The Annals of Applied Probability
2005, Vol. 15, No. 3, 1591–1651
DOI 10.1214/105051605000000430
© Institute of Mathematical Statistics, 2005

RANDOM PARTITIONS APPROXIMATING THE COALESCENCE OF LINEAGES DURING A SELECTIVE SWEEP

BY JASON SCHWEINSBERG[1] AND RICK DURRETT[2]

University of California, San Diego and Cornell University

When a beneficial mutation occurs in a population, the new, favored allele may spread to the entire population. This process is known as a selective sweep. Suppose we sample n individuals at the end of a selective sweep. If we focus on a site on the chromosome that is close to the location of the beneficial mutation, then many of the lineages will likely be descended from the individual that had the beneficial mutation, while others will be descended from a different individual because of recombination between the two sites. We introduce two approximations for the effect of a selective sweep. The first one is simple but not very accurate: flip n independent coins with probability p of heads and say that the lineages whose coins come up heads are those that are descended from the individual with the beneficial mutation. A second approximation, which is related to Kingman's paintbox construction, replaces the coin flips by integer-valued random variables and leads to very accurate results.

1. Introduction. A classical continuous-time model for a population with overlapping generations is the Moran model, which was introduced by Moran [19]. Thinking of N diploid individuals, we assume the population size is fixed at $2N$. However under the assumption that each individual is a random union of gametes, the dynamics are the same as for a population of $2N$ haploid individuals, so we will do our computation for that case. In the simplest version of the Moran model, each individual independently lives for a time that is exponentially distributed with mean 1 and then is replaced by a new individual. The parent of the new individual is chosen at random from the $2N$ individuals, including the individual being replaced.

Here we will consider a variation of the Moran model that involves two loci—one subject to natural selection, the other neutral—and with a probability r in each generation of recombination between the two loci. To begin to explain the last sentence, we assume that at the selected locus there are two alleles, B and b, and that the relative fitnesses of the two alleles are 1 and $1 - s$. The population then

---

Received June 2003; revised May 2005.

[1] Supported by an NSF Postdoctoral Fellowship while the author was at Cornell 2001–2004.

[2] Supported in part by NSF grants from the probability program (0202935) and from a joint DMS/NIGMS initiative to support research in mathematical biology (0201037).

AMS 2000 subject classifications. Primary 92D10; secondary 60J85, 92D15, 05A18.

Key words and phrases. Coalescence, random partition, selective sweep, mutation, hitchhiking.

evolves with the same rules as before, except that a replacement of an individual with a B allele by an individual with a b allele is rejected with probability s. Consequently, if at some time there are k individuals with the B allele and $2N - k$ with the b allele, then the rate of transitions that increase the number of B individuals from k to $k+1$ is $k(2N-k)/(2N)$, but the rate of transitions that reduce the number of B individuals to $k-1$ is $k(2N-k)(1-s)/(2N)$. See Chapter 3 of [6] for a summary of some work with this model.

We assume that the process starts at time zero with $2N - 1$ individuals having the b allele and one individual having the advantageous B allele. We think of the individual with the B allele as having had a beneficial mutation at time zero. There is a positive probability that eventually all $2N$ individuals will have the favorable allele. When this happens, we say that a selective sweep occurs, because the favorable allele has swept through the entire population.

If we assume that the entire chromosome containing the selected locus is passed down from one generation to the next, as is the case for the Y chromosome or mitochondrial DNA, then all $2N$ chromosomes at the end of the selective sweep will have come from the one individual that had the beneficial mutation at the beginning of the sweep. However, nonsex chromosomes in diploid individuals are typically not an identical copy of one of their parents' chromosomes. Instead, because of a process called recombination, each chromosome that an individual inherits consists of pieces of each of a parent's two chromosomes. In this case, if we are interested in the origin of a second neutrally evolving locus on the chromosome and a selective sweep occurs because of an advantageous mutation at a nearby site, then some of the lineages will be traced back to the chromosome that had the favorable allele at the beginning of the sweep, but other lineages will be traced back to different individuals because of recombination between the neutral and selected loci. When a lineage can be traced back to an individual other than the one with the beneficial mutation, we say that the lineage escapes from the selective sweep.

The combined effects of recombination and selective sweeps have been studied extensively. Maynard Smith and Haigh [18] observed that selective sweeps can alter the frequencies of alleles at sites nearby the site at which the selective sweep occurred. They referred to this as the hitchhiking effect. 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 the a allele) and the recombination probability in each generation is r, Maynard Smith and Haigh [18] showed [see (8) on page 25] that the frequency of the A allele after the selective sweep is reduced from R_0 to

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

In the calculation of Maynard Smith and Haigh, the number of individuals with the B locus grows deterministically. Kaplan, Hudson and Langley [15] used a model involving an initial phase in which the number of B's is a supercritical branching process, a middle deterministic piece where the fraction p of B's follows the logistic differential equation

$$\frac{dp}{dt} = sp(1-p), \tag{1.1}$$

and a final random piece where the number of b's follows a subcritical branching process. This process is too difficult to study analytically, so they resorted to simulation.

Stephan, Wiehe and Lenz [27] further simplified this approach by ignoring the random first and third phases and modeling the change in the frequency of B's by the logistic differential equation (1.1), which has solution

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

This approach has been popular with biologists in simulation studies (see, e.g., [23, 26]). However, as results in [2, 7] show, this can introduce substantial errors, so rather than using this approximation for our analysis, we will consider a modification of the Moran model that allows for recombination as well as beneficial mutations.

We consider two sites on each chromosome. At one site, each of the $2N$ chromosomes has either the advantageous B allele or a b allele. Our interest, however, is in the genealogy at another neutral site, at which all alleles have the same fitness. As before, we assume that each individual lives for an exponential time with mean 1 and is replaced by a new individual whose parent is chosen at random from the population, except that we disregard disadvantageous replacements of a B chromosome by a b chromosome with probability s. We will also now assume that when a new individual is born, it inherits alleles at both sites from the same individual with probability $1 - r$. With probability r, there is recombination between the two sites, and the individual inherits the allele at the neutral site from its parent's other chromosome. Since a parent's two chromosomes are considered to be two distinct individuals in the population, we model this by saying that the new individual inherits the two alleles from two ancestors chosen independently at random from the $2N$ individuals in the population.

Suppose we sample n chromosomes at the end of a selective sweep and follow their ancestral lines back until the beginning of the sweep. We will describe the genealogy of the sample by a marked partition of $\{1, \dots, n\}$, which we define to be a partition of $\{1, \dots, n\}$ in which one block of the partition may be designed as a "marked" block. We define the marked partition Θ of $\{1, \dots, n\}$ as follows. We say that two integers i and j are in the same block of Θ, denoted $i \sim_\Theta j$, if and only if the alleles at the neutral site on the ith and jth chromosomes in the sample have the same ancestor at the beginning of the sweep. Thus, if we are following the lineages associated with the allele at the neutral site, we have $i \sim_\Theta j$ if and only if the ith and jth lineages coalesce during the selective sweep. We also mark the block of Θ containing the integers i for which the ith individual is descended from the individual that had the beneficial mutation at the beginning of the sweep. Thus, to understand how a selective sweep affects the genealogy of a sample of size n, we need to understand the distribution of the random marked partition Θ.

In this paper we study two approximations to the distribution of Θ. The approximations were introduced and studied by simulation in [7]. Here we provide precise bounds on the error in the approximations. The idea behind the first approximation is that a large number of lineages will inherit their allele at the neutral site from the individual that had the beneficial mutation at the beginning of the sweep, and the corresponding integers i will be in the marked block of Θ. With high probability, the lineages that escape the selective sweep do not coalesce with one another, so the corresponding integers are in singleton blocks of Θ.

Before stating the first approximation precisely, we need a definition. Let $p \in [0, 1]$. Let $\xi_1, \dots, \xi_n$ be independent random variables such that, for $i = 1, \dots, n$, we have $P(\xi_i = 1) = p$ and $P(\xi_1 = 0) = 1 - p$. We call the random marked partition of $\{1, \dots, n\}$ such that one marked block consists of $\{i \in \{1, \dots, n\} : \xi_i = 1\}$ and the remaining blocks are singletons a p-partition of $\{1, \dots, n\}$. Let Q_p denote the distribution of a p-partition of $\{1, \dots, n\}$.

Theorem 1.1 below shows that the distribution of Θ can be approximated by the distribution of a p-partition. For this result and throughout the rest of the paper, we assume that the selective advantage s is a fixed constant that does not depend on the population size N. However, the recombination probability r is allowed to depend on N, even though we have not recorded this dependence in the notation. We will assume throughout the paper that $r \le C_0/(\log N)$ for some positive constant C_0. We denote by $\mathcal{P}_n$ the set of marked partitions of $\{1, \dots, n\}$.

THEOREM 1.1. *Fix $n \in \mathbb{N}$. Let $\alpha = r \log(2N)/s$. Let $p = e^{-\alpha}$. Then there exists a positive constant C such that $|P(\Theta = \pi) - Q_p(\pi)| \le C/(\log N)$ for all N and all $\pi \in \mathcal{P}_n$.*

In this theorem and throughout the rest of the paper, C denotes a positive constant that may depend on s, but does not depend on r or N. The value of C may change from line to line.

A consequence of Theorem 1.1 is that if $\lim_{N \to \infty} r \log(2N)/s = \alpha$ for some $\alpha \in (0, \infty)$ and $p = e^{-\alpha}$, then the distribution of Θ converges to Q_p as $N \to \infty$. However, the rate of convergence that the theorem gives is rather slow, and simulation results of Barton [2] and Durrett and Schweinsberg [7] show that the approximation is not very accurate for realistic values of N. Consequently, it is necessary to look for a better approximation. Theorem 1.2 below gives an approximation with an error term that is of order $1/(\log N)^2$ rather than $1/\log N$.

It follows from the improved approximation that the error in Theorem 1.1 is actually of order $1/\log N$.

The motivation for the second approximation comes from the observation that, at the beginning of the selective sweep, the number of B's can be approximated by a continuous-time branching process in which each individual gives birth at rate 1 and dies at rate $1 - s$. Some individuals in this supercritical branching process will have an infinite line of descent, meaning that they have descendants alive in the population at all future times. As we will show later, the individuals with an infinite line of descent can be approximated by a Yule process, a continuous-time branching process in which each individual splits into two at a constant rate s. Since our sample, taken at the end of the selective sweep, comes from lineages that have survived a long time, we can get a good approximation to the genealogy by considering only individuals with an infinite line of descent. We will also show that, during the time when there are exactly $k \geq 2$ lineages with an infinite line of descent, the expected number of recombinations along these lineages is r/s. For simplicity, we assume that the number of such recombinations is always either 0 or 1. Such a recombination causes individuals descended from the lineage with the recombination to be traced back to an ancestor at time zero different from descendants of the other $k - 1$ lineages (and therefore to belong to a different block of Θ). Well-known facts about the Yule process (see, e.g., [14]) imply that when there are k lineages, the fraction of individuals at the end of the sweep that are descendants of a given lineage has approximately a beta$(1, k - 1)$ distribution. Furthermore, we will show that with probability $(r(1-s))/(r(1-s)+s)$, there is a recombination when there is only one individual with an infinite line of descent, in which case none of the sampled lineages will get traced back to the individual with the B allele at time zero.

These observations motivate the definition of a class of marked partitions of $\{1, \dots, n\}$, which we will use to approximate the distribution of Θ. The construction resembles the paintbox construction of exchangeable random partitions due to Kingman [16]. To start the construction, assume $0 < r < s$ and let L be a positive integer. Then let $(W_k)_{k=2}^L$ be independent random variables such that W_k has a beta distribution with parameters 1 and $k - 1$. Let $(\zeta_k)_{k=2}^L$ be a sequence of independent random variables such that $P(\zeta_k = 1) = r/s$ and $P(\zeta_k = 0) = 1 - r/s$ for all k. As the reader might guess from the probabilities, $\zeta_k = 1$ corresponds to a recombination when there are k lineages with an infinite line of descent. For $k = 2, 3, \dots, L$, let $V_k = \zeta_k W_k$ and let $Y_k = V_k \prod_{j=k+1}^L (1 - V_j)$ be the fraction of individuals carried away by recombination. Let $Y_1 = \prod_{j=2}^L (1 - V_j)$. Note that $\sum_{k=1}^L Y_k = 1$. Finally, let $Q_{r,s,L}$ be the distribution of the random marked partition Π of $\{1, \dots, n\}$ constructed in the following way. Define random variables $Z_1, \dots, Z_n$ to be conditionally independent given $(Y_k)_{k=1}^L$ such that for $i = 1, \dots, n$ and $j = 1, \dots, L$, we have $P(Z_i = j | (Y_k)_{k=1}^L) = Y_j$. Here the integers i such that $Z_i = k$ correspond to lineages that recombine when there are

k members of the B population with an infinite line of descent. Then define Π such that $i \sim_\Pi j$ if and only if $Z_i = Z_j$. Independently of $(Z_i)_{i=1}^n$, we mark the block $\{i : Z_i = 1\}$ with probability $s/(r(1-s)+s)$ and, with probability $(r(1-s))/(r(1-s)+s)$, we mark no block. When the block is marked, the integers i such that $Z_i = 1$ correspond to the lineages that do not recombine and therefore can be traced back to the individual that had the beneficial mutation at time zero; otherwise, they correspond to the lineages that recombine when there is only one member of the B population with an infinite line of descent.

We are now ready to state our main approximation theorem, which says that the distribution of Θ can be approximated well by the distribution $Q_{r,s,L}$, where $L = \lfloor 2Ns \rfloor$, and $\lfloor m \rfloor$ denotes the greatest integer less than or equal to m. The choice of L comes from the fact that in a continuous-time branching process with births at rate 1 and deaths at rate $1-s$, each individual has an infinite line of descent with probability s. Therefore, the number of such individuals at the end of the selective sweep is approximately L.

THEOREM 1.2. *Fix $n \in \mathbb{N}$ and let $L = \lfloor 2Ns \rfloor$. Then there exists a positive constant C such that for all N and all $\pi \in \mathcal{P}_n$,*

$$|P(\Theta = \pi) - Q_{r,s,L}(\pi)| \le C/(\log N)^2.$$

Consider for concreteness $N = 10{,}000$, a number commonly used for the "effective size" of the human population. To explain the term in quotes, we note that although there are now 6 billion humans, our exponential population growth is fairly recent, so for many measures of genetic variability the human population is the same as a homogeneously mixing population of constant size 10,000. When $N = 10{,}000$, $\log N = 9.214$ and $(\log N)^2 = 84.8$, so Theorem 1.2 may not appear at first glance to be a big improvement. Two concrete examples, however, show that the improvement is dramatic (see Table 1). In each case $N = 10^4$ and $s = 0.1$. More extensive simulation results comparing the two approximations are given in [7].

In Table 1, $pinb$ is the probability that a lineage escapes the selective sweep. The remaining three columns pertain to two lineages: $p2inb$ is the probability that

TABLE 1

		$pinb$	$p2inb$	$p2cinb$	$p1B1b$
$r = 0.00106$	Theorem 1.1	0.1	0.01	0	0.18
	Moran	0.08203	0.00620	0.01826	0.11513
	Theorem 1.2	0.08235	0.00627	0.01765	0.11687
$r = 0.00516$	Theorem 1.1	0.4	0.16	0	0.48
	Moran	0.33656	0.10567	0.05488	0.35201
	Theorem 1.2	0.34065	0.10911	0.05100	0.36112

two lineages both escape the sweep but do not coalesce, $p2cinb$ is the probability both lineages escape but coalesce along the way and $p1B1b$ is the probability one lineage escapes the sweep but the other does not. The remaining possibility is that neither lineage escapes the sweep, but this probability can be computed by subtracting the sum of the other three probabilities from 1. The first row in each group gives the probabilities obtained from the approximation in Theorem 1.1 and the third row gives the probabilities obtained from the approximation in Theorem 1.2. The second row gives the average of 10,000 simulation runs of the Moran model described earlier. The values of the recombination rate r were chosen in the two examples to make the approximations to $pinb$ given by Theorem 1.1 equal to 0.1 and 0.4, respectively. It is easy to see from the table that the approximation from Theorem 1.2 is substantially more accurate. In particular, note that in the approximation given by Theorem 1.1, two lineages never coalesce unless both can be traced back to the individual with the beneficial mutation. Consequently, $p2cinb$ would be zero if this approximation were correct. However, in simulations, a significant percentage of pairs of lineages both coalesced and escaped from the sweep, and this probability is approximated very well by Theorem 1.2 in both examples.

The results in this paper are a first step in studying situations in which, as proposed by Gillespie [12], selective sweeps occur at times of a Poisson process in a single locus or distributed along a chromosome at different distances from the neutral locus at which data have been collected. It is well known that in the Moran model when there are no advantageous mutations, if we sample n individuals and follow their ancestors backward in time, then when time is sped up by $2N$, we get the coalescent process introduced by Kingman [17]. It is known (see [6]) that selective sweeps require an average amount of time $(2/s)\log N$, so when time is sped up by $2N$, the selective sweep occurs almost instantaneously. Durrett and Schweinsberg [8] showed that Theorem 1.1 implies that if advantageous mutations occur at times of a Poisson process then, the ancestral processes converge as $N \to \infty$ to a coalescent with multiple collisions of the type introduced by Pitman [22] and Sagitov [24]. At times of a Poisson process, multiple lineages coalesce simultaneously into one. The more accurate approximation in Theorem 1.2 suggests that a better approximation to the ancestral process can be given by a coalescent with simultaneous multiple collisions. These coalescent processes were studied by Möhle and Sagitov [20] and Schweinsberg [25].

Finally, it is important to emphasize that the results in this paper are for the case of “strong selection,” where the selective advantage s is $O(1)$. There has also been considerable interest in weak selection, where Ns is assumed to converge to a limit as $N \to \infty$, which means s is $O(1/N)$. In this case, there is a diffusion limit as $N \to \infty$. For work in this direction that incorporates the effect of recombination, see [3, 4]. Recently, Etheridge, Pfaffelhuber and Wakolbinger [9] have shown that many of the results in this paper carry over to the diffusion setting. They assume that $Ns \to \alpha$ as $N \to \infty$, so that they can work with a diffusion limit, and then obtain an approximation to the distribution of the ancestral partition Θ that has an error of order $1/(\log \alpha)^2$ as $\alpha \to \infty$, by using approximations to the genealogy similar to those used in the present paper.

2. Overview of the proofs.

Since the proofs of Theorems 1.1 and 1.2 are rather long, we outline the proofs in this section. A precise definition of the genealogy is given in Section 2.1. The proof of Theorem 1.1 is outlined in Section 2.2. In Section 2.3 we describe the coupling with a supercritical branching process and outline the proof of Theorem 1.2.

2.1. *Precise definition of the genealogy.*

We now define more precisely our model of a selective sweep. We construct a process $M = (M_t)_{t=0}^{\infty}$. The vector $M_t = (M_t(1), \dots, M_t(2N))$ contains the information about the population at the time of the tth proposed replacement, and $M_t(i) = (A_t^0(i), \dots, A_t^{t-1}(i), B_t(i))$ contains the information about the ancestors of the ith individual at time t. For $0 \le u \le t-1$, $A_t^u(i)$ is the individual at time u that is the ancestor of the ith individual at time t, when we consider the genealogy at the neutral locus. The final coordinate $B_t(i) = 1$ if the ith individual at time t has the B allele; $B_t(i) = 0$ if this individual has the b allele. Note that this is a discrete-time process, but one can easily recover the continuous-time description by replacing discrete-time steps with independent holding times, each having an exponential distribution with mean $1/(2N)$.

At time zero, only one of the chromosomes will have the B allele. We define a random variable U, which is uniform on the set $\{1, \dots, 2N\}$, and we let $B_0(U) = 1$ and $B_0(i) = 0$ for $i \neq U$. We now define a collection of independent random variables $I_{t,j}$ for $t \in \mathbb{N}$ and $j \in \{1, \dots, 5\}$. For $j \in \{1, 2, 3\}$, the random variable $I_{t,j}$ is uniform on $\{1, \dots, 2N\}$.

- The random variable $I_{t,1}$ will be the individual that dies at time t.
- The random variable $I_{t,2}$ will be the parent of the new individual at time t.
- The random variable $I_{t,3}$ will be the other parent from whom the new chromosome will inherit its allele at the neutral locus if there is recombination.
- The random variable $I_{t,4}$ will be an indicator for whether a proposed disadvantageous change will be rejected, so $P(I_{t,4} = 1) = s$ and $P(I_{t,4} = 0) = 1 - s$.
- The random variable $I_{t,5}$ will determine whether there is recombination at time t, so $P(I_{t,5} = 1) = r$ and $P(I_{t,5} = 0) = 1 - r$.

Using these random variables we can construct the process in the obvious way. Refer to Figure 1 for help with the notation.

1. If $B_{t-1}(I_{t,1}) = 1$, $B_{t-1}(I_{t,2}) = 0$ and $I_{t,4} = 1$, then the population will be the same at time t as at time $t-1$ because the proposed replacement of a B chromosome by a b chromosome is rejected. In this case, for all $i = 1, \dots, 2N$, we define $B_t(i) = B_{t-1}(i)$, $A_t^{t-1}(i) = i$ and $A_t^u(i) = A_{t-1}^u(i)$ for $u = 0, \dots, t-2$.

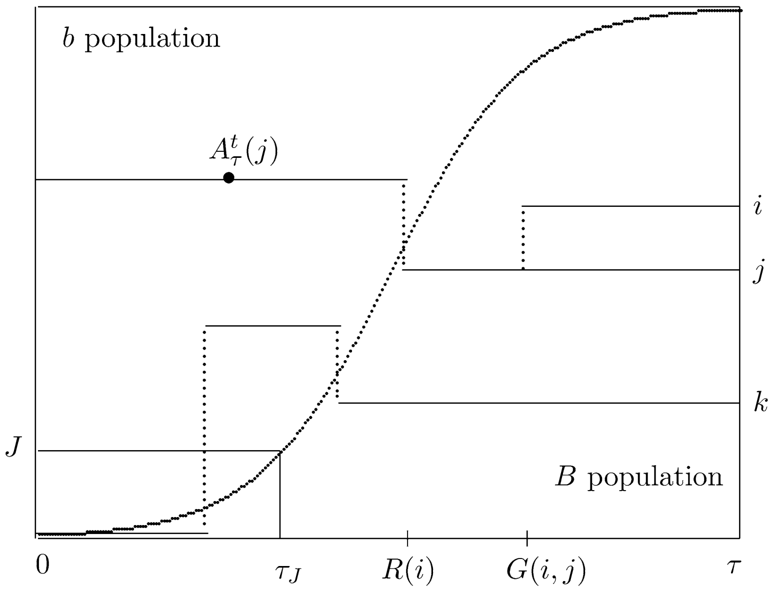

FIG. 1. *A picture to explain our notation. The lineages jump around as we move backward in time, but for simplicity we have only indicated the recombination events. Here as we work backward in time, i and j coalesce and then recombine into the b population. Proposition* 2.4 *shows that this event has probability at most $C/\log N$. Proposition* 2.1 *estimates the probability of two recombinations as shown in lineage k.*

2. If we are not in the previous case and $I_{t,5} = 0$, then there is no recombination at time t. So the individual $I_{t,1}$ dies and the new individual gets its alleles at both sites from $I_{t,2}$. For $i \neq I_{t,1}$, define $B_t(i) = B_{t-1}(i)$, $A^u_t(i) = A^u_{t-1}(i)$ for $u = 0, \dots, t-2$ and $A^{t-1}_t(i) = i$. Let $B_t(I_{t,1}) = B_{t-1}(I_{t,2})$, $A^u_t(I_{t,1}) = A^u_{t-1}(I_{t,2})$ for $u = 0, \dots, t-2$ and $A^{t-1}_t(I_{t,1}) = I_{t,2}$.
3. If we are not in either of the previous two cases, then there is recombination at time t. This means that the new individual labeled $I_{t,1}$ gets a B or b allele from $I_{t,2}$ but gets its allele at the neutral locus from $I_{t,3}$. For $i \neq I_{t,1}$, define $B_t(i) = B_{t-1}(i)$, $A^u_t(i) = A^u_{t-1}(i)$ for $u = 0, \dots, t-2$ and $A^{t-1}_t(i) = i$. Let $B_t(I_{t,1}) = B_{t-1}(I_{t,2})$, $A^u_t(I_{t,1}) = A^u_{t-1}(I_{t,3})$ for $u = 0, \dots, t-2$ and $A^{t-1}_t(I_{t,1}) = I_{t,3}$.

