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?

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

Tools for capturing chromosome conformation

Figure from:

<table>
<thead>
<tr>
<th>Method</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C</td>
<td>Dekker et al. Science 2002</td>
</tr>
<tr>
<td>ChIA-PET</td>
<td>Fullwood et al. Nature 2009</td>
</tr>
<tr>
<td>Hi-C</td>
<td>L.-Aiden et al. Science 2009</td>
</tr>
<tr>
<td></td>
<td>Duan et al. Nature 2010</td>
</tr>
</tbody>
</table>
Output of conformation capture is a contact matrix

\[ C(i,j) = \text{How many times locus } i \text{ is linked to locus } j \text{ by a paired-end read?} \]
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 $X \in R^{n \times 3}$ be the coordinates of each bead.
- Let $C \in R^{n \times n}$ be the contact count matrix and $D$ the set of non-zeros entries.
- Let $\Theta$ the count-to-distance transfer function.

**Optimization problem**

| minimize $\sigma(X, C)$ | \[ \begin{align*}
\sigma(X, C) &= \sum_{i,j} (\|x_i - x_j\|_2^2 - \Theta(c_{ij}))^2
\end{align*} \]
| subject to some bio | \[ \begin{align*}
\sum_{i,j} \frac{(\|x_i - x_j\|_2^2 - \Theta(c_{ij}))^2}{\Theta(c_{ij})^2} - \lambda \sum_{i,j \notin D} \|x_i - x_j\|_2^2
\end{align*} \]
| $x_i^T x_i \leq r_i$ |

- **MDS1** [Duan et al., 2010]

- **MDS2** [Ay et al., 2014]

- **ChromSDE** [Zhang et al., 2013]
Nonmetric MDS-based method

Idea

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

\[
c_{ij} \geq c_{k\ell} \iff \|x_i - x_j\|_2 \leq \|x_k - x_\ell\|_2
\]

(4)

\[
\begin{align*}
\text{minimize} & \quad \sigma(X, C, \Theta) \\
\text{subject to} & \quad \Theta \text{ decreasing} \\
& \quad \text{some biologically motivated constraints} \\
& \quad x_i^T x_i \leq r_{\text{max}}^2, \quad \text{(all beads should lie in the nucleus)}
\end{align*}
\]
Poisson model

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

Likelihood

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

Optimization problem

$$\min_{x_1,\ldots,x_n,\alpha,\beta} \sigma(X, C, \alpha, \beta) = -\log(\ell(X, C, \alpha, \beta))$$

subject to some biologically motivated constraints

$$x_i^T x_i \leq r_{\text{max}}^2, \quad \text{(all beads should lie in the nucleus)}$$
Default counts-to-distance transfer function

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

Fractal globule

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

Equilibrium

Default counts-to-distance transfer function

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

- **Generated Datasets**
  
  \[ c_{ij} = P(\beta d_{ij}^\alpha), \]  
  
  where
  
  - \( \alpha = -3 \) and \( \beta \) varies between 0.01 and 0.7.
  - \( \alpha \) varies between -4 and -2 and \( \beta \) 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 $\alpha$ parameter

- MDS1
- MDS2
- NMDS
- PM1
- PM2

RMSD vs $\alpha$
Mouse embryonic stem cells

- Stability across enzyme replicates

<table>
<thead>
<tr>
<th>Resolution</th>
<th>1 Mb</th>
<th>500 kb</th>
<th>200 kb</th>
<th>100 kb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSD</td>
<td>Corr</td>
<td>RMSD</td>
<td>Corr</td>
</tr>
<tr>
<td>MDS1</td>
<td>13.13</td>
<td>0.945</td>
<td>10.00</td>
<td>0.942</td>
</tr>
<tr>
<td>MDS2</td>
<td>5.54</td>
<td>0.964</td>
<td>5.68</td>
<td>0.959</td>
</tr>
<tr>
<td>NMDS</td>
<td>5.80</td>
<td>0.965</td>
<td>5.67</td>
<td>0.959</td>
</tr>
<tr>
<td>PM1</td>
<td>7.28</td>
<td>0.931</td>
<td>7.14</td>
<td>0.913</td>
</tr>
<tr>
<td>PM2</td>
<td>4.92</td>
<td>0.976</td>
<td>4.66</td>
<td>0.968</td>
</tr>
</tbody>
</table>

