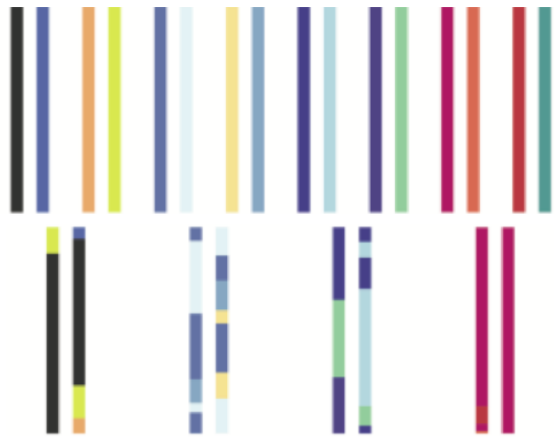
Chromosome Painting.

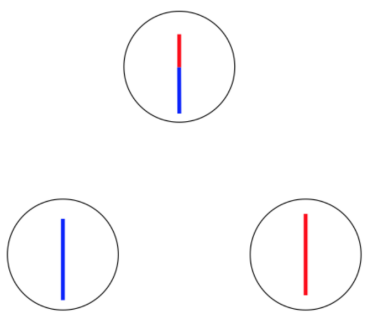
Emmanuel Scherzer (LPMA Paris 6 – SMILE Collège de France). Joint on-going work with A. Lambert and V. Miro Pina.

February 16, 2017

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

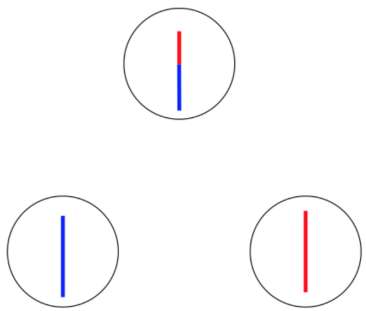
- Start with 16 individuals.
- Build a population of size $\sim 10^4$ by random intercross
- Let it evolve during during 140 generations at controlled population size.
- Genotype 180 sequences.

