Some mathematical models from population genetics

3: Selection

Alison Etheridge

University of Oxford

joint work with Nick Barton (Edinburgh), Anja Sturm (Delaware)

Peter Pfaffelhuber (Vienna), Anton Wakolbinger (Frankfurt)

Selective sweeps

The Moran model with selection.

A population of N genes occurring in two alleles, b and B, evolves in overlapping generations. At exponential rate $\binom{N}{2}$ a pair of genes is sampled (with replacement) from the population, one dies and the other splits in two. If the two genes are of different allelic types, then with probability $\frac{1+\sigma}{2}$ it is the B allele that reproduces.

Selective sweeps

The Moran model with selection.

A population of N genes occurring in two alleles, b and B, evolves in overlapping generations. At exponential rate $\binom{N}{2}$ a pair of genes is sampled (with replacement) from the population, one dies and the other splits in two. If the two genes are of different allelic types, then with probability $\frac{1+\sigma}{2}$ it is the B allele that reproduces.

Suppose that an allele conferring selective advantage σ arises (by mutation say) in an otherwise neutral population. With probability $\approx 2\sigma$ the favoured allele will become *fixed* in the population.

Selective sweeps

The Moran model with selection.
A population of N genes occurring in two alleles, b and B, evolves in overlapping generations. At exponential rate $\binom{N}{2}$ a pair of genes is sampled (with replacement) from the population, one dies and the other splits in two. If the two genes are of different allelic types, then with probability $\frac{1+\sigma}{2}$ it is the B allele that reproduces.

Suppose that an allele conferring selective advantage σ arises (by mutation say) in an otherwise neutral population. With probability $\approx 2\sigma$ the favoured allele will become *fixed* in the population.
We then say that a *selective sweep* has occurred. It takes $\mathcal{O}(\log N)$ generations to complete.

Selective sweeps

The Moran model with selection.
A population of N genes occurring in two alleles, b and B, evolves in overlapping generations. At exponential rate $\binom{N}{2}$ a pair of genes is sampled (with replacement) from the population, one dies and the other splits in two. If the two genes are of different allelic types, then with probability $\frac{1+\sigma}{2}$ it is the B allele that reproduces.

Suppose that an allele conferring selective advantage σ arises (by mutation say) in an otherwise neutral population. With probability $\approx 2\sigma$ the favoured allele will become *fixed* in the population.
We then say that a *selective sweep* has occurred. It takes $\mathcal{O}(\log N)$ generations to complete.

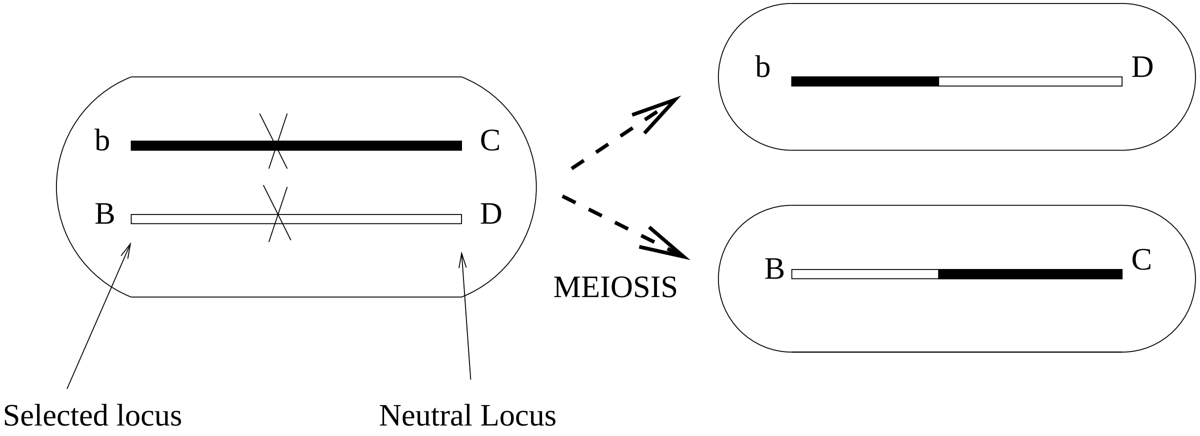
How can we detect selective sweeps?

Genetic Hitchhiking

The problem

Selection acts on a single locus. Alleles B and b.

Linked to a second neutral locus with *recombination* rate r.



What can we say about the family sizes in a sample from the neutral locus at the moment of fixation?

Durrett & Schweinsberg's model

In a population of size $2N$, individuals are labelled b and B. At exponential rate $2N$, two individuals are chosen at random (with replacement) from the population.

- If both are the same type, or if the 2nd is type B, then the first dies and the second reproduces,
- If the 1st is type B and the 2nd is type b, with probability $1 - s$ the 1st dies and the 2nd reproduces, otherwise nothing happens.

Durrett & Schweinsberg’s model

In a population of size $2N$, individuals are labelled b and B. At exponential rate $2N$, two individuals are chosen at random (with replacement) from the population.

- If both are the same type, or if the 2nd is type B, then the first dies and the second reproduces,
- If the 1st is type B and the 2nd is type b, with probability $1 - s$ the 1st dies and the 2nd reproduces, otherwise nothing happens.

Each individual has a second label, from a type space of $2N$ elements. When a new particle is born, it inherits its second label from its parent with probability $1 - r$, otherwise it inherits this label from an individual chosen at random from the population.

The frequency of B-alleles in the population is governed by

$$\begin{aligned}\mathcal{L}^{(N)}f(p) = (2N)^2\Big\{p(1-p)\left(f(p+\frac{1}{2N}) - f(p)\right)\\ + p(1-p)(1-s)\left(f(p-\frac{1}{2N}) - f(p)\right)\Big\},\end{aligned}$$

The frequency of B-alleles in the population is governed by

$$\begin{aligned}\mathcal{L}^{(N)} f(p) = (2N)^2 \Big\{ p(1-p) \left(f(p + \frac{1}{2N}) - f(p) \right) \\ + p(1-p)(1-s) \left(f(p - \frac{1}{2N}) - f(p) \right) \Big\},\end{aligned}$$

Recombination probability $r = \mathcal{O}(1/\log N)$.

