FEBS Letters 589 (2015) 2987–2995

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

# Restraint-based three-dimensional modeling of genomes and genomic domains

François Serra [a,b], Marco Di Stefano [a,b], Yannick G. Spill [a,b], Yasmina Cuartero [a,b], Michael Goodstadt [a,b], Davide Baù [a,b], Marc A. Marti-Renom [a,b,c,*]

[a] Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain
[b] Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain
[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ARTICLE INFO

*Article history:*
Received 7 April 2015
Revised 5 May 2015
Accepted 5 May 2015
Available online 14 May 2015

Edited by Wilhelm Just

*Keywords:*
Genome architecture
3D genome reconstruction
Chromosome Conformation Capture
Restraint-based modeling

ABSTRACT

**Chromosomes are large polymer molecules composed of nucleotides. In some species, such as humans, this polymer can sum up to meters long and still be properly folded within the nuclear space of few microns in size. The exact mechanisms of how the meters long DNA is folded into the nucleus, as well as how the regulatory machinery can access it, is to a large extend still a mystery. However, and thanks to newly developed molecular, genomic and computational approaches based on the Chromosome Conformation Capture (3C) technology, we are now obtaining insight on how genomes are spatially organized. Here we review a new family of computational approaches that aim at using 3C-based data to obtain spatial restraints for modeling genomes and genomic domains.**

## 1. Introduction

Genomes are often compared to libraries where the genetic information is stored in the form of books and text represents the linear sequence of the genome. Unfortunately, that linear (1D) representation omits the utterly complex three-dimensional (3D) organization of the genome. Indeed, the physical support of the genome (i.e., the books, the shelves, the corridors and the library building in our metaphor) may be as important as the functional elements it encodes [1]. It is now known that the dynamic structure of the complex gene networks in a genome regulates the orchestration of fundamental biological processes such as development [2], cell differentiation [3,4] or response to stimuli [5], among others. Moreover, most of such complex mechanisms are also among the most conserved features of our genomes [6,7]. Therefore, addressing the 3D structure of a genome may provide insights into fundamental questions like the *C*-value paradox [8] or the regulatory divergence between closely related species [9].

In the past decade, with the introduction and development of Chromosome Conformation Capture (3C) technologies (e.g., 3C [10], 4C [11], 5C [12], Hi-C [13], *in situ*-Hi-C [7], TCC [14], T2C [15] or Capture-C [16,17], which are here referred as 3C-based technologies), it has been possible to get insight into how the genome folds by interrogating physical interactions within the genome. Importantly, the combination of these 3C-based technologies with advanced imaging [18] has helped reducing the resolution gap in genome structure [19]. It is now known that the genome organizes into chromosome territories [20], which in turn are spaced into two compartment types [13] composed of finer units called Topologically Associating Domains or TADs [6,21,22]. Alongside these advances, the evidence that genome structure is tightly associated with its function was being reinforced by the comparison with chromatin epigenetic states [7,23]. However, two limitations are blurring the full picture of the genome organization. First, some of the emerging genomic features change depending on the scale at which we study the genome. For example, TADs are structural units that were shown to be robustly detectable over a large range of genomic resolutions. Yet, their existence is challenged when the genome is interrogated at finer scales [7]. Second, 3C-based experiments are usually carried out with tens of millions of cells, and thus are population-based measures superimposing millions of partial

[*] Corresponding author at: Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain.
*E-mail address:* mmarti@pcb.ub.cat (M.A. Marti-Renom).

http://dx.doi.org/10.1016/j.febslet.2015.05.012
0014-5793/© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

**Table 1**
Summary of different modeling strategies. $F_{ij}$ is the observed interaction frequency between two particles $i$ and $j$, $D_{ij}$ is the target distance usually inferred from $F_{ij}$ and $r_{ij}$ is the distance computed on the models. $N$ is the total number of particles.

