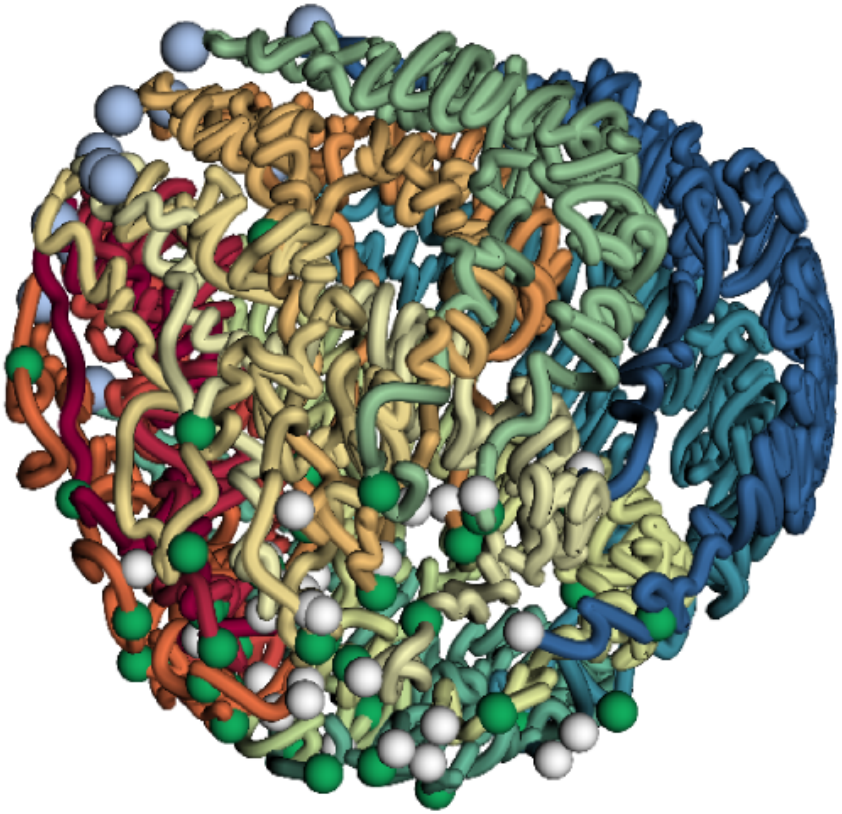
The 3D Genome



Jean-Philippe Vert

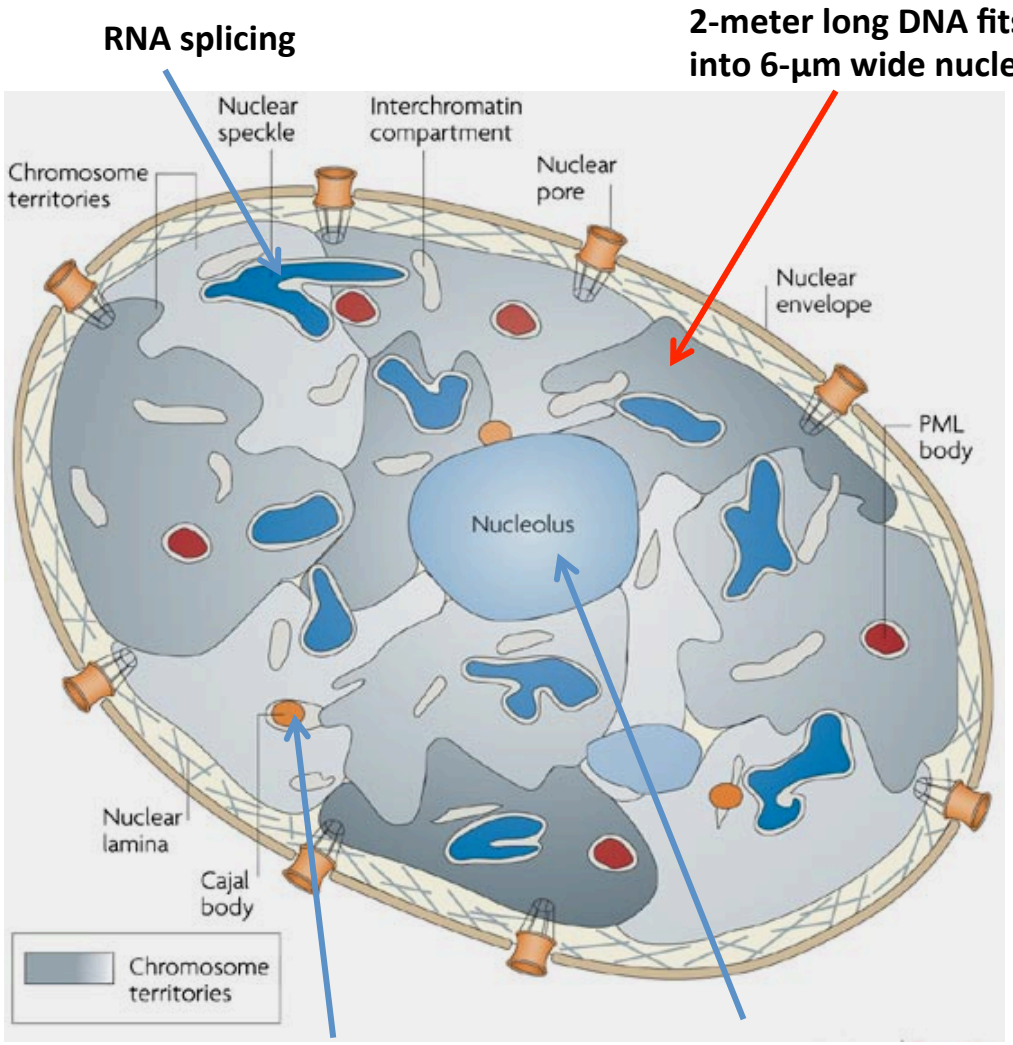
MINES ParisTech

institutCurie
Together, let's beat cancer.

1794
ENS

PSL

How does genome architecture influence genome function?



- Nuclear compartmentalization
- Nuclear lamina
- Transcription factories
- Chromosome conformation
  - Long-range looping
  - Chromatin domains
  - Chromosome territories

Lanctot et al. *Nature Rev. Genetics* 2007

Tools for capturing chromosome conformation

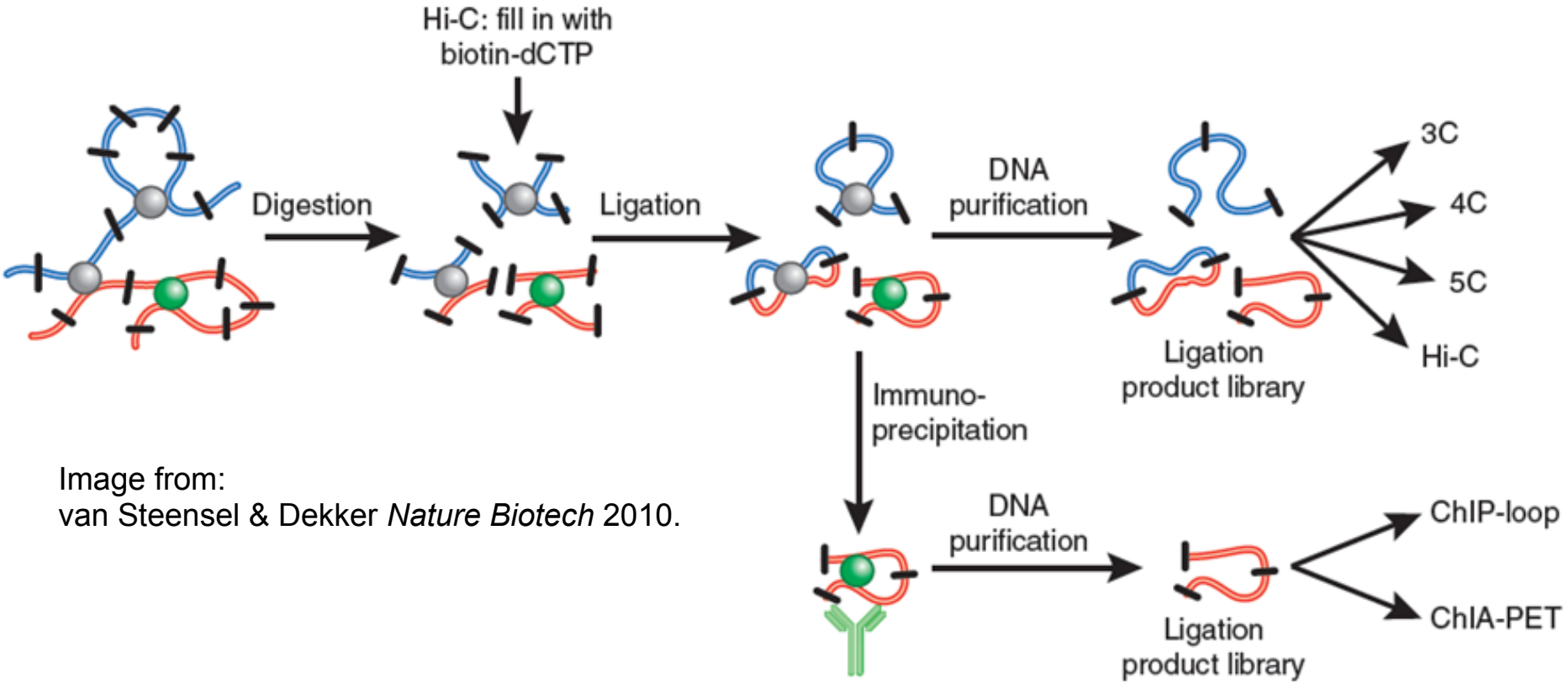
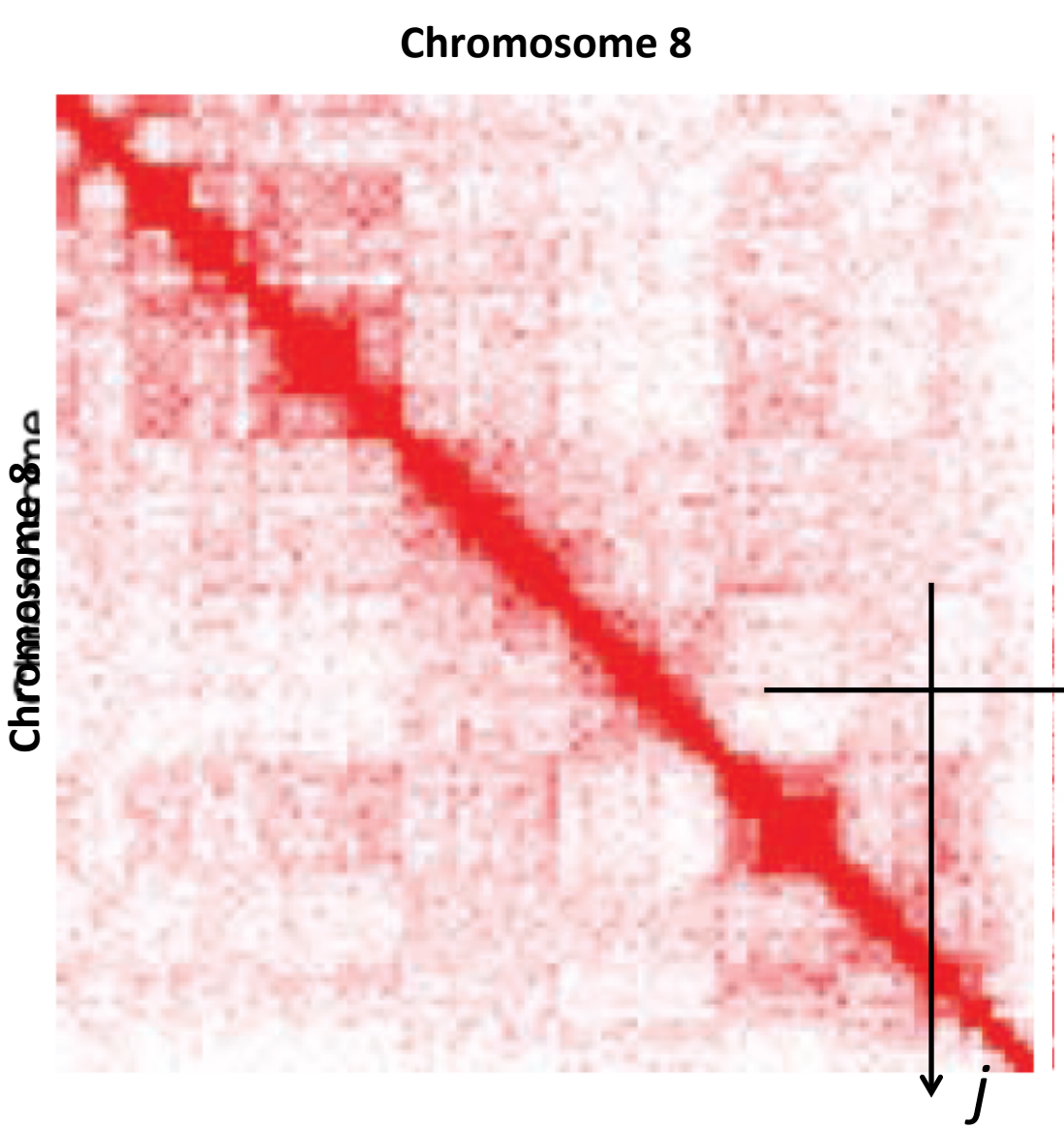


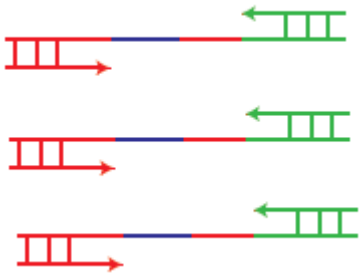
Image from:
van Steensel & Dekker *Nature Biotech* 2010.