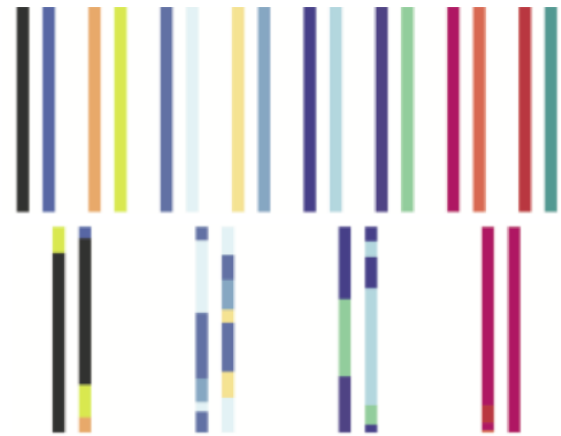




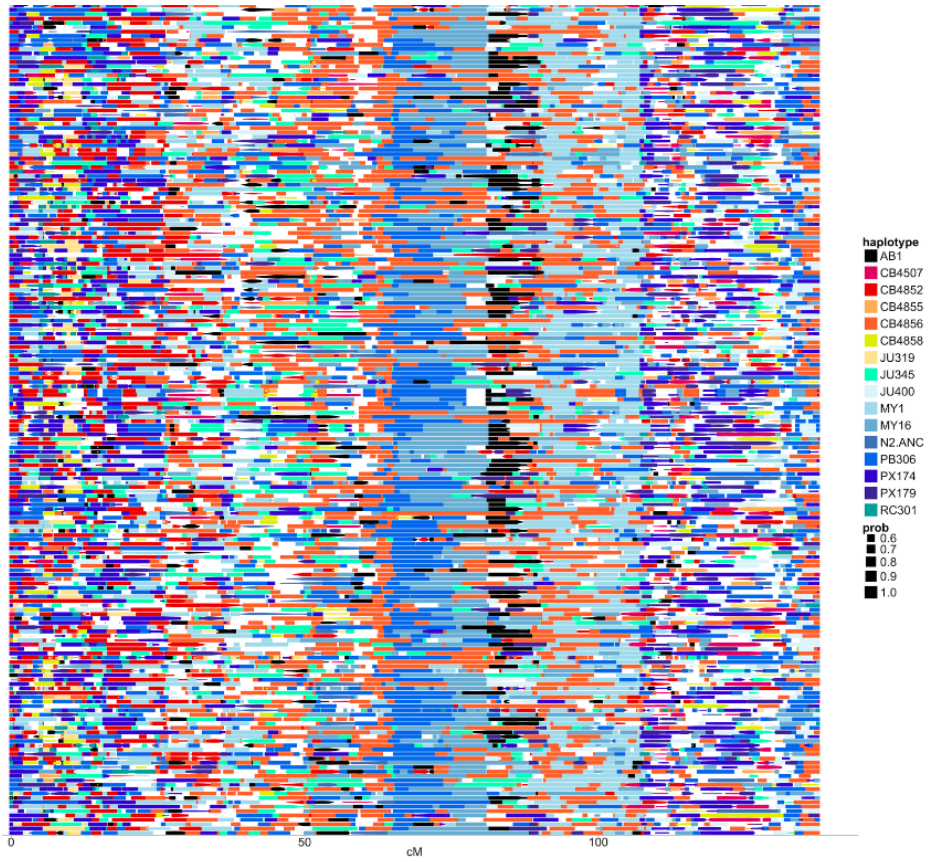
Chromosome painting: Experimental populations of Caenorhabditis elegans (Teotonio et al ('12))

- Start with 16 individuals.
- Build a population of size $\sim 10^4$ by random intercross
- Let it evolve during during 140 generations at controlled population size.
- Genotype 180 sequences.





Chromosome 3



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

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 ?
- etc.

An Haploïd Wright Fisher Model with Recombination

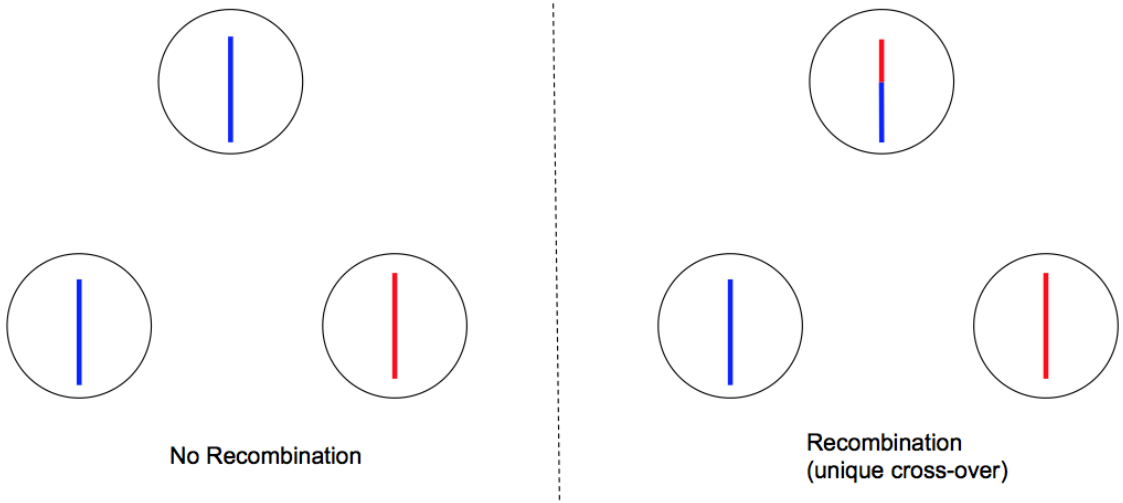
- Population of constant size N.
- Haploïd population: Each individual carries one chromosome of size R.
- Discrete time dynamics:

time 0 Each chromosome is uniformly colored with a distinct color.

time 1 Each individual chooses 2 parents from the previous generation:

