A coalescent model for the effect of advantageous mutations on the genealogy of a population

by Jason Schweinsberg

University of California at San Diego

(joint work with Rick Durrett)

Outline of Talk

1. The model
2. A simple approximation
3. An improved approximation
4. Recurrent beneficial mutations
5. Applications

The model

Population has fixed size $2N$.

Consider two sites on the chromosomes:

- One site has an A or a allele, neither is advantageous.
- One site has a B or b allele, B is advantageous.

At time zero, $2N-1$ chromosomes have the b allele and one has the B allele.

Each individual lives for an Exponential(1) time, then is replaced.

When a new individual is born:

- The B or b comes from a randomly chosen parent. A replacement of a B by a b is rejected with probability s.
- With probability $1-r$, the A or a comes from the same parent.
- With probability r, the A or a allele comes from a parent chosen independently at random.

Selective sweeps

Eventually, the number of B's reaches 0 or $2N$. If the number of B's reaches $2N$, a *selective sweep* occurs. The probability of a selective sweep is

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

Sample n individuals at the time τ when a selective sweep ends.

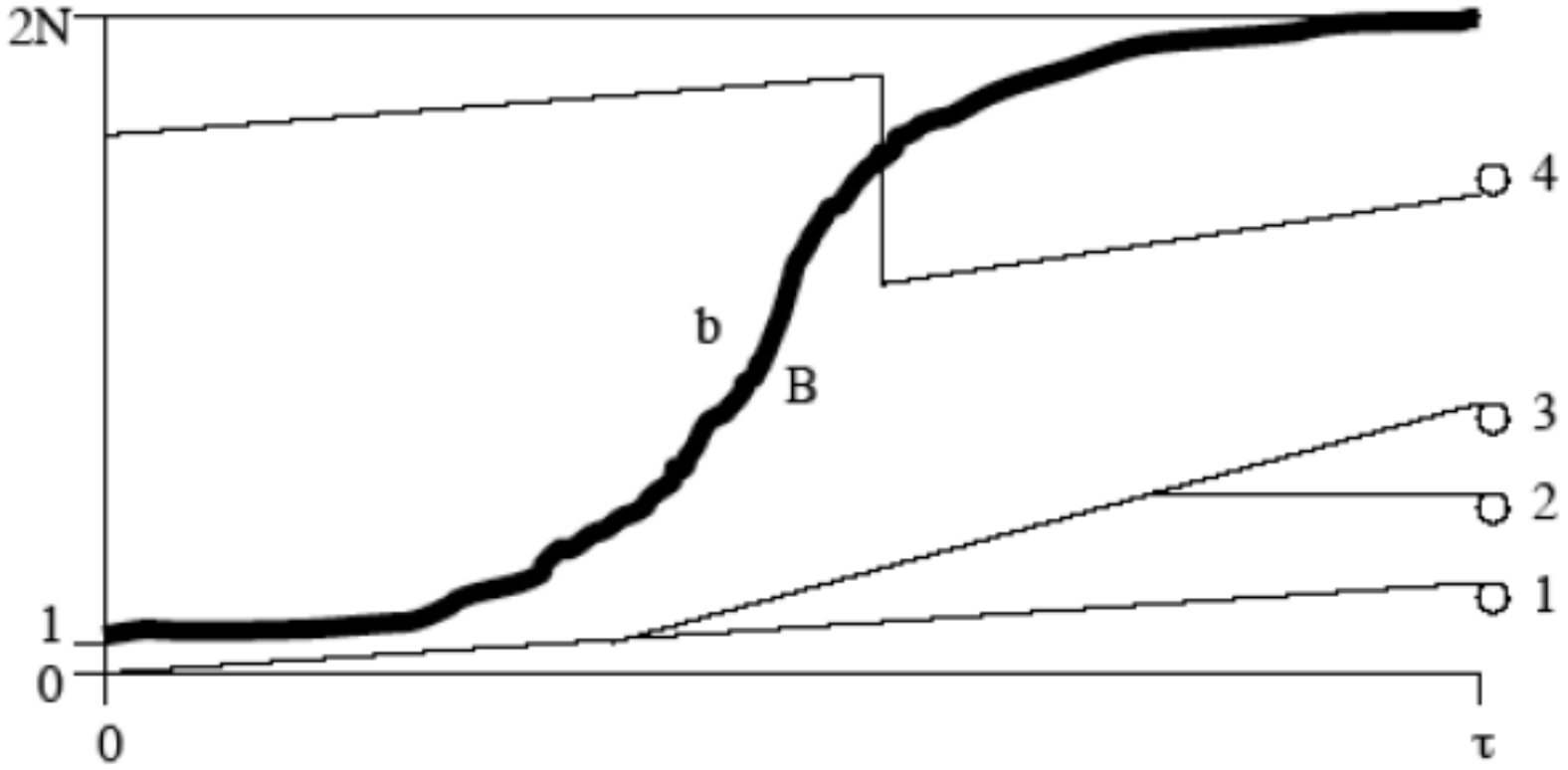
All n individuals in the sample inherited their B allele from the same individual at time 0.

