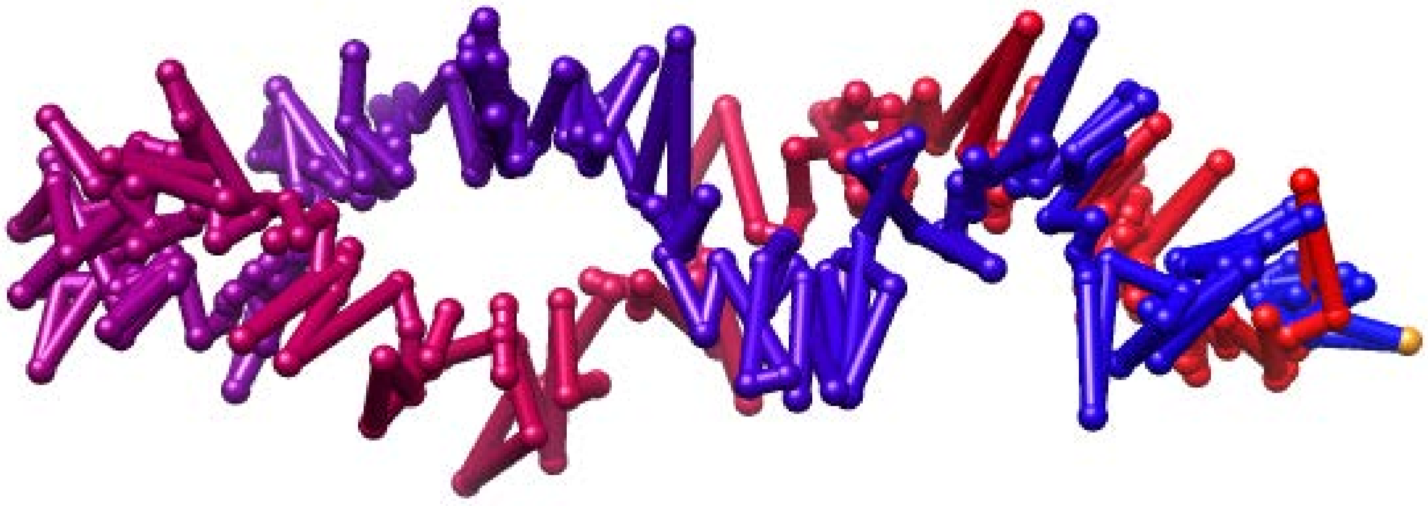
339 mers





Genome organization in Caulobacter crescentus

Arms are helical