| Method *available online | Representation | Scoring | | | | Sampling | Models |
|---|---|---|---|---|---|---|---|
| | | $U_{3C}$ | | $U_{Biol}$ | $U_{Phys}$ | | |
| | | $F_{ij} \rightarrow D_{ij}$ conversion | Functional form | | | | |
| ChromSDE* [37] | Points | $D_{ij} = \begin{cases} (\frac{1}{F_{ij}})^{\alpha} \text{ if} F_{ij} > 0 \\ \infty \text{ if } F_{ij} = 0 \end{cases}$ $\alpha$ is optimized | $\sum_{(i,j \mid D_{ij} < \infty)} \frac{(r_{ij}^2 - D_{ij}^2)}{D_{ij}} - \lambda \sum_{(i,j)} r_{ij}^2$ where $\lambda$ is set to 0.01 | N/A | N/A | Deterministic semidefinite programming to find the coordinates | Consensus |
| ShRec3D* [38] | Points | $D_{ij} = \begin{cases} \left(\frac{1}{F'_{ij}}\right)^{\alpha} & \text{if } F'_{ij} > 0 \\ \frac{N^2}{\sum_{(i,j)} F'_{ij}} & \text{if } F'_{ij} = 0 \end{cases}$ $F'_{ij}$ is the original $F_{ij}$ corrected to satisfy all triangular inequalities with the shortest path reconstruction | N/A | N/A | N/A | Deterministic transformations of $D_{ij}$ into coordinates | Consensus |
| TADbit* [43] | Spheres | $D_{ij} \propto \begin{cases} \alpha F_{ij} + \beta & \text{if } F_{ij} < \gamma' \text{ or } F_{ij} > \gamma \\ \frac{S_i + S_j}{2} & \text{if } \lvert i - j \rvert = 1 \end{cases}$ $\alpha$ and $\beta$ are estimated from the max and the min $F_{ij}$, from the optimized max distance and from the resolution. $\gamma' < \gamma$ are optimized too. $s_i$ is the radius of particle $i$ | $\sum_{(i,j)} k_{ij}(r_{ij} - D_{ij})^2$ where $k_{ij} = 5$ if $\lvert i - j \rvert = 1$ or proportional to $F_{ij}$ otherwise | Yes | $U_{excl}$ and $U_{bond}$ have harmonic forms | Monte Carlo (MC) sampling with Simulated annealing and Metropolis scheme | Resampling |
| BACH* [45] | Points | $D_{ij} \propto \frac{B_i B_j}{F_{ij}^{\alpha}}$. The biases $B_i$ and $B_j$ and $\alpha$ are optimized | $b_{ij} D_{ij}^{1/\alpha} + c_{ij} \log(D_{ij})$ where $b_{ij}$ and $c_{ij}$ are optimized parameters | No | No | Sequential importance and Gibbs sampling with hybrid MC and adaptive rejection | Population |
| Giorgetti et al. [40] | Spheres | Particles interact with pair-wise well potentials of depths $B_{ij}$ and contact radius $a$, which is larger than a hard-core radius and smaller than a maximum contact radius. The parameters are optimized over all the population of models | | No | N/A | MC sampling with metropolis scheme | Population |
| Duan et al. [41] | Spheres | $\overline{F_{\lvert i-j \rvert}} = \frac{\sum_{k=0}^{N - \lvert i-j \rvert} F_{(k,k+\lvert i-j \rvert)}}{N - \lvert i-j \rvert}$ is the average of $F_{ij}$ at genomic distance $\lvert i - j \rvert$ expressed in kb. $D_{ij} = \overline{F_{\lvert i-j \rvert}} \times 7.7 \times \lvert i - j \rvert$ assuming that $\alpha$ 1 kb maps onto 7.7 nm | $\sum_{(i,j)} (r_{ij} - D_{ij})^2$ | Yes | $U_{excl}$ and $U_{bond}$ have harmonic forms | Interior-point gradient-based method | Resampling |
| MCMC5C* [49] | Points | $D_{ij} \propto \frac{1}{F_{ij}^{\alpha}}$ where is optimized | $\sum_{(i,j)} (F_{ij} - r_{ij}^{-1/\alpha})^2$ | N/A | N/A | MC sampling with Markov chain based algorithm | Resampling |
| PASTIS* [47] | Points | $D_{ij} \propto \frac{1}{F_{ij}^{\alpha}}$ where $\alpha$ is optimized | $b_{ij} D_{ij}^{1/\alpha} + c_{ij} \log(D_{ij})$ where $b_{ij}$ and $c_{ij}$ are optimized parameters | No | No | Interior point and isotonic regression algorithms | Resampling |
| Meluzzi and Arya [48] | Spheres | $\sum_{(i,j))} k_{ij} r_{ij}^2$ where $k_{ij}$ are adjusted such that the contact probabilities computed on the models match the $F_{ij}$ | | No | $U_{excl}$ is a pure repulsive LJ potential. $U_{bond}$ and $U_{bend}$ have harmonic forms | Brownian dynamics | Resampling |
| AutoChrom3D* [44] | Points | $D_{ij} \propto \begin{cases} \alpha F_{ij} + \beta & \text{if } F_{\min} < F_{ij} < F_{\gamma} \\ \alpha' F_{ij} + \beta' & \text{if } F_{\gamma} < F_{ij} < F_{\max} \end{cases}$ where $F_{\min}$ ($F_{\max}$) are the min(max) of $F_{ij}$. The parameters $(\alpha, \beta)$, $(\alpha', \beta')$ and $F_{\gamma}$ are found using the nuclear size, the resolution and the decay of $F_{ij}$ with $\lvert i - j \rvert$ | $\sum_{(i,j)} \frac{(r_{ij} - D_{ij})^2}{D_{ij}^2}$ | Yes | N/A | Non-linear constrained | Consensus |
| Kalhor et al. [14] | Spheres | $D_{ij} = R_{\text{contact}}$ to enforce the pair contact, if the normalized contact frequency $F_{ij}$ is higher than 0.25. Otherwise the contact is not enforced | $\sum_{\text{models}} \sum_{(i,j)} k_{ij}(r_{ij} - D_{ij})^2$ where $k_{ij}$ is different for pairs of particles, on different chromosomes, on the same chromosome, or connected | Yes | $U_{excl}$ and $U_{bond}$ have harmonic forms | Conjugate gradients sampling with Simulated annealing scheme | Population |

[*] These methods are publicly available.

