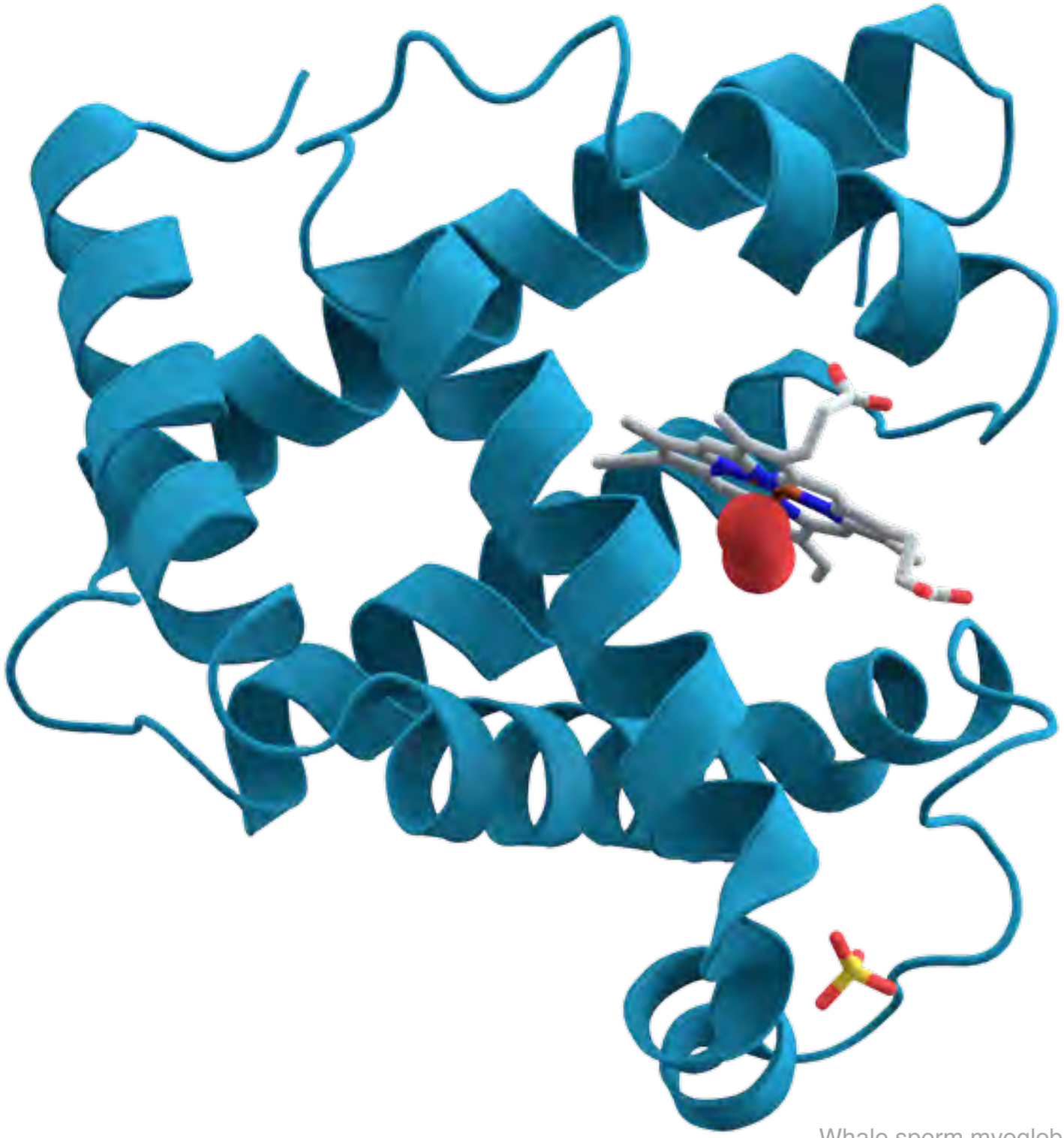
Determining the 3D structure of genomes and genomic domains.

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

*iCrea
INSTITUCIÓ CATALANA DE RECERCA I ESTUDIS AVANÇATS

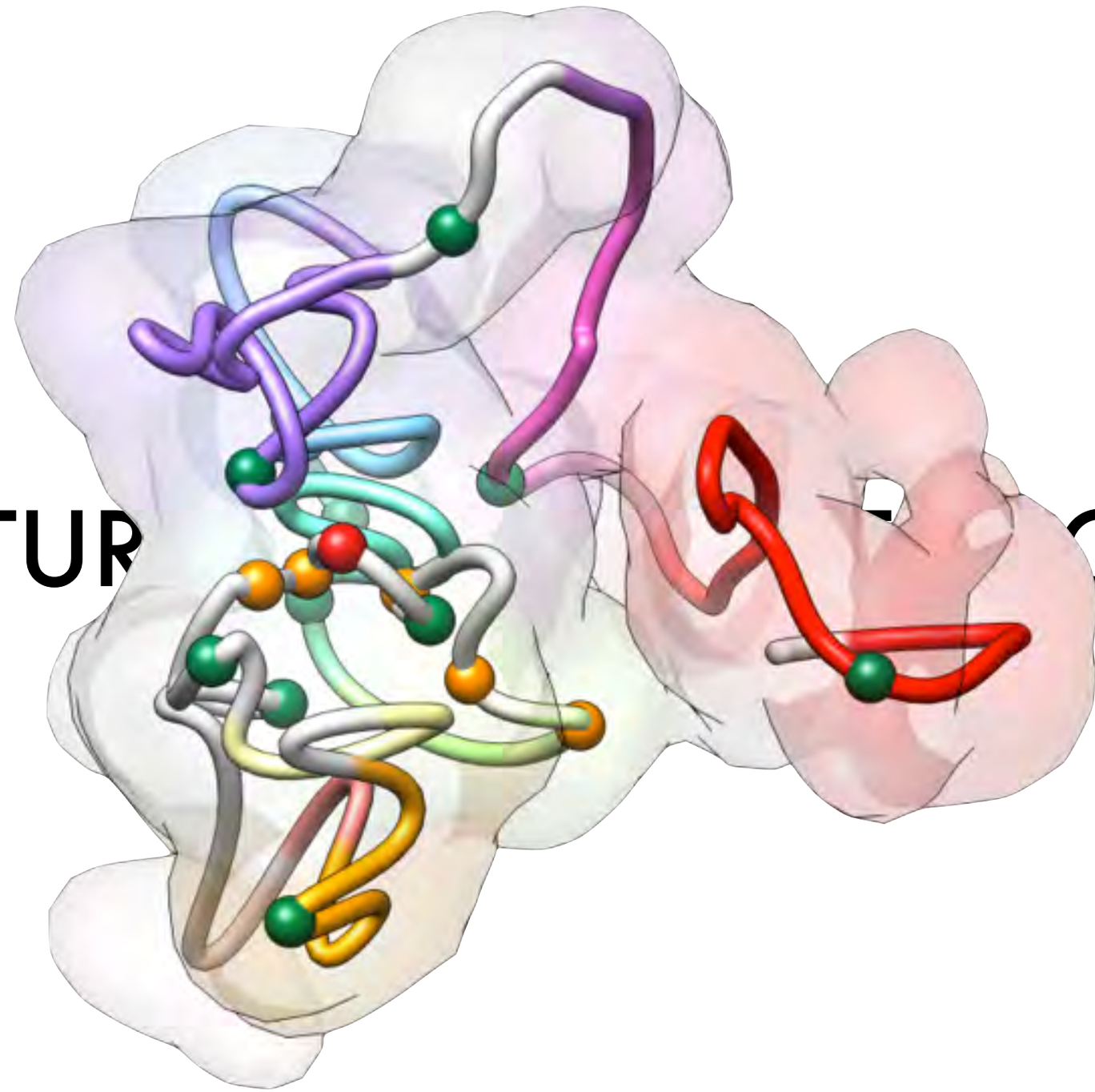
cnag

CRG
Centre for Genomic Regulation



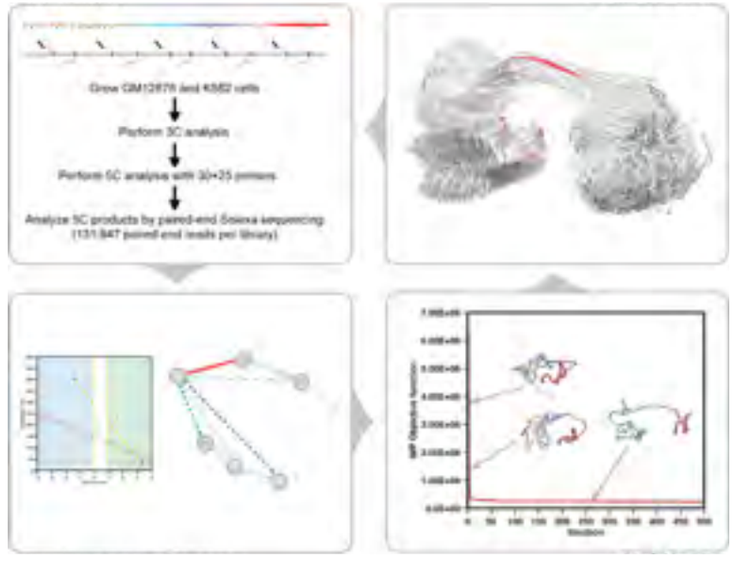
Whale sperm myoglobin structure (1960)

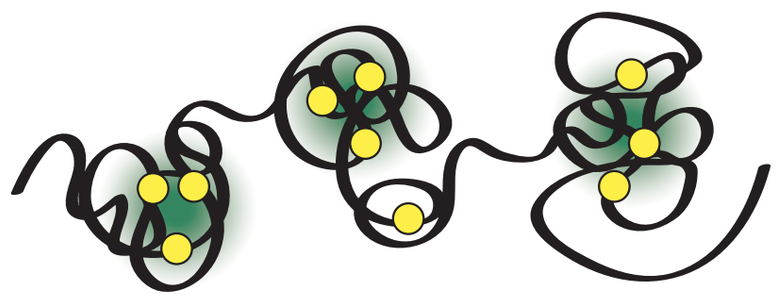
STRUCTUR CTION



alpha-globin genomic domain structure (2011)

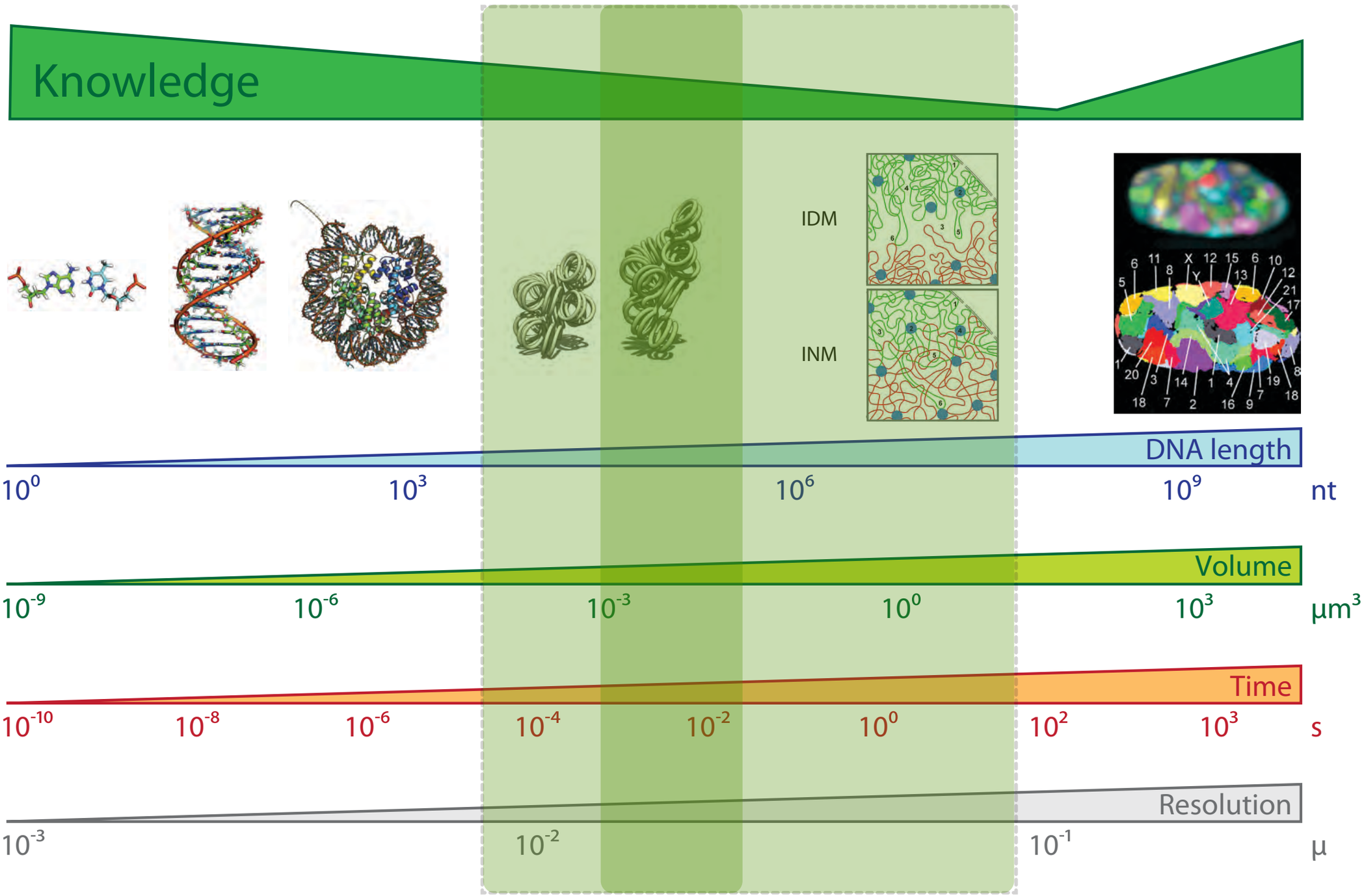
TADbit





Resolution Gap

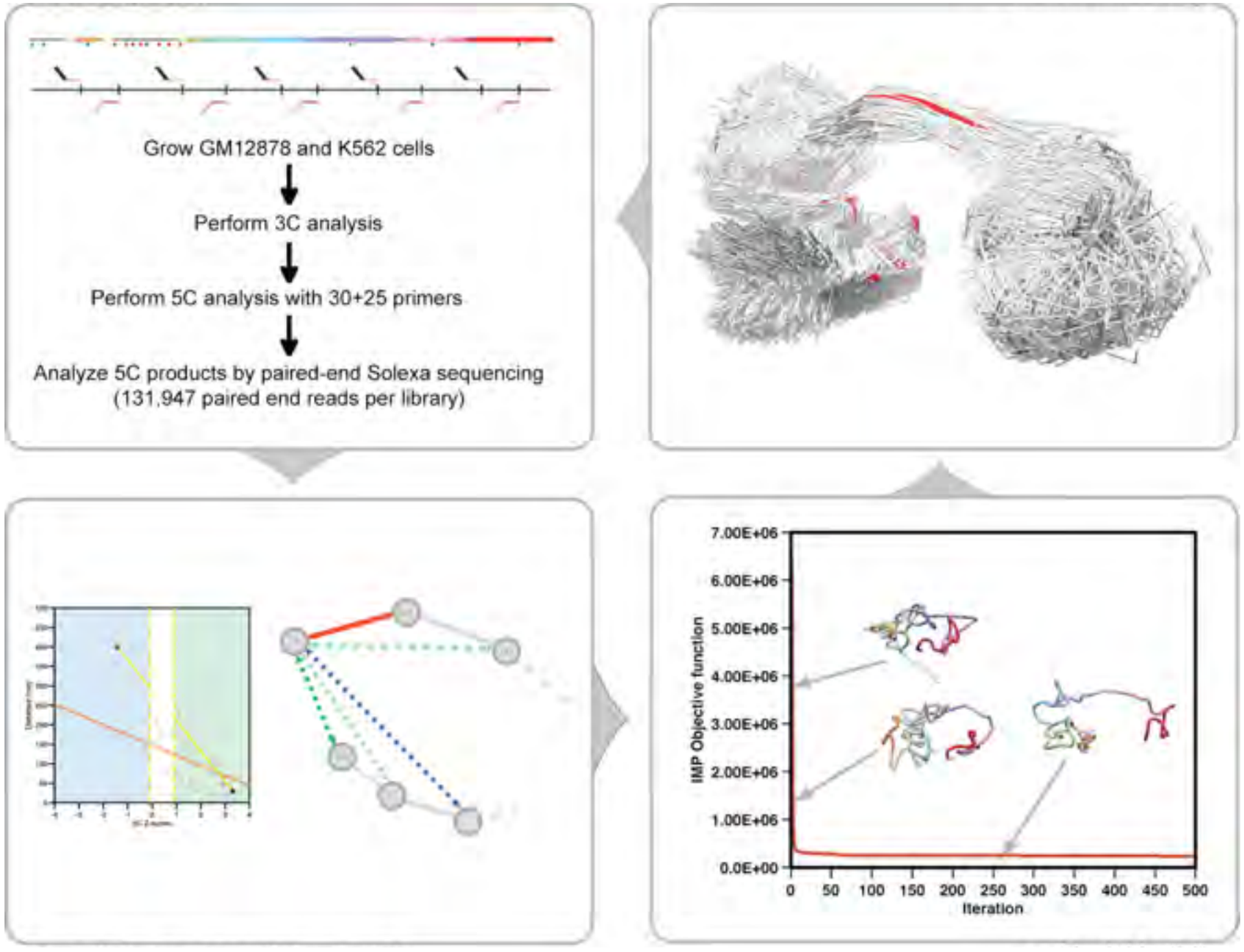
Marti-Renom, M. A. & Mirny, L. A. PLoS Comput Biol 7, e1002125 (2011)



Hybrid Method

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

Experiments

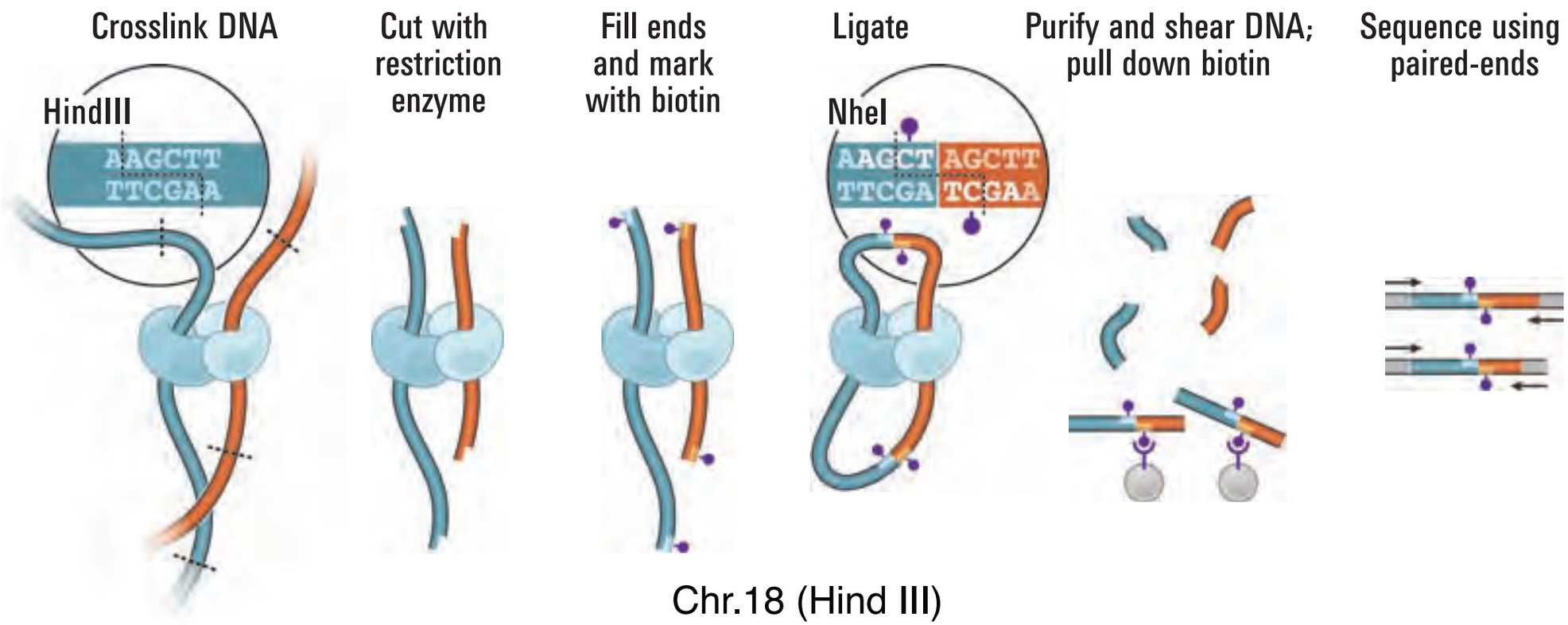


Computation

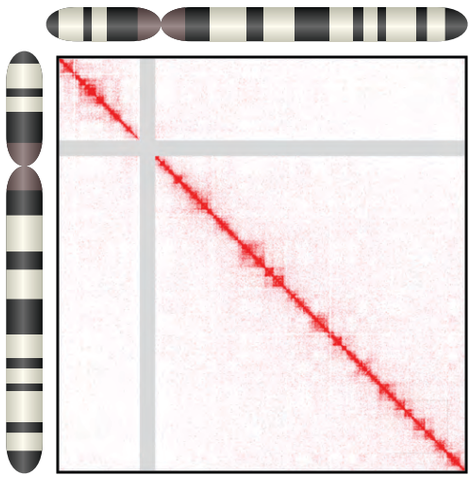
Hi-C technology

Lieberman-Aiden, E. et al. Science 326, 289–293 (2009).