The frequency of B-alleles in the population is governed by

$$\mathcal{L}^{(N)} f(p) = (2N)^2 \Big\{ p(1-p) \left(f(p + \frac{1}{2N}) - f(p) \right) + p(1-p)(1-s) \left(f(p - \frac{1}{2N}) - f(p) \right) \Big\},$$

Recombination probability $r = \mathcal{O}(1/\log N)$.

Durrett and Schweinsberg approximate ancestral sample distribution at neutral locus up to error $\mathcal{O}(1/(\log N)^2)$ in probability.

A large population limit

Measure time in units of size $2N$ and set $\alpha = 2Ns$, then

$$\begin{aligned}
\mathcal{L}^{(N)} f(p) &= (2N)^2 \Big\{ p(1-p) \left(\frac{1}{2N} f'(p) + \frac{1}{2} \frac{1}{(2N)^2} f''(p) \right) \\
&\qquad + p(1-p)(1-s) \left(\frac{-1}{2N} f'(p) + \frac{1}{2} \frac{1}{(2N)^2} f''(p) \right) \Big\} + \mathcal{O}(\\
&= 2Nsp(1-p)f'(p) + \frac{2-s}{2} f''(p) + \mathcal{O}(\frac{1}{N}) \\
&= \alpha p(1-p)f'(p) + p(1-p)f''(p) + \mathcal{O}(\frac{1}{N}).
\end{aligned}$$

A large population limit

Measure time in units of size $2N$ and set $\alpha = 2Ns$, then

$$\begin{aligned}
\mathcal{L}^{(N)} f(p) &= (2N)^2 \Big\{ p(1-p) \left(\frac{1}{2N} f'(p) + \frac{1}{2} \frac{1}{(2N)^2} f''(p) \right) \\
&\qquad + p(1-p)(1-s) \left(\frac{-1}{2N} f'(p) + \frac{1}{2} \frac{1}{(2N)^2} f''(p) \right) \Big\} + \mathcal{O}(\\
&= 2Nsp(1-p) f'(p) + \frac{2-s}{2} f''(p) + \mathcal{O}(\frac{1}{N}) \\
&= \alpha p(1-p) f'(p) + p(1-p) f''(p) + \mathcal{O}(\frac{1}{N}).
\end{aligned}$$

$$dp = \alpha p(1-p)dt + \sqrt{2p(1-p)}dW.$$

Let $\rho = 2Nr$ and write T for the time of the end of the sweep.

Backwards in time

At time T when take sample all individuals type B.

Tracing backwards in time, at time of recombination event ancestors of neutral and selective loci differ so type at *selected* locus of ancestor at neutral locus *can change*. *Effective* recombination events

$$B \rightsquigarrow b \text{ rate } \rho(1-p_{T-\beta}).$$

$$b \rightsquigarrow B \text{ rate } \rho p_{T-\beta}.$$

Backwards in time

At time T when take sample all individuals type B.
Tracing backwards in time, at time of recombination event ancestors of neutral and selective loci differ so type at *selected* locus of ancestor at neutral locus *can change*. *Effective* recombination events

$$B \rightsquigarrow b \text{ rate } \rho(1-p_{T-\beta}).$$
$$b \rightsquigarrow B \text{ rate } \rho p_{T-\beta}.$$

No mutation so two lineages can result from a common parent only if they have the same type at the selected locus.

$$\text{Two lineages in } B \text{ at time } T-\beta \text{ coalesce at rate } \frac{2}{p_{T-\beta}}.$$
$$\text{Two lineages in } b \text{ at time } T-\beta \text{ coalesce at rate } \frac{2}{(1-p_{T-\beta})}.$$

A structured coalescent

Structured coalescent in background p:

Given the path, $\{p_t\}_{0\leq t\leq T}$, of the sweep, the *structured coalescent in background* p is the system of coalescing lineages in which lineages migrate from background B to b at instantaneous rate $\rho(1-p_{T-\beta})$ and from b to B at instantaneous rate $\rho p_{T-\beta}$. Moreover, any pair of lineages in background B at time β coalesce at instantaneous rate $\frac{2}{p_{T-\beta}}$ and any pair of lineages in background b coalesce at instantaneous rate $\frac{2}{(1-p_{T-\beta})}$.