It will also be useful to have notation for the number of individuals with the favorable allele. For nonnegative integers t, define $X_t = \#\{i : B_t(i) = 1\}$, where $\#S$ denotes the cardinality of the set S. For $J = 1, 2, \dots, 2N$, let $\tau_J = \inf\{t : X_t \geq J\}$ be the first time at which the number of B's in the population reaches J. Let $\tau = \inf\{t : X_t \in \{0, 2N\}\}$ be the time at which the B allele becomes fixed in the population (in which case $X_\tau = 2N$) or disappears (in which case $X_\tau = 0$). Since our main interest is in studying a selective sweep, P' and E' will denote probabilities and expectations under the unconditional law of M, and P and E will denote probabilities and expectations under the conditional law of M given $X_\tau = 2N$. Likewise, Var and Cov will always refer to conditional variances and covariances given $X_\tau = 2N$.

To sample n individuals from the population at the time τ when the selective sweep ends, we may take the individuals $1, \dots, n$ because the distribution of genealogy of n individuals does not depend on which n individuals are chosen. Therefore, we can define Θ to be the random marked partition of $\{1, \dots, n\}$ such that $i \sim_\Theta j$ if and only if the ith and jth individuals at time τ get their allele at the neutral site from the same ancestor at time 0, with the marked block corresponding to the individuals descended from the individual U, which had the beneficial mutation at time zero. More formally, we have $i \sim_\Theta j$ if and only if $A_\tau^0(i) = A_\tau^0(j)$ with the marked block being $\{i : A_\tau^0(i) = U\}$ or, equivalently, $\{i : B_0(A_\tau^0(i)) = 1\}$.

2.2. *The first approximation.* Recall that Theorem 1.1 says that we can approximate Θ by flipping independent coins for each lineage, which come up heads with probability p, to determine which lineages fail to escape the selective sweep. These lineages are then in one block of the partition, because they are descended from the ancestor with the beneficial mutation at time zero, while the other lineages do not coalesce and correspond to singleton blocks of the partition.

The first step in establishing this picture is to calculate the probability that one lineage escapes the selective sweep. In the notation above, we need to find $P(B_0(A_\tau^0(i)) = 0)$. Define $R(i) = \sup\{t \geq 0 : B_t(A_\tau^t(i)) = 0\}$, where $\sup \varnothing = -\infty$. If we work backward in time, $R(i)$ is the first moment that the lineage of the neutral locus resides in the b population. Note that it is possible to have $R(i) \geq 0$ and $B_0(A_\tau^0(i)) = 1$ if a lineage is affected by two recombinations, one taking it from the B population to the b population and another taking it back into the B population. The next result shows that the probability of this is small.

PROPOSITION 2.1. *We have* $P(B_t(A_\tau^t(i)) = 1$ *for some* $t \leq R(i)) \leq C/(\log N)^2$.

Proposition 2.1 implies that in the proofs of Theorems 1.1 and 1.2, the probability that a lineage escapes the selective sweep can be approximated by $P(R(i) \geq 0)$. It will also be useful to have an approximation of $P(R(i) \geq \tau_J)$ for $J \geq 1$, which is the probability that a given lineage escapes into the b population after the time when the number of B's in the population reaches J. The next result gives such an approximation.

PROPOSITION 2.2. *If* $q_J = 1 - \exp(-\frac{r}{s}\sum_{k=J+1}^{2N} \frac{1}{k})$, *then*

$$P\bigl(R(i) \geq \tau_J\bigr) = q_J + O\biggl(\frac{1}{(\log N)^2} + \frac{1}{(\log N)\sqrt{J}}\biggr).$$

Propositions 2.1 and 2.2 will be proved in Section 3.

The next step is to consider two lineages. We now need to consider not only recombination, but also the possibility that the lineages may coalesce, meaning that the alleles at the neutral site on the two lineages are descended from the same ancestor at the beginning of the sweep. Let $G(i, j)$ be the time that the ith and jth lineages coalesce. More precisely, we define $G(i, j) = \sup\{t : A_\tau^t(i) = A_\tau^t(j)\}$ with $\sup\varnothing = -\infty$. Our first result regarding coalescence shows that it is unlikely for two lineages to coalesce at a given time unless both alleles at the neutral site are descended from a chromosome with the B allele at that time.

PROPOSITION 2.3. *We have* $P(G(i, j) \geq 0$ *and* $B_{G(i,j)+1}(A_\tau^{G(i,j)+1}(i)) = 0) \leq C(\log N)/N$.

Next, we bound the probability that, if we trace two lineages back through the selective sweep, the lineages coalesce and then escape from the sweep.

PROPOSITION 2.4. *We have* $P(0 \leq R(i) \leq G(i, j)) \leq C/(\log N)$.

Note that Proposition 2.3 says that, with high probability, only lineages in the B population merge, while Proposition 2.4 says that, in the first-order approximation, lineages that have merged do not escape into the b population. Together, these results will justify the approximation of Θ by a random partition in which the only nonsingleton block corresponds to lineages that fail to escape the selective sweep. The next result bounds the probability that two lineages coalesce after time τ_J.

PROPOSITION 2.5. *Let* $C' > 0$. *If* $J \leq C'N/(\log N)$, *then* $P(G(i, j) \geq \tau_J) \leq C/J$.

We prove Propositions 2.3, 2.4 and 2.5 in Section 4.

We now consider n lineages. To prove Theorem 1.1, we will need to show that the events $\{R(1) \geq 0\}, \dots, \{R(n) \geq 0\}$ are approximately independent. Let $K_t = \#\{i \in \{1, \dots, n\} : R(i) \geq t\}$. If the events that the n lineages escape the selective sweep after time t are approximately independent, then K_t should have approximately a binomial distribution. The following proposition, which we prove in Section 5, provides a binomial approximation to the distribution of K_{τ_J}. Since $\tau_1 = 0$, the $J = 1$ case will be used in the proof of Theorem 1.1, while the general case will help to prepare us for the proof of Theorem 1.2.

PROPOSITION 2.6. *Define* q_J *as in Proposition* 2.2. *If* $J \leq C'N/(\log N)$, *then*

$$\left|P(K_{\tau_J} = d) - \binom{n}{d} q_J^d (1 - q_J)^{n-d}\right| \leq \min\left\{\frac{C}{\log N}, \frac{C}{J}\right\} + \frac{C}{(\log N)^2}$$

for $d = 0, 1, \dots, n$.

PROOF OF THEOREM 1.1. Define a new partition Θ' of $\{1,\dots,n\}$ such that $i \sim_{\Theta'} j$ if and only if $R(i) = R(j) = -\infty$. We mark the block of Θ' consisting of $\{i : R(i) = -\infty\}$. In words, only the lineages that recombine and hence stay in the B population are trapped by the sweep. To do this we make the following observations:

- Proposition 2.1 implies that the probaility of two recombinations affecting a lineage can be ignored.
- Proposition 2.3 says that we can ignore coalescence in the b population.
- Proposition 2.4 says that the probability two lineages coalesce and then escape has small probability.

The results above imply that $P(\Theta \neq \Theta') \leq C/(\log N)$. Therefore, to prove Theorem 1.1, it suffices to show that $|P(\Theta' = \pi) - Q_p(\pi)| \leq C/(\log N)$ for all marked partitions π of $\{1,\dots,n\}$. It follows from Proposition 2.6 with $J = 1$ and the exchangeability of Θ' that $|P(\Theta' = \pi) - Q_{1-q_1}(\pi)| \leq C/(\log N)$ for all $\pi \in \mathcal{P}_n$. Using the definition of q_1 and $|\frac{d}{dx}e^{-x}| \leq 1$ for $x \geq 0$ gives

$$|(1-q_1) - p| = \left|\exp\left(-\frac{r}{s}\sum_{k=2}^{2N}\frac{1}{k}\right) - \exp\left(-\frac{r}{s}\log(2N)\right)\right|$$

$$\leq \frac{r}{s}\left|\sum_{k=2}^{2N}\frac{1}{k} - \log(2N)\right| \leq \frac{C}{\log N}$$

and the theorem follows. □

2.3. *Branching process coupling and the second approximation.* We now work toward improving the approximation to the distribution of Θ so that we can prove Theorem 1.2. To do this, we will break the selective sweep into two stages. Let $J = \lfloor (\log N)^a \rfloor$, where $a > 4$ is a fixed constant. We will consider separately the time intervals $[0, \tau_J)$ and $[\tau_J, \tau]$.

Part 1: $\Theta \approx \Theta_1 \approx \Theta_2$. We first establish that we can ignore coalescence involving a lineage that escapes the sweep after time τ_J. Define a random marked partition Θ_1 of $\{1,\dots,n\}$ such that $i \sim_{\Theta_1} j$ if and only if $R(i) < \tau_J$, $R(j) < \tau_J$ and $A_\tau^0(i) = A_\tau^0(j)$. Mark the block of Θ_1 consisting of $\{i : R(i) < \tau_J$ and $B_0(A_\tau^0(i)) = 1\}$. Note that $\Theta_1 = \Theta$ unless, for some i and j, we have $R(i) \geq \tau_J$ and either $i \sim_\Theta j$ or $B_0(A_\tau^0(i)) = 1$. It follows from Propositions 2.1, 2.3 and 2.5 that $P(\Theta \neq \Theta_1) \leq C/(\log N)^2$. Thus, we may now work with Θ_1.

The next step is to approximate the distribution of Θ_1. Let $K_t = \{i \in \{1,\dots,n\} : R(i) \geq t\}$, as defined before the statement of Proposition 2.6. Define $m = n - \#K_{\tau_J}$ to be the number of lineages in the B population at time τ_J. Proposition 2.5 shows that lineages are unlikely to coalesce in $[\tau_J, \tau]$. Relabel the lineages using an arbitrary bijective function f from $\{1,\dots,n\} \setminus K_{\tau_J}$ to $\{1,\dots,m\}$.

To describe the first stage of the selective sweep precisely, we define, for each $m \leq J$, a new marked partition Ψ_m of $\{1, \dots, m\}$. Let σ_m be a random injective map from $\{1, \dots, m\}$ to $\{i : B_{\tau_J}(i) = 1\}$ such that all $(J)_m = (J)(J-1) \cdots (J-m+1)$ maps are equally likely. Thus, $\sigma_m(1), \dots, \sigma_m(m)$ is a random sample from the J individuals with the B allele at time τ_J. Then define Ψ_m such that $i \sim_{\Psi_m} j$ if and only if $A^0_{\tau_J}(\sigma_m(i)) = A^0_{\tau_J}(\sigma_m(j))$. This means i and j are in the same block of Ψ_m if and only if the $\sigma_m(i)$th and $\sigma_m(j)$th individuals at time τ_J inherited their allele at the neutral locus from the same individual at the beginning of the sweep. The block $\{i : B_0(A^0_{\tau_J}(\sigma_m(i))) = 1\}$ is marked.

Define Θ_2 to be the marked partition of $\{1, \dots, n\}$ such that $i \sim_{\Theta_2} j$ if and only if $R(i) < \tau_J$, $R(j) < \tau_J$ and $f(i) \sim_{\Psi_m} f(j)$. Let the marked block of Θ_2 consist of all i such that $R(i) < \tau_J$ and $f(i)$ is in the marked block of Ψ_m. To compare Θ_1 and Θ_2, note that $f(i) \sim_{\Psi_m} f(j)$ if and only if $A^0_{\tau_J}(\sigma_m(f(i))) = A^0_{\tau_J}(\sigma_m(f(j)))$. On the other hand, $A^0_\tau(i) = A^0_\tau(j)$ if and only if $A^0_{\tau_J}(A^{\tau_J}_\tau(i)) = A^0_{\tau_J}(A^{\tau_J}_\tau(j))$. For $i \neq j$, we have $P(A^{\tau_J}_\tau(i) = A^{\tau_J}_\tau(j)) \leq C/(\log N)^4$ by Proposition 2.5. By the strong Markov property, the genealogy of the process up to time τ_J is independent of K_{τ_J}. From these observations and the exchangeability of the model, it follows that for all $\pi \in \mathcal{P}_n$, we have $|P(\Theta_1 = \pi) = P(\Theta_2 = \pi)| \leq C/(\log N)^4$.

Part 2: $\Psi_m \approx \Upsilon_m \approx Q_{r,s,\lfloor Js \rfloor}(\pi)$. Our next step is to understand the distribution of Ψ_m. The first step is to show that the first stage of a selective sweep can be approximated by a branching process. Recall that when the number of individuals with the favorable B allele is $k \ll 2N$, the rate of transitions that increase the number of B individuals from k to $k+1$ is $k(2N-k)/(2N) \approx k$, while the rate of transitions that decrease the number of B individuals from k to $k-1$ is $k(2N-k)(1-s)/(2N) \approx k(1-s)$. Therefore, the individuals with the B allele follow approximately a continuous-time branching process in which each individual gives birth at rate 1 and dies at rate $1-s$. Also, each new individual born with the B allele inherits the allele at the neutral site from its parent with probability $1-r$. We can model this recombination by considering a multitype branching process starting from one individual in which each new individual is the same type as its parent with probability $1-r$ and is a new type, different from any other member of the current population, with probability r.

Say that an individual in the branching process at time t has an infinite line of descent if it has a descendant in the population at time u for all $u > t$. Otherwise, say the individual has a finite line of descent. It is well known that the process consisting only of the individuals with an infinite line of descent is also a branching process. This is discussed, for example, in [1]. For more recent work in this direction, see [10, 11, 21]. In Section 6 we will show that when the original branching process is a continuous-time branching process with births at rate 1 and deaths at rate $1-s$, the process consisting only of the individuals with an infinite line of descent is a continuous-time branching process with no deaths in which each individual gives birth at rate s. That is, this process is a Yule process with births at rate s. The probability that a randomly chosen individual has an infinite line of descent is s, so when the original branching process has J individuals, there are approximately Js individuals with an infinite line of descent. Furthermore, since the past and future are independent by the Markov property, the genealogy of a sample will not be affected if we sample only from the individuals with infinite lines of descent.

In Section 6, we justify these approximations. This will lead to a proof of the following proposition, which explains how the genealogy of the first phase of a selective sweep can be approximated by the genealogy of a continuous-time branching process.

PROPOSITION 2.7. *Consider a continuous-time multitype branching process started with one individual at time zero such that each individual gives birth at rate* 1 *and dies at rate* $1 - s$. *Assume that each individual born has the same type as its parent with probability* $1-r$ *and a new type with probability* r. *Condition this process to survive forever. At the first time at which there are* $\lfloor Js \rfloor$ *individuals with an infinite line of descent, sample* m *of the* $\lfloor Js \rfloor$ *individuals with an infinite line of descent. Define* Υ_m *to be the marked partition of* $\{1, \dots, m\}$ *such that* $i \sim_{\Upsilon_m} j$ *if and only if the* i*th and* j*th individuals in the sample have the same type, and the marked block consists of the individuals with the same type as the individual at time zero. Then for all* $\pi \in \mathcal{P}_m$, *we have* $|P(\Psi_m = \pi) - P(\Upsilon_m = \pi)| \le C/(\log N)^2$.

Recall that in the introduction we constructed a random marked partition Π with distribution $Q_{r,s,L}$, where $L = \lfloor 2Ns \rfloor$. To compare this partition with Θ, we will consider the construction in two stages, just as we considered two stages of the selective sweep. The first stage of the construction will involve the integers i such that $Z_i \le \lfloor Js \rfloor$, and the second stage involves the integers i such that $Z_i > \lfloor Js \rfloor$. We think of $Z_i = k$ as meaning that the ith lineage escapes the selective sweep at a time when there are k individuals in the Yule process (or, equivalently, k lineages in the branching process with an infinite line of descent). We use $\lfloor Js \rfloor$ as the boundary between the two stages because, when the population size of the branching process reaches J, there are approximately Js individuals with an infinite line of descent.

The next result compares the first stage of a selective sweep to the random variables Z_i such that $Z_i \le \lfloor Js \rfloor$.

PROPOSITION 2.8. *There is a positive constant* C *such that for all partitions* π *of* $\{1, \dots, n\}$, *we have* $|P(\Upsilon_n = \pi) - Q_{r,s,\lfloor Js \rfloor}(\pi)| \le C/(\log N)^2$.

Part 3: $\Theta_2 \approx Q_{r,s,\lfloor Js \rfloor, q_J} \approx Q_{r,s,L}$. Proposition 2.6 shows that the number of lineages that escape the sweep during $[\tau_J, \tau]$ has approximately a binomial distribution with success probability q_J. This motivates the following definition:

DEFINITION 2.9. Let r, s and q be in $(0,1)$, and let L be a positive integer. Let $Q_{r,s,L,q}$ be the distribution of the random marked partition Π' of $\{1,\dots,n\}$ obtained as follows. First, let Π be a random marked partition of $\{1,\dots,n\}$ with distribution $Q_{r,s,L}$. Let $\xi_1,\dots,\xi_n$ be i.i.d. random variables such that $P(\xi_i=1)=q$ and $P(\xi_i=0)=1-q$. Then say that $i\sim_{\Pi'} j$ if and only if $i\sim_{\Pi} j$ and $\xi_i=\xi_j=0$. Mark the block of Π' consisting of all integers i in the marked block of Π such that $\xi_i=0$.

The next two propositions establish the connection between the second stage of the construction of Π and the second stage of the selective sweep. Proposition 2.10 shows that it is unlikely to have $Z_i=Z_j$ if both are at least $\lfloor Js\rfloor$, just as Proposition 2.5 shows that lineages are unlikely to coalesce during the second stage of a selective sweep. Likewise, Proposition 2.11 shows that the number of Z_i greater than $\lfloor Js\rfloor$ has approximately a binomial distribution, just as Proposition 2.6 shows that the number of lineages that escape the selective sweep during the second stage has approximately a binomial distribution.

PROPOSITION 2.10. *For all* $i\neq j$, $P(Z_i=Z_j>\lfloor Js\rfloor)\le C/(\log N)^5$.

PROPOSITION 2.11. *Let* $D=\#\{i: Z_i>\lfloor Js\rfloor\}$, *and define* q_J *as in Proposition* 2.2. *Then*

$$\left|P(D=d)-\binom{n}{d}q_J^d(1-q_J)^{n-d}\right|\le\frac{C}{(\log N)^5}\qquad \text{for } d=0,1,\dots,n.$$

Propositions 2.8, 2.10 and 2.11 are proved in Section 7. The proofs of Propositions 2.10 and 2.11 are straightforward, but the proof of Proposition 2.8 is more difficult. It involves considering marked partitions π with different numbers of blocks and doing combinatorial calculations in each case.

PROOF OF THEOREM 1.2. By Propositions 2.7 and 2.8, we have $|P(\Psi_n=\pi)-Q_{r,s,\lfloor Js\rfloor}(\pi)|\le C/(\log N)^2$ for all $\pi\in\mathcal{P}_n$. It follows from this fact, Proposition 2.6 and the construction of Θ_2 that $|P(\Theta_2=\pi)-Q_{r,s,\lfloor Js\rfloor,q_J}(\pi)|\le C/(\log N)^2$ for all $\pi\in\mathcal{P}_n$. Also, by defining $\xi_i=\mathbb{1}_{\{Z_i>\lfloor Js\rfloor\}}$ and applying Propositions 2.10 and 2.11, we see that $|Q_{r,s,\lfloor Js\rfloor,q_J}(\pi)-Q_{r,s,L}(\pi)|\le C/(\log N)^5$ for all $\pi\in\mathcal{P}_n$. This observation, combined with the discussion in Part 1 of this subsection, completes the proof of Theorem 1.2. $\square$

3. Recombination of one lineage. Our goal in this section is to prove Propositions 2.1 and 2.2, which pertain to the recombination probabilities for a single lineage. The strategy will be to study the process $X=(X_t)_{t=0}^{\tau}$, which describes how the number of individuals with the B allele evolves during the selective sweep, and then calculate recombination probabilities conditional on the process X. In Section 3.1, we show that the process X behaves like an asymmetric random walk and we work out some calculations that will be needed later. We prove Proposition 2.1 in Section 3.2 and Proposition 2.2 in Section 3.3.

3.1. *Random walk calculations.* Suppose $1 \le X_{t-1} \le 2N - 1$. Then $X_t = X_{t-1} + 1$ if and only if $B_{t-1}(I_{t,1}) = 0$ and $B_{t-1}(I_{t,2}) = 1$. Also, $X_t = X_{t-1} - 1$ if and only if $B_{t-1}(I_{t,1}) = 0$, $B_{t-1}(I_{t,2}) = 1$ and $I_{t,4} = 0$. Otherwise, $X_t = X_{t-1}$. It follows that, for $1 \le k \le 2N - 1$,

$$P'(X_t = X_{t-1} + 1 | X_{t-1} = k) = \left(\frac{2N-k}{2N}\right)\left(\frac{k}{2N}\right), \tag{3.1}$$

$$P'(X_t = X_{t-1} - 1 | X_{t-1} = k) = \left(\frac{2N-k}{2N}\right)\left(\frac{k}{2N}\right)(1-s), \tag{3.2}$$

$$P'(X_t = X_{t-1} | X_{t-1} = k) = 1 - \frac{(2-s)k(2N-k)}{(2N)^2}. \tag{3.3}$$

Let $S_0 = 0$ and, for $m \ge 1$, let $S_m = \inf\{t > S_{m-1} : X_t \ne X_{S_{m-1}}\}$ be the time of the mth jump. It follows from (3.1) and (3.2) that the process $(X_{S_m})_{m=0}^{\infty}$ is a random walk on $\{0, 1, \dots, 2N\}$ that starts at 1, at each step moves to the right with probability $1/(2-s)$ and to the left with probability $(1-s)/(2-s)$, and is absorbed when it first reaches 0 or $2N$. A standard calculation for random walks (see, e.g., Section 3.1 of [6]) gives the following result.