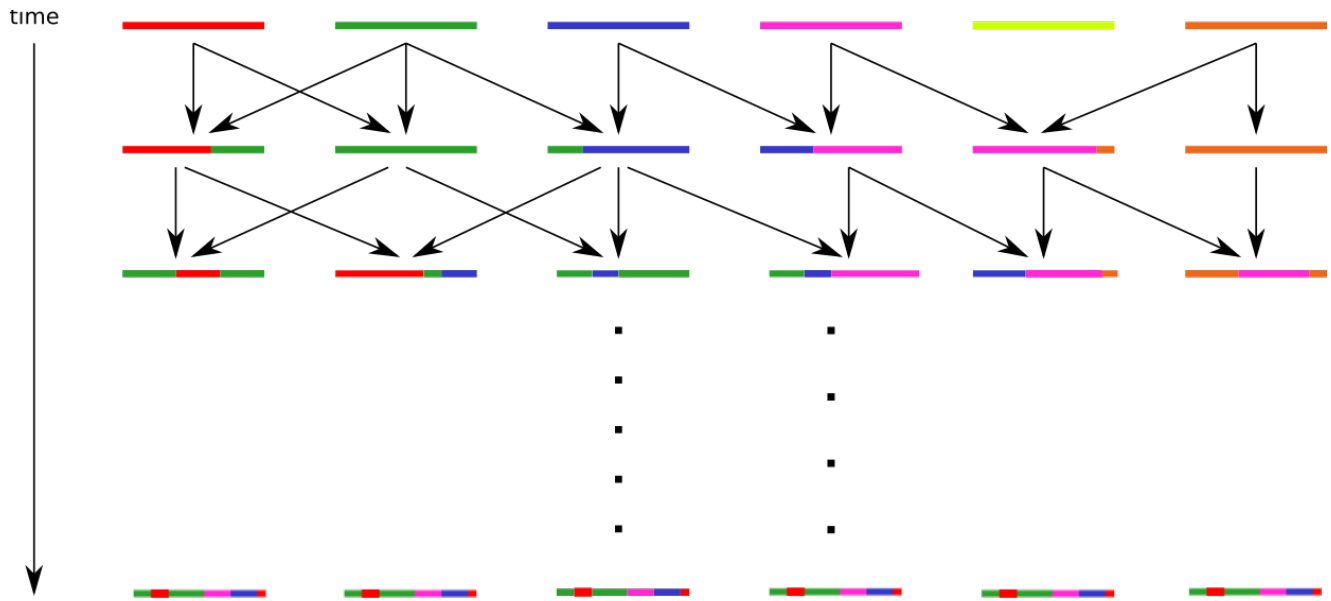
proba $1-\rho$ copies one parent chromosome.

proba ρ (Recombination event): a cross-over occurs.



An Haploïd Wright Fisher Model with Recombination

- At time 1, the population consist of N individuals, whose unique chromosome is either uniformly colored, or is partioned into two segments of distinct colors.



- After k steps, each chromosome is a mosaïc of colors, each colors corresponding to the genetic material of an ancestral individual.

- No mutation
- By genetic drift, the system a.s. reaches fixation after a finite (random) time, i.e., every individual in the system carries the same genetic material, and the system stops evolving.