3C	ChIA-PET	Hi-C	
Dekker et al. *Science* 2002	Fullwood et al. *Nature* 2009	L.-Aiden et al. *Science* 2009	Duan et al. *Nature* 2010

Output of conformation capture is a contact matrix



paired-end reads

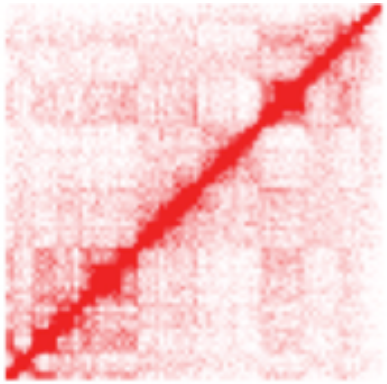


C(i,j) = How many times locus ***i*** is linked to locus ***j*** by a paired-end read?

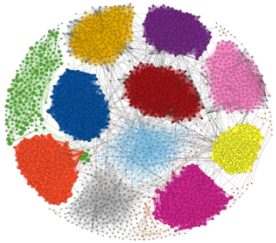
Inter-chromosomal contact

Intra-chromosomal contact

The many uses of Hi-C

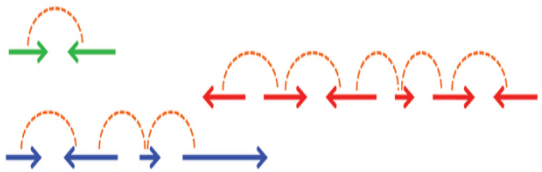


Lieberman-Aiden, *et. al*.
Science, 2009



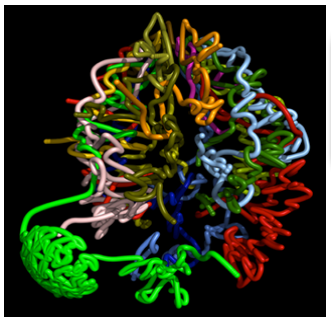
Organismal Deconvolution

Burton, Liachko, *et al*. G3, 2014



Genome scaffolding

Burton, *et al*.
Nature Biotech, 2013



3D model of genome

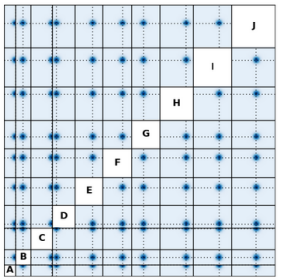
Duan, *et al*. Nature, 2010 *(S. cerevisae),*
Ay, *et al.* Genome Res., 2014a *(P.falciparum)*

Enhancer

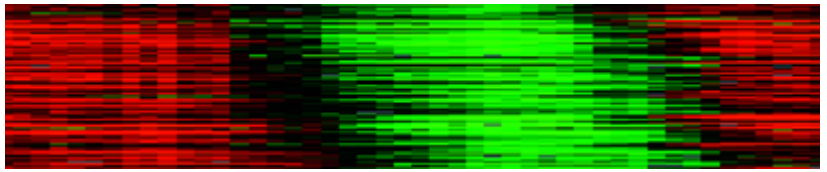
Long-range chromatin contacts

Ay, *et al*. Genome Res., 2014b

Promoter



Centromere calling

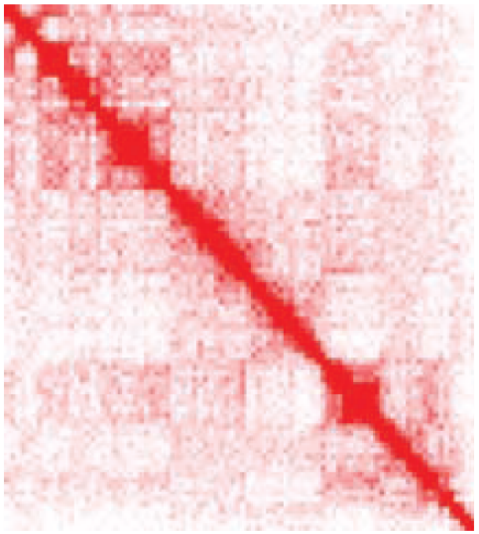


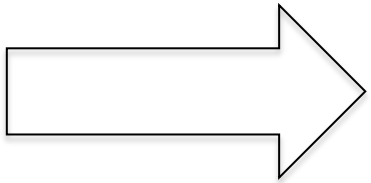
Ay*, Bunnik*, Varoquaux* et al. *Genome Research,* 2014.

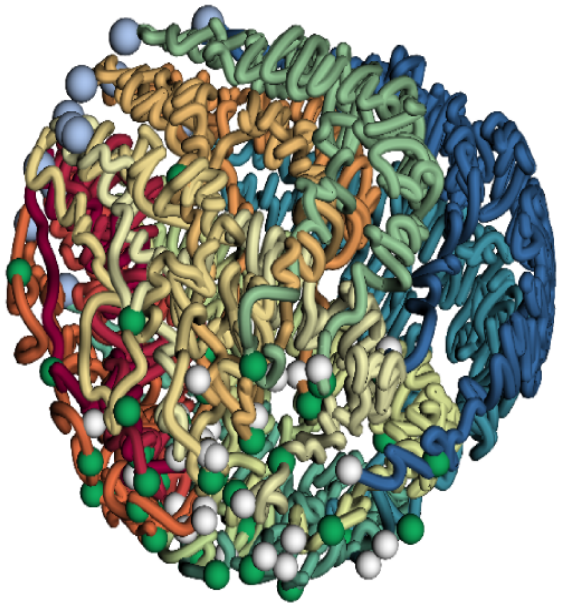
Regulation of expression

Part 1

3D model of the genome







Reconstructing the 3D structure of the genome from Hi-C data

Two main approaches:

1. **Consensus methods** that infer a unique « average » structure

   [Duan et al. 2010; Tanizawa et al. 2010; Bau et al. 2011; Zhang et al. 2013; Ben-Elazar et al. 2013]

2. **Ensemble methods** that yield a population of structures

   [Rousseau et al. 2011; Khalor et al 2011; Hu et al 2013]

Modelisation

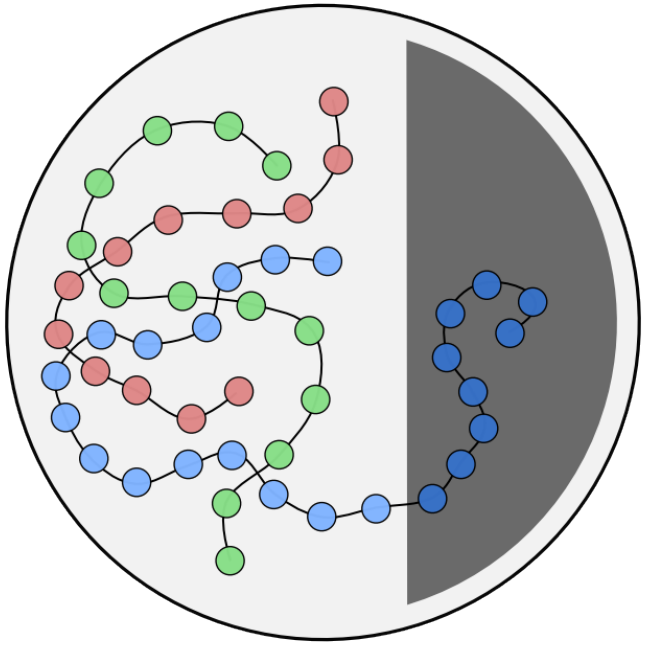
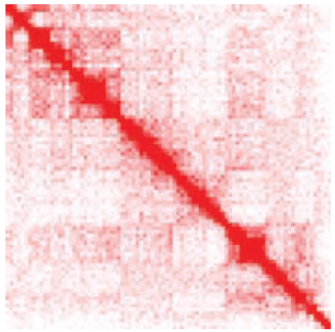


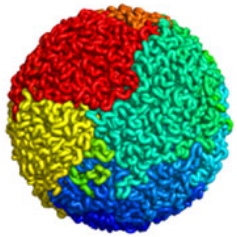
Figure: **Beads on a string model** Chromosomes are modeled as a series of beads. Nucleus is assumed to be spherical.

- Chromosomes are modeled as a series of beads.
- Each bead is spaced 10kb apart.

From interaction frequency to 3D distance (some physics...)

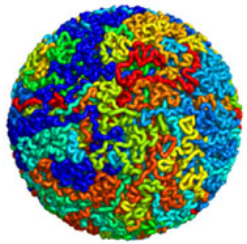


Fractal globule



- $c \sim s^{-1}$
- $d \sim s^{1/3}$
- Valid for human.

Equilibrium

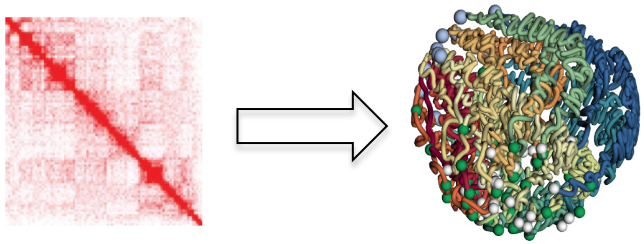


- $c \sim s^{-3/2}$
- $d \sim s^{1/2}$ for $s < s_{\max}^{2/3}$
- Valid for budding yeast, and small organism

Default counts-to-distance transfer function

$$\delta_{ij} = \gamma c_{ij}^{-1/3}, \tag{6}$$

Metric MDS-based method



- Let $\mathbf{X} \in R^{n \times 3}$ be the coordinates of each bead.
- Let $\mathbf{C} \in R^{n \times n}$ be the contact count matrix and $\mathcal{D}$ the set of non-zeros entries.
- Let Θ the count-to-distance transfer function.

Optimization problem

$$\begin{aligned} \underset{\mathbf{x}_1,\ldots,\mathbf{x}_n}{\text{minimize}} \quad & \sigma(\mathbf{X}, \mathbf{C})) \\ \text{subject to} \quad & \text{some bio} \\ & \mathbf{x}_i^T \mathbf{x}_i \leq r \end{aligned}$$

- MDS1 [Duan et al., 2010]

$$\sigma(X, C) = \sum_{i,j} (\|x_i - x_j\|_2 - \Theta(c_{ij}))^2$$

- MDS2 [Ay et al., 2014]

$$\sigma(X, C) = \sum_{i,j} \frac{(\|x_i - x_j\|_2 - \Theta(c_{ij}))^2}{\Theta(c_{ij})^2}$$

- ChromSDE [Zhang et al., 2013]

$$\sigma(X, C) = \sum_{i,j} \frac{(\|x_i - x_j\|_2^2 - \Theta(c_{ij})^2)^2}{\Theta(c_{ij})} - \lambda \sum_{i,j \notin \mathcal{D}} \|x_i - x_j\|_2^2$$

Nonmetric MDS-based method

Idea

If two loci i and j are observed to be in contact more often than loci k and ℓ, then i and j should be closer in 3D space than k and ℓ

$$c_{ij} \geq c_{k\ell} \Leftrightarrow \|x_i - x_j\|_2 \leq \|x_k - x_\ell\|_2 \tag{4}$$

$$\begin{aligned} \underset{\mathbf{x}_1,\ldots,\mathbf{x}_n,\Theta}{\text{minimize}} \quad & \sigma(\mathbf{X}, \mathbf{C}, \Theta)) \\ \text{subject to} \quad & \Theta \quad \text{decreasing} \\ & \text{some biologically motivated constraints} \\ & \mathbf{x}_i^T \mathbf{x}_i \leq r_{\max}^2, \quad \text{(all beads should lie in the nucleus)} \end{aligned}$$

(Ben-Eleazar et al, 2013)

Poisson model

The idea

Let's assume that $c \sim Poisson(\beta d^{\alpha})$, where c is the interaction count, d the euclidean distance, and β and α unknown parameter.