LEMMA 3.1. *Let $p(a, b, k) = P'(\inf\{s > t : X_s = b\} < \inf\{s > t : X_s = a\} | X_t = k)$ be the probability that if the number of B's is k, then the number of B's will reach b before a. For $0 \le a < k < b \le 2N$,*

$$p(a, b, k) = \frac{1 - (1-s)^{k-a}}{1 - (1-s)^{b-a}}$$

and

$$P(X_\tau = 2N) = p(0, 2N, 1) = \frac{s}{1 - (1-s)^{2N}}.$$

Given $1 \le j \le 2N - 1$ and $1 \le k \le 2N - 1$, we define the quantities

$$\begin{aligned} \text{up jumps} \quad & U_{k,j} = \#\{t \ge \tau_j : X_t = k \text{ and } X_{t+1} = k+1\}, \\ \text{down jumps} \quad & D_{k,j} = \#\{t \ge \tau_j : X_t = k \text{ and } X_{t+1} = k-1\}, \\ \text{holds} \quad & H_{k,j} = \#\{t \ge \tau_j : X_t = k \text{ and } X_{t+1} = k\}, \\ \text{total} \quad & T_{k,j} = U_{k,j} + D_{k,j} + H_{k,j}. \end{aligned}$$

Also, let $U_k = U_{k,1}$, $D_k = D_{k,1}$, $H_k = H_{k,1}$ and $T_k = T_{k,1}$. The expected values of these quantities are given in the following lemma.

LEMMA 3.2. *Suppose* $1 \le j \le 2N - 1$ *and* $1 \le k \le 2N - 1$. *Define*

$$q_k = \frac{p(k, 2N, k+1)}{p(0, 2N, k+1)} = \frac{s}{(1 - (1-s)^{2N-k})} \cdot \frac{(1 - (1-s)^{2N})}{(1 - (1-s)^{k+1})} \ge s.$$

Also define $q_0 = 1$. *Define* $r_{k,j} = 1$ *for* $j \le k$ *and let* $r_{0,j} = 0$. *If* $j > k$, *let*

$$r_{k,j} = 1 - \frac{p(k, 2N, j)}{p(0, 2N, j)} = 1 - \frac{(1-(1-s)^{j-k})}{(1-(1-s)^{2N-k})} \cdot \frac{(1-(1-s)^{2N})}{(1-(1-s)^{j})} \le (1-s)^{j-k}.$$

Then $E[U_{k,j}] = r_{k,j}/q_k$. *Also*, $E[D_{k,j}] = (1/q_{k-1}) - 1$ *for* $k > j$ *and* $E[D_{k,j}] = r_{k-1,j}/q_{k-1}$ *for* $k \le j$. *Furthermore*,

$$E[H_{k,j}] = E[U_{k,j} + D_{k,j}]\left(\frac{1}{2-s}\right)\frac{1}{\beta_k} \le \frac{\min\{(1-s)^{j-k}, 1\}}{s\beta_k}, \tag{3.4}$$

where $\beta_k = k(2N-k)/(k^2 + (2N-k)^2 + sk(2N-k))$.

PROOF. First, suppose $k \ge j$. On the event $\{X_\tau = 2N\}$, we have $X_t = k$ and $X_{t+1} = k + 1$ for some $t \ge \tau_j$. Note that $P'(X_s > k \text{ for all } s > t | X_t = k+1) = p(k, 2N, k+1)$ for all t, so $P(X_s > k \text{ for all } s > t | X_t = k+1) = p(k, 2N, k+1)/p(0, 2N, k+1) = q_k$. It follows that the distribution of $U_{k,j}$ is Geometric(q_k), so $E[U_{k,j}] = 1/q_k$. If instead $k < j$, then $P(X_t > k$ for all $t > \tau_j) = p(k, 2N, j)/p(0, 2N, j)$. Therefore, $P(T_{k,j} \ge 1) = r_{k,j}$. It follows from the strong Markov property that, conditional on $T_{k,j} \ge 1$, the distribution of $U_{k,j}$ is Geometric(q_k), so $E[U_{k,j}] = r_{k,j}/q_k$.

To find $E[D_{k,j}]$, note that if $k > j$, then X takes a downward step from k to $k-1$ after each step from $k-1$ to k except the last one, so $D_{k,j} = U_{k-1,j} - 1$. If $k \le j$, then the number of steps after τ_j from k to $k-1$ is the same as the number of steps from $k-1$ to k, so $D_{k,j} = U_{k-1,j}$. The formulas for $E[D_{k,j}]$ follow immediately from these observations.

Let $p_k = P(X_t \ne X_{t-1} | X_{t-1} = k)$. To prove (3.4), note that (3.3) gives

$$p_k = \frac{k(2N-k)(2-s)}{(2N)^2}.$$

It follows that the conditional distribution of $T_{k,j}$ given $U_{k,j}$ and $D_{k,j}$ is the same as the distribution of the sum of $U_{k,j} + D_{k,j}$ independent random variables with a Geometric(p_k) distribution. Therefore,

$$E[H_{k,j}] = E[T_{k,j}] - E[U_{k,j}] - E[D_{k,j}] = E[U_{k,j} + D_{k,j}]\left(\frac{1}{p_k} - 1\right).$$

Straightforward algebraic manipulations give $1/p_k - 1 = 1/[\beta_k(2-s)]$, which implies the equality in (3.4). To check the inequality in (3.4), note that if $k > j$, then

$$E[U_{k,j} + D_{k,j}] = \frac{1}{q_k} + \frac{1}{q_{k-1}} - 1 \le \frac{1}{s} + \frac{1}{s} - 1 = \frac{2-s}{s},$$

and if $k \le j$, then

$$\begin{aligned} E[U_{k,j} + D_{k,j}] &\le \frac{(1-s)^{j-k}}{q_k} + \frac{(1-s)^{j-k+1}}{q_{k-1}} \\ &\le \frac{(1-s)^{j-k}}{s}\bigl(1 + (1-s)\bigr) \\ &= \frac{(2-s)(1-s)^{j-k}}{s}. \end{aligned} \qquad \square$$

We will now calculate the probability that the ancestor at time t has the opposite B or b allele from the ancestor at time $t-1$, given that $X_{t-1} = k$ and $X_t = l$, where $1 \le k \le 2N-1$, $1 \le l \le 2N$ and $|k-l| \le 1$. All of these recombination probabilities are the same under P' and P because of the conditioning on X_{t-1} and X_t. We define

$$\begin{aligned} p_B^r(k,l) &= P\bigl(B_{t-1}(A_t^{t-1}(i)) = 0 | X_{t-1} = k, X_t = l, B_t(i) = 1\bigr), \\ p_b^r(k,l) &= P\bigl(B_{t-1}(A_t^{t-1}(i)) = 1 | X_{t-1} = k, X_t = l, B_t(i) = 0\bigr). \end{aligned}$$

LEMMA 3.3. *We have*

$$\begin{aligned} &p_B^r(k,k-1) = p_b^r(k,k+1) = 0, \\ &p_B^r(k,k+1) = \frac{r(2N-k)}{(k+1)(2N)}, \\ &p_b^r(k,k-1) = \frac{rk}{(2N-k+1)(2N)}, \\ &p_B^r(k,k) = p_b^r(k,k) = \frac{rk(2N-k)}{2N[k^2 + (2N-k)^2 + sk(2N-k)]} = \frac{r\beta_k}{2N}. \end{aligned}$$

PROOF. We will prove three of the six results; the others are similar. If $X_{t-1} = k$ and $X_t = k+1$, then the new individual born at time t has a B allele. Therefore, if $B_t(i) = 0$, then $B_{t-1}(A_t^{t-1}(i)) = 0$, so $p_b^r(k,k+1) = 0$. Suppose instead $B_t(i) = 1$. Then $B_{t-1}(A_t^{t-1}(i)) = 0$ if and only if $I_{t,1} = i$ (meaning that the ith individual is the new one born), $I_{t,5} = 1$ (meaning that there is recombination) and $B_{t-1}(I_{t,3}) = 0$ (meaning that the new individual gets its allele at the neutral site from the member of the b population). Conditional on $X_{t-1} = k$, $X_t = k+1$ and $B_t(i) = 1$, the probabilities of $I_{t,1} = i$, $I_{t,5} = 1$ and $B_{t-1}(I_{t,3}) = 0$ are $1/(k+1)$, r and $(2N-k)/(2N)$, respectively. Multiplying them gives the expression for $p_B^r(k,k+1)$. To calculate $p_B^r(k,k)$ we use the fact that, conditional on $X_{t-1} = k$ and $X_t = k$, the probability that $B_{t-1}(I_{t,1}) = B_{t-1}(I_{t,2}) = 1$ is $k^2/[k^2 + (2N-k)^2 + sk(2N-k)]$. Multiplying by $1/k$, r and $(2N-k)/(2N)$ gives $p_B^r(k,k)$. $\square$

3.2. *Bounding the probability of two recombinations.* We now begin working toward a proof of Proposition 2.1, which shows that it is unlikely that a lineage will go from the B population to the b population and then back to the B population because of two recombination events. We begin by proving two simple lemmas. Lemma 3.4 bounds the probability that the number of individuals with the B allele is k at the recombination time $R(i)$. Lemma 3.5 is a useful deterministic result, which can be proved easily by splitting the sum into terms with $j \le N/2$ and $j > N/2$.

LEMMA 3.4. *We have* $P(X_{R(i)} = k) \le r/ks$.

PROOF. Considering the cases $X_{R(i)+1} = k + 1$ and $X_{R(i)+1} = k$ and using Lemmas 3.2 and 3.3,

$$\begin{aligned} P\big(X_{R(i)} = k\big) &\le p_B^r(k, k+1)E[U_k] + p_B^r(k,k)E[H_k] \\ &\le \frac{r(2N-k)}{(k+1)(2Ns)} + \frac{r}{2Ns} \le \frac{r(2N-k)+rk}{2Nks} = \frac{r}{ks}. \qquad \square \end{aligned}$$

LEMMA 3.5. *If* $a > 1$, *there is a* C *depending on* a *but not on* N *so that* $\sum_{j=1}^{N} a^j/j \le Ca^N/N$.

PROOF OF PROPOSITION 2.1. Denote the time of the second recombination event by $R_2(i) = \sup\{t \le R(i) : B_t(A_\tau^t(i)) = 1\}$, where $\sup\varnothing = -\infty$. Our goal is to show $P(R_2(i) \ge 0) \le C/(\log N)^2$. Note that by symmetry, the conditional distribution of $(X_t)_{t=0}^{\tau-1}$ given $X_\tau = 2N$ is the same as the conditional distribution of $(2N - X_{\tau-t})_{t=1}^{\tau}$ given $X_\tau = 2N$. It follows from this fact and the strong Markov property that

$$\begin{aligned} E\big[\#\big\{t < R(i) : X_t = k \text{ and } X_{t+1} = k+1 | X_{R(i)} = j\big\}\big] &= E[U_{2N-k-1,2N-j}], \\ E\big[\#\big\{t < R(i) : X_t = k \text{ and } X_{t+1} = k-1 | X_{R(i)} = j\big\}\big] &= E[D_{2N-k+1,2N-j}], \\ E\big[\#\big\{t < R(i) : X_t = k \text{ and } X_{t+1} = k | X_{R(i)} = j\big\}\big] &= E[H_{2N-k,2N-j}]. \end{aligned}$$

Therefore, by Lemmas 3.2 and 3.3,

$$\begin{aligned} &P\big(X_{R_2(i)} = k | X_{R(i)} = j\big) \\ &\quad \le p_b^r(k, k-1)E[D_{2N-k+1,2N-j}] + p_b^r(k,k)E[H_{2N-k,2N-j}] \\ &\quad \le \frac{rk}{(2N-k+1)(2Ns)} \min\{(1-s)^{k-j}, 1\} + \frac{r}{2Ns} \min\{(1-s)^{k-j}, 1\} \\ &\quad \le \frac{r}{(2N-k)s} \min\{(1-s)^{k-j}, 1\}. \end{aligned}$$

Using Lemma 3.4,

$$\begin{aligned} P\big(R_2(i) \geq 0\big) &\leq \sum_{j=1}^{2N-1} \frac{r}{js}\left(\sum_{k=1}^{2N-1} \frac{r\min\{(1-s)^{k-j}, 1\}}{(2N-k)s}\right) \\ &= \frac{r^2}{s^2}\sum_{j=1}^{2N-1}\frac{1}{j}\left(\sum_{k=j}^{2N-1}\frac{(1-s)^{k-j}}{2N-k} + \sum_{k=1}^{j-1}\frac{1}{2N-k}\right). \end{aligned} \tag{3.5}$$

Since $r^2/s^2 \leq C/(\log N)^2$, it suffices to show that the sum on the right-hand side of (3.5) is bounded as $N \to \infty$. We will handle the two terms separately. For the first term, we change variables $\ell = 2N - k$ and use Lemma 3.5 to get the bound

$$\begin{aligned} \sum_{j=1}^{2N-1}\frac{1}{j}\left(\sum_{k=j}^{2N-1}\frac{(1-s)^{k-j}}{2N-k}\right) &= \sum_{j=1}^{2N-1}\frac{(1-s)^{2N-j}}{j}\left(\sum_{\ell=1}^{2N-j}\left(\frac{1}{1-s}\right)^{\ell}\frac{1}{\ell}\right) \\ &\leq C\sum_{j=1}^{2N-1}\frac{1}{j(2N-j)} \leq \frac{2C}{N}\sum_{j=1}^{N}\frac{1}{j} \\ &\leq \frac{2C(1+\log N)}{N}. \end{aligned} \tag{3.6}$$

The second term in the sum on the right-hand side of (3.5) can be bounded by

$$\sum_{j=1}^{N}\frac{1}{j}\left(\frac{j}{N}\right) + \sum_{j=N+1}^{2N-1}\frac{1}{N}\left(\sum_{\ell=2N-j}^{2N-1}\frac{1}{\ell}\right) = 1 + \frac{1}{N}\sum_{\ell=1}^{2N-1}\frac{1}{\ell}\sum_{j=2N-l}^{2N-1}(1) \leq 3. \qquad \square$$

3.3. *Estimating the recombination probability.* Our next goal is to prove Proposition 2.2, which gives an approximation for $P(R(i) \geq \tau_J)$. The idea behind the proof is that every time there is a change in the population, there is some probability that a lineage will escape the selective sweep at that time, given that it has not previously escaped. Since the individual probabilities are small, if they sum to S, we will be able to approximate the probability that the lineage never escapes by e^{-S}. It will be easier to work with conditional escape probabilities given X, so to justify the approximation, it will be necessary to show that the sum of the conditional probabilities has low variance.

For $1 \leq t \leq \tau$, let $\theta_t = p_B^r(X_{t-1}, X_t)$. Now, θ_t is the conditional probability, given X, that a lineage escapes at time t if it has not previously escaped, so we have

$$P\big(R(i) \geq \tau_J | X\big) = 1 - \prod_{t=\tau_J+1}^{\tau}[1 - p_B^r(X_{t-1}, X_t)] = 1 - \prod_{t=\tau_J+1}^{\tau}(1-\theta_t). \tag{3.7}$$

To estimate the probability that a lineage escapes after time τ_J, we will consider the sum of these conditional probabilities, which we denote by $\eta_J = \sum_{t=\tau_J+1}^{\tau}\theta_t$.

The next lemma shows that to estimate $P(R(i) \geq \tau_J)$ to within an error of $O((\log N)^{-2})$, it suffices to calculate $E[e^{-\eta_J}]$.

LEMMA 3.6. *For all J, we have* $|P(R(i) \geq \tau_J) - (1 - E[e^{-\eta_J}])| \leq C/(\log N)^2$.

PROOF. It follows from the Poisson approximation on page 140 of [5] that

$$\bigl|P\bigl(R(i) \geq \tau_J | X\bigr) - (1 - e^{-\eta_J})\bigr| \leq \sum_{t=\tau_J+1}^{\tau} \theta_t^2. \tag{3.8}$$

By taking expectations, we get

$$\bigl|P\bigl(R(i) \geq \tau_J\bigr) - (1 - E[e^{-\eta_J}])\bigr| \leq E\Biggl[\sum_{t=\tau_J+1}^{\tau} \theta_t^2\Biggr].$$

It now remains to bound $E[\sum_{t=1}^{\tau} \theta_t^2]$. By Lemma 3.3,

$$\sum_{t=1}^{\tau} \theta_t^2 = \sum_{k=1}^{2N-1} \biggl(U_k \frac{r^2(2N-k)^2}{(k+1)^2(2N)^2} + H_k \frac{r^2\beta_k^2}{(2N)^2}\biggr).$$

Therefore, by Lemma 3.2,

$$\begin{aligned} E\Biggl[\sum_{t=1}^{\tau} \theta_t^2\Biggr] &\leq \sum_{k=1}^{2N} \biggl(\frac{r^2(2N-k)^2}{s(k+1)^2(2N)^2} + \frac{r^2\beta_k}{(2N)^2 s}\biggr) \\ &\leq \frac{r^2}{s} \sum_{k=1}^{2N} \biggl(\frac{1}{(k+1)^2} + \frac{1}{(2N)^2}\biggr) \leq Cr^2 \leq \frac{C}{(\log N)^2}, \end{aligned} \tag{3.9}$$

which completes the proof. □

The next result will allow us to work with a truncated version of the sum.

LEMMA 3.7. *If* $\eta_J' = \sum_{t=\tau_J+1}^{\tau} \theta_t \mathbb{1}_{\{X_{t-1} \geq J\}}$, *then* $E[\eta_J - \eta_J'] \leq C/J(\log N)$.

PROOF. Using Lemmas 3.2, 3.3 and 3.5, we get

$$\begin{aligned} E[\eta_J - \eta_J'] &= \sum_{k=1}^{J-1} \bigl(p_B^r(k,k+1)E[U_{k,J}] + p_B^r(k,k)E[H_{k,J}]\bigr) \\ &\leq \sum_{k=1}^{J-1} \frac{r(2N-k)}{(k+1)(2N)} \cdot \frac{(1-s)^{J-k}}{s} + \frac{r\beta_k}{2N} \cdot \frac{(1-s)^{J-k}}{s\beta_k} \\ &\leq \sum_{k=1}^{J-1} (1-s)^{J-k} \biggl(\frac{r}{ks}\biggr) = \frac{r}{s} \sum_{k=1}^{J-1} \frac{1}{k} \biggl(\frac{1}{1-s}\biggr)^{k-J} \leq \frac{Cr}{sJ}. \end{aligned}$$

□

We will work with η'_J rather than η_J because we can obtain a rather precise estimate of its expected value, which is given in the next lemma. We will also be able to obtain a bound on its variance, which will enable us to approximate $E[e^{-\eta'_J}]$ by $e^{-E[\eta'_J]}$.

LEMMA 3.8. $E[\eta'_J] = \frac{r}{s}\sum_{k=J+1}^{2N}\frac{1}{k} + O(\frac{1}{N} + \frac{(1-s)^J}{J\log N})$.

PROOF. It follows from Lemma 3.2 and a straightforward calculation that

$$\begin{aligned} E[H_k] &= \left(\frac{1}{q_k} + \frac{1}{q_{k-1}} - 1\right)\left(\frac{1}{\beta_k(2-s)}\right) \\ &= \frac{1}{s\beta_k}\left(\frac{(1-(1-s)^k)(1-(1-s)^{2N-k})}{1-(1-s)^{2N}}\right). \end{aligned}$$

Therefore,

$$\begin{aligned} E[\eta'_J] &= \sum_{k=J}^{2N-1}\left(\frac{r(2N-k)}{(k+1)(2N)}E[U_k] + \frac{r\beta_k}{2N}E[H_k]\right) \\ &= \sum_{k=J}^{2N-1}\left(\frac{r(2N-k)(1-(1-s)^{k+1})(1-(1-s)^{2N-k})}{(k+1)(2Ns)(1-(1-s)^{2N})}\right. \\ &\qquad\qquad \left. + \frac{r(1-(1-s)^k)(1-(1-s)^{2N-k})}{(2Ns)(1-(1-s)^{2N})}\right) \\ &= \frac{r}{s}\sum_{k=J}^{2N-1}\left(\frac{1-(1-s)^{2N-k}}{1-(1-s)^{2N}}\right) \\ &\qquad \times\left(\frac{(2N-k)(1-(1-s)^{k+1})}{(2N)(k+1)} + \frac{1-(1-s)^k}{2N}\right). \end{aligned}$$

Now

$$\begin{aligned} &\sum_{k=J}^{2N-1}\left(1 - \frac{1-(1-s)^{2N-k}}{1-(1-s)^{2N}}\right)\left(\frac{(2N-k)(1-(1-s)^{k+1})}{(2N)(k+1)} + \frac{1-(1-s)^k}{2N}\right) \\ &\quad \le \sum_{k=J}^{2N-1}(1-s)^{2N-k}\left(\frac{1}{k} + \frac{1}{2N}\right) \\ &\quad \le \sum_{k=1}^{N}(1-s)^N\left(\frac{2}{k}\right) + \sum_{k=N+1}^{2N}(1-s)^{2N-k}\left(\frac{2}{N}\right) \\ &\quad \le 2(1+\log N)(1-s)^N + \frac{2}{Ns} \le \frac{C}{N}. \end{aligned}$$

Therefore,

$$E[\eta'_J] = \frac{r}{s}\sum_{k=J}^{2N-1}\left(\frac{(2N-k)(1-(1-s)^{k+1})}{(2N)(k+1)} + \frac{1-(1-s)^k}{2N}\right) + O\left(\frac{1}{N}\right).$$

Also, note that

$$\sum_{k=J}^{2N-1}\frac{(1-(1-s)^{k+1})-(1-(1-s)^k)}{2N} = \sum_{k=J}^{2N}\frac{(1-s)^k s}{2N} \leq \frac{1-s}{2N}.$$

Therefore, since $r \leq C_0/\log N$,

$$\begin{aligned}E[\eta'_J] &= \frac{r}{s}\sum_{k=J}^{2N-1}\left(\frac{2N-k}{(2N)(k+1)} + \frac{1}{2N}\right)(1-(1-s)^{k+1}) + O\left(\frac{1}{N}\right)\\ &= \frac{r}{s}\left(\frac{2N+1}{2N}\right)\sum_{k=J}^{2N-1}\frac{1-(1-s)^{k+1}}{k+1} + O\left(\frac{1}{N}\right) \qquad (3.10)\\ &= \frac{r}{s}\sum_{k=J+1}^{2N}\frac{1-(1-s)^k}{k} + O\left(\frac{1}{N}\right).\end{aligned}$$

Since

$$\frac{r}{s}\sum_{k=J+1}^{2N}\frac{(1-s)^k}{k} \leq \frac{r}{s(J+1)}\sum_{k=J+1}^{\infty}(1-s)^k = \frac{r(1-s)^{J+1}}{s^2(J+1)},$$

the desired result follows from (3.10). □

The key remaining step is to bound $\mathrm{Var}(\eta'_J)$. The necessary bound is given in Lemma 3.10. The proof uses Lemma 3.9, which can easily be proved by conditioning on M and N.

