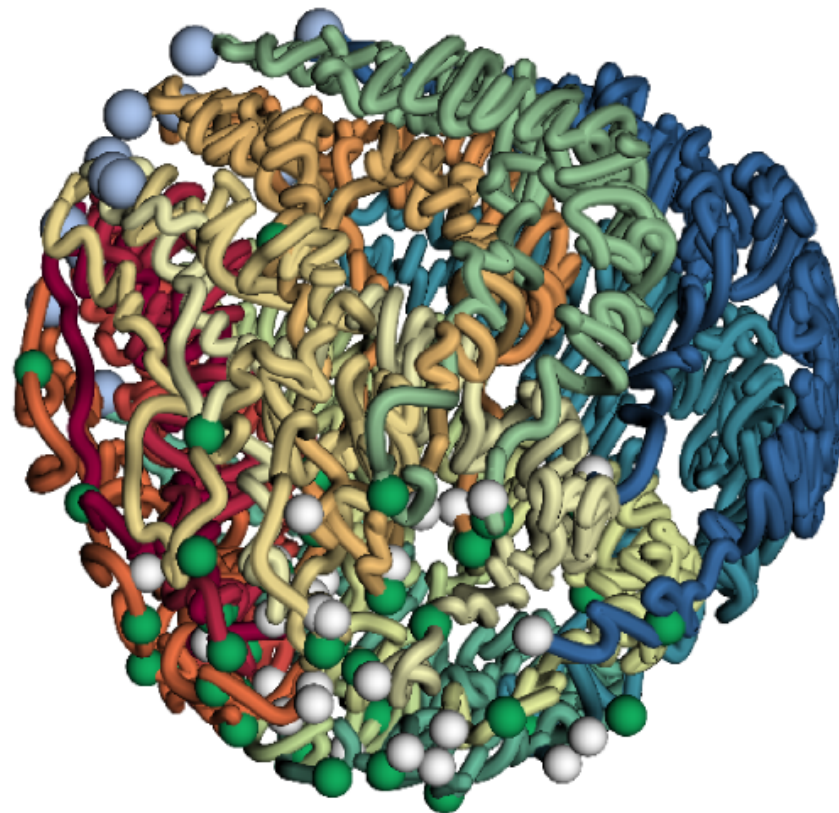# Inferring a genome 3D structure and
How it is related to gene regulation in malaria parasite

Jean-Philippe Vert

Joint work with
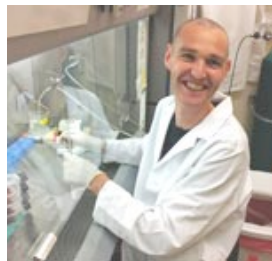
University of Washington
William Noble
Ferhat Ay

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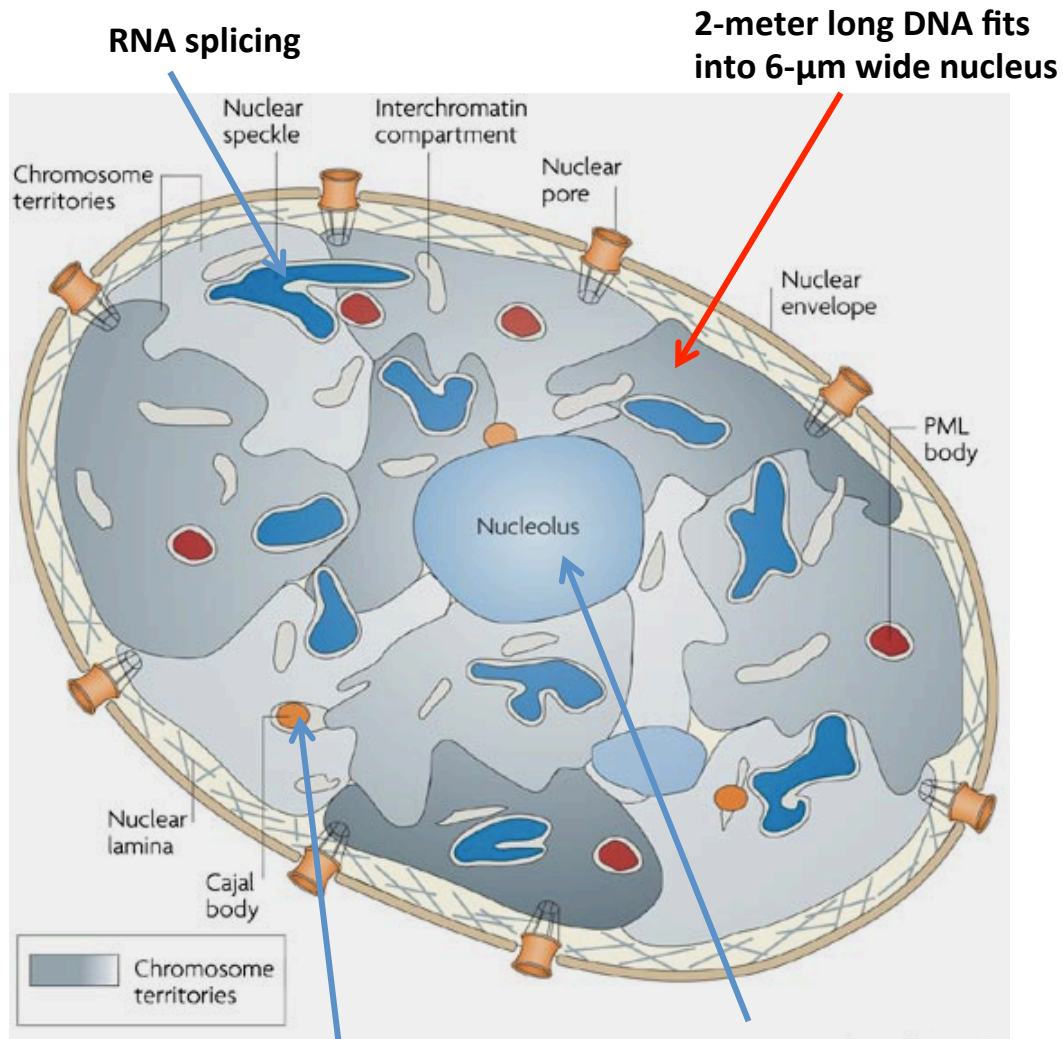Evelien Bunnik
Sebastiaan Bol
Jacques Prudhomme

Mines ParisTech
Nelle Varoquaux

How does genome architecture influence genome function?

RNA splicing

2-meter long DNA fits into 6-μm wide nucleus

Nuclear speckle

Interchromatin compartment

Nuclear pore

Chromosome territories

Nuclear envelope

Nuclear lamina

PML body

Nucleolus

Cajal body

Chromosome territories

Processing of nuclear RNA

rDNA transcription / ribosome assembly

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

Tools for capturing chromosome conformation

Hi-C: fill in with biotin-dCTP

Digestion → Ligation → DNA purification → 3C, 4C, 5C, Hi-C (Ligation product library)

Immuno-precipitation → DNA purification → ChIP-loop, ChIA-PET (Ligation product library)
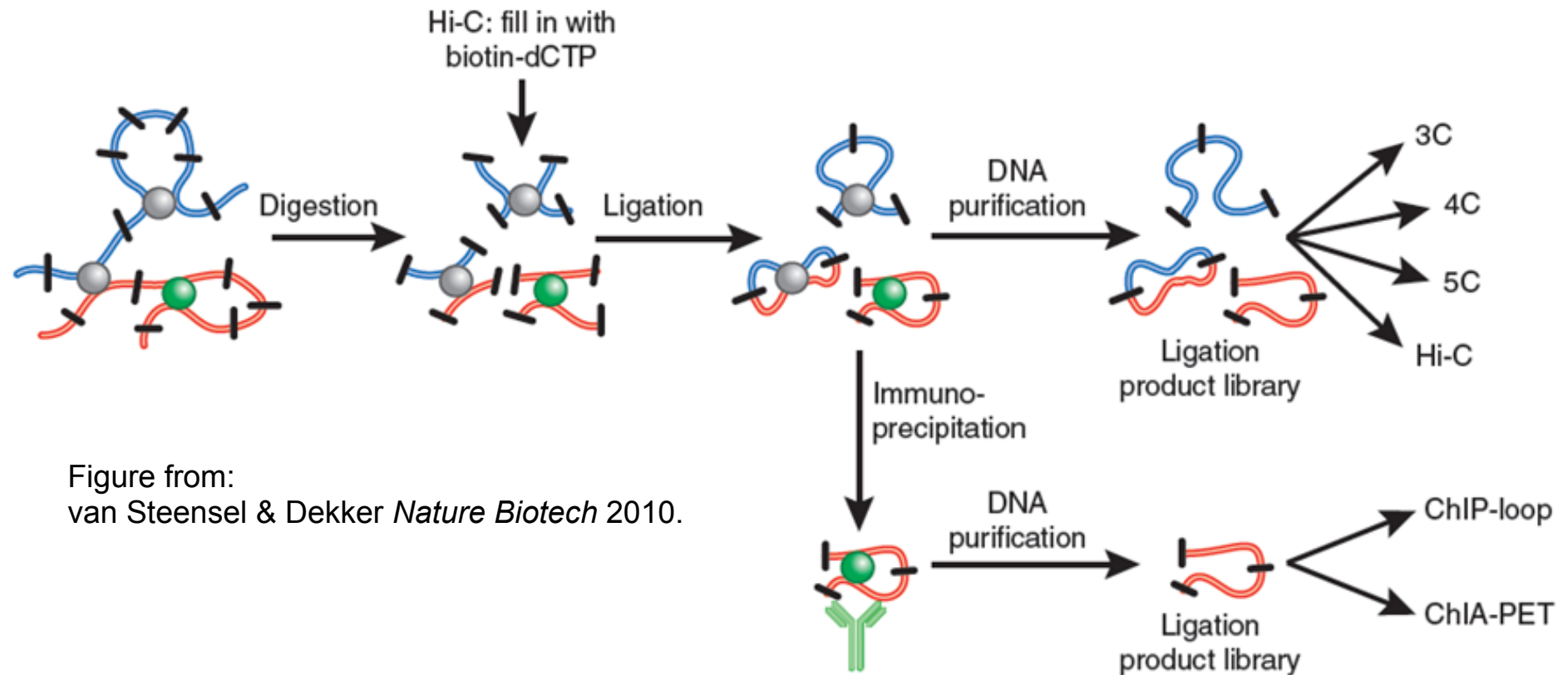
Figure from:
van Steensel & Dekker *Nature Biotech* 2010.

3C

Dekker et al.
Science 2002

ChIA-PET

Fullwood et al.
Nature 2009

Hi-C

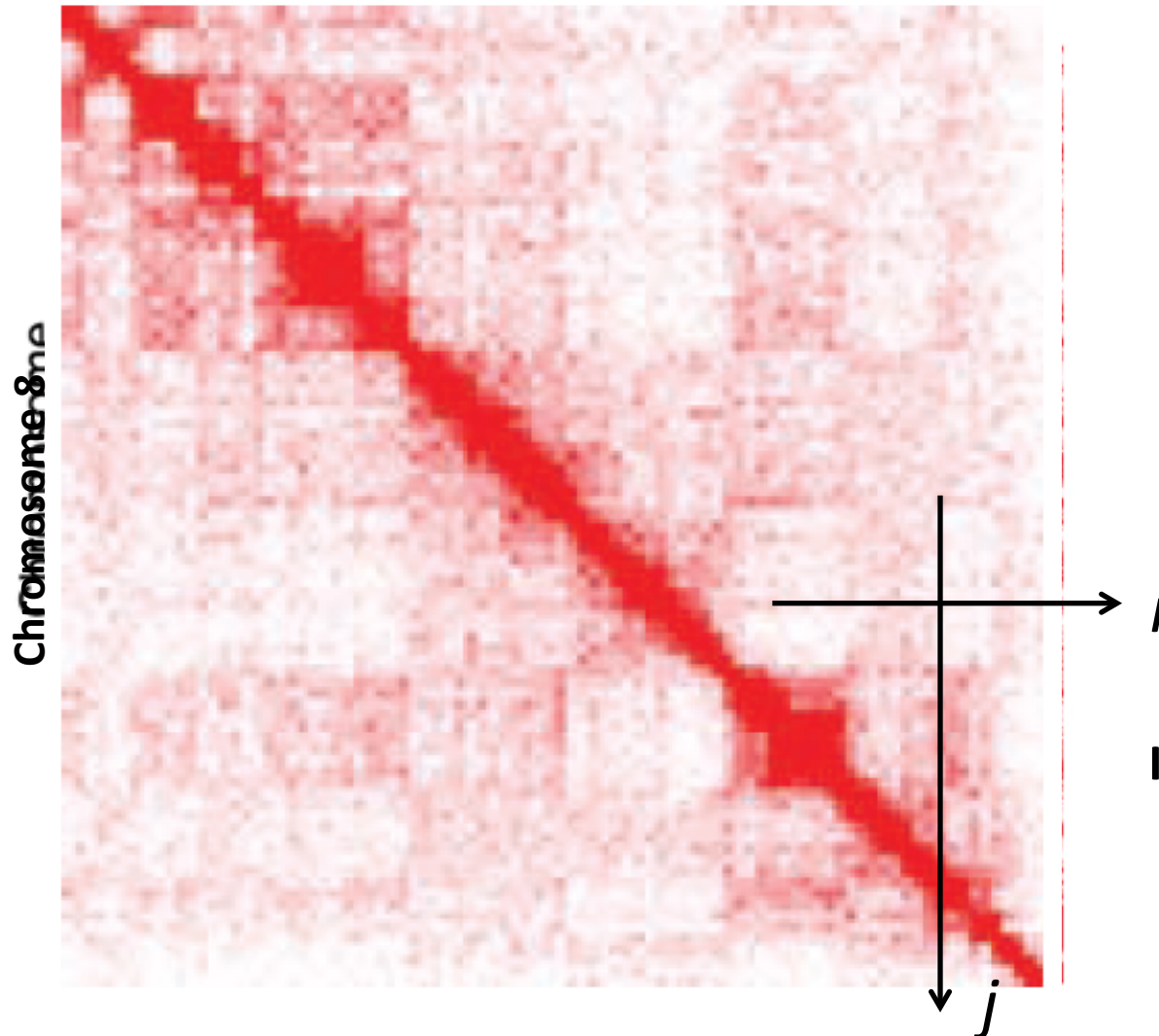L.-Aiden et al.
Science 2009

Duan et al.
Nature 2010

Output of conformation capture is a contact matrix

Chromosome 8

paired-end reads
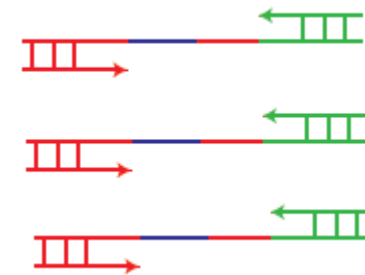
Chromosome 8

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

i

j

Inter-chromosomal contact

**Intra-chromosomal contact**

5

Part 1

PASTIS, a tool to infer the 3D structure of the genome from HiC data

Reconstructing the 3D structure from contact counts maps

Two main types of approaches:

- *Consensus* methods that aim at inferring a unique mean structure representative of the data. [Duan et al., 2010, Tanizawa et al., 2010, Bau et al., 2011, Zhang et al., 2013, Ben-Elazar et al., 2013]

- *Ensemble* methods that yield a population of structures [Rousseau et al., 2011, Hu et al., 2013, Kalhor et al., 2011]

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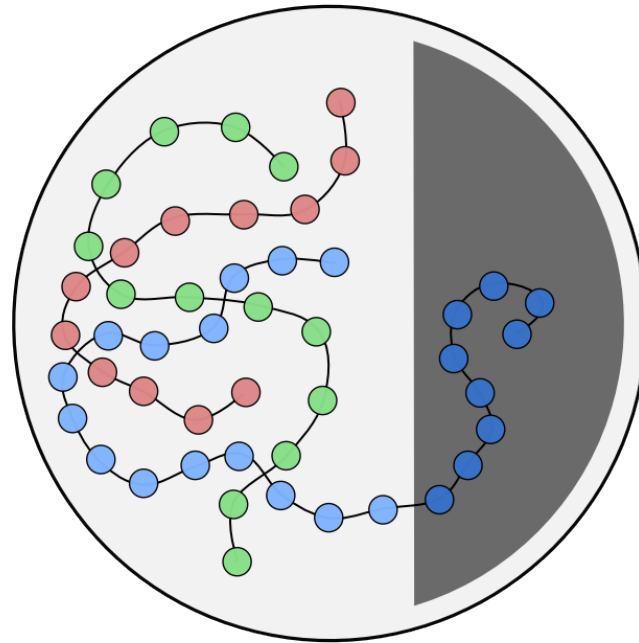
Figure: **Beads on a string model** Chromosomes are modeled as a series of beads. Nucleus is assumed to be spherical.