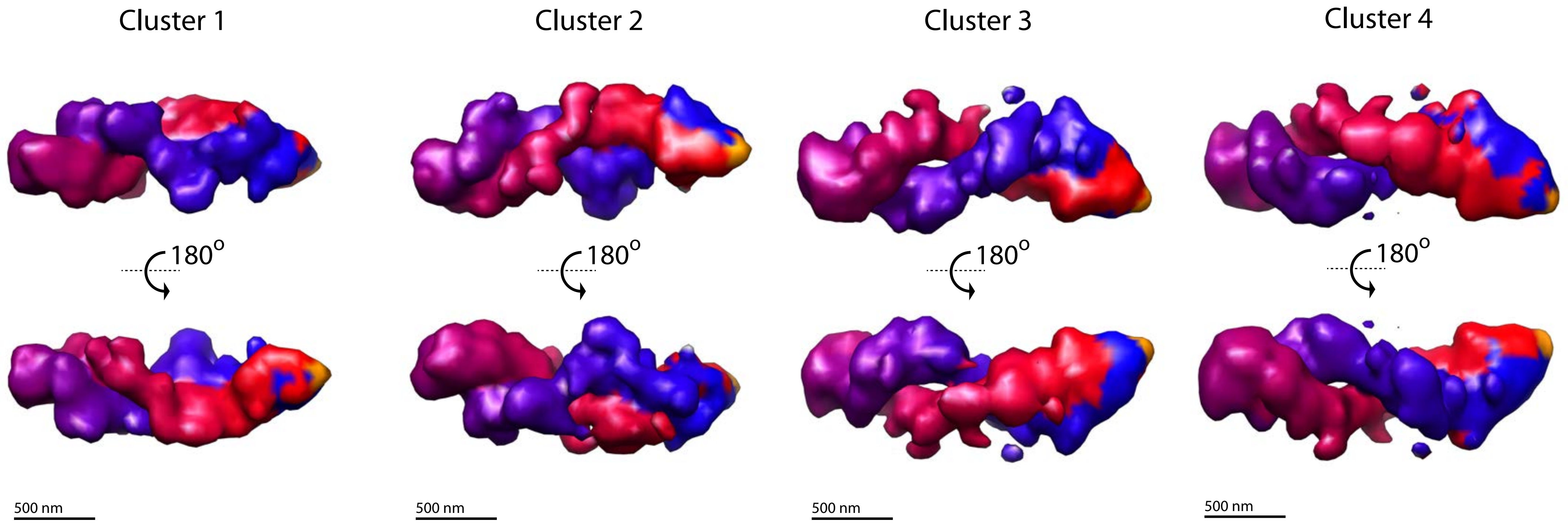


Resolution

dif site 47±17Kb from Ter

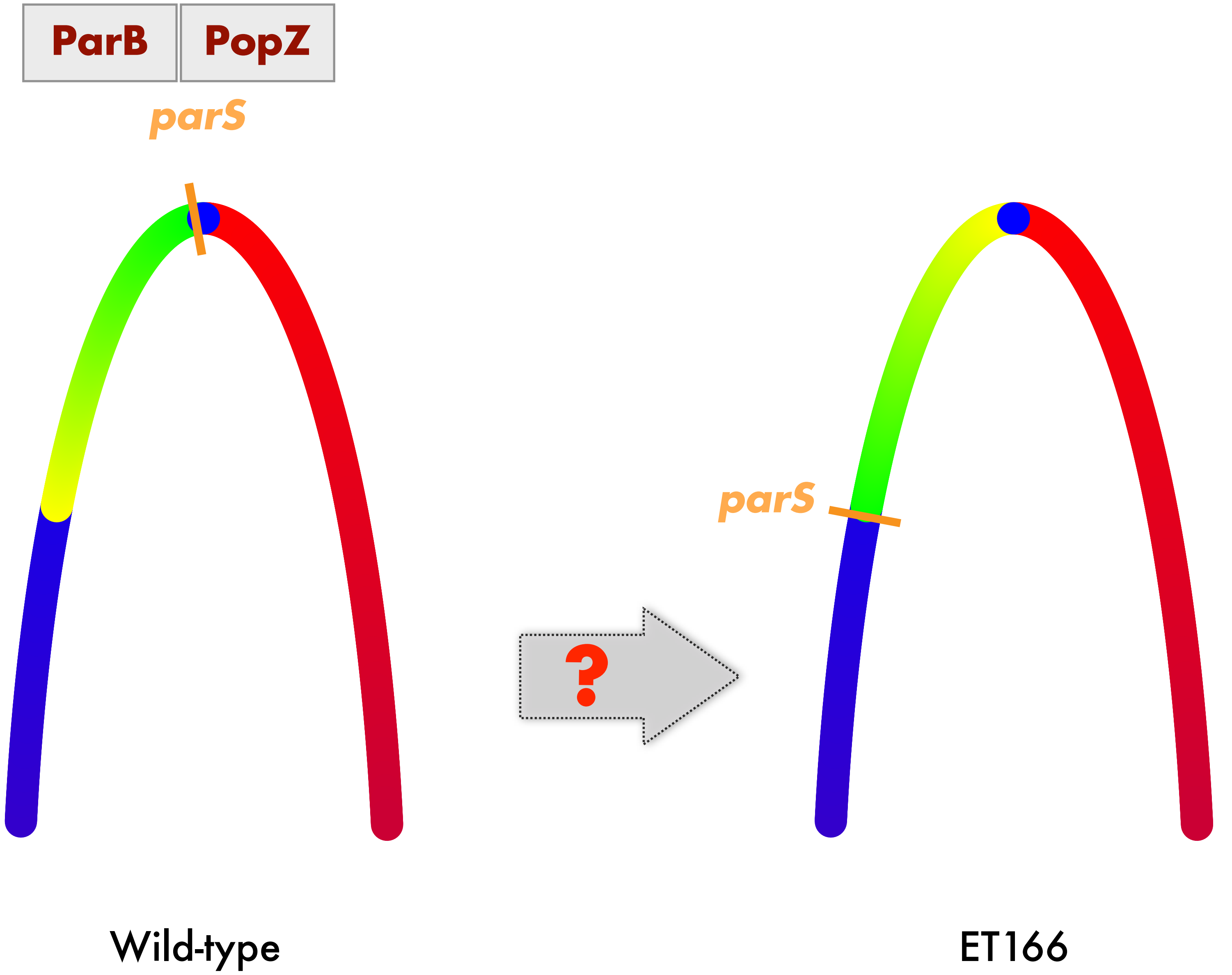
Centromer-like