LEMMA 3.9. *Suppose $(X_i)_{i=1}^{\infty}$ and $(Y_i)_{i=1}^{\infty}$ are independent i.i.d. sequences such that $E[X_1] = \mu$ and $E[Y_1] = \gamma$. Suppose M and N are integer-valued random variables that are independent of these sequences. Then* $\mathrm{Cov}(X_1 + \cdots + X_M, Y_1 + \cdots + Y_N) = \mu\gamma\,\mathrm{Cov}(M, N)$.

LEMMA 3.10. *There exists a constant C such that* $\mathrm{Var}(\eta'_J) \leq C/J(\log N)^2$.

PROOF. Let

$$a_k = \frac{2N-k}{(k+1)(2N)} \leq \frac{1}{k}, \qquad (3.11)$$

$$b_k = \frac{k(2N-k)}{2N[k^2+(2N-k)^2+sk(2N-k)]} \leq \frac{k(2N-k)}{2N^3}. \qquad (3.12)$$

Then $\eta'_J = \sum_{t=\tau_J+1}^{\tau} \theta_t \mathbb{1}_{\{X_{t-1} \ge J\}} = r \sum_{k=J}^{2N-1} (a_k U_k + b_k H_k)$. For any random variables X and Y,

$$\begin{aligned} \mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) \\ &\le \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \le 4\max\{\mathrm{Var}(X), \mathrm{Var}(Y)\}. \end{aligned}$$

Therefore,

$$\mathrm{Var}(\eta'_J) \le 4r^2 \max\left\{ \mathrm{Var}\left(\sum_{k=J}^{2N-1} a_k U_k \right), \mathrm{Var}\left(\sum_{k=J}^{2N-1} b_k H_k \right) \right\}. \tag{3.13}$$

We will bound $\mathrm{Var}(\sum_{k=J}^{2N-1} a_k U_k)$ and $\mathrm{Var}(\sum_{k=J}^{2N-1} b_k H_k)$ by C/J, which will prove the lemma.

To bound $\mathrm{Var}(\sum_{k=J}^{2N-1} a_k U_k)$, we will need to bound $\mathrm{Cov}(U_k, U_l)$. To do this, we will break up U_l into jumps from l to $l+1$ that occur before the last visit to k and those that occur after the last visit to k. More formally, let $\zeta_k = \sup\{t : X_t = k\}$. If $k \le l$, then $U_l = U'_{k,l} + \bar{U}_{k,l}$, where

$$\begin{aligned} U'_{k,l} &= \#\{t \ge \zeta_k : X_t = l \text{ and } X_{t+1} = l+1\}, \\ \bar{U}_{k,l} &= \#\{t < \zeta_k : X_t = l \text{ and } X_{t+1} = l+1\}. \end{aligned}$$

The processes $(X_t)_{0 \le t \le \zeta_k}$ and $(X_t)_{\zeta_k \le t \le \tau}$ are independent. Therefore, U_k and $U'_{k,l}$ are independent, and $\bar{U}_{k,l}$ and $U'_{k,l}$ are independent. As observed in the proof of Lemma 3.2, U_l has a Geometric(q_l) distribution. Likewise, note that $P'(X_s > l$ for all $s > t | X_t = l+1) = p(l, 2N, l+1)$ and $P'(X_s > k$ for all $s > t | X_t = l+1) = p(k, 2N, l+1)$. Therefore,

$$P(X_s > l \text{ for all } s > t | X_t = l+1, X_s > k \text{ for all } s > t) = \frac{p(l, 2N, l+1)}{p(k, 2N, l+1)}.$$

It follows that if we let $v_{k,l} = p(l, 2N, l+1)/p(k, 2N, l+1)$, then $U'_{k,l}$ has a Geometric($v_{k,l}$) distribution. Using Lemmas 3.1 and 3.2 and the fact that $q_l = p(l, 2N, l+1)/p(0, 2N, l+1)$, we have

$$\begin{aligned} \frac{1}{q_l} - \frac{1}{v_{k,l}} &= \frac{1-(1-s)^{2N-l}}{s} \left(\frac{1-(1-s)^{l+1}}{1-(1-s)^{2N}} - \frac{1-(1-s)^{l+1-k}}{1-(1-s)^{2N-k}} \right) \\ &\le \frac{1}{s}\bigl(1 - \bigl(1 - (1-s)^{l+1-k}\bigr)\bigr) = \frac{(1-s)^{l+1-k}}{s}. \end{aligned} \tag{3.14}$$

Also, $\mathrm{Var}(U_l) = \mathrm{Var}(U'_{k,l}) + \mathrm{Var}(\bar{U}_{k,l})$ because $\bar{U}_{k,l}$ and $U'_{k,l}$ are independent. Therefore, if $J \le k \le l < 2N$, then by the formula for the variance of a geometric distribution,

$$\begin{aligned} \mathrm{Var}(\bar{U}_{k,l}) &= \mathrm{Var}(U_l) - \mathrm{Var}(U'_{k,l}) = \frac{1-q_l}{q_l^2} - \frac{1-v_{k,l}}{v_{k,l}^2} \\ &= \left(\frac{1}{q_l} + \frac{1}{v_{k,l}} - 1 \right)\left(\frac{1}{q_l} - \frac{1}{v_{k,l}} \right) \le \frac{2}{s} \cdot \frac{(1-s)^{l-k}}{s}, \end{aligned} \tag{3.15}$$

where the inequality uses (3.14) and the facts that $q_l \geq s$ and $v_{k,l} \geq s$. Also

$$\operatorname{Var}(U_l) = \frac{1-q_l}{q_l^2} \leq \frac{1}{s^2}. \tag{3.16}$$

Since U_k and $U'_{k,l}$ are independent, it follows from (3.15) and (3.16) that if $k \leq l$, then

$$\begin{aligned}\operatorname{Cov}(U_k, U_l) &= \operatorname{Cov}(U_k, U'_{k,l} + \bar{U}_{k,l}) = \operatorname{Cov}(U_k, \bar{U}_{k,l}) \\ &\leq \sqrt{\operatorname{Var}(U_k)\operatorname{Var}(\bar{U}_{k,l})} \leq \frac{\sqrt{2}}{s^2}(1-s)^{(l-k)/2}.\end{aligned} \tag{3.17}$$

Using (3.17) and (3.11), we calculate

$$\begin{aligned}\operatorname{Var}\left(\sum_{k=J}^{2N-1} a_k U_k\right) &= \sum_{k=J}^{2N-1}\sum_{l=J}^{2N-1} a_k a_l \operatorname{Cov}(U_k, U_l) \\ &\leq \frac{2\sqrt{2}}{s^2}\sum_{k=J}^{2N-1}\sum_{l=k}^{2N-1} \frac{1}{kl}(1-s)^{(l-k)/2} \\ &\leq \frac{2\sqrt{2}}{s^2}\sum_{k=J}^{2N-1}\frac{1}{k^2}\left(\sum_{l=k}^{2N-1}(1-s)^{(l-k)/2}\right) \\ &\leq C\sum_{k=J}^{2N-1}\frac{1}{k^2} \leq \frac{C}{J}.\end{aligned} \tag{3.18}$$

It remains to bound $\operatorname{Var}(\sum_{k=J}^{2N-1} b_k H_k)$. Recall from the proof of Lemma 3.2 that

$$p_k = P(X_t \neq X_{t-1} | X_{t-1} = k) = \frac{k(2N-k)(2-s)}{(2N)^2}$$

and that $D_k + U_k = U_{k-1} - 1 + U_k$, using the convention that $U_0 = 1$. Therefore, we can write $H_k = G_1 + G_2 + \cdots + G_{U_k + U_{k-1} - 1}$, where $(G_i)_{i=1}^{\infty}$ is an i.i.d. sequence of random variables such that $G_i + 1$ has a Geometric(p_k) distribution for all i. Thus, $E[G_i] = p_k^{-1} - 1$. If $k \leq l$, then by Lemma 3.9,

$$\begin{aligned}\operatorname{Cov}(H_k, H_l) &= \left(\frac{1}{p_k} - 1\right)\left(\frac{1}{p_l} - 1\right)\operatorname{Cov}(U_k + U_{k-1} - 1, U_l + U_{l-1} - 1) \\ &\leq \frac{1}{p_k p_l}\operatorname{Cov}(U_k + U_{k-1}, U_l + U_{l-1}) \\ &\leq \frac{4\sqrt{2}}{s^2 p_k p_l}(1-s)^{(l-k-1)/2} \leq \frac{C}{p_k p_l}(1-s)^{(l-k)/2}.\end{aligned}$$

Note that (3.12) implies

$$\frac{b_k}{p_k} \leq \frac{k(2N-k)}{2N^3} \cdot \frac{(2N)^2}{k(2N-k)(2-s)} = \frac{2}{(2-s)N} \leq \frac{2}{N}.$$

Therefore,

$$\begin{aligned}
\operatorname{Var}\left(\sum_{k=J}^{2N-1} b_k H_k\right) &= \sum_{k=J}^{2N-1}\sum_{l=J}^{2N-1} b_k b_l \operatorname{Cov}(H_k, H_l) \\
&\leq C \sum_{k=J}^{2N-1}\sum_{l=k}^{2N-1} \frac{b_k b_l}{p_k p_l}(1-s)^{(l-k)/2} \\
&\leq \frac{C}{N^2} \sum_{k=J}^{2N-1}\sum_{l=k}^{2N-1} (1-s)^{(l-k)/2} \\
&\leq \frac{C}{N^2} \sum_{k=J}^{2N-1} \frac{1}{1-\sqrt{1-s}} \leq \frac{C}{N}.
\end{aligned} \tag{3.19}$$

The lemma follows from (3.13), (3.18) and (3.19). □

PROOF OF PROPOSITION 2.2. Lemma 3.6 gives

$$\left|P\left(R(i) \geq \tau_J\right) - (1 - E[e^{-\eta_J}])\right| \leq E\left[\sum_{t=\tau_J+1}^{\tau} \theta_t^2\right] \leq \frac{C}{(\log N)^2}.$$

Since $|\frac{d}{dx}e^{-x}| \leq 1$ for $x \geq 0$, Lemma 3.7 gives

$$E\left[e^{-\eta'_J} - e^{-\eta_J}\right] \leq E[\eta_J - \eta'_J] \leq \frac{C}{J(\log N)}.$$

Using Jensen's inequality and Lemma 3.10,

$$\begin{aligned}
\left|E\left[e^{-\eta'_J}\right] - e^{-E[\eta'_J]}\right| &\leq E\left|e^{-\eta'_J} - e^{-E[\eta'_J]}\right| \\
&\leq E|\eta'_J - E[\eta'_J]| \leq \operatorname{Var}(\eta'_J)^{1/2} \leq \frac{C}{\sqrt{J}(\log N)}.
\end{aligned}$$

Furthermore, it follows from Lemma 3.8 that

$$1 - e^{-E[\eta'_J]} = q_J + O\left(\frac{1}{N} + \frac{(1-s)^J}{J\log N}\right).$$

Combining the last four equations gives the proposition. □

4. Coalescence of two lineages. In this section, we prove Propositions 2.3, 2.4 and 2.5, all of which pertain to the probabilities that two lineages in the sample coalesce. We begin by computing the following coalescence probabilities for integers k and l such that $1 \le k \le 2N-1$, $1 \le l \le 2N$ and $|k-l| \le 1$:

$$\begin{aligned} p_{BB}^c(k,l) &= P\big(A_t^{t-1}(i) = A_t^{t-1}(j) | X_{t-1} = k, X_t = l, B_t(i) = 1, B_t(j) = 1\big), \\ p_{bb}^c(k,l) &= P\big(A_t^{t-1}(i) = A_t^{t-1}(j) | X_{t-1} = k, X_t = l, B_t(i) = 0, B_t(j) = 0\big), \\ p_{Bb}^c(k,l) &= P\big(A_t^{t-1}(i) = A_t^{t-1}(j) | X_{t-1} = k, X_t = l, B_t(i) = 1, B_t(j) = 0\big). \end{aligned}$$

As with the recombination probabilities in the previous section, the Markov property implies that the coalescence probabilities are the same under P' as under P.

LEMMA 4.1. *We have*

$$\begin{aligned} p_{BB}^c(k,k-1) &= p_{bb}^c(k,k+1) = 0, \\ p_{BB}^c(k,k+1) &= \frac{2}{k(k+1)}\left(1 - \frac{r(2N-k)}{2N}\right), \\ p_{bb}^c(k,k-1) &= \frac{2}{(2N-k)(2N-k+1)}\left(1 - \frac{rk}{2N}\right), \\ p_{bb}^c(k,k) &= \frac{2\beta_k}{k(2N-k)}\left(1 - \frac{rk}{2N}\right), \\ p_{BB}^c(k,k) &= \frac{2\beta_k}{k(2N-k)}\left(1 - \frac{r(2N-k)}{2N}\right), \\ p_{Bb}^c(k,k) &= \frac{r\beta_k}{k(2N-k)}, \qquad p_{Bb}^c(k,k+1) = \frac{r}{2N(k+1)}, \\ p_{Bb}^c(k,k-1) &= \frac{r}{2N(2N-k+1)}. \end{aligned}$$

PROOF. This result follows from a series of straightforward calculations, similar to those used to prove Lemma 3.3. We explain the idea behind some of these calculations. When $X_{t-1} = k$ and $X_t = k-1$, the new individual born at time t has the b allele. Therefore, two B lineages cannot coalesce at this time, so $p_{BB}^c(k,k-1) = 0$. By the same reasoning, $p_{bb}^c(k,k+1) = 0$. When $X_{t-1} = k$ and $X_t = k+1$, the new individual born at time t has the B allele. With probability $r(2N-k)/(2N)$, this individual inherits its allele at the neutral site from a member of the b population because of recombination. If this does not happen, then two of the B individuals get their allele at the neutral site from the same parent. Thus, conditional on $B_t(i) = B_t(j) = 1$, the probability that the ith and jth individuals get their allele at the neutral site from the same parent is $2/[k(k+1)]$, which implies the formula for $p^c_{BB}(k,k+1)$. The calculation of $p^c_{bb}(k,k-1)$ is similar.

Now suppose $X_{t-1} = X_t = k$. Conditional on this event, a B replaces a B with probability $k^2/[k^2 + (2N-k)^2 + sk(2N-k)]$. If the new B gets its allele at the neutral site from a member of the B population, which has probability $1 - r(2N-k)/(2N)$, and if $B_t(i) = B_t(j) = 1$, then the probability that the ith and jth lineages coalesce is $2/k^2$, because there are k possibilities both for the individual who dies and the parent of the new individual. The formula for $p^c_{BB}(k,k)$ follows, and $p^c_{bb}(k,k)$ can be calculated in the same way. Next, to find $p^c_{Bb}(k,k)$, note that if a B replaces a B, and $B_t(i) = 1$ and $B_t(j) = 0$, then the probability of coalescence is $r/(2kN)$, because there must be recombination, and there are k choices for the B individual that is just born and $2N$ choices for the parent from which it gets its allele at the neutral site. If instead a b replaces a b, which happens with probability $(2N-k)^2/[k^2 + (2N-k)^2 + sk(2N-k)]$ conditional on $X_{t-1} = X_t = k$, the probability of coalescence is $r/[(2N-k)(2N)]$. Adding the probabilities for the two cases gives the formula for $p^c_{Bb}(k,k)$.

Finally, to calculate $p^c_{Bb}(k,k+1)$ and $p^c_{Bb}(k,k-1)$, note that when a B replaces a b, the probability that a B lineage coalesces with a b lineage is $r/[(k+1)(2N)]$, because there must be recombination, and there are $k+1$ choices for the B individual that was just born and $2N$ choices for its parent. Likewise, the coalescence probability is $r/[(2N-k+1)(2N)]$ when a b replaces a B. $\square$

PROOF OF PROPOSITION 2.3. We consider first the case in which the jth lineage is descended from a member of the B population at the time of coalescence. Summing over the possible values k for $X_{G(i,j)}$ and applying Lemmas 3.2 and 4.1, we get

$$\begin{aligned}
&P\big(G(i,j) \geq 0,\, B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(i)\big) = 0, \text{ and } B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(j)\big) = 1\big) \\
&\quad \leq \sum_{k=1}^{2N-1} \big(p^c_{Bb}(k,k+1)E[U_k] + p^c_{Bb}(k,k-1)E[D_k] + p^c_{Bb}(k,k)E[H_k]\big) \\
&\quad \leq \sum_{k=1}^{2N-1} \left(\frac{r}{2N(k+1)s} + \frac{r}{2N(2N-k+1)s} + \frac{r}{sk(2N-k)}\right) \\
&\quad \leq \frac{r}{2Ns} \sum_{k=1}^{2N-1} \left(\frac{1}{k} + \frac{1}{2N-k} + \frac{2N}{k(2N-k)}\right) \\
&\quad = \frac{2r}{s} \sum_{k=1}^{2N-1} \frac{1}{k(2N-k)} \leq \frac{4r}{Ns} \sum_{k=1}^{N} \frac{1}{k} \leq \frac{4r(1+\log N)}{Ns} \leq \frac{C}{N}.
\end{aligned}$$

It remains to consider the case in which the ith and jth lineages are both descended from a member of the b population at the coalescence time. By summing over the possible values of $X_{R(i)}$ and $X_{G(i,j)}$, we see that it suffices to show

$$\begin{aligned}&\sum_{\ell=1}^{2N-1}\sum_{k=1}^{2N-1} P\bigl(X_{R(i)}=\ell\bigr)P\bigl(X_{G(i,j)}=k, B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(i)\bigr)=0,\\ &\qquad\qquad\qquad\text{and } B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(j)\bigr)=0|X_{R(i)}=\ell\bigr) \qquad (4.1)\\ &\quad\le \frac{C(\log N)}{N}.\end{aligned}$$

If $B_{G(i,j)+1}(A_\tau^{G(i,j)+1}(i))=0$, then $G(i,j)\le R(i)$. Therefore, it follows from Lemmas 3.2 and 4.1 and the time-reversal argument in the proof of Proposition 2.1 that

$$\begin{aligned}&P\bigl(X_{G(i,j)}=k \text{ and}\\ &\quad B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(i)\bigr)=B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(j)\bigr)=0|X_{R(i)}=\ell\bigr)\\ &\quad\le p_{bb}^c(k,k-1)E[D_{2N-k+1,2N-\ell}]+p_{bb}^c(k,k)E[H_{2N-k,2N-\ell}]\\ &\quad\le \frac{2}{(2N-k)(2N-k+1)s}\min\{(1-s)^{k-\ell},1\}\\ &\qquad+\frac{2}{sk(2N-k)}\min\{(1-s)^{k-\ell},1\}\\ &\quad\le\left(\frac{2k+2(2N-k)}{sk(2N-k)^2}\right)\min\{(1-s)^{k-\ell},1\}\\ &\quad=\frac{4N\min\{(1-s)^{k-l},1\}}{sk(2N-k)^2}.\end{aligned}$$

Combining this result with Lemma 3.4, we get that the left-hand side of (4.1) is at most

$$\begin{aligned}&\sum_{\ell=1}^{2N-1}\frac{r}{\ell s}\left(\sum_{k=1}^{2N-1}\frac{4N\min\{(1-s)^{k-\ell},1\}}{sk(2N-k)^2}\right)\\ &\qquad\le\frac{4r}{s^2}\sum_{\ell=1}^{2N-1}\frac{1}{\ell}\left(\sum_{k=\ell}^{2N-1}\frac{N(1-s)^{k-\ell}}{k(2N-k)^2}+\sum_{k=1}^{\ell-1}\frac{N}{k(2N-k)^2}\right).\end{aligned}\tag{4.2}$$

Using (3.6) and the fact that $N/[k(2N-k)]\le 1$ for $1\le k\le 2N-1$, we get

$$\frac{4r}{s^2}\sum_{\ell=1}^{2N-1}\frac{1}{\ell}\left(\sum_{k=\ell}^{2N-1}\frac{N(1-s)^{k-\ell}}{k(2N-k)^2}\right)\le\frac{4r}{s^2}\left(\frac{2C(1+\log N)}{N}\right)\le\frac{C}{N}.\tag{4.3}$$

For the second term in (4.2), we have

$$\begin{aligned}
&\frac{4r}{s^2}\sum_{\ell=1}^{2N-1}\frac{1}{\ell}\left(\sum_{k=1}^{\ell-1}\frac{N}{k(2N-k)^2}\right)\\
&\quad\le \frac{4r}{s^2}\sum_{\ell=1}^{N}\frac{1}{\ell}\left(\sum_{k=1}^{\ell}\frac{N}{kN^2}\right)+\frac{4r}{s^2}\sum_{l=N+1}^{2N-1}\frac{1}{N}\left(\sum_{k=1}^{\ell}\frac{N}{k(2N-k)^2}\right)\\
&\quad\le \frac{4r}{Ns^2}\left(\sum_{\ell=1}^{N}\frac{1}{\ell}\right)^2+\frac{4r}{s^2}\sum_{k=1}^{2N-1}\sum_{\ell=k}^{2N-1}\frac{1}{k(2N-k)^2}\\
&\quad\le \frac{4r(1+\log N)^2}{Ns^2}+\frac{4r}{s^2}\cdot 2\sum_{k=1}^{N}\frac{1}{k(2N-k)}\le\frac{C(\log N)}{N}.
\end{aligned}\tag{4.4}$$

Using (4.3) and (4.4) in (4.2) proves (4.1). □

The next lemma, which bounds the probability that there are k individuals with the B allele at the time the ith and jth lineages coalesce, will be used in the proofs of Propositions 2.4 and 2.5.

