Wald Lecture 2
My Work in Genetics with Jason Schweinsbreg

Rick Durrett

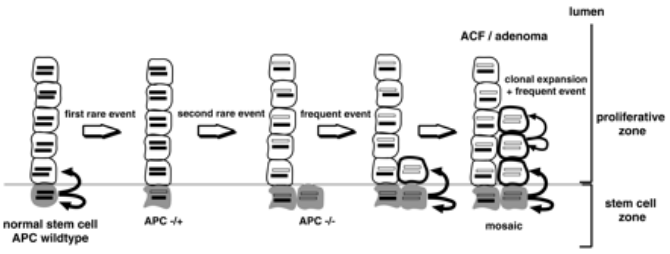
The Problem

Given a population of size N, how long does it take until τ_k the first time we have an individual with a prespecified sequence of k mutations?
We use the Moran model.

- Initially all individuals are type 0.
- Each individual is subject to replacement at rate 1.
- A copy is made of an individual chosen at random from the population.
- Type $j-1$ mutates to type j at rate u_j.

Progression to Colon Cancer

Luebeck and Moolgavakar (2002) PNAS fit a four stage model to incidence of colon cancer by age.

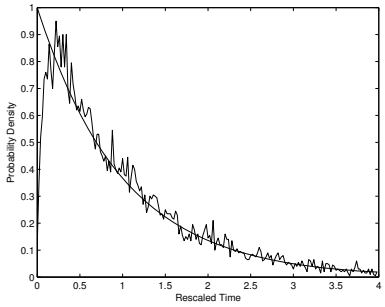


k=2 : Iwasa, Michor, Nowak (2004) Genetics

Theorem. If $Nu_1 \to 0$ and $N\sqrt{u_2} \to \infty$

$$P(\tau_2 > t/Nu_1\sqrt{u_2}) \to e^{-t}$$

10,000 simulations of $n = 10^3$, $u_1 = 10^{-4}$, $\sqrt{u_2} = 10^{-2}$



Idea of Proof

Since 1's mutate to 2's at rate u_2, τ_2 will occur when there have been $O(1/u_2)$ births of individuals of type 1.

The number of 1's is roughly a symmetric random walk, so τ_2 will occur when the number of 1's reaches $O(1/\sqrt{u_2})$.

$N \gg 1/\sqrt{u_2}$ guarantees that up to τ_2 the number of 1's is $o(N)$, so 1 mutations occur at rate Nu_1.

The waiting time from the 1 mutation until the 2 mutant appears is of order $1/\sqrt{u_2}$. For this to be much smaller than the overall waiting time $1/Nu_1\sqrt{u_2}$ we need $Nu_1 \ll 1$.

Waiting for k mutations

Total progeny of a critical binary branching process has $P(\xi > k) \sim Ck^{-1/2}$, so the sum of M such random variables is $O(M^2)$.

To get 1 individual of type 4, we need of order

$1/u_4$ births of type 3.

$1/\sqrt{u_4}$ mutations to type 3.

$1/u_3\sqrt{u_4}$ births of type 2.

$1/u_3^{1/2}u_4^{1/4}$ mutations to type 2.

$1/u_2u_3^{1/2}u_4^{1/4}$ births of type 1.

$1/u_2^{1/2}u_3^{1/4}u_4^{1/8}$ mutations to type 1.

Durrett, Schmidt, Schweinsberg, Ann Prob.

Probability type j has a type k descendant.

$$\sim r_{j,k} = u_{j+1}^{1/2} u_{j+2}^{1/4} \cdots u_k^{1/2^{k-j}} \quad \text{for } 1 \le j < k$$

Theorem. Let $k \ge 2$. Suppose that:
(i) $Nu_1 \to 0$.
(ii) For $j = 1, \ldots, k-1$, $u_{j+1}/u_j > b_j$ for all N.
(iii) There is an $a > 0$ so that $N^a u_k \to 0$.
(iv) $Nr_{1,k} \to \infty$.
Then for all $t > 0$, $\lim_{N\to\infty} P(\tau_k > t/Nu_1 r_{1,k}) = \exp(-t)$.

When $Nr_{1,k} \not\to \infty$

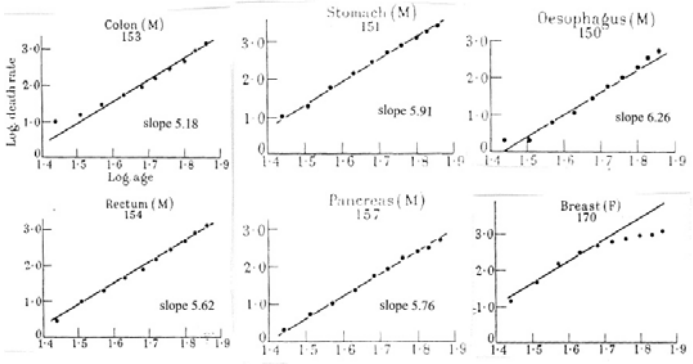
Fixation of 1 before τ_k and stochastic tunneling each have positive probability. Using convergence to the Wright-Fisher diffusion and the Feynman-Kac formula we can prove.

