# Approximating Selective Sweeps

by Richard Durrett and Jason Schweinsberg

Dept. of Math, Cornell U.

Corresponding Author: Richard Durrett

Dept. of Mathematics

523 Malott Hall

Cornell University

Ithaca NY 14853

Phone: 607-255-8282

FAX: 607-255-7149

email: rtd1@cornell.edu

ABSTRACT

The fixation of advantageous mutations in a population has the effect of reducing variation in the DNA sequence near that mutation. Kaplan, Hudson, and Langley (1989) used a three-phase simulation model to study the effect of selective sweeps on genealogies. However, most subsequent work has simplified their approach by assuming that the number of individuals with the advantageous allele follows the logistic differential equation. We show that the impact of a selective sweep can be accurately approximated by a random partition created by a stick-breaking process. Our simulation results show that ignoring the randomness when the number of individuals with the advantageous allele is small can lead to substantial errors.

Key words: selective sweep, hitchhiking, coalescent, random partition, paintbox construction

When a selectively favorable mutation occurs in a population and is subsequently fixed (i.e., its frequency rises to 100%), the frequencies of alleles at closely linked loci are altered. Alleles present on the chromosome on which the original mutation occurred will tend to increase in frequency, and other alleles will decrease in frequency. Maynard Smith and Haigh (1974) referred to this as the 'hitchhiking effect,' because an allele can get a lift in frequency from selection acting on a neighboring allele. They considered a situation with a neutral locus with alleles $A$ and $a$ and a second locus where allele $B$ has a fitness of $1+s$ relative to $b$. Suppose $p_0$ is the initial frequency of the $B$ allele, and $Q_n$ and $R_n$ are the frequencies in generation $n$ of the $A$ allele on chromosomes containing $B$ and $b$ respectively. If $Q_0 = 0$ (i.e., the advantageous mutation arises on a chromosome with $a$) and the recombination probability per generation is $r$, Maynard Smith and Haigh (1974) showed (see (8) on page 25) that the frequency of the $A$ allele after the selective sweep is reduced from $R_0$ to

$$\lim_{n\to\infty} Q_n = R_0 \sum_{n=0}^{\infty} r(1-r)^n \cdot \frac{1-p_0}{1-p_0+p_0(1+s)^{n+1}} \tag{1}$$

This is the frequency of $A$ in the entire population since after the sweep all individuals have the $B$ allele.

Kaplan, Hudson, and Langley (1989) investigated the effect of selective sweeps on genealogies. The model they analyzed is equivalent to the coalescent in a subdivided population that consists of one subpopulation with the favored $B$ allele and another with the $b$ allele. For the size of the $B$ population they used a model running forward in time that consists of an initial phase in which the number of $B$s is a supercritical branching process, a middle deterministic piece where the frequency of $B$s follows the logistic differential equation

$$\frac{dp}{dt} = sp(1-p) \tag{2}$$

and a final random piece where the number of $b$s follows a subcritical branching process.

To describe this process in detail, consider a population of $N$ diploid individuals. We will find it convenient to ignore the fact that these individuals have other chromosomes

that do not have the two loci of interest, and refer to the population as consisting of $2N$ chromosomes. Suppose we trace $k$ copies of the neutral locus backwards in time through the selective sweep. At the end of the sweep all $k$ lineages will belong to the $B$ population, however as we go back in time, some of the lineages will jump to the $b$ population because of recombination. Let $M(t)$ be the number of chromosomes with the $B$ allele at time $t$, and let $(i, j)$ be the number of lineages in the $B$ and $b$ populations respectively. Then we get a coalescent with the following transition rates at time $t$:

| transition | rate |
|---|---|
| $(i,j) \to (i-1, j+1)$ | $ir\frac{2N-M(t)-j}{2N}$ |
| $(i,j) \to (i-1, j)$ | $i\left((1-r)\frac{i-1}{M(t)} + r\frac{i+j-1}{2N}\right)$ |
| $(i,j) \to (i+1, j-1)$ | $jr\frac{M(t)-i}{2N}$ |
| $(i,j) \to (i, j-1)$ | $j\left((1-r)\frac{j-1}{2N-M(t)} + r\frac{i+j-1}{2N}\right)$ |

To check the rates, note that to have the transition $(i, j) \to (i-1, j+1)$, one of the $i$ lineages in the $B$ population must be chosen, a recombination must occur, and the parent must be chosen from the $b$ population but not be one of the $j$ existing lineages. The transition $(i, j) \to (i-1, j)$ can happen in two ways. First, one of the $i$ lineages in the $B$ population must be chosen. Then, either we have no recombination and choose one of the other $i-1$ lineages from the $B$ population as the parent, or a recombination occurs and we choose one of the existing $i+j-1$ lineages as the parent. The last two cases are similar with the populations reversed. These rates are different from the ones in formula (8) on page 891 of Kaplan, Hudson, and Langley (1989) since we do not ignore the possibility that recombination and coalescence can both occur in one jump. This probability is significant when both lineages are in the $B$ population and it is small.

For a picture of this coalescent see Figure 1, which gives a possible genealogy of a sample of size 5. Lineages 1 and 2 escape from the sweep due to recombination, while lineages 3, 4, and 5 coalesce. Analytical results are difficult to obtain for this temporally inhomogeneous

process, so Kaplan, Hudson, and Langley resorted to simulation. Stephan, Wiehe, and Lenz (1992) and Wiehe and Stephan (1993) simplified the approach of Kaplan, Hudson, and Langley (1989) by ignoring the random first and third phases and modeling the change in the frequency of $B$'s by the logistic differential equation (2). This approach has also been popular in simulation studies; see e.g., Simonsen, Churchill, and Aquadro (1995) and Przeworski (2002).

The results that we present in this section and the next pertain to this model in which the fraction $p(t) = M(t)/2N$ of individuals with the $B$ allele at time $t$ deterministically follows (2), which implies that

$$p(t) = \frac{p(0)}{p(0) + (1 - p(0))e^{-st}} \tag{3}$$

We will assume that initially there is just one individual with the $B$ allele, so $p(0) = 1/2N$. We denote by $\tau$ the duration of the selective sweep, which we define to be the time such that $p(\tau) = 1 - 1/2N$. It follows from (3) that $\tau = (2/s)\ln(2N - 1)$. We assume that $k$ lineages are sampled at time $\tau$, and these lineages are traced back to the beginning of the sweep. We refer to this model for a selective sweep as the logistic sweep model.

