Chromosome painting

Verónica Miró Pina

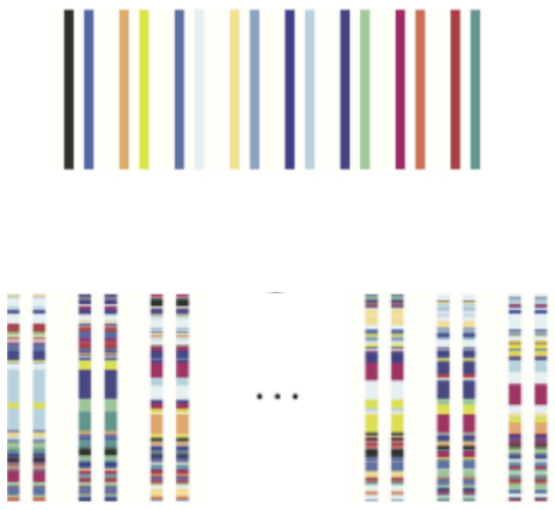
joint work with Emmanuel Schertzer & Amaury Lambert

Chromosome painting: Experimental populations of Caenorhabditis elegans (Teotonio et al ('12))

- Start with 180 individuals sampled from distinct sub-populations.

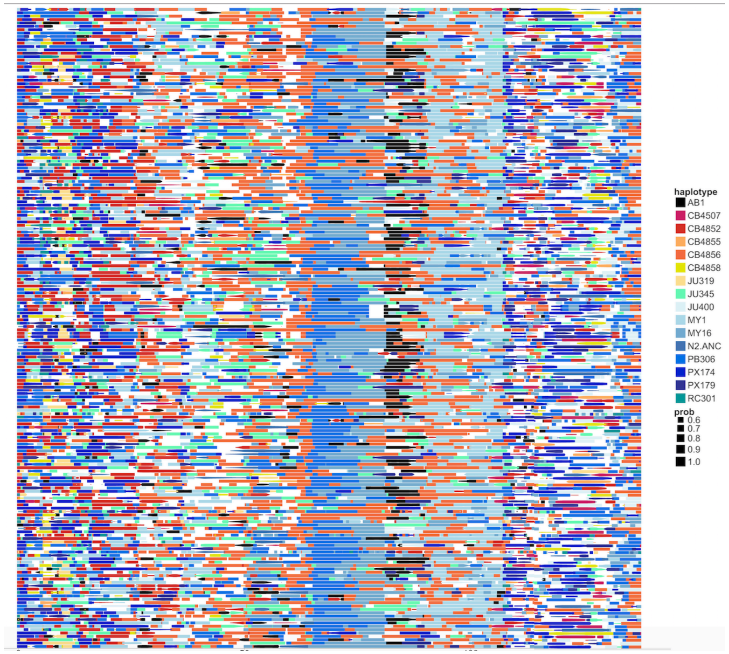
Chromosome painting: Experimental populations of Caenorhabditis elegans (Teotonio et al ('12))

- Start with 180 individuals sampled from distinct sub-populations.
- Let it evolve during during 140 generations at controlled population size.



Chromosome painting: Experimental populations of Caenorhabditis elegans (Teotonio et al ('12))

- Start with 180 individuals sampled from distinct sub-populations.
- Let it evolve during during 140 generations at controlled population size.
- Genotype these 180 sequences.



Chromosome painting

Segment = maximal connected set of of points sharing the same color.

Cluster = maximal set of points sharing the same color.

