Some mathematical models from population genetics

2: Recombination

Alison Etheridge

University of Oxford

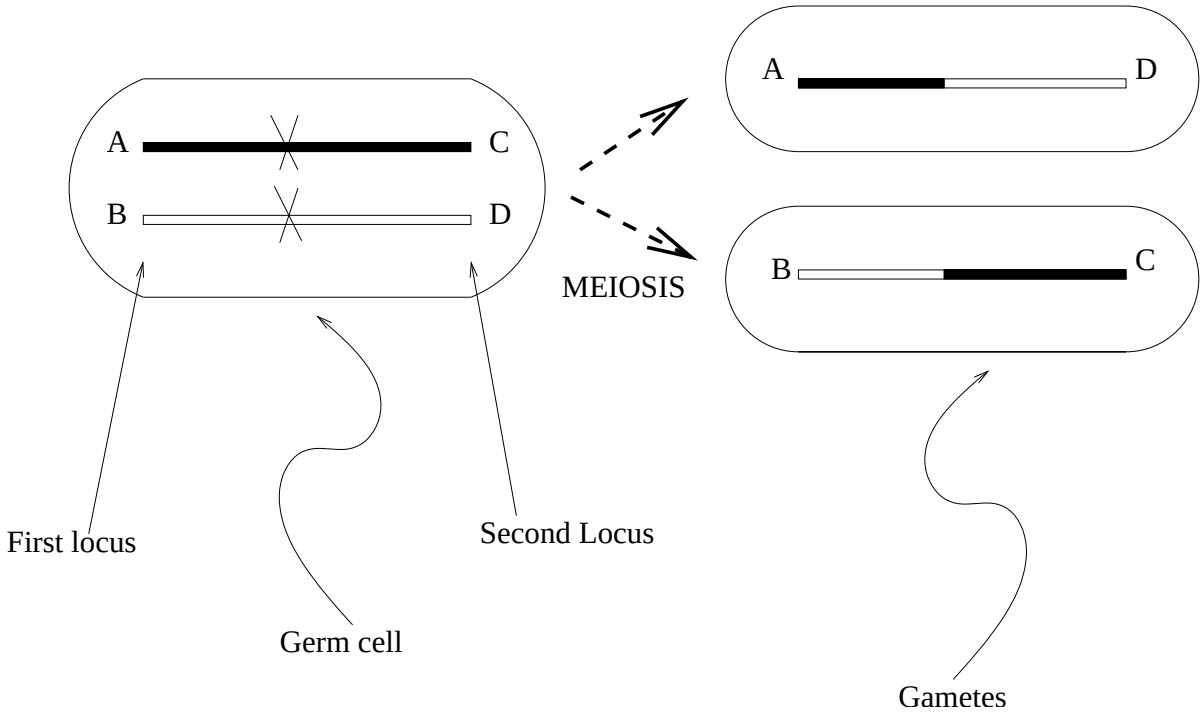
joint work with Stuart Baird (Montpellier) and Nick Barton (Edinburgh)

What is recombination?

In a diploid population, chromosomes are carried in pairs, one inherited from the mother, one from the father. But the chromosomes are not faithful copies of the parental chromosomes. One reason is recombination.

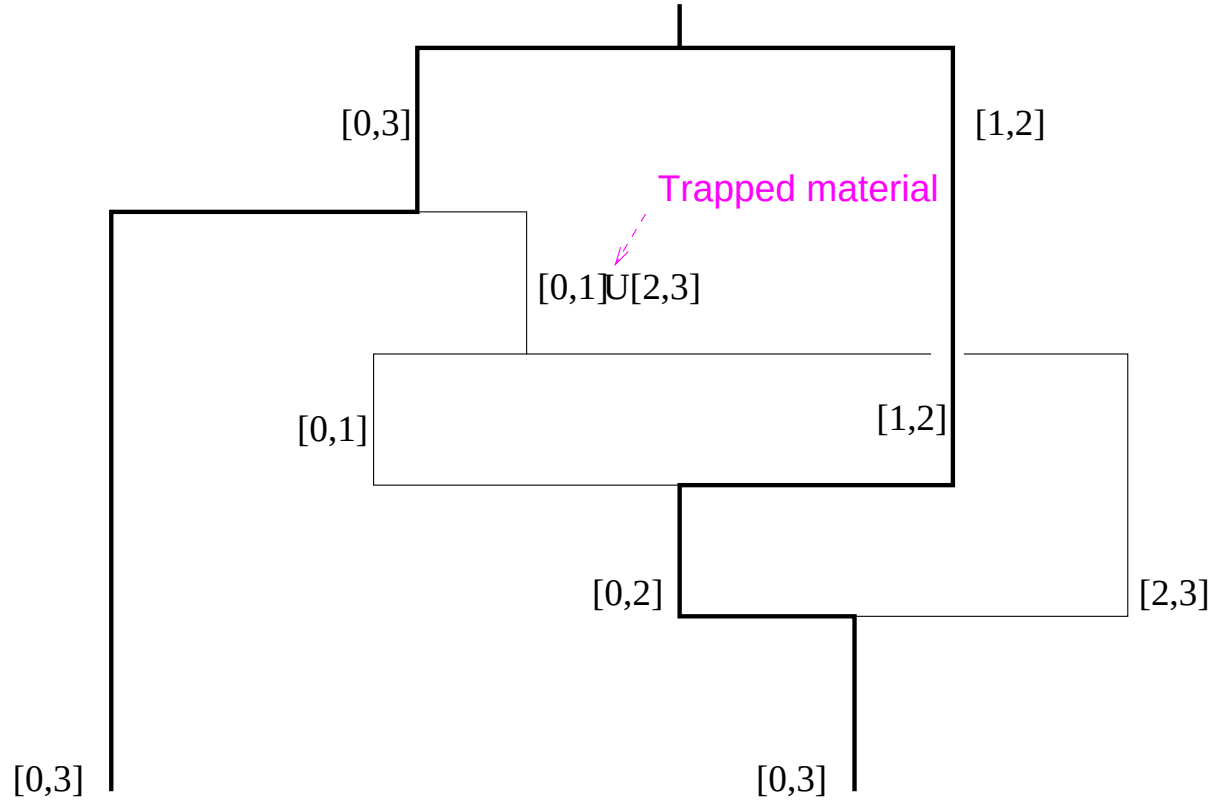
What is recombination?

In a diploid population, chromosomes are carried in pairs, one inherited from the mother, one from the father. But the chromosomes are not faithful copies of the parental chromosomes. One reason is recombination.