LEMMA 4.2. *We have*

$$\begin{aligned}
&P\big(X_{G(i,j)}=k \text{ and}\\
&\quad B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(i)\big)=B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(j)\big)=1\big)\\
&\quad\le\frac{4N}{sk^2(2N-k)}.
\end{aligned}\tag{4.5}$$

PROOF. By Lemmas 3.2 and 4.1, the probability on the left-hand side of (4.5) is at most

$$\begin{aligned}
&p_{BB}^c(k,k+1)E[U_k]+p_{BB}^c(k,k)E[H_k]\\
&\quad\le\frac{2}{sk(k+1)}+\frac{2}{sk(2N-k)}\\
&\quad\le\frac{2(2N-k)+2k}{sk^2(2N-k)}=\frac{4N}{sk^2(2N-k)}.
\end{aligned}$$

□

PROOF OF PROPOSITION 2.4. By Proposition 2.3, it suffices to show that

$$\begin{aligned}
&P\big(0\le R(i)\le G(i,j) \text{ and}\\
&\quad B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(i)\big)=B_{G(i,j)+1}\big(A_\tau^{G(i,j)+1}(j)\big)=1\big)\le\frac{C}{\log N}.
\end{aligned}$$

By Lemmas 3.2 and 3.3 and the time-reversal argument in the proof of Proposition 2.1,

$$\begin{aligned}
&P\bigl(X_{R(i)}=\ell \text{ and } 0\le R(i)\le G(i,j)|X_{G(i,j)}=k\bigr)\\
&\quad\le p_B^r(\ell,\ell+1)E[U_{2N-\ell-1,2N-k}]+p_B^r(\ell,\ell)E[H_{2N-\ell,2N-k}]\\
&\quad\le \frac{r(2N-\ell)}{(\ell+1)(2Ns)}\min\{(1-s)^{\ell+1-k},1\}+\frac{r}{2Ns}\min\{(1-s)^{\ell-k},1\}\\
&\quad\le \frac{r}{\ell s}\min\{(1-s)^{\ell-k},1\}.
\end{aligned}$$

Combining this result with (4.5), we get

$$\begin{aligned}
&P\bigl(0\le R(i)\le G(i,j) \text{ and}\\
&\qquad B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(i)\bigr)=B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(j)\bigr)=1\bigr)\\
&\qquad\le \sum_{k=1}^{2N-1}\frac{4N}{sk^2(2N-k)}\left(\sum_{\ell=1}^{2N-1}\frac{r}{\ell s}\min\{(1-s)^{\ell-k},1\}\right)\\
&\qquad\le \frac{4r}{s^2}\sum_{k=1}^{2N-1}\frac{N}{k^2(2N-k)}\left(\sum_{\ell=k}^{2N-1}\frac{(1-s)^{\ell-k}}{\ell}+\sum_{\ell=1}^{k-1}\frac{1}{\ell}\right).
\end{aligned}\tag{4.6}$$

The first term in the sum on the right-hand side of (4.6) is at most

$$\sum_{k=1}^{2N-1}\frac{N}{k^3(2N-k)}\left(\sum_{\ell=k}^{2N-1}(1-s)^{\ell-k}\right)\le\frac{1}{s}\left(\sum_{k=1}^{N}\frac{1}{k^3}+\sum_{k=N+1}^{2N-1}\frac{1}{N^2(2N-k)}\right),$$

which is bounded by a constant. The other term in the sum in (4.6) is at most

$$\sum_{k=1}^{2N-1}\frac{N(1+\log k)}{k^2(2N-k)}\le\sum_{k=1}^{N}\frac{1+\log k}{k^2}+\sum_{k=N+1}^{2N-1}\frac{1+\log(2N)}{N(2N-k)},$$

which is also bounded by a constant. Since $4r/s^2\le C/(\log N)$, the proposition follows. □

PROOF OF PROPOSITION 2.5. By reasoning similar to that used to prove Lemma 4.2, we have

$$\begin{aligned}
&P\bigl(G(i,j)\ge\tau_J \text{ and}\\
&\qquad B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(i)\bigr)=B_{G(i,j)+1}\bigl(A_\tau^{G(i,j)+1}(i)\bigr)=1\bigr)\\
&\qquad\le\sum_{k=1}^{2N-1}\bigl(p_{BB}^c(k,k+1)E[U_{k,J}]+p_{BB}^c(k,k)E[H_{k,J}]\bigr).
\end{aligned}\tag{4.7}$$

However, this time we keep the factor $\min\{(1-s)^{J-k}, 1\}$ from Lemma 3.2 to bound the right-hand side of (4.7) by

$$\sum_{k=1}^{J}(1-s)^{J-k}\frac{4N}{sk^2(2N-k)}+\sum_{k=J+1}^{2N-1}\frac{4N}{sk^2(2N-k)}. \tag{4.8}$$

Using the fact that $N/[k(2N-k)] \le 1$ for $1 \le k \le 2N-1$ and then Lemma 3.5, we have

$$\sum_{k=1}^{J}(1-s)^{J-k}\frac{4N}{sk^2(2N-k)} \le \frac{4}{s}(1-s)^J\sum_{k=1}^{J}\left(\frac{1}{1-s}\right)^k\frac{1}{k} \le \frac{C}{J}.$$

For the second term in (4.8), we observe

$$\begin{aligned}\sum_{k=J+1}^{2N-1}\frac{4N}{sk^2(2N-k)} &\le \sum_{k=J+1}^{N-1}\frac{4}{sk^2}+\sum_{k=N}^{2N-1}\frac{4}{sN(2N-k)}\\ &\le \frac{4}{sJ}+\frac{4(1+\log N)}{Ns}.\end{aligned}$$

Since $J \le C'N/(\log N)$, the bounds in the last two equations add up to C/J, and the desired result follows from these bounds and Proposition 2.3. ☐

5. Approximate independence of *n* lineages.

In this section, we prove Proposition 2.6. We first establish a lemma that involves the coupling of two $\{0, 1, \dots, n\}$-valued random variables.

LEMMA 5.1. *Let V and V' be $\{0, 1, \dots, n\}$-valued random variables such that $E[V] = E[V']$. Then, there exist random variables $\tilde{V}$ and $\tilde{V}'$ on some probability space such that V and $\tilde{V}$ have the same distribution, V' and $\tilde{V}'$ have the same distribution, and*

$$P(\tilde{V} \ne \tilde{V}') \le n \max\{P(\tilde{V} \ge 2), P(\tilde{V}' \ge 2)\}.$$

PROOF. It is clear that $\tilde{V}$ and $\tilde{V}'$ can be constructed such that they have the same distributions as V and V', respectively, and $P(\tilde{V} = \tilde{V}') \ge \min\{P(V = 0), P(V' = 0)\} + \min\{P(V = 1), P(V' = 1)\}$. Note that $P(V = 0) \ge 1 - E[V]$. Since $E[V] = E[V']$, it follows that $\min\{P(V = 0), P(V' = 0)\} \ge 1 - E[V]$. Also, $P(V = 1) = E[V] - \sum_{k=2}^{n} kP(V = k)$, so $P(V = 1) \ge E[V] - nP(V \ge 2)$. Likewise, $P(V' = 1) \ge E[V] - nP(V' \ge 2)$. It follows that

$$P(\tilde{V} = \tilde{V}') \ge 1 - n \max\{P(\tilde{V} \ge 2), P(\tilde{V}' \ge 2)\}. \qquad \square$$

Recall that $K_t = \#\{i \in \{1, \dots, n\} : R(i) \ge t\}$ for $0 \le t \le \tau$. Define $\theta_t = p_B^r(X_{t-1}, X_t)$ as in Section 3, and define $\eta_J = \sum_{t=\tau_J+1}^{\tau}\theta_t$ and $\eta'_J = \sum_{t=\tau_J+1}^{\tau}\theta_t \times \mathbb{1}_{\{X_{t-1} \ge J\}}$ as in Lemma 3.7. Finally, let $F_J = P(R(i) \ge \tau_J | X)$, which is shown in (3.7) to be equal to $1 - \prod_{t=\tau_J+1}^{\tau}(1 - \theta_t)$.

LEMMA 5.2. *If $J \le C'N/(\log N)$, then for all $d \in \{0, 1, \dots, n\}$,*

$$\left| P(K_{\tau_J} = d) - \binom{n}{d} E[F_J^d (1 - F_J)^{n-d}] \right| \le \min\left\{ \frac{C}{\log N}, \frac{C}{J} \right\} + \frac{C}{(\log N)^2}.$$

PROOF. Note that $K_\tau = 0$. Also, $K_{t-1} - K_t \in \{0, 1, \dots, n\}$ for all $1 \le t \le \tau$ and

$$E[K_{t-1} - K_t | X, (K_u)_{u=t}^{\tau}] = (n - K_t)\theta_t.$$

Define another process $(K'_t)_{t=0}^{\tau}$ such that $K'_\tau = 0$ and the conditional distribution of $K'_{t-1} - K'_t$ given X and $(K'_u)_{u=t}^{\tau}$ is Binomial$(n - K'_t, \theta_t)$. Note that $E[K'_{t-1} - K'_t | X, (K'_u)_{u=t}^{\tau}] = (n - K'_t)\theta_t$. We will show that the processes $(K_t)_{t=0}^{\tau}$ and $(K'_t)_{t=0}^{\tau}$ can be coupled so that

$$P(K_t \ne K'_t \text{ for some } t \ge \tau_J) \le \min\left\{ \frac{C}{\log N}, \frac{C}{J} \right\} + \frac{C}{(\log N)^2}. \tag{5.1}$$

Equation (5.1) implies the lemma because the conditional distribution of K'_{τ_J} given X is binomial with parameters n and $1 - \prod_{t=\tau_J+1}^{\tau}(1 - \theta_t) = F_J$.

By applying Lemma 5.1 with $V = K_{t-1} - K_t$ and $V' = K'_{t-1} - K'_t$, we can construct the process $(K'_t)_{t=0}^{\tau}$ on the same probability space as $(K_t)_{t=0}^{\tau}$ such that

$$\begin{aligned} &P(K_t \ne K'_t \text{ for some } t \ge \tau_J | X) \\ &\quad \le n \sum_{t=\tau_J+1}^{\tau} P\bigl(K_{t-1} - K_t \ge 2 | X, (K_u)_{u=t}^{\tau}\bigr) \\ &\qquad + n \sum_{t=\tau_J+1}^{\tau} P\bigl(K'_{t-1} - K'_t \ge 2 | X, (K'_u)_{u=t}^{\tau}\bigr). \end{aligned} \tag{5.2}$$

If $K_{t-1} - K_t \ge 2$ for some $t \ge \tau_J$, then $\tau_J \le R(i) \le G(i, j)$ for some i and j. We have $P(\tau_J \le R(i) \le G(i, j)) \le C/(\log N)$ for all J by Proposition 2.4 and $P(\tau_J \le R(i) \le G(i, j)) \le C/J$ for all $J \le C'N/(\log N)$ by Proposition 2.5. Therefore, for $J \le C'N/(\log N)$,

$$\begin{aligned} &E\left[\sum_{t=\tau_J+1}^{\tau} P\bigl(K_{t-1} - K_t \ge 2 | X, (K_u)_{u=t}^{\tau}\bigr) \right] \\ &\quad \le \sum_{t=1}^{\tau} P(K_{t-1} - K_t \ge 2 \text{ and } t \ge \tau_J) \\ &\quad \le \frac{n}{2} P(K_{t-1} - K_t \ge 2 \text{ for some } t \ge \tau_J) \\ &\quad \le \min\left\{ \frac{C}{\log N}, \frac{C}{J} \right\}. \end{aligned} \tag{5.3}$$

Now a binomial random variable will be at least 2 if and only if there is some pair of successful trials, so $P(K'_{t-1} - K'_t \geq 2 | X, (K'_u)_{u=t}^{\tau}) \leq \binom{n}{2}\theta_t^2$ and

$$\sum_{t=\tau_J+1}^{\tau} P\big(K'_{t-1} - K'_t \geq 2 | X, (K'_u)_{u=t}^{\tau}\big) \leq \binom{n}{2} \sum_{t=\tau_J+1}^{\tau} \theta_t^2. \tag{5.4}$$

By taking expectations in (5.2) and applying (5.3), (5.4) and (3.9), we get (5.1), which completes the proof. ☐

PROOF OF PROPOSITION 2.6. In view of Lemma 5.2, it suffices to show that

$$|E[F_J^d(1 - F_J)^{n-d}] - q_J^d(1 - q_J)^{n-d}| \leq \min\left\{\frac{C}{\log N}, \frac{C}{J}\right\} + \frac{C}{(\log N)^2} \tag{5.5}$$

for all $d \in \{0, 1, \ldots, n\}$. If $0 \leq a_1, \ldots, a_n \leq 1$ and $0 \leq b_1, \ldots, b_n \leq 1$, then $|a_1 \cdots a_n - b_1 \cdots b_n| \leq \sum_{i=1}^n |a_i - b_i|$, as shown in Lemma 4.3 of Chapter 2 of [5]. Therefore,

$$\begin{aligned}
&|E[F_J^d(1 - F_J)^{n-d}] - q_J^d(1 - q_J)^{n-d}| \\
&\quad \leq E[d|F_J - q_J| + (n - d)|(1 - F_J) - (1 - q_J)|] \\
&\quad = nE[|F_J - q_J|].
\end{aligned}$$

Note that

$$\begin{aligned}
|F_J - q_J| &\leq |F_J - (1 - e^{-\eta_J})| \\
&\qquad + \big|e^{-\eta'_J} - e^{-\eta_J}\big| + \big|e^{-\eta'_J} - e^{-E[\eta'_J]}\big| + \big|(1 - e^{-E[\eta'_J]}) - q_J\big|.
\end{aligned} \tag{5.6}$$

It follows from (3.8) and (3.9) that $E[|F_J - (1 - e^{-\eta_J})|] \leq C/(\log N)^2$. The expectations of the second, third and fourth terms on the right-hand side of (5.6) can be bounded as in the conclusion of the proof of Proposition 2.2 at the end of Section 3. All of those error estimates are smaller than the right-hand side of (5.5), so the desired result follows. ☐

6. A branching process approximation.

In this section we will show how the evolution of the individuals with the B allele during the first stage of the selective sweep can be approximated by a supercritical branching process. This will lead to a proof of Proposition 2.7. Recall that the first stage of the sweep consists of the times $0 \leq t \leq \tau_J$, where $J = \lfloor(\log N)^a\rfloor$ for some fixed constant $a > 4$. We will assume throughout this section that N is large enough that $J \leq N$. In Section 6.1, we explain the coupling between the branching process and the population model. In Section 6.2, we consider the lineages in the branching process with an infinite line of descent. Proposition 2.7 is proved using these ideas in Section 6.3.

6.1. *Coupling the population model with a branching process.* We begin by constructing a multitype branching process with the properties mentioned in Proposition 2.7. That is, the process will start with one individual at time zero, and each individual will give birth at rate 1 and die at rate $1-s$. Each new individual has the same type as its parent with probability $1-r$ and a new type, different from all other types, with probability r. We now explain how to construct this branching process so that until the number of individuals reaches J, the branching process will be coupled with the population process $(M_t)_{t=0}^{\infty}$ with high probability.

Define random variables $0=\xi_0<\xi_1<\cdots$ such that $(\xi_i-\xi_{i-1})_{i=1}^{\infty}$ is an i.i.d. sequence of random variables, each having an exponential distribution with mean $1/(2N)$. The branching process will start with one individual at time zero. Until the population size reaches J, there will be no births during the intervals (ξ_{t-1},ξ_t), but births and deaths can occur at the times $\xi_1,\xi_2,\ldots.$ This branching process will be coupled with $(M_t)_{t=0}^{\infty}$ so that, with high probability, the number of individuals with the B allele at time t will be the same as the number of individuals in the branching process at time ξ_t. To facilitate this coupling, we will also assign to each individual in the branching process a label such that all the individuals alive at a given time have distinct labels. We denote by L_t the set of all i such that there is an individual labeled i in the population at time ξ_t. When $L_t=\{i:B_t(i)=1\}$, meaning that the labels are the same as the individuals in the population model with the B allele at time t, we say the coupling holds at time t. The label of the individual at time zero will be U, where U is the random variable with a uniform distribution on $\{1,\ldots,2N\}$ defined at the beginning of Section 2. We have $B_0(U)=1$, so the coupling holds at time zero.

For the branching process to have the desired properties, each individual must have probability $1/(2N)$ of giving birth at time ξ_t and probability $(1-s)/(2N)$ of dying at time ξ_t. Also, at most one birth or death event can occur at a time. Suppose the coupling holds at time ξ_{t-1} and $i\in L_{t-1}$. Also, assume $X_{t-1}=k$. In the population model, the number of B's increases by 1 at time t, with i being the parent of the new individual, if $I_{t,2}=i$ and $B_{t-1}(I_{t,1})=0$, which has probability $(2N-k)/(2N)^2$. Also, the ith individual in the population dies at time t, causing the B population to decrease in size by 1, if $I_{t,1}=i$, $B_{t-1}(I_{t,2})=0$ and $I_{t,4}=1$, which has probability $(2N-k)(1-s)/(2N)^2$. Consequently, we can define the branching process such that the individual labeled i gives birth at time ξ_t if and only if $I_{t,2}=i$, which has probability $1/(2N)$. We give the new individual the label $I_{t,1}$, unless one of the other individuals already has this label. As a result, the coupling will hold at time t if $B_{t-1}(I_{t,1})=0$, but not if $B_{t-1}(I_{t,1})=1$. The individual labeled i will die with probability $(1-s)/(2N)$, and will die whenever $I_{t,1}=i$, $B_{t-1}(I_{t,2})=0$ and $I_{t,4}=1$. Then the probability that the coupling fails to hold at time t is

$$k\biggl(\frac{1}{2N}-\frac{2N-k}{(2N)^2}\biggr)+k\biggl(\frac{1-s}{2N}-\frac{(2N-k)(1-s)}{(2N)^2}\biggr)=\frac{k^2(2-s)}{(2N)^2}. \tag{6.1}$$

If a new individual in the branching process is born at time t, we say that it has a new type whenever $I_{t,5} = 1$, which has probability r. This means that births of individuals with new types correspond to recombinations in the population model.

Fix a positive integer m. On the event that the branching process has at least J individuals at some time, we define a random marked partition $\tilde{\Psi}_m$ as follows. Define κ such that ξ_κ is the first time at which there are J individuals. Define a random injective map $\tilde{\sigma}: \{1, \dots, m\} \to L_\kappa$ such that all $(J)_m$ possible maps are equally likely. Then say that $i \sim_{\tilde{\Psi}_m} j$ if and only if the individuals labeled $\tilde{\sigma}(i)$ and $\tilde{\sigma}(j)$ are of the same type. Mark the block of $\tilde{\Psi}_m$ consisting of all i such that the individual labeled $\tilde{\sigma}(i)$ has the same type as the individual at time zero. Furthermore, we can define $\tilde{\sigma}$ such that $\sigma = \tilde{\sigma}$ on the event that $\kappa = \tau_J$ and $L_{\tau_J} = \{i : B_{\tau_J}(i) = 1\}$, where $\sigma: \{1, \dots, m\} \to \{i : B_{\tau_J}(i) = 1\}$ is the map defined in Section 2 that is used in the construction of the random marked partition Ψ_m. Recall that $i \sim_{\Psi_m} j$ if and only if $A^0_{\tau_J}(\sigma(i)) = A^0_{\tau_J}(\sigma(j))$, and the block $\{i : B_0(A^0_{\tau_J}(\sigma(i))) = 1\}$ is marked.

Suppose $X_t = J$ for some t and the coupling holds for all $t \le \tau_J$, so $\kappa = \tau_J$. Then the genealogy of the branching process is the same as the genealogy of the B's in the population up to time τ_J. Furthermore, groups of individuals in the branching process with the same type correspond to groups of lineages in the population that escape the selective sweep at the same time and, therefore, get their allele at the neutral site from the same ancestor. Therefore, we will have $\tilde{\Psi}_m = \Psi_m$ unless one of the following events happens to a sampled lineage during the first stage of the selective sweep:

1. One of the B lineages experiences recombination, but the allele at the neutral site comes from another B individual.
2. Two recombinations cause a lineage to go from the B population to the b population and then back into the B population.
3. There is a coalescence event involving at least one lineage in the b population.

More formally, the lemma below is a consequence of our construction. Note that the events Λ_3^c, Λ_4^c and Λ_5^c correspond to the three possibilities mentioned above.

LEMMA 6.1. *Let* $R_J(i) = \sup\{t \ge 0 : B_t(A^t_{\tau_J}(i)) = 0\}$ *and* $G_J(i, j) = \sup\{t \ge 0 : A^t_{\tau_J}(i) = A^t_{\tau_J}(j)\}$. *We have* $\Psi_m = \tilde{\Psi}_m$ *on the event* $\Lambda_1 \cap \dots \cap \Lambda_5$, *where*:

Λ_1 *is the event that* $X_t = J$ *for some* t;

Λ_2 *is the event that the coupling holds for all* $t \le \tau_J$;

Λ_3 *is the event that for all* $t \le \tau_J$ *for which* $B_{t-1}(I_{t,2}) = 1$, *we have* $B_{t-1}(I_{t,3}) = 0$;

Λ_4 *is the event that for* $i \in \{1, \dots, m\}$, *we have* $B_t(A^t_{\tau_J}(\sigma(i))) = 0$ *for all* $t \le R_J(i)$;

Λ_5 *is the event that for all* $i, j \in \{1, \dots, m\}$ *with* $G_J(\sigma(i), \sigma(j)) \geq 0$, *we have*

$$\begin{aligned} &B_{G_J(\sigma(i),\sigma(j))+1}\bigl(A_{\tau_J}^{G_J(\sigma(i),\sigma(j))+1}(\sigma(i))\bigr) \\ &\qquad = B_{G_J(\sigma(i),\sigma(j))+1}\bigl(A_{\tau_J}^{G_J(\sigma(i),\sigma(j))+1}(\sigma(j))\bigr) = 1. \end{aligned}$$

PROOF. We have seen that when Λ_1 and Λ_2 occur, we have $L_{\tau_J} = \{i : B_{\tau_J}(i) = 1\}$ and $\sigma = \tilde{\sigma}$. For integers $u \leq t$ and $i \in L_t$, let $\tilde{A}_t^u(i)$ be the label of the individual in the branching process at time ξ_u that is the ancestor of the individual labeled i at time ξ_t, unless the ancestor is of a different type than the individual labeled i at time t, in which case we define $\tilde{A}_t^u(i) = 0$. Note that when Λ_1 and Λ_2 occur, we have $i \sim_{\tilde{\Psi}_m} j$ if and only if $\tilde{A}_{\tau_J}^t(\tilde{\sigma}(i)) = \tilde{A}_{\tau_J}^t(\tilde{\sigma}(j)) \neq 0$ for some t.

Since $\sigma = \tilde{\sigma}$ when Λ_1 and Λ_2 occur, we have $i \sim_{\tilde{\Psi}_m} j$ if and only if $\tilde{A}_{\tau_J}^t(\sigma(i)) = \tilde{A}_{\tau_J}^t(\sigma(j)) \neq 0$ for some t. Suppose $j \in L_t$. It follows from the constructions that $A_t^{t-1}(j) = \tilde{A}_t^{t-1}(j)$ unless $j = I_{t,1}$ and $I_{t,5} = 1$. In this case, $\tilde{A}_t^{t-1}(j) = 0$ and if Λ_3 occurs, then $B_{t-1}(A_t^{t-1}(j)) = 0$. It follows that if Λ_4 also occurs, then $\tilde{A}_{\tau_J}^t(\sigma(i)) = \tilde{A}_{\tau_J}^t(\sigma(j)) \neq 0$ if and only if we have both $A_{\tau_J}^t(\sigma(i)) = A_{\tau_J}^t(\sigma(j))$ and $B_t(A_{\tau_J}^t(\sigma(i))) = B_t(A_{\tau_J}^t(\sigma(j))) = 1$. Furthermore, when Λ_5 occurs, we have both $A_{\tau_J}^t(\sigma(i)) = A_{\tau_J}^t(\sigma(j))$ and $B_t(A_{\tau_J}^t(\sigma(i))) = B_t(A_{\tau_J}^t(\sigma(j))) = 1$ for some t if and only if $A_{\tau_J}^0(\sigma(i)) = A_{\tau_J}^0(\sigma(j))$, which is exactly the condition for $i \sim_{\Psi_m} j$. Thus, when $\Lambda_1, \dots, \Lambda_5$ all occur, we have $i \sim_{\Psi_m} j$ if and only if $i \sim_{\tilde{\Psi}_m} j$.