http://3dg.umassmed.edu

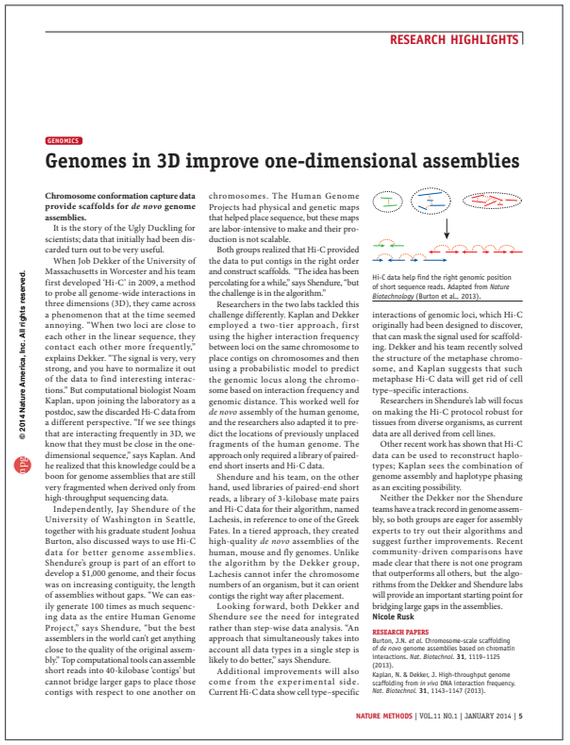


Chr.18 (Hind III)



Hi-C technology

BTW, de novo assembly!



RESEARCH HIGHLIGHTS

GENOMICS

Genomes in 3D improve one-dimensional assemblies

Chromosome conformation capture data provide scaffolds for *de novo* genome assemblies.

It is the story of the Ugly Duckling for scientists; data that initially had been discarded turn out to be very useful.

When Job Dekker of the University of Massachusetts in Worcester and his team first developed 'Hi-C' in 2009, a method to probe all genome-wide interactions in three dimensions (3D), they came across a phenomenon that at the time seemed annoying. "When two loci are close to each other in the linear sequence, they contact each other more frequently," explains Dekker. "The signal is very, very strong, and you have to normalize it out of the data to find interesting interactions." But computational biologist Noam Kaplan, upon joining the laboratory as a postdoc, saw the discarded Hi-C data from a different perspective. "If we see things that are interacting frequently in 3D, we know that they must be close in the one-dimensional sequence," says Kaplan. And he realized that this knowledge could be a boon for genome assemblies that are still very fragmented when derived only from high-throughput sequencing data.

Independently, Jay Shendure of the University of Washington in Seattle, together with his graduate student Joshua Burton, also discussed ways to use Hi-C data for better genome assemblies. Shendure's group is part of an effort to develop a $1,000 genome, and their focus was on increasing contiguity, the length of assemblies without gaps. "We can easily generate 100 times as much sequencing data as the entire Human Genome Project," says Shendure, "but the best assemblers in the world can't get anything close to the quality of the original assembly." Top computational tools can assemble short reads into 40-kilobase 'contigs' but cannot bridge larger gaps to place those contigs with respect to one another on chromosomes. The Human Genome Projects had physical and genetic maps that helped place sequence, but these maps are labor-intensive to make and their production is not scalable.

Both groups realized that Hi-C provided the data to put contigs in the right order and construct scaffolds. "The idea has been percolating for a while," says Shendure, "but the challenge is in the algorithm."

Researchers in the two labs tackled this challenge differently. Kaplan and Dekker employed a two-tier approach, first using the higher interaction frequency between loci on the same chromosome to place contigs on chromosomes and then using a probabilistic model to predict the genomic locus along the chromosome based on interaction frequency and genomic distance. This worked well for *de novo* assembly of the human genome, and the researchers also adapted it to predict the locations of previously unplaced fragments of the human genome. The approach only required a library of paired-end short inserts and Hi-C data.

Shendure and his team, on the other hand, used libraries of paired-end short reads, a library of 3-kilobase mate pairs and Hi-C data for their algorithm, named Lachesis, in reference to one of the Greek Fates. In a tiered approach, they created high-quality *de novo* assemblies of the human, mouse and fly genomes. Unlike the algorithm by the Dekker group, Lachesis cannot infer the chromosome numbers of an organism, but it can orient contigs the right way after placement.

Looking forward, both Dekker and Shendure see the need for integrated rather than step-wise data analysis. "An approach that simultaneously takes into account all data types in a single step is likely to do better," says Shendure.

Additional improvements will also come from the experimental side. Current Hi-C data show cell type–specific interactions of genomic loci, which Hi-C originally had been designed to discover, that can mask the signal used for scaffolding. Dekker and his team recently solved the structure of the metaphase chromosome, and Kaplan suggests that such metaphase Hi-C data will get rid of cell type–specific interactions.

Hi-C data help find the right genomic position of short sequence reads. Adapted from *Nature Biotechnology* (Burton et al., 2013).

Researchers in Shendure's lab will focus on making the Hi-C protocol robust for tissues from diverse organisms, as current data are all derived from cell lines.

Other recent work has shown that Hi-C data can be used to reconstruct haplotypes; Kaplan sees the combination of genome assembly and haplotype phasing as an exciting possibility.

Neither the Dekker nor the Shendure teams have a track record in genome assembly, so both groups are eager for assembly experts to try out their algorithms and suggest further improvements. Recent community-driven comparisons have made clear that there is not one program that outperforms all others, but the algorithms from the Dekker and Shendure labs will provide an important starting point for bridging large gaps in the assemblies.

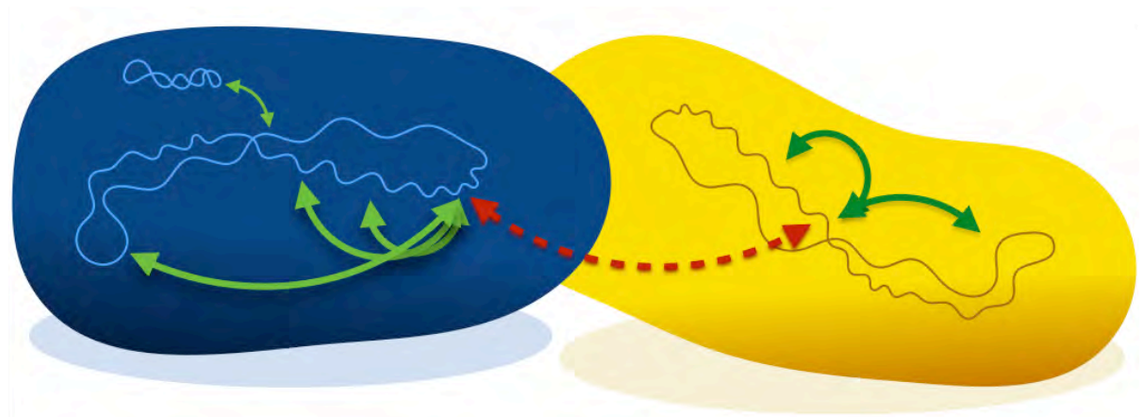
Nicole Rusk

RESEARCH PAPERS

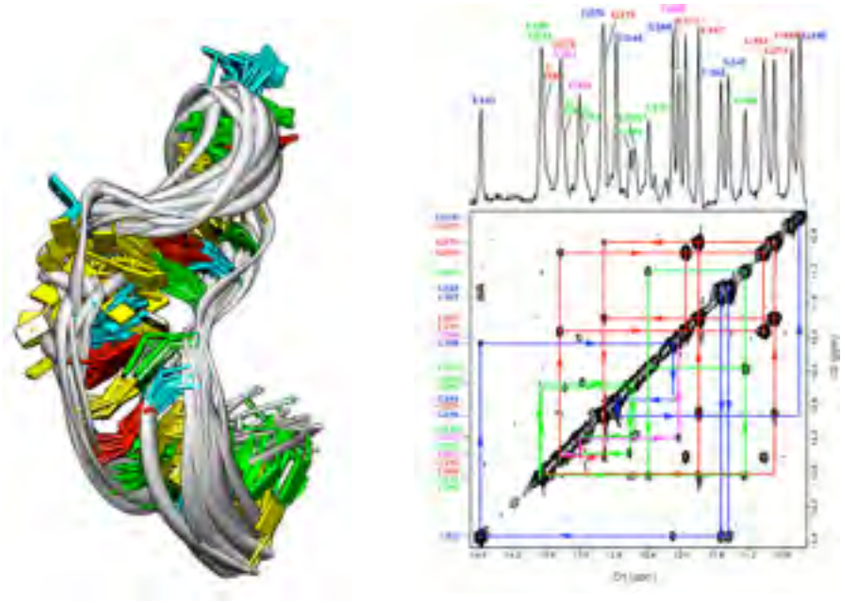
Burton, J.N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).

Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).

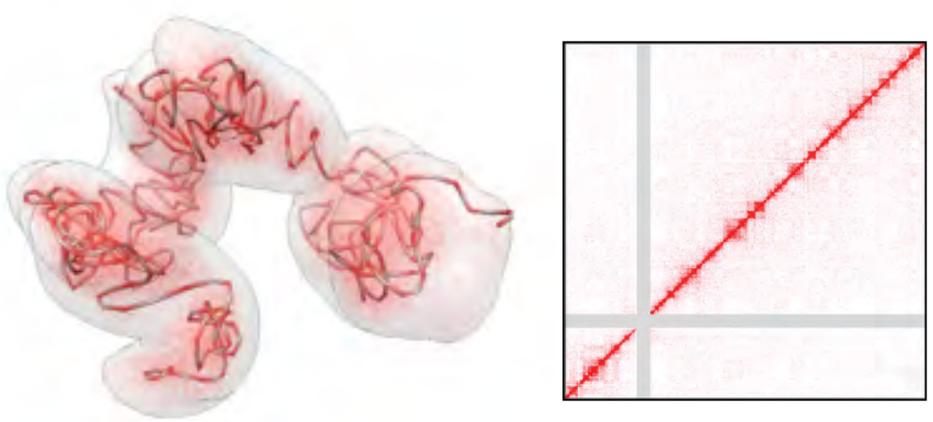
Rusk, N. (2013).
Genomics: Genomes in 3D improve one-dimensional assemblies. Nature Methods, 11(1), 5–5. doi:10.1038/nmeth.2795



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

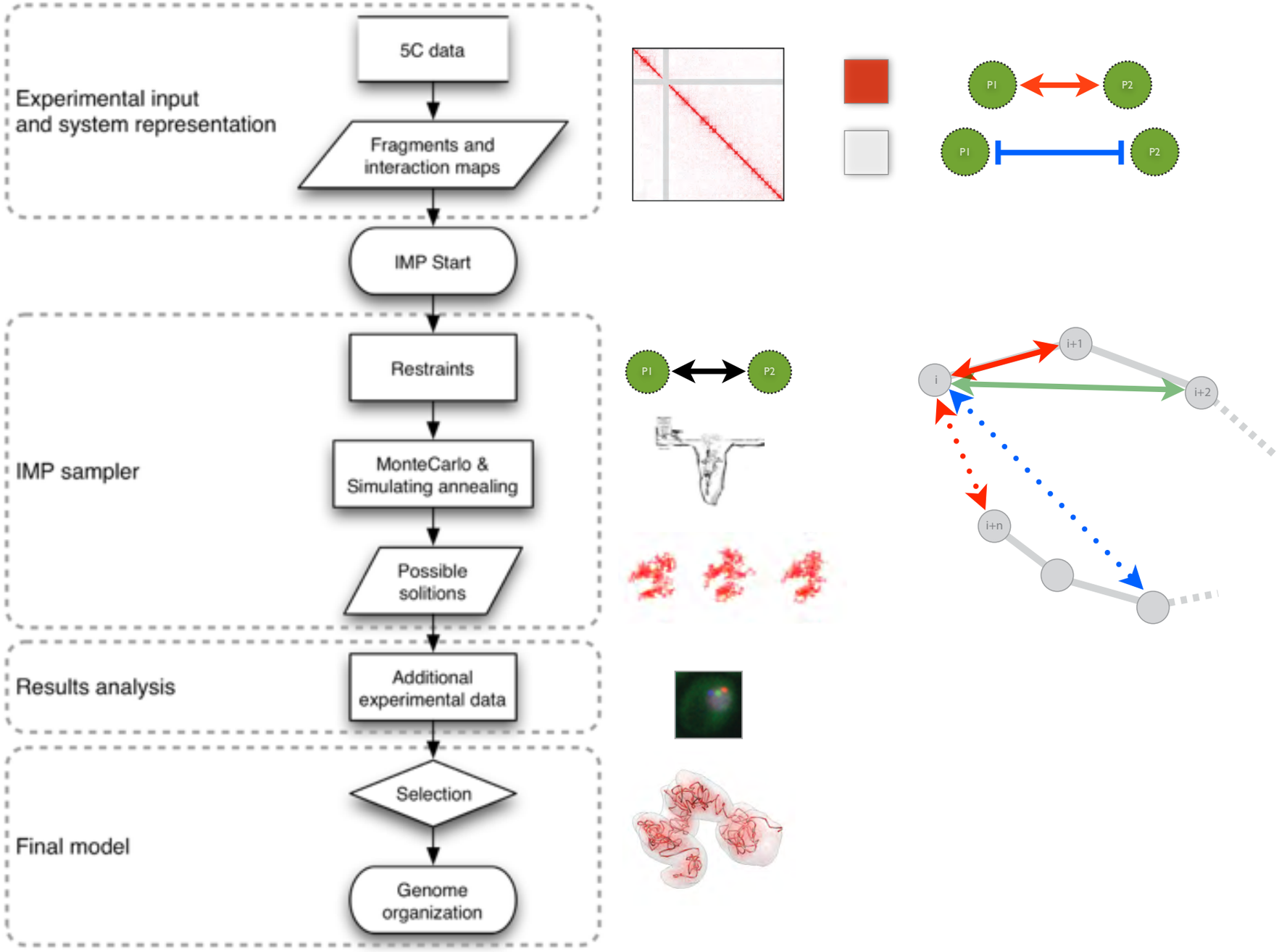


Biomolecular structure determination
2D-NOESY data



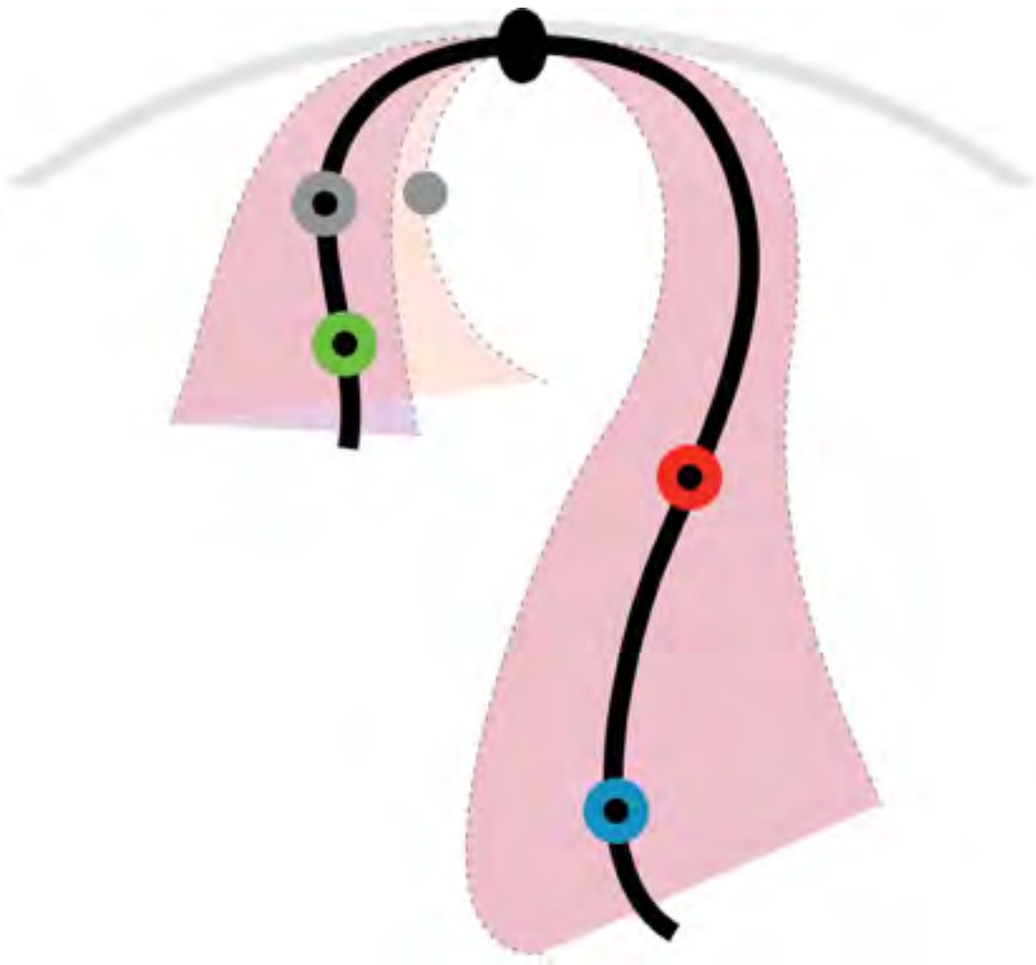
Chromosome structure determination
3C-based data

TADbit


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
P1
P2
P1
P2
P1
P2
i
i+1
i+2
i+n

TADbit

Mating-specific structure for yeast chrIII?