Let $Q(t)$ and $R(t)$ be the expected frequencies of the $A$ allele in chromosomes containing $B$ and $b$ respectively at time $t$. Suppose that the single individual with the $B$ allele at time zero has the $a$ allele, so $Q(0) = 0$. Stephan, Wiehe and Lenz (1992) derived the following analog of (1):

$$Q(\tau) = R(0) \int_0^\tau re^{-rt} \cdot \frac{(1 - 1/2N)}{(1 - 1/2N) + (1/2N)e^{st}}\, dt \tag{4}$$

Let $Q_1(t)$ be the solution to (4) with $R(0) = 1$. A little thought reveals that $Q_1(\tau)$ is the probability that the neutral locus of an individual in the $B$ population at time $\tau$ is a descendant of one in the $b$ population at time 0. In words, it is the probability that recombination allows the lineage to escape from the selective sweep. Formula (4) can be simplified considerably for large populations. Here and in what follows, $r$ and $s$ may depend

on $N$ even though we have not recorded that dependence in the notation.

**Proposition 1.** *If $N \to \infty$ and $r \ln(2N)/s \to a$ then $Q_1(\tau) \to 1 - e^{-a}$.*

Numerical results show that this simple approximation is very accurate. For example if $N = 10^4$, $s = 0.1$, and $r = 0.001064$ then $1 - e^{-a} = 0.1$ while the value from (4) is 0.099832. The reader should note that this result is a little different from the rule of thumb that "hitchhiking of the neighboring neutral locus is efficient if $r < s$ and becomes negligible if $r \approx s$," see e.g. Nurminsky (2001).

Proposition 1 concerns the effect of a sweep on a single lineage. As Kaplan, Hudson, and Langley (1989) observed in their equation (16), the heterozygosity (i.e., the probability two randomly chosen individuals differ at the $A/a$ locus) after the sweep, $H_\infty$, is related to that before the sweep, $H_0$, by

$$H_\infty / H_0 = p_{22}$$

where $p_{22}$ is the probability that two lineages sampled from the $B$ population at time $\tau$ are distinct at time 0. As Stephan, Wiehe and Lenz (1992) observed, see their formula (14a), the reduction in heterozygosity can be approximated for large $N$ by

$$p_{22} \approx 1 - (1 - Q_1(\tau))^2$$

This formula comes from the fact that, for large $N$, $1 - p_{22}$ is approximately the probability that both lineages get trapped in the $B$ population, and these events are approximately independent for large $N$.

Kaplan, Hudson, and Langley (1989), see page 892, developed numerical methods for computing the probabilities $p_{k,j}$ that $k$ lineages at the end of a selective sweep have $j$ distinct ancestors at the beginning of the sweep. Our next result extends Stephan, Wiehe and Lenz's observation to samples of size $k$. The reader should note that we are considering the case of

strong selection, where for example $s$ is held fixed or goes to 0 slowly, which is much different from the usual diffusion limit in which $2Ns$ and $2Nr$ tend to limits.

**Proposition 2.** *For the logistic sweep model, if $N \to \infty$ with $r\ln(2N)/s \to a$ and $s(\ln N)^2 \to \infty$ then for $j \geq 2$*

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

In words, the number of lineages is reduced to $k-j+1$ if $j$ lineages are trapped in the $B$ population and these events become independent as $N$ gets large. The restriction to $j \geq 2$ in the formula above comes from the fact that the number of lineages does not change if the number of trapped lineages is 0 or 1. It follows from Proposition 2 that $p_{k,k} \to (1-p)^k + kp(1-p)^{k-1}$.

## SIMULATIONS

To evaluate the quality of the approximation provided by Proposition 2, we will use simulation and numerical computation. We are interested in the probabilities of five events associated with a single selective sweep: a lineage escapes the sweep (pinb), two lineages both escape the sweep and do not coalesce (p2inb), two lineages coalesce and end up in $b$ (p2cinb), exactly one of the two lineages escapes the sweep (p1B1b), and lineages end up coalesced in B (p2inB). These can be computed for the logistic sweep model by numerically integrating the associated differential equations.

We will compare our results for the logistic sweep model to those for the Moran model. In our formulation of that model, we assume that the relative fitnesses of $B$ and $b$ are 1 and $1-s$. Rather than assuming that the fraction of chromosomes with the $B$ allele deterministically follows the logistic curve, we allow the number of $B$ chromosomes to be random. Following the dynamics of the Moran model with selection, potnetial replacements occur at times of a rate $2N$ Poisson processes. We pick an individual to be replaced, and another individual

(possibly the same as the first) to be the parent of the new individual. If a $b$ is proposed to replace a $B$ then to account for the selective advantage of $B$ that replacement only occurs with probability $1-s$. This leads to the following transition rates:

| a chromosome with | is replaced by one with | with probability |
|---|---|---|
| $B$ | $B$ | $(k/2N)(k/2N)$ |
| $B$ | $b$ | $(k/2N)(1-k/2N)(1-s)$ |
| $b$ | $B$ | $(1-k/2N)(k/2N)$ |
| $b$ | $b$ | $(1-k/2N)(1-k/2N)$ |

and nothing happens with probability $(k/2N)(1-k/2N)s$.

To simulate the Moran model it is sufficient to simulate the embedded Markov chain $X_n$, i.e., the sequence of states it visits when the successive replacements are made. There is no reason to generate the exponential waiting times between jumps. Kaplan, Hudson, and Langley (1989) do their simulations (see page 889) by producing a large number of sweeps and only keeping the successful ones. Since the probability of a successful sweep is approximately $s$, this is rather inefficient. We avoid this problem by considering the chain conditional on the event of fixation, $F$. The first step is to note that

$$h(x) = \frac{1-(1-s)^x}{1-(1-s)^{2N}}$$

is the probability that fixation of $B$ will occur when there are $x$ chromosomes with $B$ (see e.g., Durrett (2002), (1.2) on page 118). The Markov property implies that if $p(x,y)$ is the transition probability for $X_n$ and $P_x$ denotes the probability distribution of the Markov chain started at $x$ then

$$P_x(X_1 = y|F) = \frac{P_x(\{X_1 = y\} \cap F)}{P_x(F)} = \frac{p(x,y)P_y(F)}{P_x(F)} = \frac{p(x,y)h(y)}{h(x)}$$

so the last formula gives the transition probability for the conditioned chain. Note that $h(x+1)/h(x-1) > 1$ so this conditioning causes the number of $B$ individuals in the Moran model initially to rise faster than in the logistic sweep model. See Figure 2 for the results of

one simulation with $N = 10^4$ and $s = 0.1$. This effect is known and its magnitude has been estimated by Barton (1998). On page 125 he says that the "expected frequency of an allele destined for fixation is accelerated by a factor $1/2s$ relative to that expected in the absence of stochastic effects."