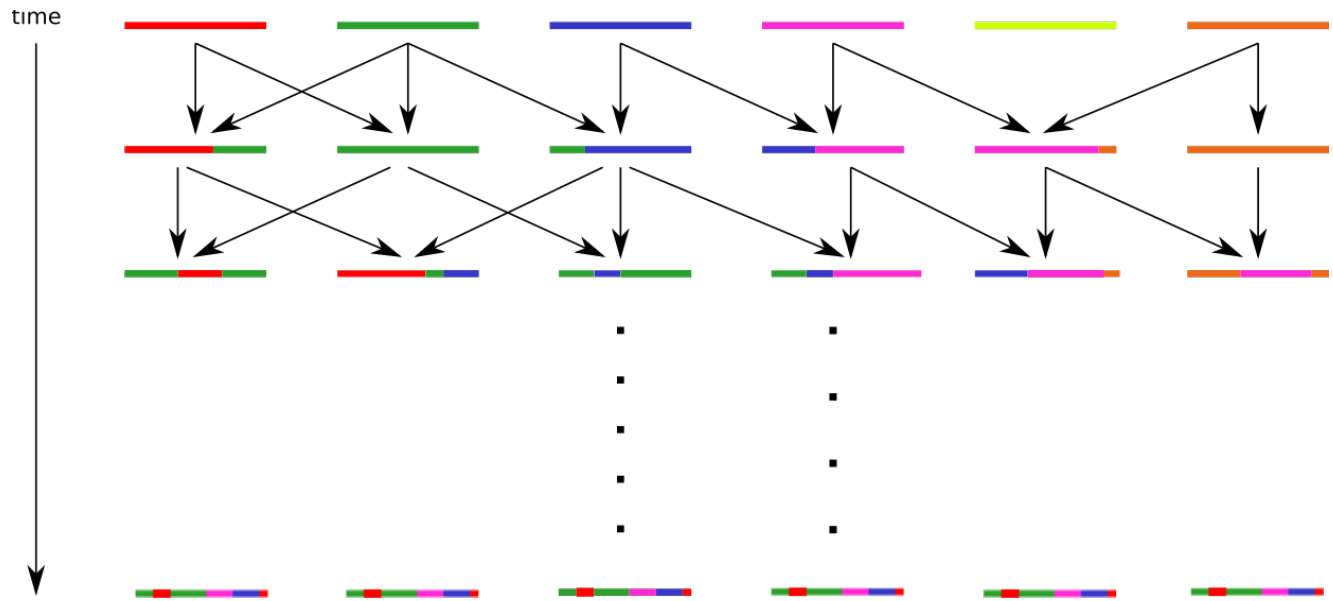


Figure: 6 segments. 4 clusters

- (N, R)-Partitioning process Π_N^R: partition of colors 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

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

(the longer the chromosome, the higher the probability of recombination).

Proposition

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

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

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

Define

$$\mathcal{L}_R \;=\; \int_0^1 1_{0\sim x}\, dx$$

the length of the cluster covering the origin.

Theorem (Lambert, Miro Pina, S.)

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

Cluster covering the origin

For every $a < b$ in $[0,1]$, let

$$\nu^R([a,b]) \;=\; \frac{1}{\log(R)} \int_{R^a}^{R^b} 1_{0\sim x}\, dx$$

Corollary of the previous result

$$\lim_{R\to\infty} \quad \nu^R([0,x]) \;=\; x\mathcal{E}(1) \;\text{ in law.}$$

Conjecture

Consider m^∞ the PPP on $[0,1]\times\mathbb{R}^+$ with intensity measure $\frac{1}{x}\exp(-y/x)dxdy$ then

$$\nu^R \implies \nu^\infty \;=\; \int_0^\infty m^\infty(dxdy)dy$$

Number of segments and clusters

Theorem (Lambert, Miro Pina, S.)

Let S_R be the number of segments in the interval $[0, R]$. Then

$$\lim_{R\to\infty} \frac{1}{R} S_R = 1 \text{ in probability.}$$

Typical size of a cluster on $[0, R]$ is of the order $\log(R)$. Thus, the number of clusters M_R should be of the order $R/\log(R)$.

Theorem (Lambert, Miro Pina, S.)

Let $\epsilon > 0$ and let $M_{R,\epsilon}$ be the number of clusters in the interval $[0, R]$ whose length is greater than $\epsilon \log(R)$. Then

$$\lim_{\epsilon\to 0} \lim_{R\to\infty} \frac{\ln(R)}{R} M_{R,\epsilon} = 1 \text{ in probability.}$$

Number of Clusters Continued

Conjecture (Wiuf and Hein 97)