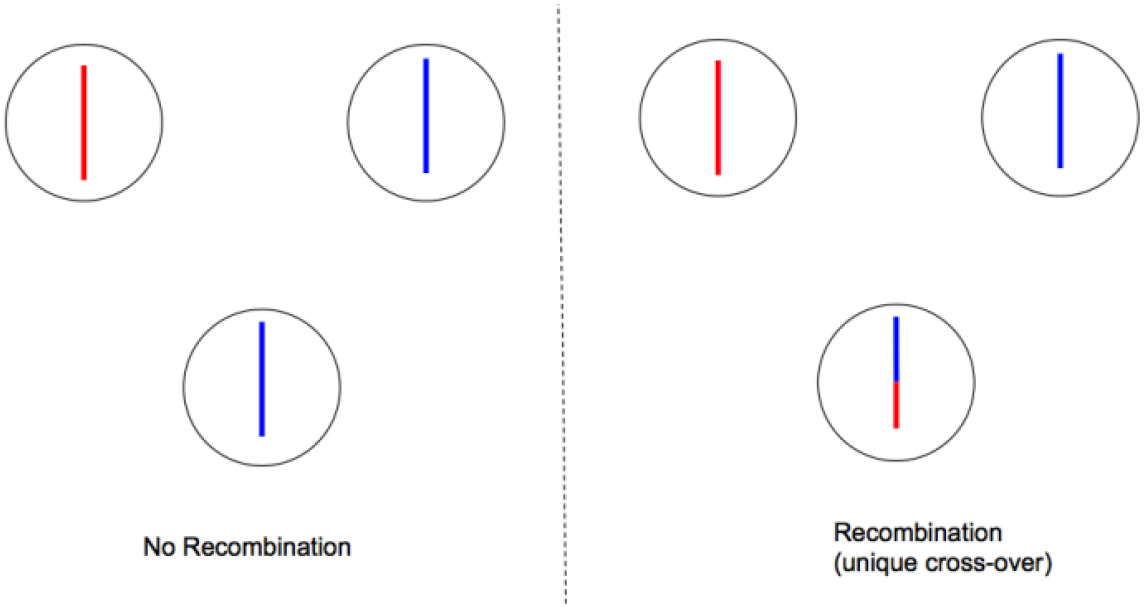
- What is the size of a typical segment ?
- What is the length, diameter of a typical cluster ?
- How many segments, clusters on a given interval ?

An haploid W-F model with recombination

- Population of constant size N.
- Each individual carries 1 chromosome of size R.
- Wright-Fisher dynamics: at each time step each individual choses two parents from the previous generation. With probability:

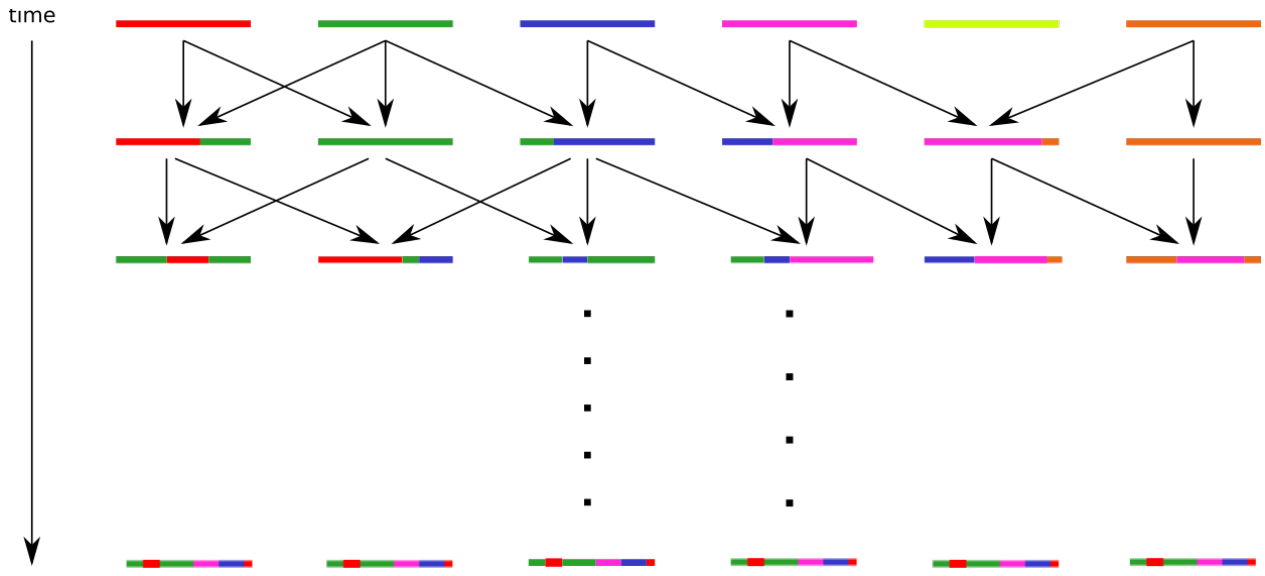
$1 - \rho$ Copies one parent chromosome.

ρ Recombination event: a cross-over occurs.