Table 1 gives results of a number of simulations. For the moment we will concentrate on the first two groups of results which both have $N = 10^4$ and $s = 0.1$, but have different recombination probabilities $r = 0.001064$ and $r = 0.005158$, these being chosen so that the approximations from Proposition 1 of the probability that a lineage ends in b are 0.1 and 0.4 respectively. The approximation provided by Proposition 1 is close to the exact value of pinb for the logistic sweep model. Turning to the predictions of Proposition 2, we see that while that approximation says there is no chance that the two lineages will end up coalesced in b, this has probabilities 0.034 and 0.096 under the logistic sweep model. Since p2cinb is underestimated, it should be no surprise that p2inb is overestimated. In both cases, p1B1b is also overestimated, leading to predictions of $p_{22}$ of 0.19 and 0.64, compared with the values of 0.1239 and 0.4646 for the logistic sweep model.

The third line of the results has more bad news: the logistic sweep model is not a very accurate approximation of the Moran model. The values of pinb for the logistic sweep model differ by 20% from those in the Moran model. The reason for the discrepancy in estimating pinb can be understood by plotting the probability of a lineage ending up in b versus the time it takes for the population with the advantageous allele to reach 1000 chromosomes in the Moran model. As Figure 3 shows there is a strong correlation between these two quantities. When 1000 chromosomes is reached quickly there is less time for a lineage to escape from the sweep, and there are fewer recombinations in the late stages of the process that produce two coalesced lineages in $b$. Thus, the randomness in the size of the $B$ population at the beginning of the sweep can cause significant variability in the fraction of lineages that end up in the $b$ population. Since the logistic sweep model ignores any randomness in the size of

the $B$ population, it is unable to provide a good approximation of the Moran model.

The values of p2cinb for the logistic sweep model of 0.034 and 0.096 are almost twice the values in the Moran model of 0.018 and 0.055. Again this comes from the fact that the initial growth of the $B$ population is faster than predicted by the logistic curve so when we work backwards there is less time for coalesced lineages to escape at the last minute. Turning to the last column, we see some surprising good news: values of $p_{22}$ predicted by the logistic sweep model never differ by more than 2.2% from the value for the Moran model. This is consistent with Kaplan, Hudson, and Langley (1989) who report that simulation results for $p_{22}$ differ from the analytical ones by at most 2%. However from the rest of the table we can see this accuracy is a result of fortuitous cancellations which in one case combine a 66% overestimate of p2cinb with a 14% underestimate of p1B1b to obtain a result for $p_{22}$ with a 1.4% error.

## A BETTER APPROXIMATION

A significant problem with the approximation provided by Proposition 2 is that it predicts that two lineages will never end up coalesced in the $b$ population. Our simulation results for the Moran model show that this indeed does occur, typically when two lineages coalesce while in the $B$ population and then recombine into the $b$ population. In this section, we explain an improved approximation that allows for this possibility. Our approach is quite different from Barton's which is based on differential equations, and gives a simple explicit approximation for the effect of a sweep on the genealogy of a sample of size $k$.

We will describe the genealogy of $k$ lineages during a sweep by using a random partition of $\{1, \ldots, k\}$. The integers $i$ and $j$ will be in the same block of the partition if and only if the $i$th and $j$th lineages in the sample coalesce. In the approximation of Proposition 2, this partition has only one block with more than one integer, because only the lineages that get trapped in the $B$ population coalesce. Here we will allow multiple blocks to have more than

one integer.

To define the random partition, we will use a stick-breaking construction, which is essentially the same as the paintbox construction of Kingman (1978). The construction is simple to describe, but it will take us a while to explain the intuition behind it. The ingredients for the construction are as follows:

- Let $M = \lfloor 2Ns \rfloor$, where $\lfloor m \rfloor$ denotes the greatest integer less than or equal to $m$.
- Let $\xi_l$, $2 \le l \le M$ be independent Bernoulli random variables that are 1 with probability $r/s$ and 0 otherwise.
- Let $W_l$, $2 \le l \le M$ be independent random variables with $W_l$ having a beta$(1, l-1)$ distribution.
- For $2 \le l \le M$, let $V_l = \xi_l W_l$, and let $T_l = V_l \prod_{i=l+1}^{M} (1 - V_i)$.
- Let $T_1 = \prod_{l=2}^{M} (1 - V_l)$.

Now, divide the interval $[0, 1]$ into $M$ subintervals (some of which may be empty) as follows. Let $a_{M+1} = 1$ and for $1 \le l \le M$, let $a_l = a_{l+1} - T_l$. Since $\sum_{l=1}^{M} T_l = 1$, we have $a_1 = 0$. Let $I_l = [a_l, a_{l+1}]$. To obtain a partition of $\{1, \ldots, k\}$, let $U_1, \ldots, U_k$ be i.i.d. random variables with a uniform distribution on $[0, 1]$. We declare $i$ and $j$ to be in the same block of the partition if and only if $U_i$ and $U_j$ are both in the interval $I_l$ for some $l$.

Since we wish also to keep track of which lineages are descended from the $B$ population and which come from the $b$ population, we will mark, with probability $s/(r(1-s)+s)$, the block of the partition containing all of the $i$ such that $U_i$ is in $I_1$ to indicate that these lineages did not escape from the sweep.

Our approximation is based on the following ideas:

- We ignore the possibility that a lineage experiences two recombinations during the sweep, taking it from the $B$ population to the $b$ population and back to the $B$ population.

- When the number of chromosomes with the $B$ allele is much smaller than the population size, the number of individuals with the $B$ allele can be approximated by a continuous-time branching process, in which each individual splits into two at rate 1 and dies at rate 1-s.

- It is known, see for example O'Connell (1993), that the lineages in a branching process that do not die out are themselves a branching process. In our case the lineages that don't die out are a Yule process, a continuous time branching process in which each particle splits into two at rate $s$. Since each lineage has an infinite line of descent with probability $s$, the number of such lineages at the end of the sweep is approximately $M = \lfloor 2Ns \rfloor$.