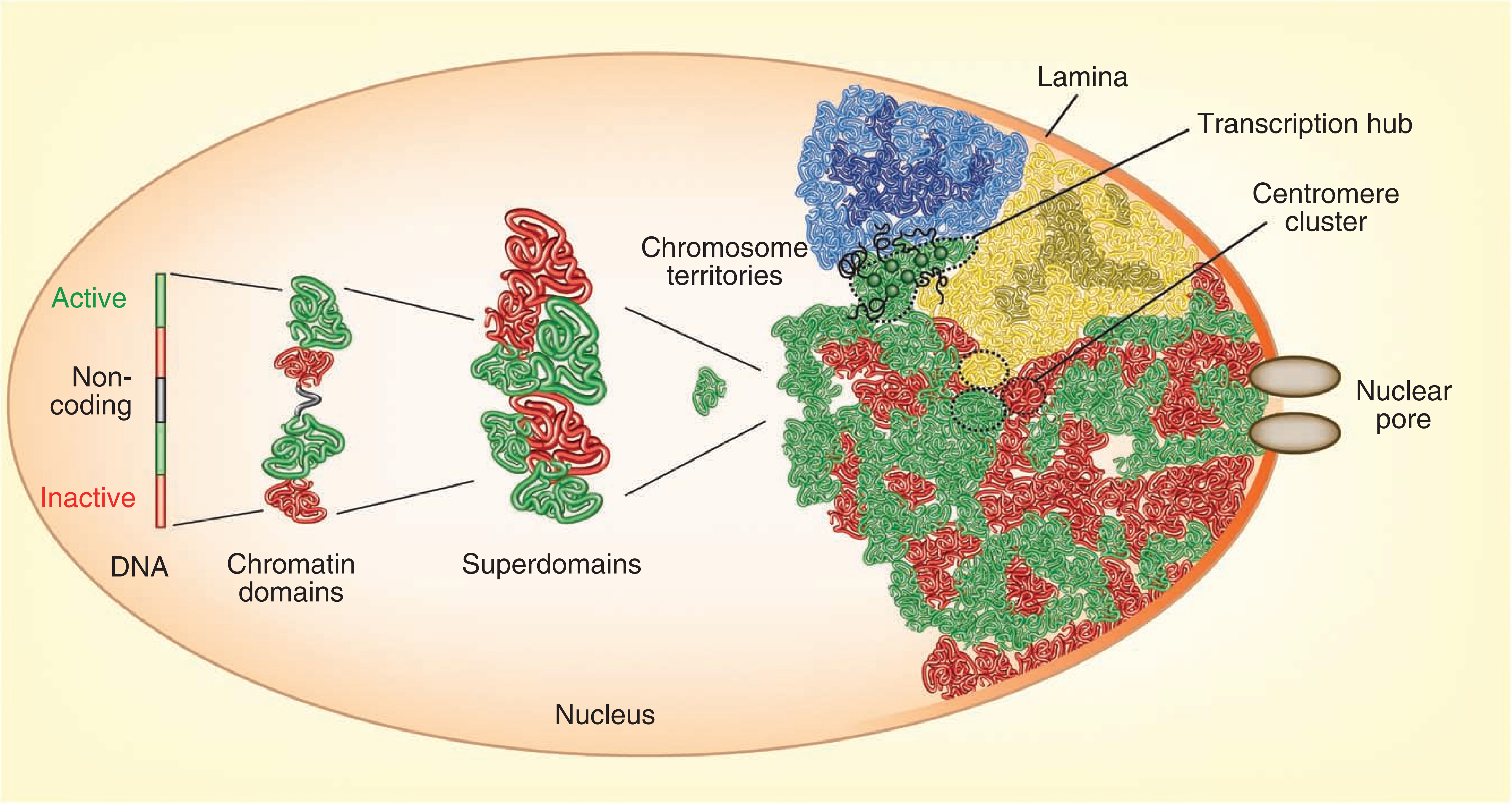
# 3DAROC16

# Summary day #2

David Castillo, François Serra &
Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