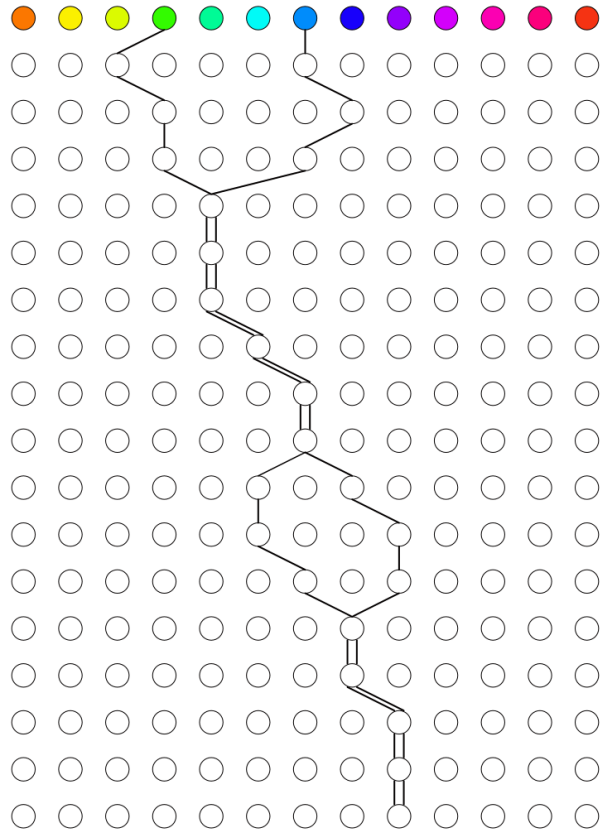
There exists a constant c such that $\frac{\ln(R)}{R} M_R \to c$ (in law, a.s. ?), with $c \approx 1.38 > 1$

For humans chromosome 1: $R \approx 5 \times 10^4$, and thus, the number of ancestors for chromosome 1 is approximatively $M_R \approx 6400$.

Idea of the proofs.

The Ancestral Recombination Graph (ARG): two sites

- Consider two sites x and y at distance l and 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 $-\infty$

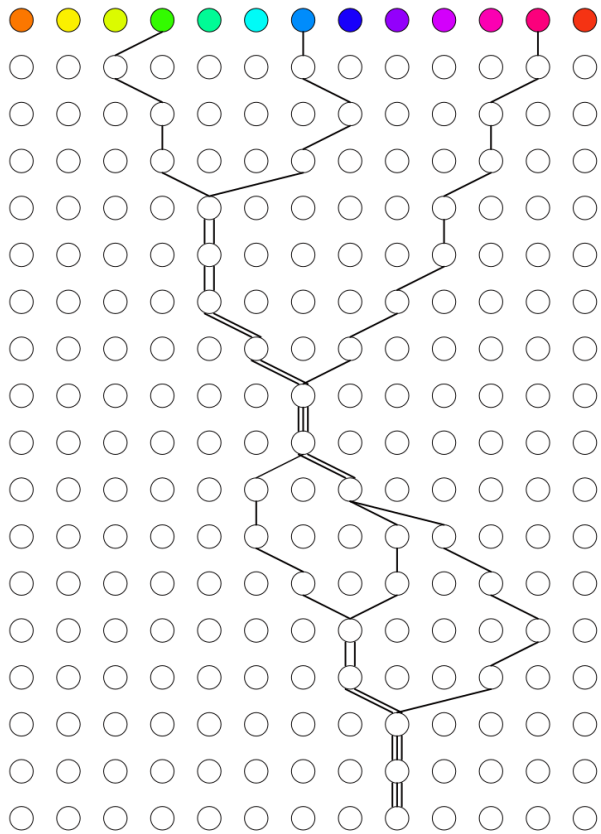


The Ancestral Recombination Graph (ARG): three sites

- Consider three sites $\{x, y, z\}$ with $x < y < z$ and

$$d(x, y) = l_1 \quad \text{and} \quad d(y, z) = l_2$$

- At each generation, the three lines of ascent split
  - into $\{x, y\}$ and $\{z\}$ with probability l_2/N.
  - into $\{x\}$ and $\{y, z\}$ with probability l_1/N.
- At each generation, each pair of lines coalesce with probability $1/N$.
- x, y, z carry the same color iff their lines coincide at $-\infty$



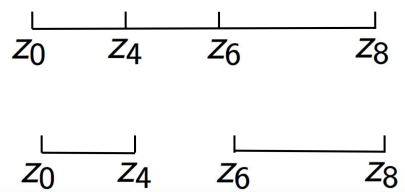
Ancestral Recombination Graph (Griffiths, Hudson)

- 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:

coalescence groups of lineages coalesce at rate 1.

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 $z_{\sigma(j+1)} - z$



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

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

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

Proof for the Cluster Size at the Origin

- Take $z_0 < z_1 < \cdots < z_n$ with $z_{i+1} - zi = R \times u_i$, $u_i > 0$.
- In the ancestral recombination graph, the most likely configuration is $\{0\} \cdots \{n\}$.