Likelihood

$$\ell(\mathbf{X}, \alpha, \beta) = \prod_{i<j\leq n} \frac{(\beta d_{ij}^{\alpha})^{c_{ij}}}{c_{ij}!} exp(-\beta d_j^{\alpha}) \tag{5}$$

Optimization problem

$$\underset{\mathbf{x}_1,\ldots,\mathbf{x}_n,\alpha,\beta}{\text{minimize}} \quad \sigma(\mathbf{X}, \mathbf{C}, \alpha, \beta) = -\log(\ell(\mathbf{X}, \mathbf{C}, \alpha, \beta))$$

subject to some biologically motivated constraints

$$\mathbf{x}_i^T \mathbf{x}_i \leq r_{\max}^2, \quad \text{(all beads should lie in the nucleus)}$$

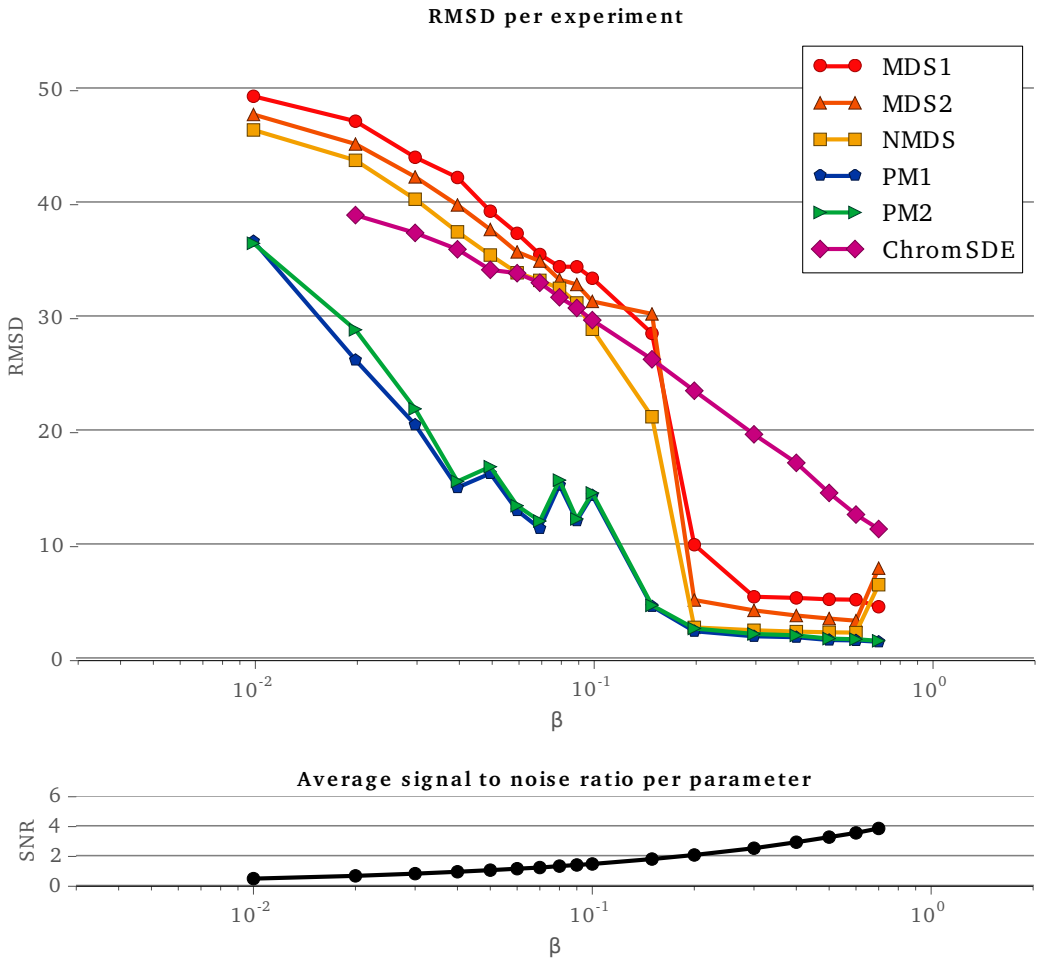
Data

- **Generated Datasets**

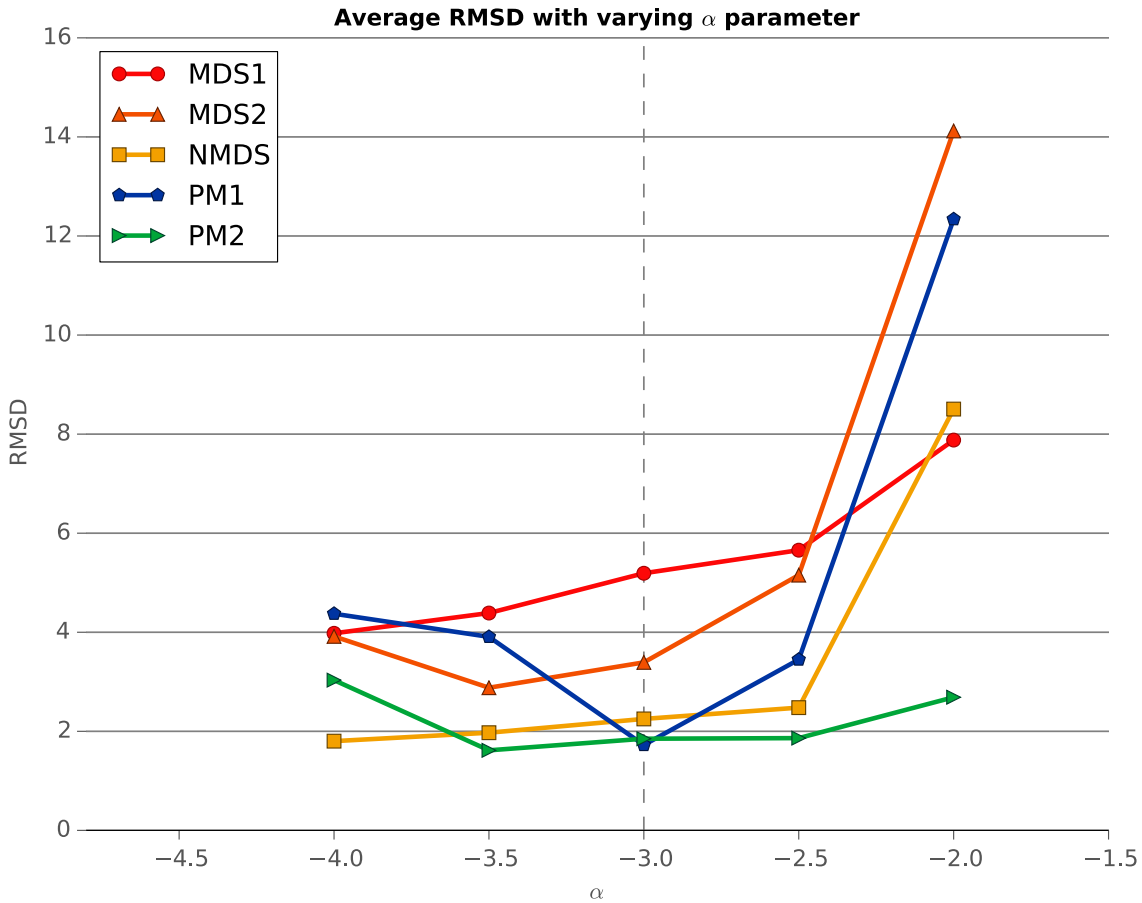
$$c_{ij} = P(\beta d_{ij}^{\alpha}), \tag{7}$$

  where
  - $\alpha = -3$ and β varies between 0.01 and 0.7.
  - α varies between -4 and -2 and β between 0.4 and 0.7.
- **Publicly available datasets** mouse embryonic stemcells at 100 kb, 200 kb, 500 kb, 1 Mb, normalized using ICE [Imakaev et al., 2012]

Performance as a function of coverage



Robustness to parameter misspecification



Mouse embryonic stem cells

- Stability across enzyme replicates

Resolution	1 Mb		500 kb		200 kb		100 kb	
	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
MDS1	13.13	0.945	10.00	0.942	5.64	0.940	5.07	0.736
MDS2	5.54	0.964	5.68	0.959	3.74	0.945	2.53	0.676
NMDS	5.80	0.965	5.67	0.959	3.73	0.946	2.52	0.666
PM1	7.28	0.931	7.14	0.913	4.01	0.891	**2.51**	0.664
PM2	**4.92**	**0.976**	**4.66**	**0.968**	**3.42**	**0.958**	2.76	**0.771**

- Stability across resolution

	MDS1	MDS2	NMDS	PM1	PM2
RMSD	14.86	12.92	12.98	13.03	**11.48**
Correlation	0.781	0.754	0.738	0.737	**0.807**

Try it?

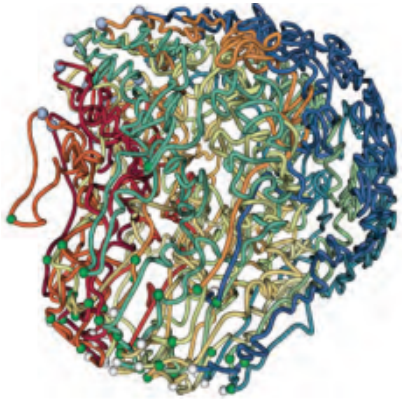
- http://cbio.mines-paristech.fr/pastis

- `$ pip install --user pastis`

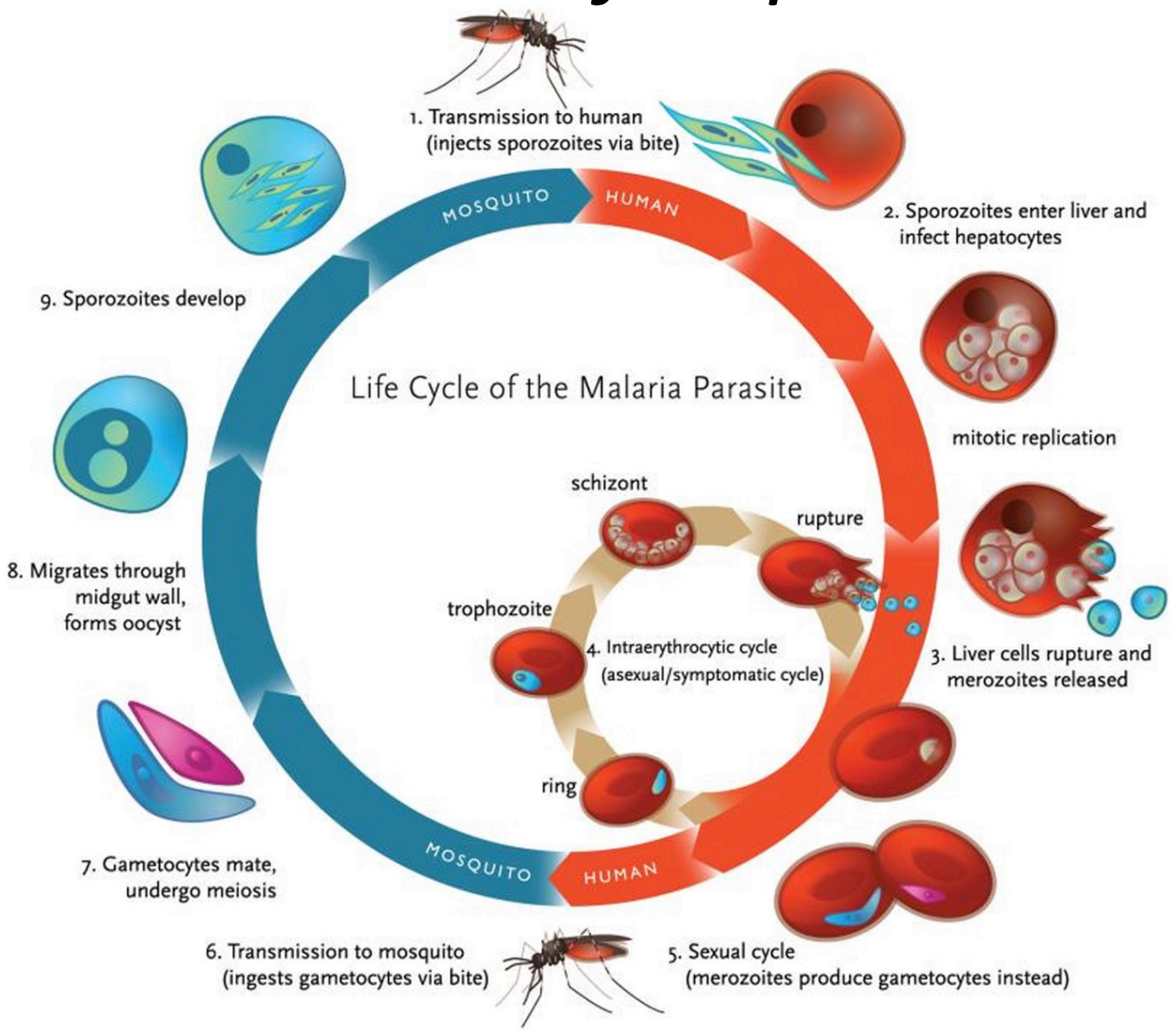
- N. Varoquaux, F. Ay, W. S. Noble and J.-P. Vert, "A statistical approach for inferring the three-dimensional structure of the genome », B*ioinformatics*, 30(12):i26-i33, 2014.