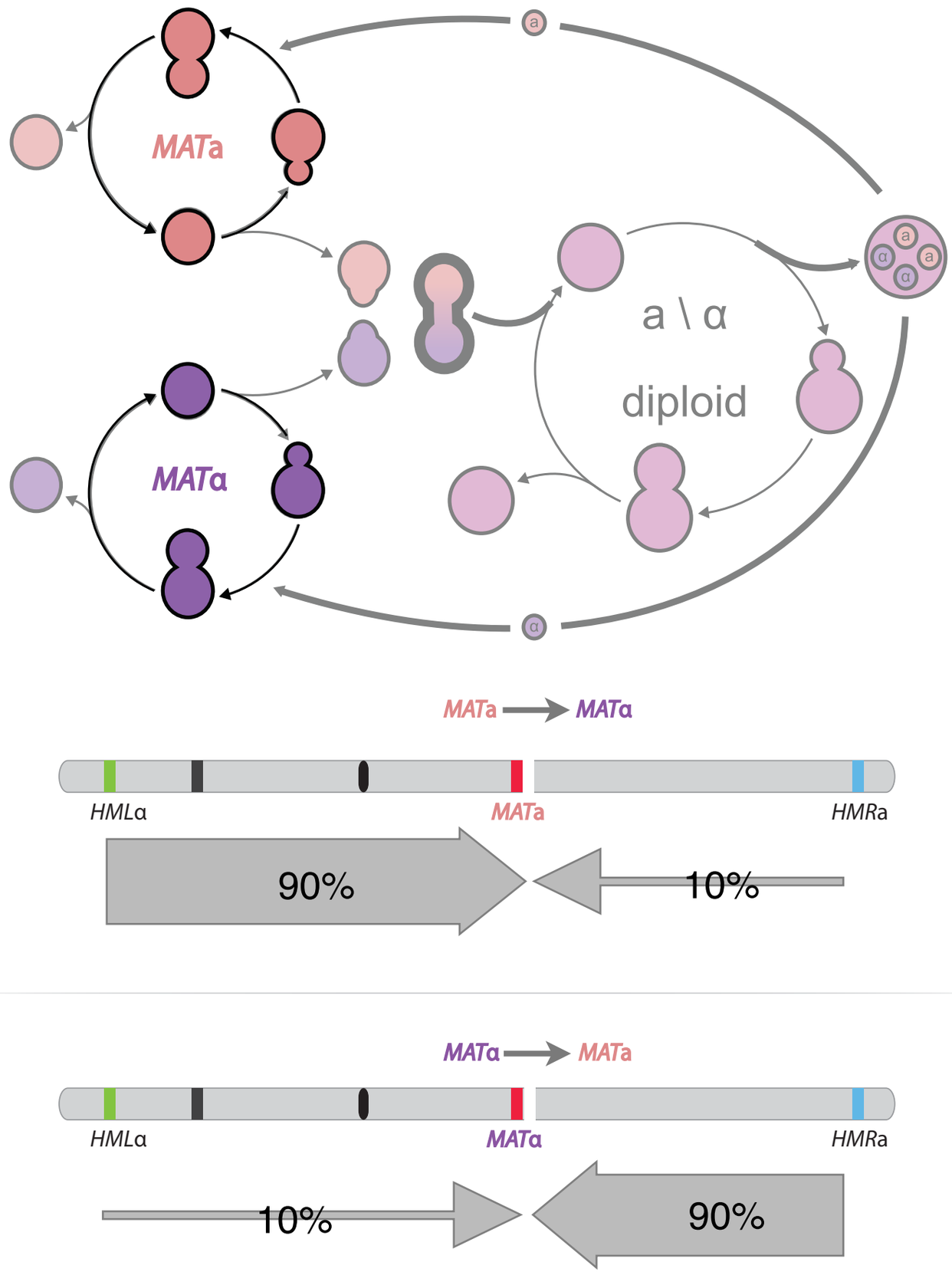


Jon-Matthew Belton
UMASS

Davide Baù
CNAG/CRG

Job Dekker
Program in Systems Biology
Department of Biochemistry and Molecular Pharmacology
University of Massachusetts Medical School
Worcester, MA, USA

Kerstin Bystricky
Chromatin and gene expression
Laboratoire de Biologie Moléculaire Eucaryote - CNRS
Toulouse, France



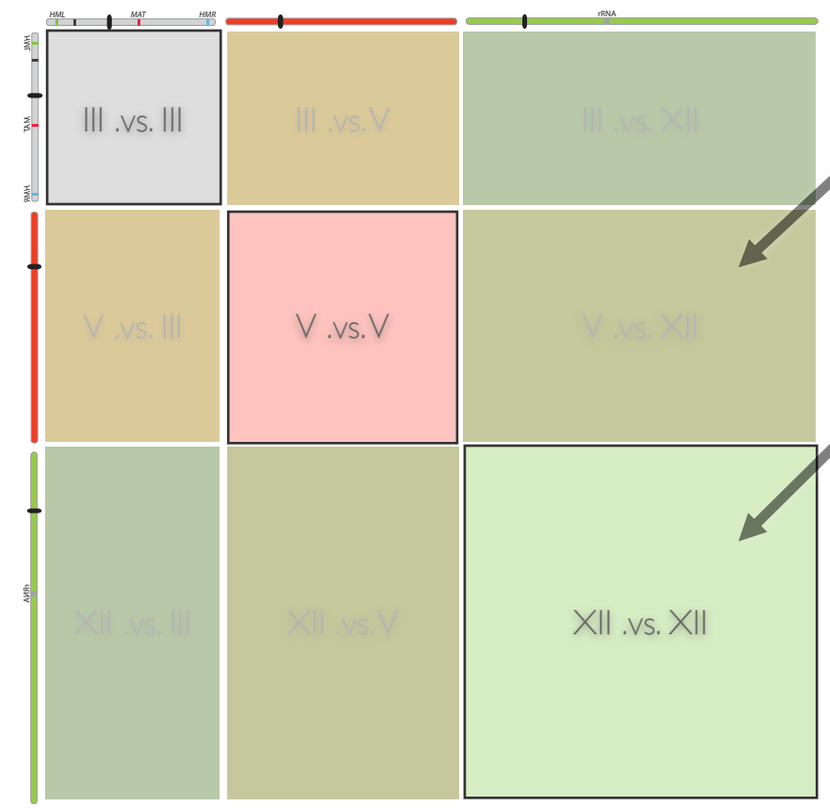
Wu, X. H., C. Wu, et al. Genetics (1997).

5C chromosome conformation

Chr. III - 317 kb: Mating Type Switching

Chr. V - 577 kb: Control

Chr. XII - 1 Mb: rDNA array



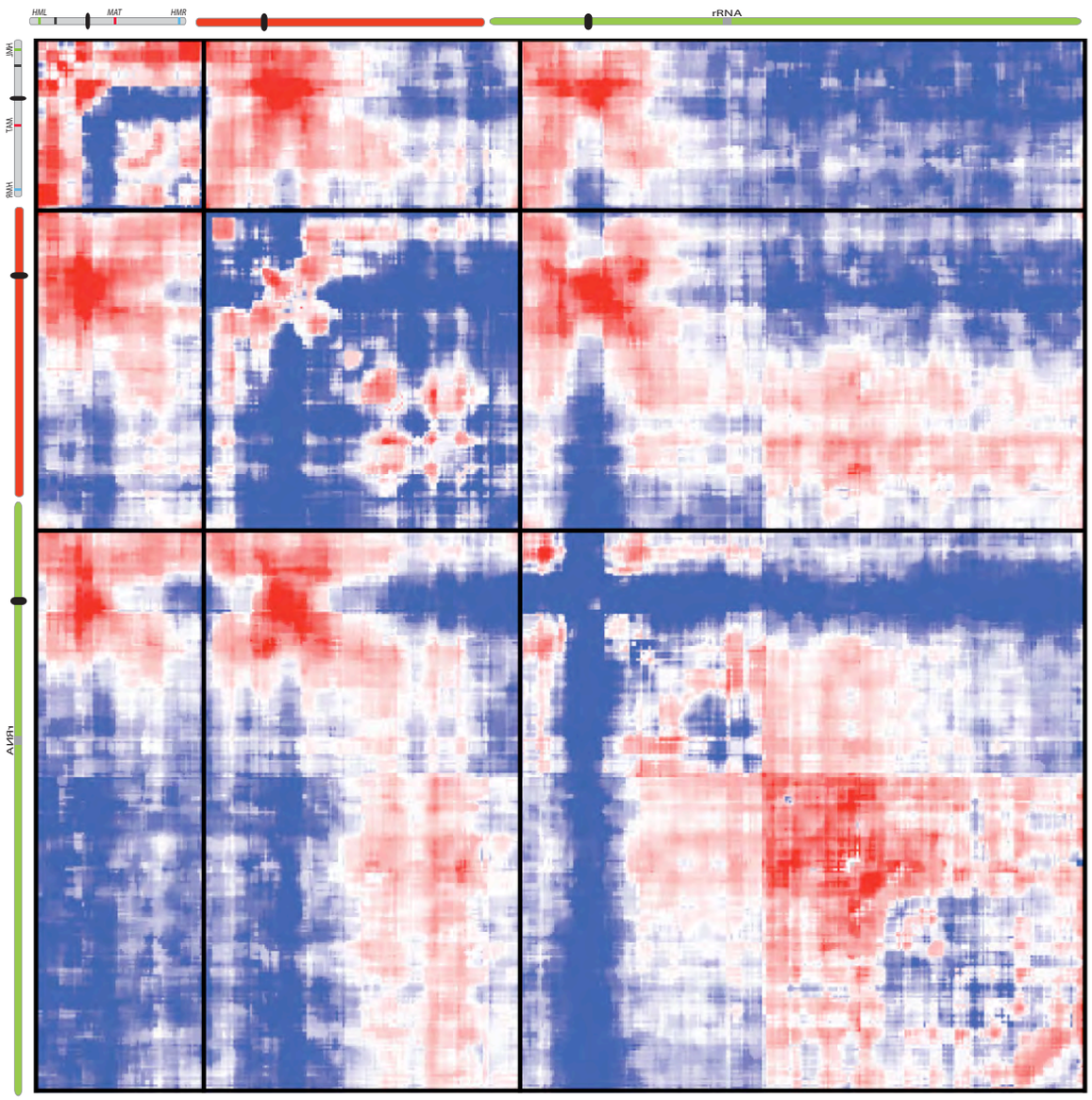
Interactions between chromosomes

Interactions within a chromosome

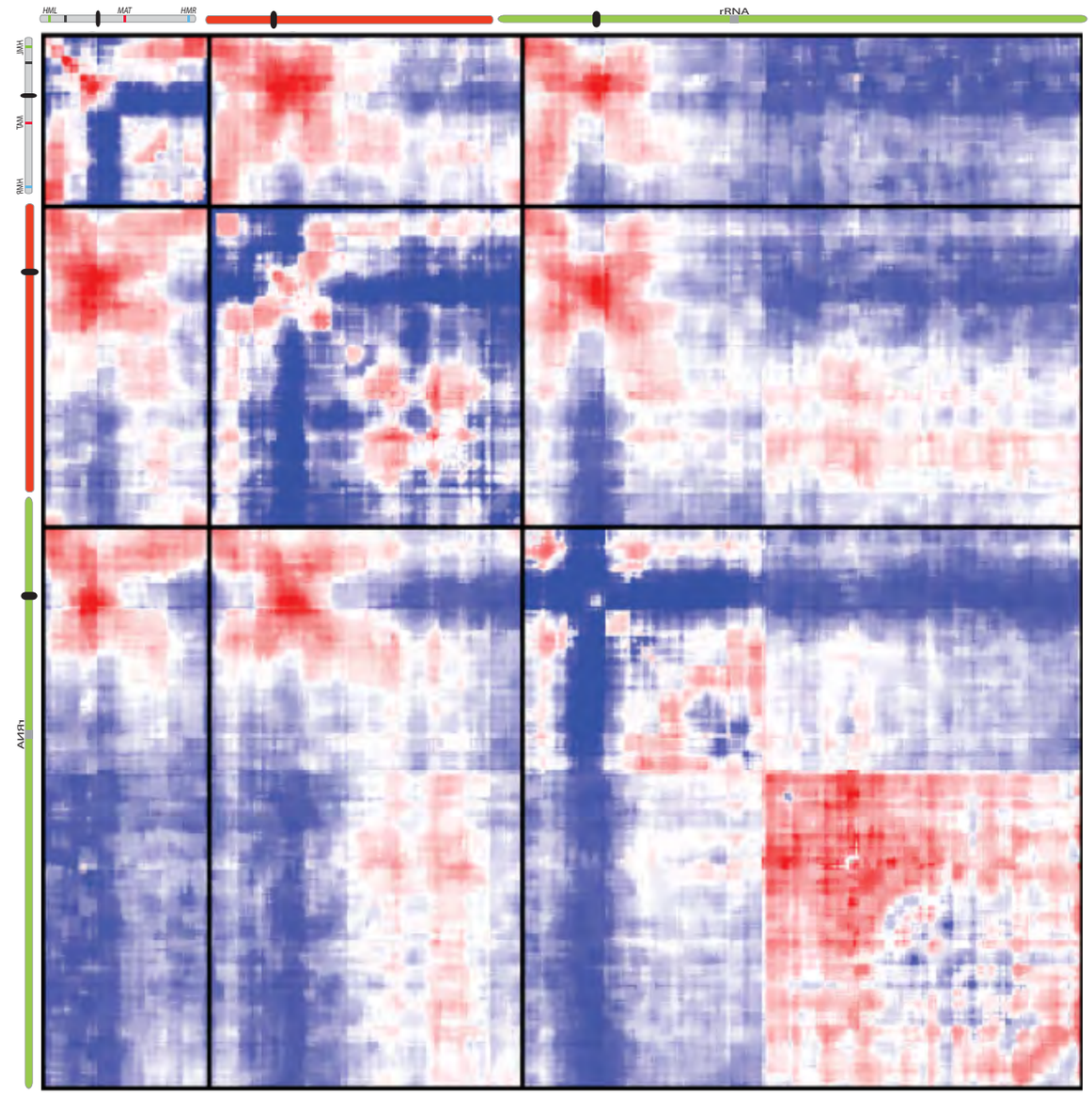
~100,000 possible interactions!

Global structure is *similar* between mating types

*MAT*a



*MAT*α

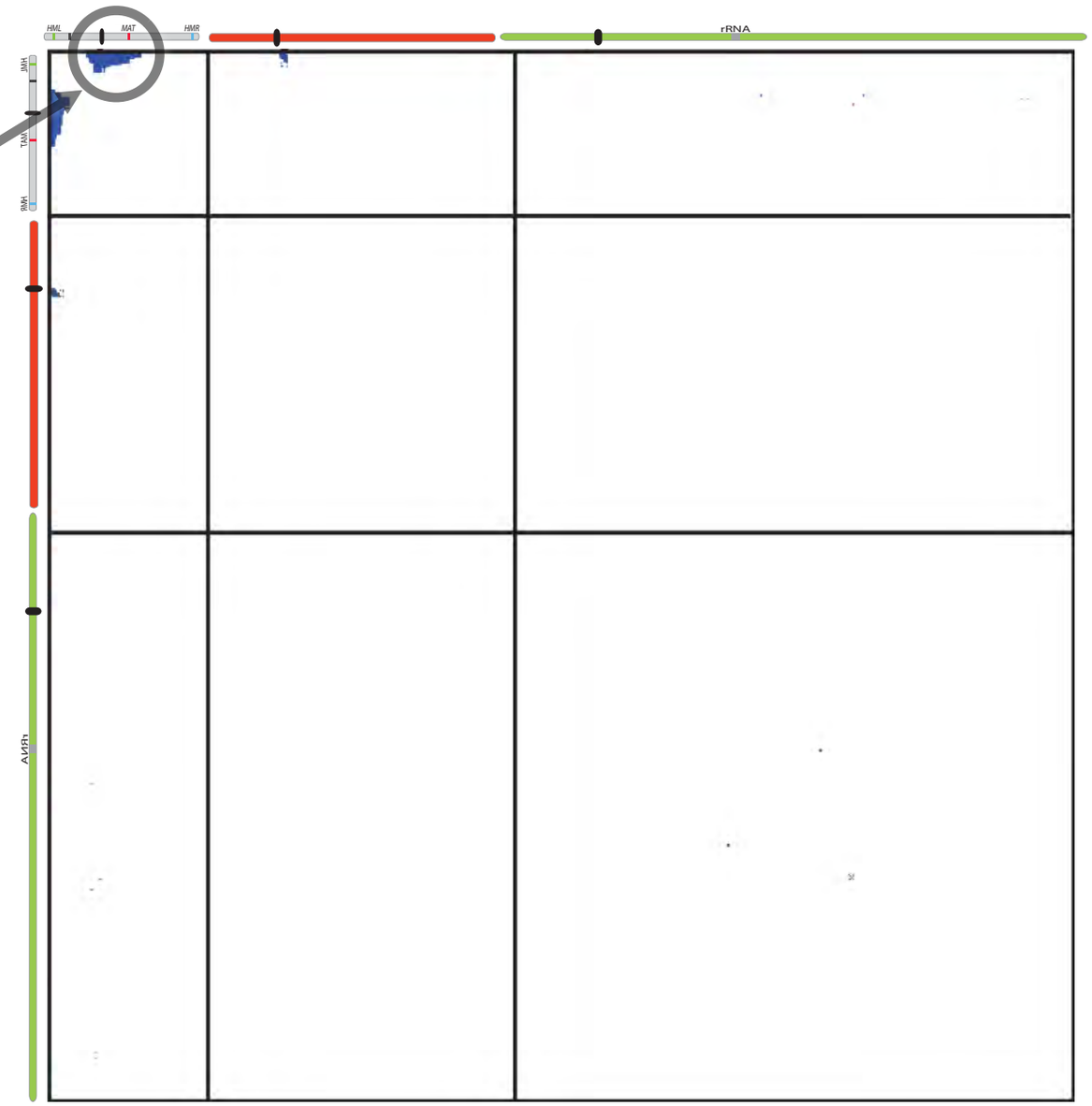


Difference in chromosome conformation

$\text{Log}_2(MAT\alpha\ /\ MAT\boldsymbol{a})$

= Enrichment of interaction in *MAT*α

= Enrichment of interaction in *MAT**a***



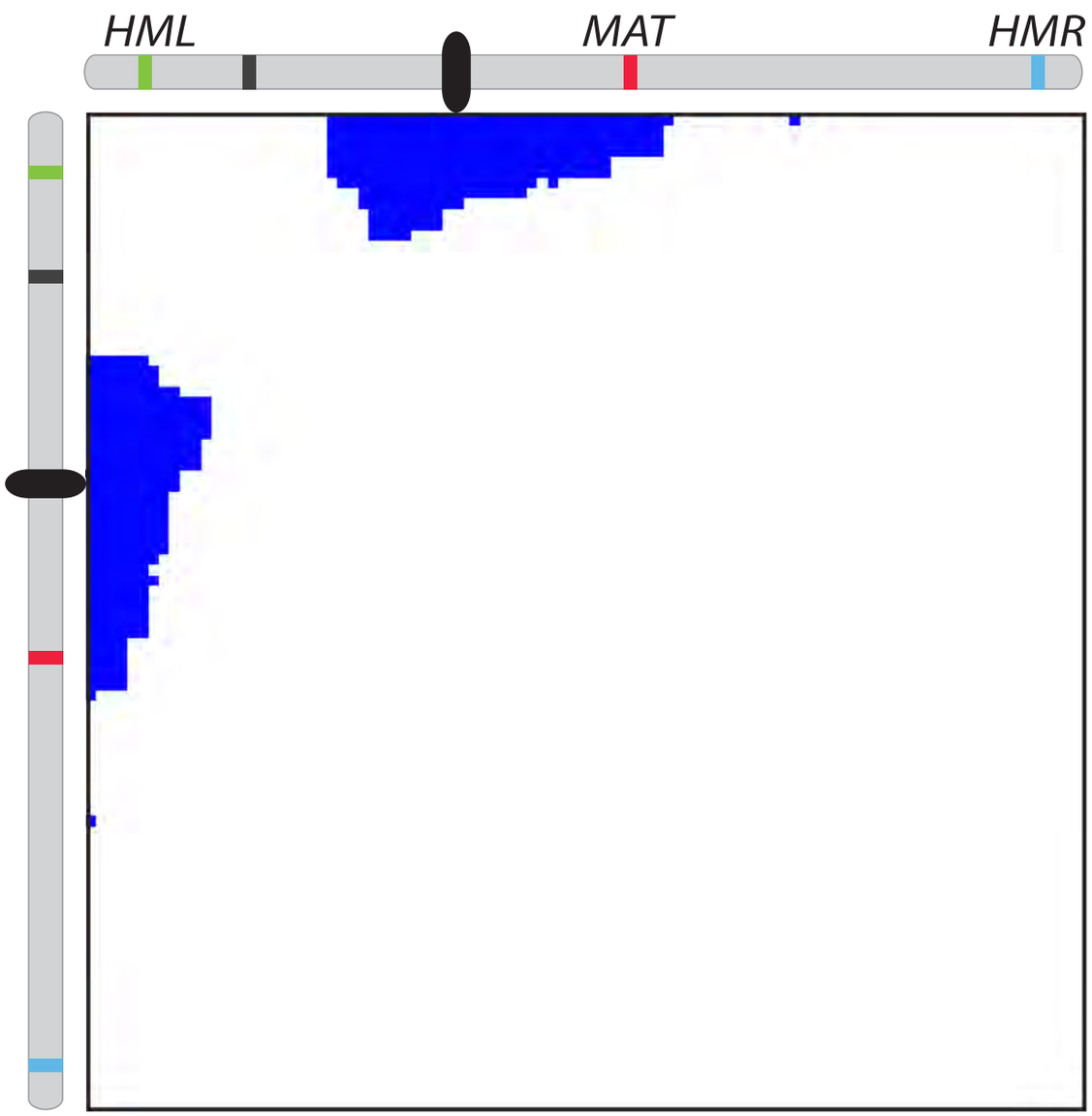
Only major difference in conformation is on chromosome III

