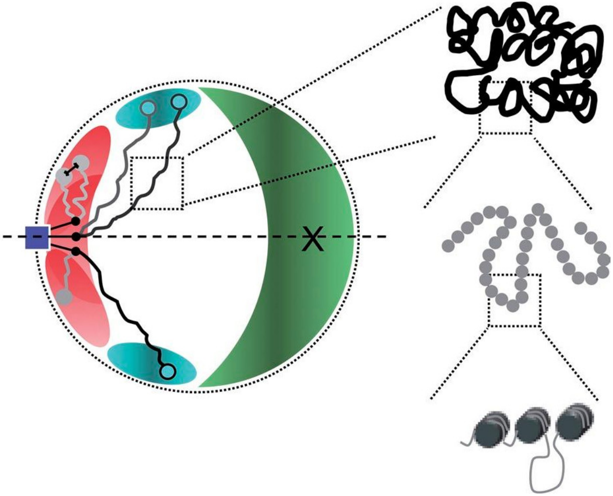
# Studying the 3D structure of the *P. falciparum*'s genome by modeling contact counts as random Negative Binomial variables.

Nelle Varoquaux

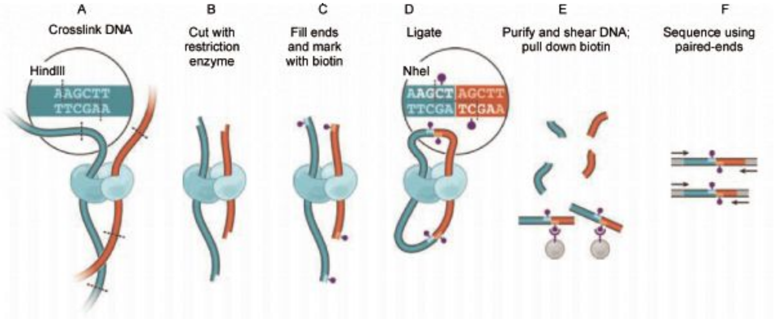
*with Kate Cook, Evelien Bunnik, Ferhat Ay,
Karine LeRoch, William Stafford Noble,
and Jean-Philippe Vert*

# The 3D structure of the genome is thought to play an important role in many biological processes



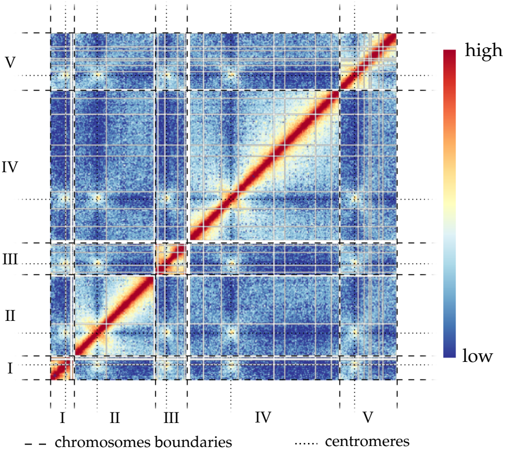
**The genome of *S. cerevisiae* is highly organized** [Zimmer and Fabre, 2011]

# The Hi-C protocol identifies physical contacts between pairs of loci genome-wide


A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
B
Cut with restriction enzyme
C
Fill ends and mark with biotin
D
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
E
Purify and shear DNA; pull down biotin
F
Sequence using paired-ends

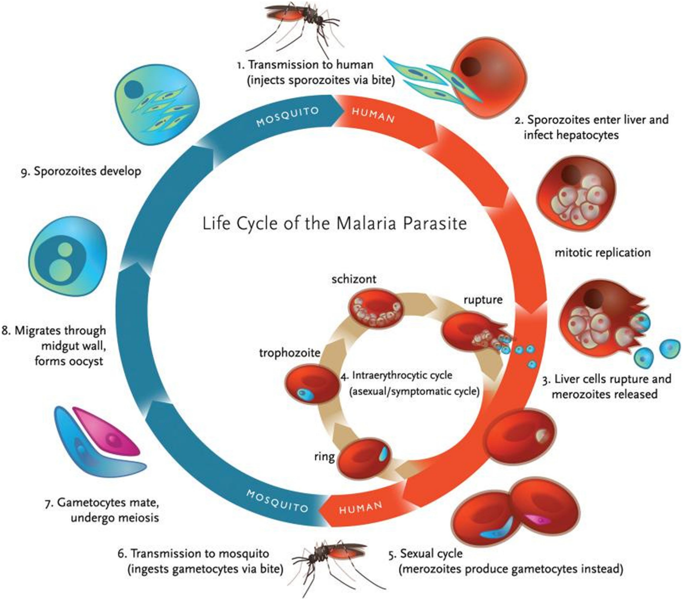
**Hi-C paves the way for a systematic and genome-wide analysis of genome architecture** [Rao et al., 2014]