- When there are $l \geq 2$ lineages in the Yule process, the time to the next birth is exponentially distributed with mean $1/sl$, and recombinations occur at rate $lr$, so the expected number of recombination events is $r/s$. We assume that the number of such events is always 0 or 1. The Bernoulli variables $\xi_l$, $2 \leq l \leq M$ tell us whether one occurs or not.

- In the first period the probability a recombination happens before the first birth is $r(1-s)/(r(1-s)+s)$. In this case, no lineage for the neutral locus comes from the $B$ population.

- As time tends to infinity the number of individuals in the Yule process divided by its mean converges to an exponential distribution with mean 1. (See e.g., Joyce and Tavaré 1987.) This implies that when there are $l$ lineages, the fraction of individuals

at the end of the sweep that are descendants of a given lineage has roughly the same distribution as $\xi_1/(\xi_1 + \cdots + \xi_l)$ where $\xi_i$ are independent exponentails. The ratio has a beta$(1, l-1)$ distribution. Thus, the $W_l$, $2 \le l \le M$ represent the fraction of descendants of an individual in the Yule process when there are $l$ individuals.

- If $\xi_l = 0$ there is no recombination and $V_l = 0$. If $\xi_l = 1$ there is a recombination that removes a fraction $V_l$ of the remaining population, i.e., the fraction of individuals that recombine at time $l$ is $T_l$. $T_1 = \prod_{l=2}^{M}(1 - V_l)$ is the fraction of the initial population that trace their ancestry back to the $B$ population at the time when there was one lineage with an infinite line of descent.

Turning to the definition of the partition, the integers $i$ such that $U_i$ is in $I_l$ correspond to the lineages in the sample that recombine when there are $l$ members of the $B$ population with an infinite line of descent. If $U_i$ and $U_j$ are both in the interval $I_l$ then they have the same parent so they belong in the same block.

The procedure described above translates directly into a procedure for simulating genealogies. However, in the case of one or two lineages one can compute the probabilities of interest analytically.

**Proposition 3.** *For the above approximation, we have*

$$\begin{aligned}
\text{pinB} &= \frac{s}{r(1-s)+s}\prod_{l=2}^{M}\left(1-\frac{r}{sl}\right)\\
\text{p2inB} &= \frac{s}{r(1-s)+s}\prod_{l=2}^{M}\left(1-\frac{2r}{s(l+1)}\right)\\
\text{p2cinb} &= \frac{r(1-s)}{r(1-s)+s}\prod_{l=2}^{M}\left(1-\frac{2r}{s(l+1)}\right)+\sum_{i=2}^{M}\frac{2r}{sl(l+1)}\prod_{l=i+1}^{M}\left(1-\frac{2r}{s(l+1)}\right)
\end{aligned}$$

From pinB, p2inB, and p2cinb, we can calculate the remaining quantities of interest. Note, in particular, that pinb $= 1 -$ pinB, and since pinB $=$ p2inB $+ \frac{1}{2}$p1B1b, we have p1B1b $=$

2(pinB − p2inB). Finally, p2inb = 1 − p1B1b − p2inB − p2cinb.

The expressions in Proposition 3 are exact but can be simplified without much loss of accuracy by using $1 - x \approx e^{-x}$. Consider for example pinB. Rewriting the first fraction and dropping the $r(1-s)$ from the denominator,

$$\begin{aligned}\text{pinB} &= \left(1 - \frac{r(1-s)}{r(1-s)+s}\right)\prod_{l=2}^{M}\left(1-\frac{r}{sl}\right) \approx \exp\left(-\frac{r(1-s)}{s} - \sum_{l=2}^{M}\frac{r}{sl}\right)\\ &= \exp\left(-\frac{r}{s}\left[-s + \sum_{l=1}^{M}\frac{1}{l}\right]\right) \approx \exp\left(-\frac{r}{s}\left[\ln(2Ns) + \gamma - s\right]\right)\end{aligned}$$

where $\gamma = \lim_{k\to\infty}\sum_{j=1}^{k}\frac{1}{j} - \ln k \approx 0.57721$ is Euler's constant.

The approximation we have just described is somewhat complicated and relies on a number of simplifications. In the companion paper Schweinsberg and Durrett (2003), we investigate its properties mathematically. Here we content ourselves to demonstrate by simulation that it works quite well. We have considered five combinations of population size $n$ and selective advantage $s$ of the newly introduced allele. For each we have chosen values of the recombination rate $r$ to make the value given by Proposition 1 for probability a lineage escapes the sweep equal to 0.1 and to 0.4. In all cases the coalescent with simultaneous multiple collisions provides an excellent approximation to the results for the Moran model.

The worst result that occurs for Proposition 3 is the 12% relative error in the approximation of p2cinb for the case $N = 10^6$, $s = 0.01$, $r = 7.3 \times 10^{-5}$. However, in this case individual replicates take 6 hours so our estimates of the value for the Moran model are based on only 100 simulations. Also, this is a relative error and the probability being estimated is small, so the absolute error is only 0.00167. In contrast the relative errors from the logistic approximation range from 75-183% for this quantity. The easy quantity to estimate is pinb since it involves only one lineage. The relative errors from Proposition 3 range from 0.1-1.8% while those from the logistic sweep model range from 18-47%.

To verify that we are computing things correctly for the logistic sweep model we have

considered the case treated in Table 1 on page 245 in Stephan, Wiehe, and Lenz (1992): $N = 10^8$, $s = 0.001$, and various values of $r$. The results are recorded in our Table 2. Readers who look at Table 1 in Stephan, Wiehe, and Lenz (1992) will note that they have $2N = 10^8$. The difference in population size comes from the fact that we are considering the Moran model while Stephan, Wiehe, and Lenz (SWL) and Kaplan, Hudson, and Langley (KHL) used the Wright-Fisher model. A second difference is that KHL and SWL only follow the logistic from the time that the frequency of the favored allele is $\epsilon = 10^{-6}$ to the time that it is $1 - \epsilon$. Thus we have done our Euler method for this set-up as well as our usual choice of $\epsilon = 1/2N$. Finally the fifth row gives the approximation that comes from Proposition 3. The five sets of numbers agree remarkably well. However, as the last four rows show the logistic with $\epsilon = 1/2N$ and the Proposition 3 approximation arrive at similar answers by giving very different values to p2inb and p1B1b. The pattern of the logistic overestimating p2inb and underestimating p1B1b is the same as in the simulations in Table 1.

