

## Λ Coalescents: Theory and Applications

Rick Durrett

## Moran model

- Each individual is replaced at rate 1. That is, individual  $x$  lives for an exponentially distributed amount with mean 1 and then is “replaced.”
- To replace individual  $x$ , we choose an individual at random from the population (including  $x$  itself) to be the parent of the new individual.

Suppose that we have two alleles  $A$  and  $a$ , and let  $X_t$  be the number of copies of  $A$ . The transition rates for  $X_t$  are

$$\begin{aligned}i \rightarrow i + 1 & \quad \text{at rate } b_i = (2N - i) \cdot \frac{i}{2N} \\i \rightarrow i - 1 & \quad \text{at rate } d_i = i \cdot \frac{2N - i}{2N}\end{aligned}$$

## Kingman's coalescent

**Theorem** When time is run at rate  $N$ , the genealogy of a sample of size  $n$  from the Moran model converges to Kingman's coalescent.

**Proof.** If we look backwards in time, then when there are  $k$  lineages, each replacement leads to a coalescence with probability  $(k - 1)/2N$ . If we run time at rate  $N$ , then jumps occur at rate  $N \cdot k/2N = k/2$ , so the total rate of coalescence is  $k(k - 1)/2$ , the right rate for Kingman's coalescent.

## Directional Selection

*Fecundity selection.* Suppose  $b$ 's are born at a rate  $1 - s$  times that of  $B$ 's. The transition rates for  $X_t$  for the number of  $B$ 's is now:

$$\begin{aligned}i \rightarrow i + 1 & \quad \text{at rate } b_i = (2N - i) \cdot \frac{i}{2N} \\i \rightarrow i - 1 & \quad \text{at rate } d_i = i \cdot \frac{2N - i}{2N}(1 - s)\end{aligned}$$

Embedded jump chain is a simple random walk that jumps up with probability  $p = 1/(2 - s)$  and down with probability  $1 - p$ .

Started with  $X_0 = i$ ,  $B$  becomes fixed in the population (reaches  $2N$ ) with probability:

$$\frac{1 - (1 - s)^i}{1 - (1 - s)^{2N}}$$

## Three phases of the fixation process

- 1 While the advantageous  $B$  allele is rare, the number of  $B$ 's can be approximated by a supercritical branching process.
- 2 While the frequency of  $B$ 's is in  $[\epsilon, 1 - \epsilon]$  there is very little randomness and it follows the solution of the logistic differential equation:  $du/dt = su(1 - u)$ .
- 3 While the disadvantageous  $b$  allele is rare, the number of  $a$ 's can be approximated by a subcritical branching process.

## Hitchhiking

Due to recombination, each chromosome you inherit from each parent is a mixture of their two chromosomes, with transitions between the two at points of a nonhomogeneous Poisson process.

In the absence of recombination, fixation of an allele would result in every individual in the population having a copy of the associated chromosome. With recombination, changes in allele frequency occur only near the allele that went to fixation.

## Maynard-Smith and Haigh (1974)

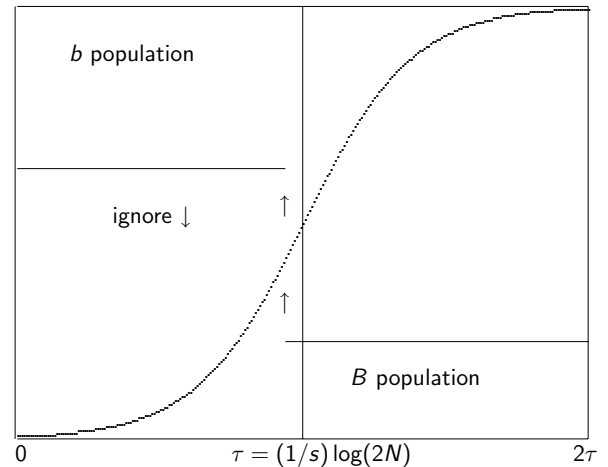
Alleles  $B$  and  $b$  have relative fitnesses 1 and  $1-s$ , neutral locus with alleles  $A$  and  $a$ , recombination between the two has probability  $p$ .

Let  $p_0 =$  frequency of  $B$  before the sweep ( $1/2N$ ).  
 $Q_t = P(A|B)$ .  $R_t = P(A|b)$ .