Part 2

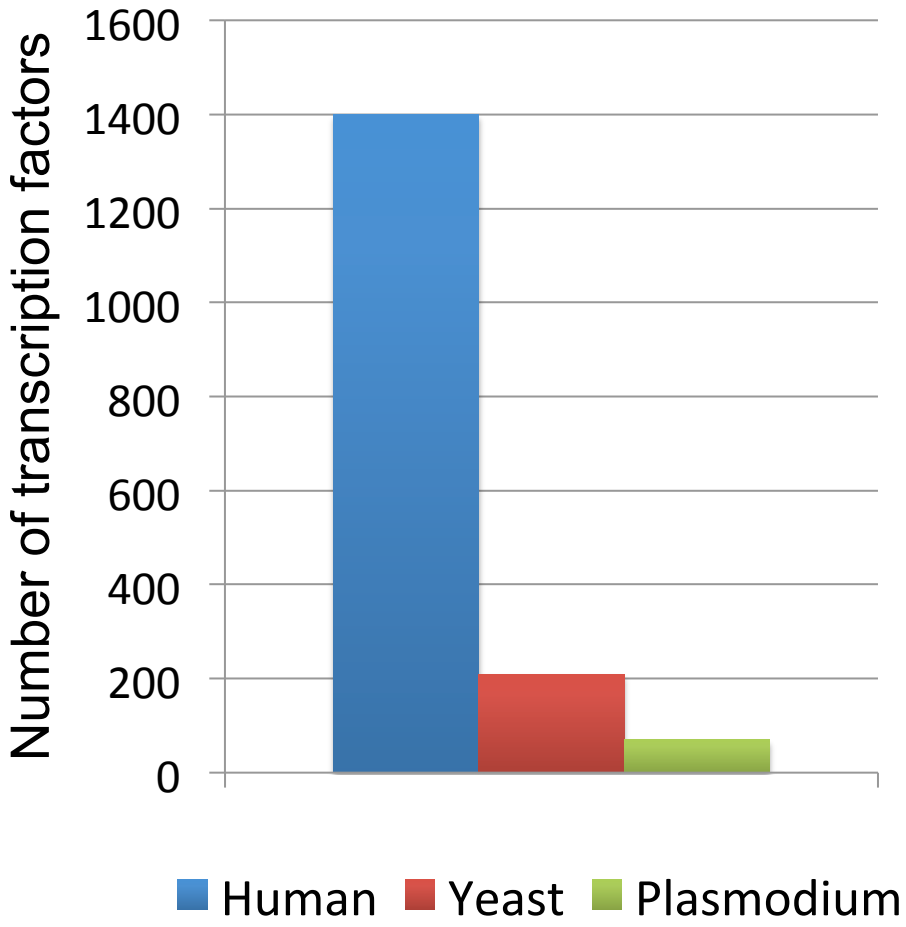
The spatial organization of the *P. falciparum* genome



The human malaria parasite *Plasmodium falciparum*



How *Plasmodium* regulates gene expression is mysterious



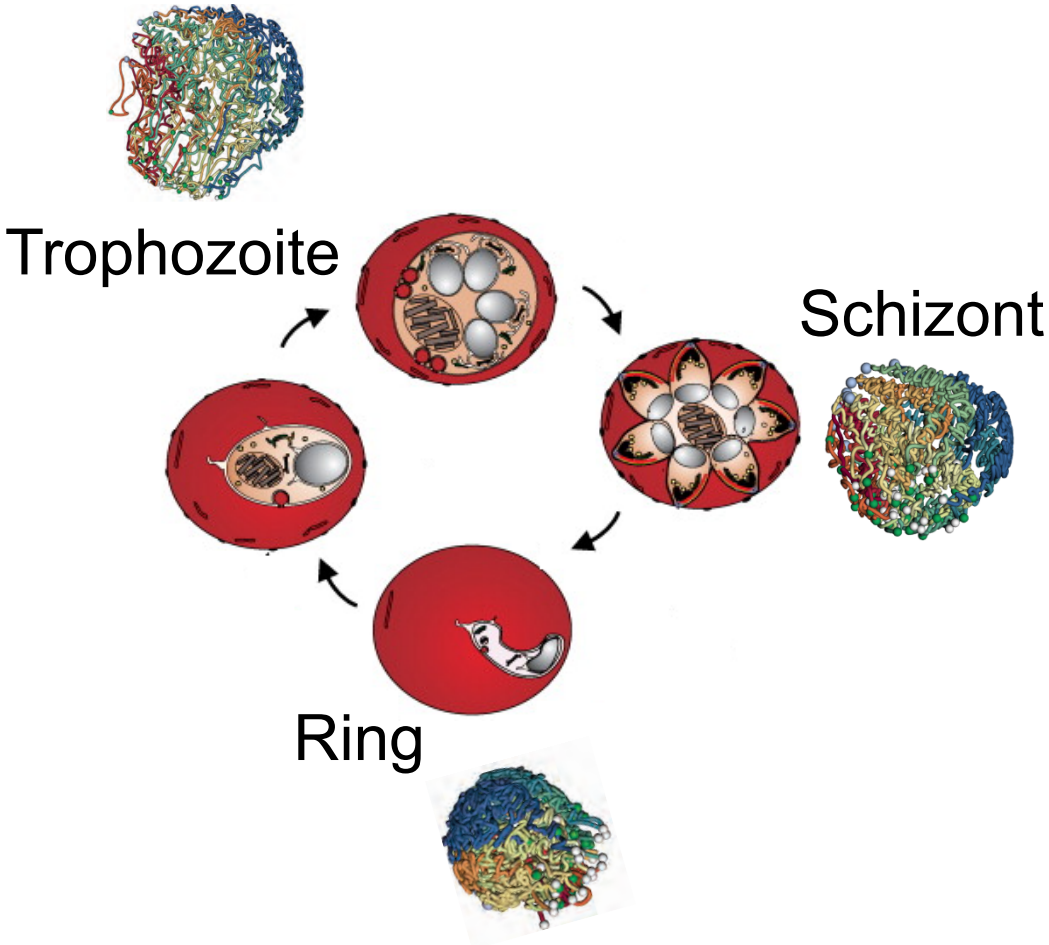
Very few transcription factors.

- 27 ApiAP2 plant-like TFs
  (Balaji et al. *NAR* 2005)
- 71 hits from homolog protein sequence search using HMMER

(Coulson et al. *Genome Research* 2004)

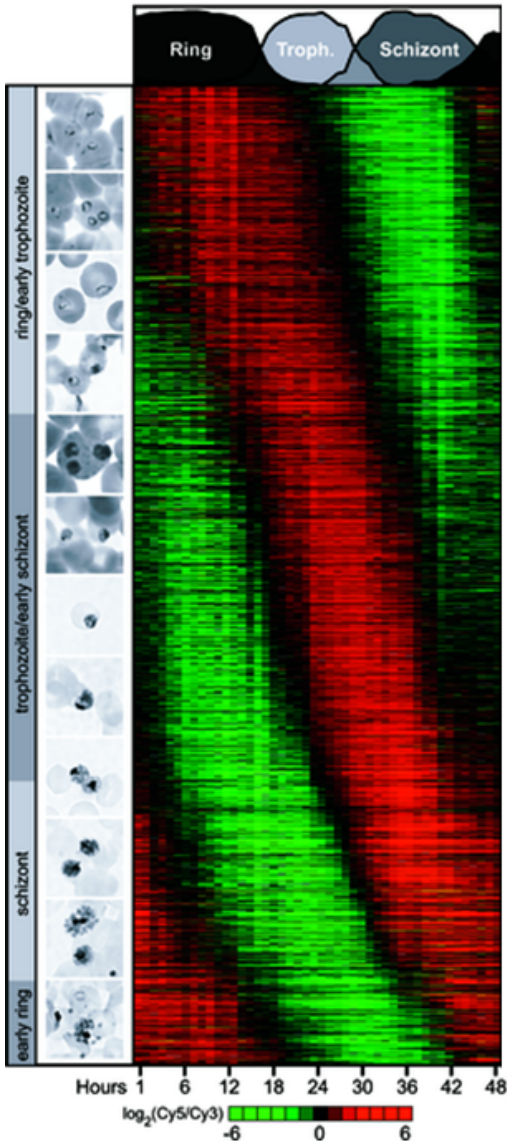
Hughes & de Boer *Genetics* 2013.

Genome architecture as an alternative mechanism for regulating gene expression?

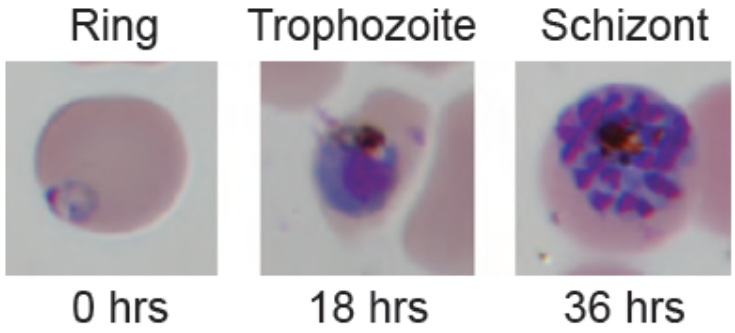


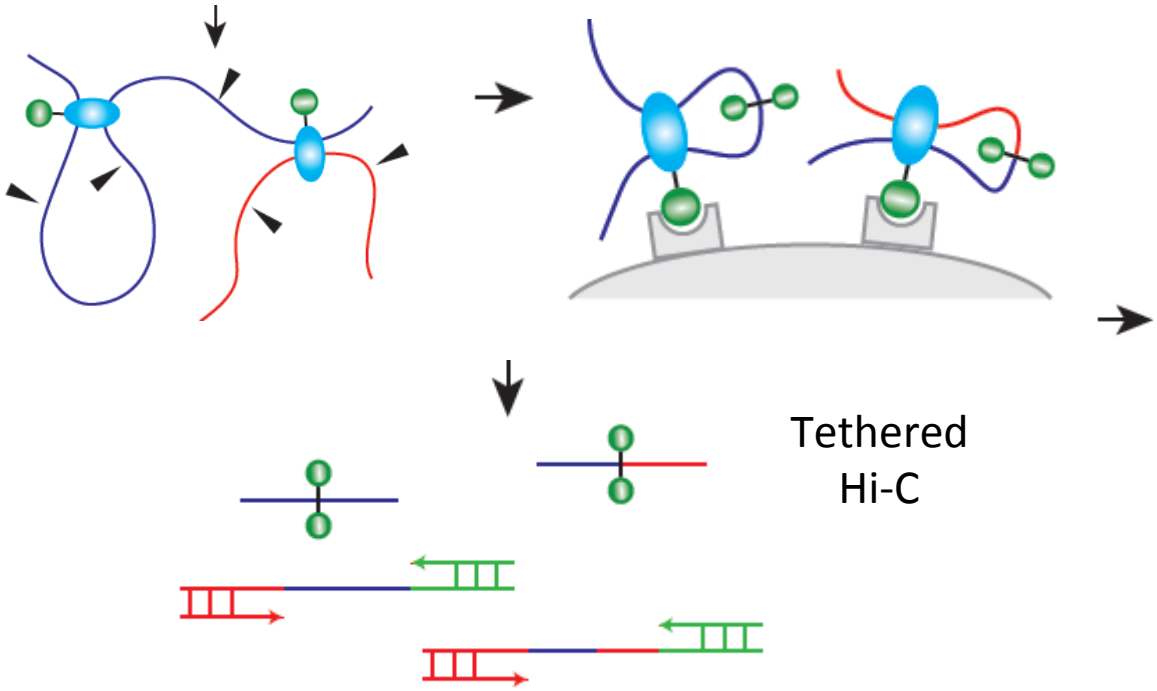
Erythrocytic cycle

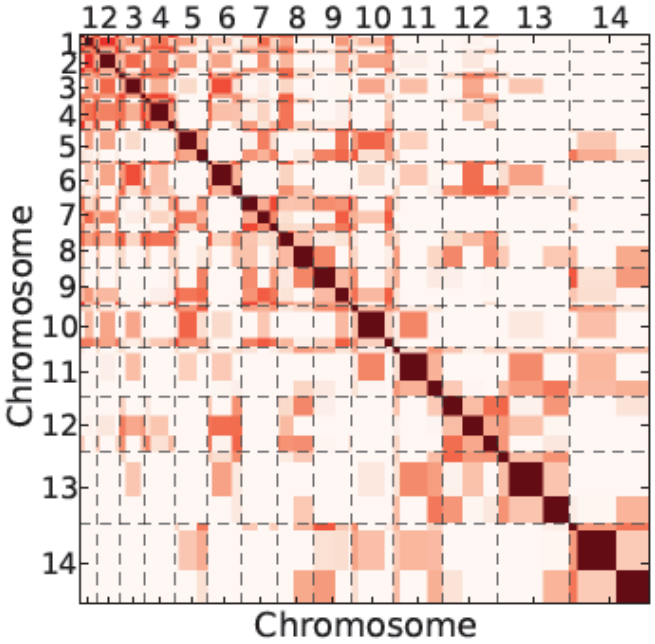
?



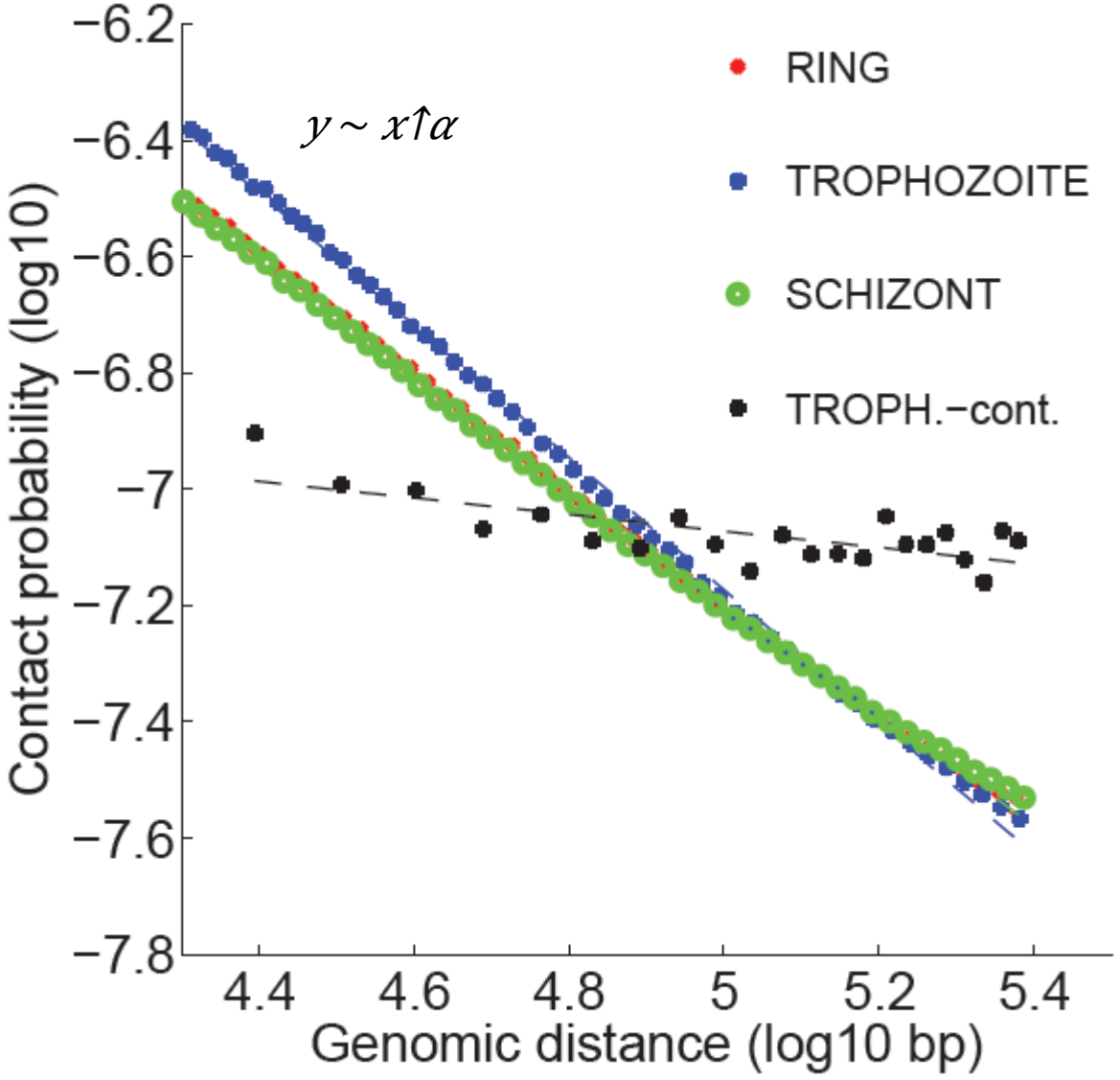
We assayed genome architecture at 3 time points in the erythrocytic cycle