# The contact count matrix recapitulates the hallmarks of genome architecture



Contact counts for the first 5 chromosomes of *S. cerevisiae*

# The human malaria parasite *P. falciparum*



1. Transmission to human
(injects sporozoites via bite)
MOSQUITO
HUMAN
2. Sporozoites enter liver and
infect hepatocytes
9. Sporozoites develop
Life Cycle of the Malaria Parasite
mitotic replication
schizont
rupture
8. Migrates through
midgut wall,
forms oocyst
trophozoite
4. Intraerythrocytic cycle
(asexual/symptomatic cycle)
3. Liver cells rupture and
merozoites released
ring
7. Gametocytes mate,
undergo meiosis
MOSQUITO
HUMAN
6. Transmission to mosquito
(ingests gametocytes via bite)
5. Sexual cycle
(merozoites produce gametocytes instead)

# Motivation: the 3D structure of *P. falciparum*

## Motivation

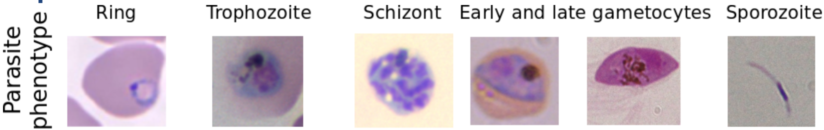
- One of the main limiting factors for the development of therapies is the poor understanding of complex gene regulation of the parasite.
- Relative paucity of specific transcription factors points towards complementary regulatory mechanisms to control gene expression.
- Chromatin remodeling enzymes are abundant in Plasmodium genomes.

## Hypothesis

- Both local and global genome architecture play an important role in *P. falciparum*'s gene regulation.

# Assessing the 3D structure changes across timepoints

### Experiments



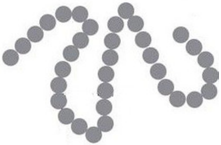
### Idea

- Inferring 3D models by modeling overdispersion of Hi-C data.
- Finding relationships between 3D models and gene expression.

# Inferring 3D structures of genome by modeling overdispersion of Hi-C data

*joint work with William S. Noble*
*and Jean-Philippe Vert.*

# Inferring 3D models of genome architecture



## Notations

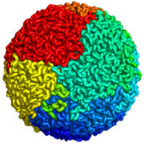
- Let $\mathbf{X} \in R^{n \times 3}$ be the coordinates of each bead.
- Let $\mathbf{C^A_{ij}} \in R^{n \times n}$ be the contact count between loci $i$ and $j$.
- Let $d_{ij} = \|x_i - x_j\|_2$

## Optimization problem

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

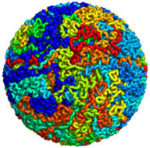
# Relationships between contact counts $c$, genomic distances $s$ and Euclidean distances $d$

**Fractal globule**



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

**Equilibrium**



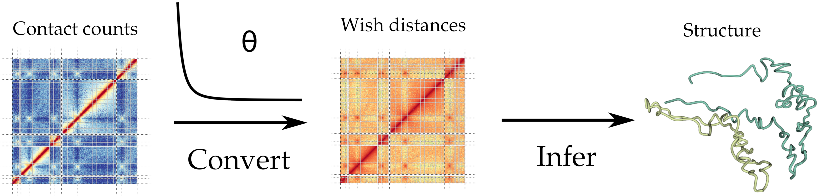
- $c \sim s^{-3/2}$
- $d \sim s^{1/2}$ for $s < s_{\max}^{2/3}$