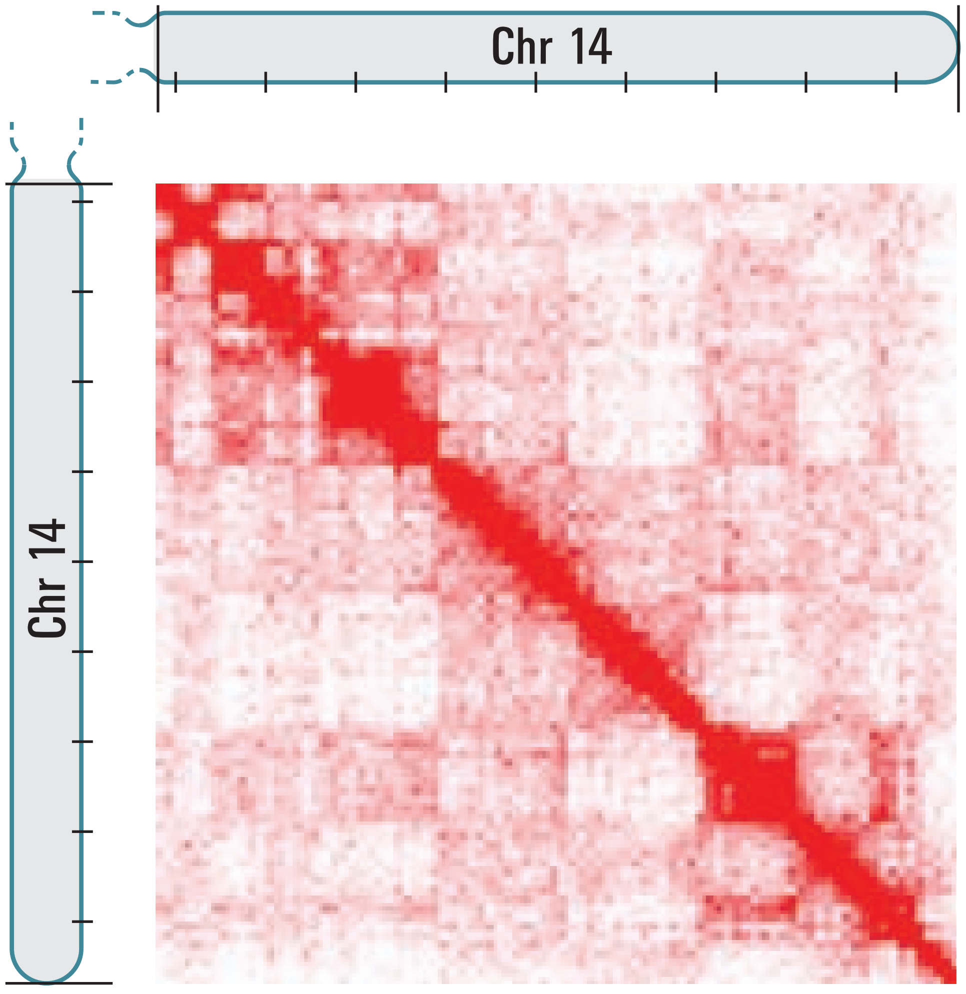
# Complex genome organization

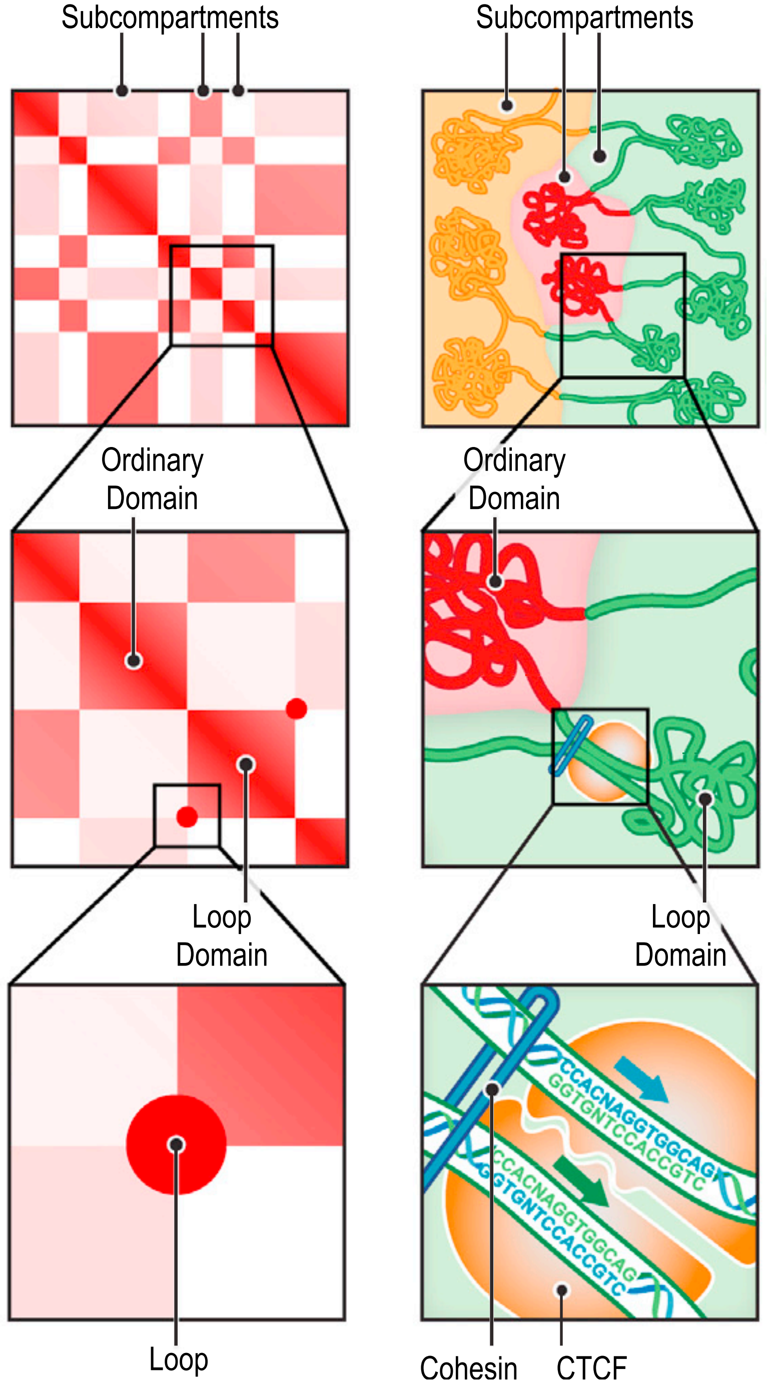
Cavalli, G. & Misteli, T. Functional implications of genome topology. Nat Struct Mol Biol 20, 290–299 (2013).