An haploid W-F model with recombination

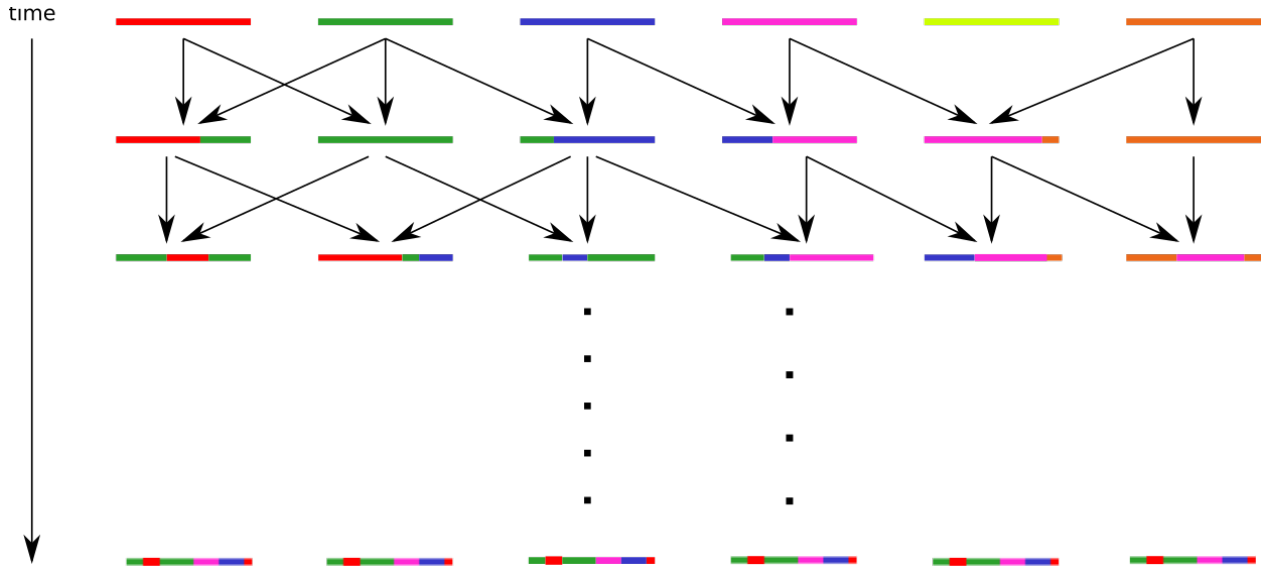
- At time 0 each chromosome is painted in a distinct color.



- After *k* steps, each chromosome is a mosaic of colors.

An haploid W-F model with recombination

- At time 0 each chromosome is painted in a distinct color.



- After *k* steps, each chromosome is a mosaic of colors.

- (N, R)-Partitioning process Π_N^R: color partition of the system at equilibrium (for a population of size *N* with chromosomes of size *R*.)

Large Population, Long Chromosome

- Let Π_N^R be the random (finite) partition of $[0, R]$ corresponding to fixation.
- Let $N \to \infty$ and let the probability of recombination $\rho_{N,R}$ depends on N and R in such a way that

$$\lim_{N\to\infty} N\,\rho_{N,R} = R.$$

Proposition

For every $R > 0$, there exists a random finite partition Π^R of $[0, R]$ such that

$$\Pi_N^R \to \Pi^R \text{ in law.}$$

Question: What can we say about Π^R on an interval of large size? (For humans $R \approx 5 \times 10^4$)