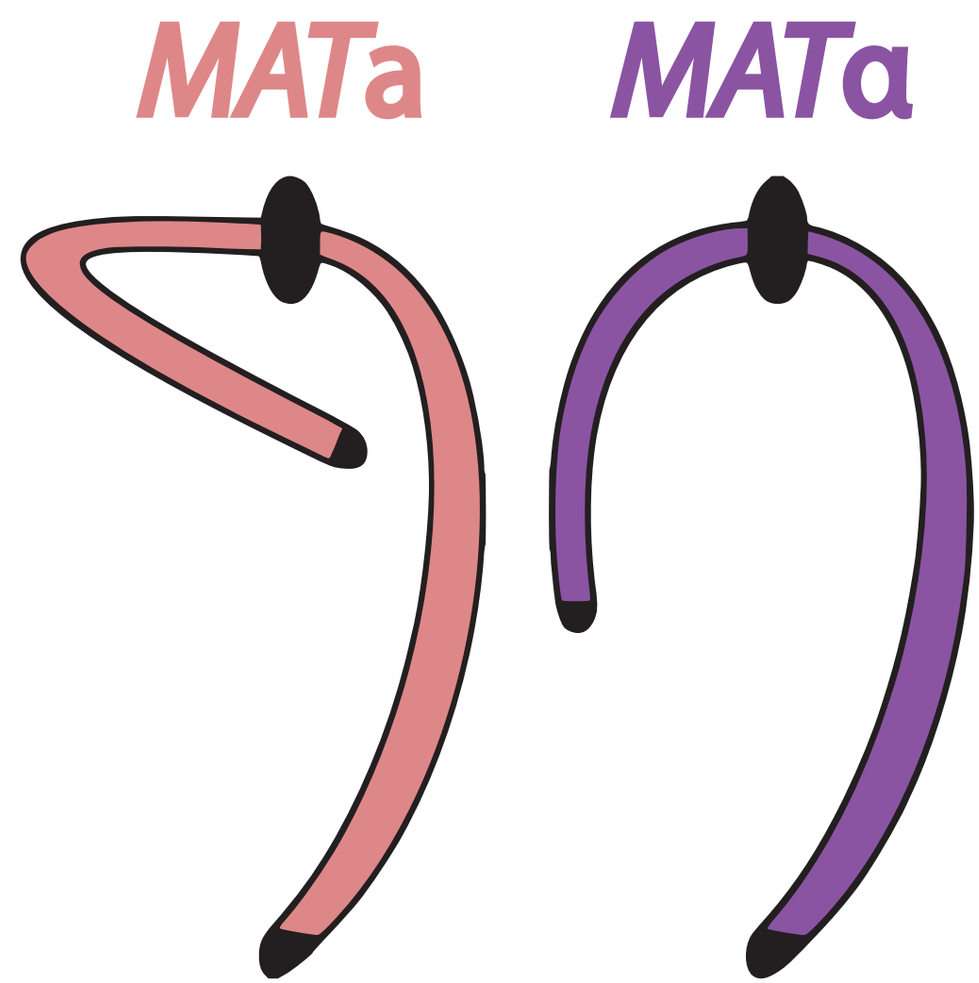
Difference in conformation of the left arm of chromosome III

$\text{Log}_2(MAT\alpha\ /\ MAT\mathbf{a})$

= Enrichment of interaction in *MAT*α

= Enrichment of interaction in *MAT**a***





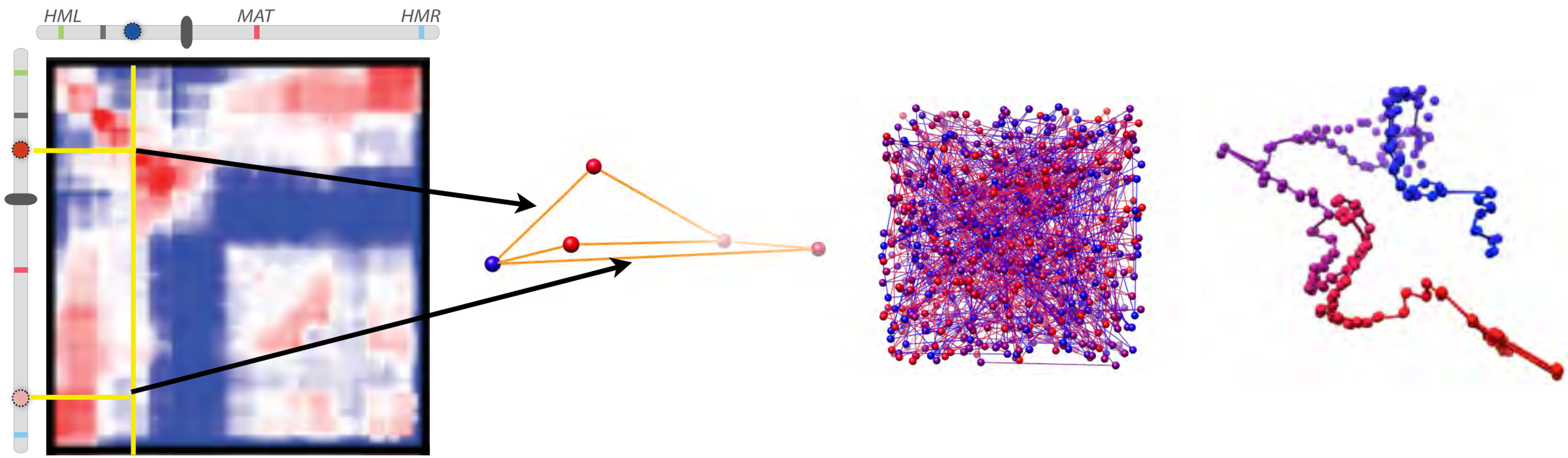
Average 3D models of ChrIII using TADbit

5C Contact probabilities

5C data converted into distance restraints

Random initial organization

Fitting to distance constraints

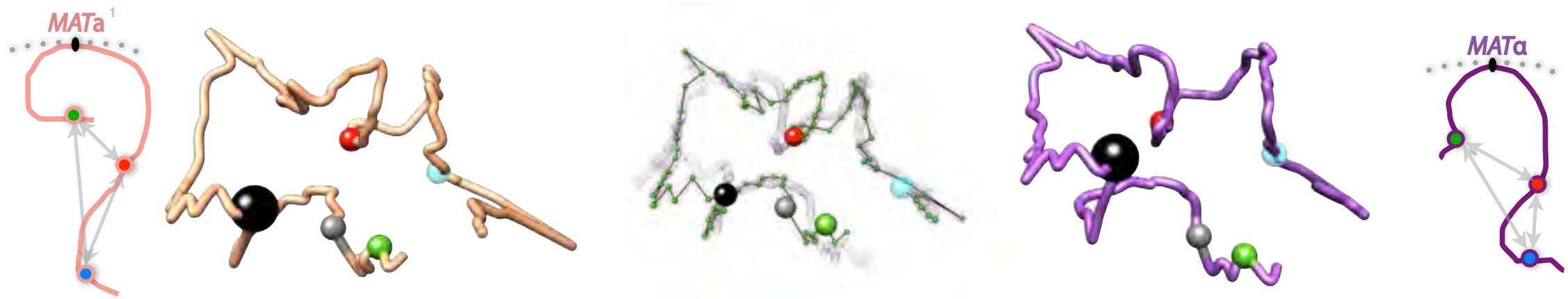


MATa

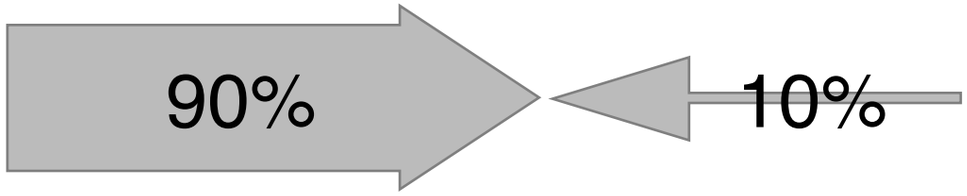
5,000 models
1,000 selected

MATα

Mating type-specific conformation of chromosome III

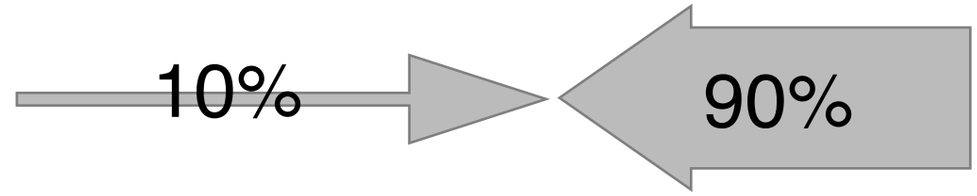


MATa → MATα

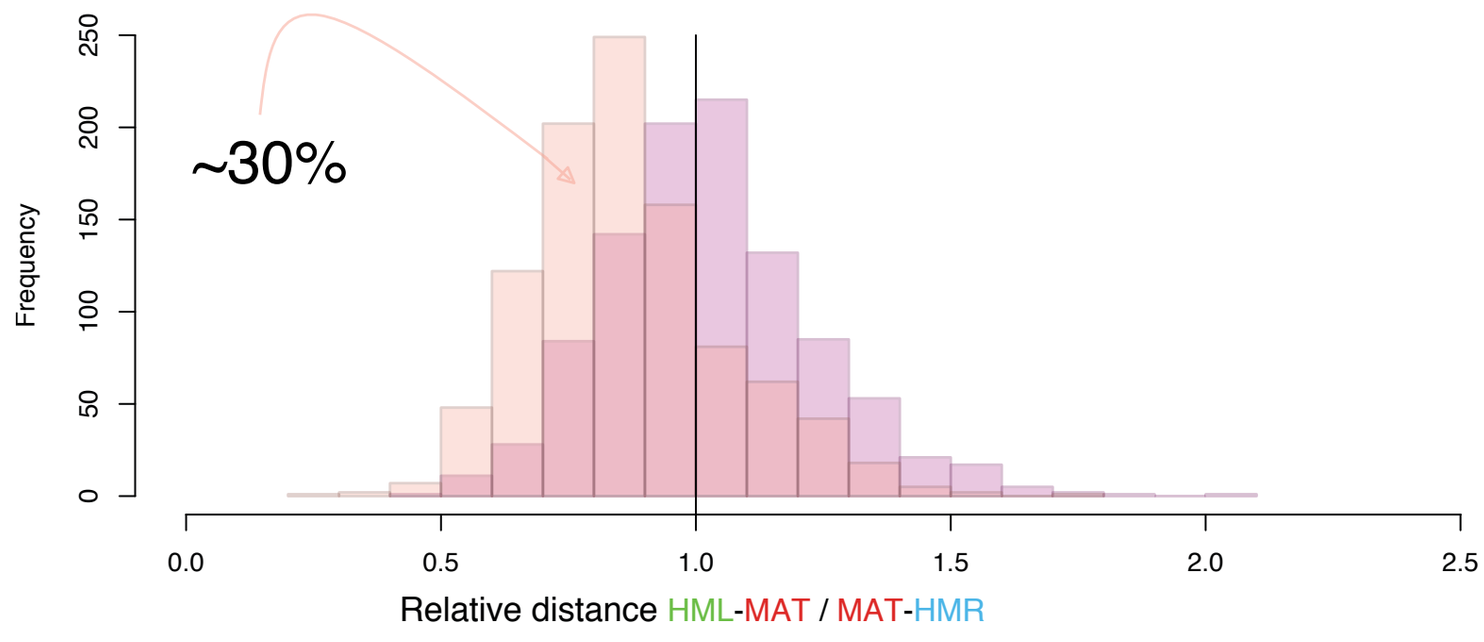


589nm 691nm

MATα → MATa

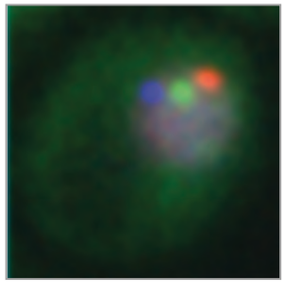


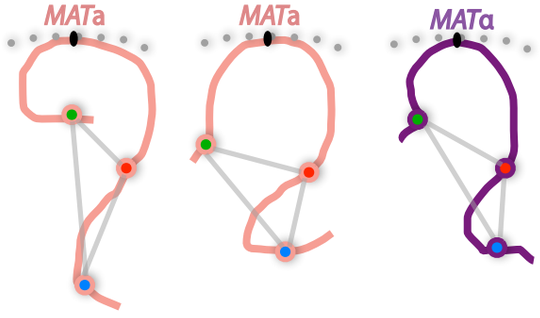
726nm 719nm



3D chrIII for mating in yeast

Sub-population in MATa responsible of mating-type recombination





Imen Lassardi
LBME/CNRS

Structuring the COLORs of chromatin

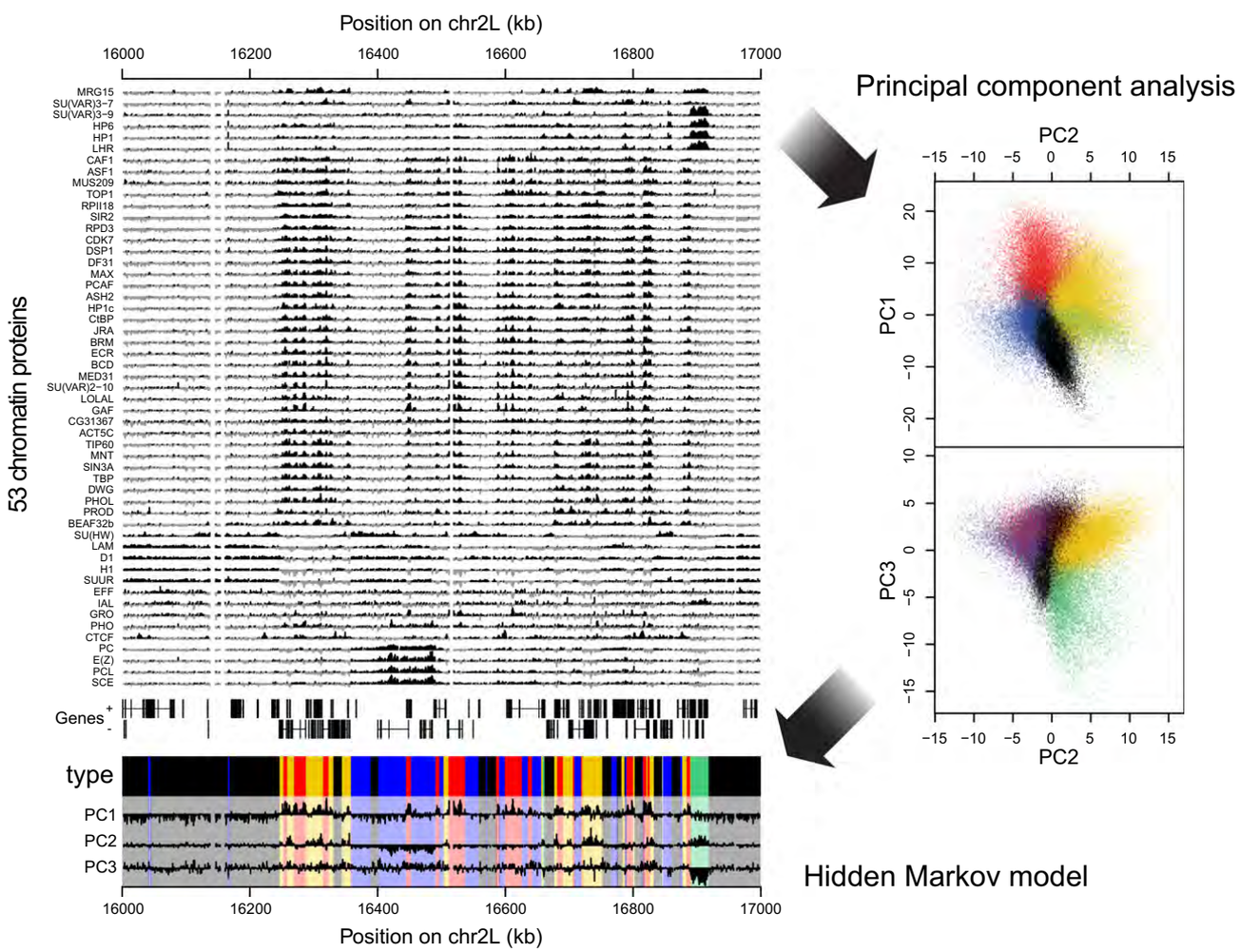
Davide Baù

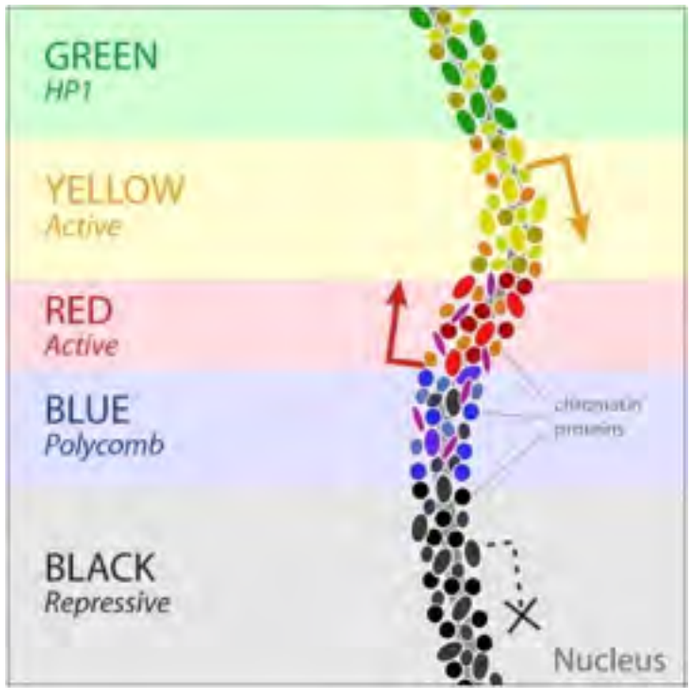
François Serra

Guillaume Filion
Gene Regulation, Stem Cells and Cancer
Centre de Regulació Genòmica
Barcelona, Spain

The COLORs

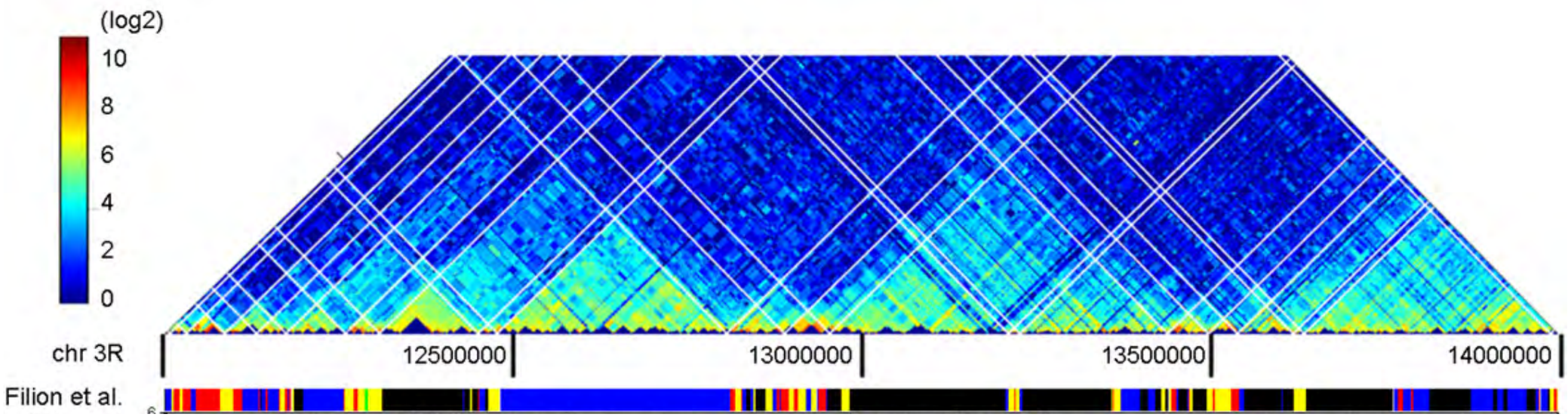
Filion et al. (2010). Cell, 143(2), 212–224.