parS sites 25±17Kb from Ori

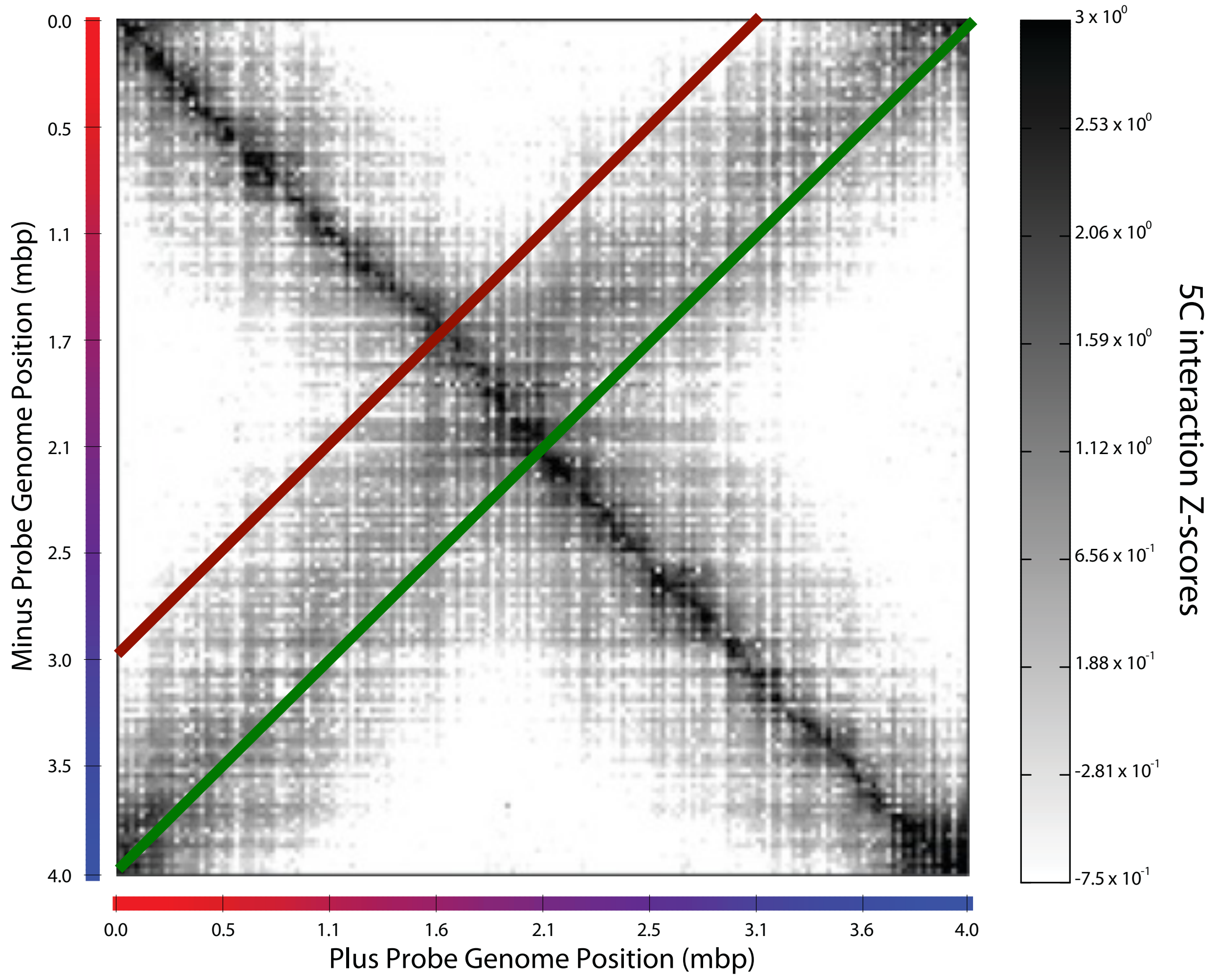


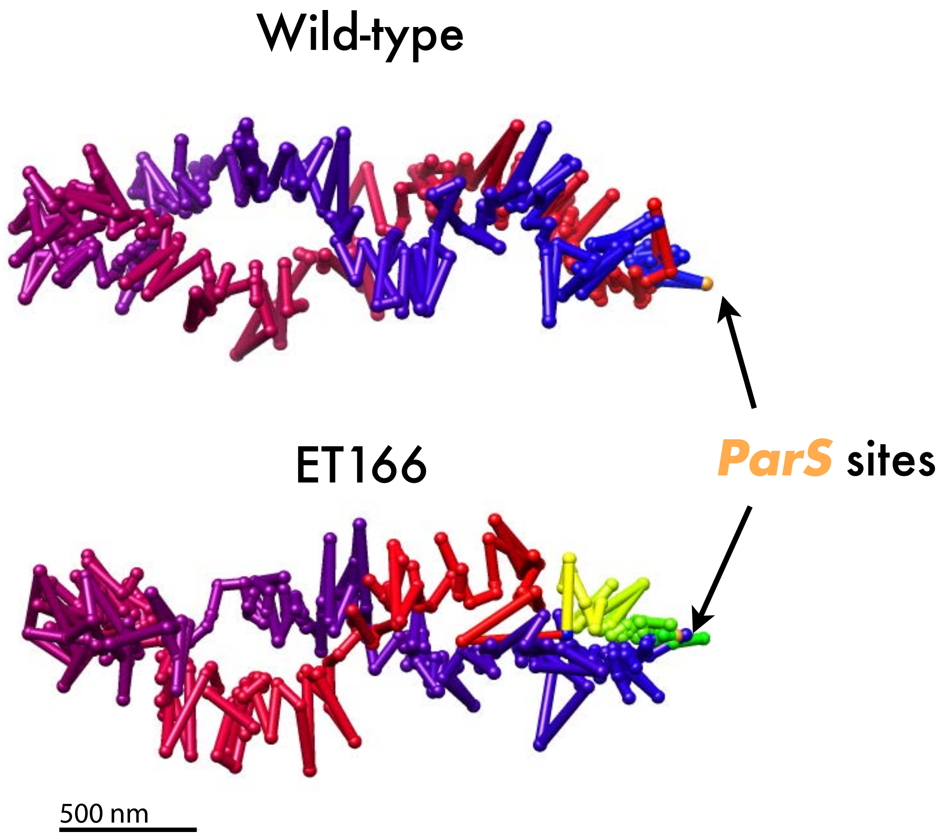
MIRRORS!