The ancestral recombination graph

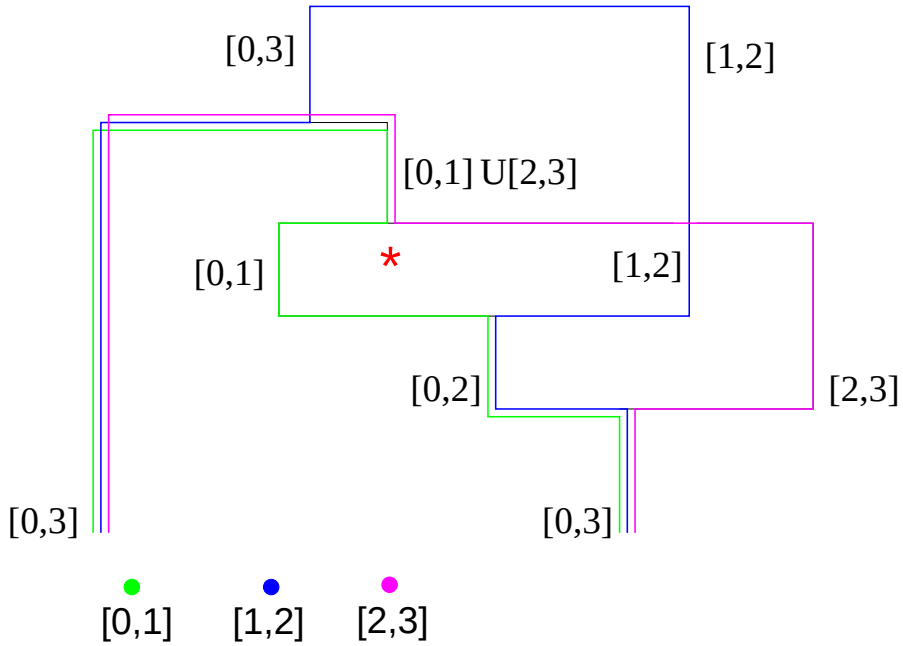
At a recombination event, we must trace *two* ancestral lineages: we see branches as well as coalescences in the genealogy.



Ancestry of the block denoted $[0, 3]$ for a sample of size two.

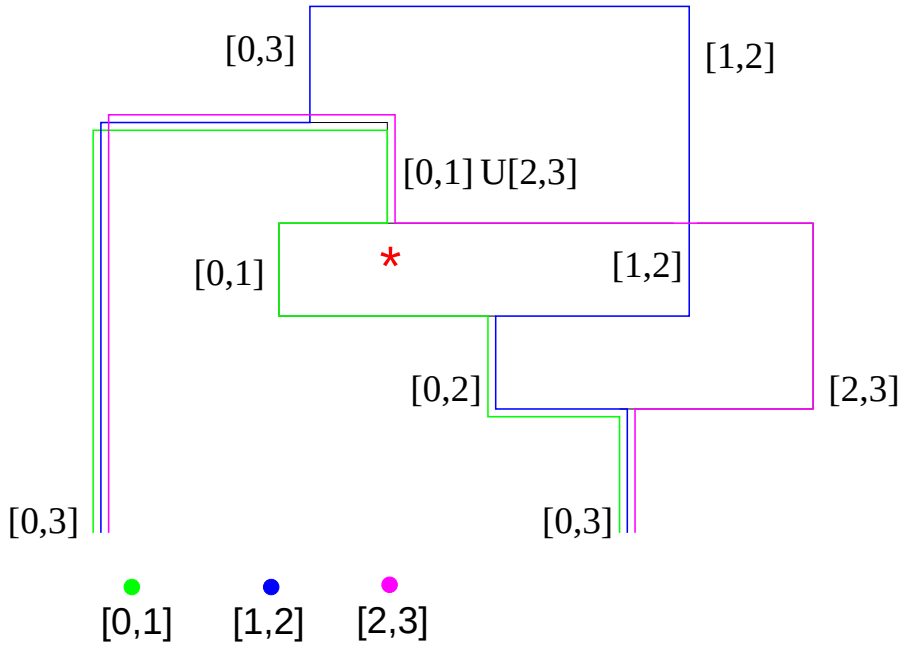
Local trees

Wiuf and Hein scan along the genome and study the process of ‘local trees’.



Local trees

Wiuf and Hein scan along the genome and study the process of ‘local trees’.



Knowing only the local tree for $[1,2]$, would not see the coalescence $*$.

Local trees do not form a Markov process.

A diversion

To trace the ancestry of a block of genome we must trace the joint location of multiple blocks.

A diversion

To trace the ancestry of a block of genome we must trace the joint location of multiple blocks.

Analytic results are hard to find. We consider a simpler process: the descent of a block of genome *forwards* in time.

The model

Each individual mates with an unrelated individual to produce a Poiss$(2(1+s))$ number of offspring. $0 \leq s \ll 1$.

The model

Each individual mates with an unrelated individual to produce a Poiss$(2(1+s))$ number of offspring. $0 \leq s \ll 1$.

Genome of map length $y \leq 1$. With probability y there is one crossover at a uniformly distributed point on the block.

The model

Each individual mates with an unrelated individual to produce a Poiss$(2(1+s))$ number of offspring. $0 \leq s \ll 1$.

Genome of map length $y \leq 1$. With probability y there is one crossover at a uniformly distributed point on the block.
Descendant of genome inherits:

Block length	probability
0	$\frac{1}{2}(1-y)$
y	$\frac{1}{2}(1-y)$
$U(0,y)$	y

The model

Each individual mates with an unrelated individual to produce a Poiss$(2(1+s))$ number of offspring. $0 \leq s \ll 1$.

Genome of map length $y \leq 1$. With probability y there is one crossover at a uniformly distributed point on the block.
Descendant of genome inherits:

Block length	probability
0	$\frac{1}{2}(1-y)$
y	$\frac{1}{2}(1-y)$
$U(0,y)$	y

Note: If $s = 0$, the expected total block length is conserved.

Some questions

- How long does the ancestral genome persist?

Some questions

- How long does the ancestral genome persist?
- What is the distribution of surviving blocks at time t?

Some questions

- How long does the ancestral genome persist?
- What is the distribution of surviving blocks at time t?

Application: statistical framework for interpreting data arising from sporadic hybridisation.

Some questions

- How long does the ancestral genome persist?
- What is the distribution of surviving blocks at time t?

Application: statistical framework for interpreting data arising from sporadic hybridisation.

Notation: $Q_t(y) =$ probability total loss by time t of ancestral block of length y. $P_t(y) = 1 - Q_t(y)$.