Marina Corral

# Hierarchical genome organisation

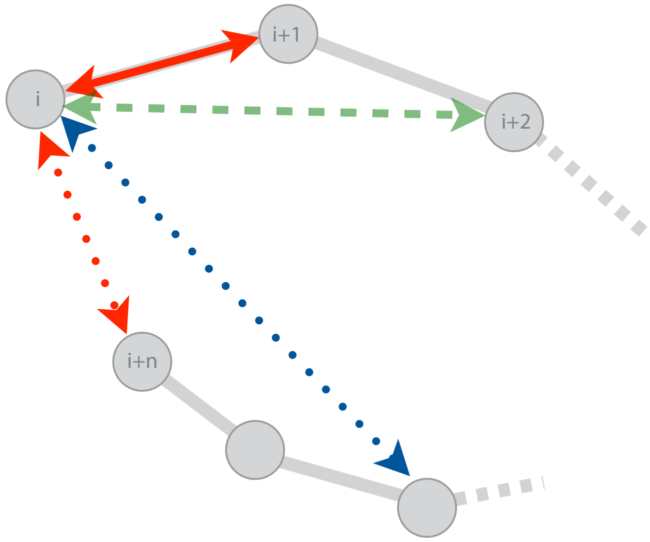


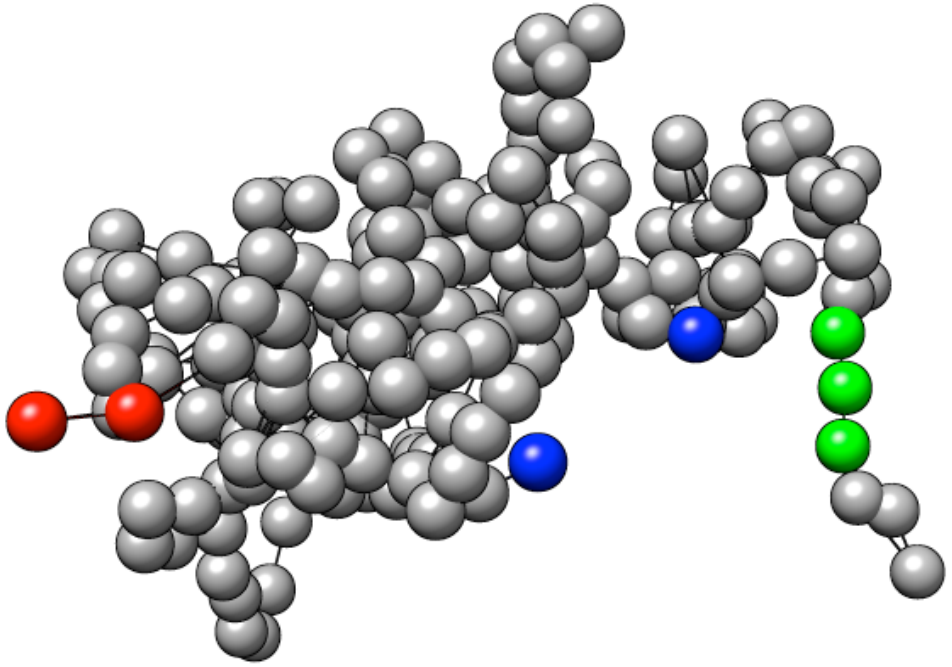


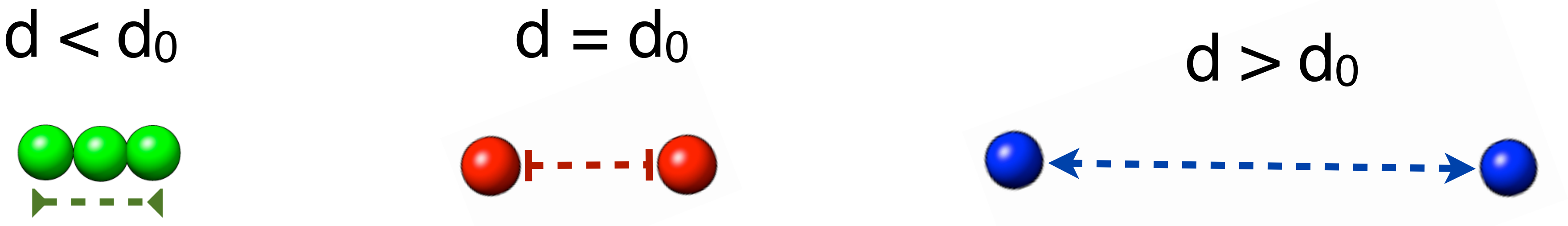
Lieberman-Aiden, E., et al. (2009). Science, 326(5950), 289–293.
Rao, S. S. P., et al. (2014). Cell, 1–29.