Plasmodium contact frequencies suggest a fractal globule architecture



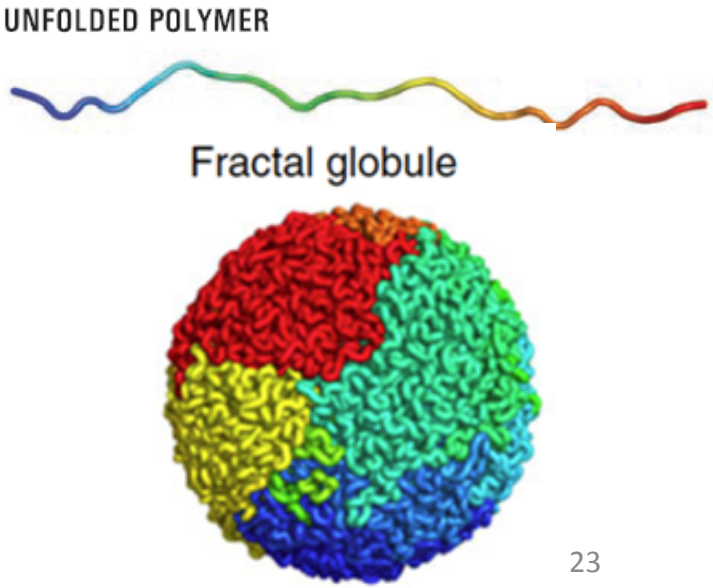
$\alpha =$

-0.98

-1.14

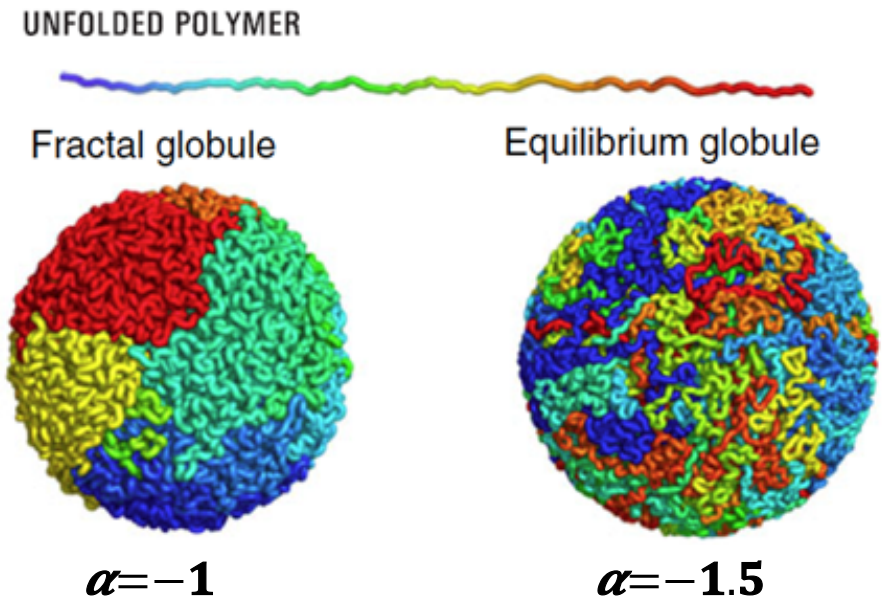
-0.96

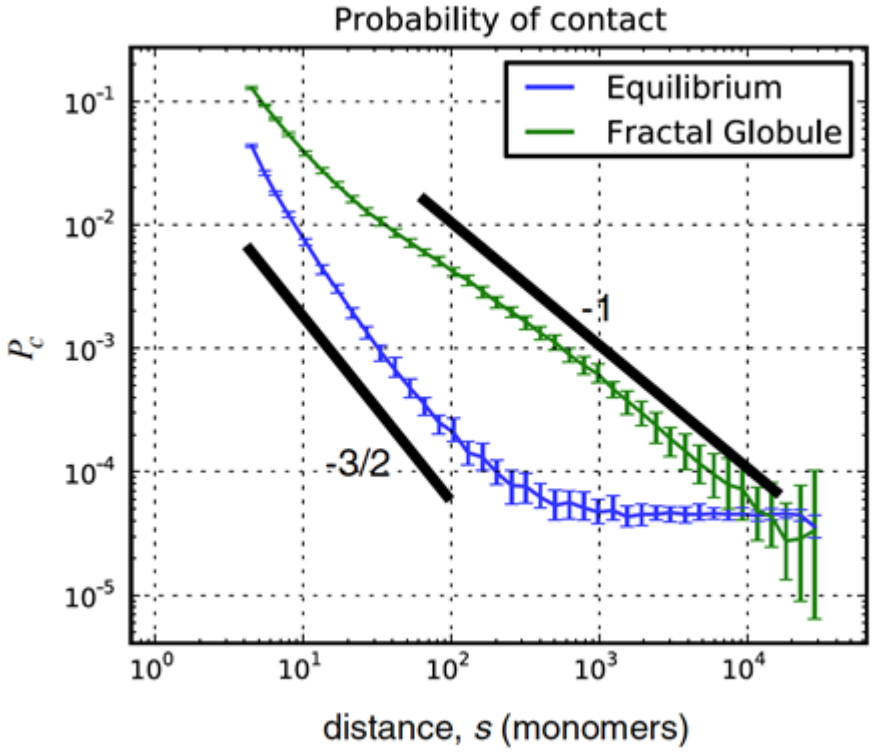
-0.14



Scaling parameter for the Trophozoite stage is indicative of more intermingled chromatin