Loss of ancestral genome by time t

Condition on number of offspring of ancestral genome:

Loss of ancestral genome by time t

Condition on number of offspring of ancestral genome:

$$Q_{t+1}(y) = \Phi\left[\frac{1-y}{2} + \frac{1-y}{2}Q_t(y) + \int_0^y Q_t(z)dz\right], \quad Q_0(y) = 0.$$

Loss of ancestral genome by time t

Condition on number of offspring of ancestral genome:

$$Q_{t+1}(y) = \Phi\left[\frac{1-y}{2} + \frac{1-y}{2}Q_t(y) + \int_0^y Q_t(z)dz\right], \quad Q_0(y) = 0.$$

Substituting for Φ,

$$Q_{t+1}(y) = \exp\left[-2(1+s)\left(\frac{1-y}{2}P_t(y) + \int_0^y P_t(z)dz\right)\right],$$

Loss of ancestral genome by time t

Condition on number of offspring of ancestral genome:

$$Q_{t+1}(y) = \Phi\left[\frac{1-y}{2} + \frac{1-y}{2}Q_t(y) + \int_0^y Q_t(z)dz\right], \quad Q_0(y) = 0.$$

Substituting for Φ,

$$Q_{t+1}(y) = \exp\left[-2(1+s)\left(\frac{1-y}{2}P_t(y) + \int_0^y P_t(z)dz\right)\right],$$

or, in differential form,

$$\frac{d}{dy}P_{t+1}(y) = (1+s)\,(1-P_{t+1}(y))\left(P_t(y) + (1-y)\frac{d}{dy}P_t(y)\right).$$

Probability of survival for ever

$$\frac{d\tilde{P}}{dy}(y) = (1+s)\tilde{Q}(y)\left(\tilde{P}(y) + (1-y)\frac{d\tilde{P}}{dy}(y)\right).$$

Probability of survival for ever

$$\frac{d\tilde{P}}{dy}(y) = (1+s)\tilde{Q}(y)\left(\tilde{P}(y) + (1-y)\frac{d\tilde{P}}{dy}(y)\right).$$

General solution

$$\tilde{P}_C = \frac{y^*}{y^* + \pi\left(Cy^* e^{-y^*}\right)}, \quad y^* = y - s(1-y),$$

where the *product log function*, π, is defined by $z = \pi(z)e^{\pi(z)}$.

Probability of survival for ever

$$\frac{d\tilde{P}}{dy}(y) = (1+s)\tilde{Q}(y)\left(\tilde{P}(y) + (1-y)\frac{d\tilde{P}}{dy}(y)\right).$$

General solution

$$\tilde{P}_C = \frac{y^*}{y^* + \pi\left(Cy^* e^{-y^*}\right)}, \quad y^* = y - s(1-y),$$

where the *product log function*, π, is defined by $z = \pi(z)e^{\pi(z)}$.
$\tilde{P}(0)$ is survival probability of a branching process with Poiss$(1+s)$ offspring distribution:

Probability of survival for ever

$$\frac{d\tilde{P}}{dy}(y) = (1+s)\tilde{Q}(y)\left(\tilde{P}(y) + (1-y)\frac{d\tilde{P}}{dy}(y)\right).$$

General solution

$$\tilde{P}_C = \frac{y^*}{y^* + \pi\left(Cy^*e^{-y^*}\right)}, \quad y^* = y - s(1-y),$$

where the *product log function*, π, is defined by $z = \pi(z)e^{\pi(z)}$.
$\tilde{P}(0)$ is survival probability of a branching process with Poiss$(1+s)$ offspring distribution: (using $\pi(z) \sim z$ as $z \downarrow 0$)

$$\tilde{P}(y) = \frac{y^*}{y^* + \pi\left(\frac{\tilde{Q}(0)}{\tilde{P}(0)}y^*e^{-y^*}\right)}.$$

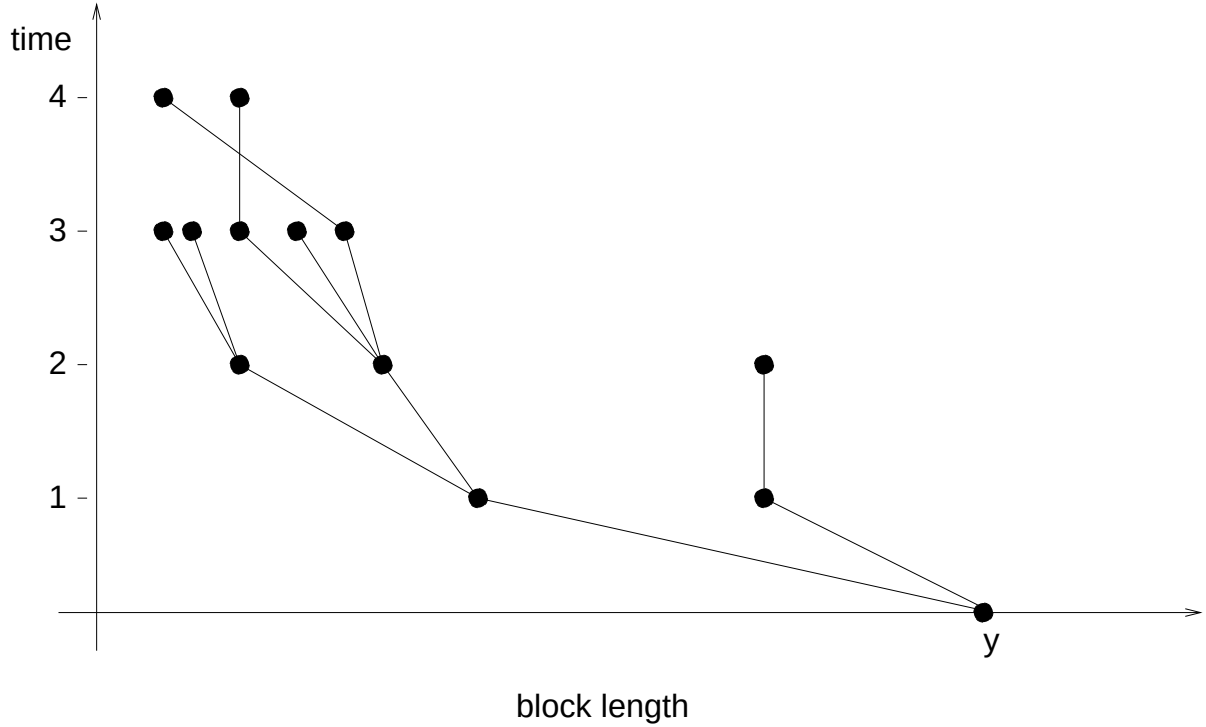
Back to finite times

Back to finite times

Think of process as branching random walk.