Cluster covering the origin

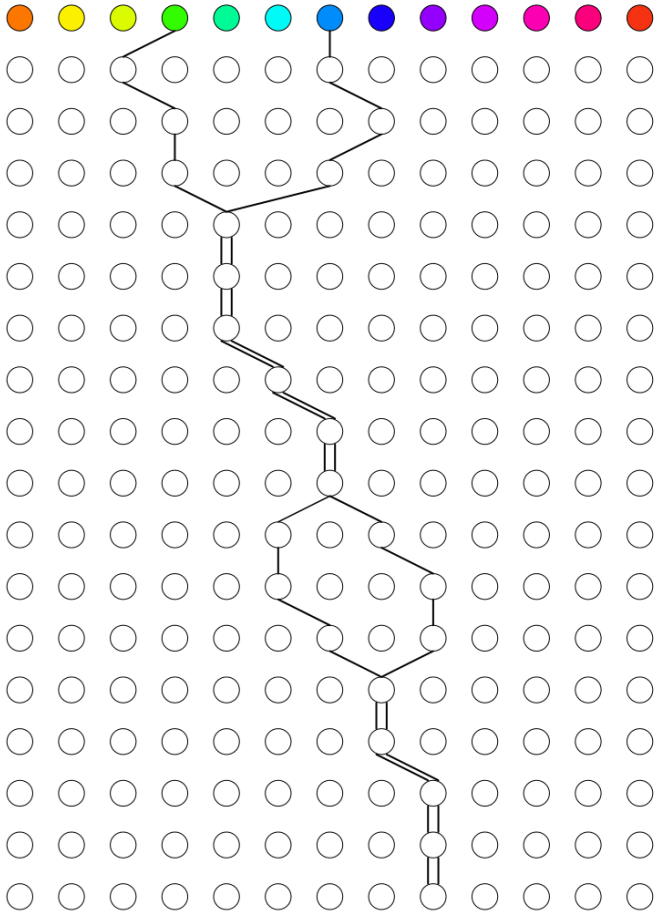
Theorem (Lambert, M. P., Schertzer)

Define $\mathcal{L}_R$ to be the length of the cluster covering 0 on the interval $[0, R]$. Then

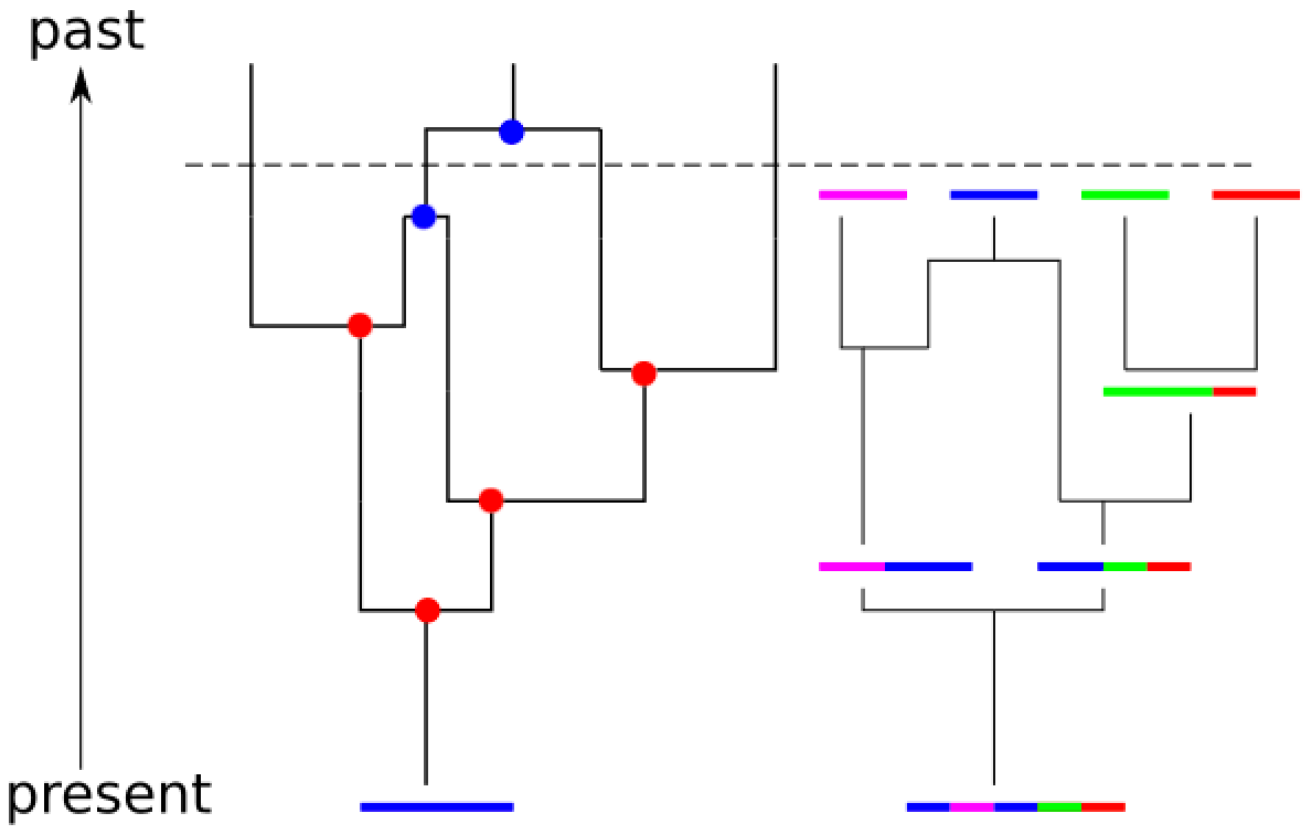
$$\lim_{R\to\infty} \frac{1}{\log(R)} \mathcal{L}_R = \mathcal{E}(1) \text{ in law.}$$