# Model representation and scoring

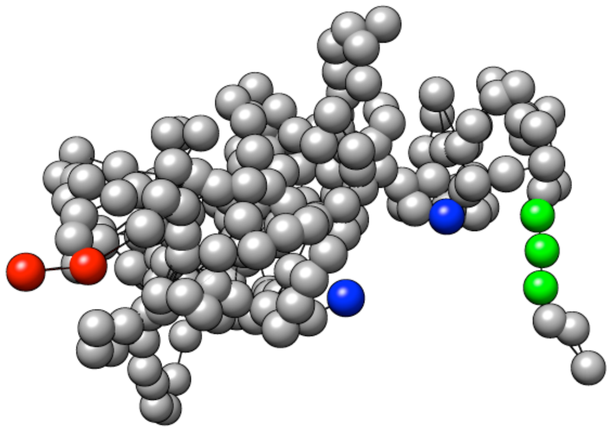
Constituent parts of the molecule



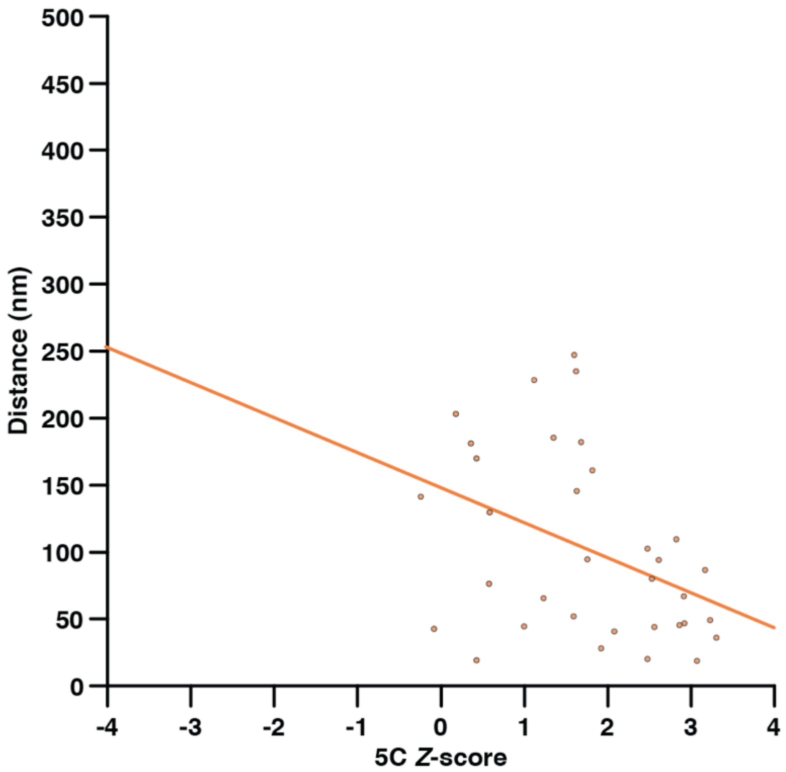




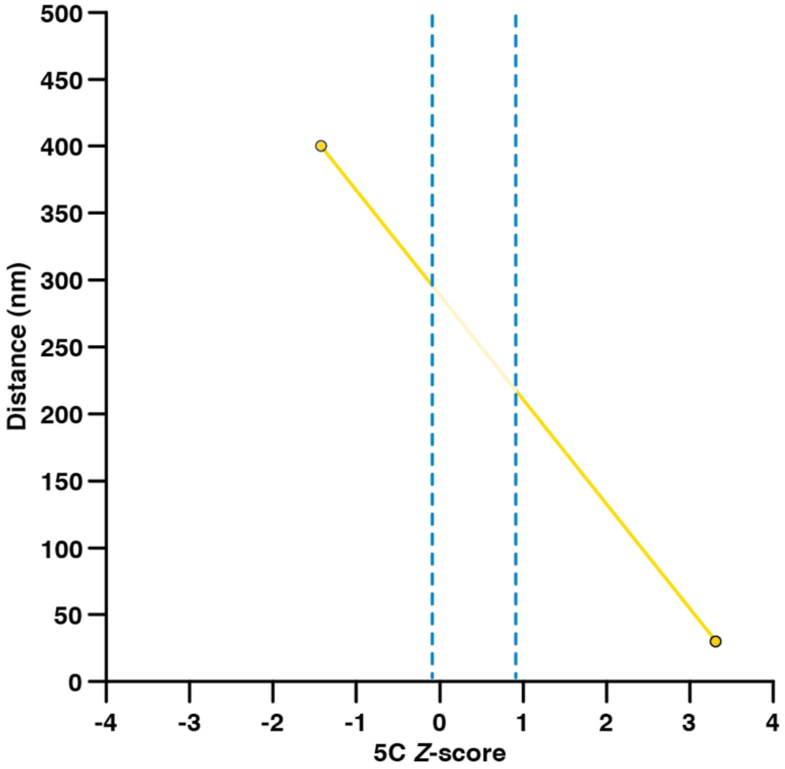
# From 5C data to spatial distances



Neighbor fragments

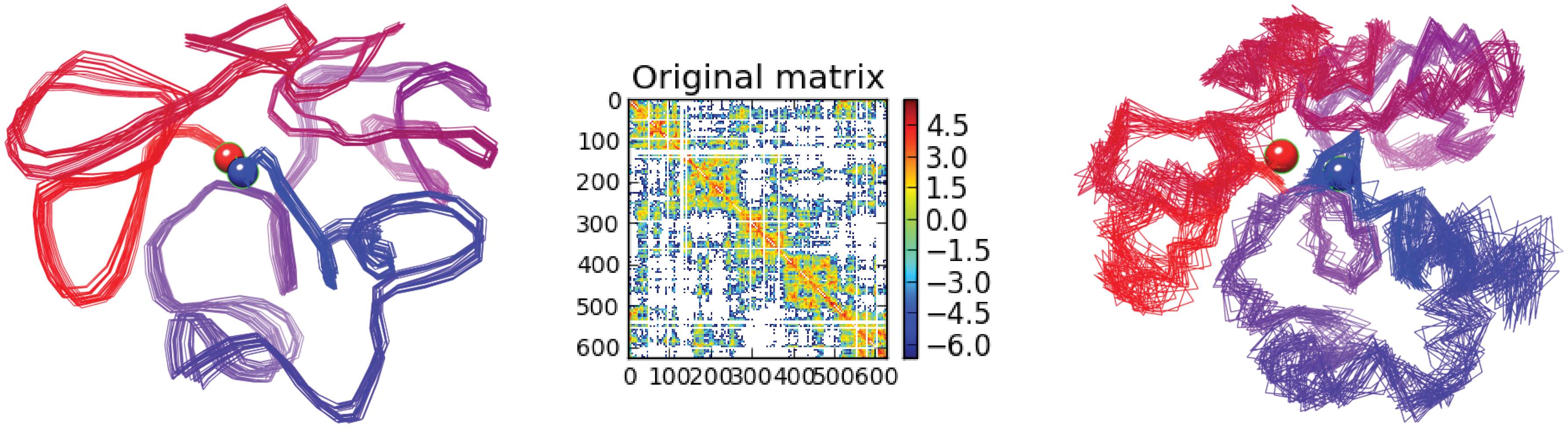


Non-Neighbor fragments



cnag CRG Centre for Genomic Regulation

# Model quality



chr40_TAD

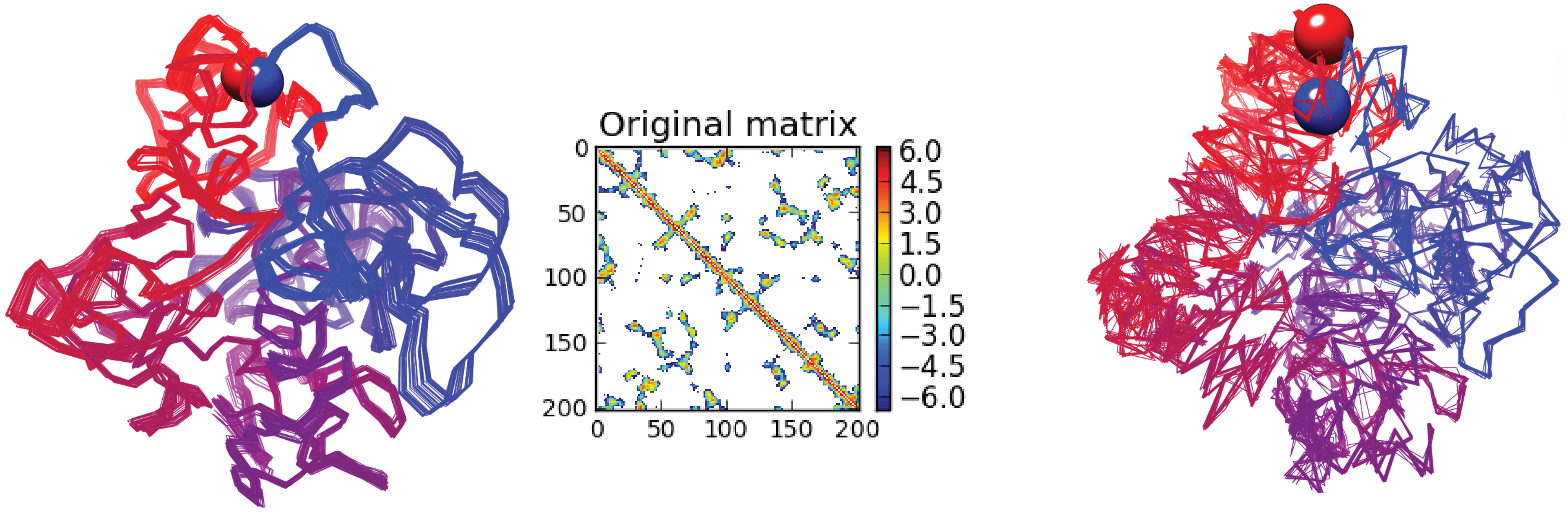
α=100

Δts=10

TADbit-SCC: 0.91

<dRMSD>: 32.7 nm

<dSCC>: 0.94



chr150_TAD

α=50

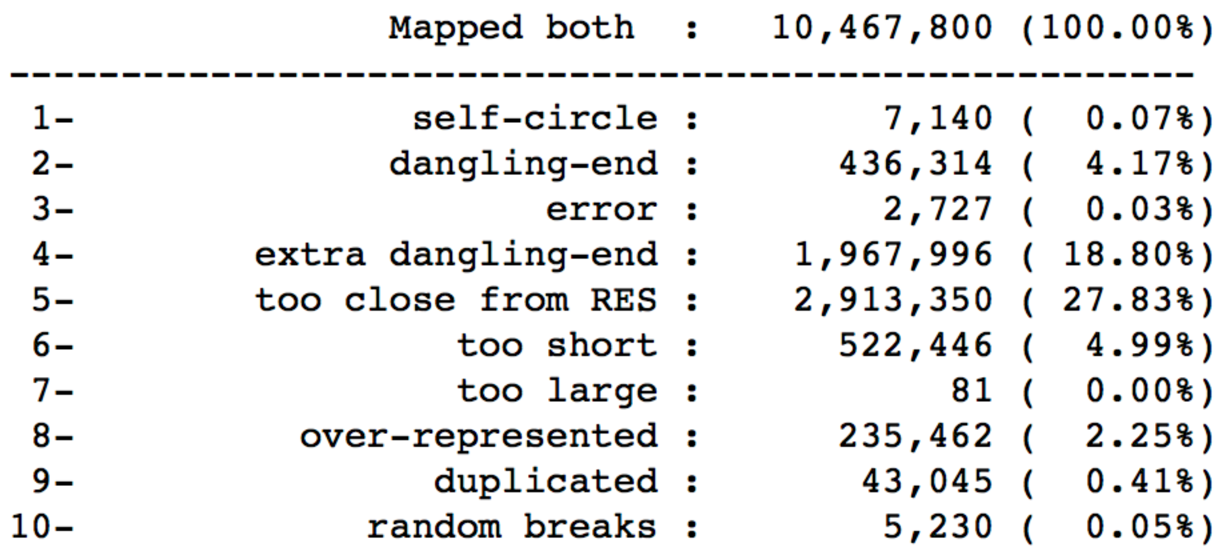
Δts=1