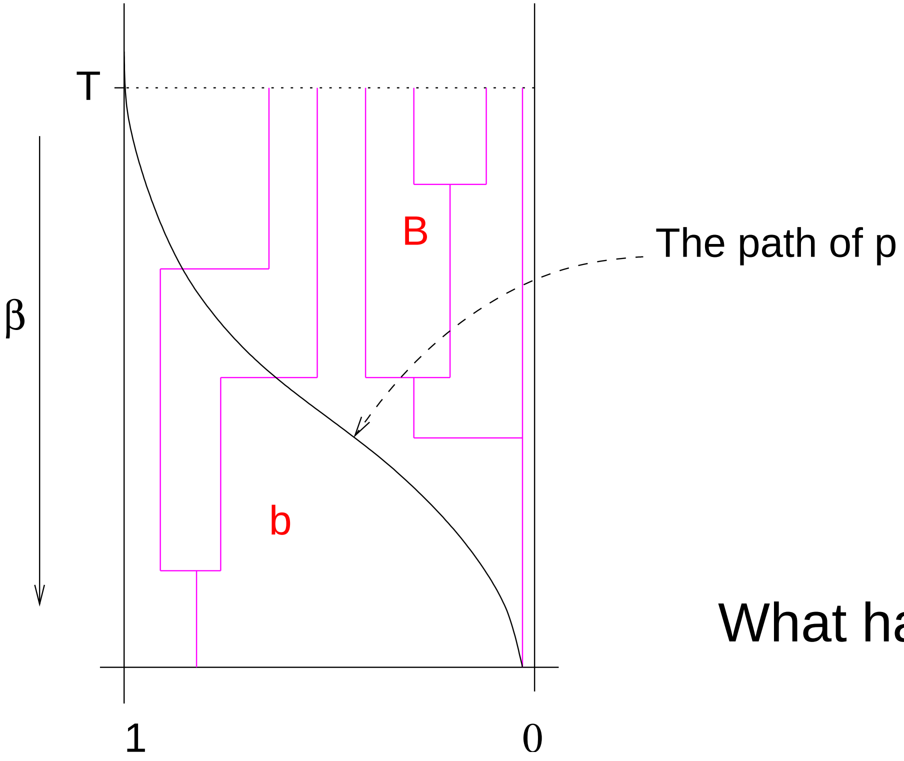
A structured coalescent

Structured coalescent in background p:
Given the path, $\{p_t\}_{0\leq t\leq T}$, of the sweep, the *structured coalescent in background* p is the system of coalescing lineages in which lineages migrate from background B to b at instantaneous rate $\rho(1-p_{T-\beta})$ and from b to B at instantaneous rate $\rho p_{T-\beta}$. Moreover, any pair of lineages in background B at time β coalesce at instantaneous rate $\frac{2}{p_{T-\beta}}$ and any pair of lineages in background b coalesce at instantaneous rate $\frac{2}{(1-p_{T-\beta})}$.

Given that a sweep takes place,

$$dp = \alpha p(1-p)\coth(\frac{\alpha}{2}p)dt + \sqrt{2p(1-p)}dW.$$

Genealogy at the *neutral* locus: structured n-coalescent in background p.

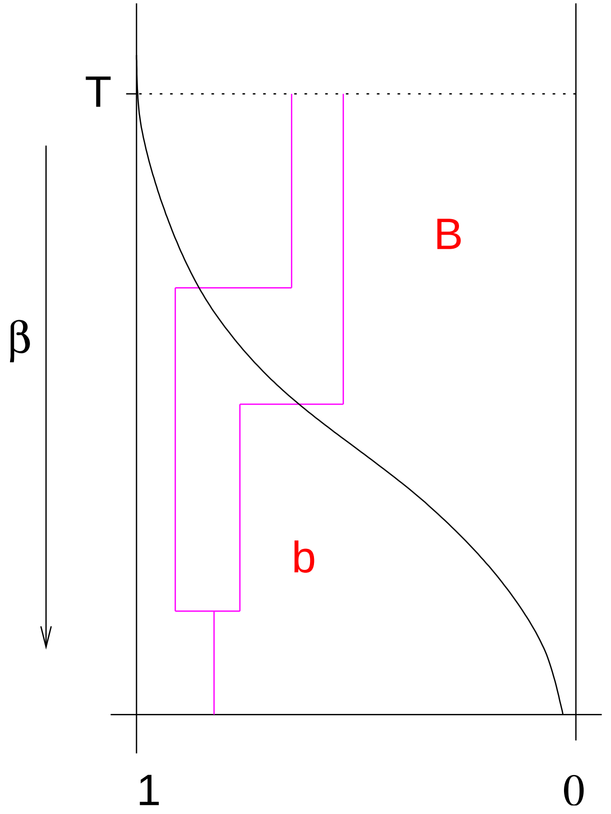


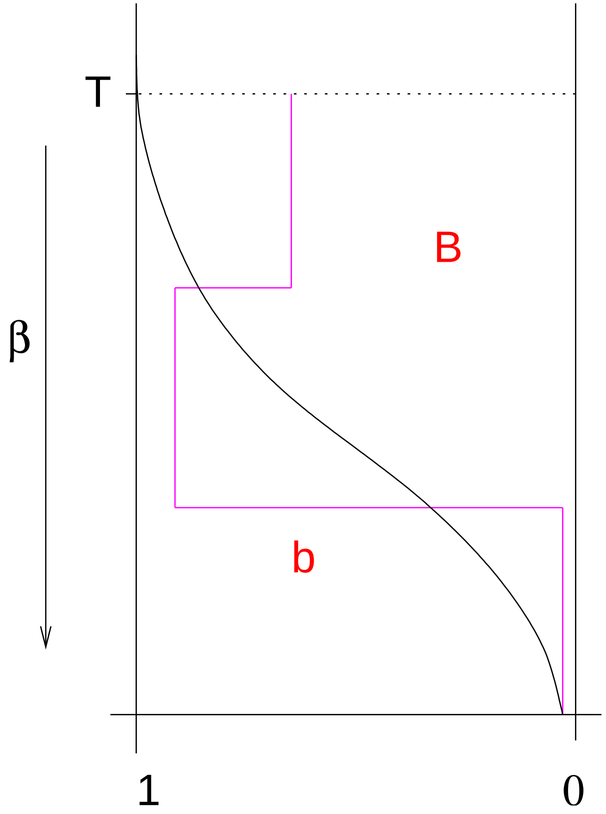
What happens as $\alpha \to \infty$?

$$\rho = 2Nr = \gamma \frac{\alpha}{\log(\alpha)}.$$

Approximation step 1

- From the structured to a marked coalescent





Each have probability of $\mathcal{O}\left(1/(\log\alpha)^2\right)$.

Approximation step 2

- From the marked coalescent to a marked Yule tree

Recall that

$$dp = \alpha p(1-p)\coth(\frac{\alpha}{2}p)dt + \sqrt{2p(1-p)}dW_t.$$

Approximation step 2

- From the marked coalescent to a marked Yule tree

Recall that

$$dp = \alpha p(1-p)\coth(\frac{\alpha}{2}p)dt + \sqrt{2p(1-p)}dW_t.$$

Time change $t \mapsto \tau$ given by $d\tau = (1-p_t)dt$. Then $p \rightsquigarrow Z$

$$dZ = \alpha Z \coth(\frac{\alpha}{2}Z)d\tau + \sqrt{2Z}d\tilde{W}_\tau.$$

Feller diffusion conditioned on non-extinction.