**Theorem.** Suppose  $Q_0 = 0$ . Under the *logistic sweep model*, which ignores the branching process phases 1 and 3,

$$Q_\infty = R_0(1 - p_0) \int_0^{2\tau} \frac{re^{-rt}}{(1 - p_0) + p_0 e^{st}} ds$$

**Proof.**  $R_0(1 - p_0)$  is the frequency of  $A$  before the sweep. In order for a sampled individual to have the  $A$  allele, its lineage must escape the sweep due to recombination.



## Durrett and Schweinsberg (2004) Theor. Pop. Biol.

From the previous theorem, the probability a lineage escapes from the sweep by recombination is

$$pinb = \int_0^{2\tau} \frac{re^{-rt}}{(1 - p_0) + p_0 e^{st}} ds$$

**Theorem.** Under the *logistic sweep model*, if  $N \rightarrow \infty$  and  $r \log(2N)/s \rightarrow a$ ,  $pinb \rightarrow 1 - e^{-a}$ .

Biologists rule of thumb:

"hitchhiking is efficient if  $r < s$  and negligible if  $r \approx s$ ."  
 (should be efficient if  $r \approx s/(\log(2N))$ )

## Effect on genealogies

**Approximation 1** Let  $p_{k,j} =$  probability  $k$  lineages reduced to  $i$  by the sweep. Under the *logistic sweep model*, if  $N \rightarrow \infty$  with

$$r \ln(2N)/s \rightarrow a \text{ and } s(\ln N)^2 \rightarrow \infty$$

then for  $j \geq 2$

$$p_{k,k-j+1} \rightarrow \binom{k}{j} p^j (1-p)^{k-j} \text{ where } p = e^{-a}$$

and  $p_{k,k} \rightarrow (1-p)^k + kp(1-p)^{k-1}$ .

*p*-merger. Flip coins with probability  $p$  of heads for each lineage and coalesce all of those with heads. Need at least two heads to get a coalescence.

## Simulation results

$N = 10,000$ ,  $s = 0.1$ . Set  $r = 0.00516$  so  $pinb \approx 0.4$ .

$p2inb = P(\text{both lineages escapes the sweep and do not coalesce})$ .

$p2cinb = P(\text{both lineages escapes the sweep but coalesce})$ .

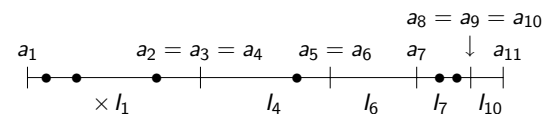
$p1B1b = P(\text{one lineage escapes but the other does not})$ .

$p22 = P(\text{no coalescence}) = p2inb + p1B1b$

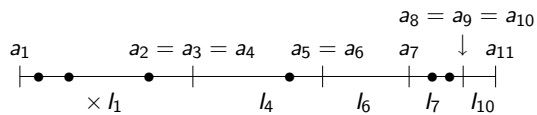
	$pinb$	$p2inb$	$p2cinb$	$p1B1b$	$p22$
Approx. 1	0.4	0.16	0	0.48	0.64
logistic ODE	0.39936	0.13814	0.09599	0.32646	0.46460
Moran sim	0.33656	0.10567	0.05488	0.35201	0.45769
Approx. 2	0.34065	0.10911	0.05100	0.36112	0.47203

## Approximation 2

A stick breaking construction that leads to a coalescent with simultaneous multiple collisions.



Pieces of stick are coalesced lineages that escape due to recombination. Sampled individuals = points random on  $(0,1)$ . Two in the same piece coalesce.  $l_1$  may be marked ( $\times$ ) or not (escapes sweep).