It remains only to show that the marked blocks of Ψ_m and $\tilde{\Psi}_m$ are the same. Note that i is in the marked block of $\tilde{\Psi}_m$ if and only if $\tilde{\sigma}(i) = \sigma(i)$ has the same type as the individual at time zero or, equivalently, if and only if $\tilde{A}_{\tau_J}^0(\sigma(i)) \neq 0$. The fact that this condition is equivalent to $B_0(A_{\tau_J}^0(\sigma(i))) = 1$ follows from the coupling and conditions Λ_3 and Λ_4. $\square$

We now use this coupling to show that the partition $\tilde{\Psi}_m$ conditioned on the survival of the branching process has almost the same distribution as Ψ_m.

LEMMA 6.2. *Let* π *be a partition of* $\{1, \dots, m\}$. *Then there exists a constant* C *such that*

$$|P'(\tilde{\Psi}_m = \pi \,|\, \#L_t > 0 \textit{ for all } t \in \mathbb{N}) - P(\Psi_m = \pi)| \leq C/(\log N)^2.$$

PROOF. We will show that if Λ_1 occurs, then $\Lambda_2 \cap \dots \cap \Lambda_5$ occurs with high probability. Conditional on the event that $X_{t-1} = k$ and that the coupling holds at time $t - 1$, it follows from (6.1) that the probability that the coupling fails to hold at time t is $k^2(2 - s)/(2N)^2$. Likewise, conditional on these same events, the probability that $B_{t-1}(I_{t,2}) = B_{t-1}(I_{t,3}) = 1$ is $(k/(2N))^2$. Thus, if D_t is the event that t is the first integer such that either the coupling fails at time t or $B_{t-1}(I_{t,2}) = B_{t-1}(I_{t,3}) = 1$, then $P'(D_t | X_t = k) \leq (3-s)k^2/(2N)^2$, where we use P' because we are not conditioning on the event that $X_t = 2N$ for some t. Therefore,

$$\begin{aligned}
P'\bigl(\Lambda_1 \cap (\Lambda_2^c \cup \Lambda_3^c)\bigr) &\leq \sum_{t=1}^{\infty} P'(D_t \cap \{t \leq \tau_J < \infty\}) \\
&= \sum_{t=1}^{\infty} E'[P'(D_t \cap \{t \leq \tau_J < \infty\} | X_{t-1})] \\
&\leq \sum_{t=1}^{\infty} E'\biggl[\frac{(3-s)X_{t-1}^2}{(2N)^2} \mathbb{1}_{\{X_{t-1} \leq J\}}\biggr] \\
&= \frac{3-s}{(2N)^2} \sum_{t=1}^{\infty} E'[X_{t-1}^2 \mathbb{1}_{\{X_{t-1} \leq J\}}] \\
&\leq \frac{3-s}{(2N)^2} \sum_{k=1}^{J} k^2 E'[T_k].
\end{aligned}$$

Since $P'(X_t \neq X_{t-1} | X_{t-1} = k) = P(X_t \neq X_{t-1} | X_{t-1} = k) = p_k = k(2N - k)(2-s)/(2N)^2$ and $E'[U_k + D_k] \leq C$, it follows that

$$\begin{aligned}
P'\bigl(\Lambda_1 \cap (\Lambda_2^c \cup \Lambda_3^c)\bigr) &\leq \frac{3-s}{(2N)^2} \sum_{k=1}^{J} k^2 \frac{E'[U_k + D_k]}{p_k} \\
&\leq \frac{C}{N^2} \sum_{k=1}^{J} \frac{k^2 (2N)^2}{k(2N-k)} \\
&\leq C \sum_{k=1}^{J} \frac{k}{2N-k} \leq \frac{CJ^2}{N}.
\end{aligned}$$

To handle Λ_4 and Λ_5, note that

$$P'(X_\tau = 2N | \Lambda_1) = p(0, 2N, J) = \frac{1-(1-s)^J}{1-(1-s)^{2N}} \geq 1-(1-s)^J. \tag{6.2}$$

It follows from (6.2) and the proof of Proposition 2.1 that $P'(\Lambda_1 \cap \Lambda_4^c) \leq C/(\log N)^2$. Likewise, it follows from (6.2) and the proof of Proposition 2.3 that $P'(\Lambda_1 \cap \Lambda_5^c) \leq C(\log N)/N$.

Since $P'(\Lambda_1) = s/(1-(1-s)^J)$ by Lemma 3.1, it follows from the above calculations that $|P'(\Lambda_1 \cap \cdots \cap \Lambda_5) - s| \leq C/(\log N)^2$. Recall that $P'(X_\tau = 2N) = s/(1-(1-s)^{2N})$ by Lemma 3.1. Since $\{\#L_t > 0 \text{ for all } t \in \mathbb{N}\}$ is the event that the branching process survives, it is well known that $P'(\#L_t > 0$ for all $t \in \mathbb{N}) = s$. Furthermore, if $\Lambda_1 \cap \cdots \cap \Lambda_5$ occurs, then $X_t = J$ for some t and $\#L_t = J$ for some t. Note that $P'(X_\tau = 2N | X_t = J \text{ for some } t) \geq 1 - (1-s)^J$ as in (6.2) and $P'(\#L_t > 0 \text{ for all } t | \#L_t = J \text{ for some } t) = 1 - (1-s)^J$. Thus, the events $\Lambda_1 \cap \cdots \cap \Lambda_5$, $\{X_\tau = 2N\}$ and $\{\#L_t = 0 \text{ for all } t\}$ agree closely enough that the probability, under P', that either all or none of these three events occurs is at least $1 - C/(\log N)^2$. It follows from this observation, Lemma 6.1 and the fact that P is the conditional probability measure of P' given $X_\tau = 2N$ that

$$\begin{aligned} P'(\tilde{\Psi}_m = \pi | \#L_t > 0 \text{ for all } t \in \mathbb{N}) &= P'(\tilde{\Psi}_m = \pi | \Lambda_1 \cap \cdots \cap \Lambda_5) + O\big((\log N)^{-2}\big) \\ &= P'(\Psi_m = \pi | \Lambda_1 \cap \cdots \cap \Lambda_5) + O\big((\log N)^{-2}\big) \\ &= P'(\Psi_m = \pi | X_\tau = 2N) + O\big((\log N)^{-2}\big) \\ &= P(\Psi_m = \pi) + O\big((\log N)^{-2}\big), \end{aligned}$$

which proves the lemma. □

6.2. *Infinite lines of descent.* Consider a continuous-time branching process in which each individual gives birth at rate 1 and dies at rate $1-s$. Equivalently, each individual lives for an exponentially distributed time with mean $1/(2-s)$ and then has some number of offspring, which is 0 with probability $(1-s)/(2-s)$ and 2 with probability $1/(2-s)$. Say that an individual at time t has an infinite line of descent if it has a descendant in the population at time u for all $u > t$. Otherwise, say that the individual has a finite line of descent.

Define the process $(Y_t^{(1)}, Y_t^{(2)})_{t \geq 0}$ such that $Y_t^{(1)}$ is the number of individuals at time t having an infinite line of descent and $Y_t^{(2)}$ is the number of individuals having a finite line of descent. Gadag and Rajarshi [11] showed that this process is a two-type Markov branching process. They also showed that the behavior of the process can be described as follows. Let p_k be the probability that an individual has k offspring and let $f(x) = \sum_{k=0}^\infty p_k x^k$ be the generating function of the offspring distribution. Let $u(x) = b[f(x) - x]$, where b^{-1} is the mean lifetime of an individual. Let $f^{(1)}(x, y) = \sum_{j=0}^\infty \sum_{k=0}^\infty p_{jk}^{(1)} x^j y^k$, where $p_{jk}^{(1)}$ is the probability that an individual with an infinite line of descent has j offspring with an infinite line of descent and k offspring with a finite line of descent. Let $f^{(2)}(x, y) = \sum_{j=0}^\infty \sum_{k=0}^\infty p_{jk}^{(2)} x^j y^k$, where $p_{jk}^{(2)}$ is the probability that an individual with a finite line of descent has j offspring with an infinite line of descent and k offspring with a finite line of descent. Let $u^{(1)}(x, y) = b[f^{(1)}(x, y) - x]$ and let $u^{(2)}(x, y) = b[f^{(2)}(x, y) - y]$. Let q be the smallest nonnegative solution of the equation $u(x) = 0$, which is also the probability that the branching process dies out. Then, by equation (4) of [11],

$$u^{(1)}(x, y) = \frac{u(x(1-q) + yq) - u(yq)}{1-q} \quad \text{and} \quad u^{(2)}(x, y) = \frac{u(yq)}{q}.$$

In the case of interest to us, we have $f(x) = \frac{1-s}{2-s} + \frac{1}{2-s}x^2$ and, therefore,

$$u(x) = (2-s)[f(x) - x] = (1-s) + x^2 - (2-s)x.$$

Since $u(x) = x$ if and only if $x \in \{1-s, 1\}$, we have $q = 1-s$. It follows that

$$\begin{aligned} u^{(1)}(x,y) &= \bigl\{[xs + y(1-s)]^2 - (2-s)[xs + y(1-s)] \\ &\qquad - [y(1-s)]^2 + (2-s)[y(1-s)]\bigr\}/s \\ &= sx^2 + 2(1-s)xy - (2-s)x. \end{aligned}$$

Thus, an individual with an infinite line of descent lives for an exponentially distributed time with mean $1/(2-s)$. It is replaced by two individuals with infinite lines of descent at rate s, and it is replaced by one individual with an infinite line of descent and another individual with a finite line of descent at rate $2(1-s)$.

Now, consider the process $(Y_t^{(1)}, Y_t^{(2)})$ started with one individual and conditioned to survive forever, which is equivalent to assuming that $Y_0^{(1)} = 1$ and $Y_0^{(2)} = 0$. Assume, as in Proposition 2.7, that the individuals are assigned types, and that each new individual born is the same type as its parent with probability $1-r$ and is a new type with probability r. Define $\lambda^* = \inf\{t : Y_t^{(1)} = \lfloor Js \rfloor\}$. Let $\lambda_k = \inf\{t : Y_t^{(1)} + Y_t^{(2)} = k\}$. Let $J_1 = \lfloor J(1 + s^{-1}\sqrt{(\log J)/J}\,)^{-1} \rfloor$ and $J_2 = \lceil J(1 - s^{-1}\sqrt{(\log J)/J}\,)^{-1} \rceil$.

LEMMA 6.3. *We have* $1 - P(\lambda_{J_1} \le \lambda^* \le \lambda_{J_2}) \le C/(\log N)^8$.

PROOF. If S has a Binomial(n, p) distribution and $p < c < 1$, then we have the large deviations result that $P(S \ge cn) \le e^{-2n(c-p)^2}$ (see [13]).

Let S_1 have a Binomial(J_1, s) distribution and let S_2 have a Binomial(J_2, s) distribution. Let $c = s + \sqrt{(\log J)/J}$. Then $J_1 = \lfloor Js/c \rfloor$, so $cJ_1 \le Js$ and, therefore,

$$\begin{aligned} P\bigl(\lambda^* \le \lambda_{J_1}\bigr) &= P(S_1 \ge \lfloor Js \rfloor | S_1 > 0) \\ &= \frac{P(S_1 \ge \lfloor Js \rfloor)}{P(S_1 > 0)} \le \frac{P(S_1 \ge \lfloor cJ_1 \rfloor)}{1 - (1-s)^{J_1}} \\ &\le \frac{P(S_1 \ge (c - 1/J_1)J_1)}{1 - (1-s)^{J_1}}. \end{aligned}$$

Recalling $J = \lfloor (\log N)^a \rfloor$ with $a > 4$, it follows that if $\varepsilon > 0$ is small, then for large N,

$$\begin{aligned} P\bigl(\lambda^* \le \lambda_{J_1}\bigr) &\le 2e^{-2J_1(\sqrt{(\log J)/J} - J_1^{-1})^2} \\ &\le Ce^{-2(J_1/J)\log J} \le CJ^{-(2-\varepsilon)} \le C/(\log N)^8. \end{aligned}$$

Likewise, if $d = (1-s) + \sqrt{(\log J)/J}$, then $J_2 = \lceil Js/(1-d) \rceil$, so $(1-d)J_2 \geq Js$ and thus

$$\begin{aligned} P(\lambda^* > \lambda_{J_2}) &= P(S_2 < \lfloor Js \rfloor | S_2 > 0) \leq P(S_2 < \lfloor Js \rfloor) \\ &= P(J_2 - S_2 > J_2 - \lfloor Js \rfloor) \leq P(J_2 - S_2 \geq dJ_2). \end{aligned}$$

Therefore, $P(\lambda^* > \lambda_{J_2}) \leq e^{-2(J_2/J)\log J} \leq J^{-2} \leq C/(\log N)^8$ and the lemma follows. □

6.3. *Proof of Proposition* 2.7. We now prove Proposition 2.7. Recall that Υ_m is the marked partition obtained by sampling m of the $\lfloor Js \rfloor$ individuals at time λ^* that have an infinite line of descent, and then declaring i and j to be in the same block of Υ_m if and only if the ith and jth individuals in the sample have the same type. The marked block of Υ_m consists of the individuals in the sample with the same type as the individual at time zero. We now define three other random marked partitions $\Upsilon_m^{(1)}$, $\Upsilon_m^{(2)}$ and $\Upsilon_m^{(3)}$ in the same way, except that the sample of m individuals is taken differently for each partition. Namely, to obtain $\Upsilon_m^{(1)}$, we sample m of the individuals at time λ_J. To get $\Upsilon_m^{(2)}$, we sample m of the individuals at time λ_{J_2}. To get $\Upsilon_m^{(3)}$, we sample m of the individuals at time λ_{J_2} that have an infinite line of descent, assuming that m such individuals exist (otherwise, sample from all individuals at time λ_{J_2}).

Since the branching process has been conditioned to survive forever, $\Upsilon_m^{(1)}$ has the same distribution as the conditional distribution of $\tilde{\Psi}_m$ given $\#L_t > 0$ for all $t \in \mathbb{N}$. Thus, by Lemma 6.2, it suffices to show that for all marked partitions $\pi \in \mathcal{P}_m$, we have

$$\left| P(\Upsilon_m^{(1)} = \pi) - P(\Upsilon_m = \pi) \right| \leq \frac{C}{(\log N)^2}.$$

Note also that $\Upsilon_m^{(2)}$ and $\Upsilon_m^{(3)}$ have the same distribution by the strong Markov property.

We can couple $\Upsilon_m^{(1)}$ and $\Upsilon_m^{(2)}$ such that the sample at time λ_J used to construct $\Upsilon_m^{(1)}$ includes all of the individuals in the sample at time λ_{J_2} that were born before time λ_J. If there are fewer than m such individuals, the rest of the sample at time λ_J can be picked from the remaining individuals. By the strong Markov property, this way of picking the sample at time λ_J does not change the distribution of $\Upsilon_m^{(1)}$. Therefore, $\Upsilon_m^{(1)} = \Upsilon_m^{(2)}$ if the m individuals sampled when constructing $\Upsilon_m^{(2)}$ were all born before time λ_J. Likewise, we can couple the partitions Υ_m and $\Upsilon_m^{(3)}$ such that on the event $\lambda^* \leq \lambda_{J_2}$, all of the individuals sampled at time λ_{J_2} that were born before time λ^* are part of the sample at time λ^* used to construct Υ_m. Note that λ^* is a stopping time with respect to the process $(Y_t^{(1)}, Y_t^{(2)})_{t \geq 0}$, so the strong Markov property implies that, conditional on $(Y_t^{(1)}, Y_t^{(2)})_{0 \le t \le \lambda^*}$, all $\binom{\lfloor Js \rfloor}{m}$ m-tuples of individuals with an infinite line of descent at time λ^* are equally likely to form the sample used to construct Υ_m. With this coupling, $\Upsilon_m^{(3)} = \Upsilon_m$ if $\lambda^* \le \lambda_{J_2}$ and all individuals sampled when constructing $\Upsilon_m^{(3)}$ were born before time λ^*.

Since $\Upsilon_m^{(2)} =_d \Upsilon_m^{(3)}$, Proposition 2.7 will be proved if the couplings described in the previous paragraph work well enough that $P(\Upsilon_m^{(3)} \ne \Upsilon_m)$ and $P(\Upsilon_m^{(1)} \ne \Upsilon_m^{(2)})$ can both be bounded by $C/(\log N)^2$. These bounds follow from Lemma 6.3 and Lemma 6.5 below.

LEMMA 6.4. *Let* $(\xi_t')_{t=0}^{\infty}$ *be a random walk on* $\mathbb{Z}$ *such that* $\xi_0' = 1$ *and, for all* k, $P(\xi_{t+1}' = k+1 | \xi_t' = k) = 1/(2-s)$ *and* $P(\xi_{t+1}' = k-1 | \xi_t' = k) = (1-s)/(2-s)$. *Let* $\xi = (\xi_t)_{t=0}^{\infty}$ *be the Markov process whose law is the same as the conditional law of* $(\xi_t')_{t=0}^{\infty}$ *given* $\xi_t' \ge 1$ *for all* t. *Let* $\kappa_n = \inf\{t : \xi_t = n\}$. *For all positive integers* n, *we have* $E[\kappa_{n+1} - \kappa_n] \le (2-s)/s$.

PROOF. Note that $\kappa_1 = 0$ and $\kappa_2 = 1$. Therefore, $E[\kappa_2 - \kappa_1] = 1$. Suppose $E[\kappa_n - \kappa_{n-1}] \le (2-s)/s$. Let $D_n = \#\{t : \kappa_n \le t < \kappa_{n+1}, \xi_t = n \text{ and } \xi_{t+1} = n-1\}$ be the number of times that ξ goes from n to $n-1$ before hitting $n+1$. Since $l_n = P(\xi_t = n+1 | \xi_{t-1} = n) \ge 1/(2-s)$, we have that $D_n + 1$ follows a geometric distribution with parameter $l_n \ge 1/(2-s)$. Therefore, $E[D_n] = (1/l_n) - 1 \le 1-s$. Note that each time that ξ goes from n to $n-1$, it must eventually return to n, which takes expected time $E[\kappa_n - \kappa_{n-1}]$. Thus, $E[\kappa_{n+1} - \kappa_n] = 1 + E[D_n](1 + E[\kappa_n - \kappa_{n-1}]) \le 1 + (1-s)[1 + (2-s)/s] = (2-s)/s$. The lemma now follows by induction. □

LEMMA 6.5. *The probability that an individual chosen at random at time* λ_{J_2} *was born after* λ_{J_1} *is at most* $C/(\log N)^2$.

PROOF. Define $(\tilde{Y}_t)_{t=0}^{\infty}$ such that if $0 = \tau_0 < \tau_1 < \cdots$ are the jump times of $(Y_t^{(1)} + Y_t^{(2)})_{t \ge 0}$, then $\tilde{Y}_t = Y_{\tau_t}^{(1)} + Y_{\tau_t}^{(2)}$. Let $\tilde{\lambda}_k = \inf\{t : \tilde{Y}_t = k\}$. The number of births between λ_{J_1} and λ_{J_2} is at most $\tilde{\lambda}_{J_2} - \tilde{\lambda}_{J_1}$. We have $E[\tilde{\lambda}_{J_2} - \tilde{\lambda}_{J_1}] \le [(2-s)/s](J_2 - J_1)$ by Lemma 6.4. Note that

$$\frac{J_2 - J_1}{J_2} \le \frac{J(1 - s^{-1}\sqrt{(\log J)/J})^{-1} - J(1 + s^{-1}\sqrt{(\log J)/J})^{-1} + 2}{J(1 + s^{-1}\sqrt{(\log J)/J})^{-1}}$$
$$\le C\sqrt{\frac{\log J}{J}},$$

so the probability that a randomly chosen individual at time λ_{J_2} was born after λ_{J_1} is at most

$$\left(\frac{2-s}{s}\right)\left(\frac{J_2 - J_1}{J_2}\right) \le C\sqrt{\frac{\log J}{J}} \le \frac{C}{(\log N)^2},$$

where the last inequality holds because $J = \lfloor (\log N)^a \rfloor$ for some $a > 4$. $\square$

7. Approximating the distribution of Θ.

In this section we complete the proof of Theorem 1.2 by proving Propositions 2.10, 2.11 and 2.8. We will use the notation W_k, ζ_k, Y_k and Z_i introduced before the statement of Theorem 1.2 in the Introduction. Recall also that $L = \lfloor 2Ns \rfloor$.

In Section 7.1, we prove Propositions 2.10 and 2.11, which pertain to the random variables Z_i introduced in the paint-box construction given in the Introduction. The rest of the section is devoted to the proof of Proposition 2.8. In Section 7.2 we introduce random variables Z_i' using the branching process. In Section 7.3 we state some lemmas comparing the Z_i and Z_i', and explain how these lemmas imply Proposition 2.8. In Section 7.4 we present some results related to Pólya urns that are needed to prove these lemmas and, finally, the lemmas are proved in Section 7.5.

7.1. *Proofs of Propositions* 2.10 *and* 2.11.

PROOF OF PROPOSITION 2.10. Since $P(Z_1 = Z_2 = k | V_k) \le V_k^2$, we have $P(Z_1 = Z_2 = k) \le E[V_k^2] = E[\zeta_k^2 W_k^2] = E[\zeta_k^2] E[W_k^2]$. Since $E[\zeta_k^2] = E[\zeta_k] = r/s$ and $E[W_k^2] = 2/k(k+1)$, it follows that

$$P(Z_1 = Z_2 > \lfloor Js \rfloor) \le \sum_{k=\lfloor Js \rfloor + 1}^{L} \frac{2r}{sk(k+1)} \le \frac{2r}{s\lfloor Js \rfloor} \le \frac{C}{(\log N)^{1+a}}. \qquad \square$$

We next prove Proposition 2.11, which says that the distribution of the number of i such that $Z_i > \lfloor Js \rfloor$ is approximately binomial. We begin with a lemma that gives an approximation to $P(Z_i > \lfloor Js \rfloor)$.

LEMMA 7.1. *We have* $P(Z_i > \lfloor Js \rfloor) = q_J + O(1/(\log N)^5)$.