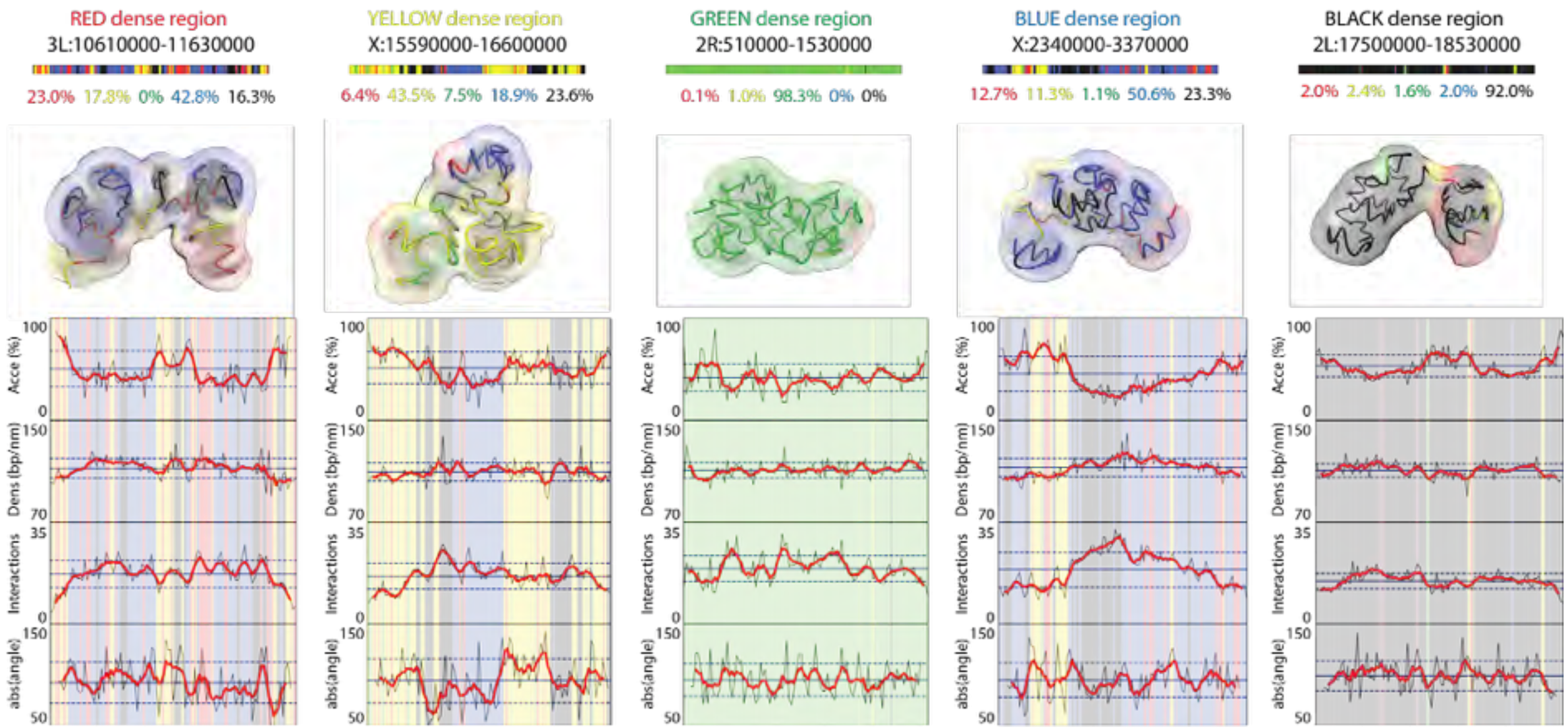
Functional COLORs

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

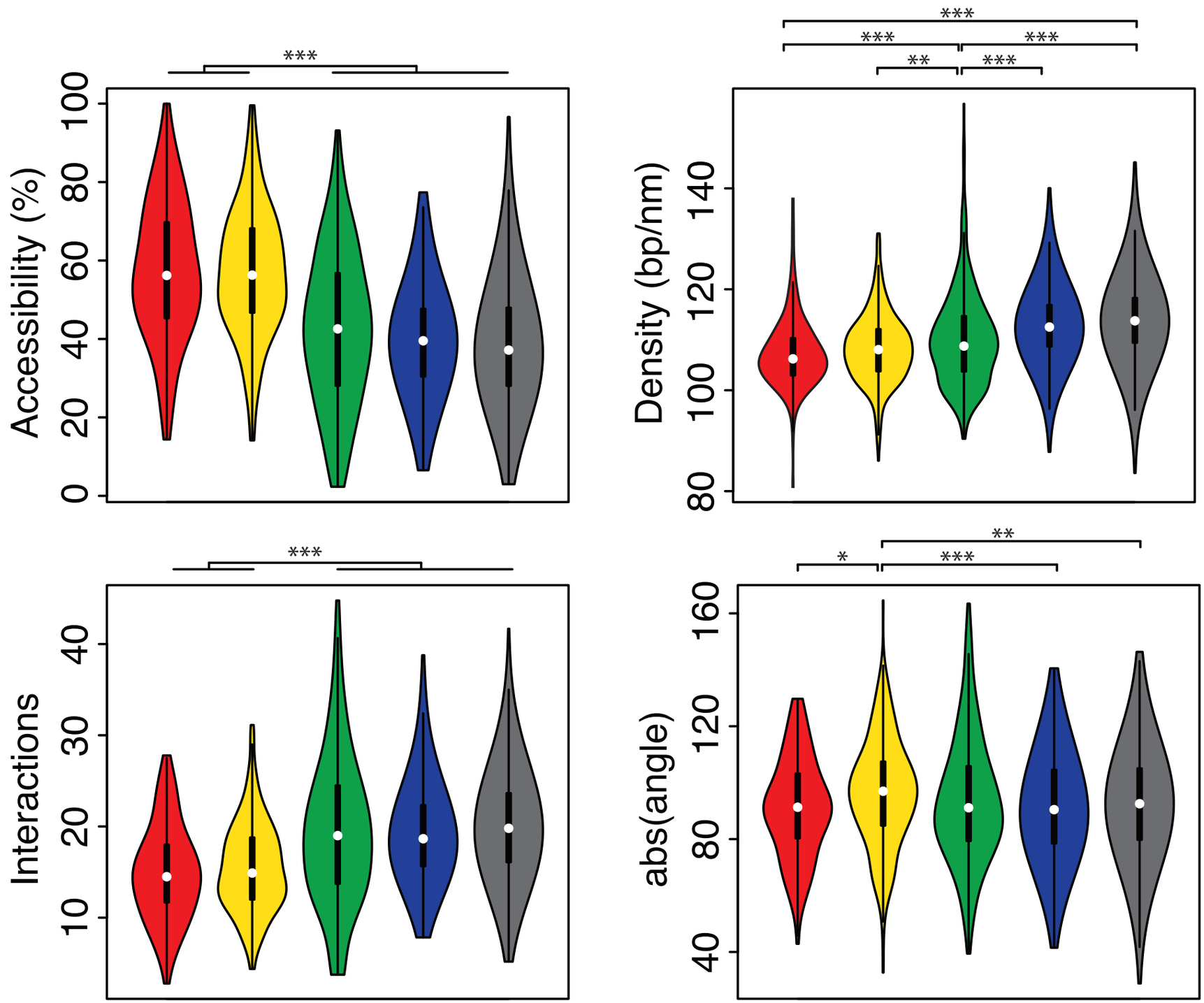


50 ~1Mb regions
10 for each color

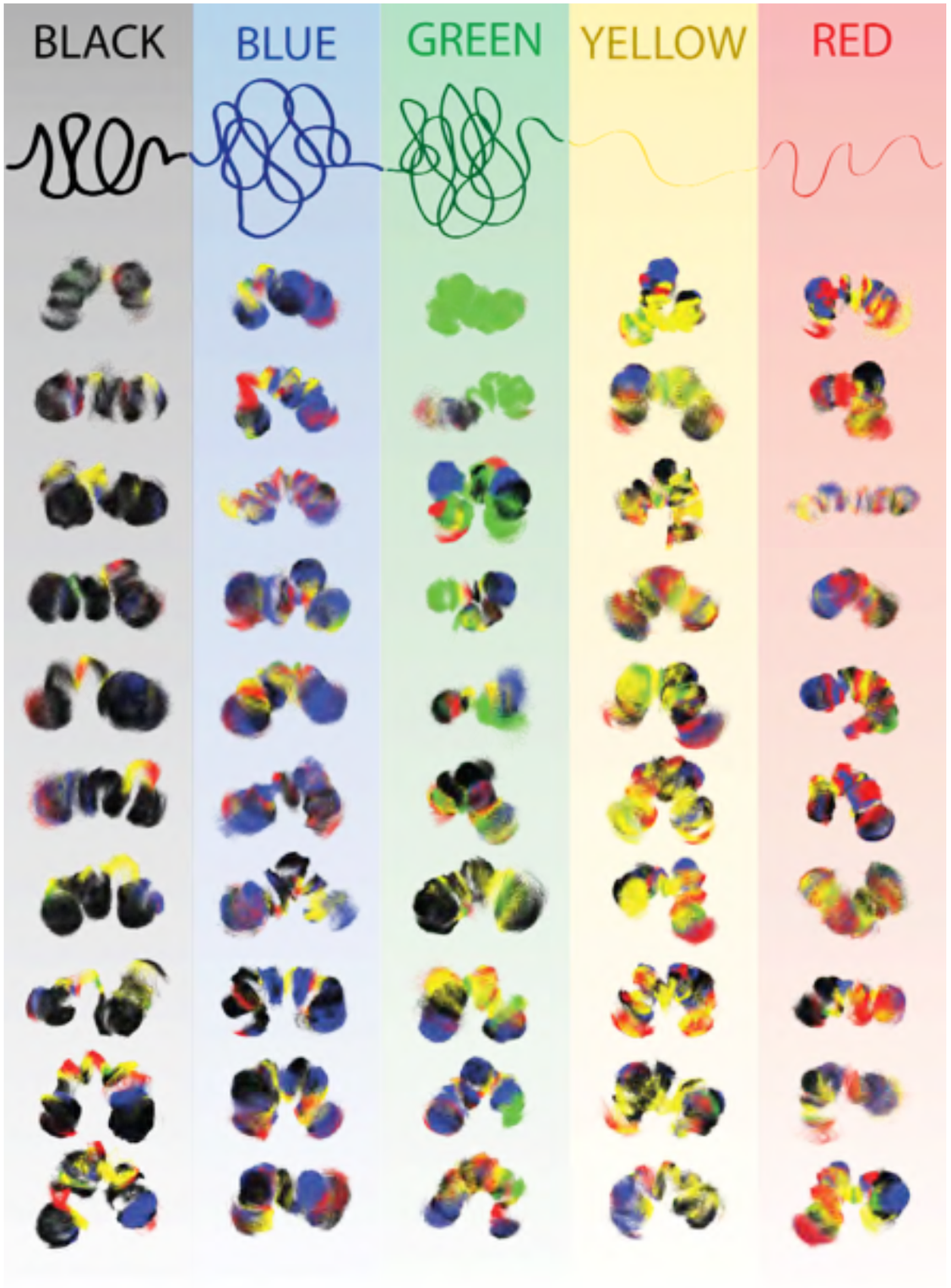
Structural COLORs

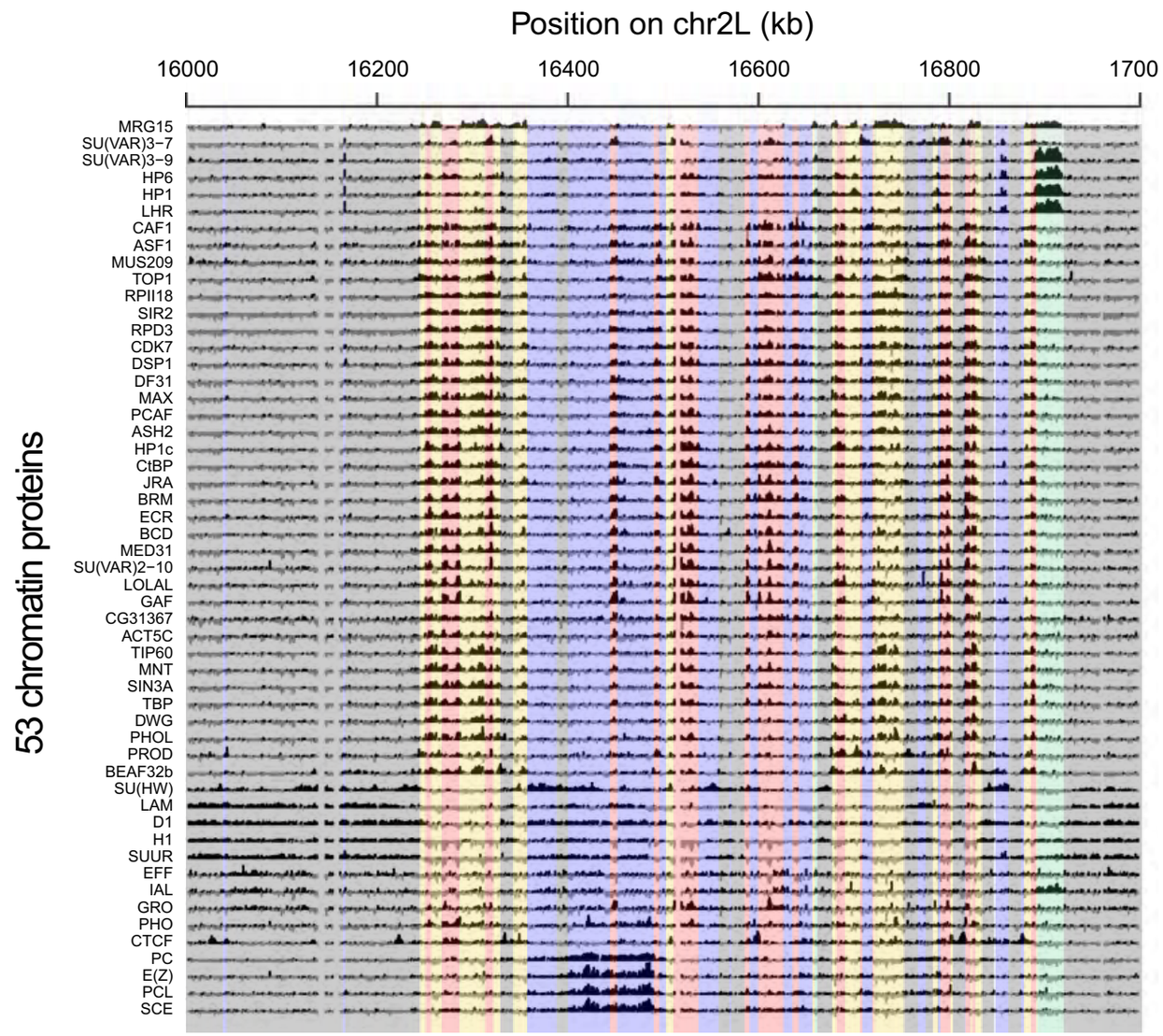


Structural COLORs



Structural COLORs





cnag CRG Centre for Genomic Regulation

On TADs and hormones

François le Dily

Davide Baù

François Serra

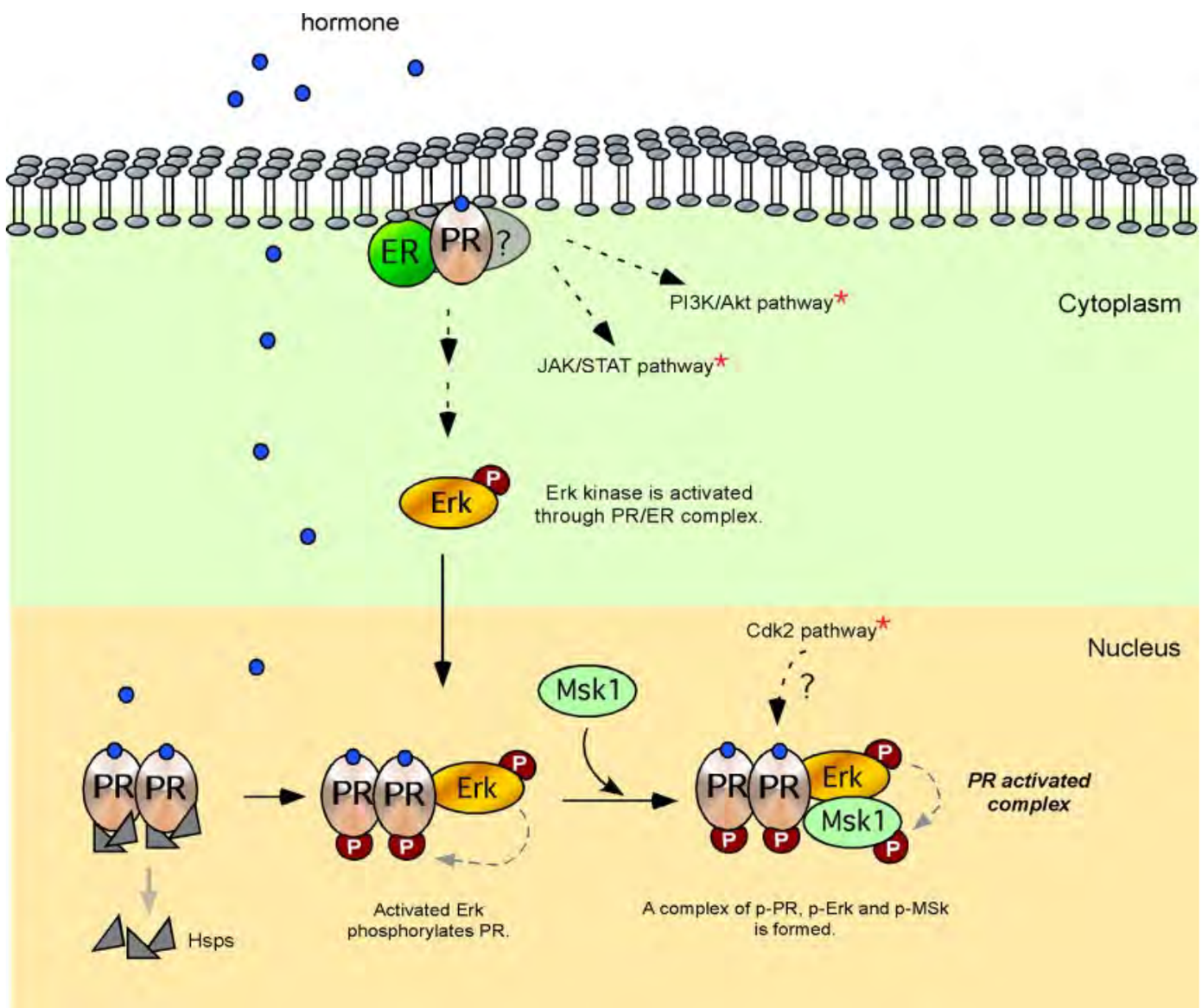
Miguel Beato & Guillaume Filion

Gene Regulation, Stem Cells and Cancer

Centre de Regulació Genòmica

Barcelona, Spain

Progesterone-regulated transcription in breast cancer



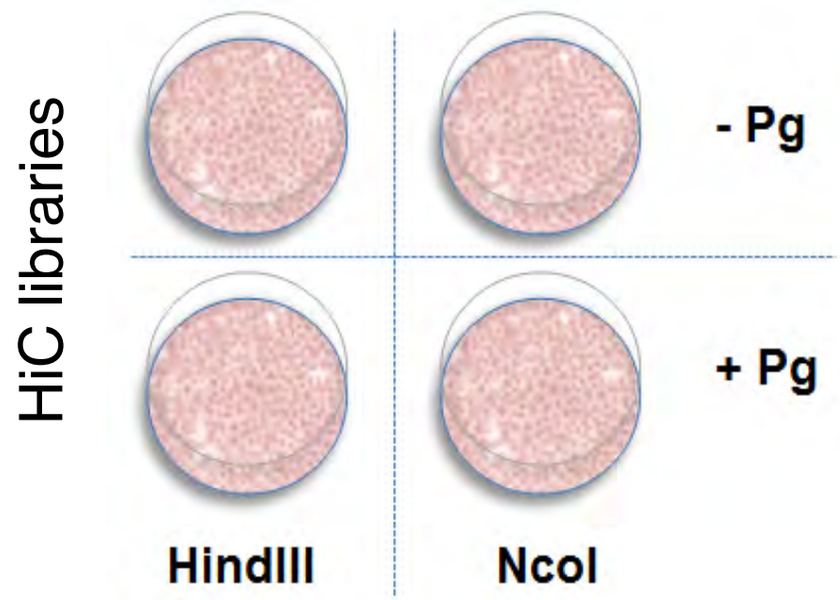