## Relationship between contact counts and Euclidean distances

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

[Mirny, 2011]

# Metric MDS-based methods



## Formulation

$$\underset{\mathbf{x}_1,\ldots,\mathbf{x}_n}{\text{minimize}} \qquad \sigma(\mathbf{X}, C) = \sum_{i,j|c_{ij}\neq 0} w_{ij}(\|x_i - x_j\|_2 - \Theta(c_{ij}))^2$$

- **X** : 3D coordinates
- **C** : normalized contact counts.
- $w_{ij}$ are weights (set to $\frac{1}{\Theta(c_{ij}^N)^2}$ in *pastis*-**MDS2**)

- $\Theta(c) = \beta c^{\alpha}$: count-to-distance function

# Statistical approaches for inferring the 3D structure of the genome

- MDS-based methods minimize an arbitrary stress function that measures the discrepancy between wish distances and 3D distances of the model.

## Statistical approach for stable inference of genome structure

- replace the arbitrary MDS loss function with a better-motivated likelihood function
- define a probabilistic model of contact counts parametrized by the 3D model.

# Inferring 3D models of genome architecture

**The idea** Let's assume that $c \sim NegativeBinomial(\beta d^{\alpha}, r)$, where $c$ is the interaction count, $d$ the pairwise euclidean distance, $r$ the dispersion parameter, $\alpha$ unknown parameters, and $\beta$ a scale coefficient.

**Likelihood**

$$\ell(\mathbf{X}, C) = \prod_{i,j} \frac{\Gamma(c_{ij} + r)}{\Gamma(c_{ij} + 1)\Gamma(r)} (\frac{\beta d_{ij}^{\alpha}}{r + \beta d_{ij}^{\alpha}})^{c_{ij}} (1 - \frac{\beta d_{ij}^{\alpha}}{r + d_{ij}^{\alpha}})^{r} \tag{1}$$

**The optimization problem**

$$\max_{\alpha,\beta,\mathbf{X}} \quad \mathcal{L}(\mathbf{X}, \alpha, \beta) = \sum_{i<j\leq n} c_{ij}\alpha \log d_{ij} - (c_{ij} + r) \log(r + \beta d_{ij}^{\alpha}) \tag{2}$$

# Estimating the dispersion *r*

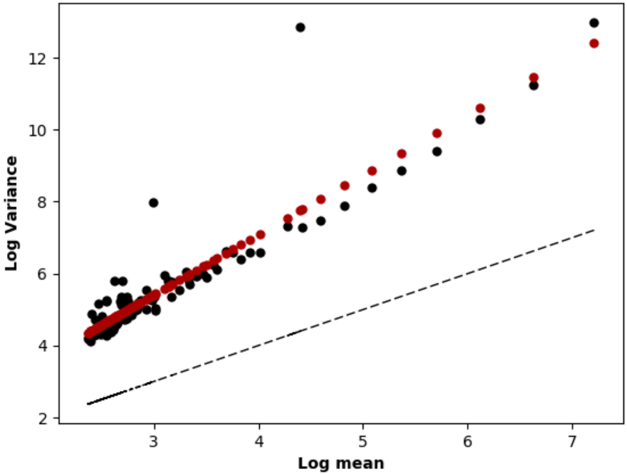
### Assumptions

- Contact counts for pairs of loci are of the same order of magnitude.
- The variance is a smooth function of the mean.