TADbit-SCC: 0.82

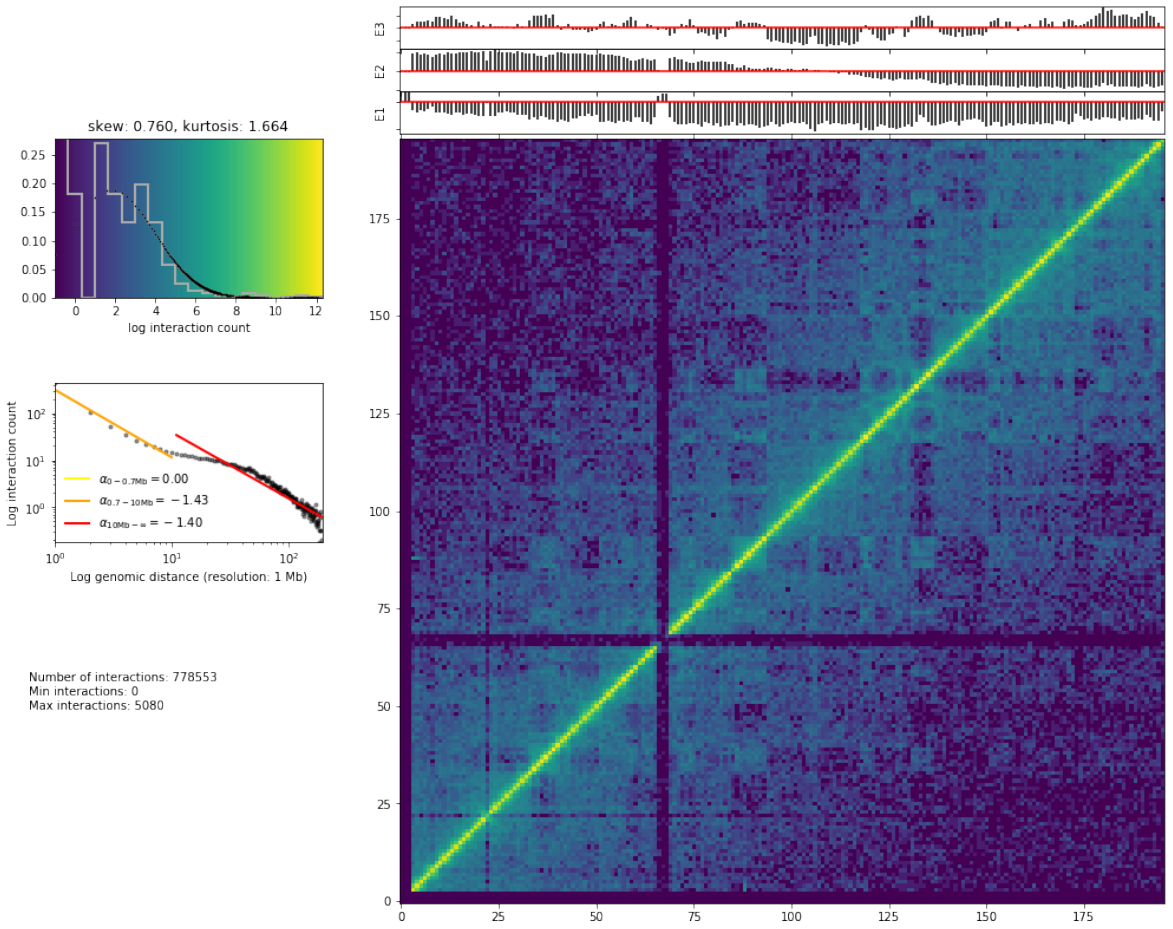
<dRMSD>: 45.4 nm

<dSCC>: 0.86

# Hi-C map generation and filtering



```
                       Mapped both  :   10,467,800 (100.00%)
-----------------------------------------------------
 1-              self-circle :        7,140 (   0.07%)
 2-             dangling-end :      436,314 (   4.17%)
 3-                    error :        2,727 (   0.03%)
 4-       extra dangling-end :    1,967,996 (  18.80%)
 5-       too close from RES :    2,913,350 (  27.83%)
 6-                too short :      522,446 (   4.99%)
 7-                too large :           81 (   0.00%)
 8-          over-represented :      235,462 (   2.25%)
 9-               duplicated :       43,045 (   0.41%)
10-            random breaks :        5,230 (   0.05%)
```



Number of interactions: 778553
Min interactions: 0
Max interactions: 5080

# How confortable are you with…

- what 3C-based methods have told us about the genome?
- the three levels of organization (A/B, TAD, Loops)?
- modeling 3D genomes (XYZ coordinates)?
- the limits of 3D modeling (MMP Score)?
- TADbit filtering/normalization?