Let Θ be a random partition of $\{1,\ldots,n\}$ such that i and j are in the same block if and only if the ith and jth sampled individuals inherited their A/a allele from the same individual at time zero.

Goal: to describe the distribution of the random partition Θ.

Previous work: Maynard Smith-Haigh (1974), Kaplan-Hudson-Langley (1989), Stephan-Wiehe-Lenz (1992), Barton (1998, 2000).

Illustration of a selective sweep



$$\Theta = \{\{1, 2, 3\}, \{4\}\}.$$

If the A/a allele of one individual comes from an individual that had the b allele at time zero, we say the lineage *escapes* the selective sweep.

A simple approximation

Define a random partition Θ_p of $\{1, \ldots, n\}$ as follows:

- Flip n independent coins with probability p of heads.
- One block of Θ_p is $\{i : \text{the } i\text{th coin is heads}\}$.
- The other blocks are singletons.

Theorem 1: Let $a = r \log(2N)/s$. Let $p = e^{-a}$. Suppose s is constant and $r \leq A/(\log N)$ for some constant A. Then there exists a positive constant C such that

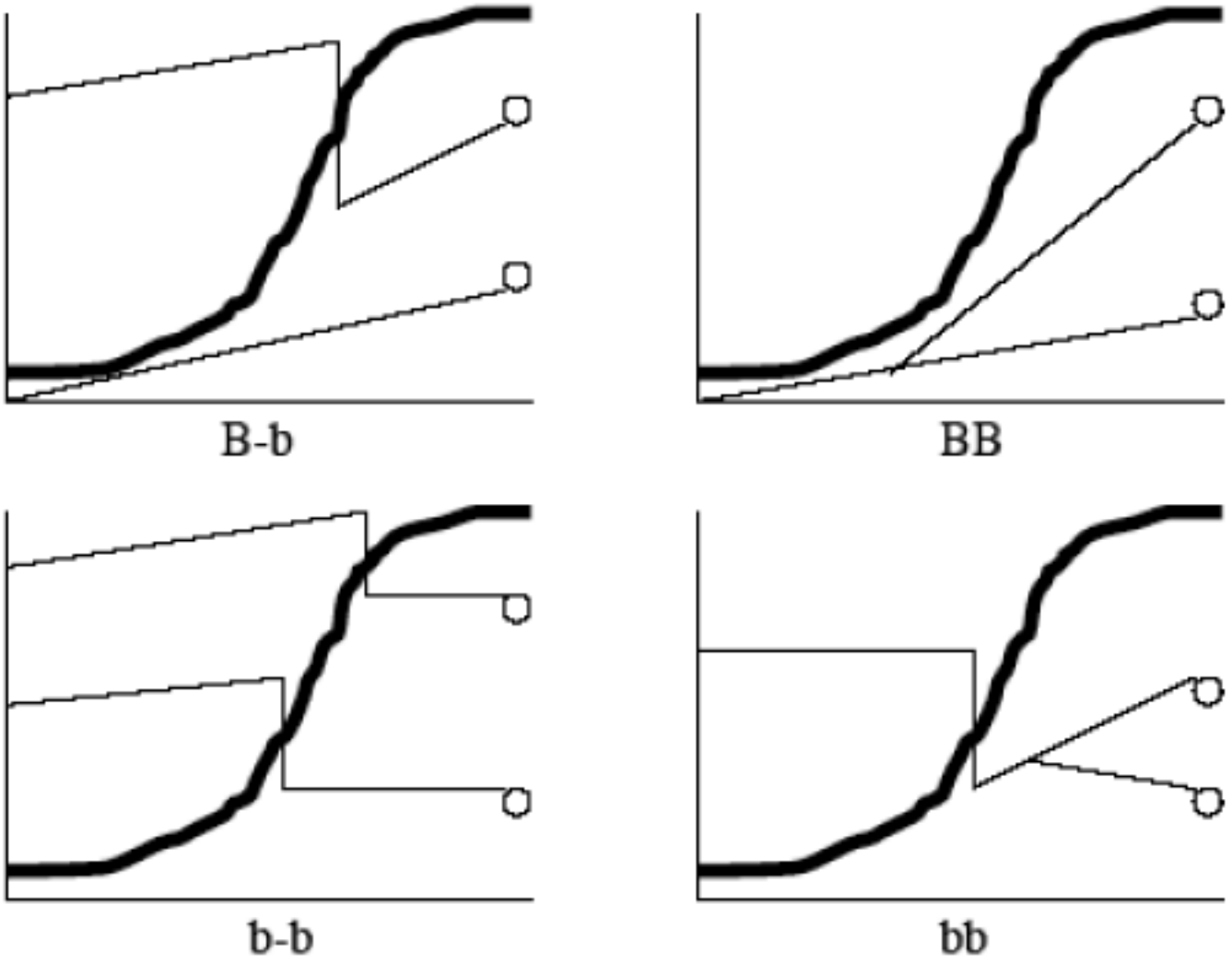
$$|P(\Theta = \pi) - P(\Theta_p = \pi)| \leq \frac{C}{\log N}$$

for all N and all partitions π of $\{1, \ldots, n\}$.

Simulations

Keep track of the fraction of lineages that escape the sweep.

Also, we have the following possibilities for two lineages:



Simulation results

Choose r so that $1 - e^{-a} = 0.4$, where $a = r \log(2N)/s$.

N = 10,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.295	.303	.553	.067	.077