## CONCLUSIONS

The evaluation of probabilities associated with a selective sweep via simulation is time consuming for large populations. Here we have shown that a stick breaking construction provides an easily computed and accurate approximation. The simulations we have used to investigate our approximation have shown that the common practice of using the logistic curve to model population size changes during the sweep, and ignoring the randomness in the size of the $B$ population at the beginning of the sweep, leads to substantial errors in the approximation of some probabilities. Remarkably, these errors approximately cancel out when the probability two lineages are not coalesced during the sweep, $p_{22}$, is computed.

## ACKNOWLEDGEMENTS

The authors would like to thank Yuseob Kim for calling our attention to Barton (1998) and for several useful comments. An anonymous referee read the manuscript thoroughly and

helped to clarify the presentation. R.D. was partially supported by NSF grants from the probability program (0202935) and from a joint DMS/NIGMS initiative to support research in mathematical biology (0201037). J.S. was supported by an NSF Postdoctoral Fellowship.

## LITERATURE CITED

Barton, N. H. 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. Camb. **72:** 123–133

Durrett, R. 2002 Probability Models for DNA Sequence Evolution. Springer-Verlag, New York.

Joyce, P., and S. Tavaré, 1987 Cycles, permutations, and the structure of the Yule process with immigration. Stoch. Proc. Appl. **25:** 309–314

Kaplan, N .L., R. R. Hudson, and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899

Kingman, J. F. C., 1978 The representation of partition structures. J. London Math. Soc. **18:** 374–380

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. Camb. **23:** 23–35

Nurminsky, D.I., 2001 Genes in sweeping competition. Cell. Mol. Life Sci. **58:** 125–134

O'Connell, N., 1993 Yule process approximation for the skeleton of a branching process. J. Appl. Prob. **30:** 725–729

Przeworski, M. (2002) The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189

Schweinsberg, J. and R. Durrett, 2003. Random partitions approximating the coalescence of lineages during a selective sweep. Preprint available at www.math.cornell.edu/~durrett

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, C.F. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429

Stephan, W., T. Wiehe, T., and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor. Pop. Biol. **41:** 237–254

Wiehe, T., and W. Stephan, 1993 Analysis of genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–854

## APPENDIX: PROOFS

**Proof of Proposition 1.** By (4) the probability a lineage escapes from the selective sweep is

$$Q_1(\tau) = (1 - 1/2N) \int_0^\tau \frac{re^{-rt}}{(1 - 1/2N) + (1/2N)e^{st}}\, dt$$

Writing $\sim$ to denote that the ratio of the two sides tends to 1 as $N \to \infty$, we have $r \sim as/\ln(2N)$ and therefore

$$Q_1(\tau) \sim \int_0^\tau \frac{\frac{as}{\ln(2N)} e^{-ast/\ln(2N)}}{(1 - 1/2N) + (1/2N)e^{st}}\, dt$$

Changing variables $u = st/\ln(2N)$, $du = sdt/\ln(2N)$ and recalling $t = 2\ln(2N)/s \sim \tau$ corresponds to $u = 2$, the above becomes

$$Q_1(\tau) \sim \int_0^2 \frac{ae^{-au}}{(1 - 1/2N) + (2N)^{u-1}}\, du$$

When $u > 1$ the denominator tends to $\infty$, and when $u < 1$ it approaches 1. Therefore,

$$\lim_{N\to\infty} Q_1(\tau) = \int_0^1 ae^{-au}\, du = 1 - e^{-a}$$ □

**Proof of Proposition 2.** Supposing that the sweep takes place between times 0 and $\tau = (2/s)\ln(2N - 1)$, let $\sigma = (2/s)\ln\ln(2N)$. We will argue that working backwards from time $\tau$ the lineages that are in the $B$ population at time $\sigma$ do not escape and all coalesce between times 0 and $\sigma$, while none of the ones that are in the $b$ population at time $\sigma$ coalesce. From (3) with $p(0) = 1/2N$, we get

$$p(t) = \frac{1}{1 + (2N - 1)e^{-st}}$$

There are $p(t)N$ lineages in the $B$ population at time $t$, so the probability that two lineages in the $B$ population coalesce between times $\sigma$ and $\tau$ is at most

$$\begin{aligned}\int_\sigma^\tau \frac{1}{p(t)N}\, dt &= \frac{1}{N}\int_\sigma^\tau 1 + (2N - 1)e^{-st}\, dt \\ &\le \frac{1}{N}\left(\tau + \frac{2N - 1}{s}e^{-s\sigma}\right) \le \frac{2\ln 2N}{sN} + \frac{2}{s(\ln 2N)^2}\end{aligned}$$

Thus, we can ignore coalescence in the $B$ population during the time interval $[\sigma, \tau]$. By a symmetric argument, we can ignore coalescence in the $b$ population in the time interval $[0, \tau - \sigma]$.

The expected number of recombinations between times 0 and $\sigma$ for a sample of size $k$, regardless of whether the lineages are in the $B$ or $b$ population, is at most

$$kr\sigma \sim ka\frac{s}{\ln 2N} \cdot \frac{2\ln\ln(2N)}{s}$$

Therefore, we can ignore recombinations during $[0, \sigma]$ and, by the same argument, we can ignore recombinations during $[\tau - \sigma, \tau]$. It follows from these observations that we may ignore the possibility that two lineages may coalesce in the $B$ population and then recombine into the $b$ population, as well as the possibility that two lineages may both recombine, and then coalesce in the $b$ population. Thus, in the limit, the lineages that coalesce are precisely those that get trapped in the $B$ population. Furthermore, it follows from these observations that the events that different lineages get trapped in the $B$ population are approximately independent. Therefore, if $p$ is the probability that a lineage escapes the sweep, the probability that exactly $j$ lineages get trapped is approximately

$$\binom{k}{j} p^j (1-p)^{k-j}$$

Combining this result with Proposition 1 gives Proposition 2. □

**Proof of Proposition 3.** Since $W_l$ has a beta$(1, l-1)$ distribution, which has density function $(l-1)(1-x)^{l-2}$, integration shows that $E[W_l] = 1/l$ and $E[W_l^2] = 2/l(l+1)$. To calculate pinB, first note that pinB $= [s/(r(1-s)+s)]P(U_1 \in I_1)$ where $U_1$ is uniform on $(0, 1)$. If $U_1$ is not in any of the intervals $I_{l+1}, \ldots, I_M$, then the probability, conditional on $V_l$, that $U_1 \in I_l$ is $V_l$. Therefore, for $2 \le l \le M$, we have

$$P(U_1 \in I_l | U_1 \notin I_{l+1} \cup \cdots \cup I_M) = E[V_l] = E[\xi_l]E[W_l] = \frac{r}{sl}$$

It follows that

$$P(U_1 \in I_1) = \prod_{l=2}^{M} P(U_1 \notin I_l | U_1 \notin I_{l+1} \cup \cdots \cup I_M) = \prod_{l=2}^{M} \left(1 - \frac{r}{sl}\right)$$

which implies the first statement of Proposition 3.