Vicent *et al* 2011, Wright *et al* 2012, Ballare *et al* 2012

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

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

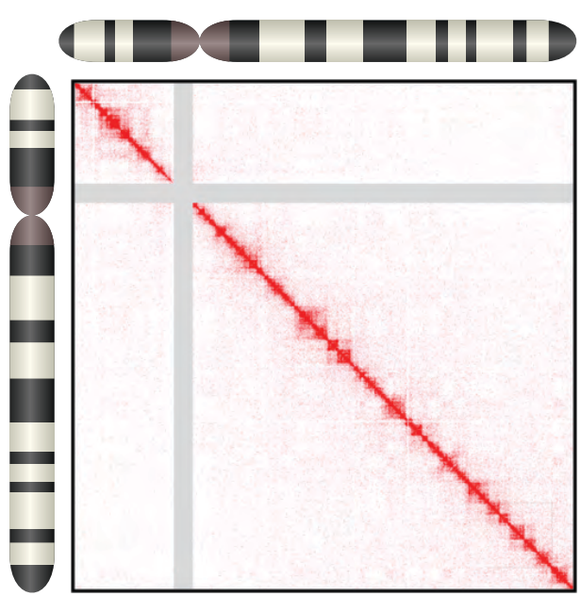
Regulation in 3D?

Experimental design

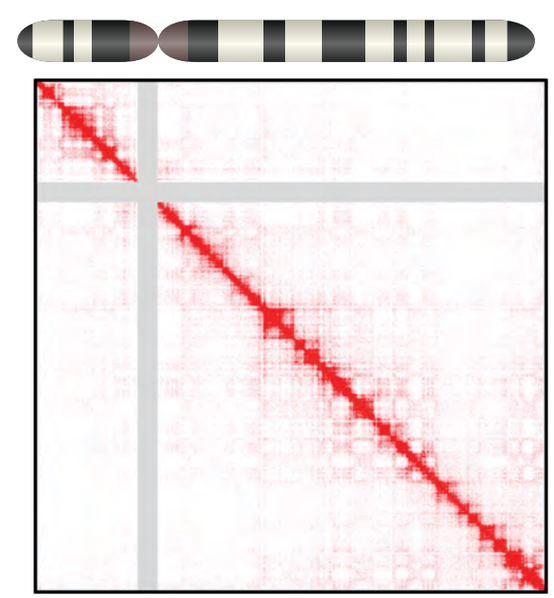


ChIP-Seq
RNA-Seq
Hi-C

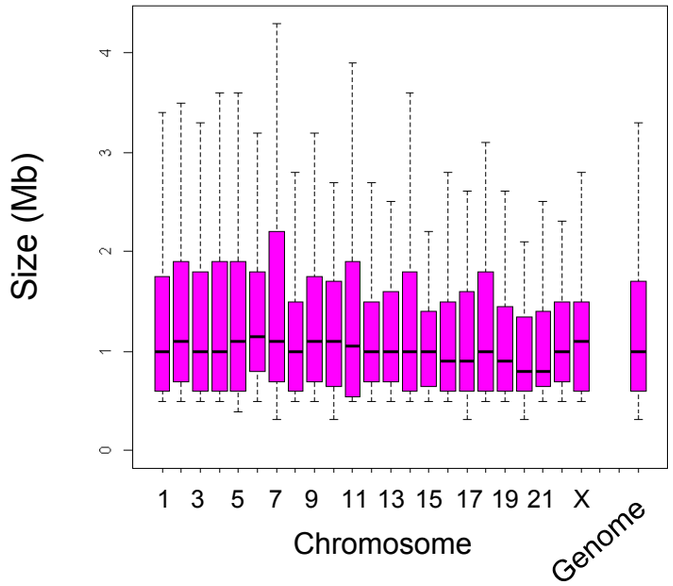
Chr.18 (Hind III)



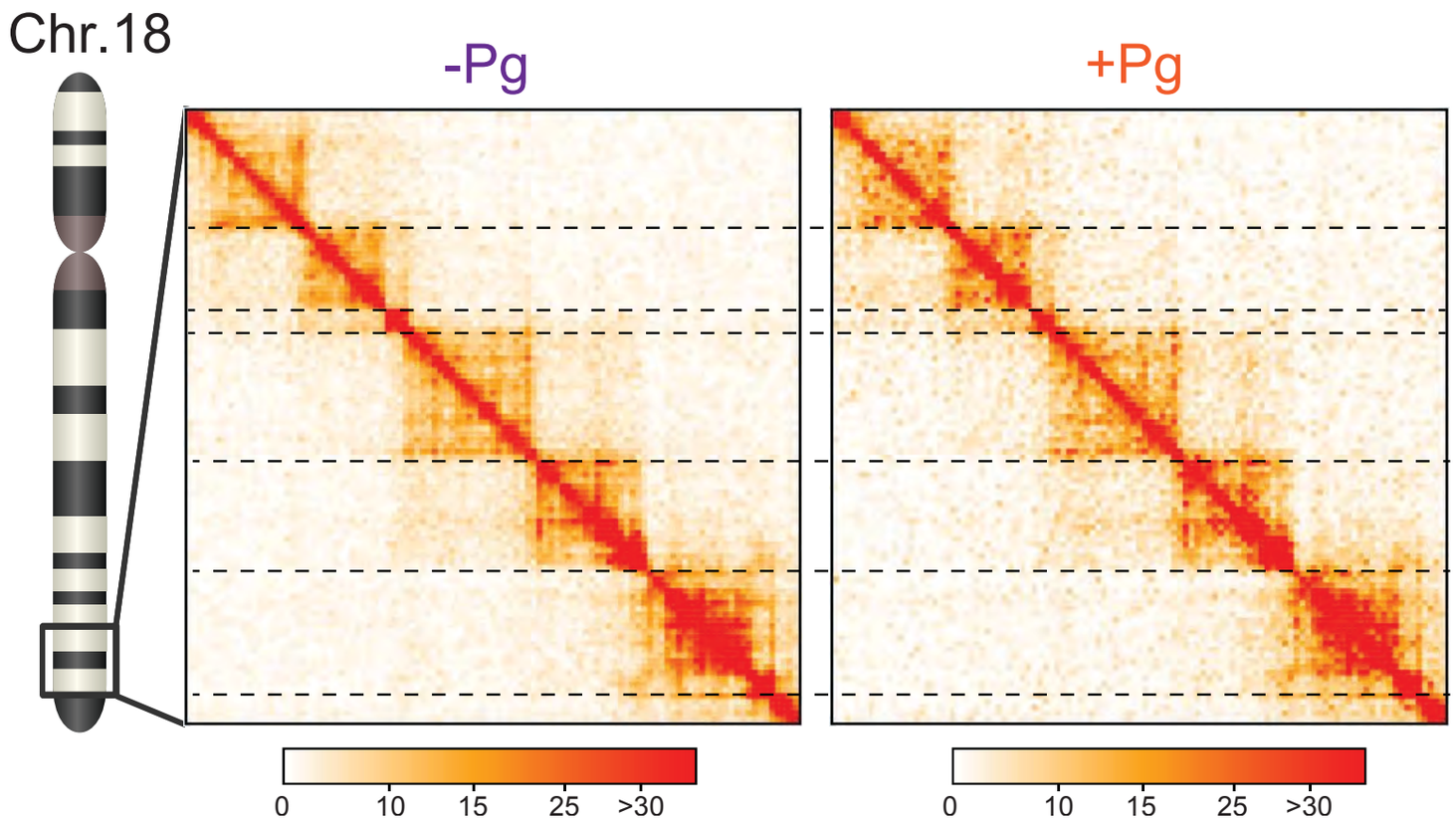
Chr.18 (NcoI)

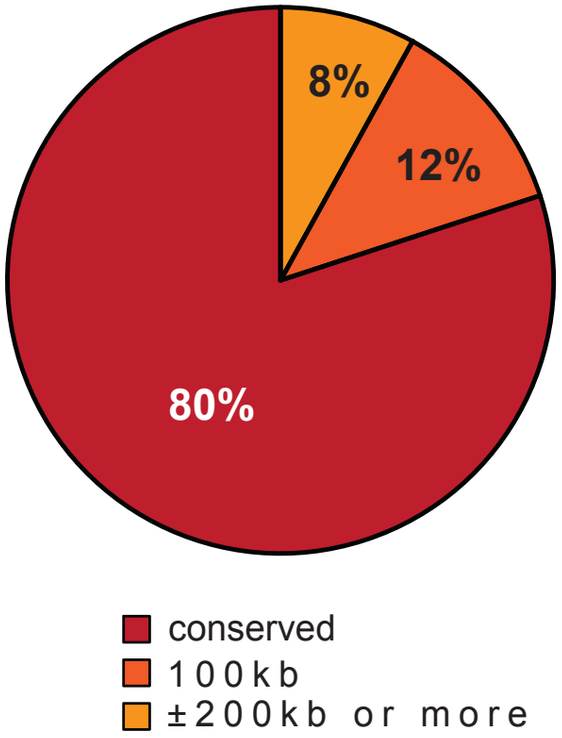


Are there TADs? how robust?

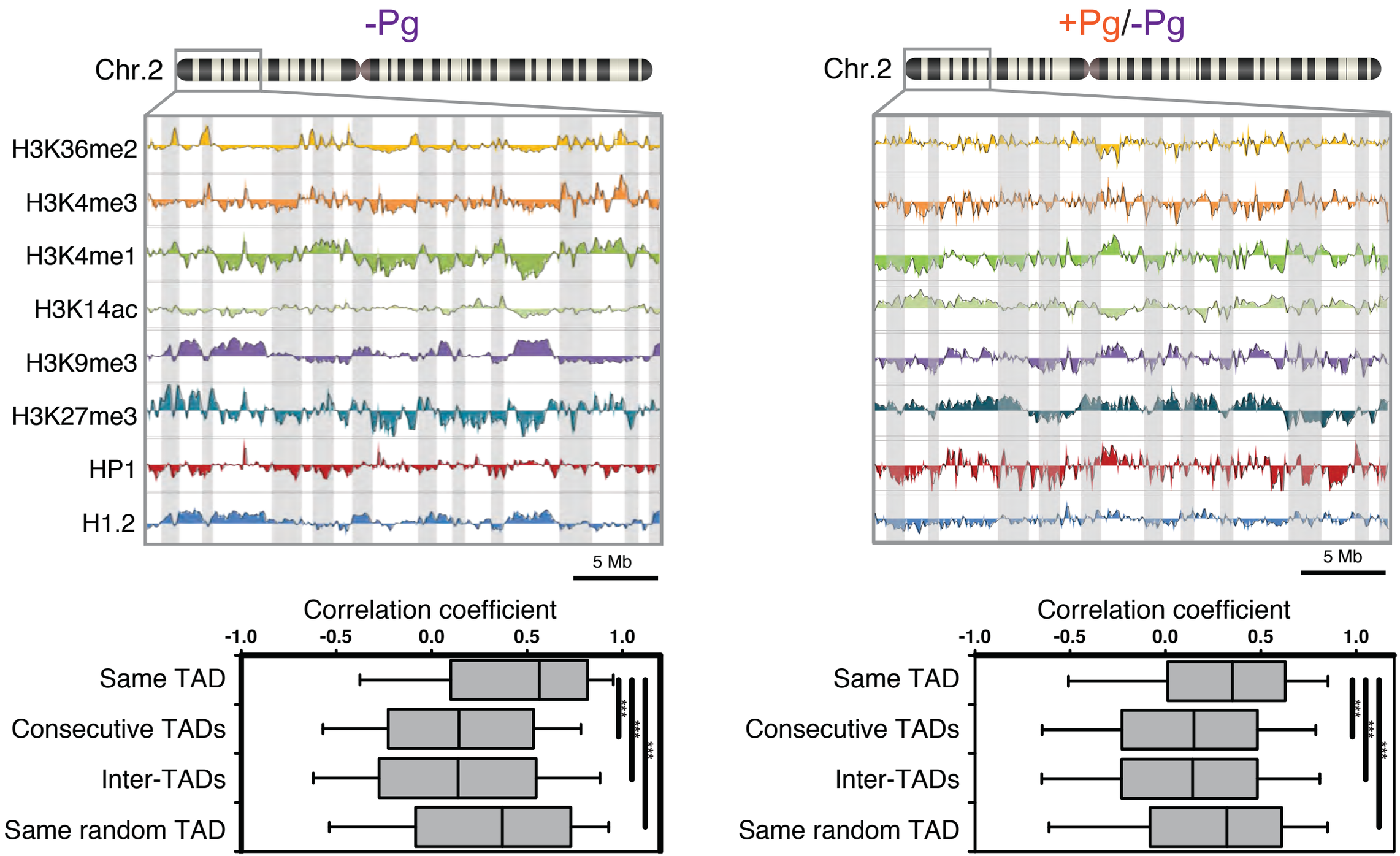


>2,000 detected TADs

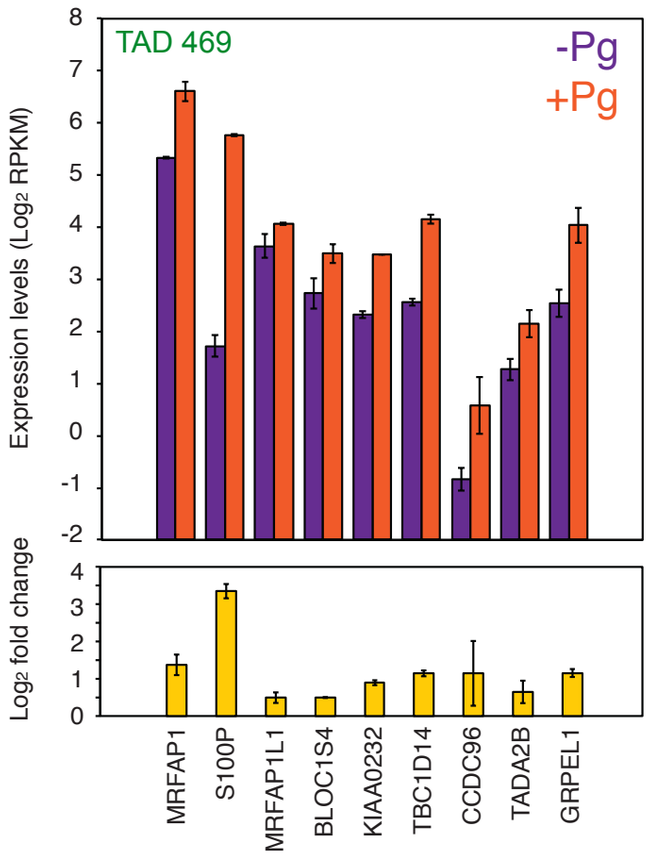


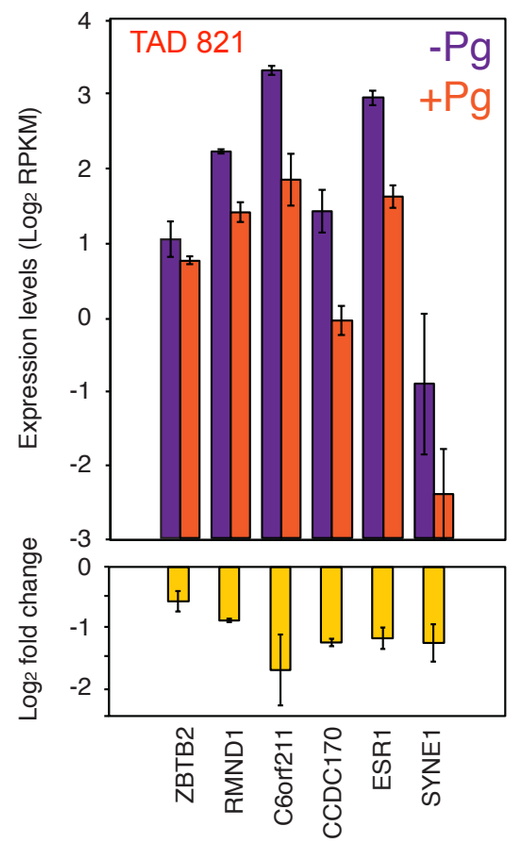


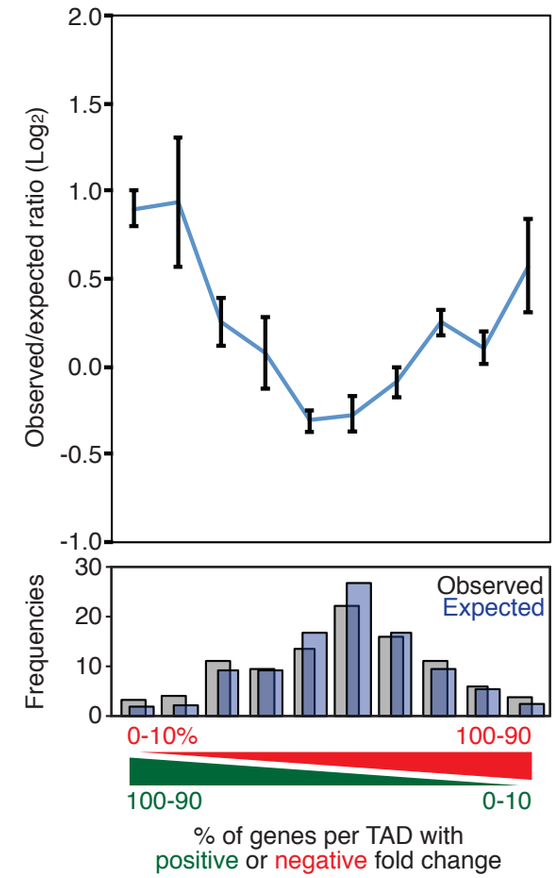
Are TADs homogeneous?