Theorem 1	.400	.480	.360	.000	.160

N = 100,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.318	.352	.505	.046	.096
Theorem 1	.400	.480	.360	.000	.160

N = 1,000,000; s = 0.01	b	B–b	BB	bb	b–b
simulations	.308	.355	.515	.039	.091
Theorem 1	.400	.480	.360	.000	.160

Approximation based on Theorem 1 is poor, error $O(1/\log N)$.

Dominant source of error (Barton, 1998): a recombination soon after the beneficial mutation may cause several lineages that have already coalesced to be descended from the same individual in the b population. Then Θ has more than one large block.

The beginning of a selective sweep

The recombinations that cause additional large blocks in Θ are those that occur when the number of B's is small.

When the B-population is small, it is approximately a continuous-time branching process in which each individual dies at rate $1-s$ and gives birth at rate 1.

The number of lineages with an infinite line of descent is a branching process with no deaths and births at rate s.

Define $0 = \tau_1 < \tau_2 < \ldots$ such that τ_k is the first time at which there are k individuals with an infinite line of descent.

If there is recombination along a lineage with an infinite line of descent between times τ_k and τ_{k+1}, descendants of that lineage will have a different ancestor at the beginning of the sweep than descendants of the other $k-1$ lineages.

What fraction of the population is descended from this lineage?