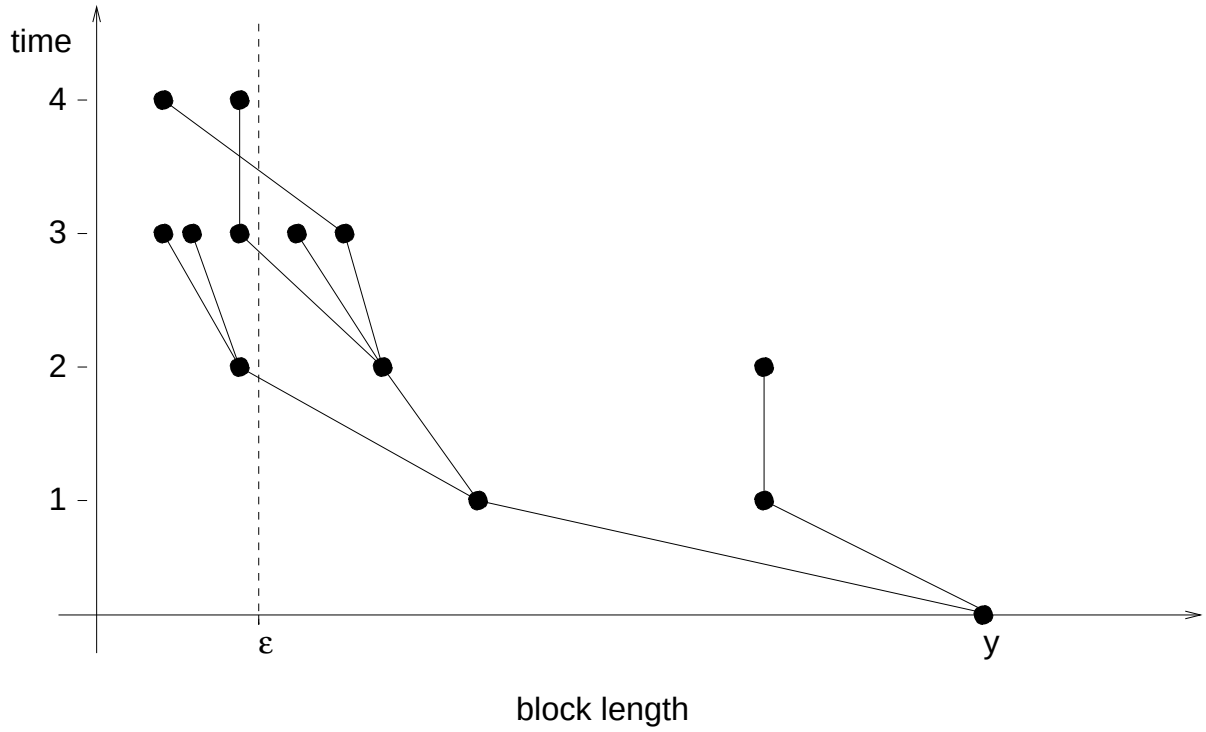
Back to finite times

Think of process as branching random walk.



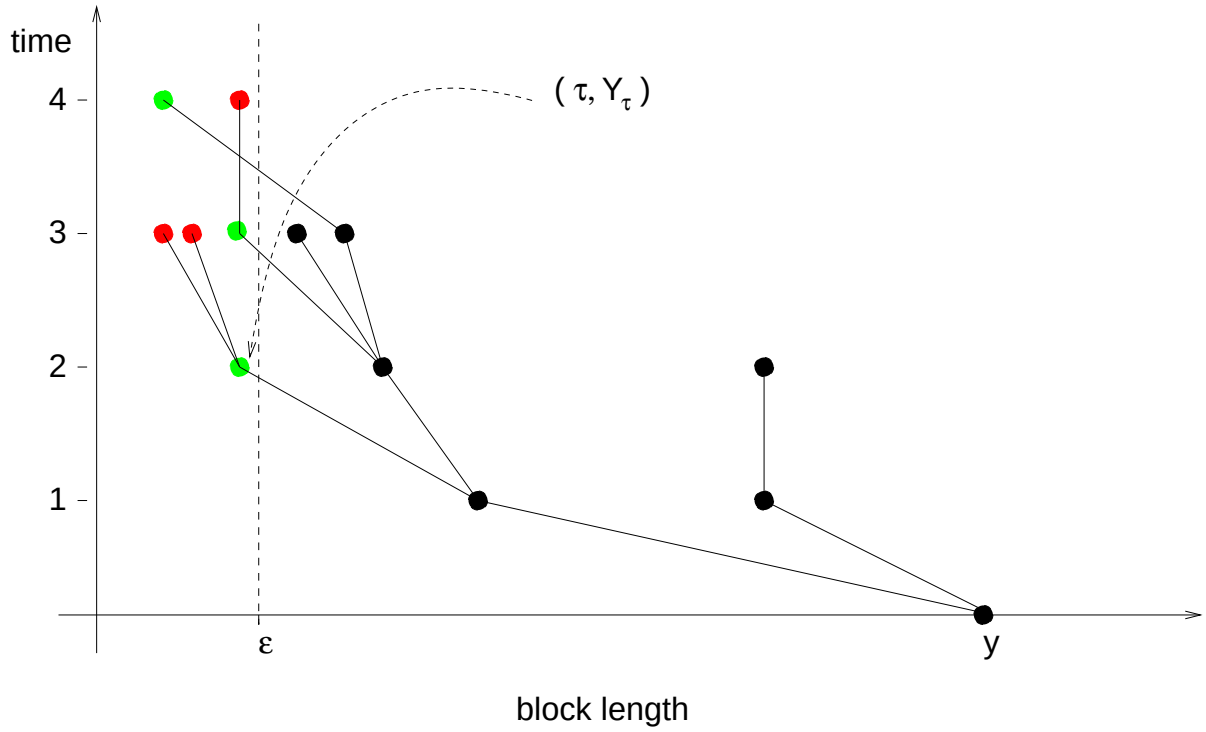
Back to finite times

Think of process as branching random walk. *Freeze* individuals on exit from $[\epsilon, y] \times [0, t]$.



