Reconstructing the 3D architecture of the genome

Jean-Philippe Vert
Collaborators

University of Washington
William Noble
Ferhat Ay

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

MINES ParisTech, France
Jean-Philippe Vert
Nelle Varoquaux

Josh Burton
Ivan Liachko
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


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

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

1. How to reconstruct the 3D structure of the genome from Hi-C data?
2. How genome architecture is related to gene expression regulation: the case of *P. falciparum*
3. Cheap identification of centromeres from metagenomic Hi-C data
Part 1

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

**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 \( s < s_{\text{max}}^{2/3} \)
- 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 $\mathbf{X} \in \mathbb{R}^{n \times 3}$ be the coordinates of each bead.
- Let $\mathbf{C} \in \mathbb{R}^{n \times n}$ be the contact count matrix and $\mathcal{D}$ the set of non-zeros entries.
- Let $\Theta$ the count-to-distance transfer function.

**Optimization problem**

<table>
<thead>
<tr>
<th>minimize $\sigma(\mathbf{X}, \mathbf{C})$</th>
<th>$\sigma(\mathbf{X}, \mathbf{C}) = \sum_{i,j} (|x_i - x_j|<em>2 \Theta(c</em>{ij}))^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>subject to $\mathbf{x}_i^T \mathbf{x}_i \leq r_i$</td>
<td>$\sigma(\mathbf{X}, \mathbf{C}) = \sum_{i,j} (|x_i - x_j|<em>2 - \Theta(c</em>{ij}))^2$</td>
</tr>
<tr>
<td>$\mathbf{x}_1, \ldots, \mathbf{x}_n$</td>
<td>$\sigma(\mathbf{X}, \mathbf{C}) = \sum_{i,j} \frac{(|x_i - x_j|<em>2 - \Theta(c</em>{ij}))^2}{\Theta(c_{ij})^2}$</td>
</tr>
<tr>
<td>some bio</td>
<td>$\sigma(\mathbf{X}, \mathbf{C}) = \sum_{i,j} \left(\frac{|x_i - x_j|<em>2}{\Theta(c</em>{ij})^2} - \lambda \sum_{i,j \notin \mathcal{D}} |x_i - x_j|_2^2\right)$</td>
</tr>
</tbody>
</table>

- **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)

minimize \( \sigma(X, C, \Theta) \)

subject to \( \Theta \) decreasing

some biologically motivated constraints

\[
x_i^T x_i \leq r^2_{\text{max}}, \quad \text{(all beads should lie in the nucleus)}
\]

(Ben-Eleazar et al, 2013)
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
\[
\ell(X, \alpha, \beta) = \prod_{i<j\leq n} \frac{(\beta d_{ij}^{\alpha})^{c_{ij}}}{c_{ij}!} \exp(-\beta d_{ij}^{\alpha})
\] (5)

Optimization problem
\[
\text{minimize} \quad \sigma(X, C, \alpha, \beta) = -\log(\ell(X, C, \alpha, \beta))
\]
\[
\text{subject to} \quad \text{some biologically motivated constraints}
\]
\[
x_i^T x_i \leq r_{\text{max}}^2, \quad (\text{all beads should lie in the nucleus})
\]
Data

- **Generated Datasets**

\[ c_{ij} = P(\beta d_{ij}^\alpha), \quad (7) \]

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

Performance as a function of SNR (RMSD)

Average signal to noise ratio per parameter

RMSD per experiment

- MDS1
- MDS2
- NMDS
- PM1
- PM2
- ChromSDE
Robustness to parameter misspecification

Average RMSD with varying $\alpha$ parameter

- MDS1
- MDS2
- NMDS
- PM1
- PM2

RMSD

$\alpha$

-4.5 -4.0 -3.5 -3.0 -2.5 -2.0 -1.5
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><strong>0.738</strong></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?
We assayed genome architecture at 3 time points in the erythrocytic cycle.
*Plasmodium* contact frequencies suggest a fractal globule architecture

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

- \( \alpha = -0.98 \)
- \( \alpha = -1.14 \)
- \( \alpha = -0.96 \)
- \( \alpha = -0.14 \)
Scaling parameter for the Trophozoite stage is indicative of more intermingled chromatin

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

\[ \alpha = -1 \]

\[ \alpha = -1.5 \]

Lieberman-Aiden et al. Science 2009
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:

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

\[\mathcal{D} = \{\delta_{ij} | \delta_{ij} \neq 0\}\]

- 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 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.jpg" alt="Image" /></td>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="image3.jpg" alt="Image" /></td>
<td><img src="image4.jpg" 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.jpg" alt="Image" /></td>
<td><img src="image6.jpg" alt="Image" /></td>
<td><img src="image7.jpg" alt="Image" /></td>
<td><img src="image8.jpg" 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
Simulated and observed chromatin contacts are highly concordant in yeast...