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

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

$$\underset{x_1,\ldots,x_n}{\text{minimize}} \quad \sigma(\mathbf{X}, \mathbf{C}))$$

$$\text{subject to} \quad \text{some bio}$$

$$\mathbf{x}_i^T \mathbf{x}_i \leq r$$

- 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

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

$$\underset{\mathbf{x}_1,\ldots,\mathbf{x}_n,\Theta}{\text{minimize}} \quad \sigma(\mathbf{X}, \mathbf{C}, \Theta))$$

$$\text{subject to} \quad \Theta \quad \text{decreasing}$$

some biologically motivated constraints

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

# 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

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

$$\text{subject to} \quad \text{some biologically motivated constraints}$$

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

Default counts-to-distance transfer function

Fractal globule
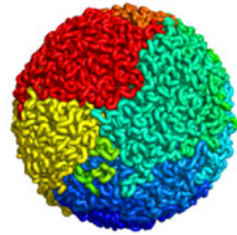
Equilibrium

- $c \sim s^{-1}$
- $d \sim s^{1/3}$

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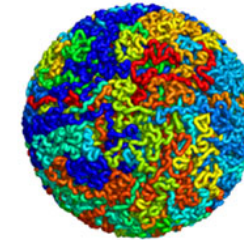
- $c \sim s^{-3/2}$
- $d \sim s^{1/2}$ for $s < s_{\max}^{2/3}$
- Valid for budding yeast, and small organism

Default counts-to-distance transfer function

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

(6)

Data

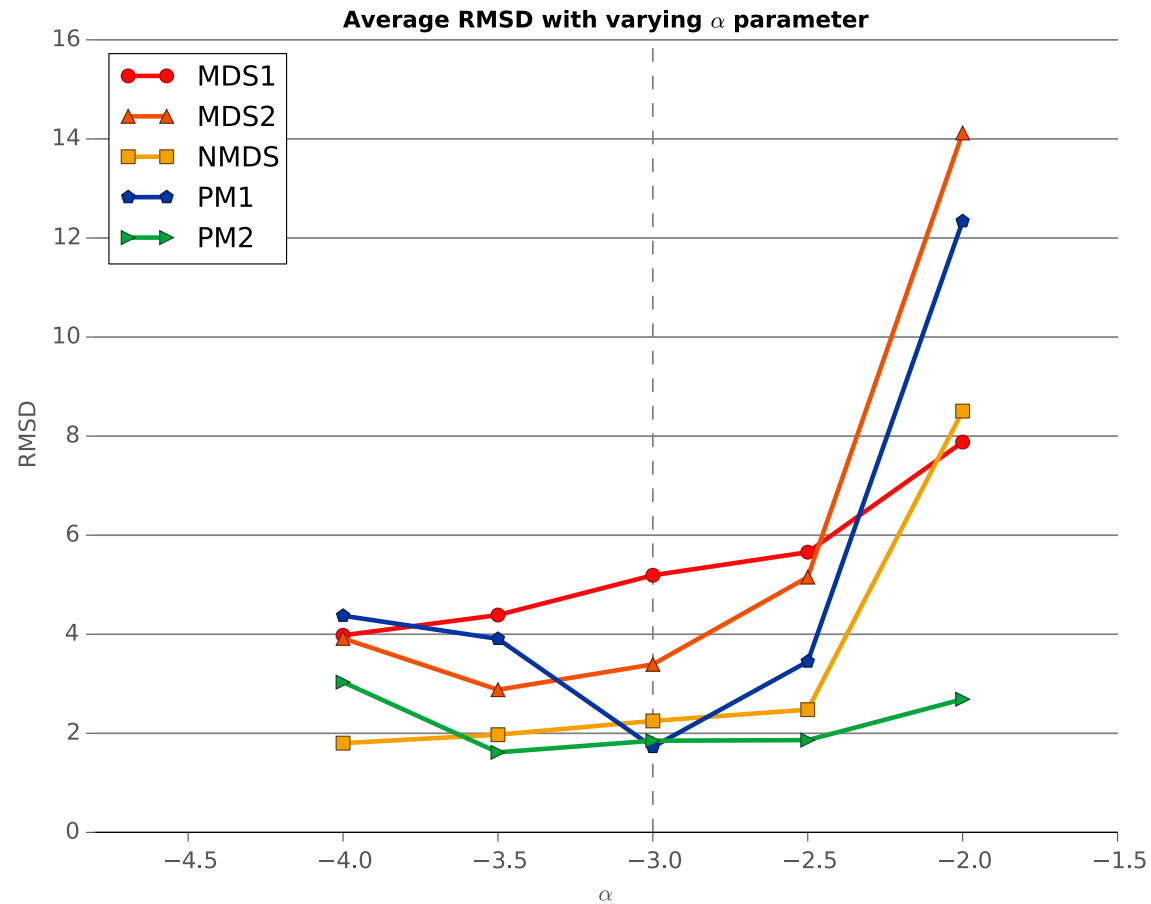- **Generated Datasets**

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

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
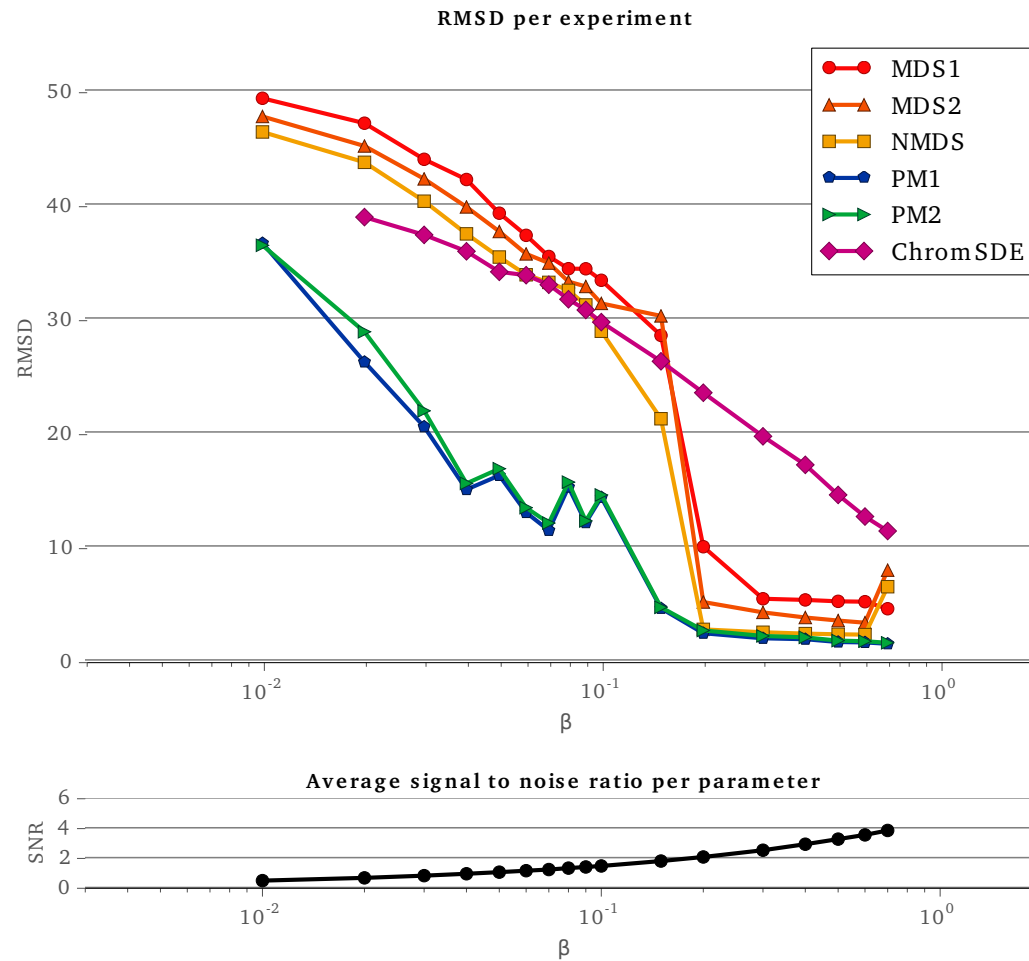
Robustness to parameter misspecification

Average RMSD with varying α parameter

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**

Conclusion

- Poisson model outperforms MDS-based models, particularly at low coverage (or high resolution)

- Code available at http://cbio.mines-paristech.fr/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.

- Extension to aneuploidy?

Chrom 1
Chrom 2
Chrom 3
Chrom 4
Chrom 5
Chrom 6
Chrom 7
Chrom 8 A
Chrom 9
Chrom 10
Chrom 11
Chrom 12
Chrom 13
Chrom 14
Chrom 15 A
Chrom 16
Chrom 17
Chrom 18
Chrom 19
Chrom 20
Chrom 21
Chrom 22
Chrom X
Chrom Y
Chrom 8 B
Chrom 15 B
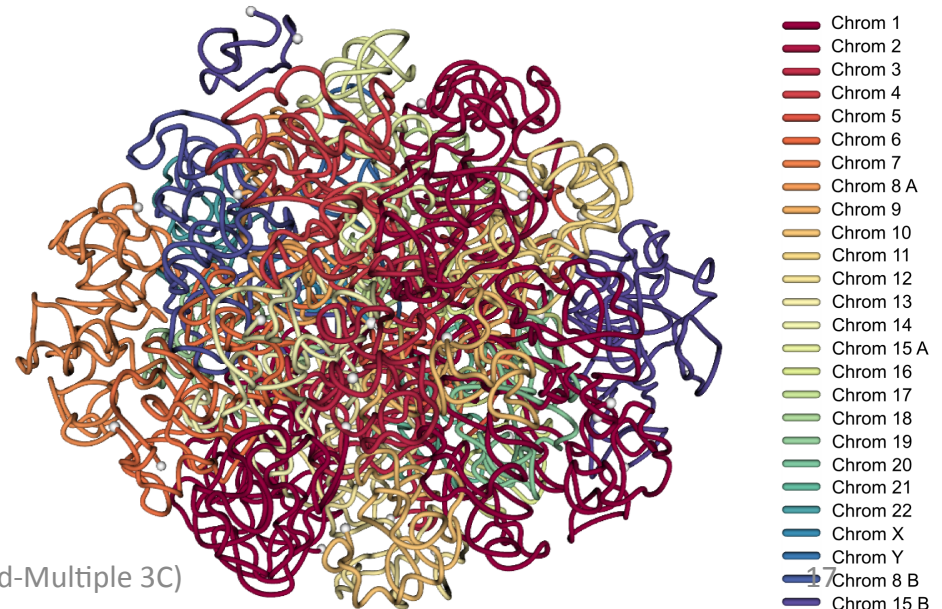
Part 2

The spatial organization of the *P. falciparum* genome

Modeling the dynamic genome architecture of a malaria parasite *(Plasmodium falciparum)*
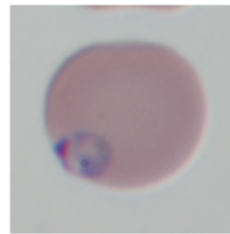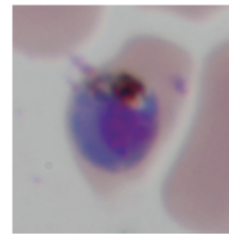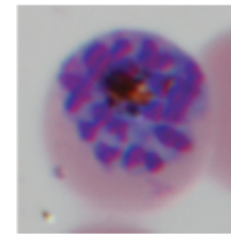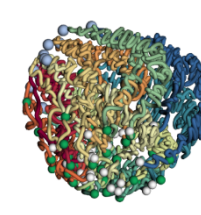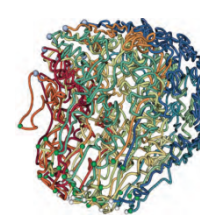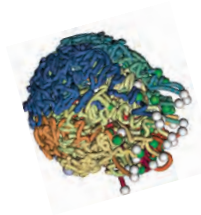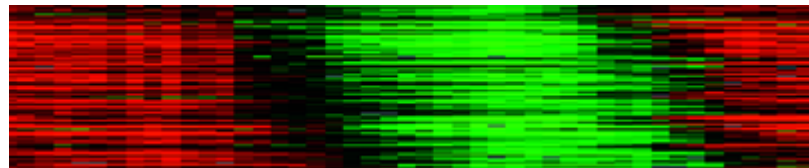
Gene expression

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

How *Plasmodium* regulates gene expression is mysterious
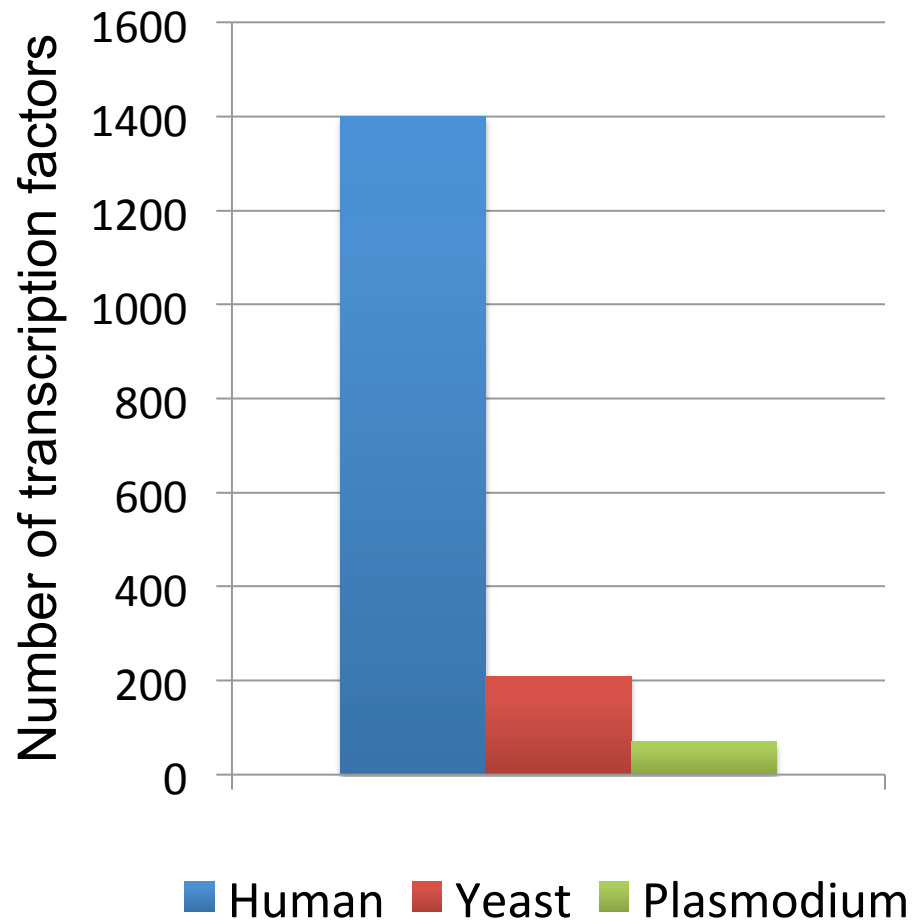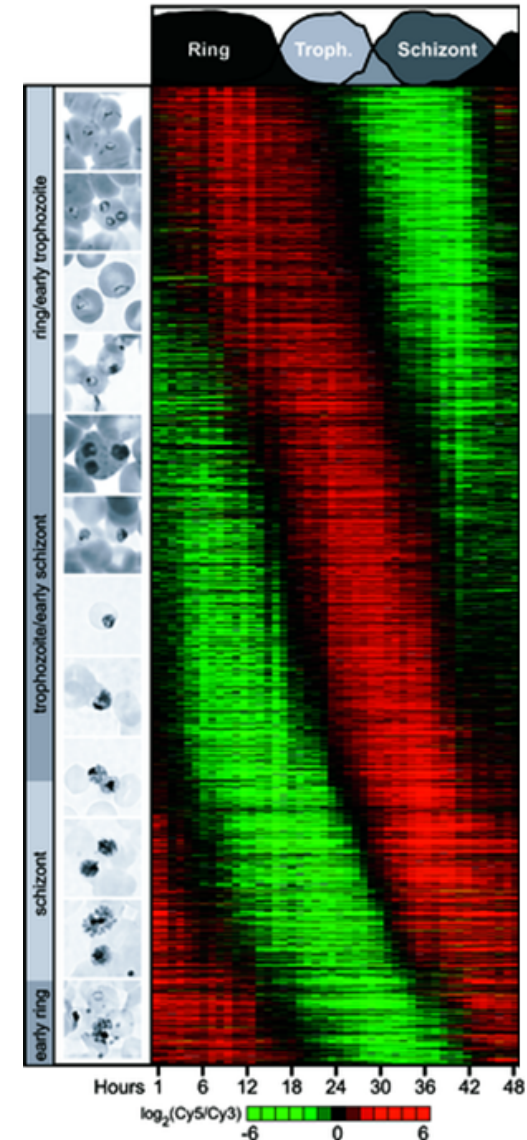

Very few transcription factors.

- 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?