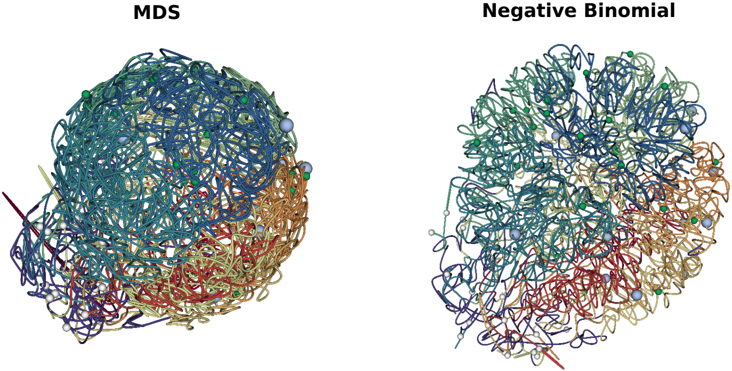
### Estimating the dispersion *r*

- For each genomic distance *l*, compute the empirical mean and variance on normalized data:
  - $\hat{q}_l = \frac{1}{|I(l)|} \sum_{(i,j) \in I(l)} c_{ij}$
  - $\hat{v}_l = \frac{1}{|I(l) - 1|} \sum_{(i,j \in I(l))} (c_{ij} - \hat{q}_l)$
- Fit a polynomial function between $\hat{q}$ and $\hat{v}$
- Or estimate a constant dispersion paramater.

# Dispersion fit on *S. cerevisiae*



# MDS versus the Negative Binomial modeling: the case of Sporozoites *P. falciparum*



MDS

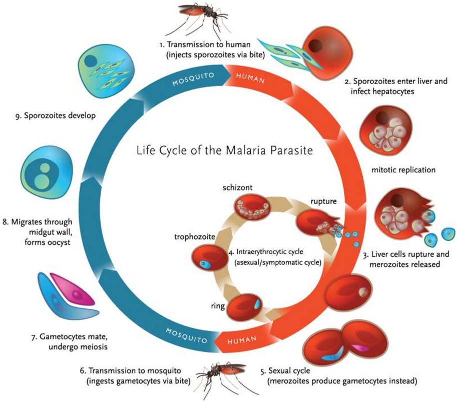
Negative Binomial

**Sporozoite stage**

# Three-dimensional modeling of the *P. falciparum* genome during reveals a strong connection between genome architecture and gene expression.

*joint work with Evelien Bunnik, Kate Cook, Ferhat Ay,*
*Sebastiaan Bol, Jacques Prudhomme, Jean-Philippe Vert,*
*William S. Noble and Karine Le Roch.*

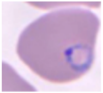
# 5 timepoints in the life cycle of *P. falciparum*


Life Cycle of the Malaria Parasite
1. Transmission to human
(injects sporozoites via bite)
2. Sporozoites enter liver and
infect hepatocytes
mitotic replication
3. Liver cells rupture and
merozoites released
4. Intraerythrocytic cycle
(asexual/symptomatic cycle)
schizont
rupture
trophozoite
ring
5. Sexual cycle
(merozoites produce gametocytes instead)
6. Transmission to mosquito
(ingests gametocytes via bite)
7. Gametocytes mate,
undergo meiosis
8. Migrates through
midgut wall,
forms oocyst
9. Sporozoites develop
MOSQUITO
HUMAN
MOSQUITO
HUMAN

## Experiments

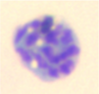
Parasite phenotype

Ring

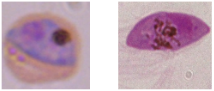


Trophozoite

Schizont



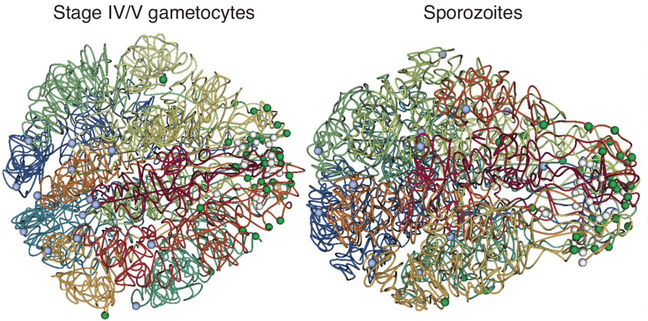
Early and late gametocytes



Sporozoite

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

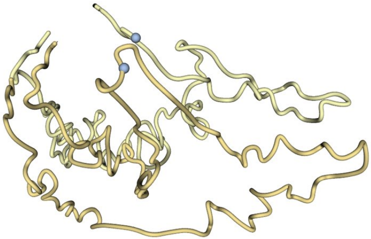
## We applied our method to the data sets thus obtaining 5 models