Polya Urns and Branching Processes

Start with one "marked" lineage and $k-1$ "unmarked" lineages. Mark individuals descended from the marked lineage. When there are x marked individuals and y unmarked,

$$P(\text{next individual is marked}) = \frac{x}{x+y}.$$

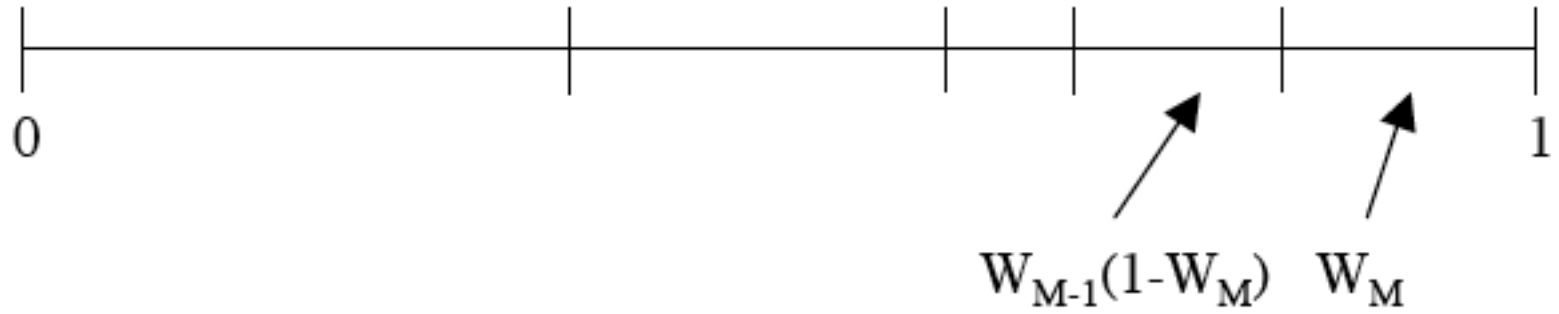
Polya urn: start with a white balls and b black balls. Repeatedly draw a ball at random, then return it to the urn along with another ball of the same color. When there are x white balls and y black balls, $P(\text{next ball is white}) = x/(x+y)$.

Equivalent description: let U have a Beta(a, b) distribution. Conditional on U, each ball is independently white with probability U, black with probability $1-U$.

The limiting fraction of marked individuals has a Beta$(1, k-1)$ distribution.

Stick-breaking construction

Stick-breaking (paintbox) construction (Kingman, 1978):



Let $M = \lfloor 2Ns \rfloor$. For $k = M, M-1, M-2, \ldots, 3, 2$, we break off a fraction W_k of the interval that is left.

W_k corresponds to the fraction of lineages that escape the sweep between times τ_k and τ_{k+1}.

Expected number of recombinations between τ_k and τ_{k+1} is r/s. Assume the number is 0 or 1.

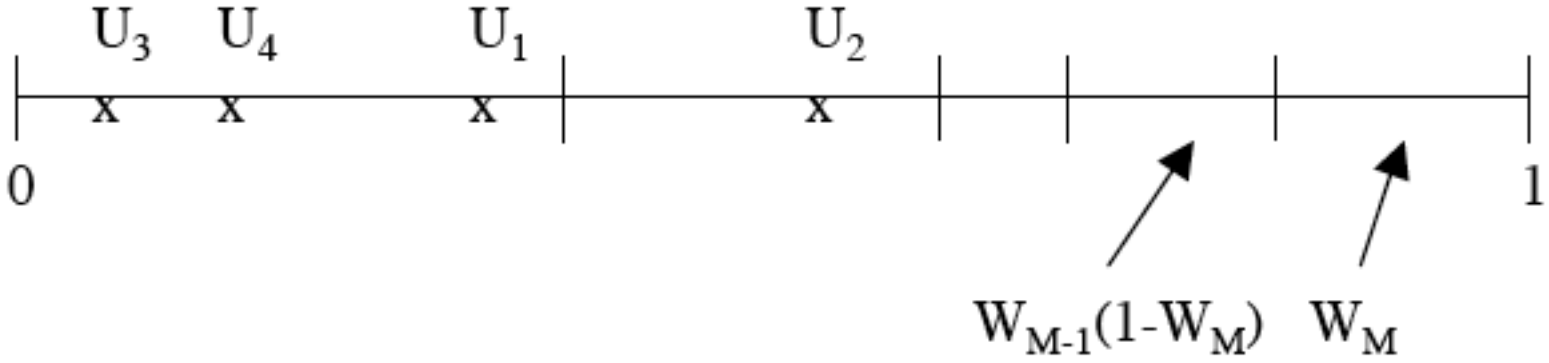
With probability r/s, W_k has the Beta$(1, k-1)$ distribution.