Definition

Let $\pi \in \mathcal{P}_n$. We say that π is of order k if it can be obtained from $\{0\} \cdots \{n\}$ by k successive coalescence events.

- $\{i,j\} +$ singletons is of order 1
- $\{i,j,k\} +$ singletons is of order 2. Three scenarios:

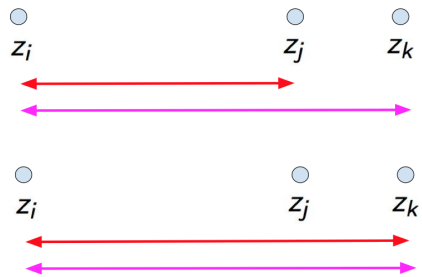
$$\begin{aligned}
\{i\}\{j\}\{k\}\cdots &\to \{i,j\}\{k\}\cdots \to \{i,j,k\}\cdots \\
\{i\}\{j\}\{k\}\cdots &\to \{i,k\}\{j\}\cdots \to \{i,j,k\}\cdots \\
\{i\}\{j\}\{k\}\cdots &\to \{k,j\}\{i\}\cdots \to \{i,j,k\}\cdots
\end{aligned}$$

- $\{i,j\},\{k,l\} +$ singletons is of order 2.
- $\cdots$
- $\{0,1,2,\cdots,n\}$ is of order n.

Energy of a coalescence scenario

$\{i,j,k\}+$ singletons is of order 2. Three scenarios:

$$
\begin{aligned}
s1: &\quad \{i\}\{j\}\{k\}\cdots \rightarrow \{i,j\}\{k\} \cdots \rightarrow \{i,j,k\}\cdots \\
s2: &\quad \{i\}\{j\}\{k\}\cdots \rightarrow \{i,k\}\{j\} \cdots \rightarrow \{i,j,k\}\cdots \\
s3: &\quad \{i\}\{j\}\{k\}\cdots \rightarrow \{k,j\}\{i\} \cdots \rightarrow \{i,j,k\}\cdots
\end{aligned}
$$



We define the energy of a coalescence scenario as the inverse of the product of the successive cover lengths at each step of the scenario.

$$
\begin{aligned}
E(s1,\mathbf{z}) &= \frac{1}{z_j - z_i} \times \frac{1}{z_k - z_i} \\
E(s2,\mathbf{z}) &= \frac{1}{z_k - z_i} \times \frac{1}{z_k - z_i}
\end{aligned}
$$

Lemma

For every $\pi, \pi' \in \mathcal{P}_n$ of order k and l with $l > k$

$$\mu^{\mathbf{z}}(\pi)/\mu^{\mathbf{z}}(\pi') \to 0$$

Corollary

Let $\pi \in \mathcal{P}_n$ of order k, then

$$\lim_{R\to\infty} R^k \mu^{\mathbf{z}}(\pi) = \lim_{R\to\infty} R^k \sum_{\mathbf{S}} E(\mathbf{S}, \mathbf{z})$$

where the sum is taken over every possible coalescence scenarii to go from $\{1\} \cdots \{n\}$ to π.

Proof.

Solve ${}^t\pi\ M^{\mathbf{z}} = 0$ using the previous approximation, where $M^{\mathbf{z}}$ is the transition matrix for the ARG process. □

$$
\begin{aligned}
\frac{1}{\log(R)^n}\mathbb{E}\left(\mathcal{L}_R^n\right) &= \frac{R^n}{\log(R)^n} \times \\
&\quad \frac{1}{R^n}\int_{[0,R]^n} \mu^{\mathbf{z}}(\{0,\cdots,n\})dV \\
&\approx \frac{R^n}{\log(R)^n} \times \\
&\quad \frac{1}{R^n}\int_{[0,R]^n} \sum_{\mathbf{S}} E(\mathbf{S},\mathbf{z})\ dV \\
&\approx n!
\end{aligned}
$$

Open Questions

- Law of large number for the number of clusters.
- Central Limit Theorems for number of clusters and segments to build confidence intervals for our null model.
- Detecting selection, epistasis etc.

Thank you !