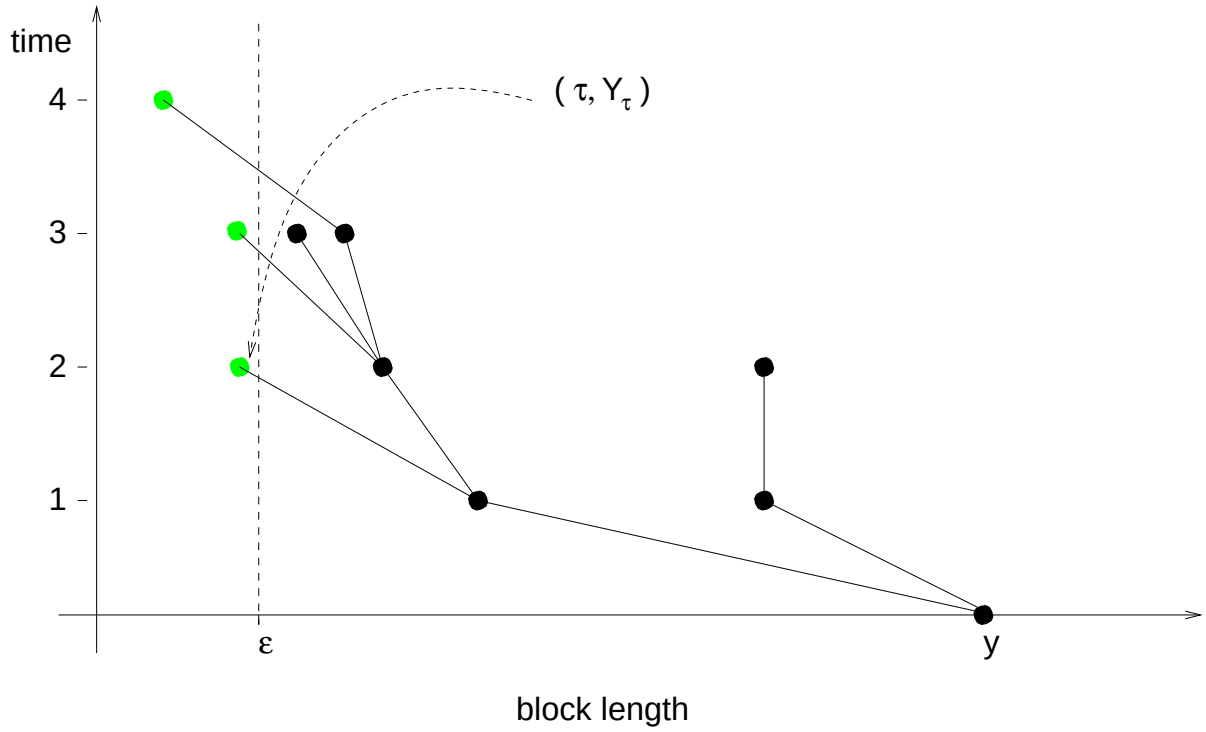
Back to finite times

Think of process as branching random walk. *Freeze* individuals on exit from $[\epsilon, y] \times [0, t]$.



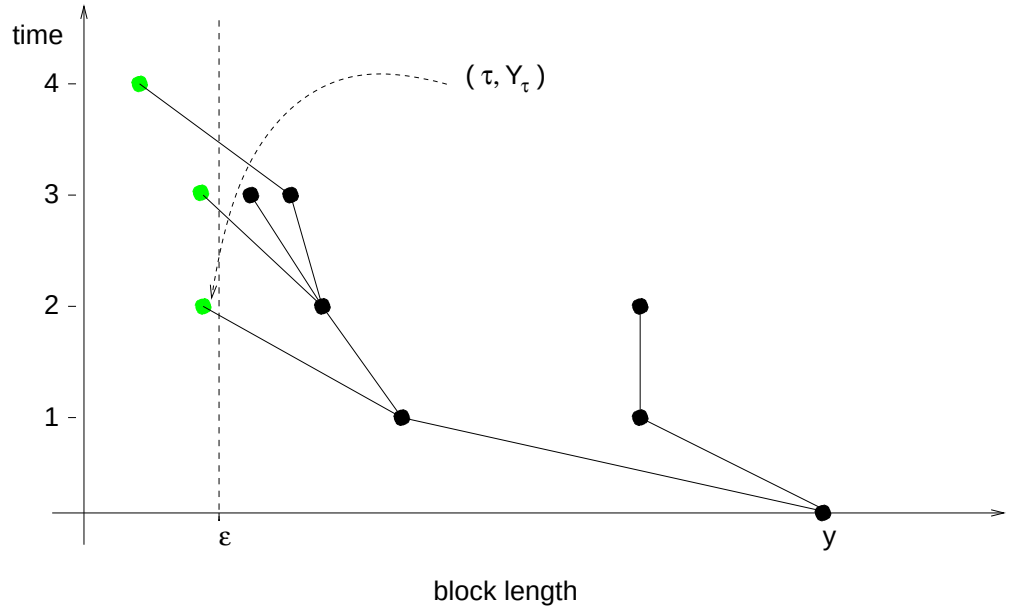
Back to finite times

Think of process as branching random walk. *Freeze* individuals on exit from $[\epsilon, y] \times [0, t]$.



A Special Markov property

Notation: $N_\tau =$ number of individuals in new process at time t. $\tau_i =$ time of freezing of ith particle. $Y_{\tau_i} =$ corresponding block length.



$$Q_t(y) = \mathbb{E}\left[\prod_{i=1}^{N_\tau} Q_{t-\tau_i}(Y_{\tau_i})\right].$$

$$Q_t(y) = \mathbb{E}\left[\prod_{i=1}^{N_\tau} Q_{t-\tau_i}(Y_{\tau_i})\right].$$

Suppose $\epsilon \ll 1$ and $\max_i \tau_i \ll t$ then can approximate $Q_{t-\tau_i}(Y_{\tau_i})$ by $Q_t(0)$.

When is this valid?

$$Q_t(y) = \mathbb{E}\left[\prod_{i=1}^{N_\tau} Q_{t-\tau_i}(Y_{\tau_i})\right].$$

Suppose $\epsilon \ll 1$ and $\max_i \tau_i \ll t$ then can approximate $Q_{t-\tau_i}(Y_{\tau_i})$ by $Q_t(0)$.

When is this valid?

Crude bound:

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \mathbb{E}\Big[\#\big\{\text{individuals carrying block length} \geq \epsilon \text{ at time } t_0\big\}\Big]$$

$$Q_t(y) = \mathbb{E}\left[\prod_{i=1}^{N_\tau} Q_{t-\tau_i}(Y_{\tau_i})\right].$$

Suppose $\epsilon \ll 1$ and $\max_i \tau_i \ll t$ then can approximate $Q_{t-\tau_i}(Y_{\tau_i})$ by $Q_t(0)$.

When is this valid?

Crude bound:

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \mathbb{E}\Big[\#\big\{\text{individuals carrying block length} \geq \epsilon \text{ at time } t_0\big\}\Big]$$

To estimate the right-hand side we superimpose recombinations on a *pedigree*.