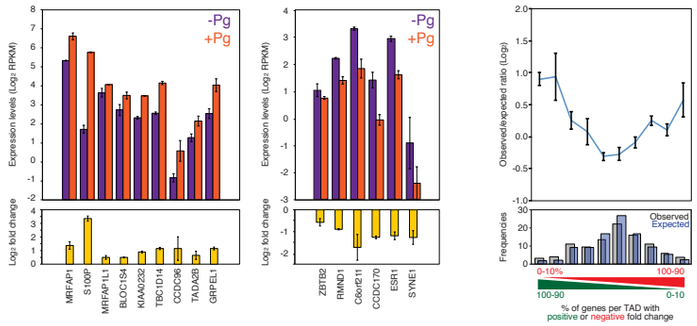
Do TADs respond differently to Pg treatment?



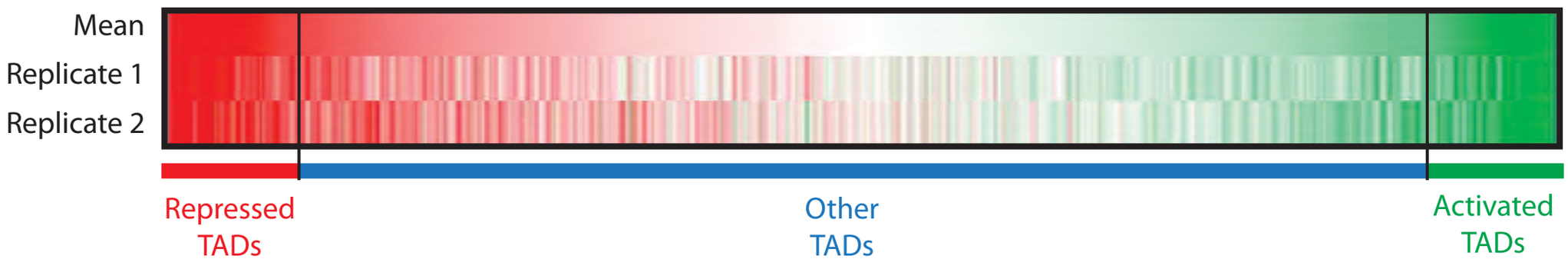


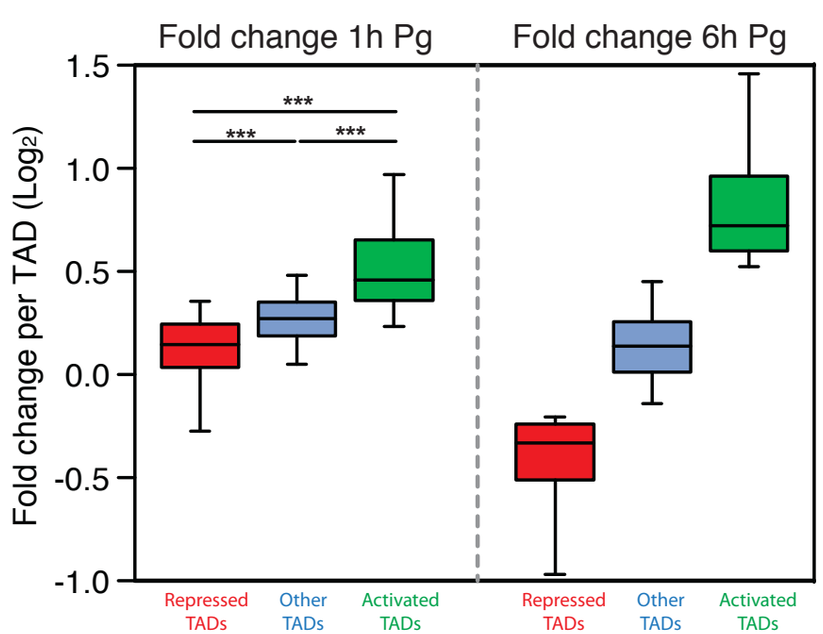


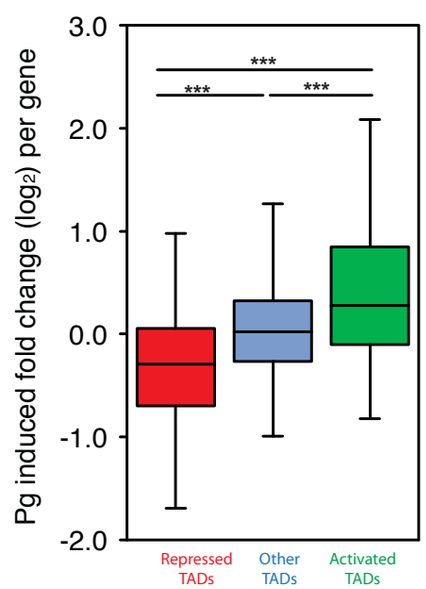
Do TADs respond differently to Pg treatment?

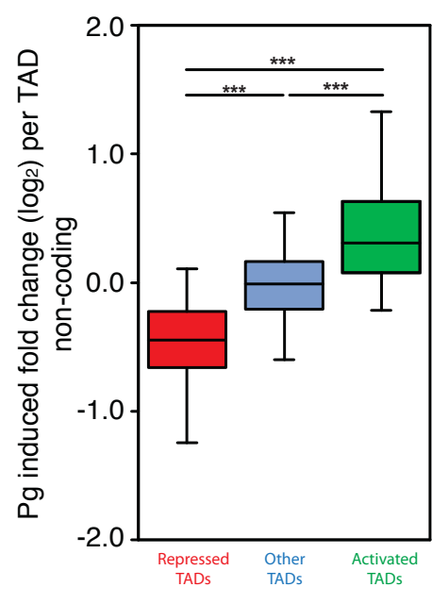


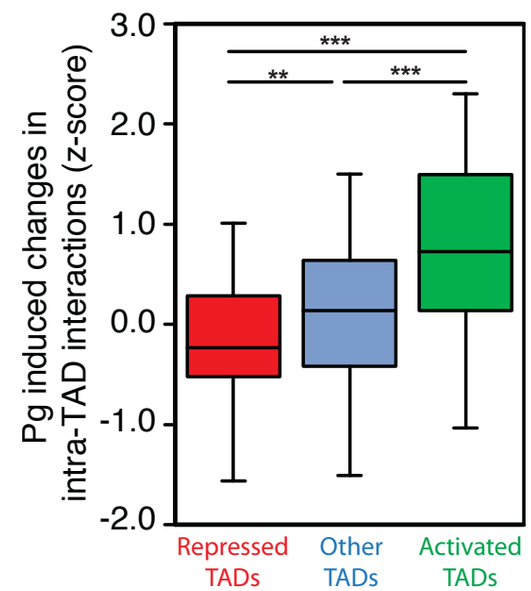
Pg induced fold change per TAD (6h)



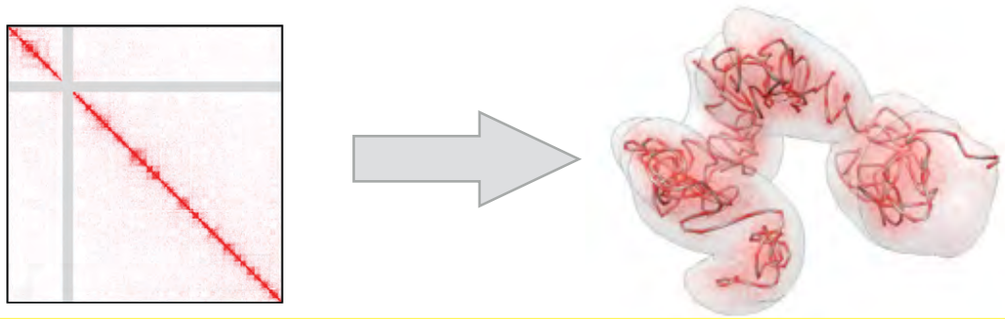




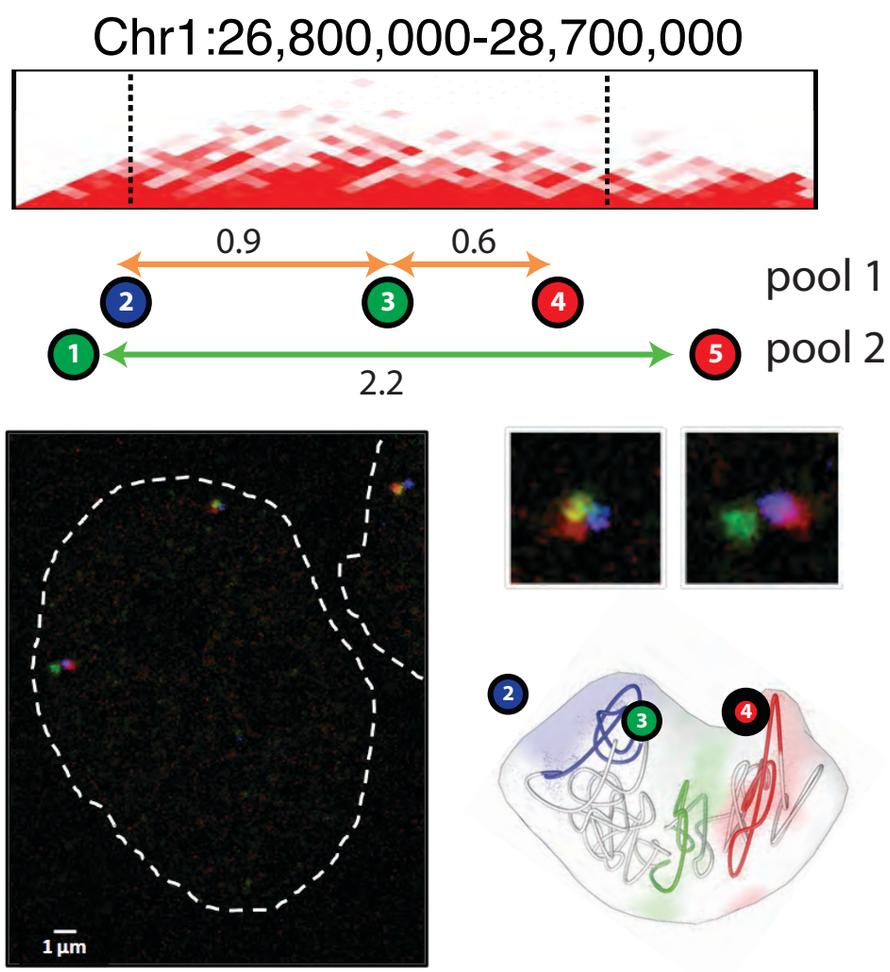


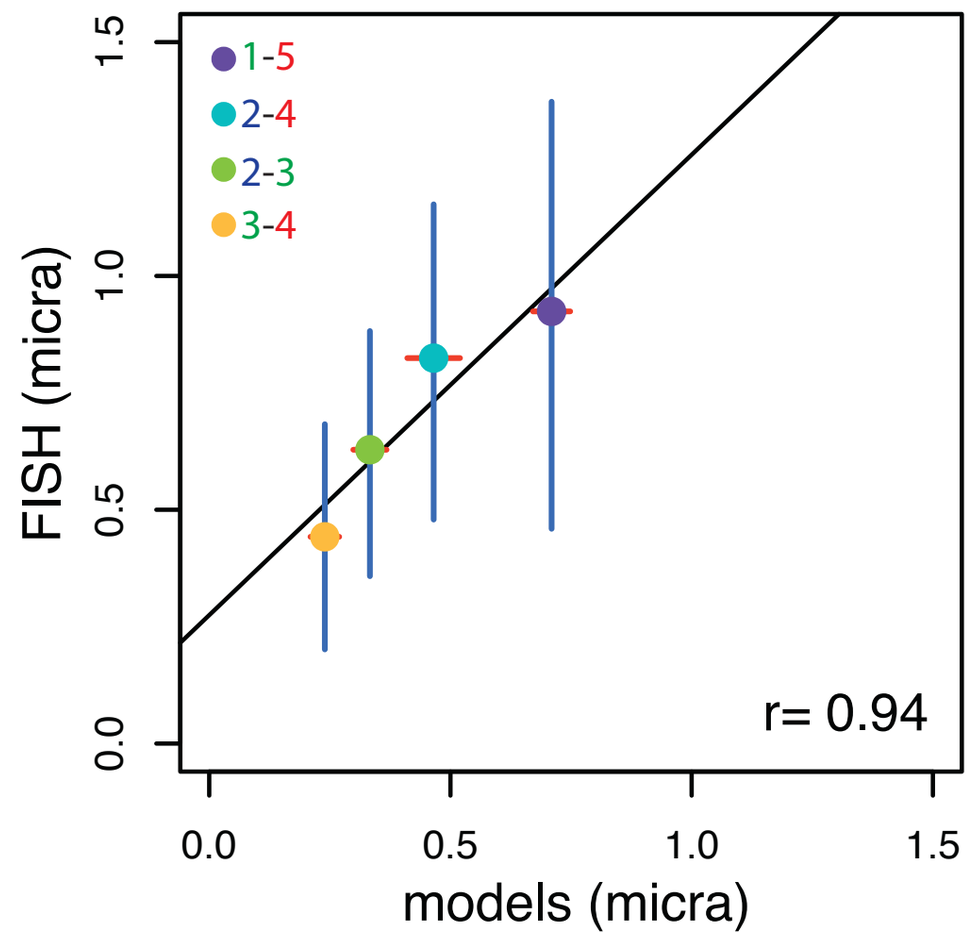


Modeling 3D TADs

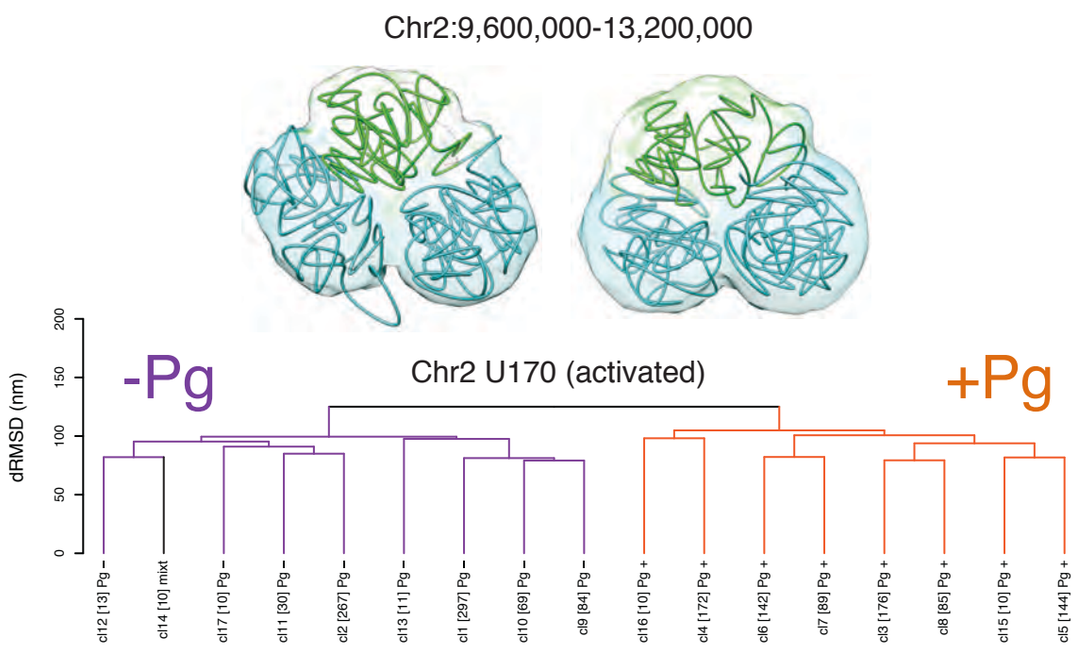


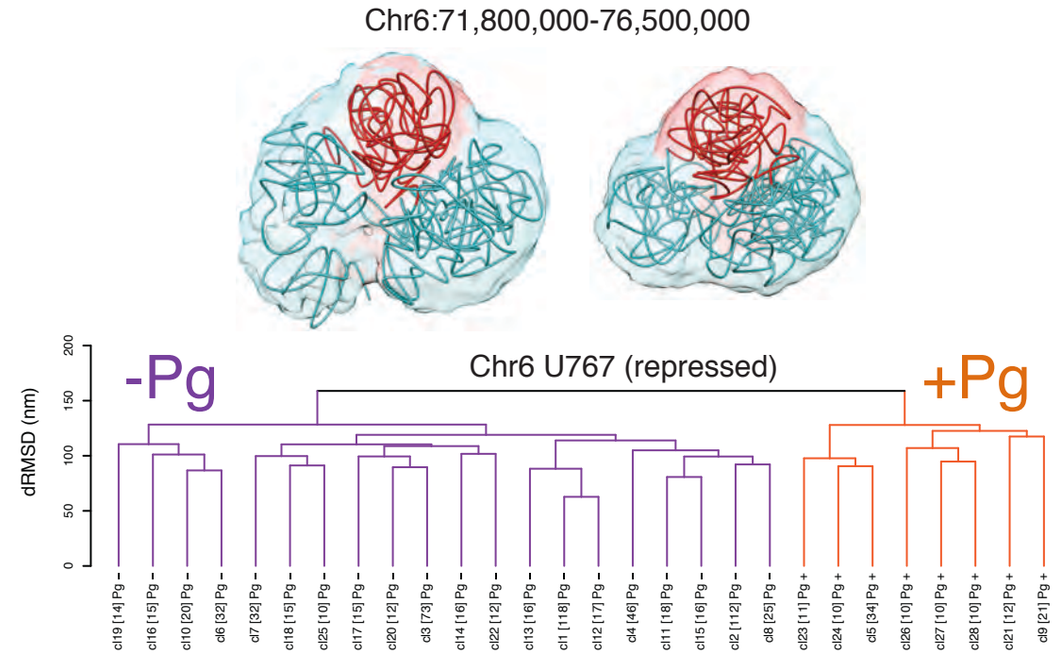
61 genomic regions containing 209 TADs covering 267Mb

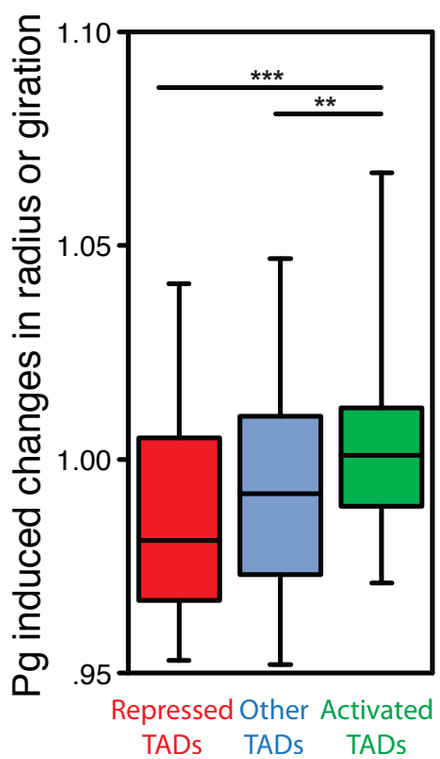


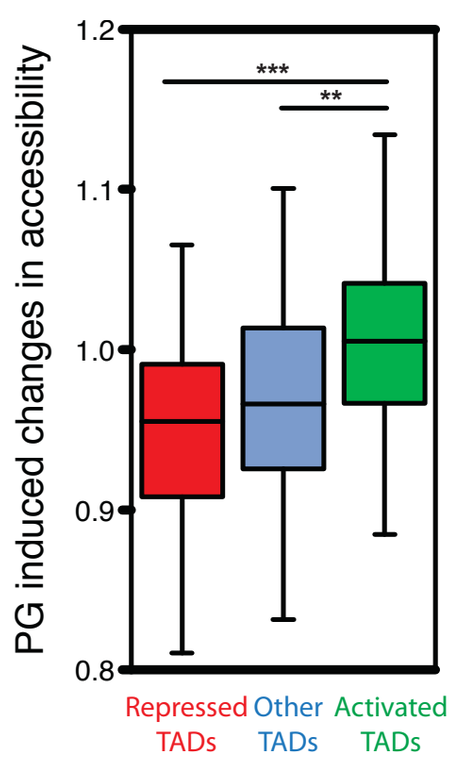


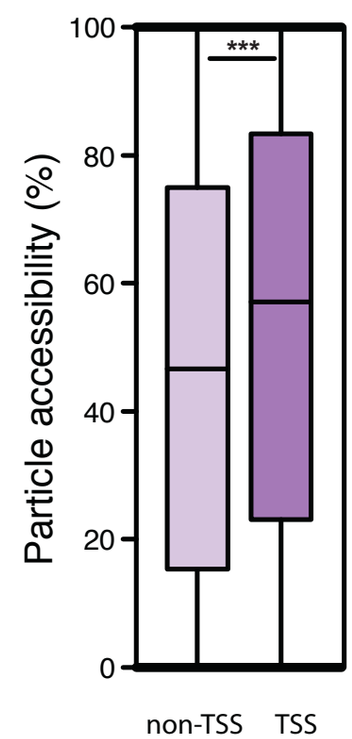
How TADs respond structurally to Pg?