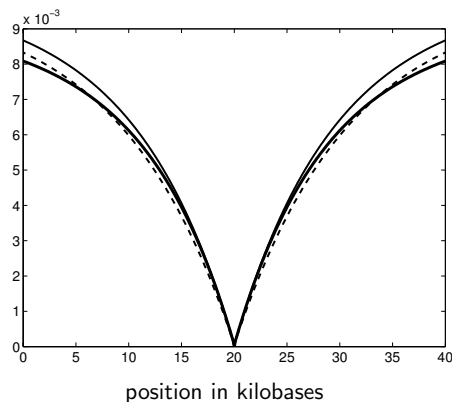


$M = [2Ns]$  number of lineages with an infinite line of descent  
 $\xi_\ell, 2 \leq \ell \leq M$  iid Bernoulli, 1 (recombination) with prob  $r/s$ .  
 $W_\ell, 2 \leq \ell \leq M$  are beta( $1, \ell - 1$ ) (fraction of lineages)  
 $V_\ell = \xi_\ell W_\ell, T_\ell = V_\ell \prod_{i=\ell+1}^M (1 - V_i)$   
 $a_\ell = a_{\ell+1} - T_\ell, l_\ell = [a_\ell, a_{\ell+1}]$

Proofs. Schweinsberg and Durrett (2005) Ann. Appl. Prob.  
 Error is  $O(1/\log^2 N)$  versus  $O(1/\log N)$  for approx 1

## Reduction of $\pi = 0.01$ due to a sweep

Kim and Stephan (2002)  $>$  D & S (dashed)  $\approx$  answer



## A Drosophila Puzzle

Begun and Aquadro (1992) observed that in *Drosophila melanogaster* there is a positive correlation between nucleotide diversity and recombination rates. Two explanations:

- Repeated episodes of hitchhiking caused by the fixation of newly arising advantageous mutations, which has a greater effect in regions of low recombination, because the average size of the region affected depends on the ratio  $s/r$ .
- Background selection (removal of deleterious alleles) which leads to a reduction of the "effective population size" has a greater impact in regions of low recombination, but does not change the site frequency spectrum.

## $\Lambda$ -coalescents. Pitman, Möhle and Sagitov

State is a partition. Sets in partition are lineages that have coalesced.  
 $\xi \rightarrow \eta$  is a  $k$ -merger if  $k$  sets in  $\xi$  collapse to one in  $\eta$ , and the rest of  $\eta$  does not change.

$$q_{\xi, \eta} = \int_0^1 p^{k-2} (1-p)^{|\xi|-k} \Lambda(dp)$$

$\Lambda(\{0\}) = 1$ . Kingman's coalescent.

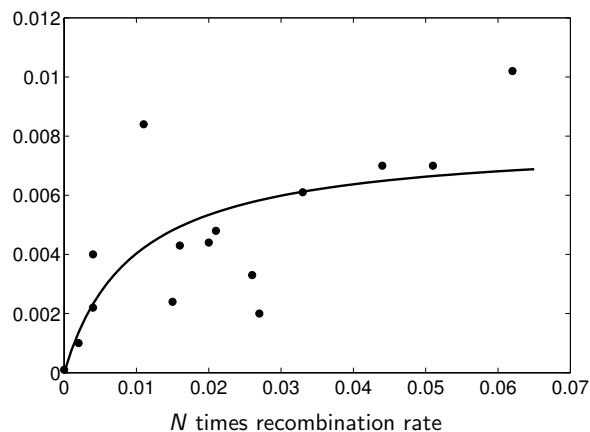
If  $\lambda = \int_0^1 p^{-2} \Lambda(dp) < \infty$ ,  $p$ -mergers with a random  $p^{-2} \Lambda(dp)$  distributed  $p$  occur at rate  $\lambda$ .

## Durrett and Schweinsberg (2005) Stoch. Proc. Appl.

Suppose that the recombination rate between 0 and  $x$  is  $\beta|x|$ . Mutations with a fixed selective advantage  $s$  occur in the population at rate  $\gamma$  per unit length.

**Theorem.** The genealogies converge to a  $\Lambda$  coalescent with  $\Lambda = \delta_0 + c \gamma dy$  where  $c = 2\gamma s^2 / \beta$ .

## Comparison with data on $\pi$ . Stephan (1995)



## Large family sizes

The original biological motivation for  $\Lambda$ -coalescents is that many species have a highly variable number of offspring.

**Cannings' model** Suppose that the  $2N$  members of the population have offspring  $(\nu_1, \dots, \nu_{2N})$ . The  $\nu_i$  are exchangeable and sum to  $2N$ . (Distribution depends on  $N$ .)