Next, note that if $U_2$ is independent of $U_1$ and uniform on $(0,1)$, then

$$\text{p2inB} = \frac{s}{(r(1-s)+s)} P(U_1 \in I_1 \text{ and } U_2 \in I_1)$$

If $U_1$ and $U_2$ are not in any of the intervals $I_{l+1}, \ldots, I_M$, then the probability, conditional on $V_l$, that either $U_1$ or $U_2$ is in $I_l$ is $1-(1-V_l)^2$. A little calculation shows

$$E[1-(1-V_l)^2] = 2E[V_l] - E[V_l^2] = E[\xi_l](2E[W_l] - E[W_l^2]) = \frac{r}{s}\left(\frac{2}{l} - \frac{2}{l(l+1)}\right) = \frac{2r}{s(l+1)}$$

and the formula for p2inB now follows by the same reasoning as the formula for pinB.

Finally, to obtain the formula for p2cinb, we note that

$$\text{p2cinb} = \frac{r(1-s)}{r(1-s)+s} P(U_1 \in I_1 \text{ and } U_2 \in I_1) + \sum_{i=2}^{M} P(U_1 \in I_i \text{ and } U_2 \in I_i).$$

From the calculation for p2inB, we know that the probability that $U_1$ and $U_2$ are not in any of the intervals $I_{i+1}, \ldots, I_M$ is

$$\prod_{l=i+1}^{M} \left(1 - \frac{2r}{s(l+1)}\right)$$

This formula when $i = 1$ gives $P(U_1 \in I_1 \text{ and } U_2 \in I_1)$. Conditional on the event that $U_1$ and $U_2$ are not in any of the intervals $I_{i+1}, \ldots, I_M$, the probability that $U_1$ and $U_2$ are both in $I_i$ is $E[V_i^2] = E[\xi_i]E[W_i^2] = 2r/[sl(l+1)]$. By combining these observations, we obtain the desired formula for p2cinb. □

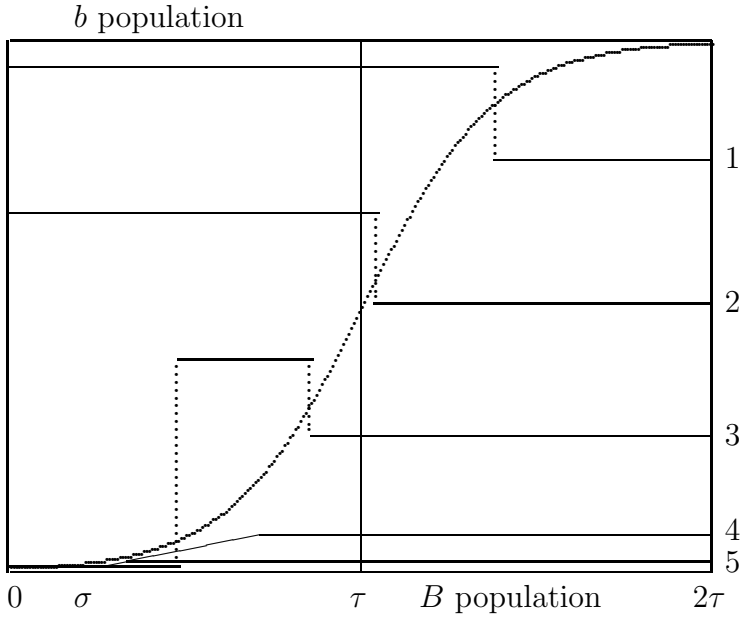


Figure 1. Hypothetical genealogy of a sample of size five. The curve is the solution of the logistic differential equation. Dotted lines mark recombination events that change the state of the selected locus. Lineages 1 and 2 undergo recombination once and escape the selective sweep. Lineage 3 recombines twice returning to the $B$ population where it coalesces with the genealogies 4 and 5 that did not experience recombination.

| | pinb | p2inb | p2cinb | p1B1b | $p_{22}$ |
|---|---|---|---|---|---|
| | $N = 10^4$ | $s = 0.1$ | $r = 1.064 \times 10^{-3}$ | | |
| Prop. 2 | 0.1 | 0.01 | 0 | 0.18 | 0.19 |
| Logistic | 0.09983(21%) | 0.00845(36%) | 0.03365(84%) | 0.11544(0.3%) | 0.12390(2.1%) |
| Moran | 0.08203(11) | 0.00620(2) | 0.01826(8) | 0.11513(6) | 0.12134 |
| Prop. 3 | 0.08235(0.4%) | 0.00627(1.1%) | 0.01765(-3.4%) | 0.11687(1.5%) | 0.12314(1.5%) |
| | $N = 10^4$ | $s = 0.1$ | $r = 5.158 \times 10^{-3}$ | | |
| Prop. 2 | 0.4 | 0.16 | 0 | 0.48 | 0.64 |
| Logistic | 0.39936(18%) | 0.13814(31%) | 0.09599(75%) | 0.32646(-7.3%) | 0.46460(1.5%) |
| Moran | 0.33656(38) | 0.10567(21) | 0.05488(23) | 0.35201(16) | 0.45769 |
| Prop. 3 | 0.34065 (1.2%) | 0.10911(3.2%) | 0.05100 (-7.1%) | 0.36112 (2.6%) | 0.47023(2.7%) |
| | $N = 10^4$ | $s = 0.03$ | $r = 3.192 \times 10^{-4}$ | | |
| Logistic | 0.09983(41%) | 0.00723(64%) | 0.04677(130%) | 0.09164(-1.1%) | 0.09888(1.8%) |
| Moran | 0.07099(12) | 0.00440(1) | 0.02026(10) | 0.09265(7) | 0.09706 |
| Prop. 3 | 0.07121(0.3%) | 0.00452(2.7%) | 0.01873(-7.6%) | 0.09592(3.5%) | 0.10044(3.5%) |
| | $N = 10^4$ | $s = 0.03$ | $r = 1.547 \times 10^{-3}$ | | |
| Logistic | 0.39763(35%) | 0.12025(55%) | 0.14497(117%) | 0.26427(-13%) | 0.38453(1.2%) |
| Moran | 0.29546(43) | 0.07734(19) | 0.06678(28) | 0.30266(17) | 0.38001 |
| Prop. 3 | 0.30084(1.8%) | 0.08238(6.5%) | 0.05945 (-11%) | 0.31803 (5.1%) | 0.40041(5.4%) |