With probability $1 - r/s$, $W_k = 0$.

A second approximation

Let $U_1, U_2, \ldots, U_n$ be i.i.d. with the uniform distribution on $[0, 1]$.

Let Π be the random partition of $\{1, \ldots, n\}$ such that i and j are in the same block if and only if U_i and U_j are in the same subinterval.



Example: $\Pi = \{\{1, 3, 4\}, \{2\}\}$.

Theorem 2. If $r \leq A/\log(2N)$, then there exists a constant C such that for all N and all partitions π of $\{1, \ldots, n\}$, we have

$$|P(\Theta = \pi) - P(\Pi = \pi)| \leq \frac{C}{(\log N)^2}.$$

Simulation results

Choose r so that $1 - e^{-a} = 0.4$, where $a = r \log(2N)/s$.

N = 10,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.295	.303	.553	.067	.077
Theorem 2	.301	.318	.540	.059	.082

N = 100,000; s = 0.03	b	B–b	BB	bb	b–b
simulations	.318	.352	.505	.046	.096
Theorem 2	.321	.358	.501	.044	.098

N = 1,000,000; s = 0.01	b	B–b	BB	bb	b–b
simulations	.308	.355	.515	.039	.091
Theorem 2	.308	.358	.513	.038	.091

The stick-breaking approximation works much better than the coin tossing approximation.

Remarks

1. Theorems 1 and 2 hold for "strong selection" when the selective advantage s is $O(1)$.

2. One can also consider "weak selection" when s is $O(1/N)$. There is diffusion limit, studied by Krone-Neuhauser (1997), Donnelly-Kurtz (1999), Barton-Etheridge-Sturm (2004).

3. Etheridge-Pfaffelhuber-Wakolbinger (2005) show that same approximations work in the diffusion limit, if we set $s = \alpha/N$ and then let $\alpha \to \infty$.

4. Eriksson-Fernström-Mehlig-Sagitov (2007) give approximation that works well even when r/s is large.

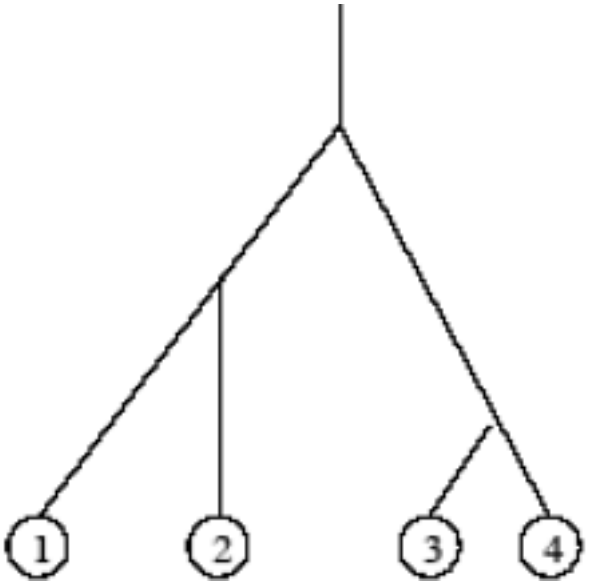
Coalescent processes

Sample n individuals at time 0.