Trophozoite

Schizont

Ring
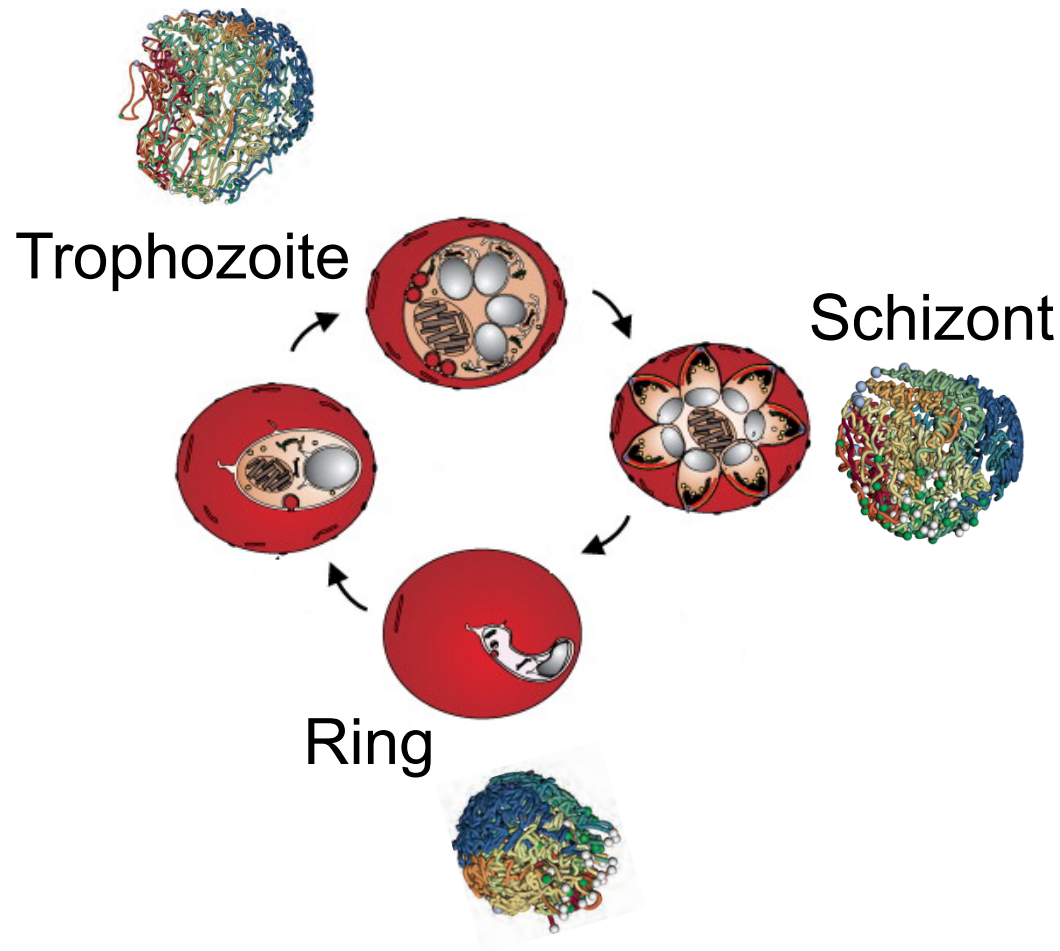
Erythrocytic cycle

We assayed genome architecture at 3 time points in the erythrocytic cycle
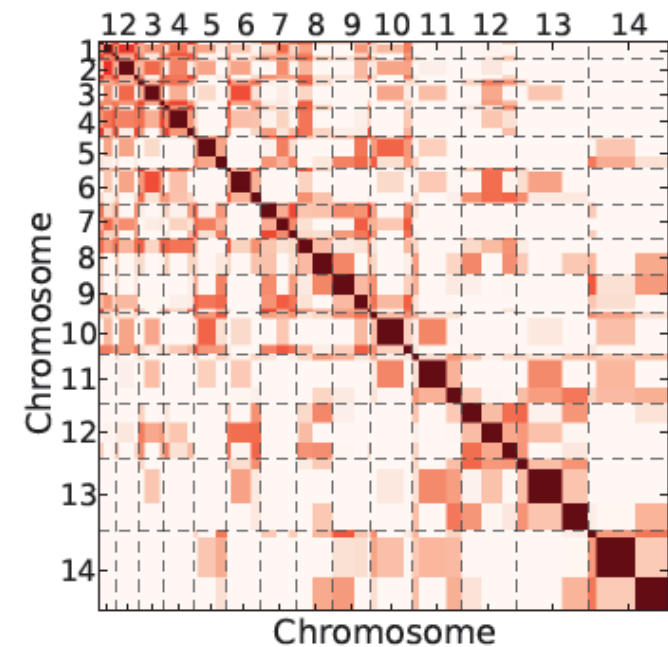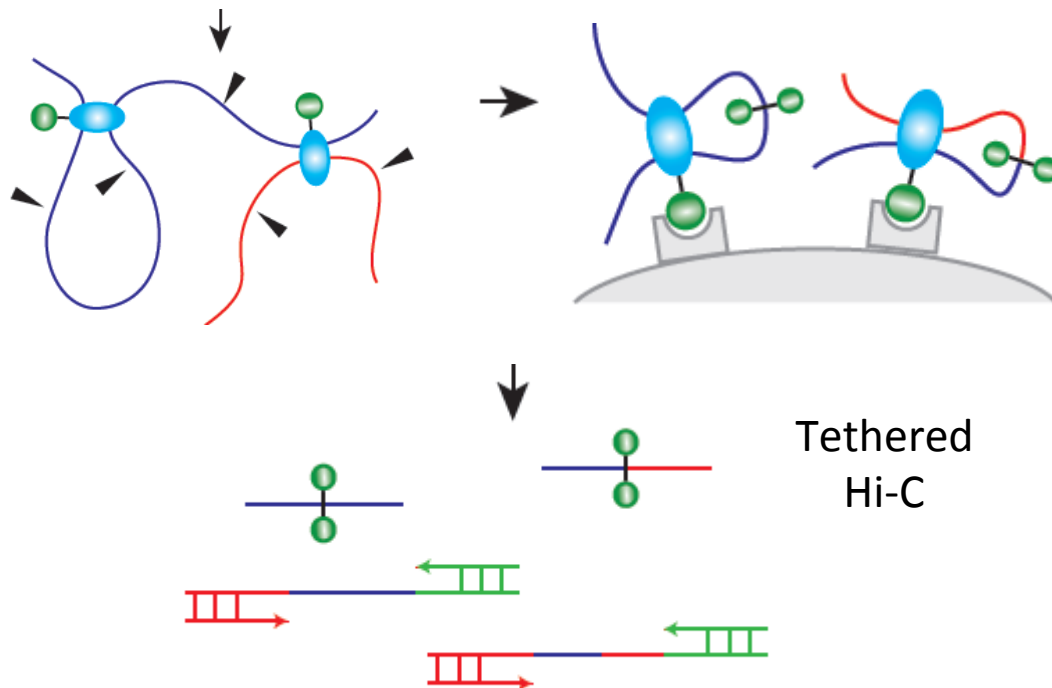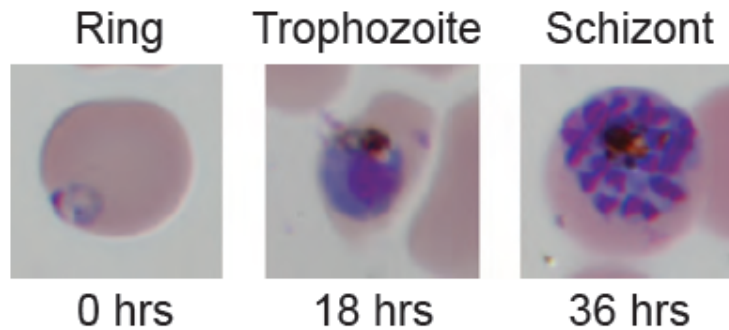
Plasmodium contact frequencies suggest a fractal globule architecture

$y \sim x\uparrow\alpha$

RING

TROPHOZOITE

SCHIZONT

TROPH.-cont.

$\alpha=$
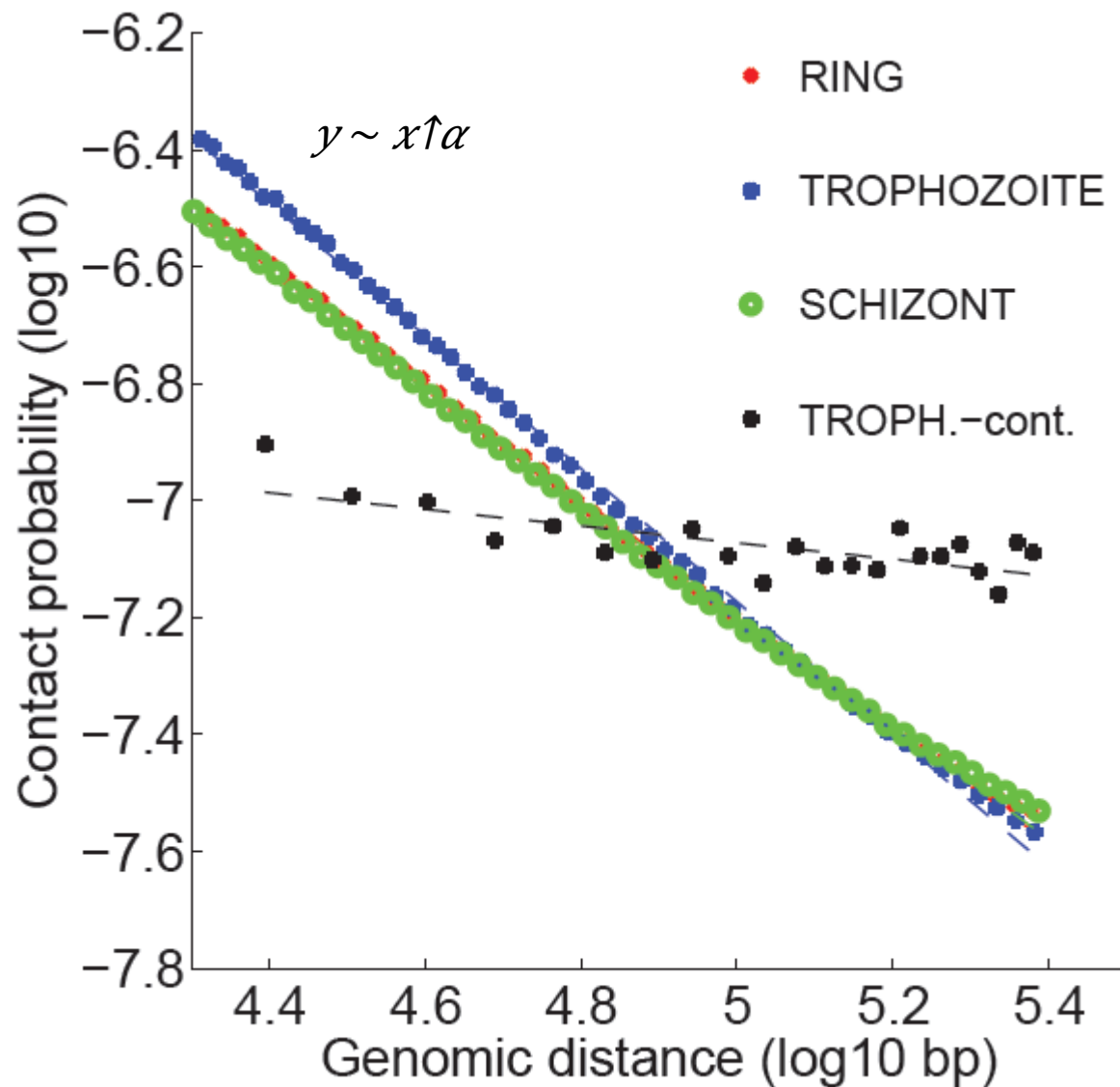
-0.98

-1.14

-0.96

-0.14

UNFOLDED POLYMER

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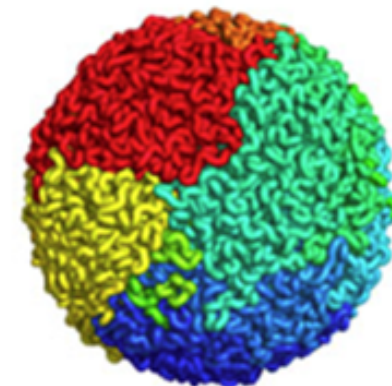
Scaling parameter for the Trophozoite stage is indicative of more intermingled chromatin

$y \sim x^{-1.14}$

Trophozoite

18 hrs

UNFOLDED POLYMER

Fractal globule

Equilibrium globule

$\alpha = -1$

$\alpha = -1.5$

Probability of contact

Equilibrium
Fractal Globule

P_c

-1

$-3/2$

distance, s (monomers)
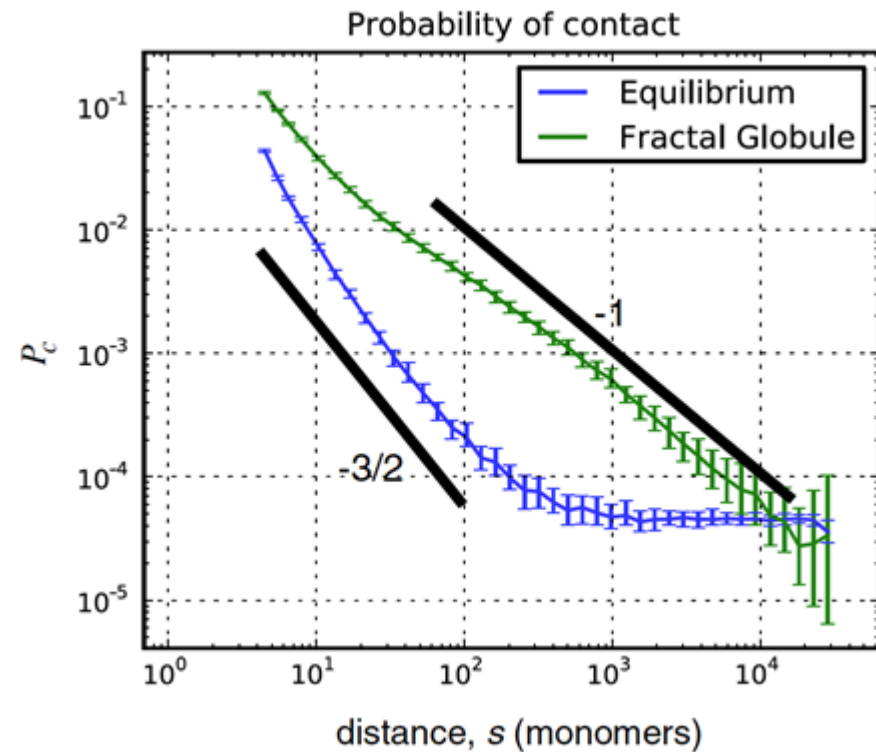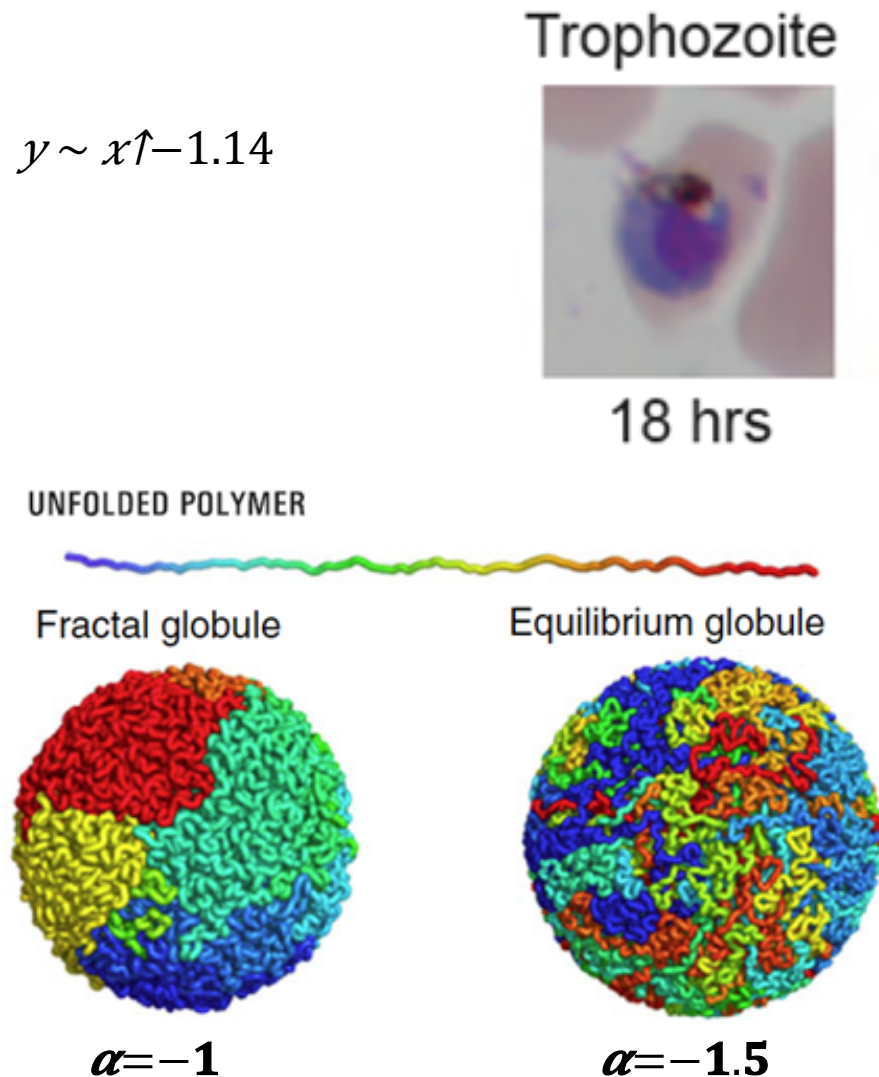
Lieberman-Aiden et al. *Science* 2009

24

We use the observed contact counts to infer a 3D model

- Model the genome as beads at 10 kbp resolution.

- Estimate Euclid[ean] distance [...] e m[...] ruler derived from intra-chromoso[mal interac]tions.

- Find 3D coordinates that yield the expected distances:

Distance according to the model

Distance according to the data

$$\underset{\mathbf{X}}{\text{minimize}} \quad \sum_{\delta_{ij} \in \mathcal{D}} \frac{1}{d_{ij}^2} \left(d_{ij} - \delta_{ij} \right)^2 \quad \mathbf{X} \in R^{3 \times n}$$

$$\mathcal{D} = \{\delta_{ij} | \delta_{ij} \neq 0\}$$

- Include constraints reflecting physical and biological prior knowledge.
  1. All loci must lie within a spherical nucleus centered on the origin.

     , , (Weiner et al. Cell Microbiology, 2011).
  2. Two adjacent loci must not to be too far apart.