| | | | | | |
|---|---|---|---|---|---|
| | $N = 10^5$ | $s = 0.03$ | $r = 2.590 \times 10^{-4}$ | | |
| Logistic | 0.09989(30%) | 0.00837(51%) | 0.03034(133%) | 0.11097(-0.46%) | 0.11935(2.0%) |
| Moran | 0.07675(33) | 0.00554(4) | 0.01545(25) | 0.11149(18) | 0.11704 |
| Prop. 3 | 0.07671(-0.1%) | 0.00553(-0.1%) | 0.01494(-3.4%) | 0.11249(-0.9%) | 0.11803(0.8%) |
| | $N = 10^5$ | $s = 0.03$ | $r = 1.256 \times 10^{-3}$ | | |
| Logistic | 0.39826(25%) | 0.13808(43%) | 0.10204(123%) | 0.31627(-10%) | 0.45437(1.2%) |
| Moran | 0.31846(106) | 0.09641(57) | 0.04581(64) | 0.35246(46) | 0.44888 |
| Prop. 3 | 0.32074(0.7%) | 0.09790(1.5%) | 0.04409 (-3.8%) | 0.35750 (1.4%) | 0.45540(1.5%) |
| | $N = 10^5$ | $s = 0.01$ | $r = 8.632 \times 10^{-5}$ | | |
| Logistic | 0.09989(47%) | 0.00753(78%) | 0.04531(183%) | 0.09409(-1.2%) | 0.10162(2.2%) |
| Moran | 0.06783(33) | 0.00422(4) | 0.01599(26) | 0.09524(18) | 0.09947 |
| Prop. 3 | 0.06807(0.3%) | 0.00426(0.9%) | 0.01537(-3.9%) | 0.09687(1.7%) | 0.10113(1.7%) |
| | $N = 10^5$ | $s = 0.01$ | $r = 4.185 \times 10^{-4}$ | | |
| Logistic | 0.39826(39%) | 0.12595(66%) | 0.13616(164%) | 0.27230(-14%) | 0.39826(1.4%) |
| Moran | 0.28575(118) | 0.07580(55) | 0.05158(77) | 0.31674(48) | 0.39255 |
| Prop. 3 | 0.28935(1.3%) | 0.07822(3.2%) | 0.04875(-5.5%) | 0.32478(2.5%) | 0.40300(2.6%) |
| | $N = 10^6$ | $s = 0.01$ | $r = 7.262 \times 10^{-5}$ | | |
| Logistic | 0.09992(34%) | 0.00835(59%) | 0.03709(163%) | 0.10898(-0.6%) | 0.11733(2.1%) |
| Moran | 0.07447(96) | 0.00526(12) | 0.01440(77) | 0.10962(47) | 0.11489 |
| Prop. 3 | 0.07321(-1.7%) | 0.00510(-3.1%) | 0.01273(-12%) | 0.11075(1.0%) | 0.11585(0.8%) |

| | $N = 10^6$ | $s = 0.01$ | $r = 3.521 \times 10^{-4}$ | | |
|---|---|---|---|---|---|
| Logistic | 0.39877(30%) | 0.13823(52%) | 0.10485(169%) | 0.31190(-12%) | 0.45015(0.9%) |
| Moran | 0.30752(318) | 0.09106(168) | 0.03895(200) | 0.35502(136) | 0.44608 |
| Prop. 3 | 0.30821(0.2%) | 0.09135(0.3%) | 0.03798 (-2.5%) | 0.35776 (0.8%) | 0.44911(0.7%) |

Table 1. We have considered five combinations of population size $n$ and selective advantage $s$ of the newly introduced allele. For each we have chosen values of the recombination rate $r$ to make the value given by Proposition 1 for the probability that a lineage escapes the sweep equal to 0.1 and to 0.4. Here we investigate five quantities: the probability a lineage escapes the sweep (pinb), the probability two lineages both escape the sweep and do not coalesce (p2inb), the probability both escape and coalesce (p2cinb), the probability exactly one of the two lineages escapes the sweep (p1B1b), and the probability that the two lineages do not coalesce ($p_{22}$). Note that p2inB is not given but can be computed by taking 1 - p2inb - p2cinb - p1B1b. Also, note that $p_{22}$ = p2inb + p1B1b. The rows marked Moran give estimates based on simulation with the numbers in parentheses being $10^5$ times the standard deviation of the estimate. 10,000 simulations were used when $N = 10^4$, 1,000 for $N = 10^5$ and 100 for $N = 10^6$. The rows marked Logistic and Proposition 3 give the approximations that come from the logistic sweep model and Proposition 3. The numbers in parentheses give the relative percentage error (i.e., the difference between the approximation and the value for the Moran model divided by the value for the Moran model and then multiplied by 100).

| $-\log_{10}(r/s)$ | 1.6 | 1.8 | 2.0 | 2.2 | 2.4 | 2.6 |
|---|---|---|---|---|---|---|
| KHL | 0.419731 | 0.291868 | 0.195660 | 0.128580 | 0.083376 | 0.053477 |
| SWL | 0.420186 | 0.291707 | 0.195883 | 0.128650 | 0.083279 | 0.053409 |
| $\epsilon = 10^{-6}$ | 0.420186 | 0.291707 | 0.195881 | 0.128648 | 0.083278 | 0.053408 |
| $\epsilon = 1/2N$ | 0.420158 | 0.291689 | 0.195870 | 0.128641 | 0.083273 | 0.053405 |
| Prop. 3 | 0.417977 | 0.289382 | 0.193937 | 0.127199 | 0.082265 | 0.052727 |
| p2inb $\epsilon = 1/2N$ | 0.124168 | 0.057335 | 0.025114 | 0.010627 | 0.004398 | 0.001795 |
| p2inb Prop. 3 | 0.073382 | 0.032632 | 0.013946 | 0.005809 | 0.002380 | 0.000965 |
| p1B1b $\epsilon = 1/2N$ | 0.295990 | 0.234354 | 0.170757 | 0.118014 | 0.078875 | 0.051610 |
| p1B1b Prop. 3 | 0.344595 | 0.256750 | 0.179991 | 0.121390 | 0.079885 | 0.051763 |

Table 2. Comparison of results for a population of size $10^8$ with $s = 0.001$ for various values of the recombination rate $r$. In the top half of the table we give values of $p_{22}$ computed by various methods. The first two rows are simulation results of Kaplan, Hudson, and Langley (1989) and numerical integration of the differential equations by Stephan, Wiehe, and Lenz (1992) using a fourth-order Runge-Kutta method. These authors only follow the sweep from an initial frequency of $B$s of $10^{-6}$ so in the third and fourth rows we give results from our Euler method for that initial density and for our usual choice of $\epsilon = 1/2N$. The fifth row gives the approximation that comes from Proposition 3. The five sets of numbers agree remarkably well. However, as the last four rows show the logistic with $\epsilon = 1/2N$ and approximation in Proposition 3 arrive at similar answers by giving very different values to p2inb and p1B1b.