Let $\Psi_N(t)$ be the partition of $\{1, \ldots, n\}$ such that i and j are in the same block iff the ith and jth individuals in the sample have the same ancestor at time $-t$.

Consider the process $\Psi_N = (\Psi_N(Nt), t \geq 0)$, which is a coalescent process taking its values in the set of partitions of $\{1, \ldots, n\}$.

For the ordinary Moran model (no selective sweeps), Ψ_N is Kingman's coalescent (each pair of blocks merges at rate 1).



{1, 2, 3, 4}

{1, 2}, {3, 4}

{1}, {2}, {3, 4}

{1}, {2}, {3}, {4}

Recurrent selective sweeps

The duration of a selective sweep is approximately $(2/s)\log(2N)$.

With strong selection, all of the lineages that coalesce during a selective sweep do so almost instantaneously for large N.

Gillespie (2000) proposed that selective sweeps happen at times of a Poisson process.

If selective sweeps happen at rate $O(N^{-1})$, then Ψ_N converges to a coalescent with multiple collisions (Pitman (1999), Sagitov (1999)) in which many blocks can merge at once.

A better approximation can be obtained using a coalescent with simultaneous multiple collisions (Möhle-Sagitov (2001), Schweinsberg (2000)) in which many mergers can occur simultaneously.

Coalescents with multiple collisions

Let π be a partition of $\{1, \ldots, n\}$ into blocks $B_1, \ldots, B_j$. Let $p \in (0, 1]$. A p-merger of π is obtained as follows:

- Let $\xi_1, \ldots, \xi_j$ be i.i.d. Bernoulli(p).
- Merge the blocks B_i such that $\xi_i = 1$.

Coalescents can be described in terms of a finite measure Λ on $[0, 1]$. Write $\Lambda = a\delta_0 + \Lambda_0$, where $\Lambda_0(\{0\}) = 0$. Transitions in the Λ-coalescent are as follows:

- Each pair of blocks merges at rate a.
- Construct a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity $dt \times p^{-2}\Lambda_0(dp)$. If (t, p) is a point of this Poisson process, then a p-merger occurs at time t.

When there are b blocks, let $\lambda_{b,k}$ denote the rate of a transition in which k blocks merge into one. Then, for $2 \leq k \leq b$,

$$\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k}\, \Lambda(dp).$$

Limiting processes

- No selection: $\Lambda = \delta_0$ (Kingman's coalescent).
- Case 1: If the mutations all occur at the same site, then $\Lambda = \delta_0 + \alpha p^2 \delta_p$.
- Case 2: If mutations and recombinations occur uniformly along the chromosome, then $\Lambda(dx) = \delta_0 + \beta x \, dx$.
- Other Λ could arise under different assumptions.

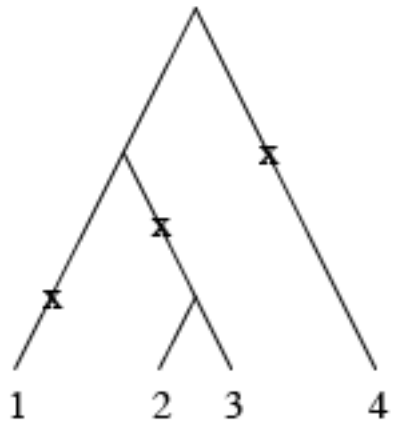
Assume that the genealogy of the population can be described by a Λ-coalescent, and that we are in either Case 1 or Case 2.

Assume neutral mutations occur along each lineage at rate $\theta/2$. Infinite sites model: each mutation happens at a different site.

Segregating sites

Let S_n be the number of segregating sites.

Example: $S_n = 3$.



Let λ_b be merger rate for the Λ-coalescent when b blocks.