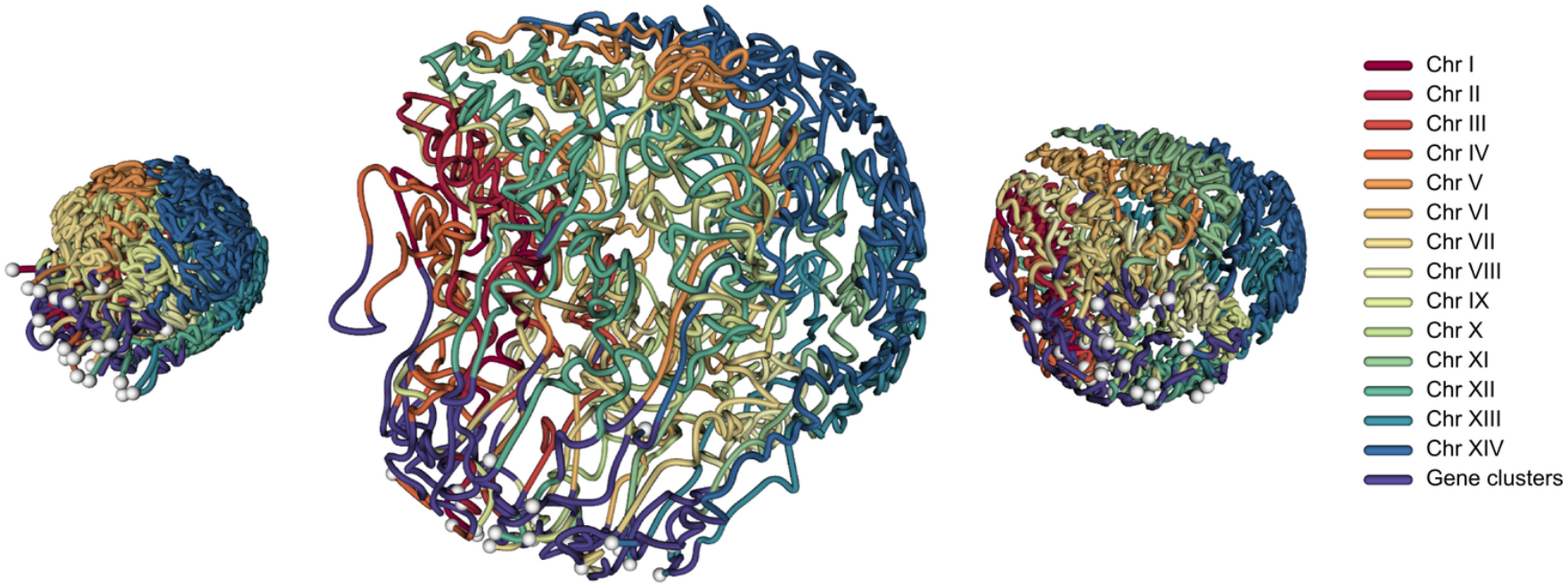
$y \sim x^{-1.14}$





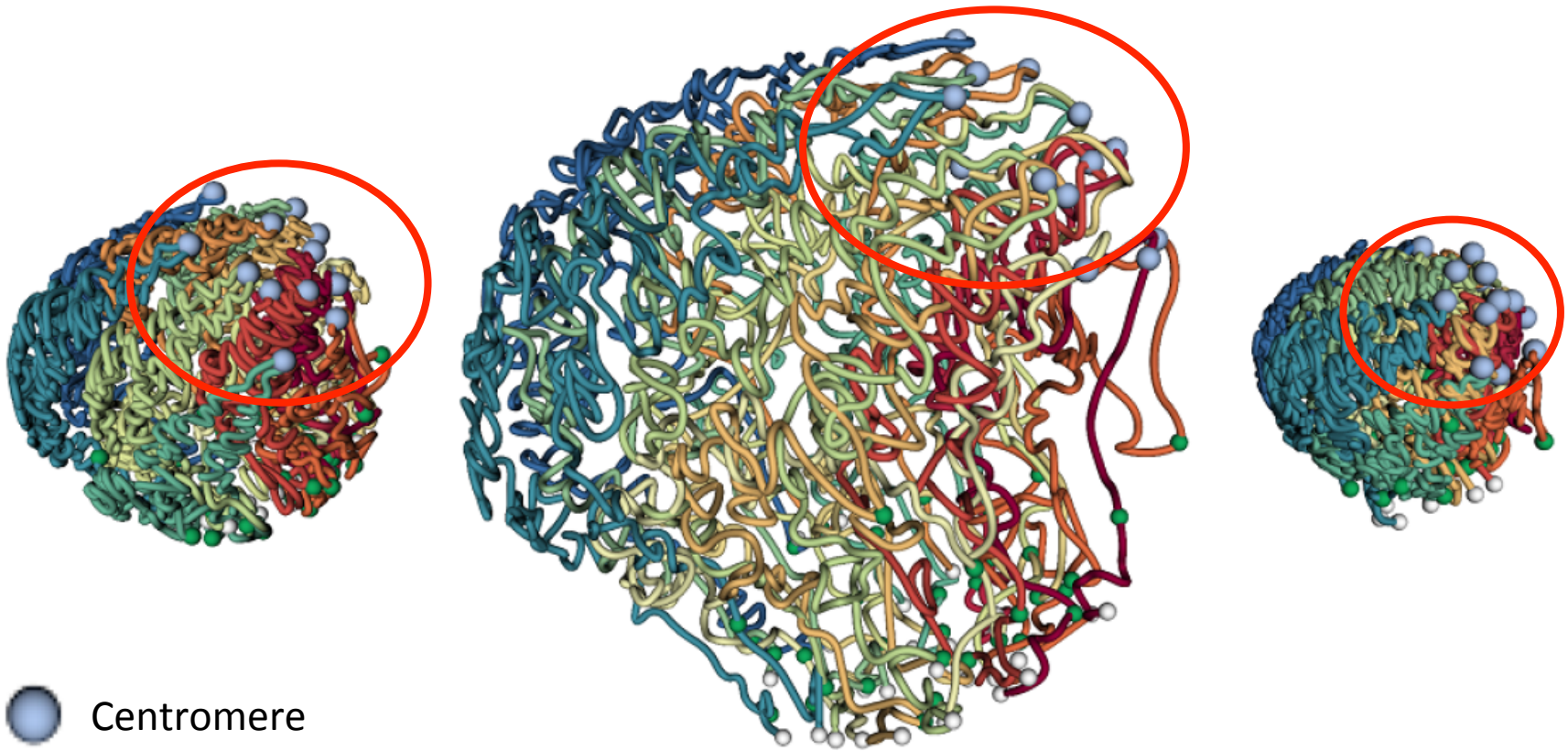
Lieberman-Aiden et al. *Science* 2009

3D modeling recapitulates known organizational principles of *Plasmodium* genome

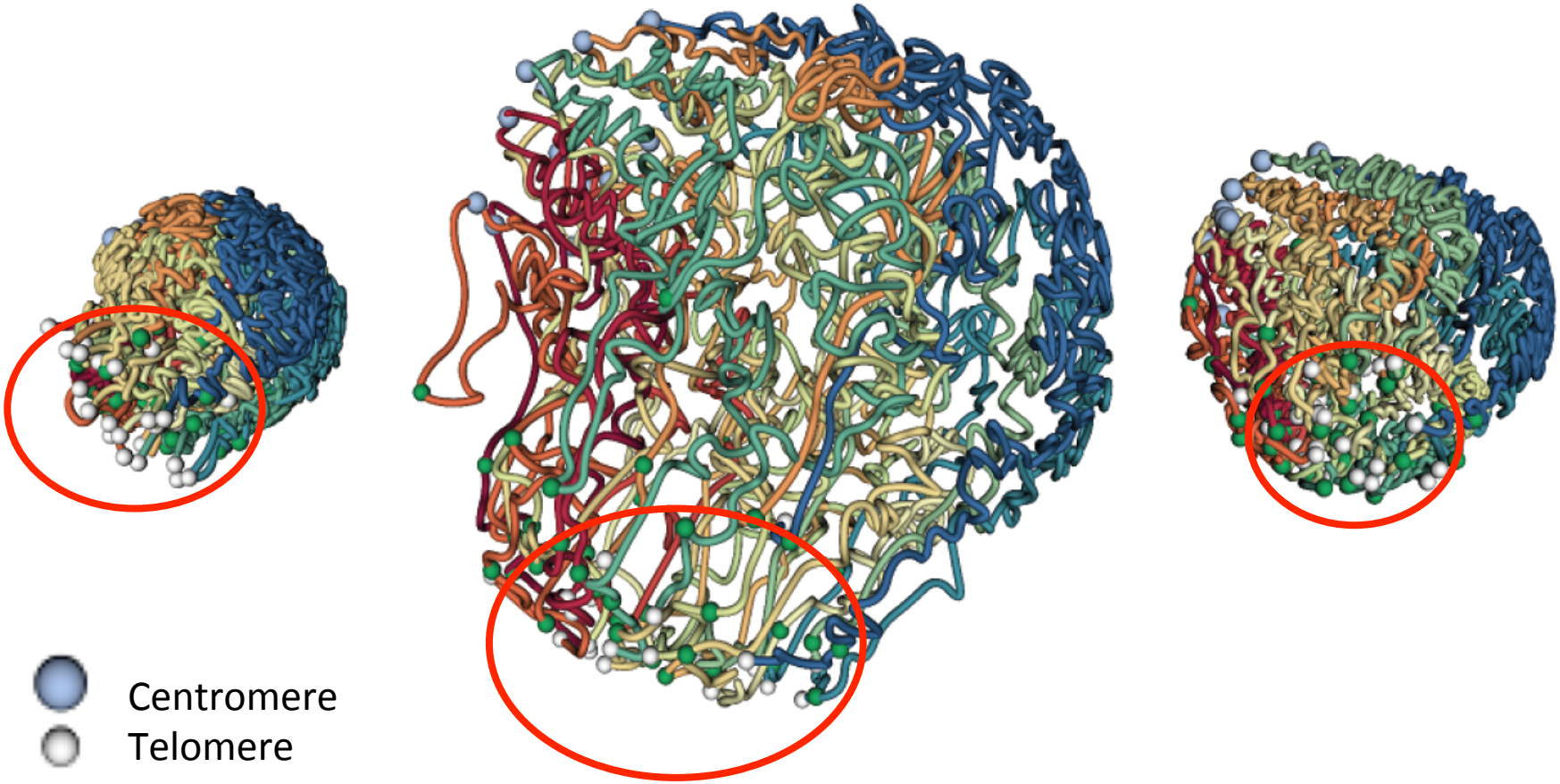


Ring **Trophozoite** **Schizont**

Centromeres colocalize in 3D



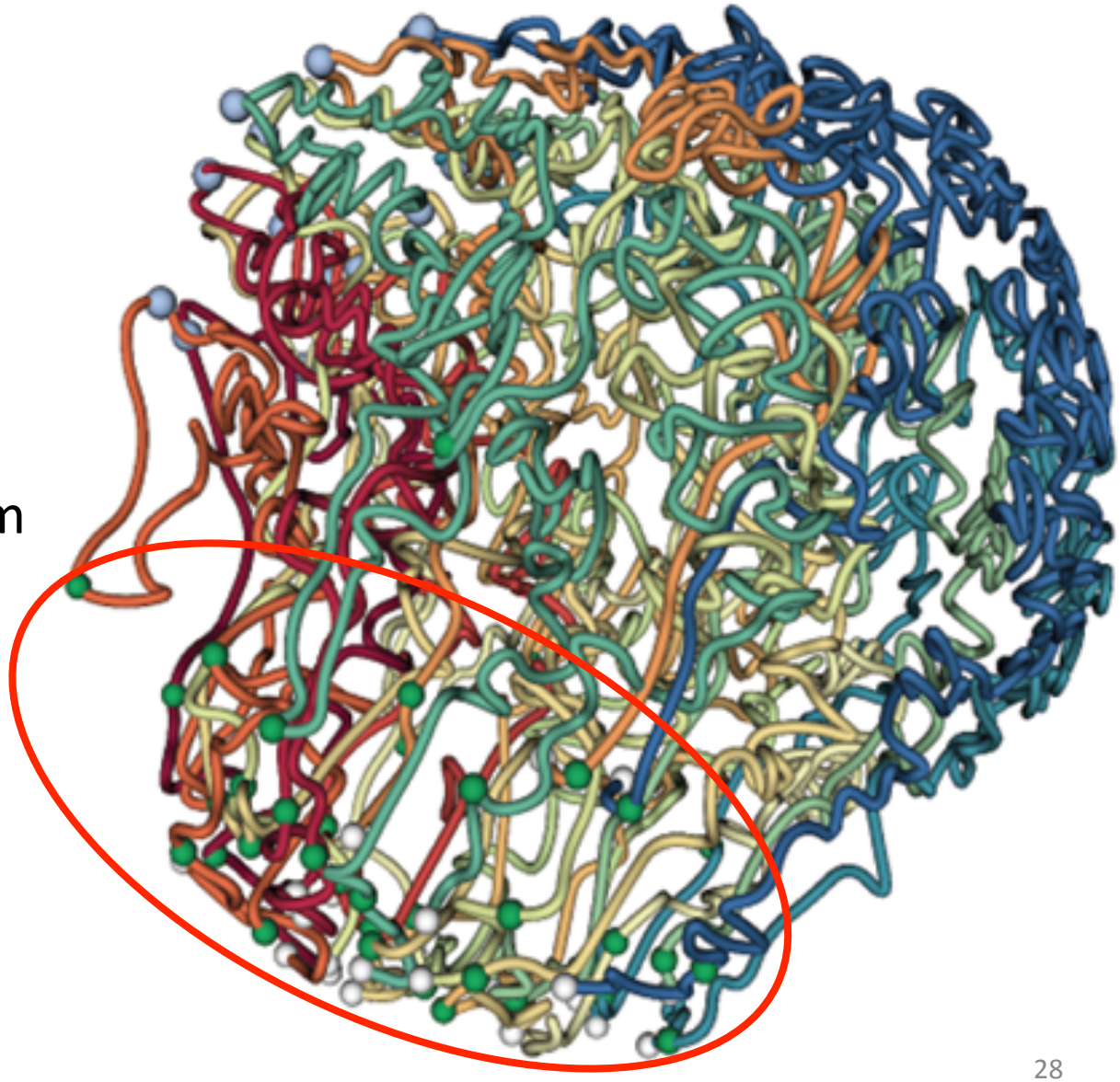
Telomeres colocalize in 3D



Virulence gene clusters colocalize in 3D

- *Plasmodium* encodes 60 virulence genes.
- Exactly one gene is expressed per cell.
- Regulatory mechanism of repression involves H3K36me3.

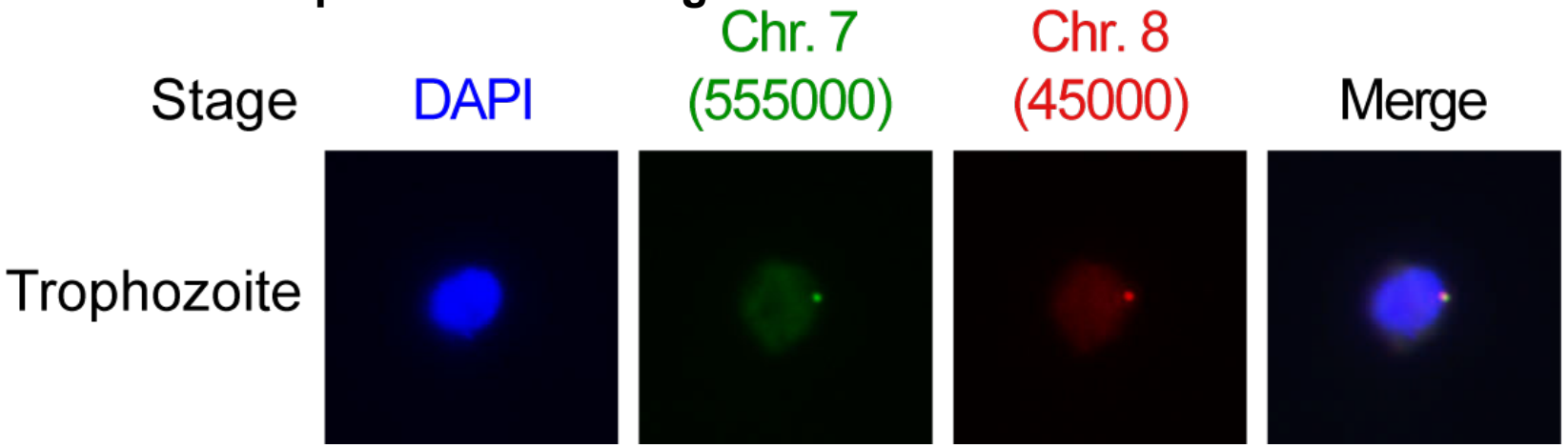
Jiang et al. *Nature* 2013.

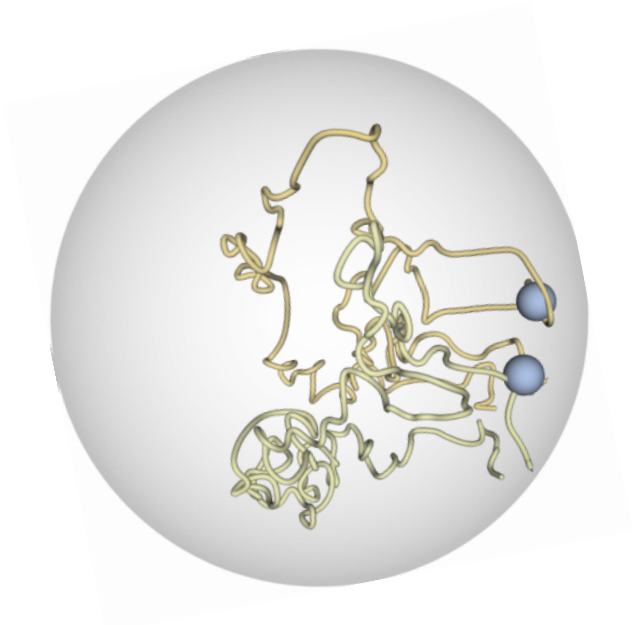


Centromere
Telomere
Virulence gene cluster

DNA FISH confirms selected contacts

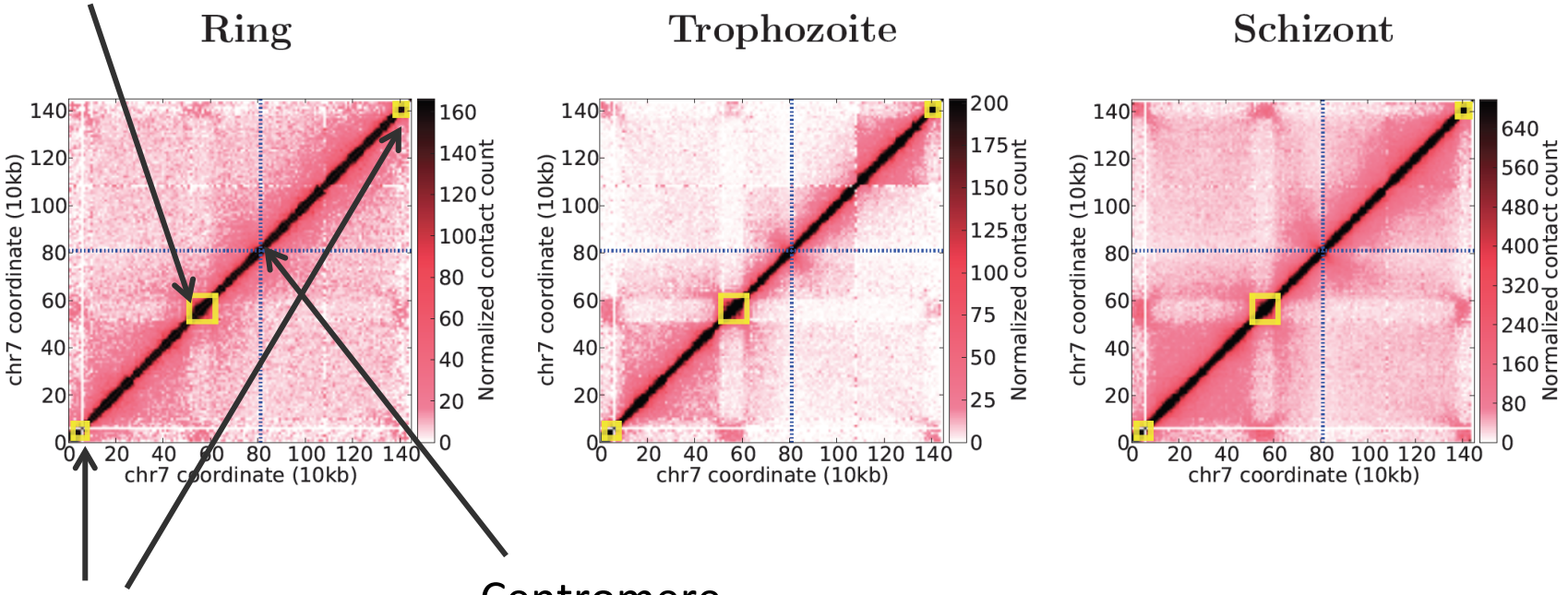
Inter-chromosomal pair of virulence genes