PROOF. By the construction in the Introduction, $P(Z_i = k | Z_i \le k) = E[V_k] = E[\zeta_k] E[W_k] = r/(sk)$. Therefore, $P(Z_i \le \lfloor Js \rfloor) = \prod_{k=\lfloor Js \rfloor + 1}^{L} (1 - r/(sk))$. This is the same as the probability that none of the events $A_{\lfloor Js \rfloor + 1}, \dots, A_L$ occurs if the events are independent and $P(A_k) = r/(sk)$. Since

$$\sum_{k=\lfloor Js \rfloor + 1}^{L} \left(\frac{r}{sk} \right)^2 \le \frac{r^2}{s^2 \lfloor Js \rfloor} \le \frac{C}{(\log N)^6},$$

it follows from the Poisson approximation result on page 140 of [5] that

$$P(Z_i > \lfloor Js \rfloor) = 1 - \exp\left(- \sum_{k=\lfloor Js \rfloor + 1}^{L} \frac{r}{sk} \right) + O\left(\frac{1}{(\log N)^6} \right).$$

If $1 \le y_1 < y_2$, then $0 \le \sum_{k=\lfloor y_1 \rfloor}^{\lfloor y_2 \rfloor} \frac{1}{k} - \log\left(\frac{y_2}{y_1}\right) \le 2/\lfloor y_1 \rfloor$. Therefore,

$$\left|\sum_{k=J+1}^{2N} \frac{1}{k} - \sum_{k=\lfloor Js \rfloor+1}^{\lfloor 2Ns \rfloor} \frac{1}{k}\right| \le \frac{1}{J} + \left|\sum_{k=J}^{2N} \frac{1}{k} - \log\left(\frac{2N}{J}\right)\right| + \left|\log\left(\frac{2Ns}{Js}\right) - \sum_{k=\lfloor Js \rfloor}^{\lfloor 2Ns \rfloor} \frac{1}{k}\right|$$
$$\le \frac{3}{J} + \frac{2}{\lfloor Js \rfloor} \le \frac{C}{(\log N)^a}.$$

It follows that

$$P(Z_i > \lfloor Js \rfloor) = 1 - \exp\left(-\sum_{k=J+1}^{2N} \frac{r}{sk}\right) + O\left(\frac{1}{(\log N)^5}\right)$$
$$= q_J + O\left(\frac{1}{(\log N)^5}\right). \qquad \square$$

PROOF OF PROPOSITION 2.11. Let $\eta_k = \#\{i : Z_i = k\}$. Then $D = \eta_{\lfloor Js \rfloor+1} + \cdots + \eta_L$. Define the sequence $(\tilde{\eta}_k)_{k=\lfloor Js \rfloor+1}^{L}$ such that $\tilde{\eta}_L$ has a Binomial$(n, r/(sL))$ distribution and, conditional on $\tilde{\eta}_{k+1}, \dots, \tilde{\eta}_L$, the distribution of $\tilde{\eta}_k$ is binomial with parameters $n - \tilde{\eta}_{k+1} - \cdots - \tilde{\eta}_L$ and $r/(sk)$. Thinking of flipping n coins and continuing to flip those that do not show tails, it is easy to see that $\tilde{D} = \tilde{\eta}_{\lfloor Js \rfloor+1} + \cdots + \tilde{\eta}_L$ has a binomial distribution with parameters n and γ, where $\gamma = P(Z_i > \lfloor Js \rfloor)$. To compare D and $\tilde{D}$ we note that

$$P(\eta_k \ge 2 | \eta_{k+1}, \dots, \eta_L) \le \binom{n}{2} E[V_k^2] = \binom{n}{2} E[\zeta_k] E[W_k^2]$$
$$= \binom{n}{2} \frac{2r}{sk(k+1)}$$

and $P(\tilde{\eta}_k \ge 2 | \tilde{\eta}_{k+1}, \dots, \tilde{\eta}_L) \le \binom{n}{2}(r/(sk))^2$. By Lemma 5.1, we can couple the η_k and $\tilde{\eta}_k$ such that $P(\eta_k \ne \tilde{\eta}_k | \eta_l = \tilde{\eta}_l$ for $l = k+1, \dots, L) \le Cr/k^2$ for all k. Therefore,

$$P(\eta_k \ne \tilde{\eta}_k \text{ for some } k > \lfloor Js \rfloor) \le \sum_{k=\lfloor Js \rfloor+1}^{L} \frac{Cr}{k^2} \le \frac{Cr}{\lfloor Js \rfloor} \le \frac{C}{(\log N)^5}.$$

This result, combined with Lemma 7.1, gives the proposition. $\square$

7.2. *Random variables Z_i' from the branching process.* It remains only to prove Proposition 2.8, which requires considerably more work. For convenience, let $H = \lfloor Js \rfloor$. From this point forward, $Z_1, \dots, Z_n$ will be random variables defined as in the Introduction but with $L = H$, so that the associated marked partition Π has the distribution $Q_{r,s,H}$. Our goal is to describe the distribution of the marked partition Υ_n from Propositions 2.7 and 2.8 using random variables

$Z'_1, \dots, Z'_n$, where Z'_i will be the number of individuals with an infinite line of descent at the time when the type of the ith individual first appeared. We will then prove Proposition 2.8 by comparing the distribution of $(Z'_1, \dots, Z'_n)$ to the distribution of $(Z_1, \dots, Z_n)$.

Define times $0 = \gamma_1 < \gamma_2 < \cdots < \gamma_H$ such that $\gamma_j = \inf\{t : Y_t^{(1)} = j\}$ is the first time that the branching process has j individuals with an infinite line of descent. Note that $(\gamma_{j+1} - \gamma_j)_{i=1}^{H-1}$ is a sequence of independent random variables and the distribution of $\gamma_{j+1} - \gamma_j$ is exponential with rate js. Whenever a new individual with an infinite line of descent is born, it has a new type with probability r. Also, each individual with an infinite line of descent is giving birth to a new individual with a finite line of descent at rate $2(1-s)$. Since a new individual has a new type with probability r, between times γ_j and γ_{j+1}, births of individuals with new types occur at rate $2jr(1-s)$. Whenever such a birth occurs, the type of the individual with an infinite line of descent changes with probability $1/2$. Thus, between times γ_j and γ_{j+1}, we can view the branching process as consisting of j lineages with infinite lines of descent, and their types are changing at rate $r(1-s)$. It follows that if, for some $j \geq 1$, we choose at random one of the j individuals at time $\gamma_{j+1}-$ with an infinite line of descent, the probability that its ancestor at time γ_j is not of the same type is

$$\frac{r(1-s)}{r(1-s)+js}. \tag{7.1}$$

Furthermore, for $j \geq 2$, the probability that its ancestor at time γ_j is not of the same type as its ancestor at time γ_j- is r/j because, with probability r, exactly one of the individuals at time γ_j is of a type that did not exist at time γ_j-. It follows that for $j \geq 2$, the probability that the individual sampled at time $\gamma_{j+1}-$ has a different type from its ancestor at time γ_j- is

$$\frac{r(1-s)}{r(1-s)+js} + \frac{js}{r(1-s)+js}\left(\frac{r}{j}\right) = \frac{r}{r(1-s)+js} \leq \frac{r}{js}. \tag{7.2}$$

Likewise, the probability that at least one of the j individuals with an infinite line of descent at time $\gamma_{j+1}-$ has a different ancestor at time γ_j- is

$$\frac{r(1-s)}{r(1-s)+s} + \frac{s}{r(1-s)+s}(r) = \frac{r}{r(1-s)+s}.$$

Let $\sigma'(1), \dots, \sigma'(n)$ represent n individuals sampled at random from those with an infinite line of descent at time γ_H. Then we can take the partition Υ_n to be defined such that $i \sim_{\Upsilon_n} j$ if and only if $\sigma'(i)$ and $\sigma'(j)$ have the same type, and the marked block is $\{i : \sigma'(i)$ has the same type as the individual at time $0\}$. Now define $Z'_1, \dots, Z'_n$ as follows. Let $Z'_i = 1$ if the ancestor at time 0 of $\sigma'(i)$ has the same type as $\sigma'(i)$. Otherwise, define

$$Z'_i = \max\{k : \sigma'(i) \text{ has a different type from its ancestor at time } \gamma_k-\}.$$

If $Z_i' \neq Z_j'$, then since each new type is different from all types previously in the population, $\sigma'(i)$ and $\sigma'(j)$ have different types. If $Z_i' = Z_j'$, then $\sigma'(i)$ and $\sigma'(j)$ have the same type unless $\sigma'(i)$ and $\sigma'(j)$ have different ancestors at time $\gamma_{Z_i'+1}-$ because they both have the same type as their ancestor at time $\gamma_{Z_i'+1}-$. We will show in Lemma 7.2 below that the probability that $Z_i' = Z_j'$, and $\sigma'(i)$ and $\sigma'(j)$ have different ancestors at time $\gamma_{Z_i'+1}-$ is $O((\log N)^{-2})$. Therefore, the probability that, for some i and j, we have $Z_i' = Z_j'$ but $\sigma'(i)$ and $\sigma'(j)$ have different types is $O((\log N)^{-2})$. Furthermore, it follows from (7.1) that the individuals $\{\sigma'(i) : Z_i' = 1\}$ have the same type as the individual at time 0 with probability $s/(r(1-s)+s)$. Define the marked partition Υ_n' of $\{1, \dots, n\}$ such that $i \sim_{\Upsilon_n'} j$ if and only if $Z_i' = Z_j'$, and independently with probability $s/(r(1-s)+s)$, mark the block $\{i : Z_i' = 1\}$. The preceding discussion implies that

$$|P(\Upsilon_n = \pi) - P(\Upsilon_n' = \pi)| \leq \frac{C}{(\log N)^2} \tag{7.3}$$

for all $\pi \in \mathcal{P}_n$. Thus, for proving Proposition 2.8, we may consider Υ_n' instead of Υ_n. This will be convenient because Υ_n' is defined from $Z_1', \dots, Z_n'$ in the same way that Π is defined from $Z_1, \dots, Z_n$. Consequently, once we establish Lemma 7.2 below, the remainder of the proof of Proposition 2.8 will just involve comparing Z_i and Z_i'.

LEMMA 7.2. *If $i \neq j$, then*

$$\begin{aligned} &P\bigl(Z_i' = Z_j' \text{ and } \sigma'(i) \text{ and } \sigma'(j) \text{ have different ancestors at time } \gamma_{Z_i'+1}-\bigr) \\ &\qquad \leq \frac{C}{(\log N)^2}. \end{aligned} \tag{7.4}$$

PROOF. First note that if $Z_i' = Z_j' = k$, then $\sigma'(i)$ and $\sigma'(j)$ have the same type as their ancestor at time $\gamma_{k+1}-$. If they have different ancestors at time $\gamma_{k+1}-$, there must be a $\gamma \in (\gamma_k, \gamma_{k+1})$ such that either $\sigma'(i)$ or $\sigma'(j)$ has an ancestor of a different type at time $\gamma-$ but not at time γ. The other of $\sigma'(i)$ and $\sigma'(j)$ must have an ancestor of a different type at time γ_k- than at time $\gamma-$. Given that $\sigma'(i)$ and $\sigma'(j)$ have different ancestors at time $\gamma_{k+1}-$, the probability that both of these things happen if $k \geq 2$ is

$$\left(\frac{2r(1-s)}{2r(1-s)+ks}\right)\left(\frac{r}{r(1-s)+ks}\right) \leq \frac{2r^2}{k^2s^2}.$$

The first factor is the probability that $\sigma'(i)$ or $\sigma'(j)$ has an ancestor of a different type at some time $\gamma-$, while the second factor is the probability from (7.2) that the other of $\sigma'(i)$ and $\sigma'(j)$ has an ancestor of a different type at time γ_k- than at time $\gamma-$. If $k=1$, then this conditional probability becomes

$$\left(\frac{2r(1-s)}{2r(1-s)+ks}\right)\left(\frac{r(1-s)}{r(1-s)+ks}\right)\le\frac{2r^2}{k^2s^2}$$

by (7.1). Therefore, if $i\neq j$, the probability that $Z_i'=Z_j'$, and $\sigma'(i)$ and $\sigma'(j)$ have different ancestors at time $\gamma_{Z_i'+1}-$ is at most $\sum_{k=1}^{H}(2r^2/(k^2s^2))\le C/(\log N)^2$, as claimed. □

7.3. *Comparison of the* Z_i *and* Z_i', *and proof of Proposition* 2.8. We first prove two fairly straightforward lemmas, one for the Z_i and one for the Z_i'. Lemma 7.3 allows us to disregard the possibility that the Z_i' may take more than two distinct values greater than 1, as well as the possibility that there may be two distinct values greater than 1, with multiple occurrences of the higher value. Lemma 7.4 rules out the same possibilities for the Z_i.

LEMMA 7.3. *We have*

$$P(Z_1'=j,Z_2'=k,Z_3'=l\text{ for some }2\le j<k<l)\le\frac{C(\log(\log N))^3}{(\log N)^3},\tag{7.5}$$

$$P(Z_1'=j,Z_2'=Z_3'=k\text{ for some }2\le j<k)\le\frac{C}{(\log N)^2}.\tag{7.6}$$

PROOF. From (7.2), we get $P(Z_3'=l)\le r/sl$, $P(Z_2'=k|Z_3'=l)\le r/(sk)$ and $P(Z_1'=j|Z_2'=k,Z_3'=l)\le r/(sj)$. Thus, the probability on the left-hand side of (7.5) is at most

$$\sum_{j=1}^{H}\sum_{k=j}^{H}\sum_{l=k}^{H}\left(\frac{r}{ls}\right)\left(\frac{r}{ks}\right)\left(\frac{r}{js}\right)\le\frac{C(\log(\log N))^3}{(\log N)^3}.$$

Conditional on the event that $\sigma'(2)$ and $\sigma'(3)$ have different ancestors at time $\gamma_{m+1}-$, the probability that they have the same ancestor at time γ_m- is $\binom{m}{2}^{-1}=2/m(m-1)$. Therefore, the probability that $\sigma'(2)$ and $\sigma'(3)$ have the same ancestor at time $\gamma_{k+1}-$ is at most $\sum_{m=k+1}^{H}2/m(m-1)\le 2/k$. The probability that $Z_2'=Z_3'=k$ given that $\sigma'(2)$ and $\sigma'(3)$ have the same ancestor at time $\gamma_{k+1}-$ is at most $r/(ks)$. Also, for $j<k$, we have $P(Z_1'=j|Z_2'=Z_3'=k)\le r/(js)$. Combining these results with Lemma 7.2, we can bound the probability on the left-hand side of (7.6) by

$$\begin{aligned}\frac{C}{(\log N)^2}+\sum_{j=1}^{H}\sum_{k=j+1}^{H}\left(\frac{r}{js}\right)\left(\frac{r}{ks}\right)\left(\frac{2}{k}\right)&\le\frac{C}{(\log N)^2}+\frac{2r^2}{s^2}\sum_{j=1}^{H}\sum_{k=j+1}^{H}\frac{1}{jk^2}\\&\le\frac{C}{(\log N)^2}.\end{aligned}$$

□

LEMMA 7.4. *We have*

$$P(Z_1 = j, Z_2 = k, Z_3 = l \text{ for some } 2 \le j < k < l) \le \frac{C(\log(\log N))^3}{(\log N)^3},$$

$$P(Z_1 = j, Z_2 = Z_3 = k \text{ for some } 2 \le j < k) \le \frac{C}{(\log N)^2}. \tag{7.7}$$

PROOF. Fix j, k, l such that $2 \le j < k < l \le H$. We have $P(Z_3 = l | Z_3 \le l) = \frac{r}{sl}$, $P(Z_2 = k | Z_3 = l, Z_2 \le k) = \frac{r}{sk}$, $P(Z_1 = j | Z_2 = k, Z_3 = l, Z_1 \le j) = \frac{r}{sj}$ and hence

$$P(Z_1 = j, Z_2 = k, Z_3 = l) \le \left(\frac{r}{sj}\right)\left(\frac{r}{sk}\right)\left(\frac{r}{sl}\right).$$

Summing as in the proof of Lemma 7.3 gives the first result. To prove (7.7), first note that

$$P(Z_2 = Z_3 = k) \le E[V_k^2] = E[\zeta_k^2]E[W_k^2] = \frac{2r}{sk(k+1)}$$

and $P(Z_1 = j | Z_2 = Z_3 = k) \le r/(sj)$. Then compute as in the proof of Lemma 7.3. ☐

Throughout the rest of this section we use the notation

$$q_{k,a,n} = \frac{(k-1)a!(n-a+k-2)!}{(n+k-1)!}.$$

We now state four more lemmas related to Z_i and Z_i'. Their proofs will be given after we explain how they imply Proposition 2.8.

LEMMA 7.5. *Suppose* $1 \le a \le n-1$. *Then*

$$P(Z_1' = l, Z_2' = \cdots = Z_{a+1}' = k, Z_{a+2}' = \cdots = Z_n' = 1 \text{ for some } 2 \le k < l)$$
$$= \frac{r^2}{s^2} \sum_{k=2}^{H} \sum_{l=k+1}^{H} \frac{q_{k,a,n-1}}{l} + O\left(\frac{1}{(\log N)^2}\right).$$

LEMMA 7.6. *Suppose* $1 \le a \le n-1$. *Then*

$$P(Z_1 = l, Z_2 = \cdots = Z_{a+1} = k, Z_{a+2} = \cdots = Z_n = 1 \text{ for some } 2 \le k < l)$$
$$= \frac{r^2}{s^2} \sum_{k=2}^{H} \sum_{l=k+1}^{H} \frac{q_{k,a,n-1}}{l} + O\left(\frac{1}{(\log N)^2}\right).$$

LEMMA 7.7. *If* $2 \le a \le n$, *then*

$$\begin{aligned}
&P(Z_1' = \cdots = Z_a' = k \text{ and } Z_{a+1}' = \cdots = Z_n' = 1 \text{ for some } k \ge 2) \\
&\quad = \frac{r}{s}\sum_{k=2}^{H} q_{k,a,n} - \frac{nr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{q_{k,a,n}}{l} + O\biggl(\frac{1}{(\log N)^2}\biggr),
\end{aligned} \tag{7.8}$$

$$\begin{aligned}
&P(Z_1' = k \text{ and } Z_2' = \cdots = Z_n' = 1 \text{ for some } k \ge 2) \\
&\quad = \frac{r}{s}\sum_{k=2}^{H} q_{k,1,n} - \frac{nr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{q_{k,1,n}}{l} \\
&\qquad - \frac{(n-1)r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{1}{k(n+l-2)} + O\biggl(\frac{1}{(\log N)^2}\biggr).
\end{aligned} \tag{7.9}$$

LEMMA 7.8. *If* $2 \le a \le n$, *then*

$$\begin{aligned}
&P(Z_1 = \cdots = Z_a = k \text{ and } Z_{a+1} = \cdots = Z_n = 1 \text{ for some } k \ge 2) \\
&\quad = \frac{r}{s}\sum_{k=2}^{H} q_{k,a,n} - \frac{nr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{q_{k,a,n}}{l} + O\biggl(\frac{1}{(\log N)^2}\biggr),
\end{aligned} \tag{7.10}$$

$$\begin{aligned}
&P(Z_1 = k \text{ and } Z_2 = \cdots = Z_n = 1 \text{ for some } k \ge 2) \\
&\quad = \frac{r}{s}\sum_{k=2}^{H} q_{k,1,n} - \frac{nr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{q_{k,1,n}}{l} \\
&\qquad - \frac{(n-1)r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{1}{k(n+l-2)} + O\biggl(\frac{1}{(\log N)^2}\biggr).
\end{aligned} \tag{7.11}$$

PROOF OF PROPOSITION 2.8. Let $\pi \in \mathcal{P}_n$. If π has four or more blocks, or three blocks of size at least 2, then $P(\Upsilon_n' = \pi) \le C/(\log N)^2$ by Lemma 7.3 and $Q_{r,s,H}(\pi) \le C/(\log N)^2$ by Lemma 7.4. If π has three blocks, at least one containing just one integer, then the fact that $|P(\Upsilon_n' = \pi) - Q_{r,s,H}(\pi)| \le C/(\log N)^2$ follows from Lemmas 7.3, 7.4, 7.5 and 7.6, as well as the fact that the probabilities that the blocks $\{i : Z_i = 1\}$ and $\{i : Z_i' = 1\}$ are marked in the two partitions are both $s/(r(1-s)+s)$. If π has just two blocks, then $|P(\Upsilon_n' = \pi) - Q_{r,s,H}(\pi)| \le C/(\log N)^2$ follows from Lemmas 7.7 and 7.8, Lemmas 7.5 and 7.6 with $a = n-1$, and equations (7.6) and (7.7). Finally, when π has just one block, $|P(\Upsilon_n' = \pi) - Q_{r,s,H}(\pi)| \le C/(\log N)^2$ follows from Lemmas 7.7 and 7.8 with $a = n$, and the fact that $P(Z_1 = \cdots = Z_n = 1)$ and $P(Z_1' = \cdots = Z_n' = 1)$ can be obtained by subtracting from 1 the remaining possibilities. Proposition 2.8 now follows from these results and (7.3). □

7.4. *Pólya urn facts.* It remains to prove Lemmas 7.5, 7.6, 7.7 and 7.8. In this section, we establish three lemmas that are related to Pólya urns. The first two lemmas are standard and straightforward, and their proofs are omitted.

LEMMA 7.9. *Suppose X has a beta distribution with parameters* 1 *and* $k-1$, *where k is an integer. Let $U_1, \dots, U_n$ be i.i.d. random variables with a uniform distribution on* $[0,1]$. *Then*

$$P(U_i \le X \text{ for } i = 1, \dots, a \text{ and } U_i > X \text{ for } i = a+1, \dots, n) = q_{k,a,n}.$$

LEMMA 7.10. *Consider an urn with one red ball and $k-1$ black balls. Suppose that n new balls are added to the urn one at a time. Each new ball is either red or black, and the probability that a given ball is red is equal to the fraction of red balls currently in the urn. Let S be any a-element subset of $\{1, \dots, n\}$. The probability that the ith ball added is red for $i \in S$ and black for $i \notin S$ is $q_{k,a,n}$. Note that this implies the sequence of draws is exchangeable.*

LEMMA 7.11. *In the setting of Lemma* 7.10, *suppose instead $l-k$ new balls are added to the urn. Then suppose we sample n of the l balls at random. Let $p_{k,l,a,n}$ be the probability that the first a balls sampled are red and the next $n-a$ are black. If $a \ge 1$, then there exists a constant C, which may depend on a and n, such that $|p_{k,l,a,n} - q_{k,a,n}| \le C/(kl)$ for all k and l.*

PROOF. It follows from Lemma 7.10 that, conditional on the event that none of the original k balls is in the sample of n, the probability that the first a balls sampled are red and the next $n-a$ are black is exactly $q_{k,a,n}$. The probability that the sample of n balls contains exactly j of the original k balls, an event we call $D_{j,k}$, is

$$\begin{aligned} \frac{\binom{k}{j}\binom{l-k}{n-j}}{\binom{l}{n}} &\le \left(\frac{k^j}{j!}\right)\left(\frac{(l-k)^{n-j}}{(n-j)!}\right)\left(\frac{n!(l-n)!}{l!}\right) \\ &\le \binom{n}{j}\frac{k^j l^{n-j}(l-n)!}{l!} \le C\left(\frac{k}{l}\right)^j, \end{aligned} \tag{7.12}$$

since n is a constant and thus so are $a \le n$ and $j \le n$.