Moving the parS sites 400 Kb away from Ori



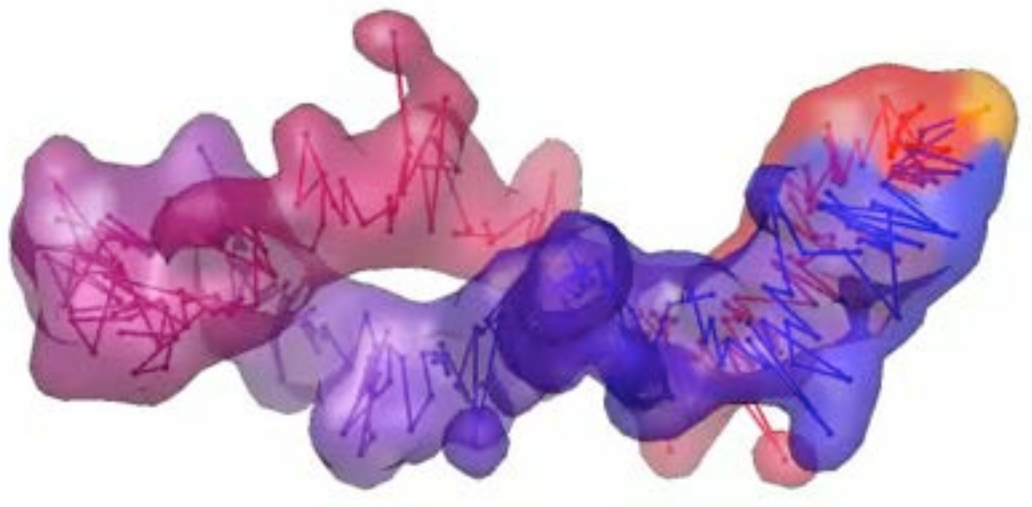
Moving the parS sites results in whole genome rotation!