Theorem. Let $k \ge 2$. Assume (i), (ii), and (iii) from before.
(iv) $(Nr_{1,k})^2 \to \gamma > 0$, and we let

$$\alpha = \sum_{k=1}^{\infty} \frac{\gamma^k}{(k-1)!(k-1)!} \Big/ \sum_{k=1}^{\infty} \frac{\gamma^k}{k!(k-1)!} > 1$$

then for all $t > 0$, $\lim_{N\to\infty} P(u_1\tau_k > t) = \exp(-\alpha t)$.

Back to reality. Armitage and Doll (1954)



Small time behavior

Our results are appropriate for the regulatory sequence application since one is interested in the typical amount of time that the process takes.

However, most cancers occur in less than 1% of the population so we are looking at the lower tail of the distribution. Let $g_k(t) = Q_1(\tau_k \le t)$ where Q_1 is the probability for the branching process started with one type 1. In the case $u_j \equiv \mu$

$$g_j'(t) = \mu g_{j-1}(t) - (1-\mu) g_j(t)^2 - 2\mu g_j(t)$$

One can inductively solve the differential equations and finds

If $t \ll \mu^{-1/2}$ then $g_k(t) \approx \mu^{k-1} t^{k-1}/(k-1)!$

Other results

Schweinsberg (2008) studies all possible limits in the case $\mu_j \equiv \mu$ paper is on the arXiv

When $m = 3$ the behavior changes at

$$\mu = N^{-2} \qquad N^{-4/3} \qquad N^{-1} \qquad N^{-2/3}$$

Thus there are five regimes and four borderline cases.

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

$$i \to i+1 \quad \text{at rate} \quad b_i = (2N-i) \cdot \frac{i}{2N}$$
$$i \to i-1 \quad \text{at rate} \quad d_i = i \cdot \frac{2N-i}{2N}$$

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:

$$i \to i+1 \quad \text{at rate} \quad b_i = (2N-i)\cdot \frac{i}{2N}$$
$$i \to i-1 \quad \text{at rate} \quad d_i = i \cdot \frac{2N-i}{2N}(1-s)$$

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 r.

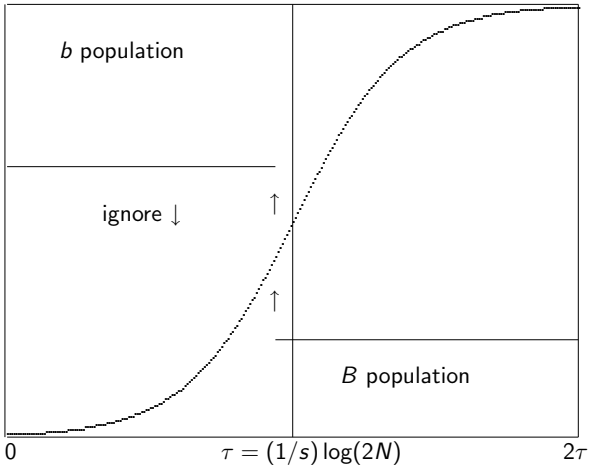
Let p_0 = frequency of B before the sweep $(1/2N)$.
$Q_t = P(A|B)$. $R_t = P(A|b)$.

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

$$Q_\infty = R_0(1-p_0)\int_0^{2\tau} \frac{re^{-rt}}{(1-p_0)+p_0e^{st}}\,ds$$

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.

Hitchhiking = Population subdivision



Durrett and Schweinsberg (2004) Th. Pop. Biol.

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

$$pinb = \int_0^{2\tau} \frac{re^{-rt}}{(1-p_0)+p_0e^{st}}\,ds$$

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

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

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

then for $j \geq 2$

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

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

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

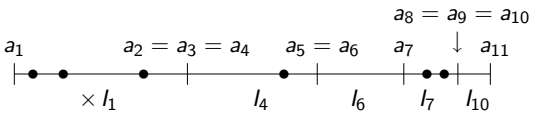
$N = 10,000$, $s = 0.1$. Set $r = 0.00516$ so $pinb \approx 0.4$.