... but volume exclusion modeling does not capture *Plasmodium* architecture

Chromosome 7, trophozoite stage
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.

Part 3

Centromere calling from metagenonomic Hi-C data
Centromeres (CEN)

- part of a chromosome that links sister chromatid
Where are yeast CENs?

- Important for bioengineering (billion dollar industry)
- Still unknown for major yeast species
- Point centromeres: hard to find using traditional methods
  - CEN consists of 2-3 binding site motifs
  - Virtually no heterochromatin
CENs are strongly clustered in 3D for yeasts and some other species


Barzel and Kupiec, Nat Rev Genet, 2008


Idea: use Hi-C data to call CENs
Data: Hi-C on metagenomic samples
Two datasets

Mixture of Yeasts (M-Y)
16 yeast strains

Mixture of 3 Domains (M-3D)
8 yeasts, 9 bacteria, and 1 archaeon

Outline of centromere identification for multiple yeasts from Hi-C contact maps

1. Align paired-end reads to a concatenated reference genome.
2. Post-process mapping results and create a contact map for each organism.
3. Pool data from all replicates (e.g., different restriction enzymes) for each organism.
4. Normalize contact maps for experimental/technical biases (10, 20, 40 kb).
5. Make “initial guesses” for each centromere using marginalized inter-chromosomal (trans) contact counts.
6. Starting from the initial guesses find a centromere position for each chromosome that best explains the observed inter-chromosomal contact count pattern.
Initial guess of CEN localization

**Ideal**

![Ideal Graph]

**Observed**

![Observed Graph]
Optimization from initial guess

Data

$$\minimize_{\mathbf{P}} \sum_{(i,j) \in \mathcal{D}} (c_{i,j} - a \sum_{k,l} 1_{[i \in \mathcal{D}_k]} 1_{[j \in \mathcal{D}_l]} e^{\frac{(i-p_k)^2 - (j-p_l)^2}{2\sigma^2}})^2$$

subject to \( p_l \in \mathcal{D}_l \) \( \forall l \in [1, \ldots, K] \)

Gaussian centered in pairs of centromeres
Results
Accuracy on S. cerevisiae

10 out of 16 predictions are within ~2 kb of the known mid-point of the centromere

## Average prediction error per genome

<table>
<thead>
<tr>
<th>Organism</th>
<th>Our method</th>
<th>Marie-Nelly et al. (Bioinformatics, 2014)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>L. kluyveri</em> (M-Y)</td>
<td>&lt; 5 kb</td>
<td>~ 312 kb</td>
</tr>
<tr>
<td><em>S. pombe</em> (M-PE)</td>
<td>&lt; 11 kb</td>
<td>~ 574 kb</td>
</tr>
<tr>
<td><em>S. mikatae</em> (M-Y)</td>
<td>&lt; 6 kb</td>
<td>~ 124 kb</td>
</tr>
<tr>
<td><em>S. bayanus</em> (M-Y)</td>
<td>&lt; 6 kb</td>
<td>~ 113 kb</td>
</tr>
<tr>
<td><em>K. lactis</em> (M-Y)</td>
<td>&lt; 9 kb</td>
<td>~ 19 kb</td>
</tr>
</tbody>
</table>

* only validated on the partial ground truth available

~5-10M paired-end reads enough per species (that’s $50-$100)!
Conclusion: The many uses of Hi-C

- **Lieberman-Aiden, et. al.** Science, 2009
- **Duan, et al.** Nature, 2010 (S. cerevisae)
- **Ay, et al.** Genome Res., 2014a (P. falciparum)
- **Burton, Liachko, et al.** G3, 2014

**3D model of genome**

- **Enhancer calling**
- **Promoter calling**
- **Long-range chromatin contacts**

**Organismal Deconvolution**

**Genome scaffolding**
Acknowledgements

University of Washington
William Noble
Ferhat Ay

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

MINES ParisTech, France
Jean-Philippe Vert
Nelle Varoquaux

Josh Burton
Ivan Liachko

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 death was caused by *P. falciparum*. ²

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

First infected person

Infected liver

First infected mosquito

© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.
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.

\[
\text{ICP} = \frac{\text{inter}}{\text{intra}(>1\text{kb})}
\]

<table>
<thead>
<tr>
<th></th>
<th>Ring</th>
<th>Trophozoite</th>
<th>Schizont</th>
</tr>
</thead>
<tbody>
<tr>
<td>Troph - control</td>
<td>3.0%</td>
<td>7.6%</td>
<td>22.0%</td>
</tr>
<tr>
<td>Random</td>
<td>9.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Graph:**
- Logarithmic scale showing contact probability decay with genomic distance.
- Equal occupancy bins represented.
- Trend line for log-linear decay: \( y \sim x^{-0.98} \).
Our data exhibits characteristics biases of Hi-C

Larger number of restriction enzyme (RE) cut sites \( \rightarrow \) 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?
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.