The Ancestral Recombination Graph (ARG): two sites

- 2 sites x and y at distance l: follow their ascendances as time goes backward.
- At each generation, the common line of ascent $\{x, y\}$ splits with probability l/N.
- At each generation, the singleton lines $\{x\}$ and $\{y\}$ coalesce with probability $1/N$.
- x, y carry the same color iff their lines coincide at $-\infty$



Ancestral Recombination Graph (Griffiths, Hudson)

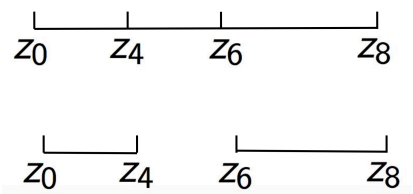


Duality: The color partition has the same law as the stationary partition of the ARG.

Ancestral Recombination Graph (Griffiths, Hudson)

- Let $z_0 < \cdots < z_n$ in $\mathbb{R}$.
- The ancestral recombination graph is the continuous time Markov process on $\mathcal{P}_n$ — the set of partitions of $\{0, \cdots, n\}$ — with following rates:
  - $\rightarrow$ Coalescence: groups of lineages coalesce at rate 1.
  - $\rightarrow$ Fragmentation: group of lineages $\{\sigma(0) < \cdots < \sigma(j) < \sigma(j+1) < \cdots < \sigma(K)\}$ splits into two parts :

    $\{\sigma(0) < \cdots < \sigma(j)\}$ and $\{\sigma(j+1) < \cdots < \sigma(K)\}$ at rate $z_{\sigma(j+1)} - z_{\sigma(j)}$.



Duality:

$$\mathbb{P}(z_0 \sim \cdots \sim z_n) = \mu^{\mathbf{z}}(\{0, \cdots, n\})$$

where $\mu^{\mathbf{z}}$ is the invariant distribution of the ancestral recombination graph corresponding to $\mathbf{z} = (z_0, z_1, \cdots, z_n)$.

Proof for the Cluster Size at the Origin

- We aim at proving that

$$\lim_{R\to\infty} \frac{1}{\log(R)} \mathcal{L}_R = \mathcal{E}(1) \text{ in law.}$$

  where $\mathcal{L}_R$ is the length of the cluster at 0 on $[0, R]$.

- Main Idea: Method of moments.

- Using Carleman's condition, it is enough to show that

$$\lim_{R\to\infty} \frac{1}{\log(R)^n} \mathbb{E}\left(\mathcal{L}_R^n\right) = n!$$

Proof for the Cluster Size at the Origin

$$
\begin{aligned}
\frac{1}{\log(R)^n}\mathbb{E}\left(\mathcal{L}_R^n\right) &= \frac{1}{\log(R)^n}\mathbb{E}\left(\left(\int_0^R 1_{0\sim z}dz\right)^n\right) \\
&= \frac{1}{\log(R)^n}\mathbb{E}\left(\int_{[0,R]^n} 1_{0\sim z_1\cdots\sim z_n}dV\right) \\
&= \frac{1}{\log(R)^n}\int_{[0,R]^n}\mathbb{P}(0\sim z_1\cdots\sim z_n)dV \\
&= \frac{R^n}{\log(R)^n}\times\frac{1}{R^n}\int_{[0,R]^n}\mu^{\mathbf{z}}(\{0,\cdots,n\})dV
\end{aligned}
$$

where $\mu^{\mathbf{z}}$ is the invariant distribution in the ancestral recombination graph corresponding to $\mathbf{z}=(z_0=0,z_1,\cdots,z_n)$.

Perspectives

- Results about the number of clusters (in progress).
- Describe the geometry of the cluster at origin.
- Work on a neutrality test based on haplotypes (without mutation): in collaboration with Mathieu Tiret and Frédéric Hospital (INRA)
- Try to apply our results to analyse real data: with Henrique Teotonio.