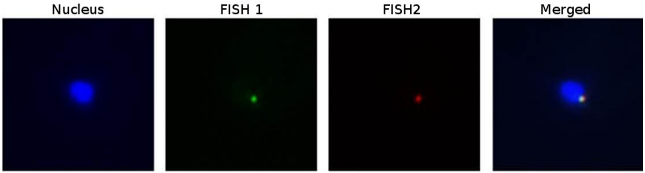


Stage IV/V gametocytes

Sporozoites

# Colocalization of loci is validated with FISH

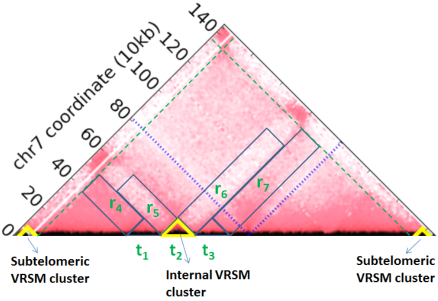




**Var genes on chromosomes VII and VIII colocolize**

# Biological insights on the 3D architecture of the genome

- Virulence gene clusters on different chromosomes colocalize in 3D.
- Highly transcribed rDNA units colocalize in 3D during the ring stage.
- Transcriptionally active trophozoite stage exhibits an open chromatin structure.
- VRSM gene clusters form domain-like structures.



# Identifying links between gene expression profiles and 3D structure

**Motivation:** Extract a gene expression profile $v \in \mathbb{R}^p$ that is:

- representative of the gene expression profiles ;
- correlated with the 3D structure;

**Data:** For each gene $g \in \mathcal{G}$

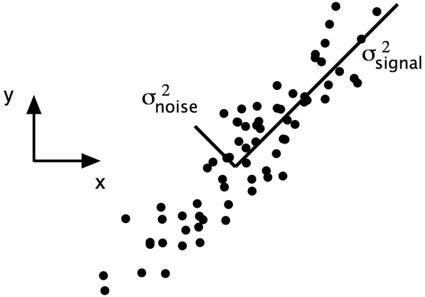
- Log expression profiles at 27 datapoints: $e(g) = (e_1(g), \dots, e_p(g)) \in \mathbb{R}^p$ .
- Gene's 3D coordinates, extracted from the inferred 3D structure: $x(g)$.

**Method**: KernelCCA [Vert and Kanehisa, 2003, Bach and Jordan, 2002]

# Extracting a vector *v* representative of the gene expression profiles

Find $v \in \mathbb{R}^p$ to maximize:

$$V(v) = \frac{\sum_{g \in \mathcal{G}} \left(v^T e(g)\right)^2}{\|v\|^2}$$



# Find $f$ such that $f$ is smooth with respect to the 3D structure

Let $f$ be a vector of scores assigned to each genes.

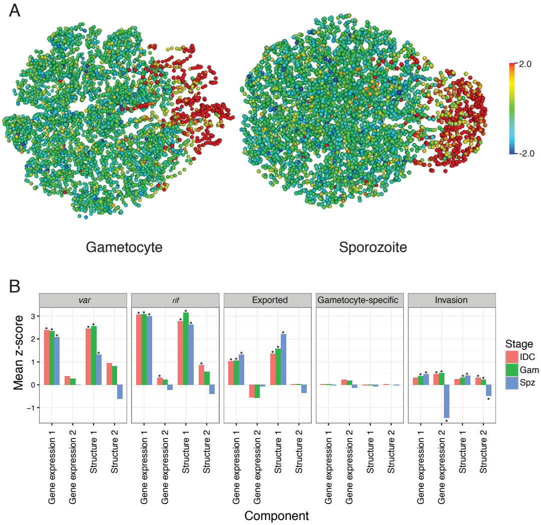
$$S(f) = \frac{f^\top K_{3D}^{-1} f}{||f||^2}$$

We want:

- $V(v)$ be large,
- $S(f)$ be small,
- $(v^\top e(g))_{g \in \mathcal{G}}$ and $f$ be as correlated as possible