     1000 bp of chromatin occupies a distance between 6.6 to 9.1 nm (Bystricky et al. *PNAS*, 2004).
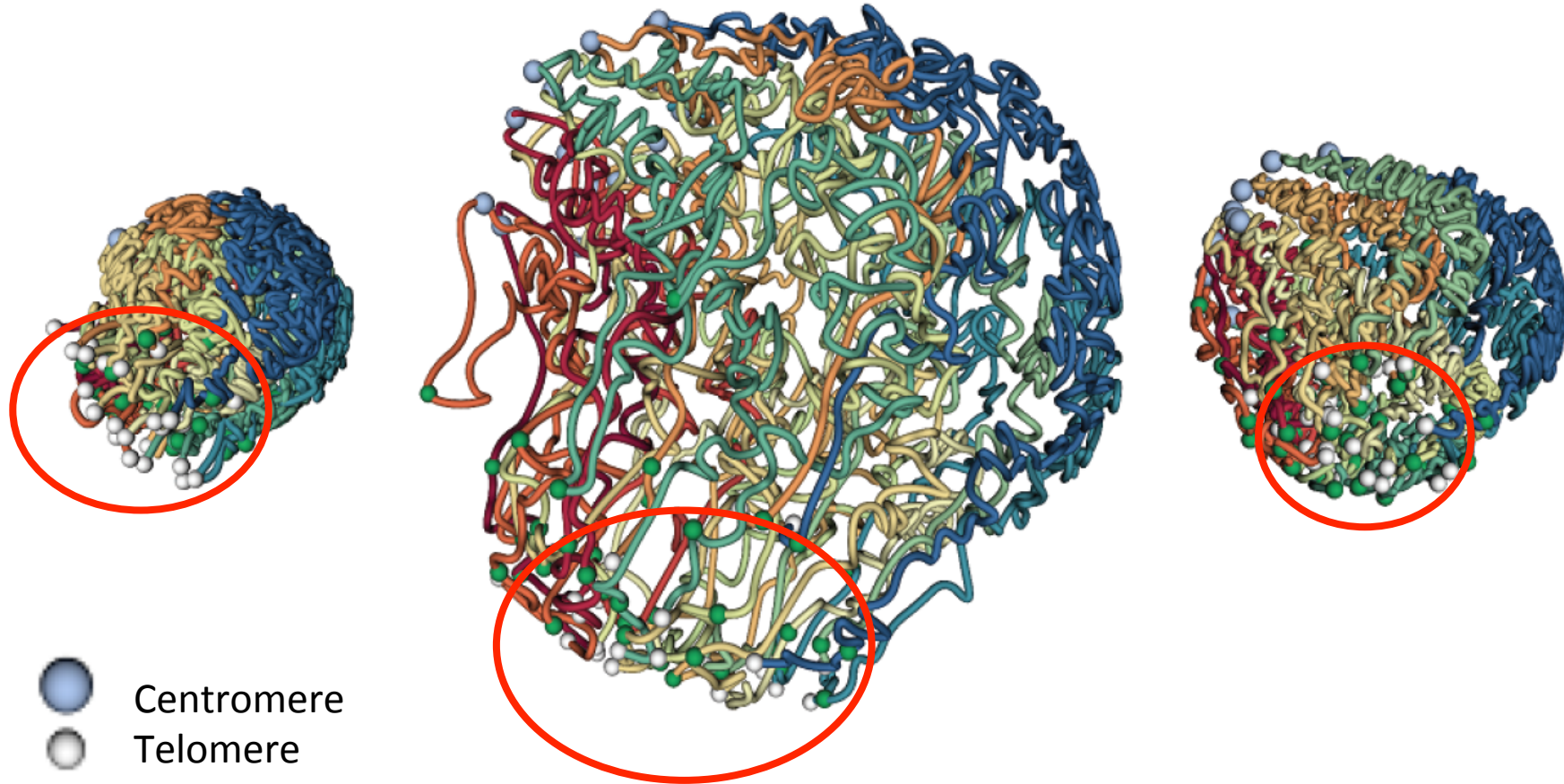
P. falciparum genome structure dynamic
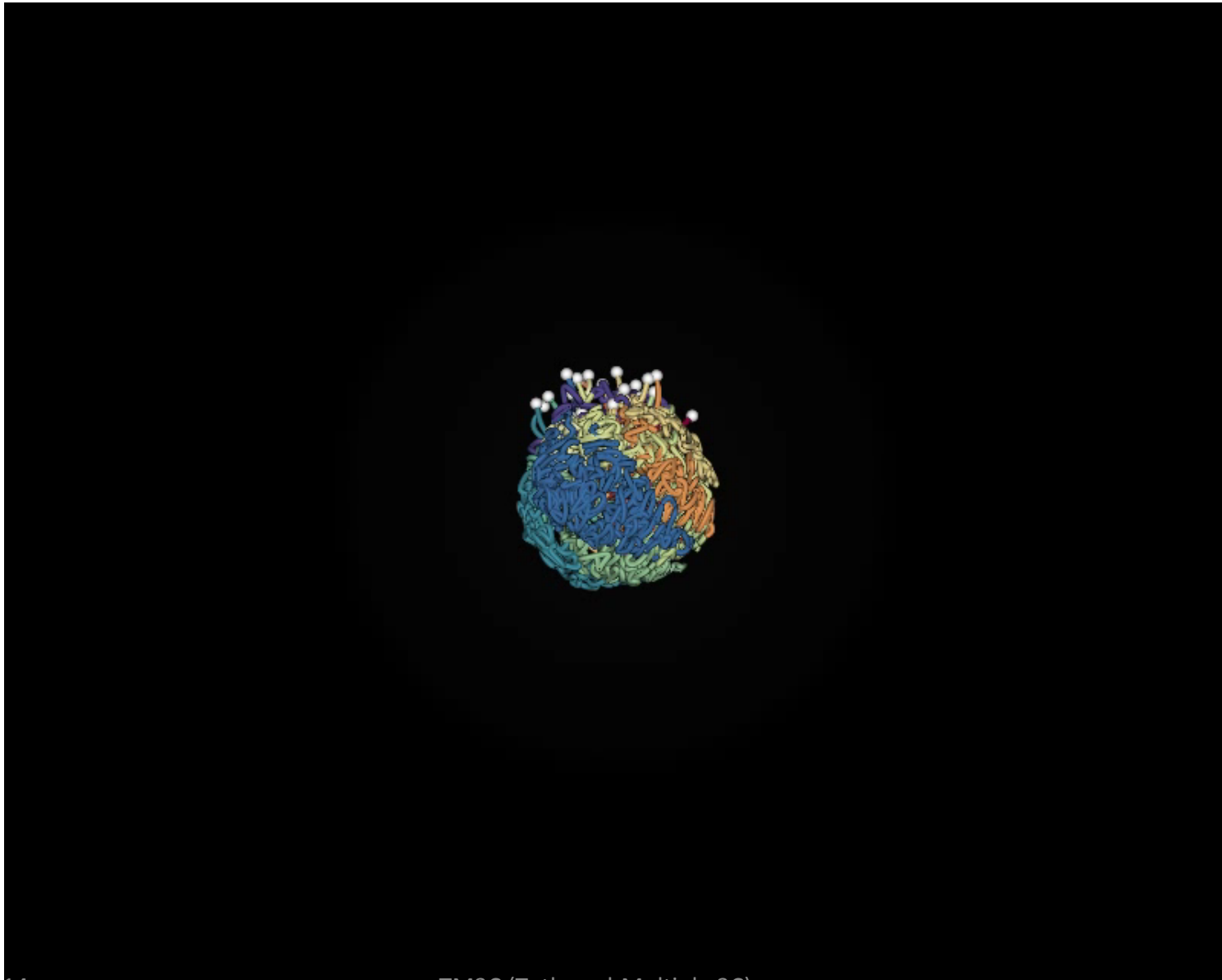
Centromeres colocalize in 3D
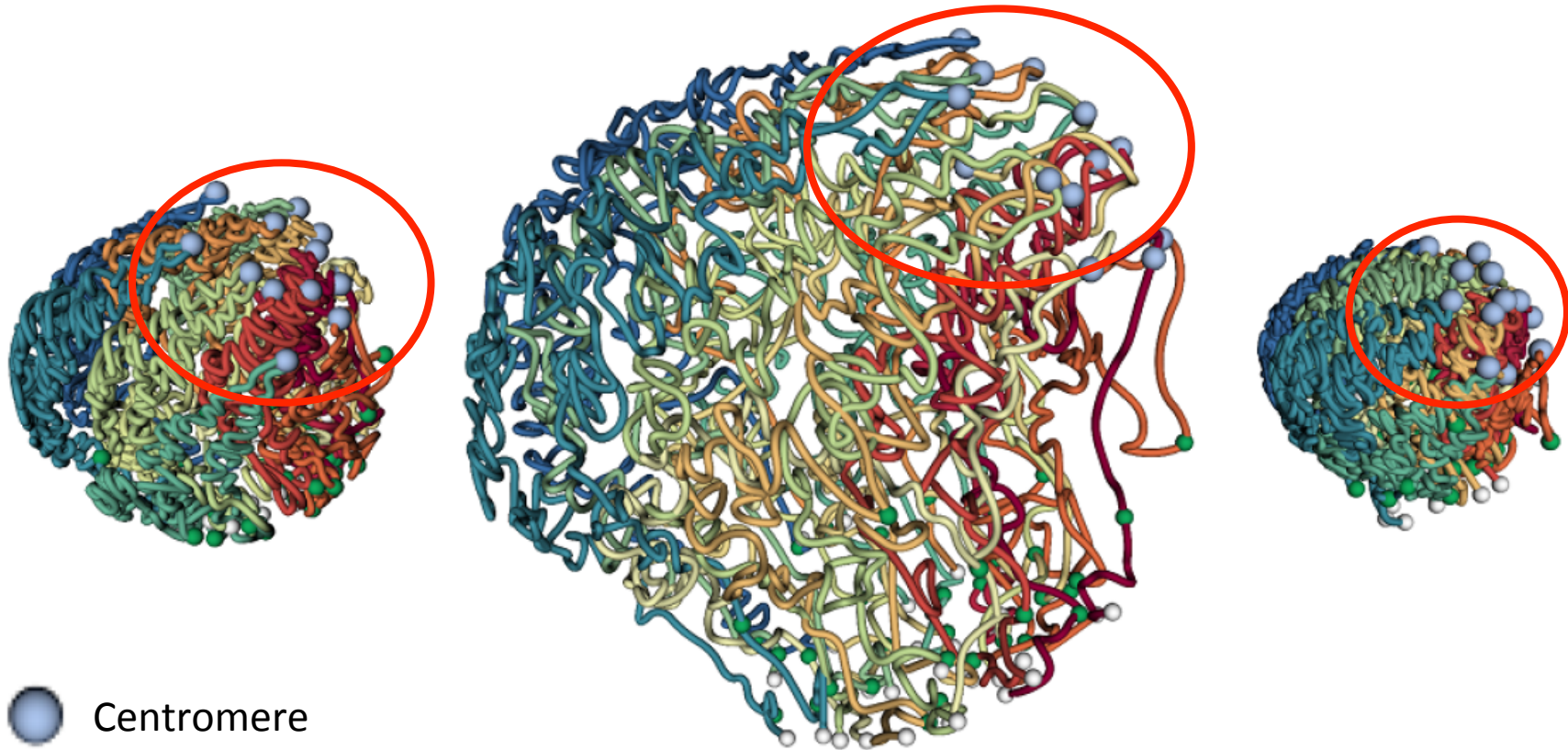
Centromere

Telomeres colocalize in 3D
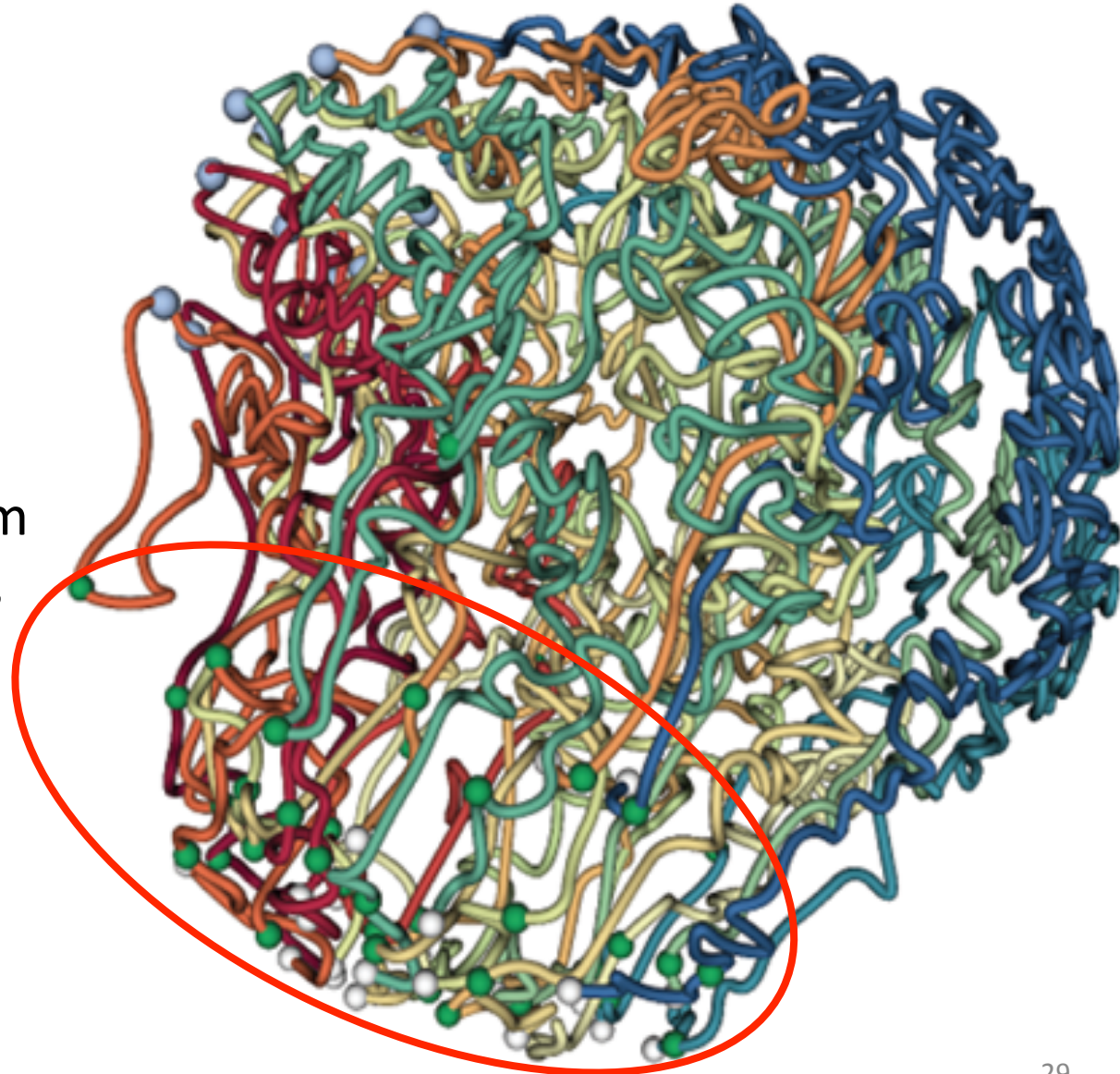

Centromere
Telomere

Virulence gene clusters colocalize in 3D

- *Plasmodium* encodes 60 virulence genes.

- Exactly one gene is expressed per cell.