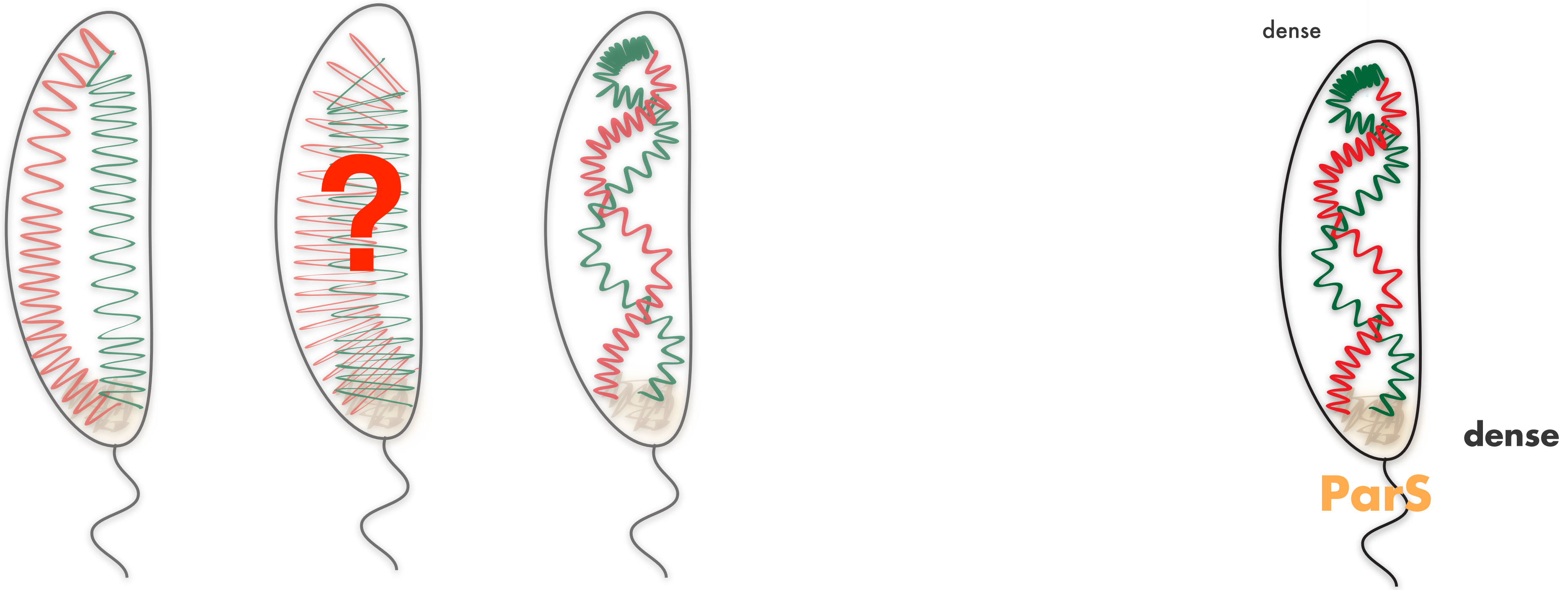




Arms are **STILL** helical

Genome architecture in Caulobacter



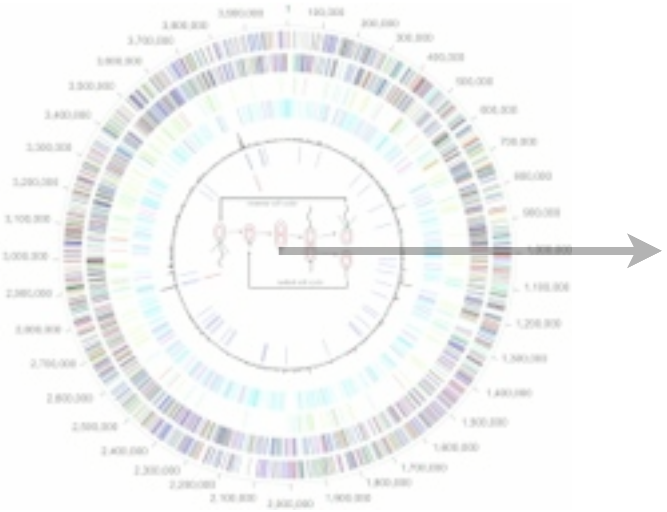


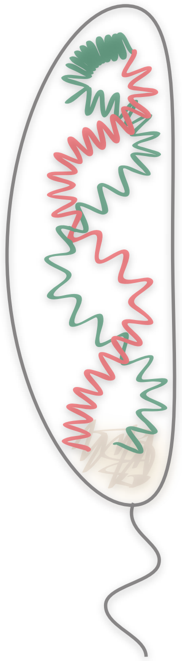
M.A. Umbarger, *et al.* **Molecular Cell** (2011) **44**:252–264