Approximation step 2

- From the marked coalescent to a marked Yule tree

Recall that

$$dp = \alpha p(1-p)\coth(\frac{\alpha}{2}p)dt + \sqrt{2p(1-p)}dW_t.$$

Time change $t \mapsto \tau$ given by $d\tau = (1-p_t)dt$. Then $p \rightsquigarrow Z$

$$dZ = \alpha Z \coth(\frac{\alpha}{2}Z)d\tau + \sqrt{2Z}d\tilde{W}_\tau.$$

Feller diffusion conditioned on non-extinction.

Marking rate $\rightsquigarrow \rho d\tau$. Coalescence rate $\rightsquigarrow \frac{2}{Z(1-Z)} \approx \frac{2}{Z}$.

Approximation step 3

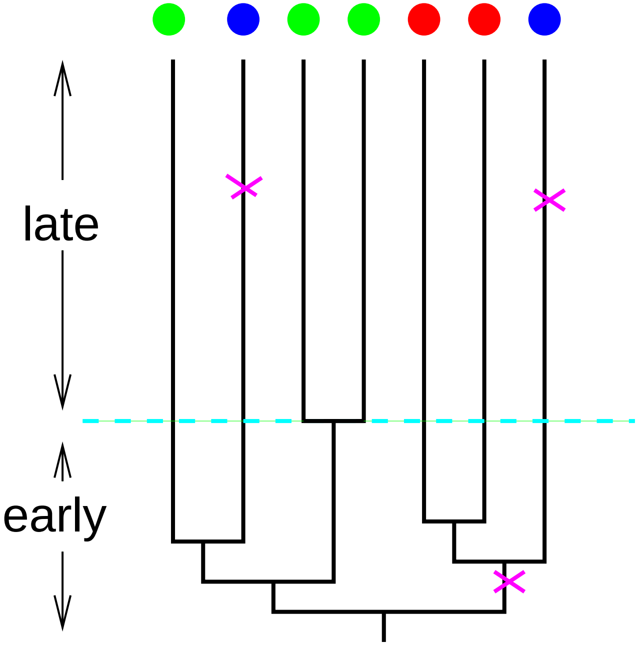
- Approximating sample partitions in marked Yule trees.

Coalescent (approximately) genealogy of a sample from a Yule tree with *constant* rate of marking.

One can construct this process *forwards* in time.

Many exact calculations are possible.

Phases of the sweep

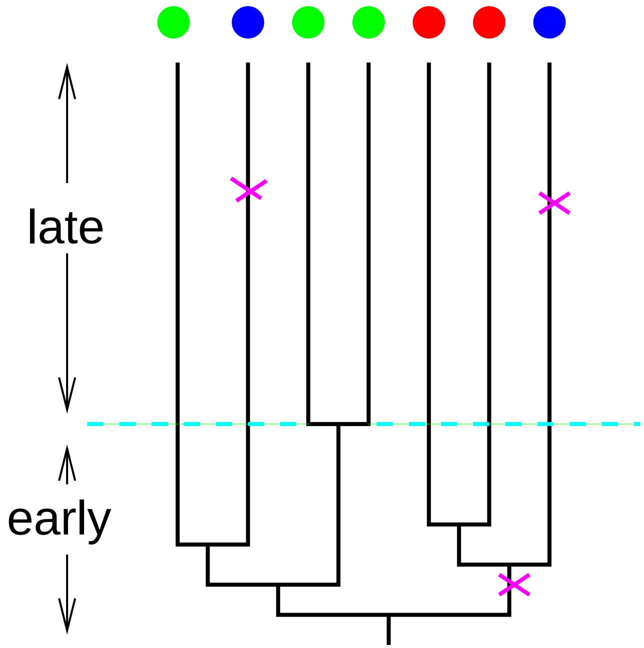


🟢 Non−recombinant

🔵 Late recombinant

🔴 Early recombinant

Phases of the sweep



🟢 Non−recombinant

🔵 Late recombinant

🔴 Early recombinant

Up to an error $\mathcal{O}(1/(\log\alpha)^2)$, we will see at most *one* early recombinant family.

Main result

Theorem

Fix n. For a selection coefficient $\alpha \gg 1$ and a recombination rate $\rho = \gamma \frac{\alpha}{\log \alpha}$, the ancestral partition of an n-sample drawn at time T consists, up to an error in probability of order $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$, of

- L late recombinant singletons
- *one* family of early recombinants of size E
- one non-recombinant family of size $n - L - E$.

Let F be an $\mathbb{N}$-valued random variable with

$$\mathbf{P}[F \leq i] = \frac{(i-(n-1))\cdots(i-1)}{(i+(n-1))\cdots(i+1)},$$

Let F be an $\mathbb{N}$-valued random variable with

$$\mathbf{P}[F \le i] = \frac{(i-(n-1))\cdots(i-1)}{(i+(n-1))\cdots(i+1)},$$

F is the number of individuals alive in the full Yule tree when the early phase ends.

Let F be an $\mathbb{N}$-valued random variable with

$$\mathbf{P}[F \leq i] = \frac{(i-(n-1))\cdots(i-1)}{(i+(n-1))\cdots(i+1)},$$

F is the number of individuals alive in the full Yule tree when the early phase ends.

Given $F = f$, let L be a binomial random variable with n trials and success probability $1 - p_f$, where

$$p_f = \exp\left(-\frac{\gamma}{\log \alpha}\sum_{i=f}^{\alpha}\frac{1}{i}\right).$$

Let F be an $\mathbb{N}$-valued random variable with

$$\mathbf{P}[F \leq i] = \frac{(i-(n-1))\cdots(i-1)}{(i+(n-1))\cdots(i+1)},$$

F is the number of individuals alive in the full Yule tree when the early phase ends.