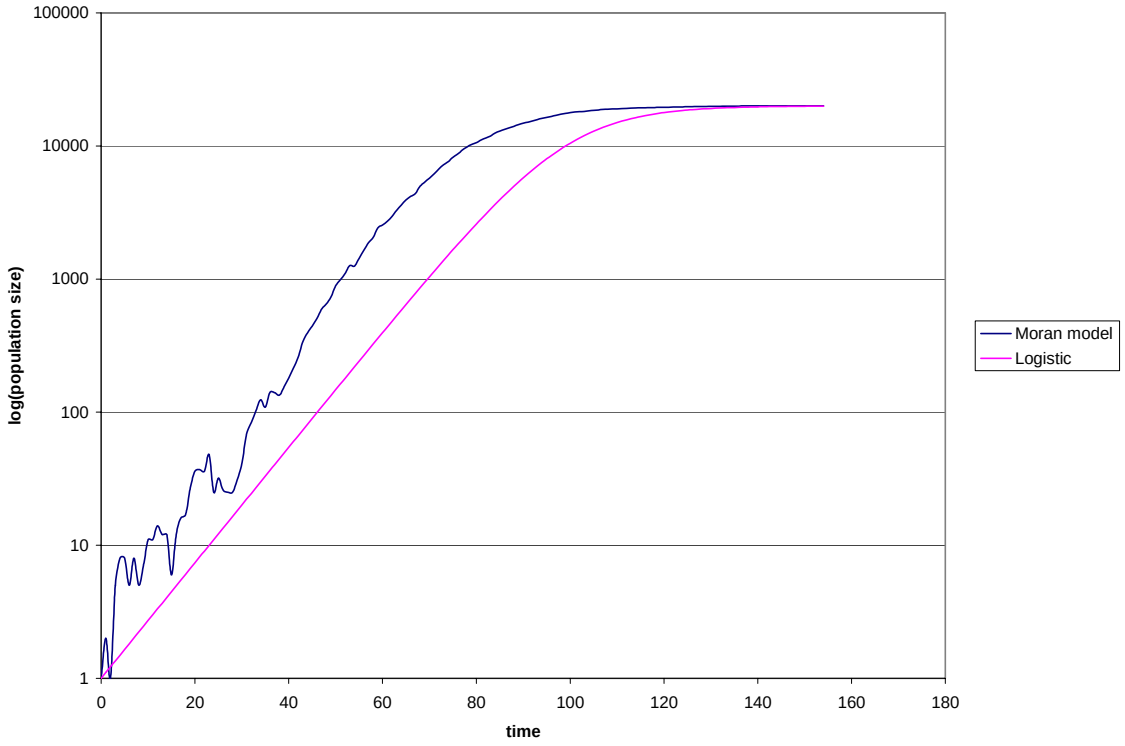


Figure 2. Number of chromosomes with the favorable allele in the Moran model compared to the logistic sweep model.

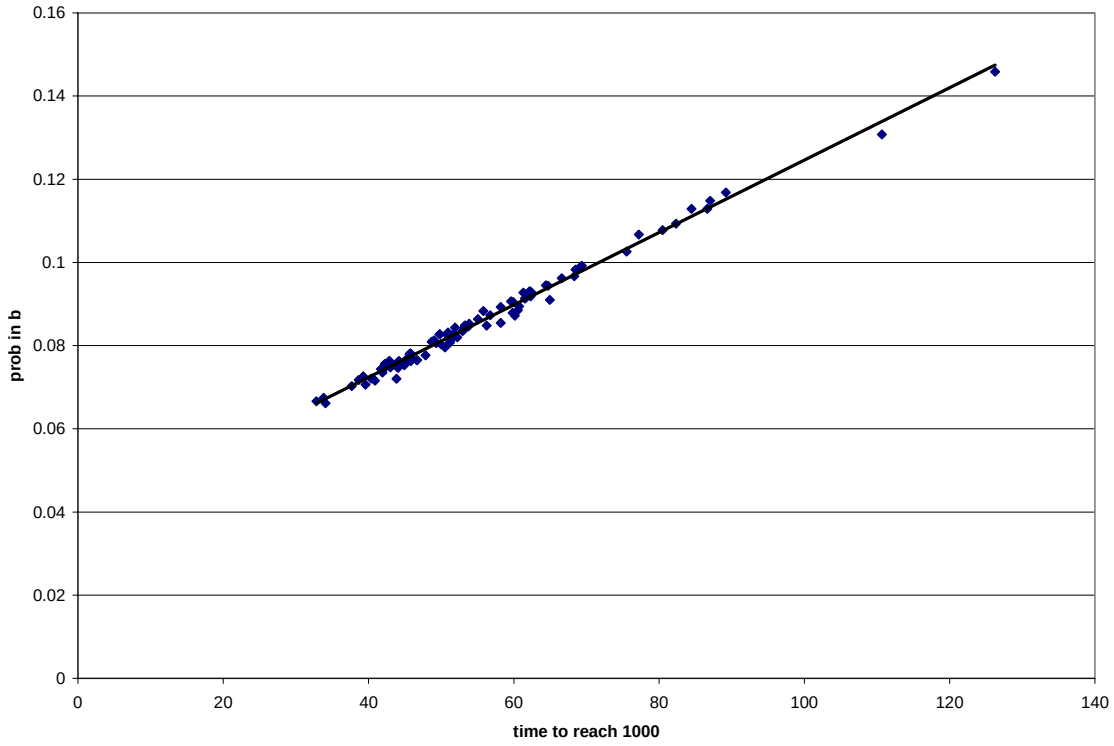


Figure 3. Probability of ending in the b population compared to the time to reach 1000 chromosomes with the favored B allele.