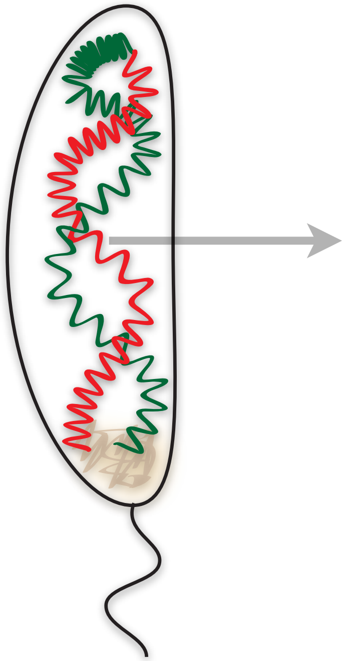
From Sequence to Function

5C + IMP

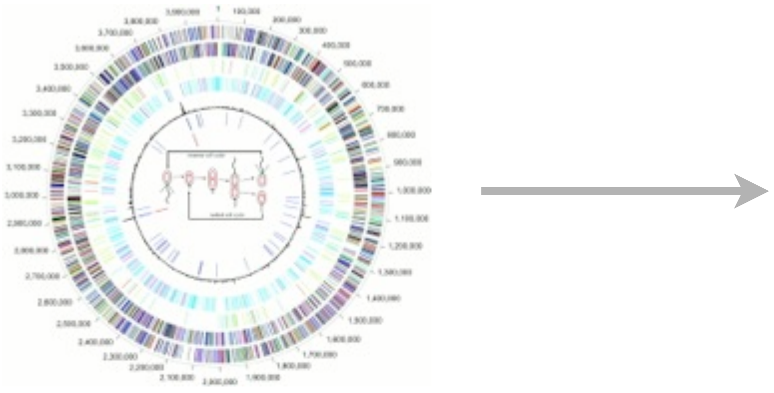
Technology







Hypothesis

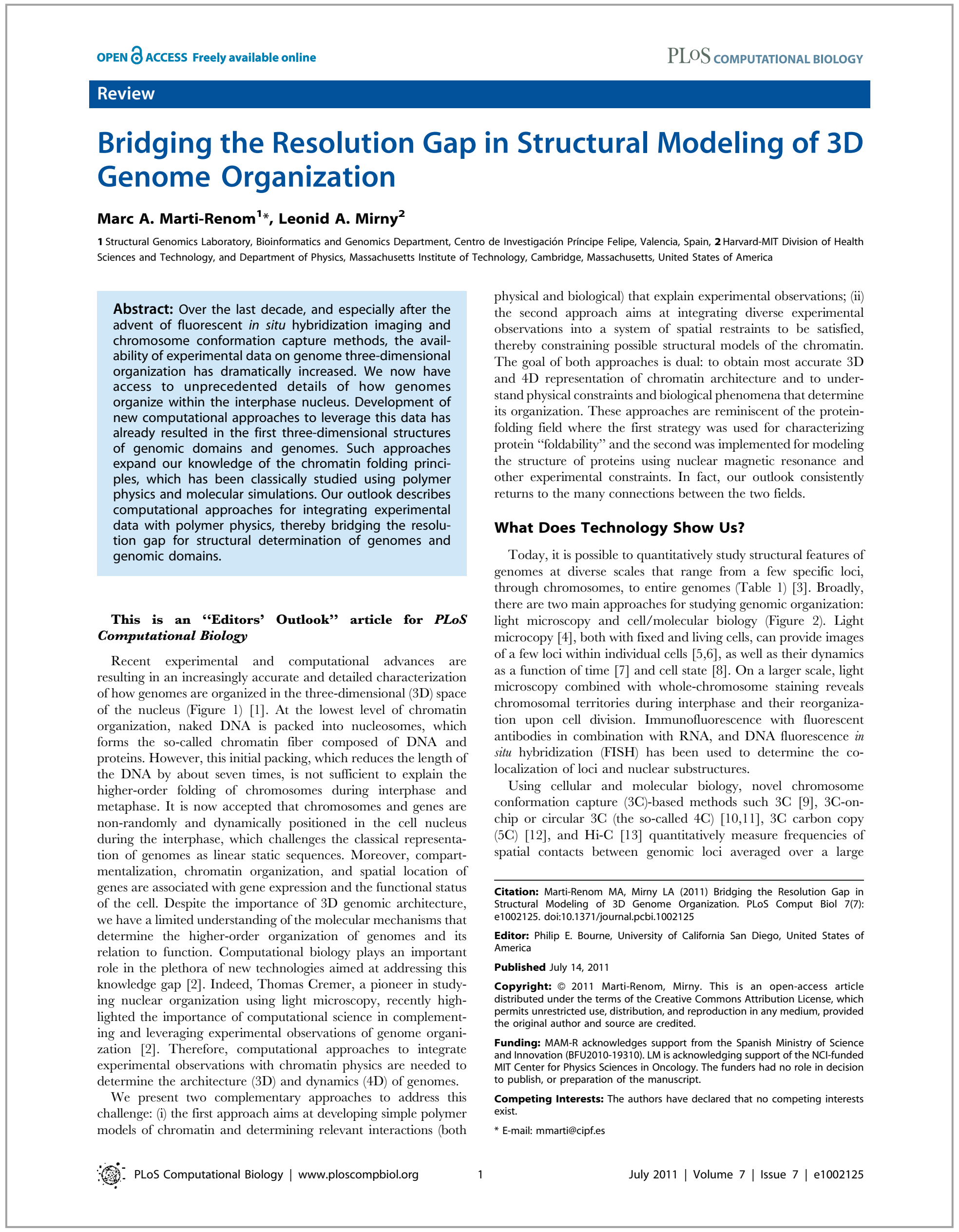


Function!

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

PLoS CB Outlook

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



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

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

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

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

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

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

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

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

What Does Technology Show Us?

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

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