Given $F = f$, let L be a binomial random variable with n trials and success probability $1 - p_f$, where

$$p_f = \exp\left(-\frac{\gamma}{\log \alpha}\sum_{i=f}^{\alpha}\frac{1}{i}\right).$$

This gives us the number of late recombinants.

Independently of all this, let S be a $\{0, 1, ..., n\}$-valued random variable with

$$\mathbf{P}[S = s] = \begin{cases} \frac{\gamma n}{\log \alpha} \sum_{i=2}^{n-1} \frac{1}{i}, & s = 1, \\ \frac{\gamma n}{\log \alpha} \frac{1}{s(s-1)}, & 2 \leq s \leq n-1 \\ \frac{\gamma n}{\log \alpha} \frac{1}{n-1}, & s = n. \end{cases}$$

Independently of all this, let S be a $\{0, 1, ..., n\}$-valued random variable with

$$\mathbf{P}[S = s] = \begin{cases} \frac{\gamma n}{\log \alpha} \sum_{i=2}^{n-1} \frac{1}{i}, & s = 1, \\ \frac{\gamma n}{\log \alpha} \frac{1}{s(s-1)}, & 2 \leq s \leq n-1 \\ \frac{\gamma n}{\log \alpha} \frac{1}{n-1}, & s = n. \end{cases}$$

S is the number of early recombinants at the end of the early phase.

Independently of all this, let S be a $\{0, 1, ..., n\}$-valued random variable with

$$\mathbf{P}[S = s] = \begin{cases} \frac{\gamma n}{\log \alpha} \sum_{i=2}^{n-1} \frac{1}{i}, & s = 1, \\ \frac{\gamma n}{\log \alpha} \frac{1}{s(s-1)}, & 2 \leq s \leq n-1 \\ \frac{\gamma n}{\log \alpha} \frac{1}{n-1}, & s = n. \end{cases}$$

S is the number of early recombinants at the end of the early phase.
Given $S = s$ and $L = l$, the random variable E is hypergeometric,

$$\mathbf{P}[E = e] = \frac{\binom{s}{e}\binom{n-s}{n-l-e}}{\binom{n}{n-l}}.$$

Independently of all this, let S be a $\{0, 1, ..., n\}$-valued random variable with

$$\mathbf{P}[S = s] = \begin{cases} \frac{\gamma n}{\log \alpha} \sum_{i=2}^{n-1} \frac{1}{i}, & s = 1, \\ \frac{\gamma n}{\log \alpha} \frac{1}{s(s-1)}, & 2 \leq s \leq n-1 \\ \frac{\gamma n}{\log \alpha} \frac{1}{n-1}, & s = n. \end{cases}$$

S is the number of early recombinants at the end of the early phase.

Given $S = s$ and $L = l$, the random variable E is hypergeometric,

$$\mathbf{P}[E = e] = \frac{\binom{s}{e}\binom{n-s}{n-l-e}}{\binom{n}{n-l}}.$$

This provides the ‘thinning’ of S to give the number of early recombinants at the end of the sweep.

Numerical results

We distinguish the number and types of ancestors of the sample at the beginning of the sweep.

$n = 1$

$$\text{pinb} \approx \mathbf{P}[L = 1].$$

$n = 2$

Two ancestors: ‘p2inb’, ‘p1B1b’

One ancestor: ‘p2cinb’ or a B allele.

$$\text{p2inb} \approx \mathbf{P}[L = 2 \text{ or } S = 2, L = 1],$$
$$\text{p2cinb} \approx \mathbf{P}[L = 0, S = 2],$$
$$\text{p1B1b} \approx \mathbf{P}[L = 1, S = 0].$$

	pinb	p2inb	p2cinb	p1B1b
	$N = 10^4$	$s = 0.1$	r=0.001064	
Moran	0.08203	0.00620	0.01826	0.11513
Logistic	0.09983(21%)	0.00845(36%)	0.03365(84%)	0.11544(0.3%)
SD03	0.08235(0.4%)	0.00627(1.1%)	0.01765(-3.4%)	0.11687(1.5%)
EPW05	0.0822(0.2%)	0.00659(6.3%)	0.01867(2.2%)	0.11515(0.0%)
	$N = 10^4$	$s = 0.1$	r=0.005158	
Moran	0.33656	0.10567	0.05488	0.35201
Logistic	0.39936(18%)	0.13814(31%)	0.09599(75%)	0.32646(-7.3%)
SD03	0.34065(1.2%)	0.10911(3.2%)	0.05100(-7.1%)	0.36112(2.6%)
EPW05	0.32973(-2.0%)	0.10857(2.7%)	0.05662(3.2%)	0.34157(-0.3%)

A new kind of data

Beginning to see data that documents genetic variation at genomic scales.

Can we identify the locations of selective sweeps by looking for long blocks of shared material?

Need to understand ‘false positives’.

- Rare neutral trees
- Bottlenecks
- Spatial subdivision

Examine the way in which diversity recovers as we move away from the shared block.

Balancing selection

Directional selection drives one allele to fixation/extinction. Other forms of selection can work to maintain alleles at non-trivial frequencies.

Example: selection in favour of heterozygosity.

Balancing selection

Directional selection drives one allele to fixation/extinction. Other forms of selection can work to maintain alleles at non-trivial frequencies.

Example: selection in favour of heterozygosity.

Parental type	PP	PQ	QQ
Relative fitness	$1-\tilde{\sigma}$	$1+\tilde{\sigma}$	$1-\tilde{\sigma}$

Balancing selection

Directional selection drives one allele to fixation/extinction. Other forms of selection can work to maintain alleles at non-trivial frequencies.

Example: selection in favour of heterozygosity.

Parental type	PP	PQ	QQ
Relative fitness	$1-\tilde{\sigma}$	$1+\tilde{\sigma}$	$1-\tilde{\sigma}$

Diploid population of size N. Model the corresponding $2N$ genomes as haploid.