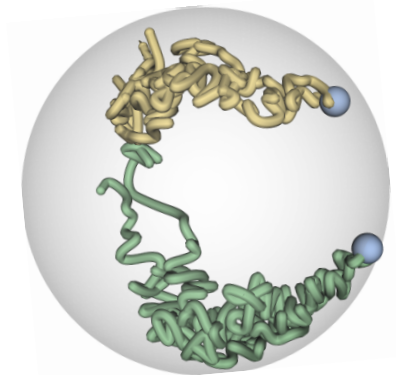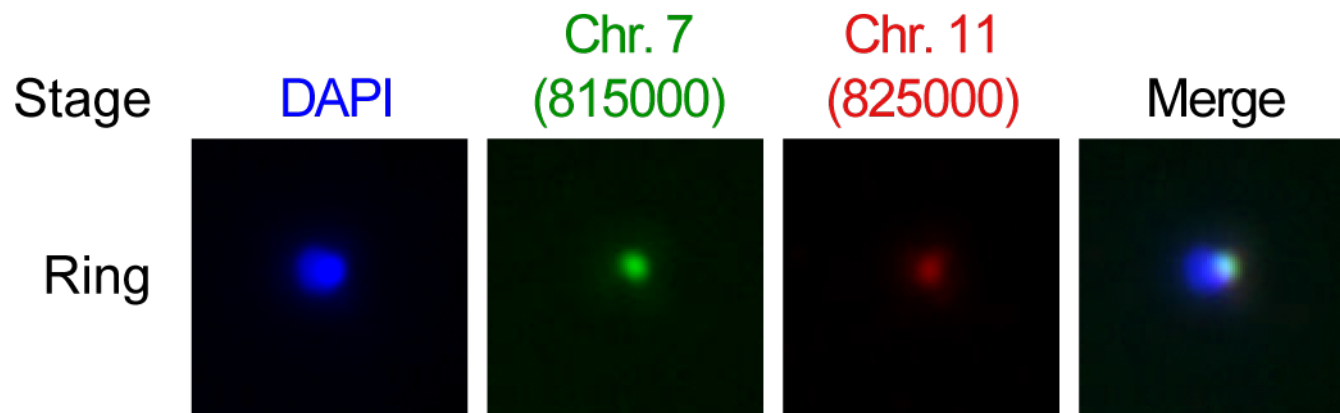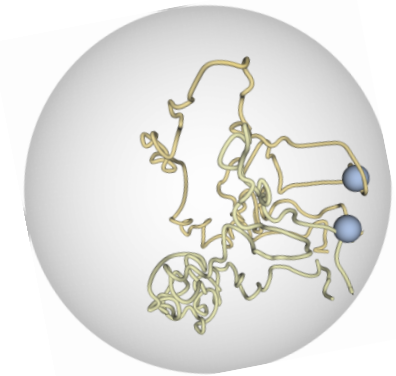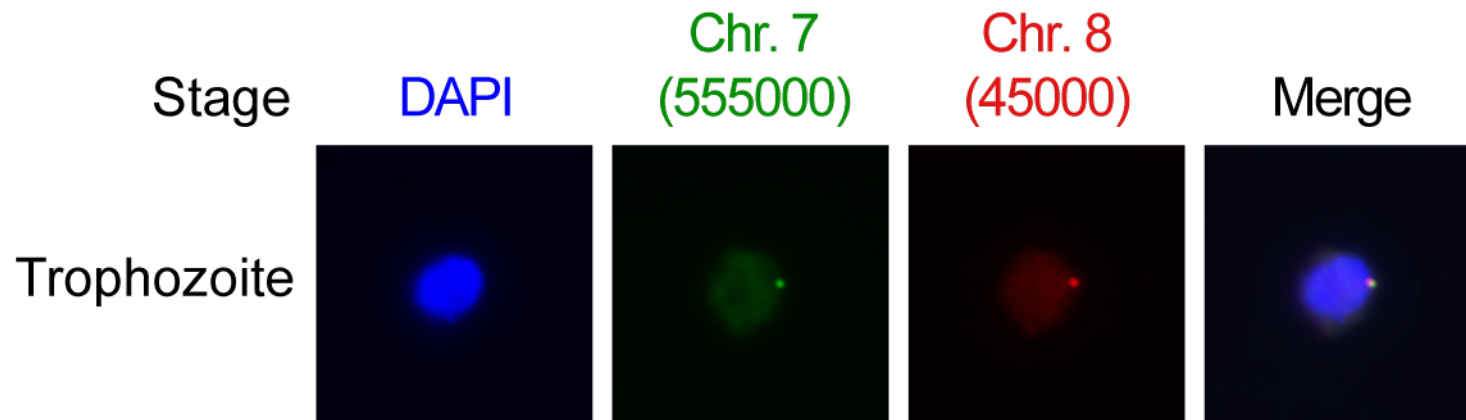- Regulatory mechanism of repression involves H3K36me3.

  Jiang et al. *Nature* 2013.
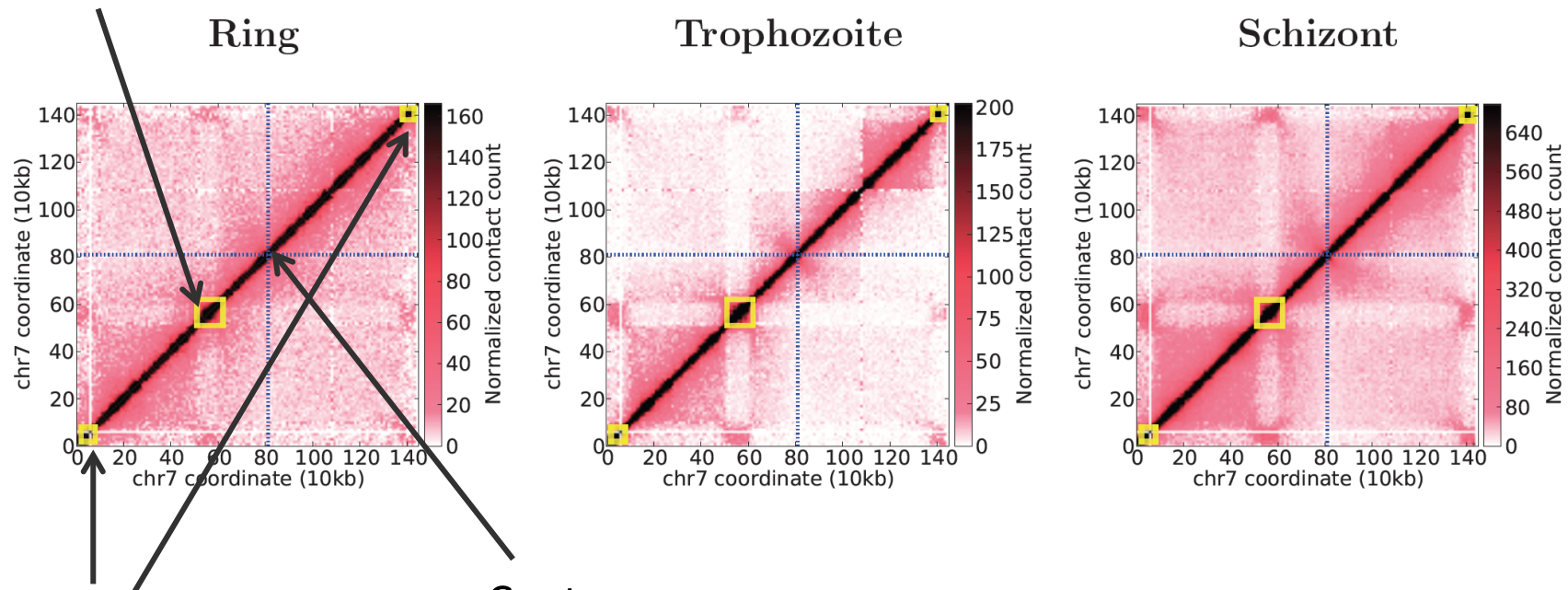

- Centromere
- Telomere
- Virulence gene cluster

DNA FISH confirms selected contacts

Inter-chromosomal pair of virulence genes

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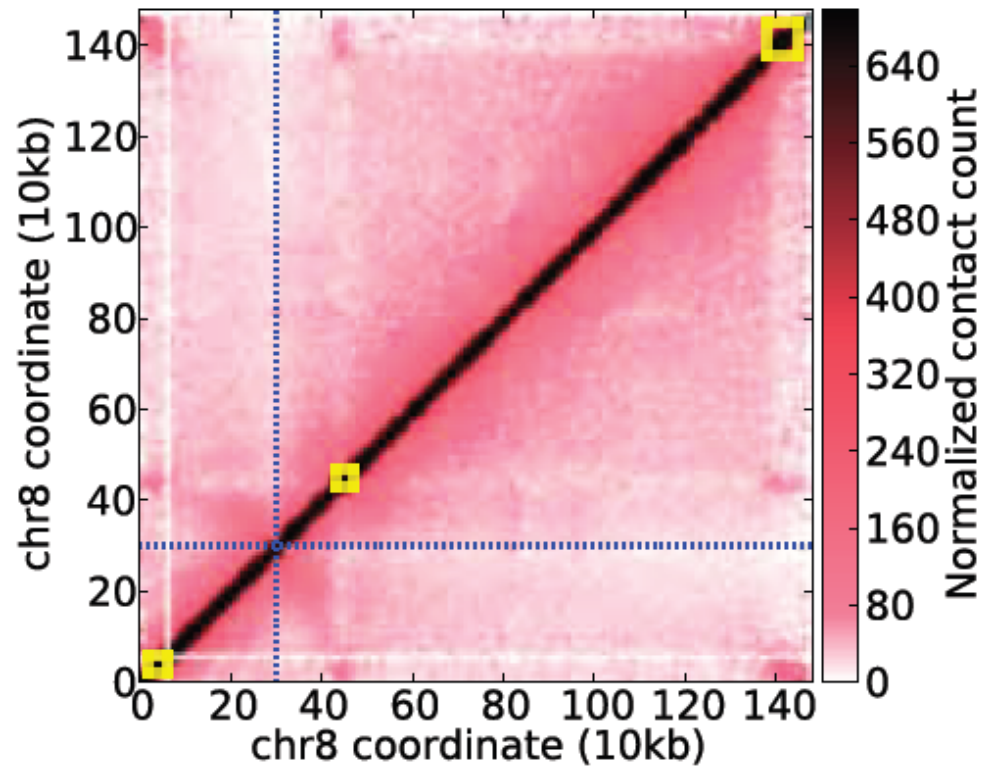
Internal virulence
gene clusters

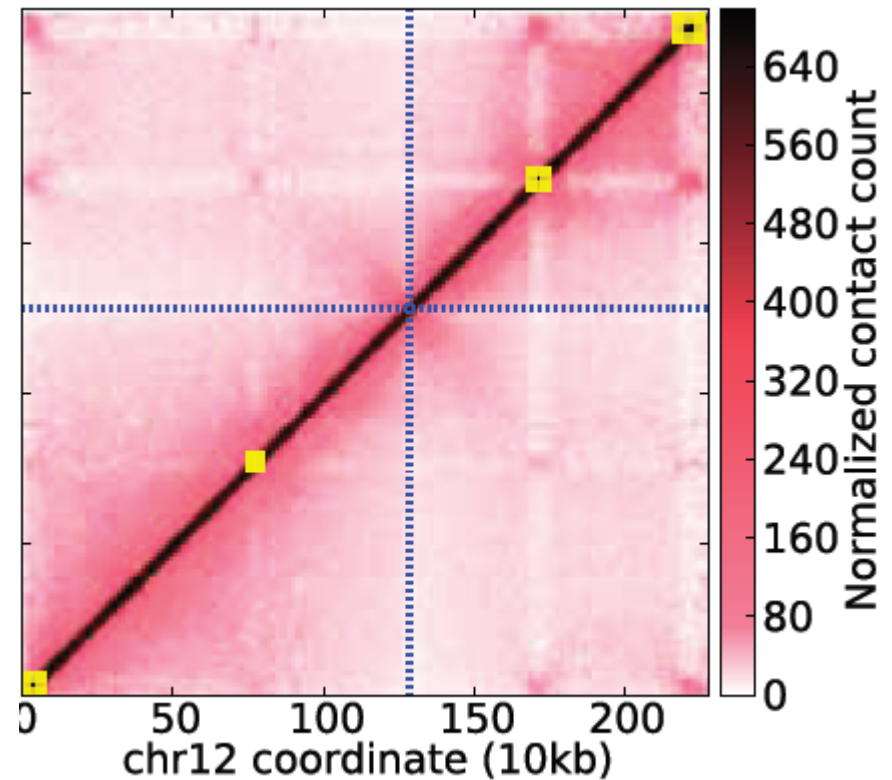Sub-telomeric virulence
gene clusters

Centromere

Chromosome 7

The pattern is consistent across chromosomes

Chromosome 8

Chromosome 12

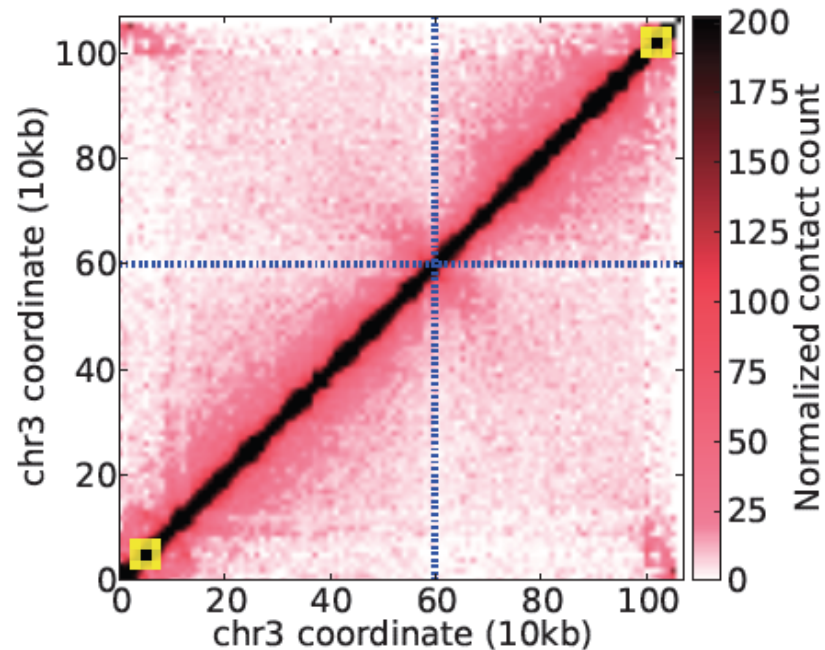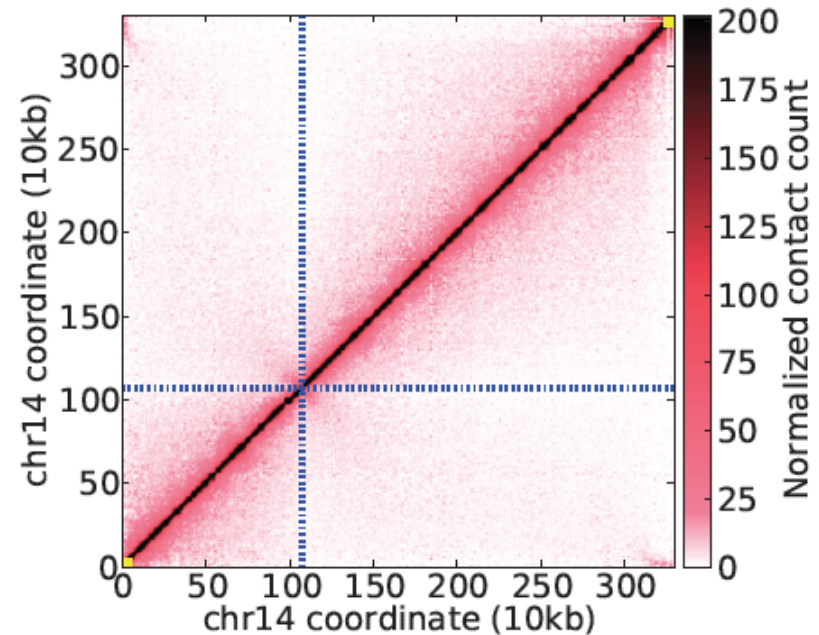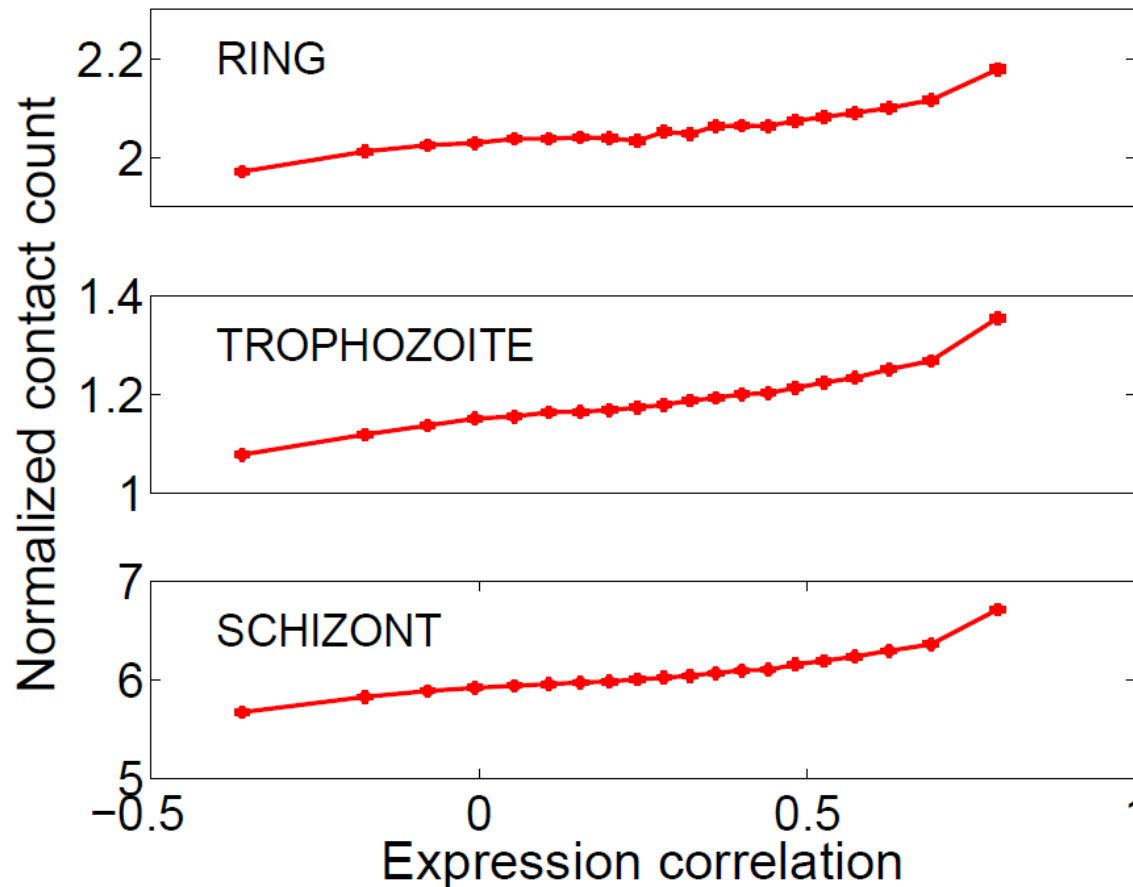# ... and absent in chromosomes with no internal virulence gene clusters

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.