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

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

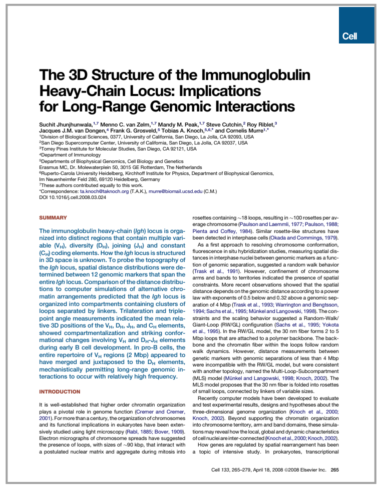
Published July 14, 2011

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

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

* E-mail: mmarti@cipf.es

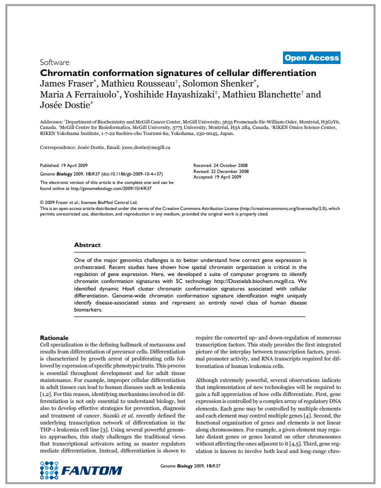
MURRE
Cell (2008) **133**:265-79



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

Suchit Jhunjhunwala, Menno C. van Zelm, Mandy M. Peak, Steve Cutchin, Roy Riblet, Jacques J.M. van Dongen, Frank G. Grosveld, Tobias A. Knoch and Cornelis Murre

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

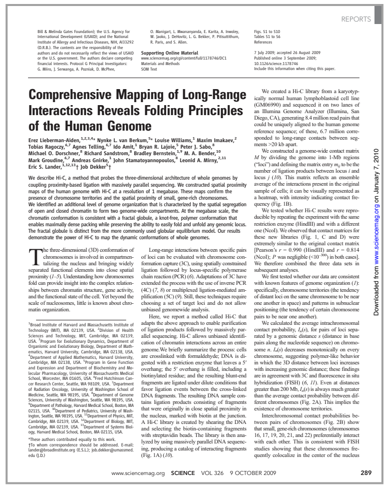


Software

Chromatin conformation signatures of cellular differentiation

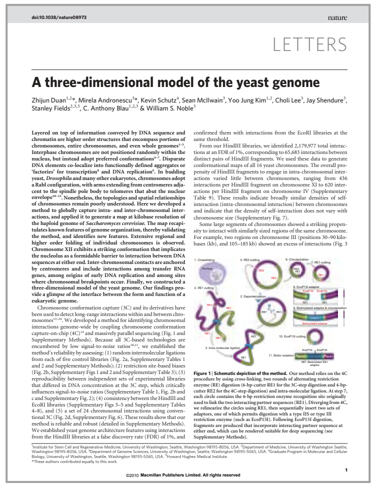
James Fraser, Mathieu Rousseau, Solomon Shenker, Maria A Ferraiuolo, Yoshihide Hayashizaki, Mathieu Blanchette and Josée Dostie