The Moran model

At exponential rate $\binom{2N}{2}$ a pair of individuals is chosen at random from the population. One dies, the other reproduces.

If the pair chosen consists of one type P and one type Q, then with probability $(1+\sigma)/2$ it is the type P individual that reproduces.

The Moran model

At exponential rate $\binom{2N}{2}$ a pair of individuals is chosen at random from the population. One dies, the other reproduces.
If the pair chosen consists of one type P and one type Q, then with probability $(1+\sigma)/2$ it is the type P individual that reproduces.

Note: σ may depend on the current frequency of P-alleles.

The Moran model

At exponential rate $\binom{2N}{2}$ a pair of individuals is chosen at random from the population. One dies, the other reproduces.

If the pair chosen consists of one type P and one type Q, then with probability $(1+\sigma)/2$ it is the type P individual that reproduces.

Note: σ may depend on the current frequency of P-alleles.

Mutation:

		Offspring	
		P	Q
Parent	P	$1-\overline{\mu}_1$	$\overline{\mu}_1$
	Q	$\overline{\mu}_2$	$1-\overline{\mu}_2$

Transition rates

		P death P birth	P death Q birth	Q death P Birth	Q death Q birth
Parents	PP	$1-\bar{\mu}_1$	$\bar{\mu}_1$	0	0
	PQ	$\frac{(1-\sigma)}{2}\bar{\mu}_2$	$\frac{(1-\sigma)}{2}(1-\bar{\mu}_2)$	$\frac{(1+\sigma)}{2}(1-\bar{\mu}_1)$	$\frac{(1+\sigma)}{2}\bar{\mu}_1$
	QQ	0	0	$\bar{\mu}_2$	$1-\bar{\mu}_2$

What is σ for balancing selection?

What is σ for balancing selection?

$$\begin{aligned}(1+\tilde{\sigma})(1-p)+(1-\tilde{\sigma})p : (1+\tilde{\sigma})p+(1-\tilde{\sigma})(1-p)\\ = 1+\tilde{\sigma}(1-2p) : 1-\tilde{\sigma}(1-2p).\end{aligned}$$

What is σ for balancing selection?

$$\begin{aligned}(1+\tilde{\sigma})(1-p)+(1-\tilde{\sigma})p : (1+\tilde{\sigma})p+(1-\tilde{\sigma})(1-p)\\ = 1+\tilde{\sigma}(1-2p) : 1-\tilde{\sigma}(1-2p).\end{aligned}$$

Take $\sigma = 2\tilde{\sigma}(\frac{1}{2}-p)$ for some $\tilde{\sigma}$.

What is σ for balancing selection?

$$(1+\tilde{\sigma})(1-p)+(1-\tilde{\sigma})p:(1+\tilde{\sigma})p+(1-\tilde{\sigma})(1-p)$$
$$=1+\tilde{\sigma}(1-2p):1-\tilde{\sigma}(1-2p).$$

Take $\sigma=2\tilde{\sigma}(\frac{1}{2}-p)$ for some $\tilde{\sigma}$.

Weak selection limit.

$$\sigma=\frac{s}{N},\quad \overline{\mu}_i=\frac{\mu_i}{N}.$$

What is σ for balancing selection?

$$(1+\tilde{\sigma})(1-p)+(1-\tilde{\sigma})p:(1+\tilde{\sigma})p+(1-\tilde{\sigma})(1-p)$$
$$=1+\tilde{\sigma}(1-2p):1-\tilde{\sigma}(1-2p).$$

Take $\sigma=2\tilde{\sigma}(\frac{1}{2}-p)$ for some $\tilde{\sigma}$.

Weak selection limit.

$$\sigma=\frac{s}{N},\quad \overline{\mu}_i=\frac{\mu_i}{N}.$$

Write $p(t)$ for the proportion of P alleles in population at time t and $\mathcal{L}^{(N)}$ for the generator of the rescaled Moran model.

Weak selection limit

Lemma

For any smooth function $f : [0,1] \to \mathbb{R}$,

$$\begin{aligned}\mathcal{L}^{(N)} f(p) = \left(2s(p)p(1-p) - \mu_1 p + \mu_2(1-p)\right) f'(p) \\ + \frac{1}{2}p(1-p)f''(p) + \mathcal{O}(\frac{1}{N}).\end{aligned}$$

Weak selection limit

Lemma

For any smooth function $f : [0,1] \to \mathbb{R}$,

$$\mathcal{L}^{(N)} f(p) = \left(2s(p)p(1-p) - \mu_1 p + \mu_2(1-p)\right) f'(p) + \frac{1}{2}p(1-p)f''(p) + \mathcal{O}(\frac{1}{N}).$$

Proof.

$$\mathcal{L}^{(N)} f(p)$$

$$= N(2N-1)\Big\{2p(1-p)\frac{(1+\sigma)}{2}(1-\bar{\mu}_1)+(1-p)^2\bar{\mu}_2\Big\}\left(f(p+\frac{1}{2N}) - f(p)\right)$$

$$+N(2N-1)\Big\{2p(1-p)\frac{(1-\sigma)}{2}(1-\bar{\mu}_2)+p^2\bar{\mu}_1\Big\}\left(f(p-\frac{1}{2N}) - f(p)\right).$$

Weak selection limit

Lemma

For any smooth function $f : [0,1] \to \mathbb{R}$,

$$\mathcal{L}^{(N)} f(p) = (2s(p)p(1-p) - \mu_1 p + \mu_2(1-p)) f'(p) + \frac{1}{2}p(1-p)f''(p) + \mathcal{O}(\frac{1}{N}).$$

Proof.

$$\mathcal{L}^{(N)} f(p)$$

$$= N(2N-1)\Big\{2p(1-p)\frac{(1+\sigma)}{2}(1-\bar{\mu}_1)+(1-p)^2\bar{\mu}_2\Big\}\left(f(p+\frac{1}{2N}) - f(p)\right)$$

$$+N(2N-1)\Big\{2p(1-p)\frac{(1-\sigma)}{2}(1-\bar{\mu}_2)+p^2\bar{\mu}_1\Big\}\left(f(p-\frac{1}{2N}) - f(p)\right).$$

Substitute for σ and $\bar{\mu}_i$ and expand f in a Taylor series about p. □

In the weak selection limit, the frequency of P-alleles follows

$$dp_t = \Big\{ s_0 p_t (1-p_t)(1-2p_t) \\ - \mu_1 p_t + \mu_2 (1-p_t) \Big\} dt + \sqrt{p_t(1-p_t)}\, dW_t,$$

where $\{W_t\}_{t \geq 0}$ is standard Brownian motion and s_0 is a constant.