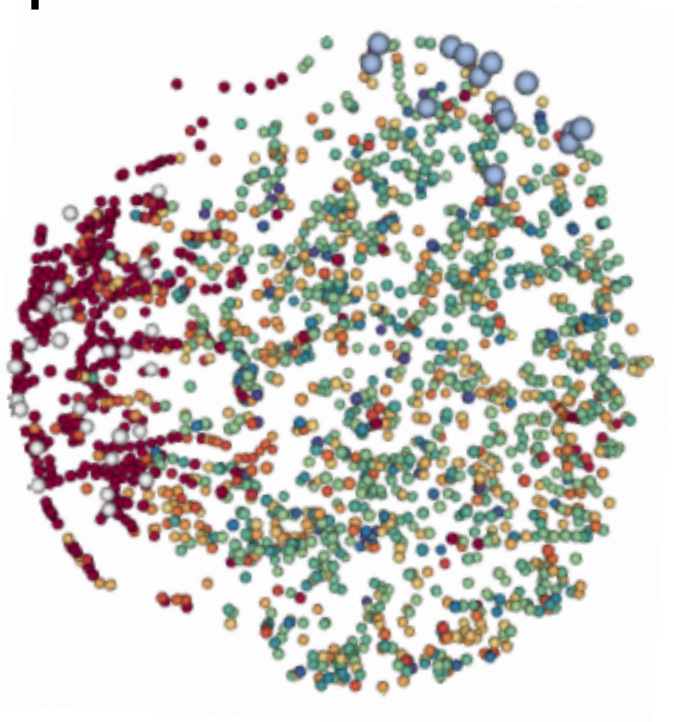# Telomeres have a repressive effect on gene expression
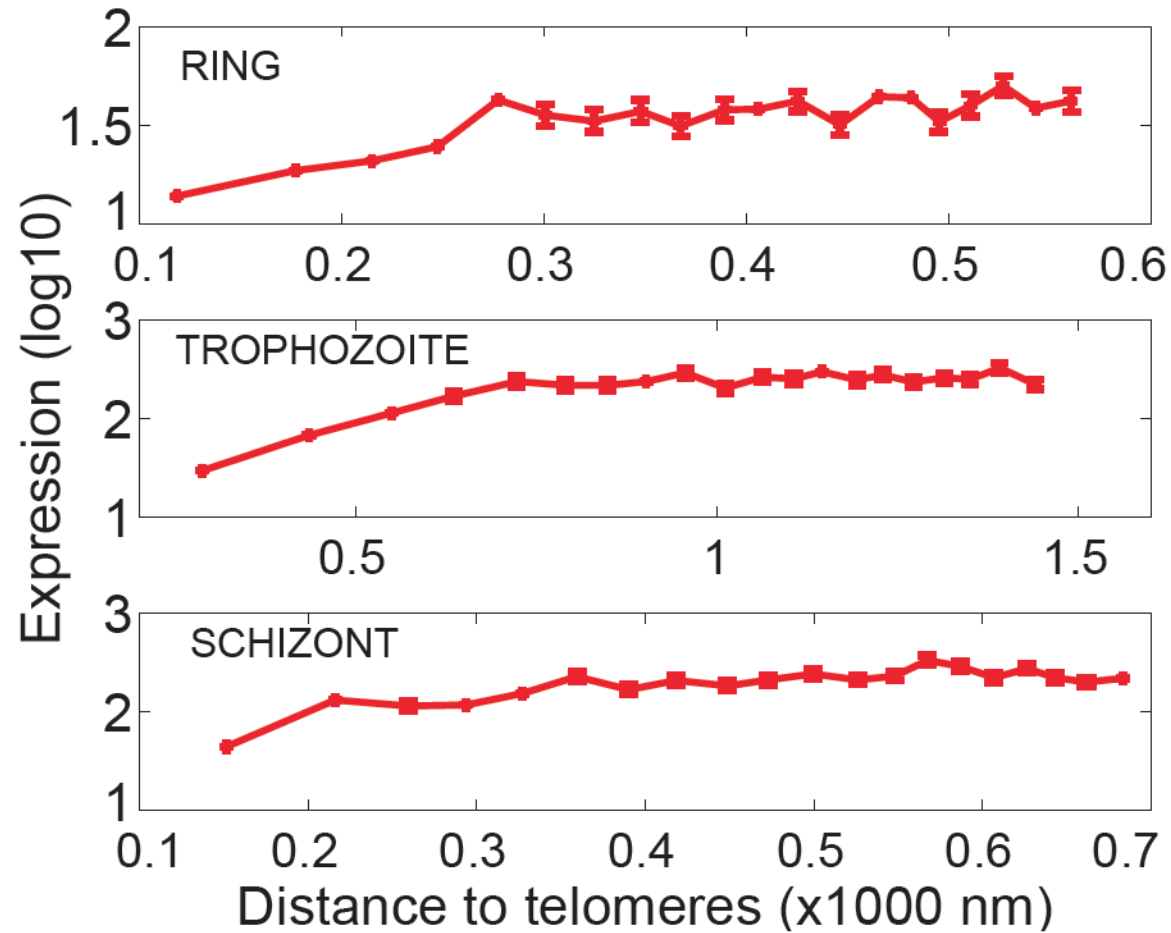
Gene expression variation exhibits a gradient across the structure
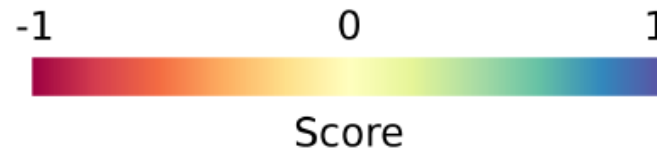
Trophozoite

Telomeric

- Antigenic variation

- Sexual stage genes

Non-telomeric

- Translation

- Trophozoite genes

Centromere   Telomere

-1   0   1

Score

Kernel Canonical Correlation
Analysis (kCCA)

36

Conclusions

✓ **Plasmodium genome architecture exhibits strong clustering of:**
  - centromeres, telomeres, virulence genes, and highly transcribed rDNA units.

✓ **Changes in power-law fits and chromosomal territories support a closed-open-closed model of the chromatin.**

✓ **Virulence gene clusters exhibit domain-like behavior.**

✓ **Genes with similar expression profiles tend to be in close spatial proximity.**

Plasmodium **species may be excellent model organisms to study the impact of genome structure on gene regulation.**

F. Ay, E. M. Bunnik, N. Varoquaux, S. M. Bol, J. Prudhomme, J.-P. Vert, W. S. Noble and K. G. Le Roch, "Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression » *Genome Research*, 24:974-988, 2014.
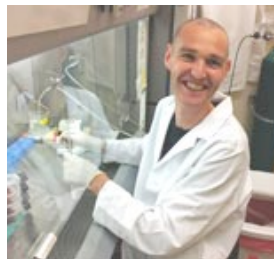
Acknowledgements

University of Washington
William Noble
Ferhat Ay

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

Ecole Des Mines, France
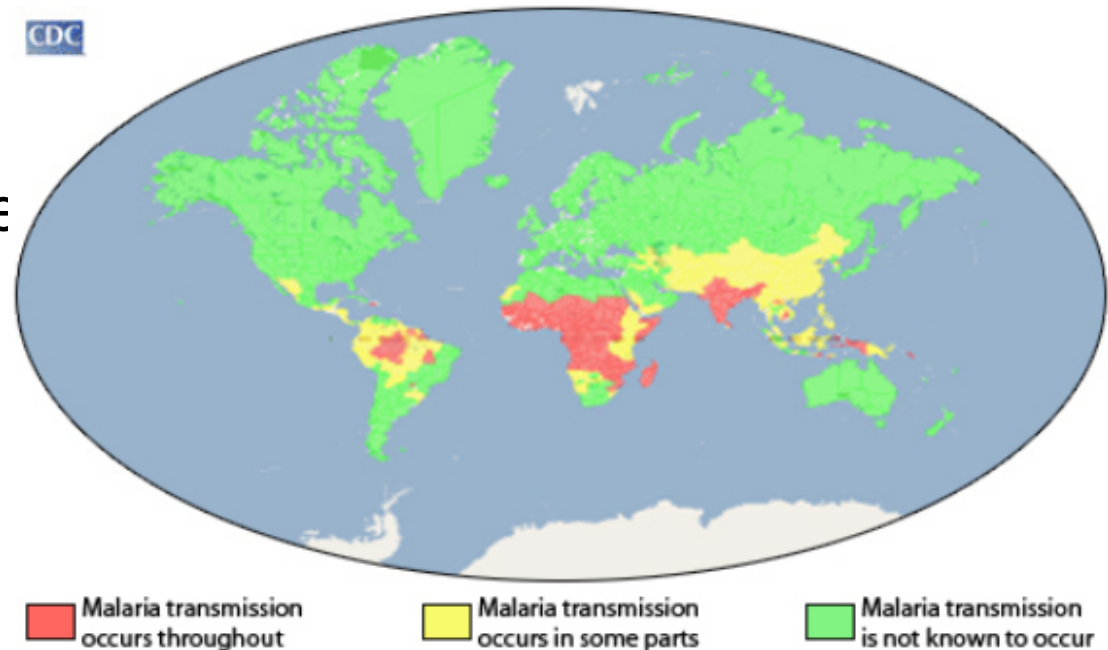Jean-Philippe Vert
Nelle Varoquaux

Funding

APPENDIX

- *Plasmo-3D*

Malaria facts

- About 3.3 billion people are at risk of malaria. [1]

- In 2010, ~219 million cases and ~660 000 deaths. [1]

- In sub-Saharan Africa over 75% of cases were due to *P. falciparum*. [2]
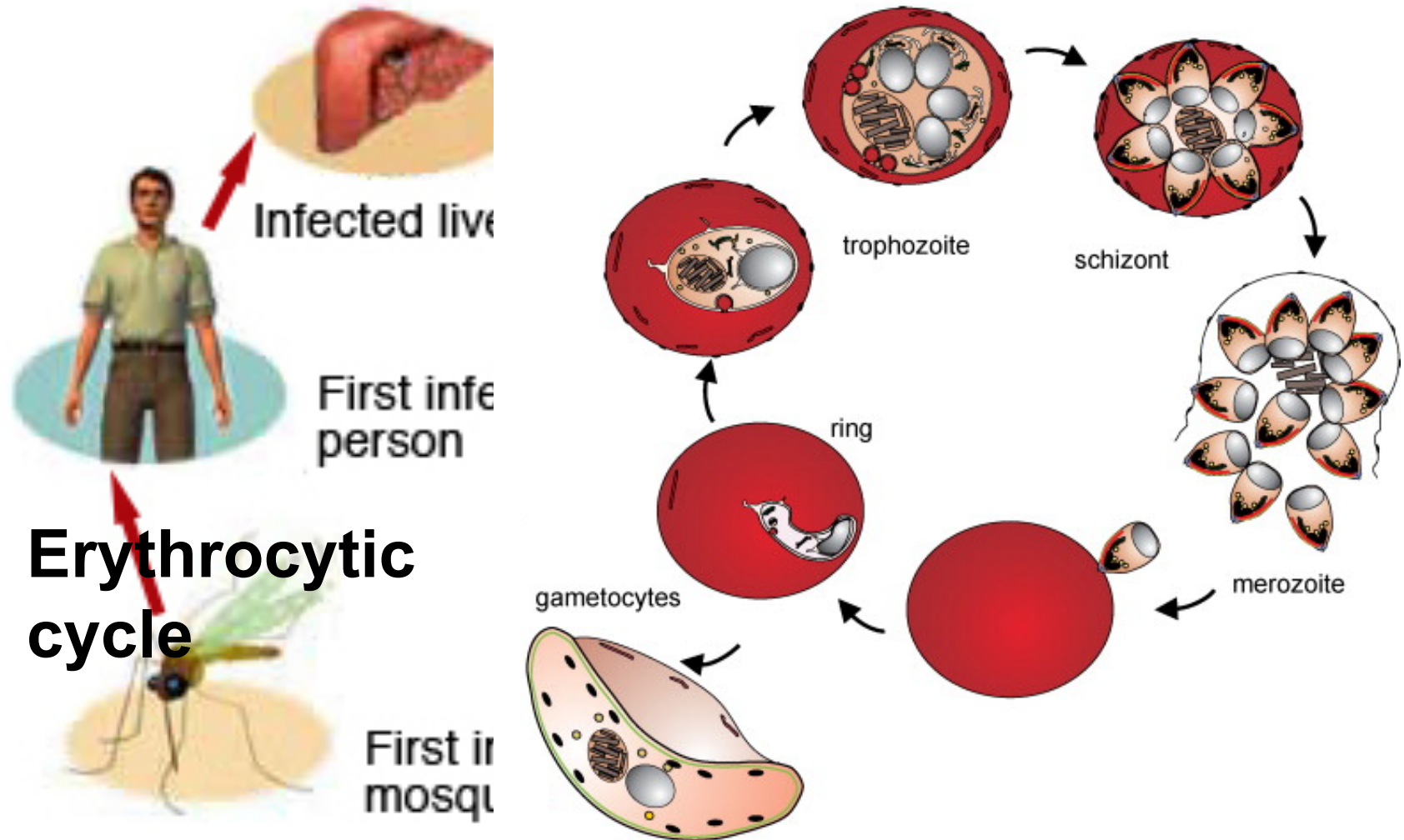
- Almost every malarial de

Malaria transmission occurs throughout

Malaria transmission occurs in some parts

Malaria transmission is not known to occur

[1] http://www.cdc.gov/malaria/about/facts.html

[2] World Malaria Report 2008, WHO.

Malaria transmission cycle

Infected live[r]

First infe[cted]
person

Erythrocytic cycle

First i[n]
mosq[uito]

trophozoite

schizont

merozoite

ring

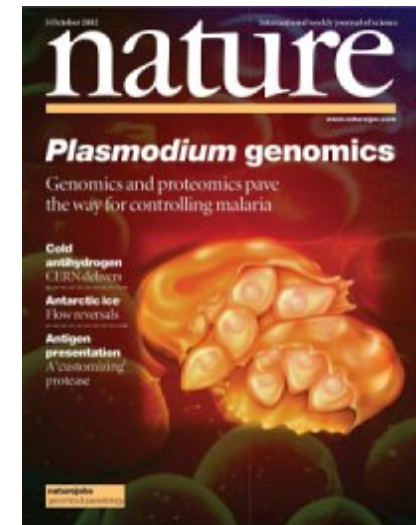gametocytes

The first malaria parasite genome was published over a decade ago

- 23.26 Mb in size.

- Haploid.

- 14 nuclear chromosomes, 1 mitochondrion, 1 plastid.

- ~6372 genes and 5524 protein coding genes.

- The most AT rich genome to date (~80%).
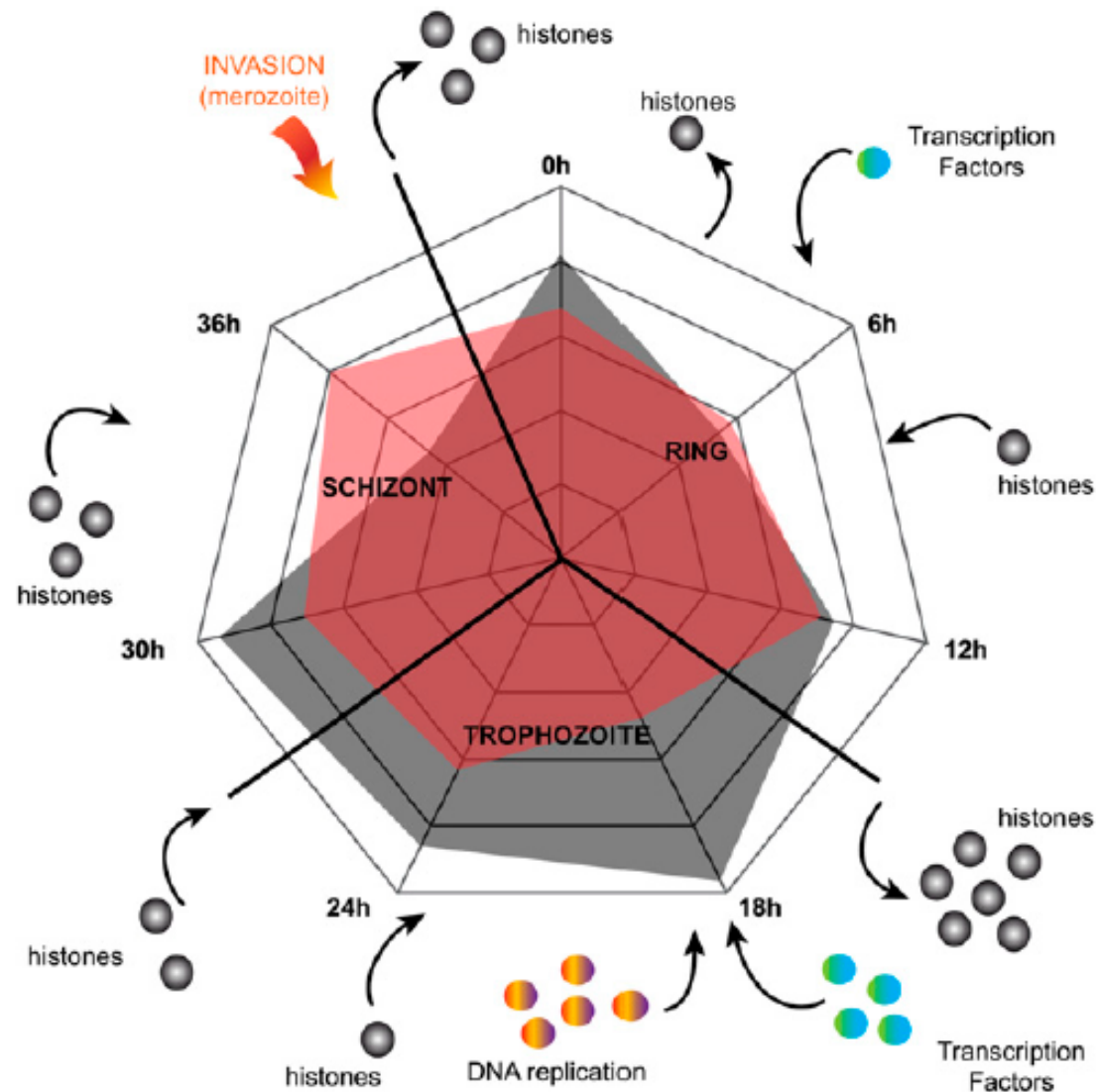
- 47% are still hypothetical proteins.