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



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

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



LETTERS

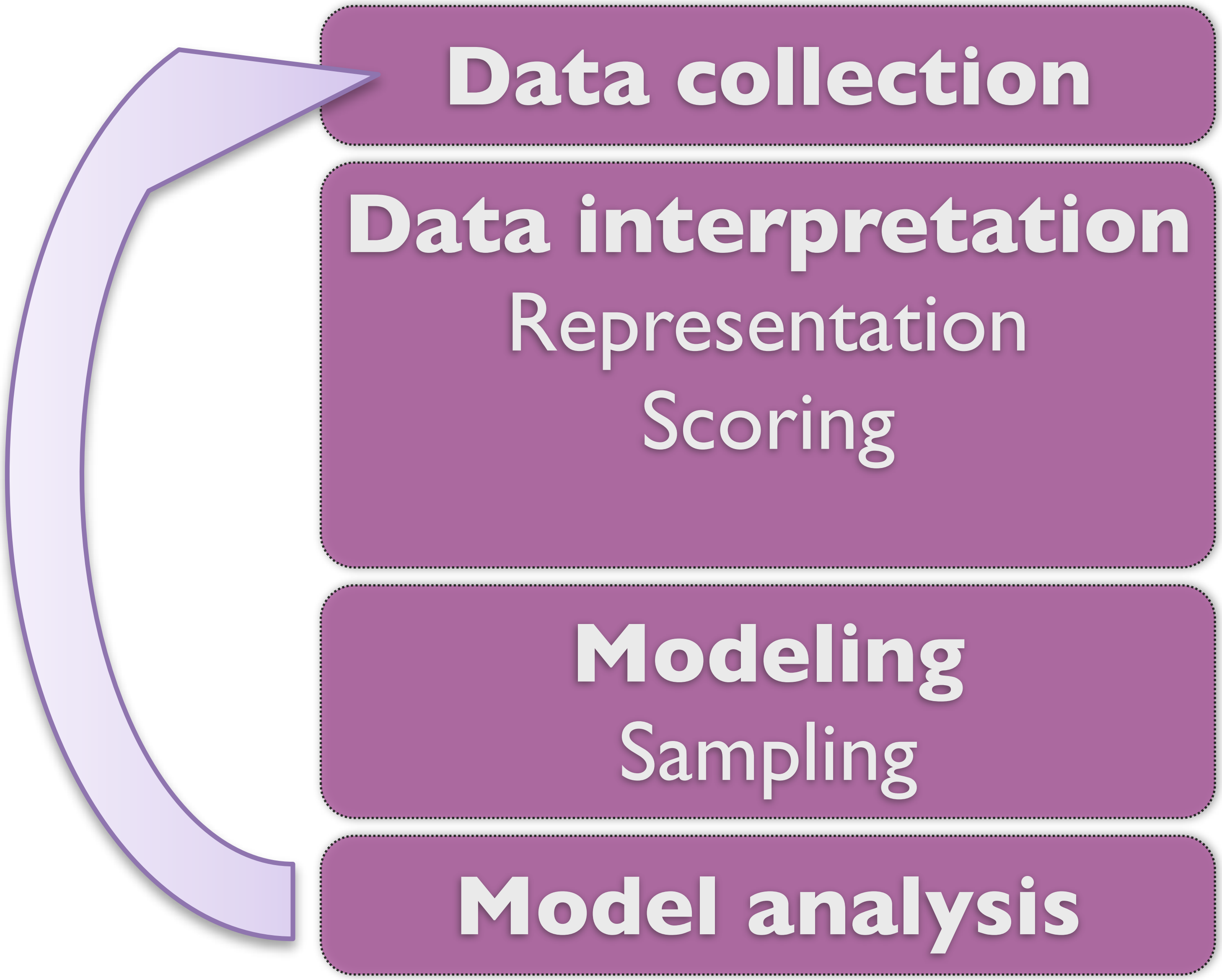
A three-dimensional model of the yeast genome

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

TECHNICAL REPORTS

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

Take home message



Data collection

Data interpretation

Representation

Scoring

Modeling

Sampling

Model analysis