$p2inb = P($ both lineages escapes the sweep and do not coalesce$)$.
$p2cinb = P($ both lineages escape the sweep but coalesce$)$.
$p1B1b = P($ one lineage escapes but the other does not$)$.
$p_{22} = P($ no coalescence $) = p2inb + p1B1b$

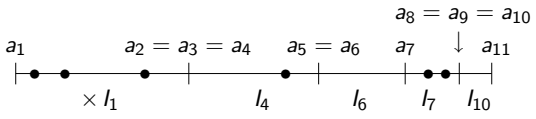
	$pinb$	$p2inb$	$p2cinb$	$p1B1b$	p_{22}
Approx. 1	0.4	0.16	0	0.48	0.64
logistic ODE	0.39936	0.13814	0.09599	0.32646	0.46460
Moran sim	0.33656	0.10567	0.05488	0.35201	0.45769
Approx. 2	0.34065	0.10911	0.05100	0.36112	0.47203

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. I_1 may be marked ($\times$) or not (escapes sweep).

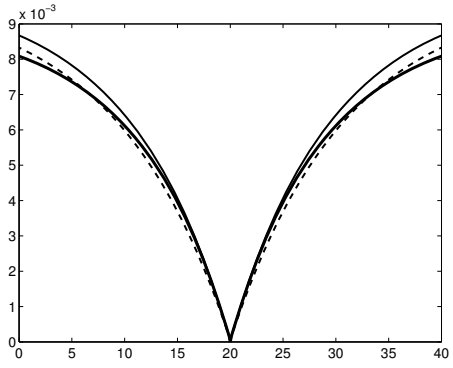


$M = [2Ns]$ number of lineages with an infinite line of descent
ξ_ℓ, $2 \leq \ell \leq M$ iid Bernoulli, 1 (recombination) with prob r/s.
W_ℓ, $2 \leq \ell \leq M$ are beta$(1, \ell-1)$ (fraction of lineages)
$V_\ell = \xi_\ell W_\ell$, $T_\ell = V_\ell \prod_{i=\ell+1}^{M}(1-V_i)$
$a_\ell = a_{\ell+1}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

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.

Λ-coalescents. Pitman, Möhle and Sagitov

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

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

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

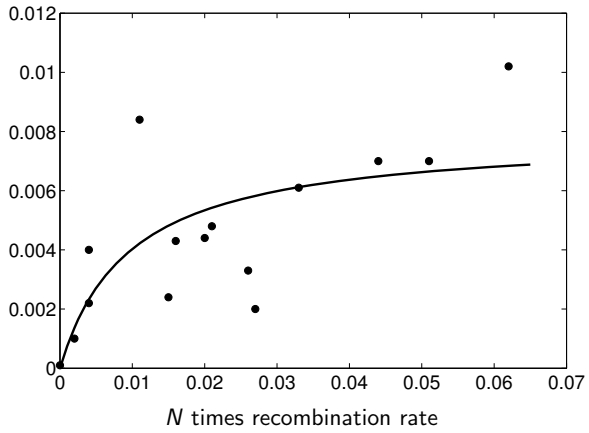
If $\lambda = \int_0^1 p^{-2}\Lambda(dp) < \infty$, p-mergers with a random $\lambda^{-1}p^{-2}\Lambda(dp)$ distributed p occur at rate λ.

Durrett and Schweinsberg (2005) SPA

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 γ per unit length.

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

Comparison with data on π. Stephan (1995)



Large family sizes

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

Cannings' model Suppose that the $2N$ members of the population have offspring $(\nu_1, \ldots \nu_{2N})$. The ν_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

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

In words, if and only if no triple mergers.

Schweinsberg (2003) Stoch. Proc. Appl.

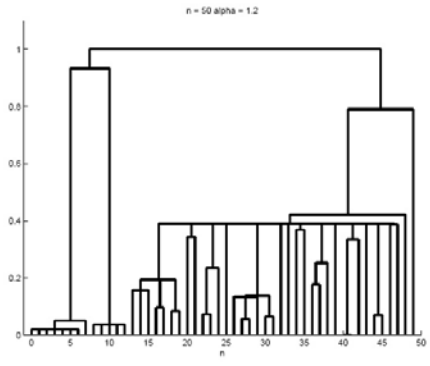
Each individual has X_i offspring (independent) then N are chosen to make the next generation. Part (c) of Theorem 4 shows

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

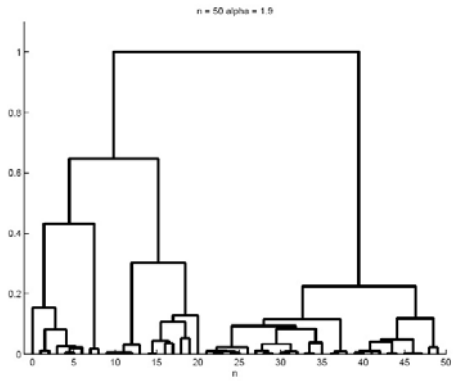
$$\Lambda(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha,\alpha)}$$