Recombination on a pedigree

The *pedigree* is the tree of *all* descendants of the ancestor.

Take initial block length $y = 1$*.*

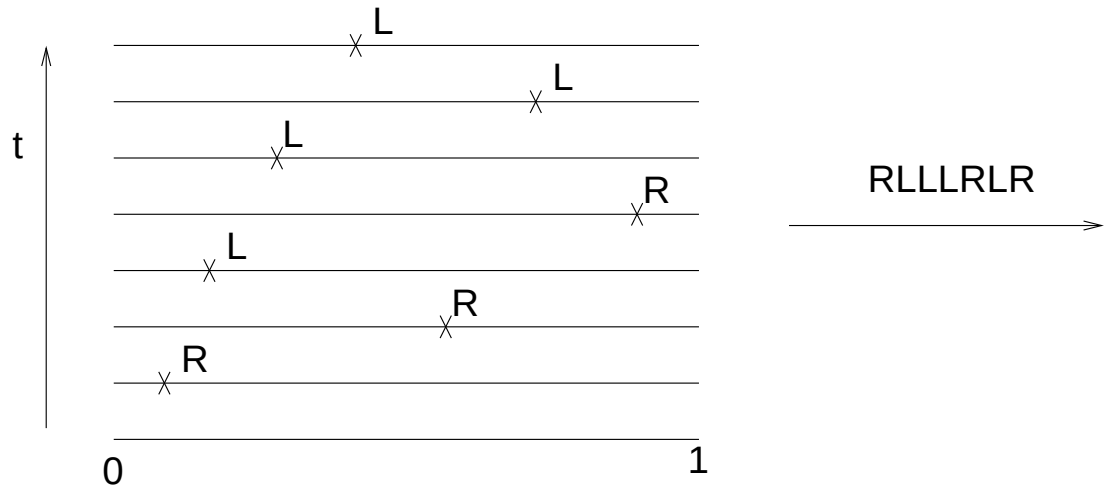
Consider one line of descent through the pedigree:

Recombination on a pedigree

The *pedigree* is the tree of *all* descendants of the ancestor.

Take initial block length $y = 1$.

Consider one line of descent through the pedigree:

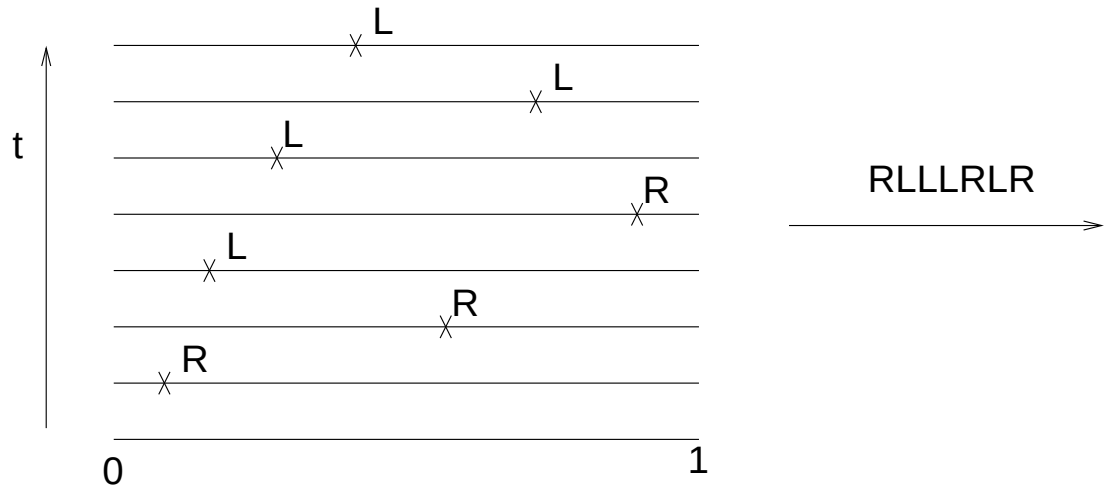


Recombination on a pedigree

The *pedigree* is the tree of *all* descendants of the ancestor.

Take initial block length $y = 1$.

Consider one line of descent through the pedigree:



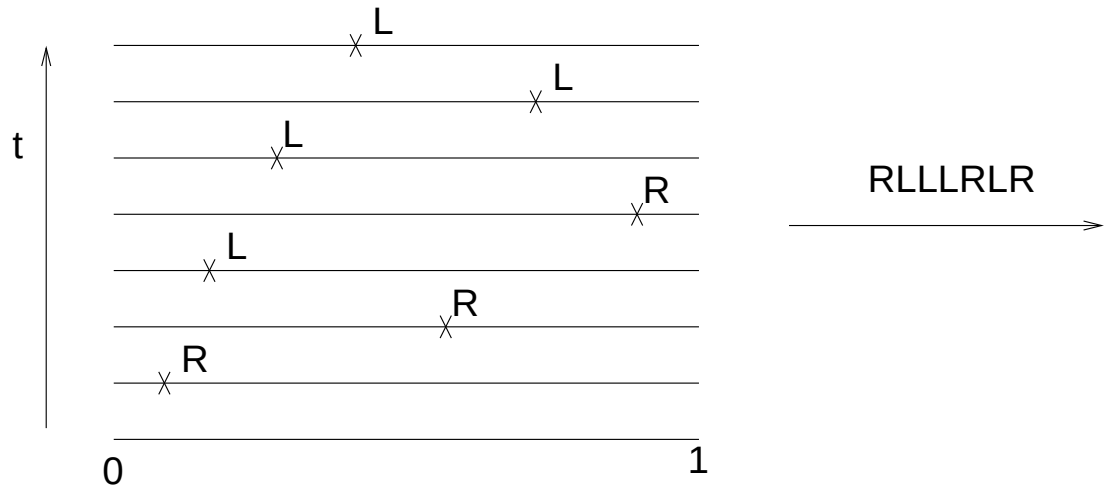
If an ‘L’ mark is followed by an ‘R’ mark, all ancestral genome is lost.

Recombination on a pedigree

The *pedigree* is the tree of *all* descendants of the ancestor.

Take initial block length $y = 1$.

Consider one line of descent through the pedigree:



If an 'L' mark is followed by an 'R' mark, all ancestral genome is lost.

Survival requires $\underbrace{RR\ldots R}_{m \text{ times}}\ \underbrace{LL\ldots L}_{t_0-m \text{ times}}$ for some $m \in \{0, 1, \ldots, t_0\}$.

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.