Clusters of virulence genes exhibit domain-like behavior at all stages

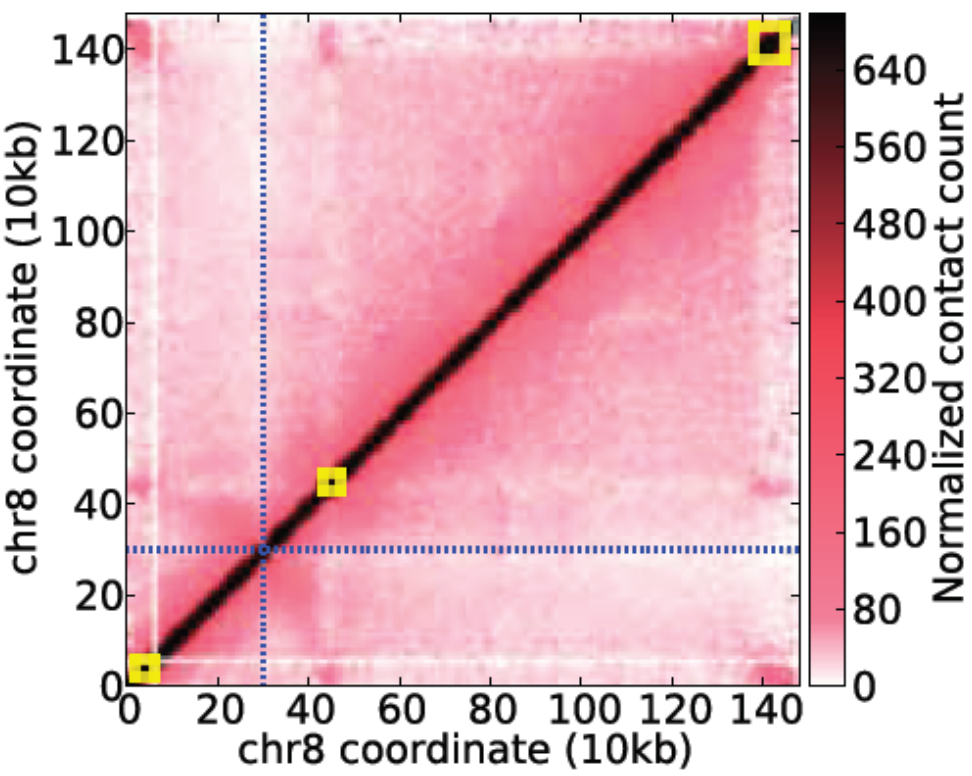
Internal virulence gene clusters



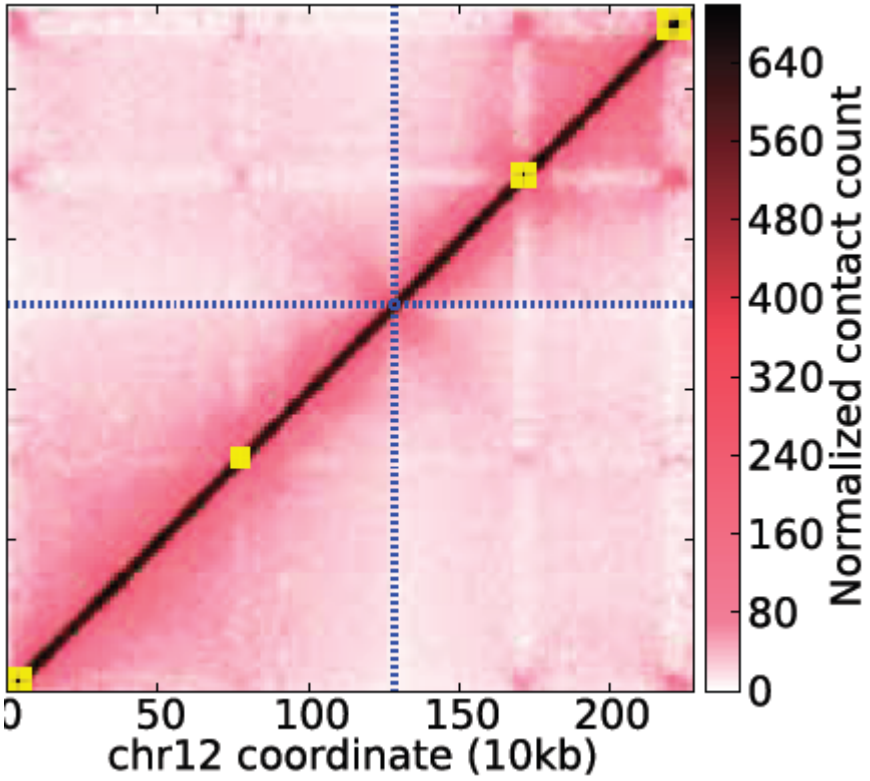
Chromosome 7

The pattern is consistent across chromosomes

Chromosome 8

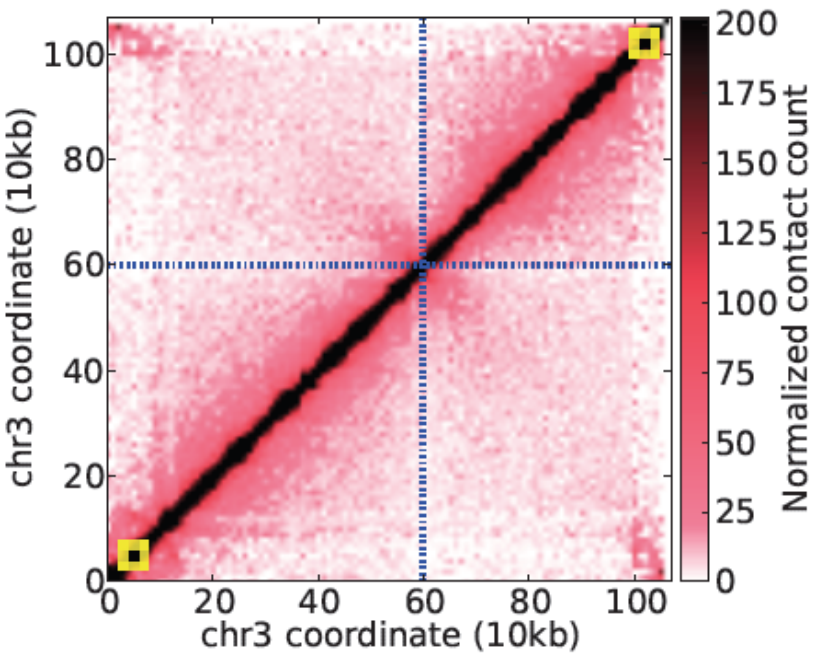


Chromosome 12

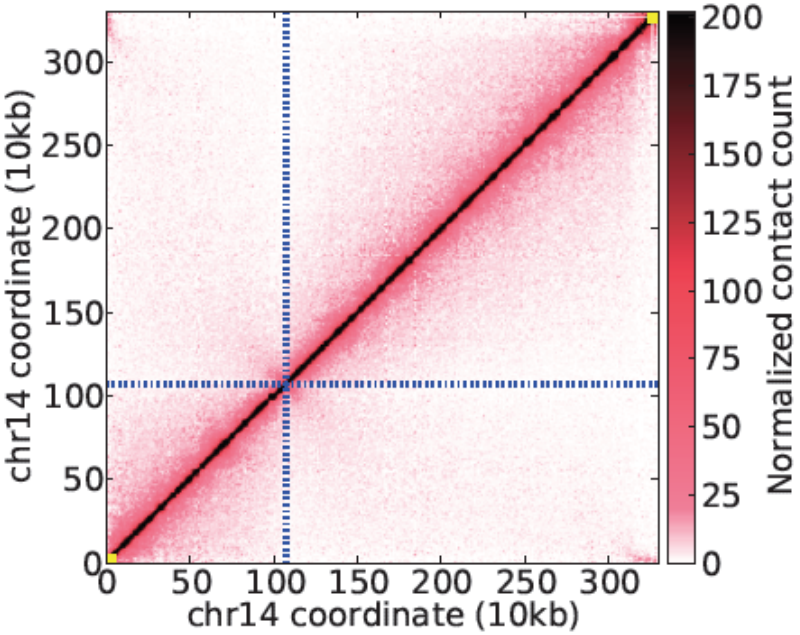


… and absent in chromosomes with no internal virulence gene clusters

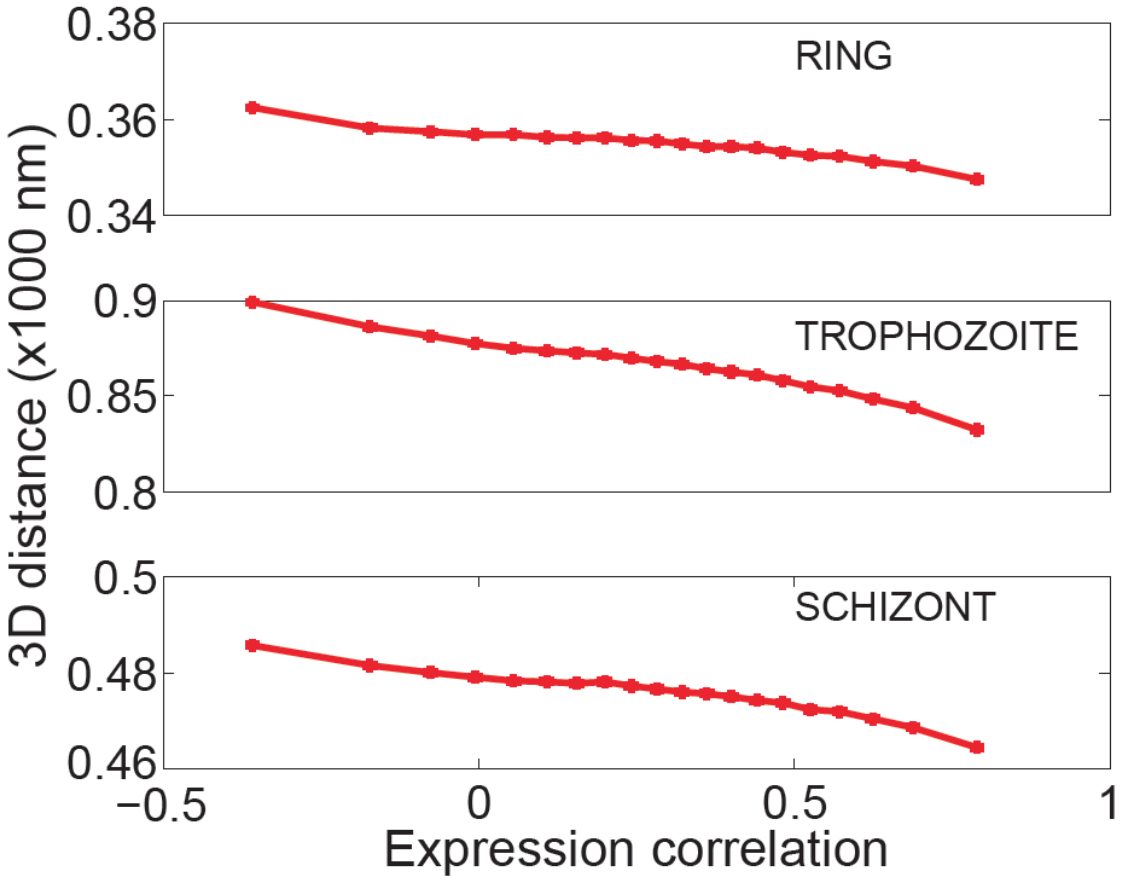
Chromosome 3



Chromosome 14



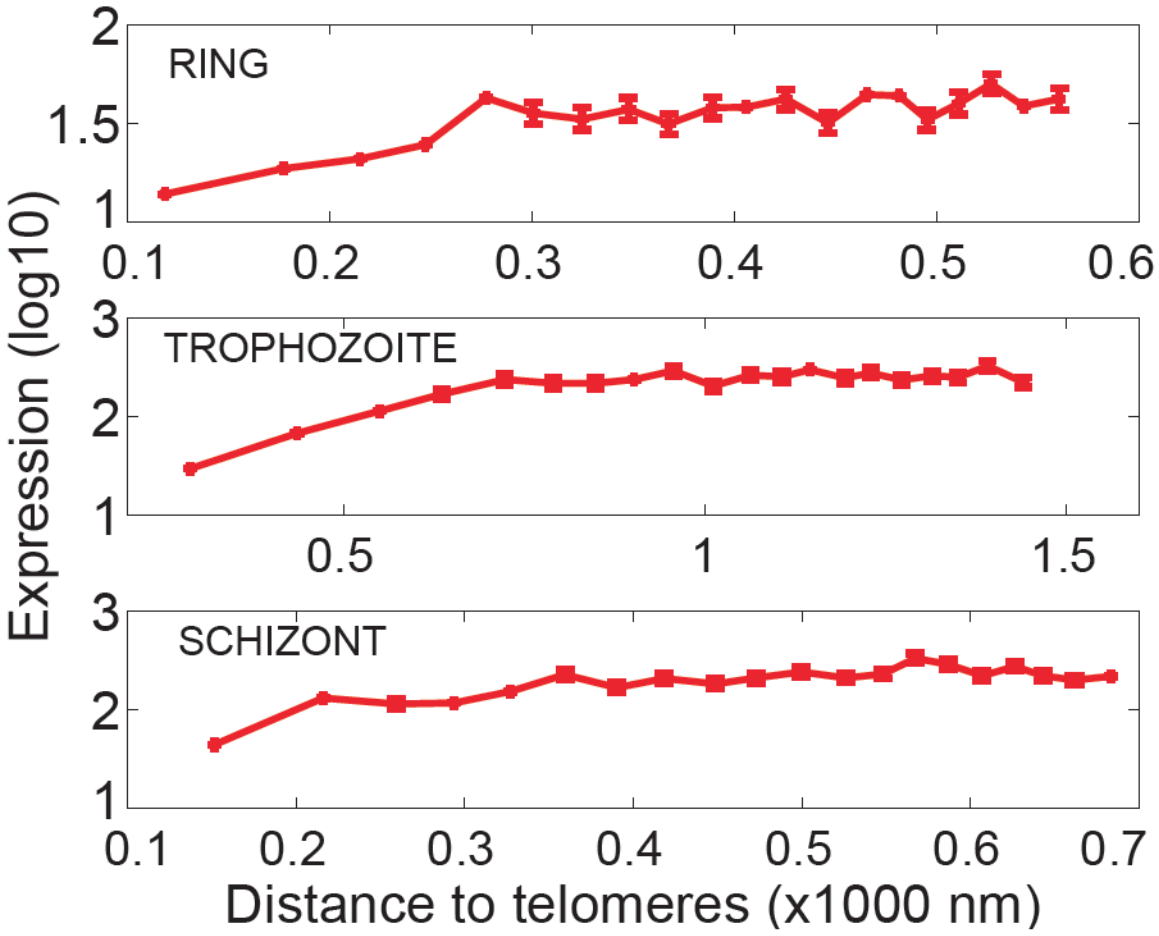
Genes that are close together exhibit correlated expression profiles



- Only inter-chromosomal gene pairs.
- Correlation between expression vectors:
  - Le Roch et al. *Science* 2003.
  - Otto et al. Mol. Microbiology 2010.
  - Lopez-B. et al. *BMC Genomics* 2011.
  - Bunnik et al. *Genome Biology* 2013.

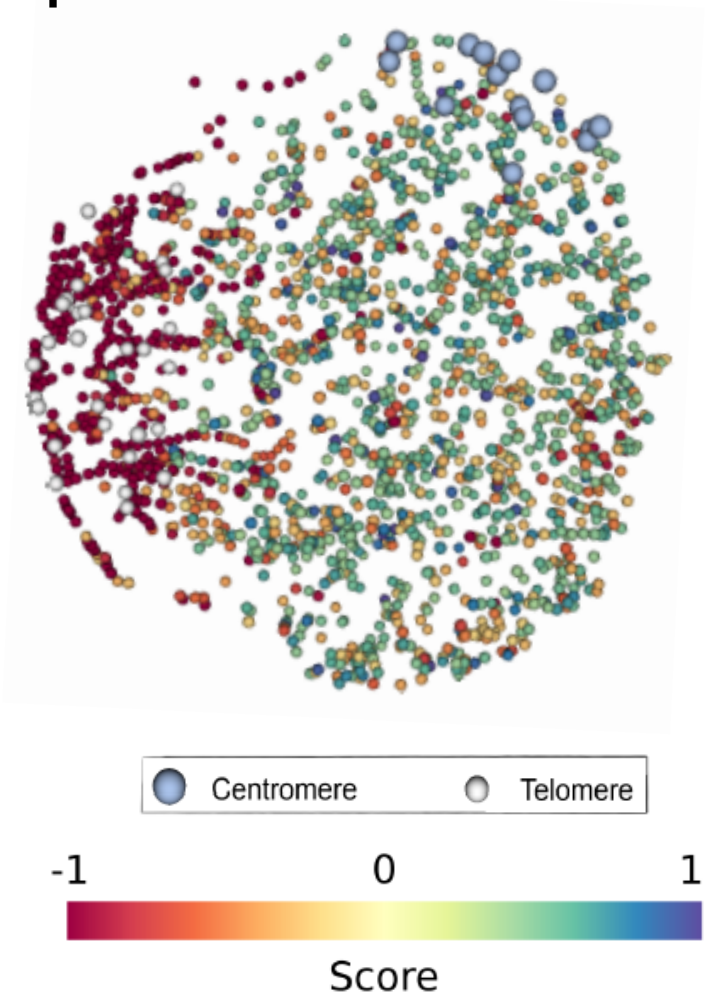
Closer in 3D distance ⇔ more similar expression profile.

Telomeres have a repressive effect on gene expression



Gene expression variation exhibits a gradient across the structure

Trophozoite



Telomeric

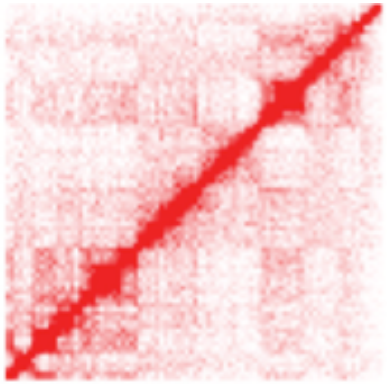
- Antigenic variation
- Sexual stage genes

Non-telomeric

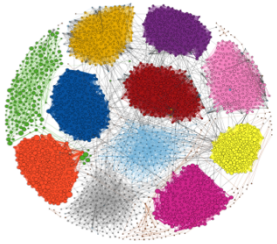
- Translation
- Trophozoite genes