**Möhle (2000)**. Run time at rate  $2N/\text{var}(\nu_i)$ . Convergence to Kingman's coalescent occurs if and only if

$$\frac{E[\nu_1(\nu_1 - 1)(\nu_1 - 2)]/N^2}{E[\nu_1(\nu_1 - 1)]/N} \rightarrow 0$$

In words, if and only if no triple mergers.

## Schweinsberg (2003) Stoch. Proc. Appl.

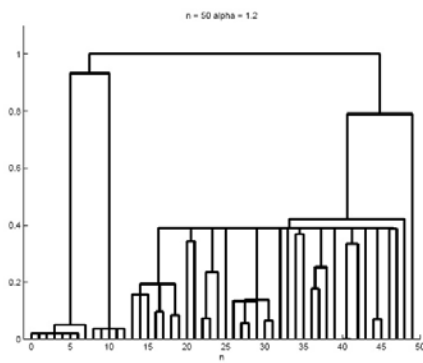
Each individual has  $X_i$  offspring (independent) then  $N$  are chosen to make the next generation. Part (c) of Theorem 4 shows

**Theorem.** Suppose  $EX_i = \mu > 1$  and  $P(X_i \geq k) \sim Ck^{-\alpha}$  with  $1 < \alpha < 2$ . Then, when time is run at rate  $2N/\text{var}(\nu_i) \approx C'N^{\alpha-1}$ , the genealogical process converges to a  $\Lambda$ -coalescent where  $\Lambda$  is the beta( $2 - \alpha, \alpha$ ) distribution, i.e.,

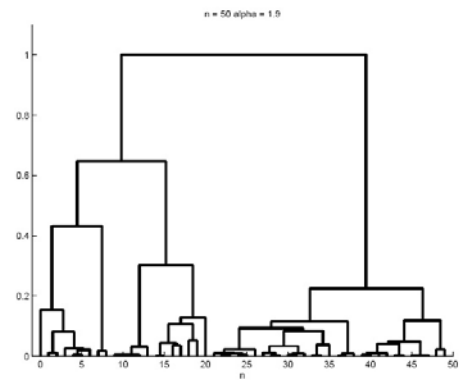
$$\Lambda(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha, \alpha)}$$

where  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ , and  $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x} dx$  is the usual gamma function.

## Genealogy when $\alpha = 1.2$



## Genealogy when $\alpha = 1.9 \approx$ Kingman



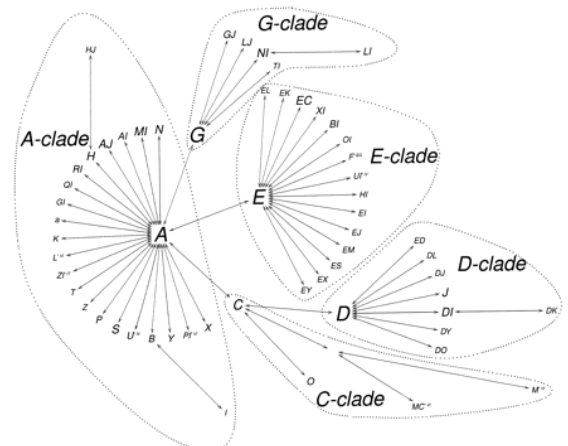
## Arnason (2004) data on cytochrome b in 1278 cod

39 mutations define 59 haplotypes (mutation patterns):

This indicates some sites were hit more than once, for if not, the number of haplotypes = 1 + the number of mutations

Haplotype frequencies:

696, 193, 124, 112, 29, 15, 9, 7, 6, 5(3), 4(2), 3(6), 2(7), 1(32)



## Site frequency spectrum

J. Berestycki, N. Berestycki, and Schweinsberg (2006a,b).

**Theorem** Suppose we introduce mutations into the beta coalescent at rate  $\theta$ , and let  $M_{n,k}$  be the number of mutations affecting  $k$  individuals in a sample of size  $n$ . Then

$$\frac{M_{n,k}}{S_n} \rightarrow a_k = \frac{(2-\alpha)\Gamma(\alpha+k-2)}{\Gamma(\alpha-1)k!} \sim C_\alpha k^{\alpha-3}$$

in probability as  $n \rightarrow \infty$ .

When  $\alpha = 2$  this reduces to the  $1/k$  behavior found in Kingman's coalescent.

When  $k = 1$ ,  $a_k = 2 - \alpha$ .