The problem

Selection acts on a single locus. Alleles P and Q.

The problem

Selection acts on a single locus. Alleles P and Q.

Strictly positive mutation rates $P \leftrightarrow Q$.

The problem

Selection acts on a single locus. Alleles P and Q.

Strictly positive mutation rates $P \leftrightarrow Q$.

Linked to a second neutral locus with *recombination* rate r.

The problem

Selection acts on a single locus. Alleles P and Q.
Strictly positive mutation rates $P \leftrightarrow Q$.
Linked to a second neutral locus with *recombination* rate r.
The neutral locus is embedded in a fluctuating genetic background.
Migration due to mutation and recombination.

What can we say about the genealogy of a sample from the neutral locus?

Some assumptions

At the neutral locus, assume mutation to a novel type at rate ν.

Some assumptions

At the neutral locus, assume mutation to a novel type at rate ν.

Assume that frequency of P-alleles has reached stationarity.

Some assumptions

At the neutral locus, assume mutation to a novel type at rate ν.

Assume that frequency of P-alleles has reached stationarity.

Let $n_t = (n_1(t), n_2(t))$ where $n_1(t)$ is the number of ancestors of our sample in background P at time t *before* the present, $n_2(t)$ is the number in background Q.

Some assumptions

At the neutral locus, assume mutation to a novel type at rate ν.

Assume that frequency of P-alleles has reached stationarity.

Let $n_t = (n_1(t), n_2(t))$ where $n_1(t)$ is the number of ancestors of our sample in background P at time t *before* the present, $n_2(t)$ is the number in background Q.

Writing p_t for the frequency of P-alleles at time t *before* the present, can write down the generator of (p_t, n_t).

A weak selection limit

The model is too special.

Pass to a diffusion approximation:

$$\overline{\mu}_i = \frac{\mu_i}{N}, \quad \overline{r} = \frac{r}{N}, \quad \overline{s} = \frac{s}{N}, \quad \overline{\nu} = \frac{\nu}{N}.$$

A weak selection limit

The model is too special.

Pass to a diffusion approximation:

$$\overline{\mu}_i = \frac{\mu_i}{N}, \quad \overline{r} = \frac{r}{N}, \quad \overline{s} = \frac{s}{N}, \quad \overline{\nu} = \frac{\nu}{N}.$$

Let $E = [0,1] \times \{1, \ldots, n_1(0) + n_2(0)\}^2$ and suppose that $f(p, n_1, n_2) : E \to \mathbb{R}$ is C^2 as a function of p.

$$
\begin{aligned}
Af &= \frac{1}{p}\binom{n_1}{2}\big(f(p,n_1-1,n_2)-f(p,n_1,n_2)\big)\\
&+ \frac{1}{q}\binom{n_2}{2}\big(f(p,n_1,n_2-1)-f(p,n_1,n_2)\big)\\
&+ \frac{p}{q}\mu_1 n_2\big(f(p,n_1+1,n_2-1)-f(p,n_1,n_2)\big)\\
&+ \frac{q}{p}\mu_2 n_1\big(f(p,n_1-1,n_2+1)-f(p,n_1,n_2)\big)\\
&+ r n_2 p\big(f(p,n_1+1,n_2-1)-f(p,n_1,n_2)\big)\\
&+ r n_1 q\big(f(p,n_1-1,n_2+1)-f(p,n_1,n_2)\big)\\
&+ (-\mu_1 p+\mu_2 q+spq)\, f' + \frac{1}{2}pqf''
\end{aligned}
$$

where $q = 1 - p$ and $'$ denotes differentiation with respect to p.

Remarks

No surprises in the *form* of the generator.

Remarks

No surprises in the *form* of the generator.

Convergence provided s Lipschitz and $\mu_i > 0$.

Remarks

No surprises in the *form* of the generator.

Convergence provided s Lipschitz and $\mu_i > 0$.

Structured coalescent with rates driven by a diffusion process.

Remarks

No surprises in the *form* of the generator.

Convergence provided s Lipschitz and $\mu_i > 0$.

Structured coalescent with rates driven by a diffusion process.

Crucially, if τ is first hitting time of zero by the diffusion p then $\int^{\tau} \frac{1}{p(s)} ds$ diverges. (Similar statement at $p = 1$).

Coalescence times

Let $F_{PP}(t, p)$ be the probability that the two lineages ancestral to our sample have coalesced by time t if both individuals in the sample are originally taken from the P background. Similarly define $F_{PQ}(t, p)$ and $F_{QQ}(t, p)$.

Coalescence times

Let $F_{PP}(t, p)$ be the probability that the two lineages ancestral to our sample have coalesced by time t if both individuals in the sample are originally taken from the P background. Similarly define $F_{PQ}(t, p)$ and $F_{QQ}(t, p)$.