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



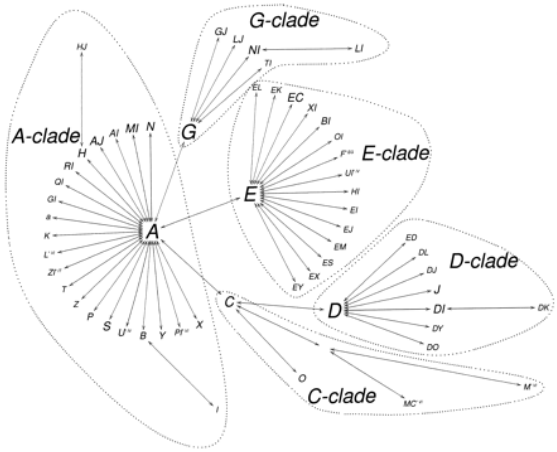
Arnason (2004) cytochrome b data, 1278 cod

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 θ, and let $M_{n,k}$ be the number of mutations affecting k individuals in a sample of size n. Then as $n \to \infty$,

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

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

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

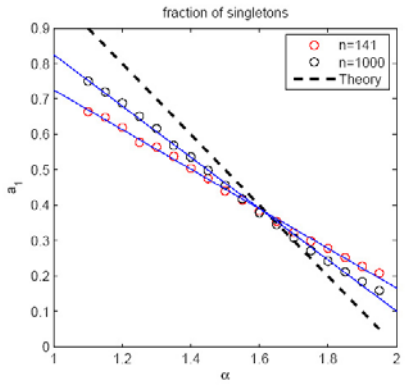
Data set 2

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 $\alpha = 1.8$.

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



Segregating sites

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

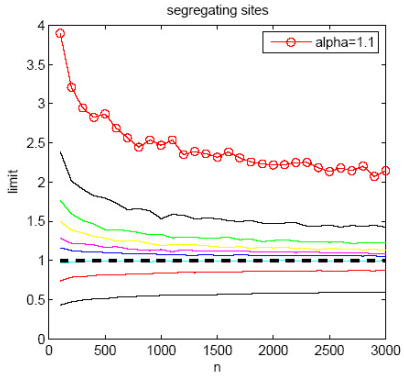
Theorem Suppose we introduce infinite sites mutations into the beta coalescent at rate θ, and let S_n be the number of segregating sites observed in a sample of size n. If $1 < \alpha < 2$ then as $n \to \infty$

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

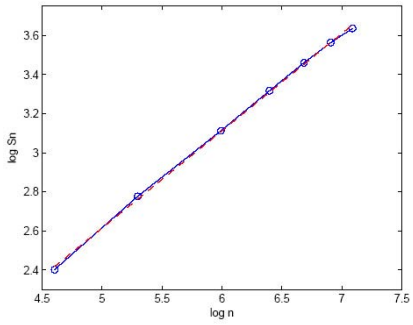
In Kingman's coalescent

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

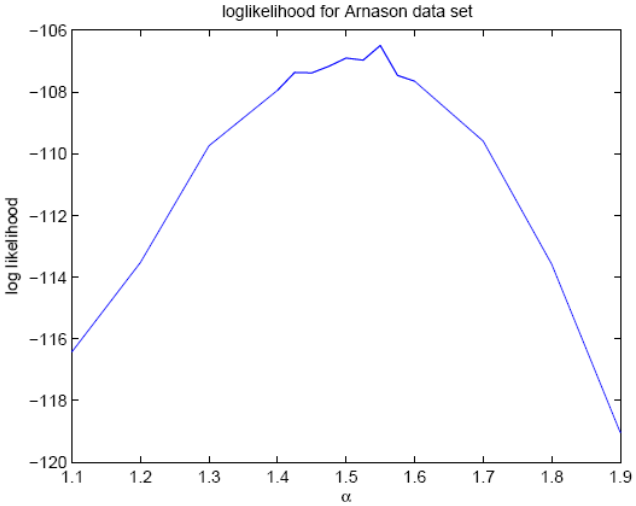
Simulation mean / formula : slow convergence



Subsampling Arnason, $\alpha \approx 1.50$ (vs. 1.54)



PRF likelihood of SFS – Carlos Bustamante



Estimation results: Emilia Huerta-Sanchez

Now VIGRE postdoc, U.C. Berkeley Statisitcs.