**This can be cast as a generalized eigenvalue problem**

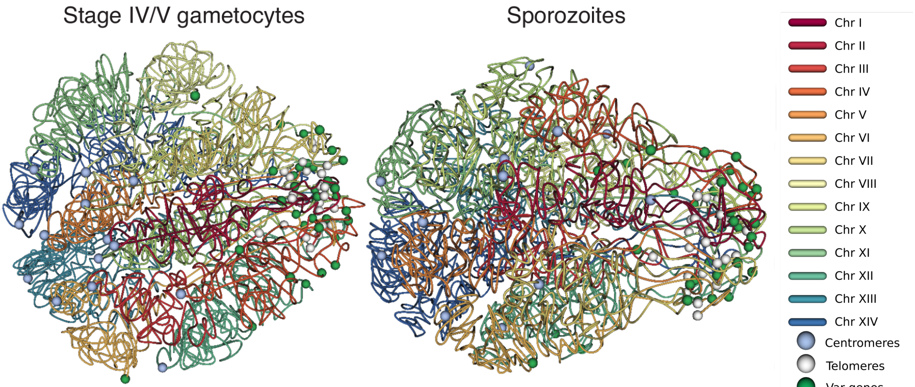
# KernelCCA reveals a strong correlation between gene expression profiles and 3D structure



# Conclusion

- We built high-resolution models of *P. falciparum*'s genome architecture at three time points.
- We observed :
  - strong clustering of centromeres, telomeres, virulance genes and rDNA, resulting in a **complex architecture**.
  - strong correlation between 3D genome architecture and gene expression.
- **Disruption of the parasite's genome organization** is likely to interfere with its life cycle, and could therefore be **lethal**.

# 3D models



Stage IV/V gametocytes

Sporozoites

# References I

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3: 1–48, 2002.

L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1): 37–51, 2011.

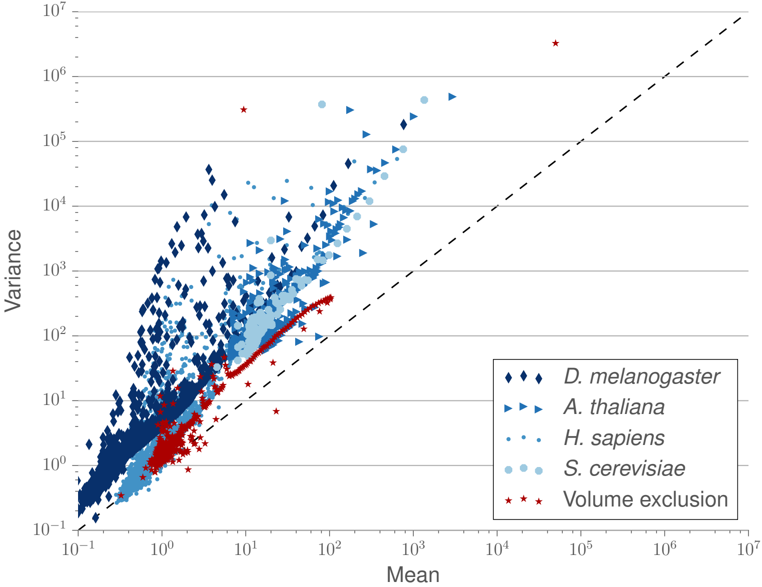
S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin v looping. *Cell*, 59(7):1665–1680, 2014.

J.-P. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In *Advances in Neural Information Processing Systems 15*, pages 1425–1432, Cambridge, MA, 2003. MIT Press.

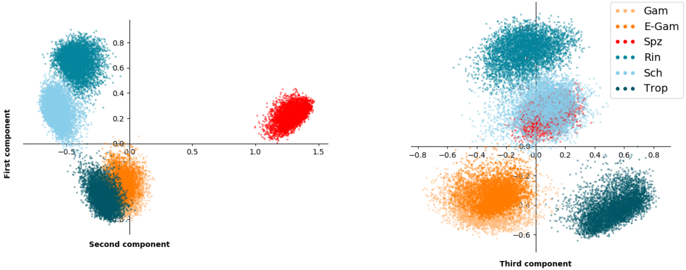
C. Zimmer and E. Fabre. Principles of chromosomal organization: lessons from yeast. *Journal of Cell Biology*, 192 (5):723–733, 2011.

# Contact counts are overdispersed

# Contact counts are overdispersed II



# Variation is greater between timepoints than between initial points


First component
Second component
Gam
E-Gam
Spz
Rin
Sch
Trop
Third component