Given that $\{p(0)\}_{t \geq 0}$ is drawn from the (reversible) stationary distribution for the process $\{p(t)\}_{t \geq 0}$, $\{F_{PP}(t, p), F_{PQ}(t, p), F_{QQ}(t, p)\}$ can be characterised as the unique *minimal* solution to the following system of differential equations subject to $F'_{PP}(t, 1) = 0$, $F'_{QQ}(t, 0) = 0$ and $F_{PP}(0, p) = F_{PQ}(0, p) = F_{QQ}(0, p) = 0$.

$$
\begin{aligned}
\dot{F}_{PP} &= \frac{1-F_{PP}}{p} + 2\left(\frac{\mu_2 q}{p} + rq\right)(F_{PQ} - F_{PP}) \\
&\quad + (-\mu_1 p + \mu_2 q + spq)\, F'_{PP} + \frac{1}{2} pq F''_{PP} \\
\dot{F}_{PQ} &= \left(\frac{p\mu_1}{q} + rp\right)(F_{PP} - F_{PQ}) \\
&\quad + \left(\frac{q\mu_2}{p} + rq\right)(F_{QQ} - F_{PQ}) \\
&\quad + (-\mu_1 p + \mu_2 q + spq)\, F'_{PQ} + \frac{1}{2} pq F''_{PQ} \\
\dot{F}_{QQ} &= \frac{1-F_{QQ}}{q} + 2\left(\frac{\mu_1 p}{q} + rp\right)(F_{PQ} - F_{QQ}) \\
&\quad + (-\mu_1 p + \mu_2 q + spq)\, F'_{QQ} + \frac{1}{2} pq F''_{QQ}.
\end{aligned}
$$

Probability of identity

At neutral locus, mutate to novel state at rate ν.

Probability of identity

At neutral locus, mutate to novel state at rate ν.

$f_{PP}(p) = \mathbb{P}[$sample size 2 from P-background identical in state$]$ etc.

Probability of identity

At neutral locus, mutate to novel state at rate ν.

$f_{PP}(p) = \mathbb{P}[$sample size 2 from P-background identical in state$]$ etc.

Integration by parts $\rightsquigarrow$

$$\begin{aligned}
0 = \quad & -2\nu f_{PP} + \frac{1-f_{PP}}{p} + 2\left(\frac{\mu_2 q}{p} + rq\right)(f_{PQ} - f_{PP}) \\
& \quad + (-\mu_1 p + \mu_2 q + spq)\, f'_{PP} + \tfrac{1}{2} pq f''_{PP} \\
0 = \quad & -2\nu f_{PQ} + \left(\frac{p\mu_1}{q} + rp\right)(f_{PP} - f_{PQ}) + \left(\frac{q\mu_2}{p} + rq\right)(f_{QQ} - f_{PQ}) \\
& \quad + (-\mu_1 p + \mu_2 q + spq)\, f'_{PQ} + \tfrac{1}{2} pq f''_{PQ} \\
0 = \quad & -2\nu f_{QQ} + \frac{1-f_{QQ}}{q} + 2\left(\frac{\mu_1 p}{q} + rp\right)(f_{PQ} - f_{QQ}) \\
& \quad + (-\mu_1 p + \mu_2 q + spq)\, f'_{QQ} + \tfrac{1}{2} pq f''_{QQ}.
\end{aligned}$$

(*minimal* solution)

A numerical example

Calculate the probability of identity of a sample of size two and thus the expected time to the most recent common ancestor, E.

A numerical example

Calculate the probability of identity of a sample of size two and thus the expected time to the most recent common ancestor, E.

Typically we *don't know* the frequency at the selected site, nor even the type. Treat it as a random sample.

A numerical example

Calculate the probability of identity of a sample of size two and thus the expected time to the most recent common ancestor, E.

Typically we *don't know* the frequency at the selected site, nor even the type. Treat it as a random sample. The graph shows $p^2 E_{PP} + 2pqE_{PQ} + q^2 E_{QQ}$ averaged over the stationary distribution of p_t. In this example,

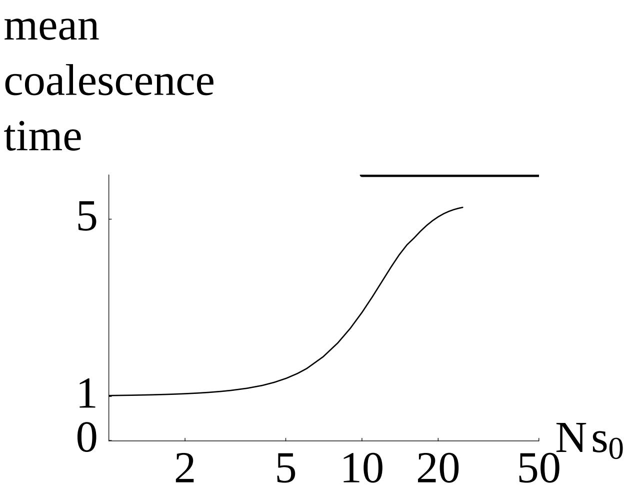
$dp = s_0 p(1-p)(1-2p)dt + \mu(1-2p)dt + \sqrt{p(1-p)}dW$.

A numerical example

Calculate the probability of identity of a sample of size two and thus the expected time to the most recent common ancestor, E.

Typically we *don't know* the frequency at the selected site, nor even the type. Treat it as a random sample. The graph shows $p^2 E_{PP} + 2pqE_{PQ} + q^2 E_{QQ}$ averaged over the stationary distribution of p_t. In this example,

$$dp = s_0 p(1-p)(1-2p)dt + \mu(1-2p)dt + \sqrt{p(1-p)}dW.$$



Pause for thought

- Fluctuations matter.

Pause for thought

- Fluctuations matter.

- The parameters in our diffusion approximation correspond to $N\overline{\mu}_i$, $N\overline{s}$ etc.

Pause for thought

- Fluctuations matter.
- The parameters in our diffusion approximation correspond to $N\overline{\mu}_i$, $N\overline{s}$ etc.
- Biological populations are *finite*. In particular $\log N \sim 10$.

Pause for thought

- Fluctuations matter.
- The parameters in our diffusion approximation correspond to $N\overline{\mu}_i$, $N\overline{s}$ etc.
- Biological populations are *finite*. In particular $\log N \sim 10$.
- Quoted numbers often for the *effective* population size.