Probability such a block has length $\geq \epsilon$ is at most $(1-\epsilon)^{t_0}$.

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.

Probability such a block has length $\geq \epsilon$ is at most $(1-\epsilon)^{t_0}$.

Combining the above,

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \left[(1+s)(1-\epsilon)\right]^{t_0} (t_0+1).$$

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.
Probability such a block has length $\geq \epsilon$ is at most $(1-\epsilon)^{t_0}$.
Combining the above,

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \left[(1+s)(1-\epsilon)\right]^{t_0} (t_0+1).$$

Choose $\epsilon > \frac{s}{1+s}$ for this to decay rapidly. Then

$$Q_t(y) \approx \mathbb{E}\left[Q_t(0)^{N_\tau}\right].$$

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.
Probability such a block has length $\geq \epsilon$ is at most $(1-\epsilon)^{t_0}$.
Combining the above,

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \left[(1+s)(1-\epsilon)\right]^{t_0} (t_0+1).$$

Choose $\epsilon > \frac{s}{1+s}$ for this to decay rapidly. Then

This gives

$$Q_t(y) \approx \mathbb{E}\left[Q_t(0)^{N_\tau}\right].$$

$$P_t(y) \approx \frac{y^*}{y^* + \pi\left(\frac{Q_t(0)}{P_t(0)} y^* e^{-y^*}\right)}.$$

Survival until time t

Probability of any block being passed down $= \frac{t_0+1}{2^{t_0}}$.
Probability such a block has length $\geq \epsilon$ is at most $(1-\epsilon)^{t_0}$.
Combining the above,

$$\mathbb{P}\left[\max_i \tau_i > t_0\right] \leq \left[(1+s)(1-\epsilon)\right]^{t_0} (t_0+1).$$

Choose $\epsilon > \frac{s}{1+s}$ for this to decay rapidly. Then

This gives

$$Q_t(y) \approx \mathbb{E}\left[Q_t(0)^{N_\tau}\right].$$

$$P_t(y) \approx \frac{y^*}{y^* + \pi\left(\frac{Q_t(0)}{P_t(0)} y^* e^{-y^*}\right)}.$$

Approximate $P_t(0)$ e.g. via Feller's diffusion.

An example

Suppose $s = 0$ (so $y^* = y$) and $yt \gg 1$, since $\pi(z) \sim \log z$ as $z \to \infty$,

$$P_t(y) \sim \frac{y}{\log(yt/2)}.$$

Survival until time t declines like $1/\log t$.

Compare to $1/t$ for a single locus.

An example

Suppose $s = 0$ (so $y^* = y$) and $yt \gg 1$, since $\pi(z) \sim \log z$ as $z \to \infty$,

$$P_t(y) \sim \frac{y}{\log(yt/2)}.$$

Survival until time t declines like $1/\log t$.

Compare to $1/t$ for a single locus.

Recombination rapidly breaks the ancestral genome into small blocks, but these can persist for a very long time.

Long genomes

What about long genomes?

Crossovers according to a Poisson process of rate one.

Long genomes

What about long genomes?

Crossovers according to a Poisson process of rate one.

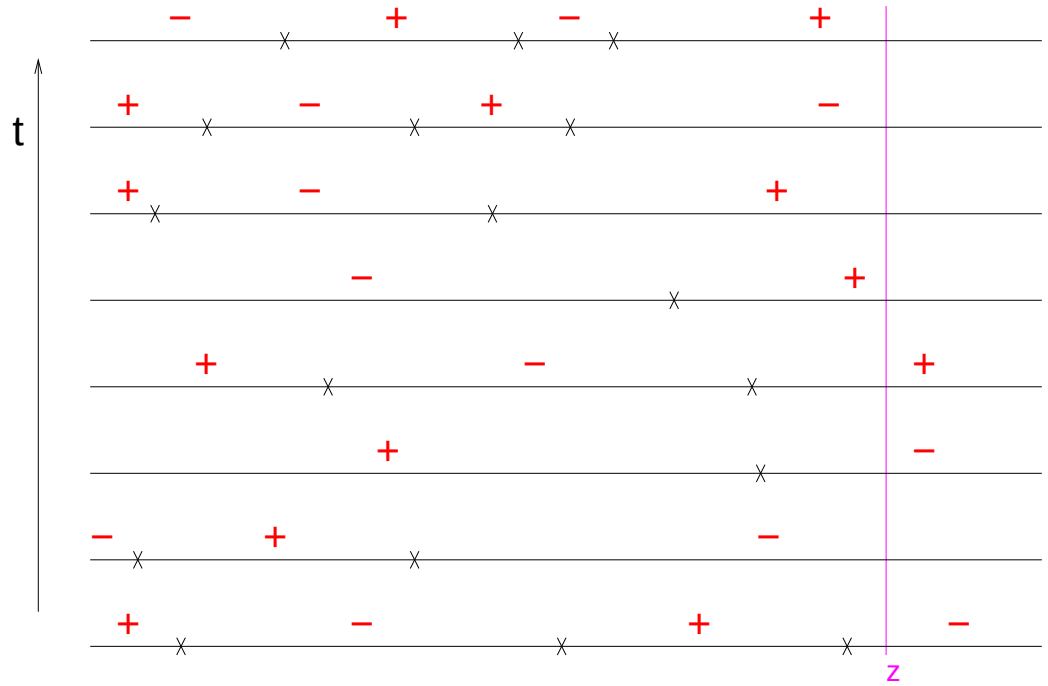
What is the mean number of individuals to inherit some ancestral material at time t?