Let $G_n(b) = P(\text{coalescent has exactly } b \text{ blocks at some time})$.

$$E[S_n] = \frac{\theta}{2} \sum_{b=2}^{n} b \lambda_b^{-1} G_n(b).$$

Kingman: $E[S_n] = \frac{\theta}{2} \sum_{b=2}^{n} b \binom{b}{2}^{-1} = \theta \sum_{b=2}^{n} \frac{1}{b-1} = \theta h_{n-1}$.

Cases 1 and 2: $\lim_{n \to \infty} (E[S_n] - \theta h_{n-1}) = -\rho$.

Pairwise differences

Let $\Delta_{i,j}$ be number of sites at which segments i and j differ.

Let $\Delta_n = \binom{n}{2}^{-1} \sum_{i<j} \Delta_{i,j}$.

$E[\Delta_n] = \theta \lambda_2^{-1}$.

Number of Singletons

Let J_n be number of mutations that affect exactly one lineage.

Kingman: $E[J_n] = \theta$.

Case 1: $E[J_n] = \theta - O((\log n)/n)$.

Case 2: $E[J_n] = \theta - O((\log n)^2/n)$.

Test Statistics

Tajima's (1989) D-statistic:

$$D = \frac{\Delta_n - S_n/h_{n-1}}{\sqrt{a_n S_n + b_n S_n^2}}.$$

Multiple mergers reduce Δ_n by $O(1)$ and S_n/h_{n-1} by $O(1/\log n)$, so D will be negative, consistent with simulations of Braverman-Hudson-Kaplan-Langley-Stephan (1995) and Simonsen-Churchill-Aquadro (1995).

Fu and Li's D-statistic (1993):

$$D = \frac{S_n - h_{n-1} J_n}{\sqrt{c_n S_n + d_n S_n^2}}.$$

Expected value of numerator goes to $-\rho$ as $n \to \infty$.

Standard deviation of numerator is $O(\log n)$ for Fu and Li's D-statistic but $O(1)$ for Tajima's D-statistic, so Tajima's D-statistic should be more powerful for detecting selective sweeps.

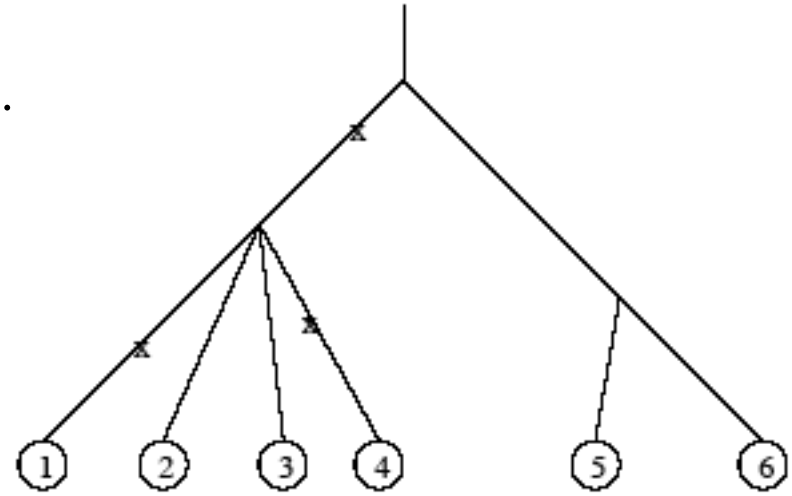
Site Frequency Spectrum

Let M_k be the number of mutations that affect k lineages. The sequence $(M_1, M_2, \ldots, M_{n-1})$ is the site frequency spectrum.

Full site frequency spectrum is needed for Fay and Wu's (2000)

$$H = \Delta_n - \sum_{k=1}^{n-1} \frac{2k^2 M_k}{n(n-1)}.$$

Kingman: $E[M_k] = \theta/k$ for all k.



A single selective sweep increases the number of high-frequency and low-frequency mutants (Fay-Wu, 2000; Kim-Stephan, 2002).

Recurrent selective sweeps lead to an excess of low-frequency mutants but not high-frequency mutants (Kim, 2006).

Analytical results for cases 1 and 2 have not yet been obtained.