M. J Gardner et al. *Nature* 2002

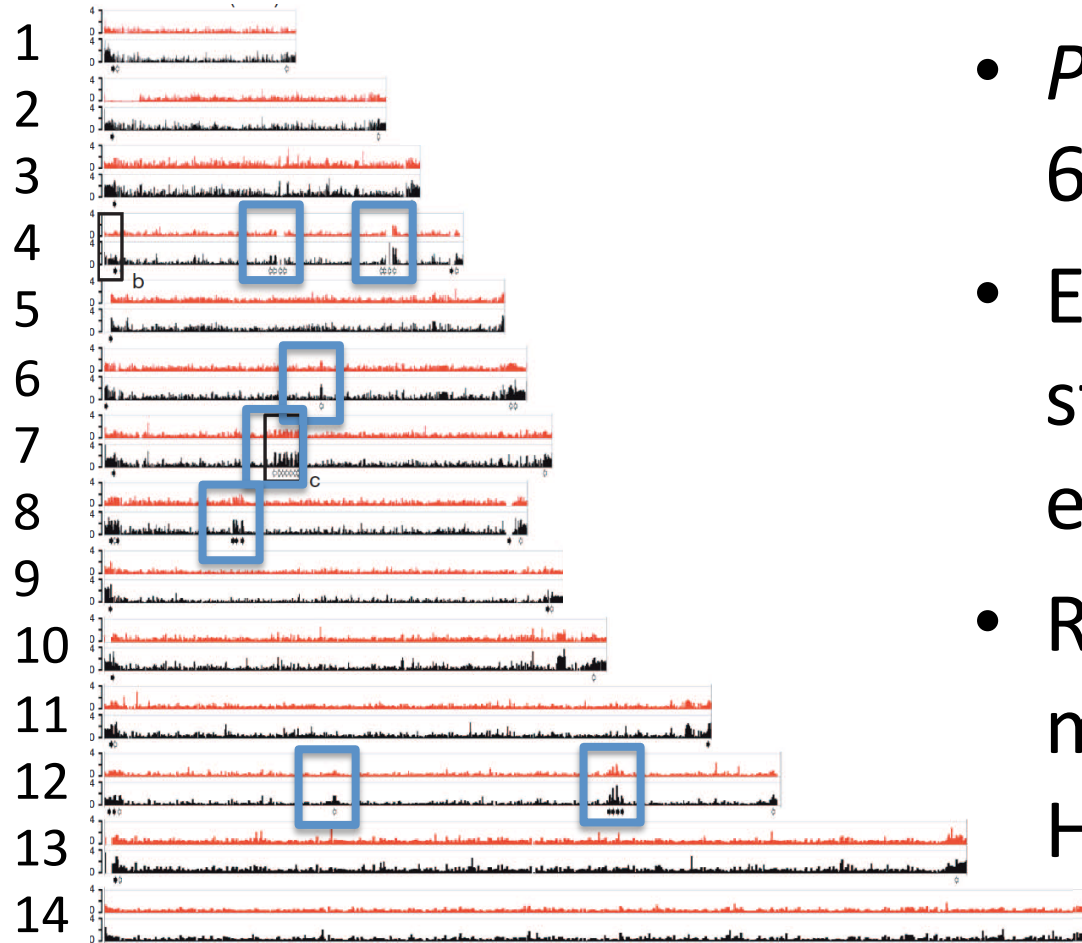# *Plasmodium* chromatin architecture goes through systematic changes

FAIRE (Open chromatin)

MAINE (Closed chromatin)

Ponts et al.
Genome Research 2010.

43

Virulence genes are tightly regulated for precise expression patterns
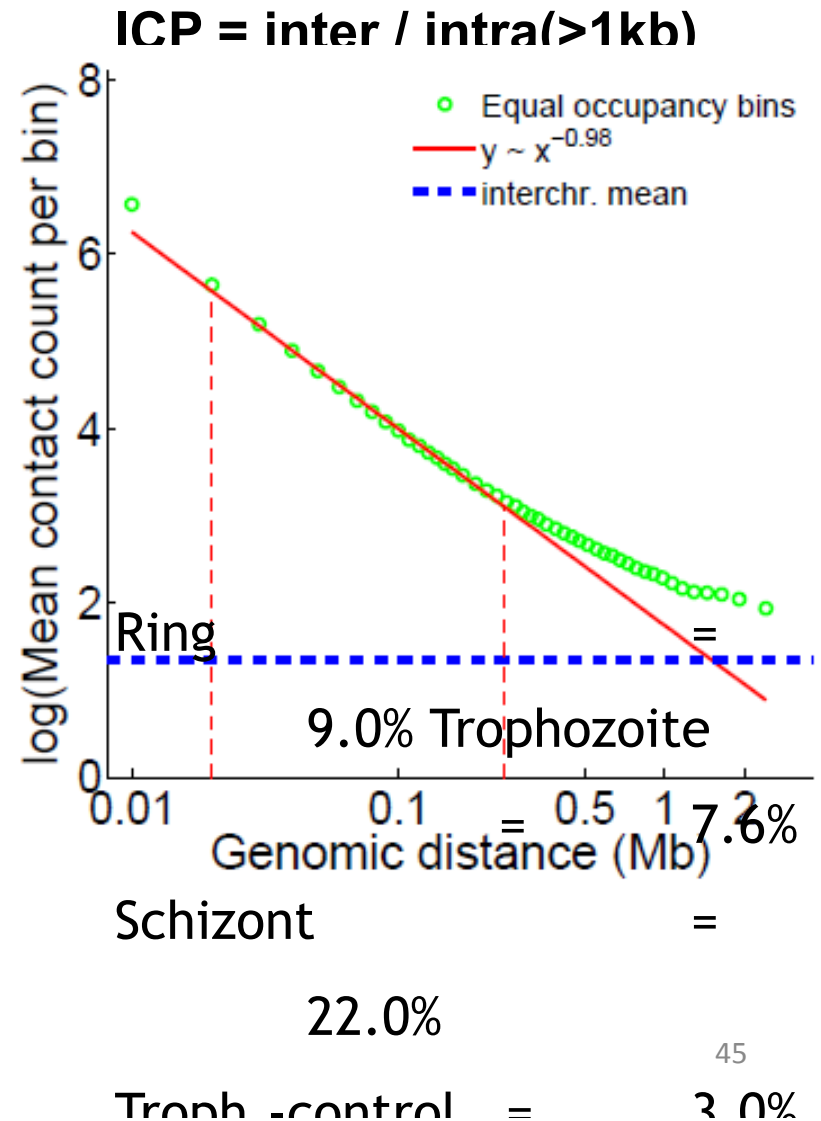

- *Plasmodium* encodes 60 virulence genes.
- Exactly one gene is stochastically expressed per cell.
- Regulatory mechanism involves H3K36me3.

Jiang et al. *Nature* 2013.

Our modified Hi-C protocol works for the AT-rich *Plasmodium* genome

1) ICP index: Relatively low numbers of **inter-chromosomal** contacts from crosslinked samples with respect to random expectation and non-crosslinked control.

2) Contact probability between two **intra-chromosomal** loci exhibits a log-linear decay with increasing genomic distance.
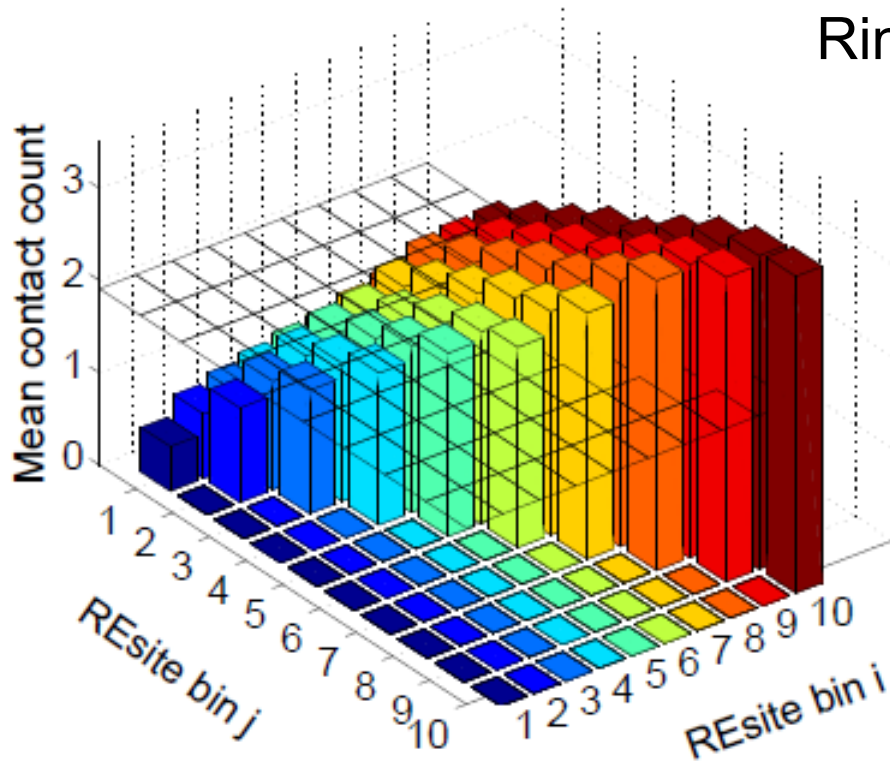
3) The percentage of long-range contacts (either interchromosomal or intrachromosomal >20 kb) is higher than control and comparable to previous Hi-C data from other organisms.

ICP = inter / intra(>1kb)

- ○ Equal occupancy bins
- — $y \sim x^{-0.98}$
- - - interchr. mean

y-axis: log(Mean contact count per bin)

x-axis: Genomic distance (Mb) — 0.01, 0.1, 0.5, 1, 2

Ring =

9.0% Trophozoite

7.6%

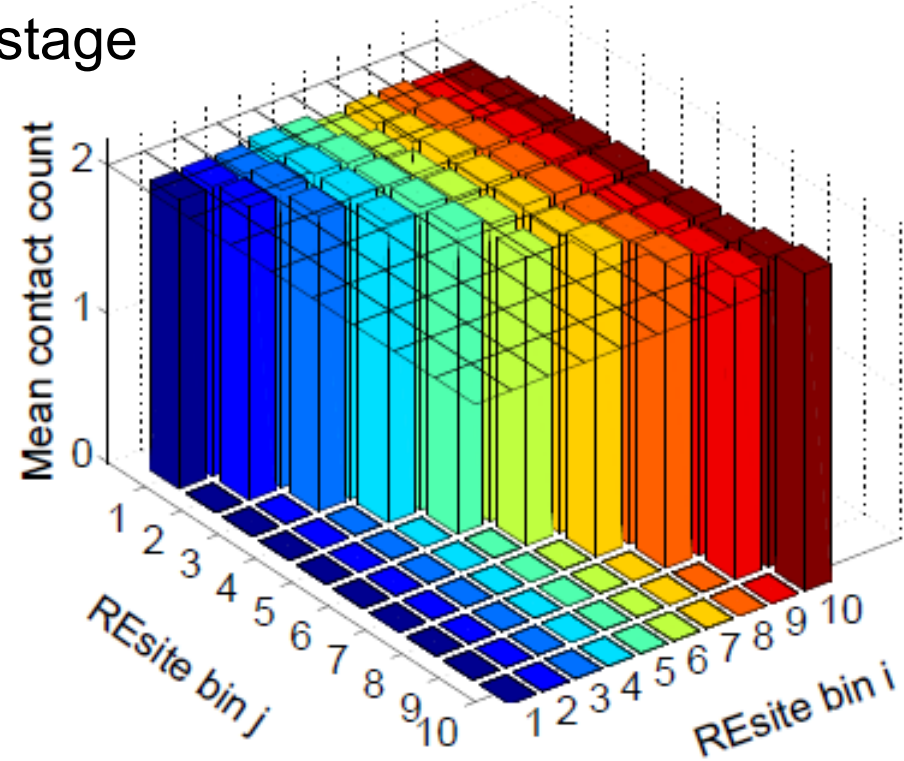Schizont =

22.0%

Troph.-control = 3.0%

Our data exhibits characteristics biases of Hi-C
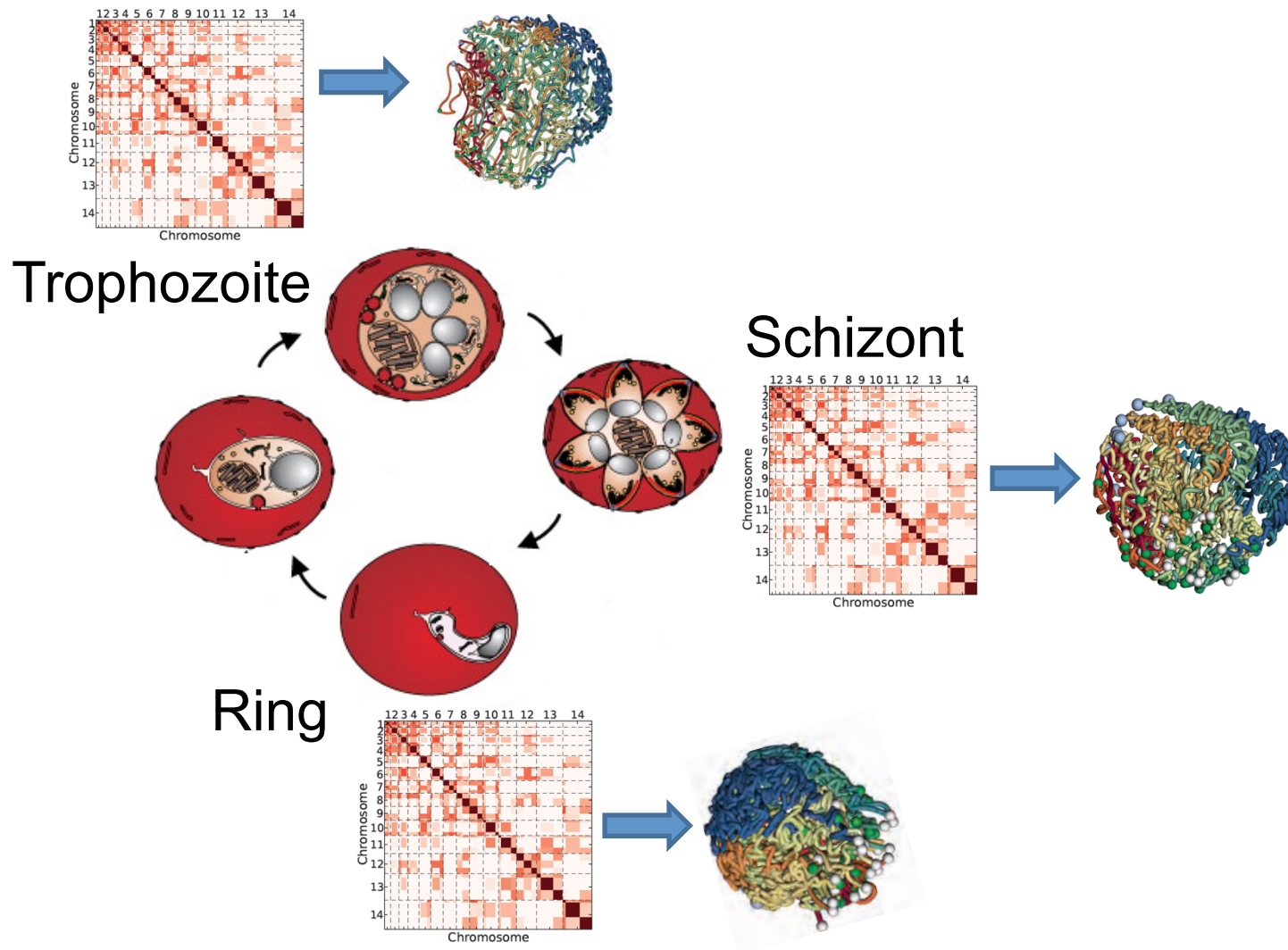
Ring stage