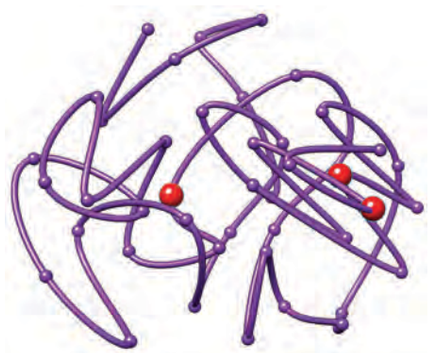


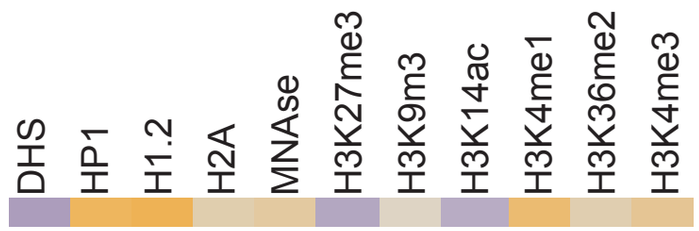


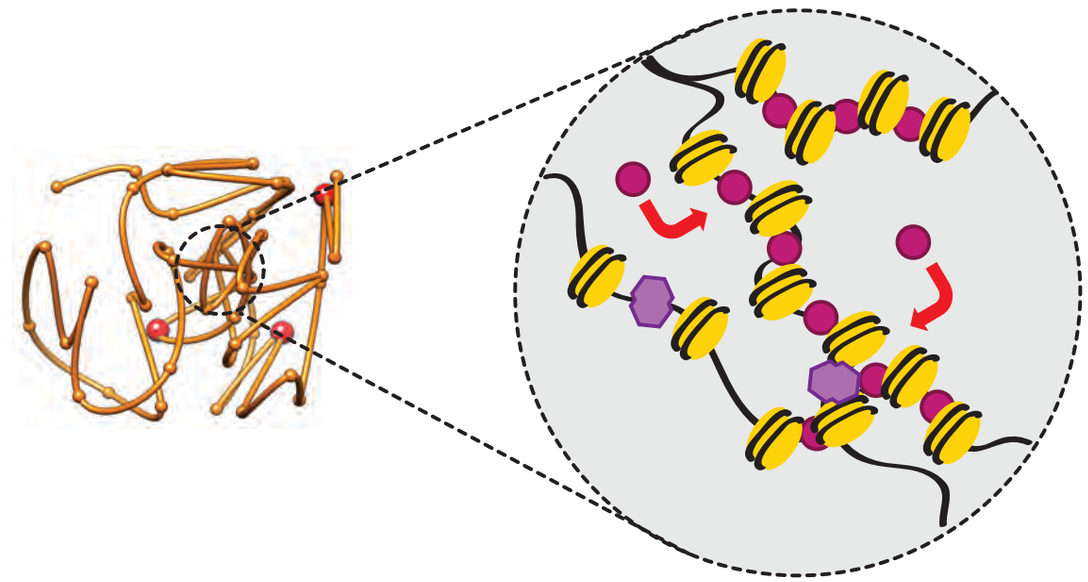


Model for TAD regulation

Repressed TAD
chr1 U41

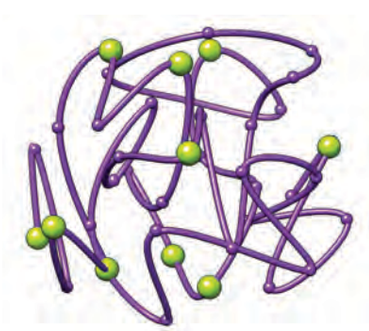


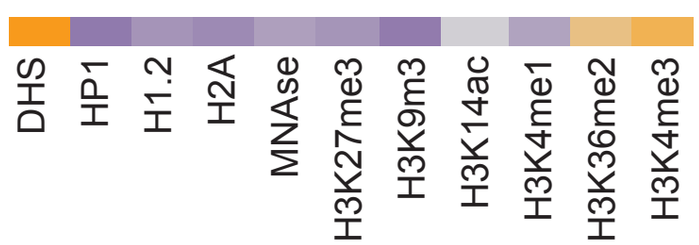


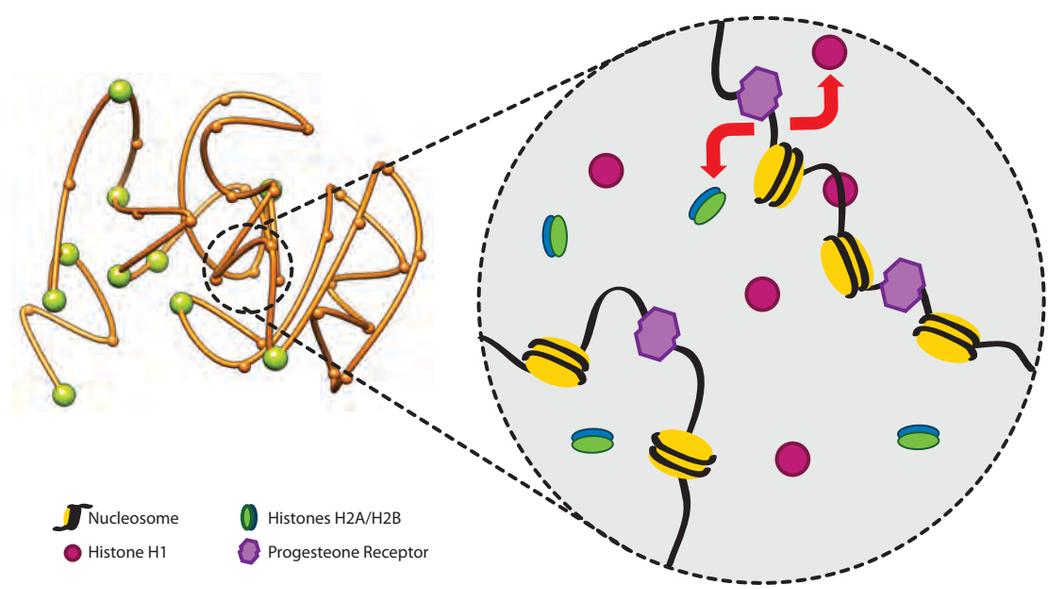


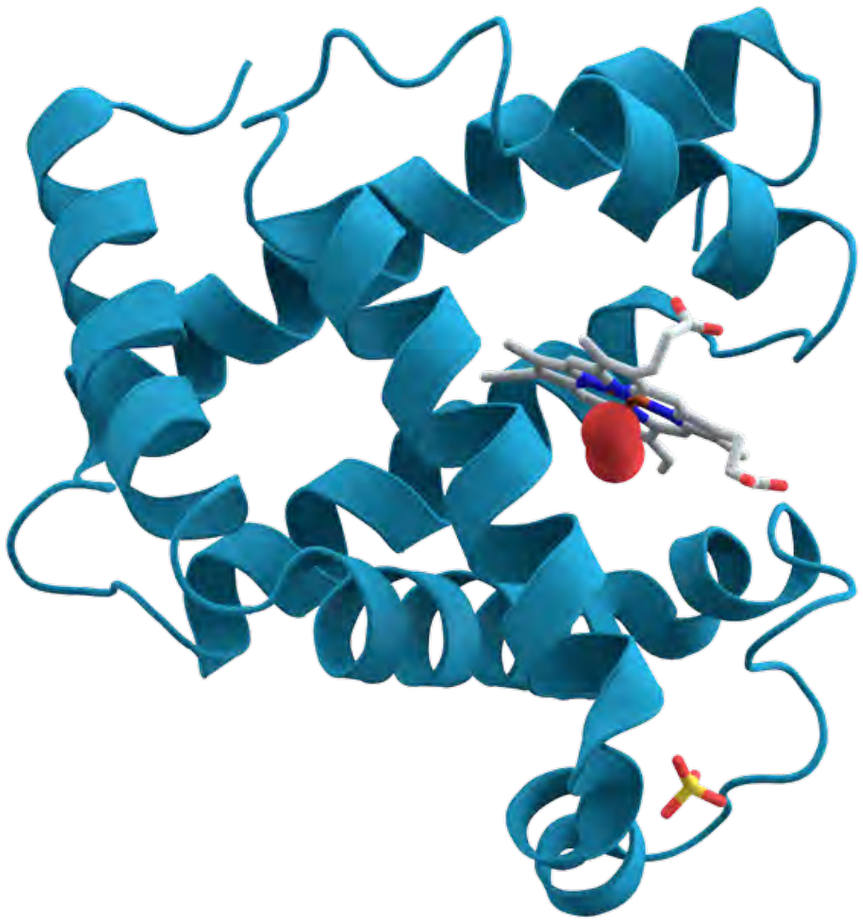
Structural transition **+Pg**

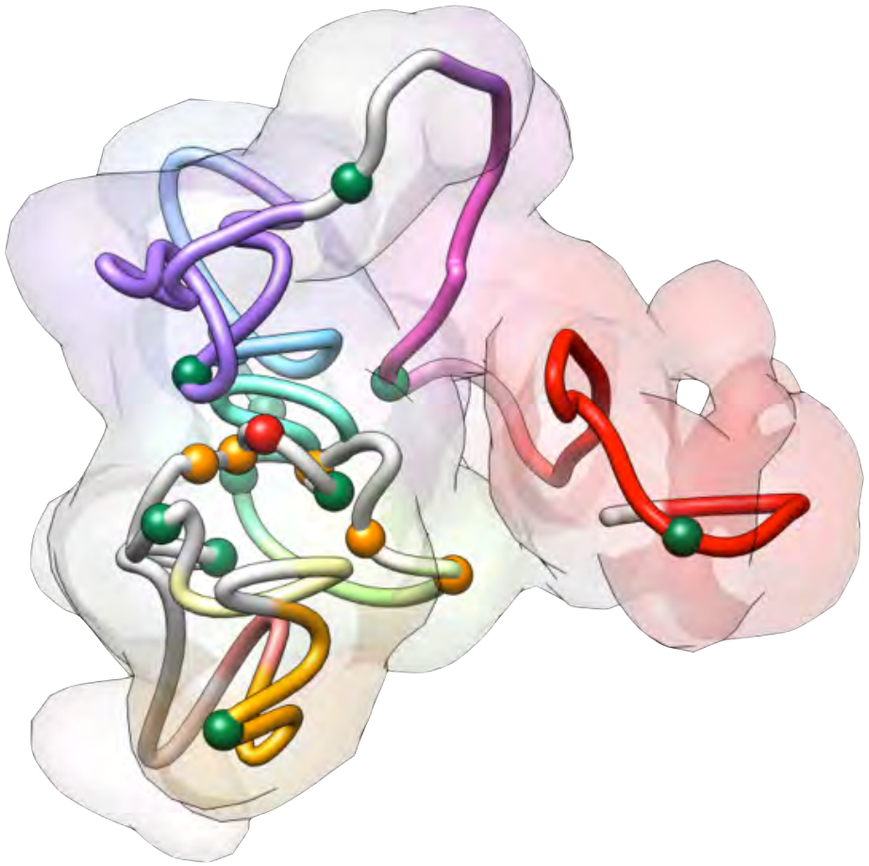
Activated TAD
chr2 U207





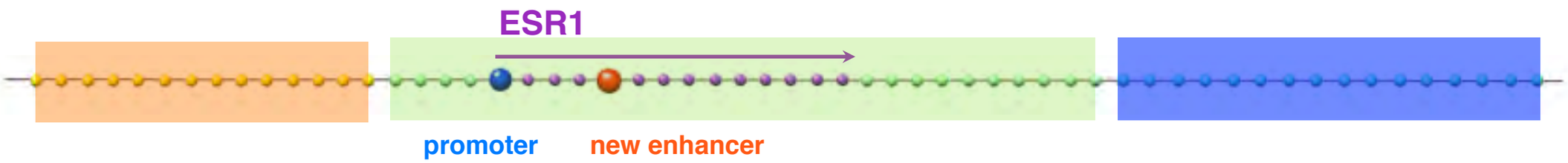


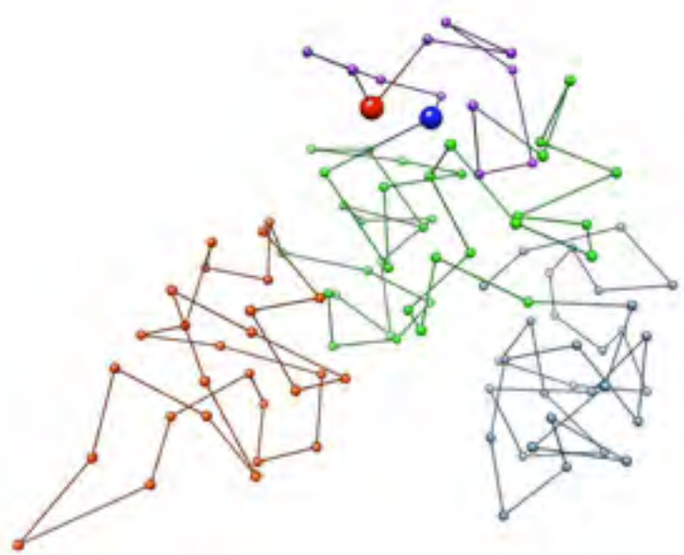




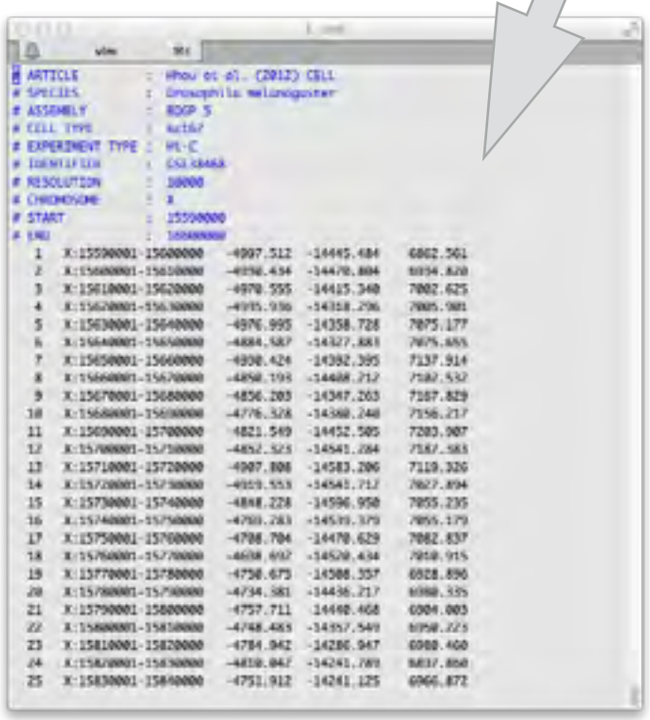
STRUCTURE ⟷ FUNCTION

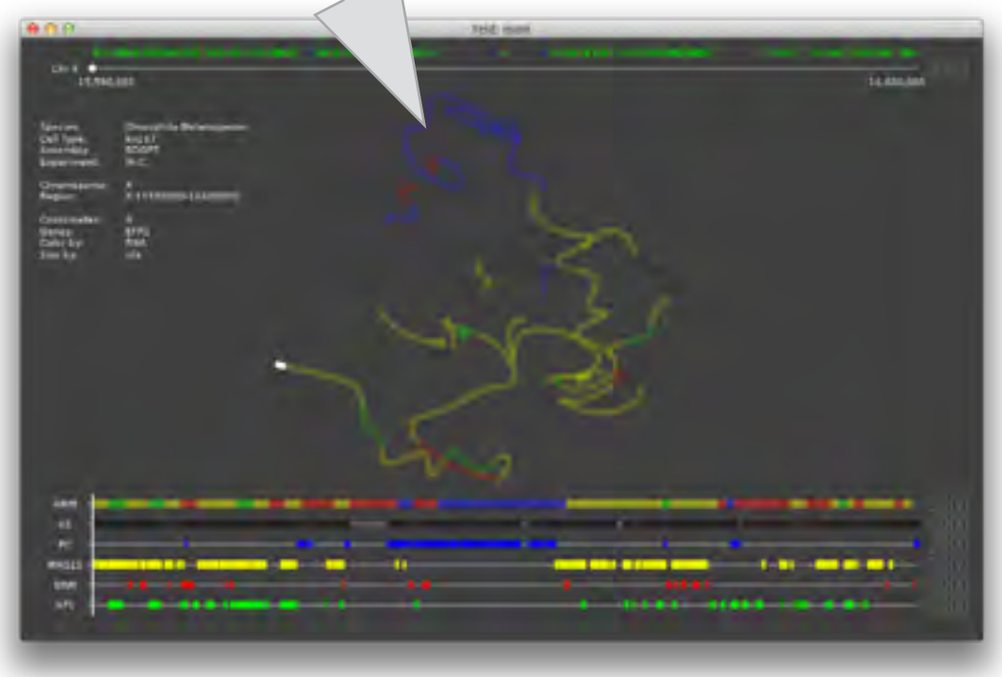
Structure >> Function!





http://3DGenomes.org





Acknowledgments

Davide Baù
François le Dily
François Serra

David Dufour
Mike Goodstadt
Gireesh Bogu
Francisco Martínez-Jiménez

Job Dekker
Program in Systems Biology
Department of Biochemistry and Molecular Pharmacology
University of Massachusetts Medical School
Worcester, MA, USA

Kerstin Bystricky
Chromatin and gene expression
Laboratoire de Biologie Moléculaire Eucaryote - CNRS
Toulouse, France

Miguel Beato & Guillaume Filion
Gene Regulation, Stem Cells and Cancer
Centre de Regulació Genòmica
Barcelona, Spain

http://marciuslab.org
http://3DGenomes.org
http://cnag.cat · http://crg.cat