Navigation icons

## Data set 2

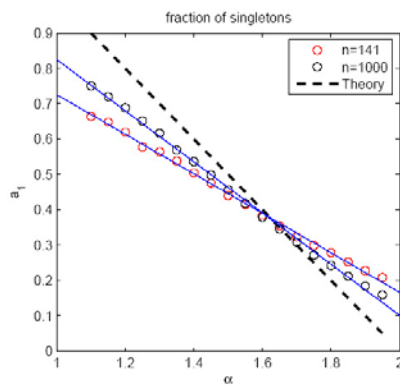
Boom, Boulding, and Beckenbach (1994) did a restriction enzyme digest of mtDNA on a sample of 141 Pacific Oysters from British Columbia. They found 51 segregating sites and 30 singleton mutations, resulting in an estimate of

$$\alpha = 2 - \frac{30}{51} = 1.41$$

However, this estimate is biased. If the underlying data was generated by Kingman's coalescent, we would expect a fraction  $1/\ln(141) = 0.202$  of singletons, resulting in an estimate of  $\alpha = 1.8$ .

Navigation icons

## BBB $\alpha = 1.19$ (uncorr: 1.41), Arnason $\alpha = 1.54$



Navigation icons

## Segregating sites

J. Berestycki, N. Berestycki, and Schweinsberg (2006a,b).

**Theorem** Suppose we introduce infinite sites mutations into the beta coalescent at rate  $\theta$ , and let  $S_n$  be the number of segregating sites observed in a sample of size  $n$ . If  $1 < \alpha < 2$  then

$$\frac{S_n}{n^{2-\alpha}} \rightarrow \frac{\theta\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}$$

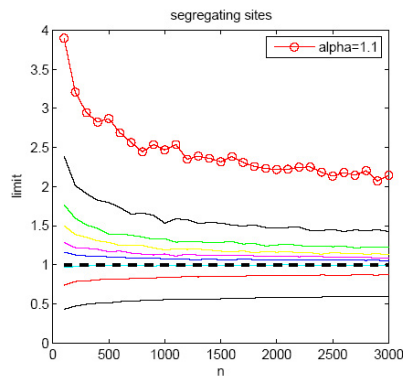
in probability as  $n \rightarrow \infty$ .

In Kingman's coalescent

$$\frac{S_n}{\log n} \rightarrow \theta$$

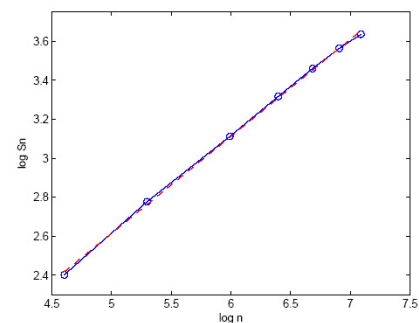
Navigation icons

## Simulation mean / formula : slow convergence



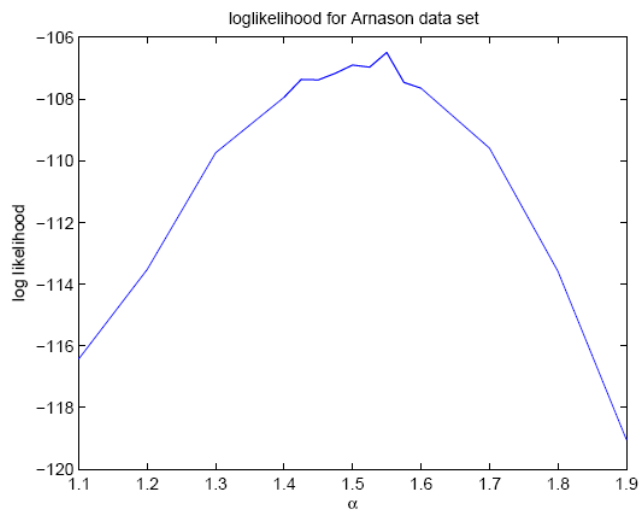
Navigation icons

## Subsampling the Arnason data, $\alpha \approx 1.50$ (prev: 1.54)



Navigation icons

## PRF likelihood of SFS – Carlos Bustamante



## Estimation results: Emilia Huerta-Sanchez

Now VIGRE postdoc, U.C. Berkeley Statistics.