Before normalization

After normalization
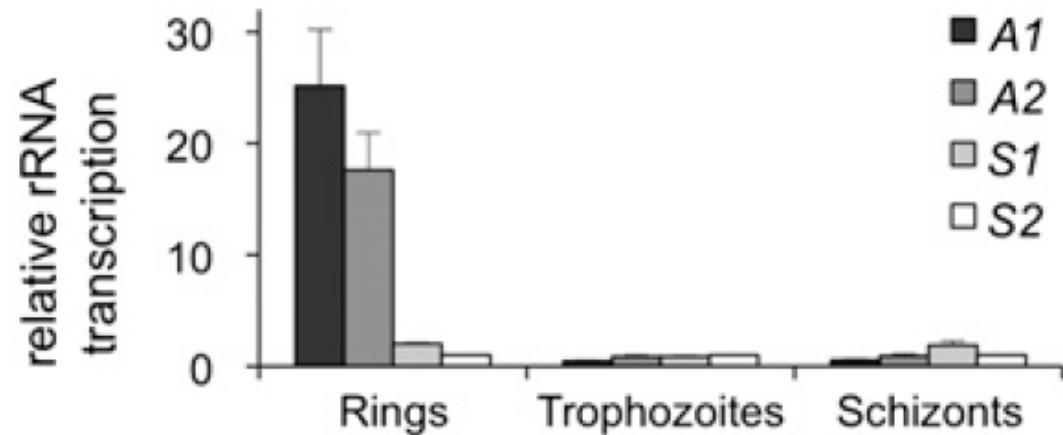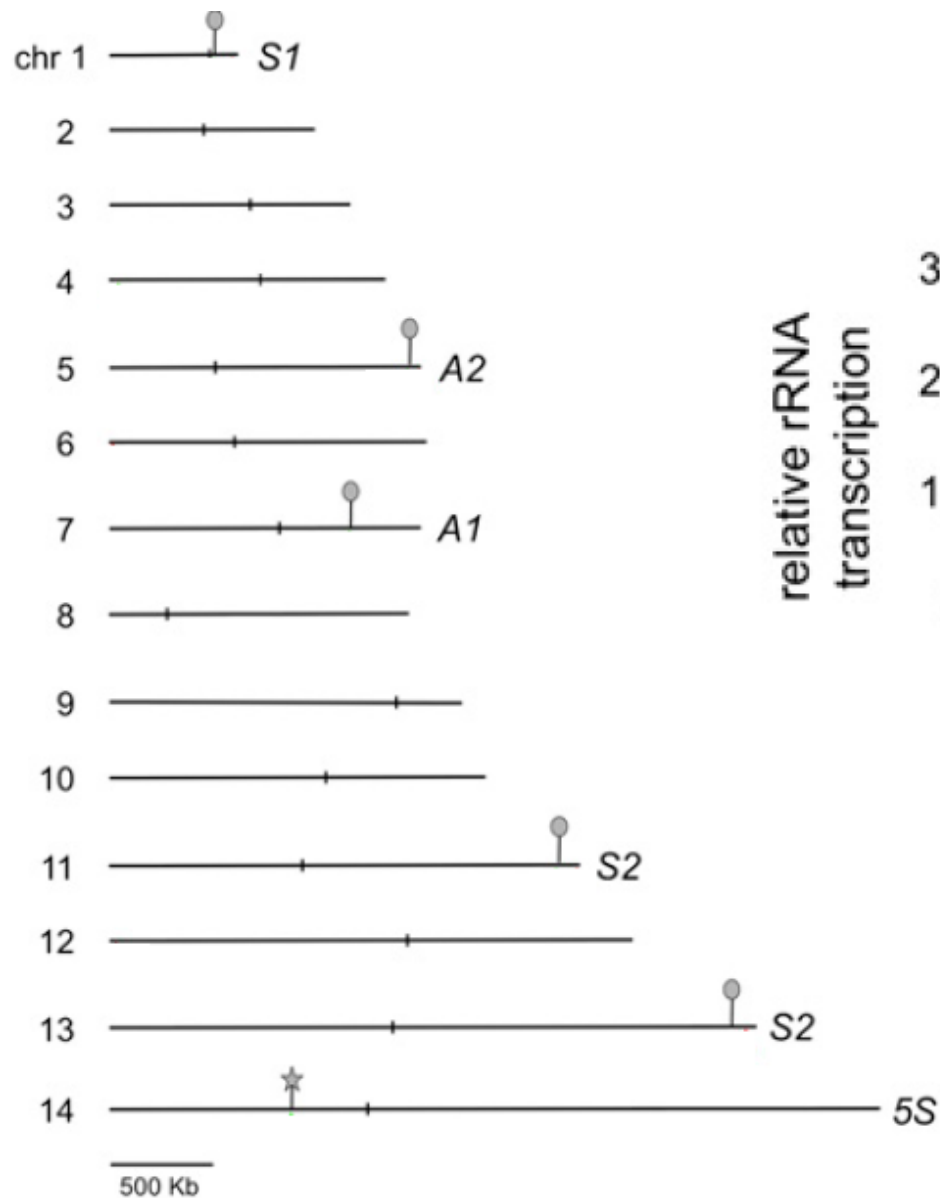
Larger number of restriction enzyme (RE) cut sites → More contacts

Other biases: GC content, mappability, visibility (in general).

Imakaev et al. *Nature Methods* 2012.

Modeling genome architecture using Hi-C contact maps

Trophozoite

Schizont

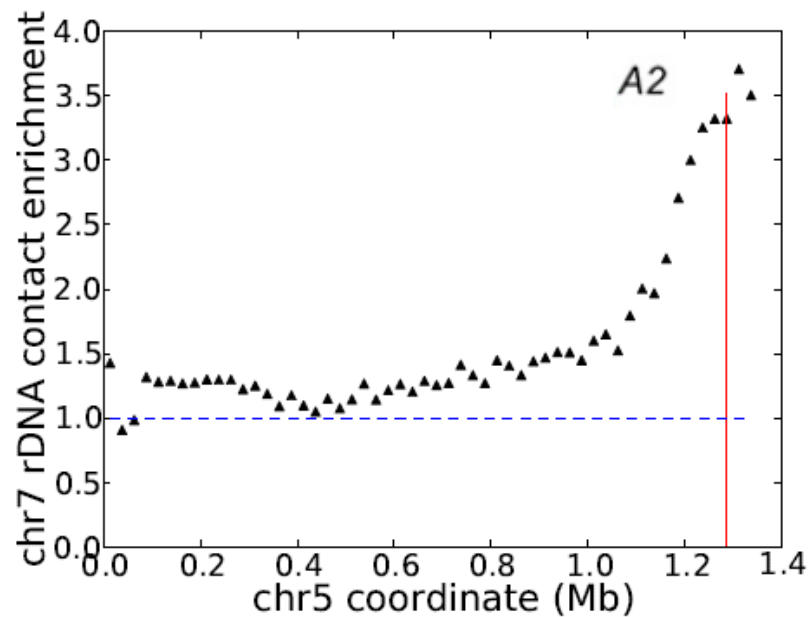Ring

Highly transcribed rRNA genes colocalize
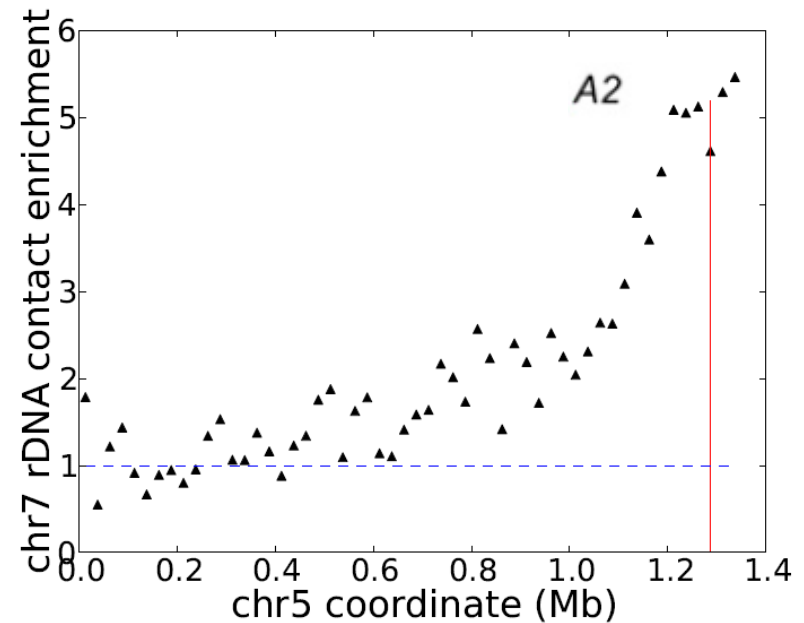

Mancio-Silva et al. *PNAS* 2010.

Highly transcribed rRNA genes colocalize

Virtual 4C plots using the A1 rDNA unit on chromosome 7 as the bait
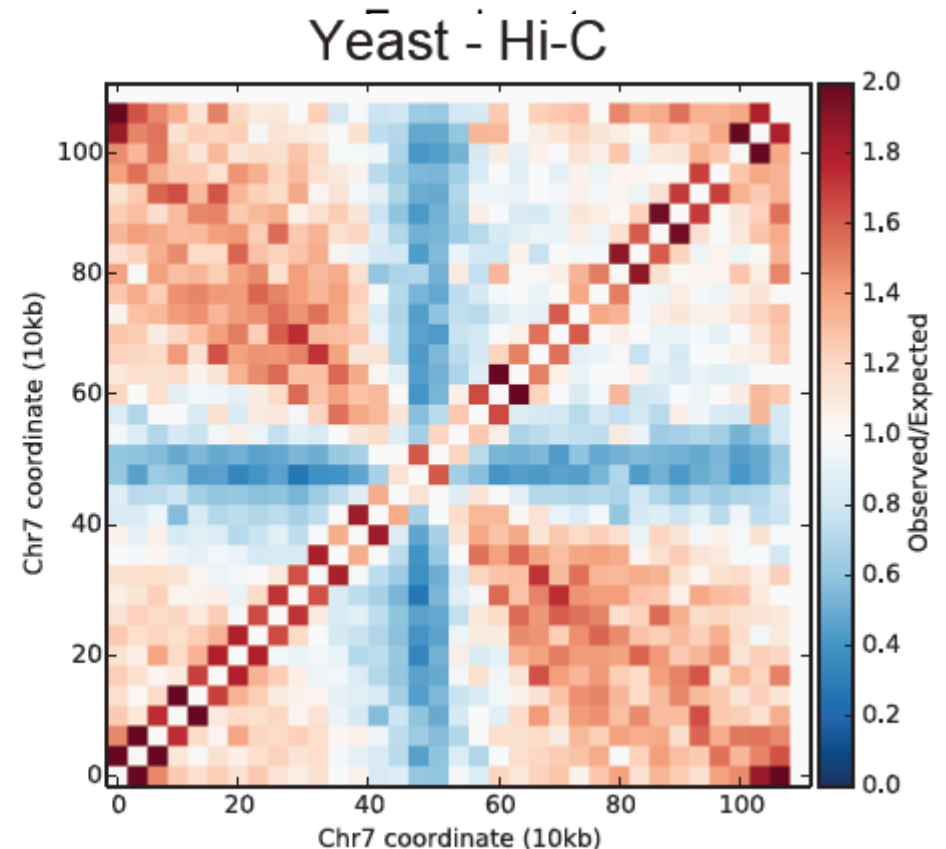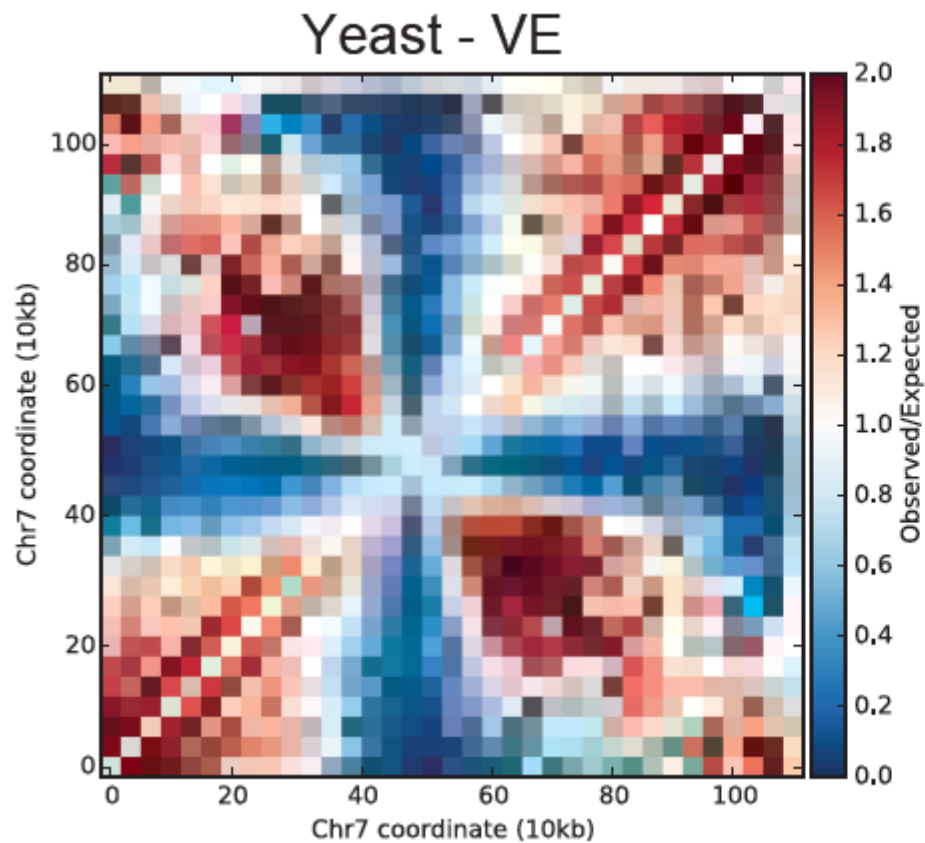

Our data – Ring stage

B15C2 cells - Ring stage
Lemieux et al. *Mol. Microbiology* 2013

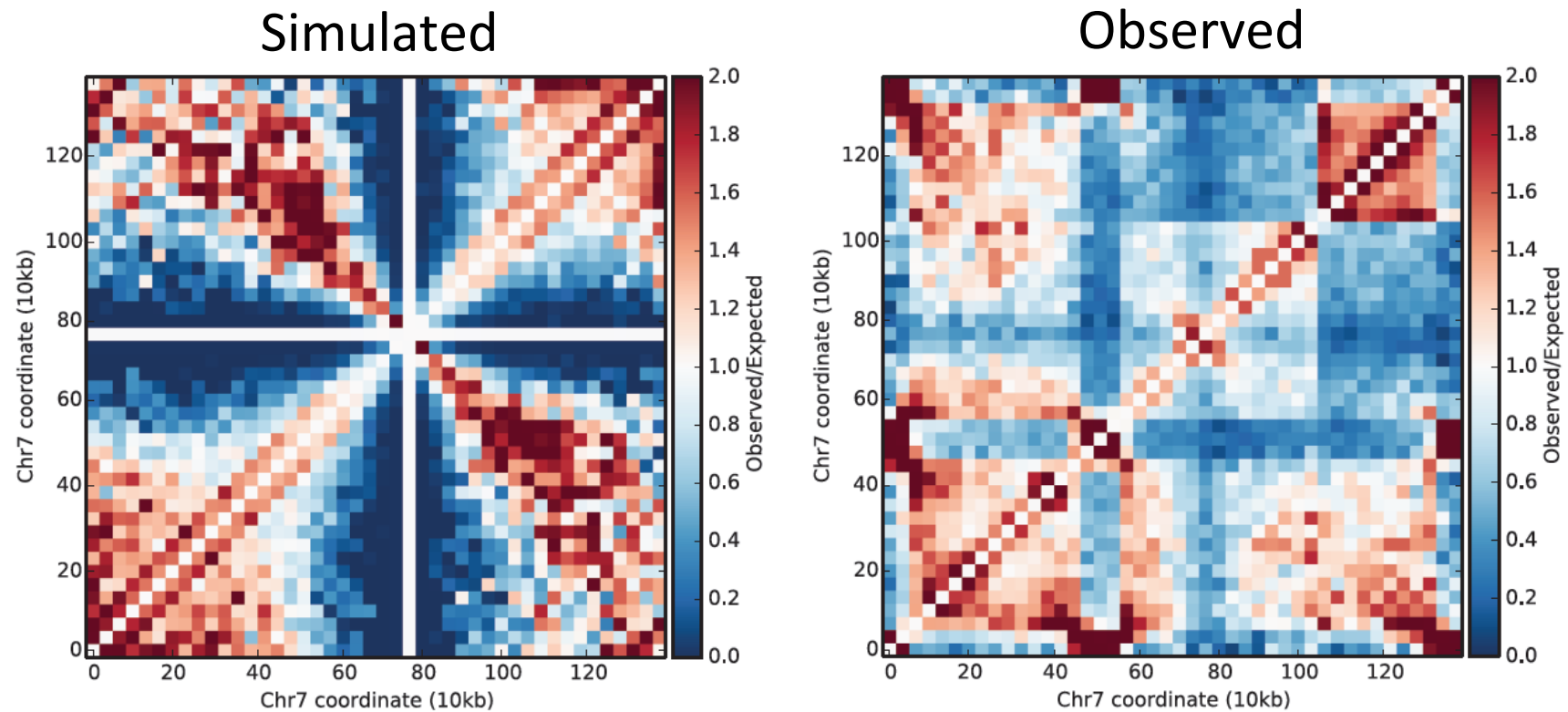# How unexpected/non-random is *Plasmodium* genome architecture?

Simulated and observed chromatin contacts are highly concordant in yeast

Tjong et al. *Genome Research* 2011.

Duan et al. *Nature* 2010.

Volume exclusion modeling does not capture *Plasmodium* architecture

Simulated

Observed

Chromosome 7, trophozoite stage

So far

✓ Assayed 3 time points using Hi-C.

✓ Generated consensus 3D models.

✓ Validated our models using FISH and prior knowledge.

✓ Showed that simple volume exclusion does not explain *Plasmodium* genome architecture.

✓ Demonstrated existence of domain-like structures shaped around virulence genes.

Videos

Future directions

✓ Mechanism behind formation of repressive/active compartments in *Plasmodium*.

✓ Causality between virulence gene clustering and transcriptional silencing.

✓ Interruption of genome architecture changes to interfere with parasite cell cycle for development of antimalarial drugs.