Long genomes

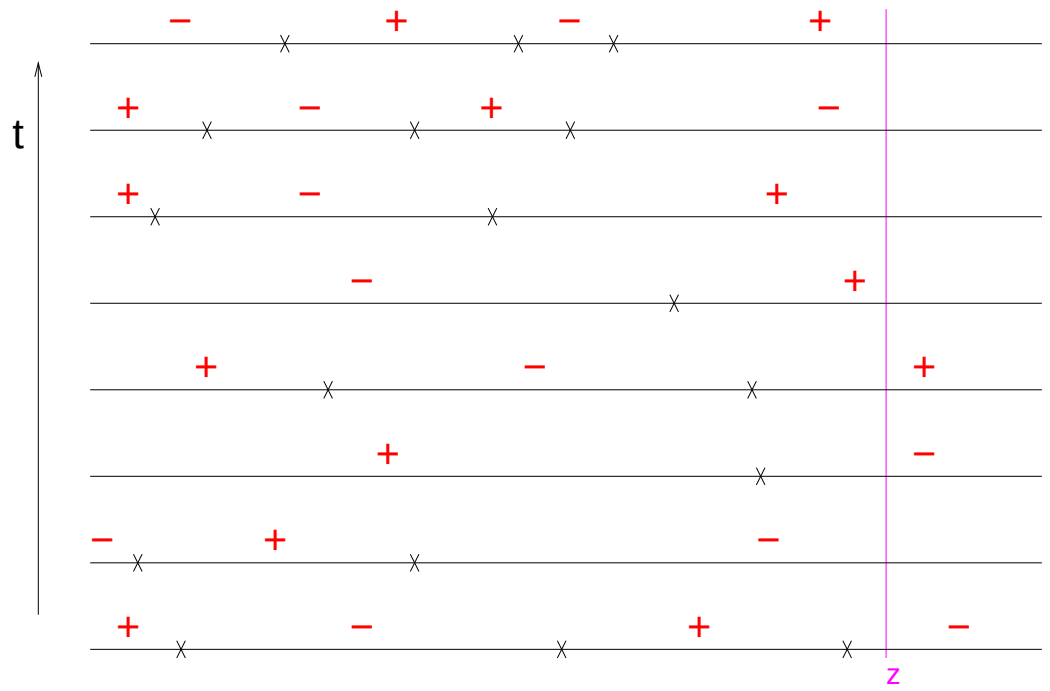
What about long genomes?
Crossovers according to a Poisson process of rate one.

What is the mean number of individuals to inherit some ancestral material at time t?

Again consider a single line of descent



Label $z \in [0, y]$ by $\underline{a}(z) = (a_1(z), a_2(z), \ldots, a_t(z)) \in \{-, +\}^t$.



a(z)=− − − + + + − +

Label $z \in [0, y]$ by $\underline{a}(z) = (a_1(z), a_2(z), \ldots, a_t(z)) \in \{-, +\}^t$. A point z is in a block that is passed down iff $\underline{a}(z) = (+, +, \ldots, +)$.

A change of perspective

Define continuous time Markov chain $\{X_z\}_{z\in[0,y]}$ by

$$X_z = \#\{i \in \{1, 2, \ldots, t\} : a_i(z) = -\}.$$

We seek $\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right]$.

A change of perspective

Define continuous time Markov chain $\{X_z\}_{z\in[0,y]}$ by

$$X_z = \#\{i \in \{1, 2, \ldots, t\} : a_i(z) = -\}.$$

We seek $\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right]$.
Transitions of X_z occur at rate t.

$$P_{ij} = \begin{cases} \frac{i}{t} & j = i - 1 \\ \frac{t-i}{t} & j = i + 1 \\ 0 & \text{otherwise} \end{cases}.$$

A change of perspective

Define continuous time Markov chain $\{X_z\}_{z\in[0,y]}$ by

$$X_z = \#\{i \in \{1, 2, \ldots, t\} : a_i(z) = -\}.$$

We seek $\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right]$.
Transitions of X_z occur at rate t.

$$P_{ij} = \begin{cases} \frac{i}{t} & j = i - 1 \\ \frac{t-i}{t} & j = i + 1 \\ 0 & \text{otherwise} \end{cases}.$$

Continuous time version of the *Ehrenfest model*. P & T Ehrenfest (1907).

Some consequences

From Bellman & Harris (1951) we deduce

$$\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right] \approx \frac{1}{2^t}(1 + ty).$$

Some consequences

From Bellman & Harris (1951) we deduce

$$\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right] \approx \frac{1}{2^t}(1 + ty).$$

- Mean number of individuals carrying any ancestral material $\approx (1 + s)^t(ty + 1)$.

Some consequences

From Bellman & Harris (1951) we deduce

$$\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right] \approx \frac{1}{2^t}(1 + ty).$$

- Mean number of individuals carrying any ancestral material $\approx (1 + s)^t(ty + 1)$.
- The length of an inherited block is distributed approx as Exp(t).

Some consequences

From Bellman & Harris (1951) we deduce

$$\mathbb{P}\left[X_z = 0, \text{for some } z \in [0, y]\right] \approx \frac{1}{2^t}(1 + ty).$$

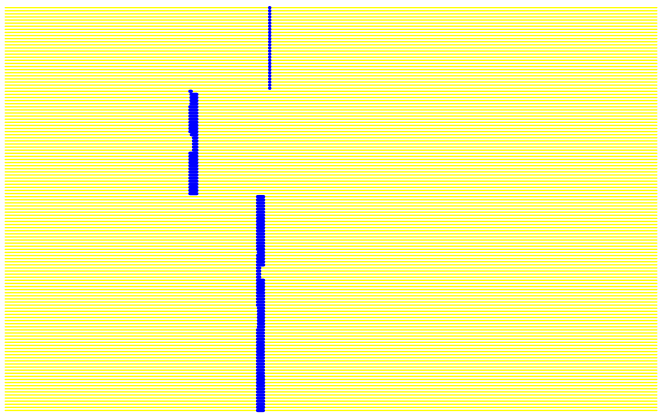
- Mean number of individuals carrying any ancestral material $\approx (1 + s)^t(ty + 1)$.
- The length of an inherited block is distributed approx as Exp(t).
- For a single line of descent, the probability of inheriting multiple blocks is at most

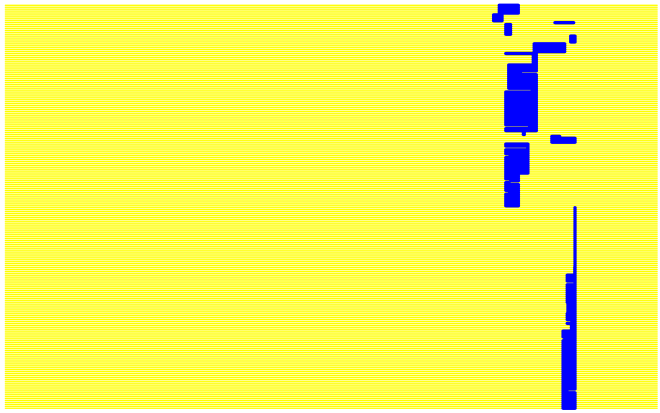
$$\mathbb{P}[\text{A single block survives}] \times \mathbb{P}[X_z = 0 \text{ for some } z \in [0, y] | X_0 = 1]$$

$$\approx \frac{(ty + 1)}{2^t} \frac{ty}{2^t}$$

For example, if $s = 0$, $y = 1$ and $t = 10$, this suggests that there is a $< 1\%$ chance of seeing multiple blocks.

We expect some portion of introgressed genome to persist for a long time, but the effect will be highly variable along the genome.





50 generations, $y = 1$.