- Stability across resolution

<table>
<thead>
<tr>
<th></th>
<th>MDS1</th>
<th>MDS2</th>
<th>NMDS</th>
<th>PM1</th>
<th>PM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSD</td>
<td>14.86</td>
<td>12.92</td>
<td>12.98</td>
<td>13.03</td>
<td><strong>11.48</strong></td>
</tr>
<tr>
<td>Correlation</td>
<td>0.781</td>
<td>0.754</td>
<td>0.738</td>
<td>0.737</td>
<td><strong>0.807</strong></td>
</tr>
</tbody>
</table>
Conclusion

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

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


• Extension to aneuploidy?
Part 2

The spatial organization of the *P. falciparum* genome
Modeling the dynamic genome architecture of a 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.

**Ring**

**Trophozoite**

**Schizont**

0 hrs  
18 hrs  
36 hrs

Tethered Hi-C

![Chromosome Heatmap](image-url)
*Plasmodium* contact frequencies suggest a fractal globule architecture.
Scaling parameter for the Trophozoite stage is indicative of more intermingled chromatin.

\[ y \sim x^{\alpha - 1.14} \]

\[ \alpha = -1 \]

\[ \alpha = -1.5 \]

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

- Model the genome as beads at 10 kbp resolution.
- Estimate Euclidean distance matrix using a ruler derived from intra-chromosomal interactions.
- Find 3D coordinates that yield the expected distances:

\[
\text{minimize} \quad \sum_{\delta_{ij} \in \mathcal{D}} \frac{1}{d_{ij}} (d_{ij} - \delta_{ij})^2 \quad \text{for} \quad X \in \mathbb{R}^{3 \times n}
\]

- Include constraints reflecting physical and biological prior knowledge.
  1. All loci must lie within a spherical nucleus centered on the origin.
  2. Two adjacent loci must not 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
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.

DNA FISH confirms selected contacts

Inter-chromosomal pair of virulence genes

<table>
<thead>
<tr>
<th>Stage</th>
<th>DAPI</th>
<th>Chr. 7 (555000)</th>
<th>Chr. 8 (45000)</th>
<th>Merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trophozoite</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage</th>
<th>DAPI</th>
<th>Chr. 7 (815000)</th>
<th>Chr. 11 (825000)</th>
<th>Merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ring</td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
</tbody>
</table>
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
Genes that are close together exhibit correlated expression profiles.

- Only inter-chromosomal gene pairs.
- Correlation between expression vectors:
Telomeres have a repressive effect on gene expression
Gene expression variation exhibits a gradient across the structure.

Telomeric
- Antigenic variation
- Sexual stage genes

Non-telomeric
- Translation
- Trophozoite genes

Kernel Canonical Correlation Analysis (kCCA)
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.

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

NSF/CRA CI Fellows award
- CIF 1136996 to FA

NIH/NIAID
- R01 AI85077 to KLR
- R01 AI106775 to WN/KLR

Human Frontier Science Program
- LT00507/2011 to EB

The European Research Council
- SMAC-ERC-280032 to JPV

The European Commission
- HEALTH-F5-2012-305626 to JPV/NV

The French National Research Agency
- ANR-11-BINF-0001 to JPV/NV
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

¹ [http://www.cdc.gov/malaria/about/facts.html](http://www.cdc.gov/malaria/about/facts.html)
² World Malaria Report 2008, WHO.
Malaria transmission cycle

Erythrocytic cycle

Infected liver

First infected person

First infected mosquito

Erythrocytic cycle:
- Ring
- Trophozoite
- Schizont
- Merozoite
- 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.

*Plasmodium* chromatin architecture goes through systematic changes

Ponts et al.  
*Genome Research* 2010.
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.

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.
Our data exhibits characteristics biases of Hi-C

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

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

Modeling genome architecture using Hi-C contact maps
Highly transcribed rRNA genes colocalize

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

Volume exclusion modeling does not capture *Plasmodium* architecture

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.