Kernel Canonical Correlation Analysis (kCCA)

Conclusion : The many uses of Hi-C

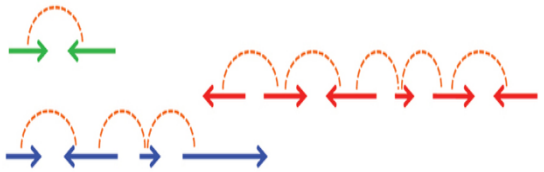


Lieberman-Aiden, *et. al*.
Science, 2009



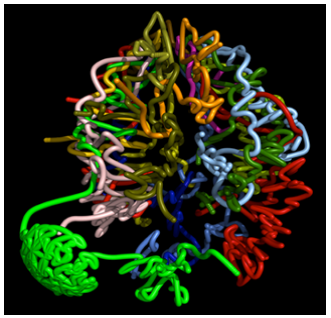
Organismal Deconvolution

Burton, Liachko, *et al*. G3, 2014



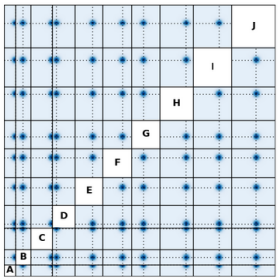
Genome scaffolding

Burton, *et al*.
Nature Biotech, 2013

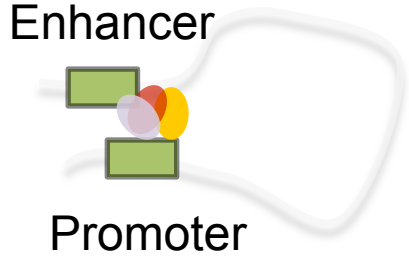


3D model of genome

Duan, *et al*. Nature, 2010 *(S. cerevisae),*
Ay, *et al*. Genome Res., 2014a *(P.falciparum)*



Centromere calling

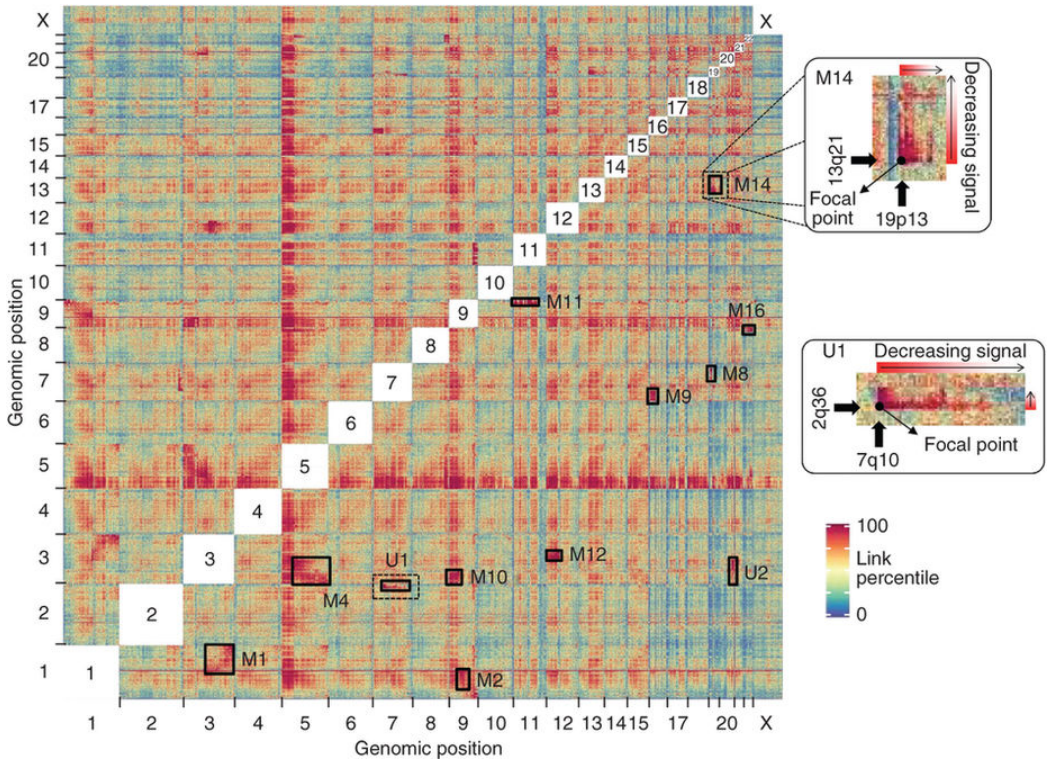


Long-range chromatin contacts

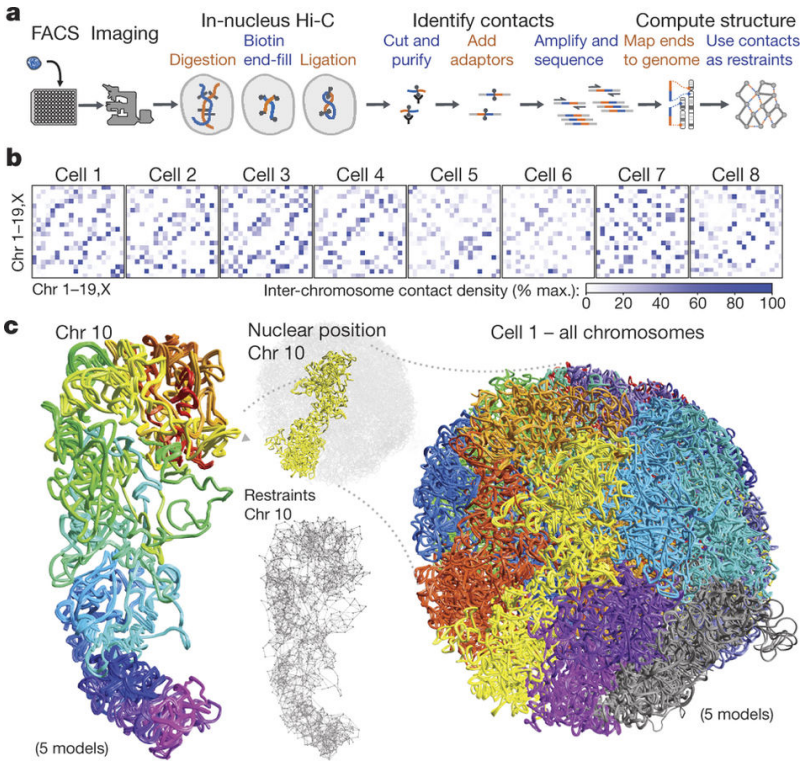
Ay, *et al*. Genome Res., 2014b

Challenges

Cancer genomes



Burton et al., Nat Biotech 2015



Single-cell Hi-C

Stevns et al., Nature 2017

Acknowledgements

University of Washington
William Noble
Ferhat Ay



Josh Burton

Ivan Liachko

University of California Riverside
Karine Le Roch
Evelien Bunnik
Sebastiaan Bol
Jacques Prudhomme



MINES ParisTech, France
Jean-Philippe Vert
Nelle Varoquaux


Funding

NATIONAL INSTITUTES OF HEALTH
COMPUTING RESEARCH ASSOCIATION CRA
NSF
RITA ALLEN FOUNDATION
AGENCE NATIONALE DE LA RECHERCHE ANR
INVESTISSEMENTS D'AVENIR
SEVENTH FRAMEWORK PROGRAMME
erc