Conditional on the event $D_{j,k}$, we can calculate the probability that we sample a red balls and $n-a$ black balls. The probability that the original red ball is in the sample is j/k. If it is, then by Lemma 7.10 the probability that $a-1$ of the other balls in the sample are red is $\binom{n-j}{a-1}q_{k,a-1,n-j}$. Likewise, conditional on the event that the original red ball is not in the sample, the probability that a of the other balls in the sample are red is $\binom{n-j}{a}q_{k,a,n-j}$. Thus, conditional on $D_{j,k}$, the probability that we sample a red balls and $n-a$ black balls is

$$\frac{j}{k}\frac{(n-j)!(k-1)(n-j-a+k-1)!}{(n-j-a+1)!(n-j+k-1)!} + \frac{k-j}{k}\frac{(n-j)!(k-1)(n-j-a+k-2)!}{(n-j-a)!(n-j+k-1)!}.$$

Our next step is to bring $\binom{n}{a}q_{k,a,n}$ out in front. Using that $(m-j)! = m!/(m)_j$ for integers $1 \le j \le m$, we get, for $k \ge 3$,

$$\begin{aligned}\binom{n}{a}\cdot a!\cdot\frac{(k-1)(n-a+k-2)!}{(n+k-1)!}\Bigg[&\frac{(n-a)_{j-1}}{(n)_j}\frac{j}{k}\frac{(n+k-1)_j}{(n-a+k-2)_{j-1}}\\ &+\frac{(n-a)_j}{(n)_j}\frac{k-j}{k}\frac{(n+k-1)_j}{(n-a+k-2)_j}\Bigg].\end{aligned}\tag{7.13}$$

Consider the expression in brackets. Each term can be written as a ratio of two polynomials in k of the same degree. Since $a \le n$ and $j \le n$, if $k \to \infty$ with n fixed, the expression in brackets is bounded by a constant. Now suppose $a = 1$. The bracketed expression becomes

$$\begin{aligned}&\frac{j(n+k-1)(n+k-2)}{nk(n+k-j-1)} + \frac{(n-j)(k-j)(n+k-1)(n+k-2)}{nk(n+k-j-1)(n+k-j-2)}\\ &\quad= \frac{j(n+k-1)(n+k-2)(n+k-j-2)}{nk(n+k-j-1)(n+k-j-2)}\\ &\qquad+\frac{(n-j)(k-j)(n+k-1)(n+k-2)}{nk(n+k-j-1)(n+k-j-2)}.\end{aligned}$$

Both the numerator and the denominator of this fraction can be written as third-degree polynomials in k whose leading term is nk^3. Consequently, this fraction minus 1 can be written as a second-degree polynomial in k divided by a third-degree polynomial in k, which can be bounded by Ck^{-1} for some constant C.

Note that

$$\begin{aligned}q_{k,a,n} &= \frac{(k-1)a!(n-a+k-2)!}{(n+k-1)!} \le \frac{a!(n-a+k-2)!}{(n+k-2)!}\\ &= \frac{a!}{(n+k-2)_a} \le \frac{C}{k^a}.\end{aligned}\tag{7.14}$$

To compare $p_{k,l,a,n}$ and $q_{k,a,n}$ when $a \ge 2$, we will break up the probability $p_{k,l,a,n}$ by conditioning on the number of the original k balls that were sampled. Conditional on sampling $j \ge 1$ of the original k balls, the probability that the first a balls sampled are red and the next $n-a$ are black is $\binom{n}{a}^{-1}$ times the probability in (7.13), which can be bounded by $Cq_{k,a,n}$. The probability of sampling j of the original k balls is at most $C(k/l)^j$ by (7.12), so

$$|p_{k,l,a,n}-q_{k,a,n}|\le C\sum_{j=1}^{n}\left(\frac{k}{l}\right)^{j}q_{k,a,n}\le Ck^{-a}\sum_{j=1}^{n}\left(\frac{k}{l}\right)^{j}\le Ck^{-a}n\cdot\frac{k}{l}\le\frac{C}{kl}.$$

Finally, when $a=1$, we have

$$|p_{k,l,a,n}-q_{k,a,n}|\le C\sum_{j=1}^{n}\left(\frac{k}{l}\right)^{j}q_{k,a,n}\frac{C}{k}\le C\sum_{j=1}^{n}\left(\frac{k}{l}\right)^{j}k^{-2}\le\frac{C}{kl}.\qquad\square$$

7.5. *Proofs of Lemmas* 7.5, 7.6, 7.7 *and* 7.8.

PROOF OF LEMMA 7.5. For $2\le k\le l$, let $A_1^{k,l}$ be the event that $\sigma'(1),\dots,\sigma'(n)$ all have distinct ancestors at time $\gamma_{l+1}-$. Let $A_2^{k,l}$ be the event that the ancestor of $\sigma'(1)$ at time γ_l- has a different type from the ancestor of $\sigma'(1)$ at time $\gamma_{l+1}-$. Let $A_3^{k,l}$ be the event that one of the k individuals at time $\gamma_{k+1}-$ is the ancestor of $\sigma'(2),\dots,\sigma'(a+1)$ but not $\sigma'(a+2),\dots,\sigma'(n)$, and let $A_4^{k,l}$ be the event that the ancestor of this individual at time γ_k- has a different type. We claim that

$$\begin{aligned}&P(Z_1'=l,Z_2'=\cdots=Z_{a+1}'=k,\\ &\qquad Z_{a+2}'=\cdots=Z_n'=1\text{ for some }2\le k<l)\\ &\quad=P\left(\bigcup_{2\le k<l}A_1^{k,l}\cap A_2^{k,l}\cap A_3^{k,l}\cap A_4^{k,l}\right)+O\left(\frac{1}{(\log N)^2}\right).\end{aligned}\tag{7.15}$$

First consider the probability that $Z_1'=l,Z_2'=\cdots=Z_{a+1}'=k$ and $Z_{a+2}'=\cdots=Z_n'=1$ for some $2\le k<l$ but that not all of $A_1^{k,l}$, $A_2^{k,l}$, $A_3^{k,l}$ and $A_4^{k,l}$ occur for any k and l. Note that this can only happen in two ways. One way would be for $A_1^{k,l}$ not to hold, which would mean $\sigma'(1),\dots,\sigma'(n)$ do not all have distinct ancestors at time $\gamma_{l+1}-$. However, it follows from the argument used to prove (7.6) that $P((A_1^{k,l})^c\cap\{Z_1'=l\}\cap\{Z_2'=k\}$ for some $2\le k<l)$ is $O((\log N)^{-2})$. The second way would be for $A_1^{k,l}$ to hold but for $\sigma'(2),\dots,\sigma'(a+1)$ not all to have the same ancestor at time $\gamma_{k+1}-$. It follows from Lemma 7.2 that this possibility also has probability $O((\log N)^{-2})$.

Next, we consider the probability that $A_1^{k,l}$, $A_2^{k,l}$, $A_3^{k,l}$ and $A_4^{k,l}$ all hold, but we do not have $Z_1'=l,Z_2'=\cdots=Z_{a+1}'=k$ and $Z_{a+2}'=\cdots=Z_n'=1$. This is only possible if there is a third time γ, other than the times between γ_l and γ_{l+1} and between γ_k and γ_{k+1}, such that the type of the ancestor of one of the individuals $\sigma'(1), \dots, \sigma'(n)$ at time γ is different from the type of the ancestor at time $\gamma-$. However, it is a consequence of (7.5) that the probability that this occurs is at most $O((\log\log N)^3/(\log N)^3)$. It follows that (7.15) holds.

Recall from the proof of Lemma 7.3 that if two individuals with an infinite line of descent are chosen at random at time $\gamma_{k+1}-$, then the probability that they will have the same ancestor at time γ_k- is $2/(k(k-1))$. Since there are $\binom{n}{2}$ pairs of individuals, we have

$$P(A_1^{k,l}) \geq 1 - \binom{n}{2} \sum_{k=l+1}^{H} \frac{2}{k(k-1)} \geq 1 - \binom{n}{2}\frac{2}{l} \geq 1 - \frac{C}{l}.$$

We have $P(A_2^{k,l} | A_1^{k,l}) = r/[r(1-s) + ls]$ by (7.2). Next, note that if we choose at random one of the k individuals between times γ_k and γ_{k+1}, then the probability that the individual born at time γ_{k+1} is a descendant of the randomly chosen individual is $1/k$, and thereafter the probability that each new individual is a descendant of the randomly chosen individual is the fraction of the current individuals that are descended from the randomly chosen individual. This is the same description as the urn problem of Lemma 7.11, so conditional on $A_1^{k,l}$, the probability that $\sigma'(2), \dots, \sigma'(a+1)$ but not $\sigma'(a+2), \dots, \sigma'(n)$ are descended from the randomly chosen individual is $p_{k,l,a,n-1}$. Therefore, $P(A_3^{k,l} | A_1^{k,l} \cap A_2^{k,l}) = kp_{k,l,a,n-1}$. By (7.2), we have $P(A_4^{k,l} | A_1^{k,l} \cap A_2^{k,l} \cap A_3^{k,l}) = r/[r(1-s) + ks]$. By the arguments used to prove (7.5), the probability that $A_1^{k,l} \cap A_2^{k,l} \cap A_3^{k,l} \cap A_4^{k,l}$ holds for more than one pair (k,l) is at most $O((\log\log N)^3/(\log N)^3)$. Thus,

$$\begin{aligned} &P\biggl(\bigcup_{2 \leq k < l} A_1^{k,l} \cap A_2^{k,l} \cap A_3^{k,l} \cap A_4^{k,l}\biggr) \\ &\quad = \sum_{k=2}^{H}\sum_{l=k+1}^{H} \biggl(\frac{r}{r(1-s)+ls}\biggr)\biggl(\frac{kr}{r(1-s)+ks}\biggr)(p_{k,l,a,n-1})P(A_1^{k,l}) \\ &\qquad + O\biggl(\frac{(\log\log N)^3}{(\log N)^3}\biggr). \end{aligned} \tag{7.16}$$

By Lemma 7.11, we can write $p_{k,l,a,n-1} = q_{k,a,n-1} + \delta$, where $|\delta| \leq C/(kl)$. Also, $P(A_1) = 1 - \eta$, where $\eta \leq C/l$. Note that $r/[r(1-s)+ls] \leq r/(ls)$ and $kr/[r(1-s)+ks] \leq r/s$. Recall from (7.14) that $q_{k,a,n} \leq C/k$ for all $a \geq 1$. To complete the proof, we will need to simplify the four factors inside the sum in (7.16) by obtaining four inequalities. First, note that

$$\biggl|\frac{r}{r(1-s)+ls} - \frac{r}{ls}\biggr| = \frac{r^2(1-s)}{(r(1-s)+ls)(ls)} \leq \frac{r^2}{l^2s^2}.$$

Therefore,

$$\begin{aligned}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left|\frac{r}{r(1-s)+ls}-\frac{r}{ls}\right|\left(\frac{r}{s}\right)\left(\frac{C}{k}\right)&\le\frac{Cr^3}{s^3}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{1}{kl^2}\\&\le Cr^3=O\left(\frac{1}{(\log N)^3}\right).\end{aligned}\tag{7.17}$$

Also,

$$\left|\frac{kr}{r(1-s)+ks}-\frac{r}{s}\right|=\frac{r^2(1-s)}{(r(1-s)+ks)s}\le\frac{r^2}{ks^2}.$$

Therefore,

$$\begin{aligned}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left|\frac{kr}{r(1-s)+ks}-\frac{r}{s}\right|\left(\frac{r}{ls}\right)\left(\frac{C}{k}\right)&\le\frac{Cr^3}{s^3}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{1}{k^2l}\\&\le Cr^3\log H=O\left(\frac{\log\log N}{(\log N)^3}\right).\end{aligned}\tag{7.18}$$

Also,

$$\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left(\frac{r}{ls}\right)\left(\frac{r}{s}\right)\left(\frac{C}{kl}\right)\le\frac{Cr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{1}{kl^2}\le Cr^2=O\left(\frac{1}{(\log N)^2}\right)\tag{7.19}$$

and

$$\begin{aligned}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left(\frac{r}{ls}\right)\left(\frac{r}{s}\right)\left(\frac{C}{k}\right)(1-P(A_1^{k,l}))&\le\frac{Cr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{1}{kl^2}\\&=O\left(\frac{1}{(\log N)^2}\right).\end{aligned}\tag{7.20}$$

It follows from (7.16)–(7.20) that

$$\begin{aligned}&P\left(\bigcup_{2\le k<l}A_1^{k,l}\cap A_2^{k,l}\cap A_3^{k,l}\cap A_4^{k,l}\right)\\&\qquad=\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left(\frac{r}{ls}\right)\left(\frac{r}{s}\right)q_{k,a,n-1}+O\left(\frac{1}{(\log N)^2}\right),\end{aligned}$$

which, combined with (7.15), implies the lemma. □

PROOF OF LEMMA 7.6. Suppose $2\le k<l$. Let $B_1^{k,l}$ be the event that $Z_i\le l$ for $i=1,\dots,n$. Let $B_2^{k,l}$ be the event that $Z_1=l$ and $Z_i\ne l$ for all $2\le i\le n$. Let $B_3^{k,l}$ be the event that $Z_i\le k$ for all $2\le i\le n$. Let $B_4^{k,l}$ be the event that $Z_2=\cdots=$

$Z_{a+1} = k$ but $Z_i \neq k$ for $a + 2 \leq i \leq n$. Let $B_5^{k,l}$ be the event that $Z_{a+2} = \cdots = Z_n = 1$. Note that $Z_1 = l$, $Z_2 = \cdots = Z_{a+1} = k$ and $Z_{a+2} = \cdots = Z_n = 1$ for some $2 \leq k < l$ if and only if, for some $2 \leq k < l$, the event $B_1^{k,l} \cap B_2^{k,l} \cap B_3^{k,l} \cap B_4^{k,l} \cap B_5^{k,l}$ occurs. Furthermore, the events $B_1^{k,l} \cap \cdots \cap B_5^{k,l}$ are disjoint for different values of k and l, so we need to calculate $\sum_{k=2}^{H} \sum_{l=k+1}^{H} P(B_1^{k,l} \cap B_2^{k,l} \cap B_3^{k,l} \cap B_4^{k,l} \cap B_5^{k,l})$. We have

$$\begin{aligned} P(B_1^{k,l}) &= \prod_{j=l+1}^{H} E[(1 - V_j)^n] \geq \prod_{j=l+1}^{H} E[1 - nV_j] \\ &\geq 1 - \sum_{j=l+1}^{H} nE[V_j] = 1 - n \sum_{j=l+1}^{H} \frac{r}{js}. \end{aligned} \tag{7.21}$$

By Lemma 7.9,

$$\begin{aligned} P(B_2^{k,l} | B_1^{k,l}) &= \frac{r}{s} q_{l,1,n} = \frac{r}{s}\left(\frac{(l-1)(n+l-3)!}{(n+l-1)!}\right) \\ &= \frac{r}{sl}\left(\frac{l(l-1)}{(n+l-1)(n+l-2)}\right) \leq \frac{r}{sl}. \end{aligned}$$

By the same reasoning used to get (7.21), we have

$$P(B_3^{k,l} | B_1^{k,l} \cap B_2^{k,l}) \geq 1 - (n-1) \sum_{j=k+1}^{l-1} \frac{r}{js}. \tag{7.22}$$

By Lemma 7.9,

$$P(B_4^{k,l} | B_1^{k,l} \cap B_2^{k,l} \cap B_3^{k,l}) = \frac{r}{s} q_{k,a,n-1}.$$

Finally, by the argument used to establish (7.21) and (7.22),

$$P(B_5^{k,l} | B_1^{k,l} \cap B_2^{k,l} \cap B_3^{k,l} \cap B_4^{k,l}) \geq 1 - (n-a-1) \sum_{j=2}^{k-1} \frac{r}{js}. \tag{7.23}$$

Note that the product of the probabilities on the right-hand side of (7.21), (7.22) and (7.23) is at least $1 - n \sum_{j=1}^{H} \frac{r}{js} \geq 1 - \frac{C \log H}{\log N}$. Since $q_{k,a,n-1} \leq C/k$ by (7.14), we have

$$\begin{aligned} \sum_{k=2}^{H} \sum_{l=k+1}^{H} \left(\frac{r}{sl}\right)\left(\frac{r}{s}\right) q_{k,a,n-1} \left(\frac{C \log H}{\log N}\right) &\leq \frac{C}{(\log N)^3} \sum_{k=2}^{H} \sum_{l=k+1}^{H} \frac{(\log H)}{kl} \\ &\leq \frac{C (\log H)^3}{(\log N)^3} \end{aligned}$$

and so

$$\sum_{k=2}^{H}\sum_{l=k+1}^{H} P(B_1^{k,l}\cap B_2^{k,l}\cap B_3^{k,l}\cap B_4^{k,l}\cap B_5^{k,l})$$
$$(7.24)\qquad = \sum_{k=2}^{H}\sum_{l=k+1}^{H}\left(\frac{r}{sl}\right)\left(\frac{r}{s}\right)\left[\frac{l(l-1)}{(n+l-1)(n+l-2)}\right]q_{k,a,n-1} + O\left(\frac{1}{(\log N)^2}\right).$$

Finally, note that $|1-\frac{l(l-1)}{(n+l-1)(n+l-2)}|\le\frac{C}{l}$ for some constant C. Since $q_{k,a,n-1}\le C/k$ and

$$\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left(\frac{r}{sl}\right)\left(\frac{r}{s}\right)\frac{C}{kl}\le\frac{C}{(\log N)^2},$$

equation (7.24) remains true if the term in brackets is replaced by 1. The lemma follows. ☐

PROOF OF LEMMA 7.7. Let A_1^k be the event that one of the k individuals at time $\gamma_{k+1}-$ is the ancestor of $\sigma'(1),\dots,\sigma'(a)$ but not $\sigma'(a+1),\dots,\sigma'(n)$, and let A_2^k be the event that the ancestor of this individual at time γ_k- has a different type. It follows from Lemma 7.2 that the probability that, for some $k\ge 2$, we have $Z_1'=\cdots=Z_a'=k$ and $Z_{a+1}'=\cdots=Z_n'=1$ but the event $A_1^k\cap A_2^k$ does not occur is at most $O((\log N)^{-2})$. We will therefore calculate the probability that the event $A_1^k\cap A_2^k\cap\{Z_1'=\cdots=Z_a'=k\}\cap\{Z_{a+1}'=\cdots=Z_n'=1\}$ occurs for some $k\ge 2$. Note that this occurs for at most one value of k, so we may sum the probabilities over $k=2,\dots,H$.

Note that $P(A_1^k)=kp_{k,H,a,n}$ and $P(A_2^k|A_1^k)=r/(r(1-s)+ks)$ by (7.2). It follows that $P(A_1^k\cap A_2^k)=[kr/(r(1-s)+ks)]p_{k,H,a,n}$. Note that $kr/(r(1-s)+ks)\le r/s$, and recall that $|p_{k,H,a,n}-q_{k,a,n}|\le C/(kH)$ by Lemma 7.11. Therefore,

$$\sum_{k=2}^{H}\left(\frac{kr}{r(1-s)+ks}\right)|p_{k,H,a,n}-q_{k,a,n}|\le\frac{Cr}{s}\sum_{k=2}^{H}\frac{1}{kH}\le\frac{Cr\log H}{H}\le\frac{C}{(\log N)^5}.$$

It follows that $\sum_{k=2}^{H}P(A_1^k\cap A_2^k)=\sum_{k=2}^{H}(\frac{kr}{r(1-s)+ks})q_{k,a,n}+O(1/(\log N)^5)$. Also, $q_{k,a,n}\le C/k$, so

$$\sum_{k=2}^{H}\left(\frac{kr}{r(1-s)+ks}-\frac{r}{s}\right)q_{k,a,n}\le\sum_{k=2}^{H}\left(\frac{r^2}{ks^2}\right)\frac{C}{k}=O\left(\frac{1}{(\log N)^2}\right).$$

Thus,

$$\sum_{k=2}^{H} P(A_1^k \cap A_2^k) = \frac{r}{s} \sum_{k=2}^{H} q_{k,a,n} + O\left(\frac{1}{(\log N)^2}\right). \tag{7.25}$$

If A_1^k and A_2^k both occur, then we will have $Z_1' = \cdots = Z_a' = k$ and $Z_{a+1}' = \cdots = Z_n' = 1$ unless either $Z_i' = l$ for some $i = 1, \ldots, n$ and $l \notin \{1, k\}$ or $Z_i' = k$ for some $i \geq a+1$. By Lemma 7.2, we have $P(A_1^k \cap A_2^k \cap \{Z_i = k\}$ for some $k \geq 2$ and $i \geq a+1) \leq C/(\log N)^2$. Therefore, we only need to consider the possibility that $Z_i' = l$ for some $i = 1, \ldots, n$ and $l \notin \{1, k\}$. We will treat separately the cases $l < k$ and $l > k$. Note that by (7.5), the probability that A_1^k and A_2^k both occur, $Z_i' = l_1$ and $Z_j' = l_2$, where l_1 and l_2 are distinct integers not in $\{1, k\}$, is at most $O((\log \log N)^3/(\log N)^3)$.

We first consider $l > k$. By (7.6) the probability that A_1^k and A_2^k both occur and $Z_i' = Z_j' = l$ for some $i \neq j$ is $O((\log N)^{-2})$. By the same argument used to prove Lemma 7.5, the probability that $A_1^k \cap A_2^k$ for some k but $Z_i' = l$ for some $l > k$ is

$$\frac{nr^2}{s^2} \sum_{k=2}^{H} \sum_{l=k+1}^{H} \frac{q_{k,a,n}}{l} + O\left(\frac{1}{(\log N)^2}\right). \tag{7.26}$$

There are two differences between this formula and the result of Lemma 7.5, which can be explained as follows. First, in place of the event $A_2^{k,l}$, we need the event that, for some $i = 1, \ldots, n$, the ancestor of $\sigma'(i)$ at time γ_l- has a different type from the ancestor of $\sigma'(i)$ at time $\gamma_{l+1}-$. This is why the double summation is multiplied by n. Second, instead of $A_3^{k,l}$, we need one of the individuals at time $\gamma_{k+1}-$ to be the ancestor of $\sigma'(1), \ldots, \sigma'(a)$ but not $\sigma'(a+1), \ldots, \sigma'(n)$, rather than $\sigma'(2), \ldots, \sigma'(a+1)$ but not $\sigma'(a+2), \ldots, \sigma'(n)$. This is why we have $q_{k,a,n}$ in the formula rather than $q_{k,a,n-1}$. Otherwise, the calculation proceeds as before.

If $a \geq 2$, a consequence of (7.6) is that the probability that $A_1^k \cap A_2^k$ for some k but $Z_i' = l$ for some $l < k$ is $O((\log N)^{-2})$. Thus, (7.8) follows by subtracting (7.26) from (7.25). Now, consider the case $a = 1$. Let S be a d-element subset of $\{2, \ldots, n\}$. By the argument used to prove Lemma 7.5, the probability that, for some $2 \leq l < k$, the events $A_{1,k}$ and $A_{2,k}$ occur but $Z_i' = l$ for $i \in S$ and $Z_i' = 1$ for $i \in \{2, \ldots, n\} \setminus S$ is

$$\frac{r^2}{s^2} \sum_{k=2}^{H} \sum_{l=2}^{k-1} \frac{q_{l,d,n-1}}{k} + O\left(\frac{1}{(\log N)^2}\right).$$

Summing this over $d = 1, \ldots, n-1$ and all subsets S of size d, we get that the probability that $A_{1,k}$ and $A_{2,k}$ occur but $Z_i' = l$ for $i \in S$ and $Z_i' = 1$ for $i \in \{2, \ldots, n\} \setminus S$ for some nonempty $S \subset \{2, \ldots, n\}$ is

$$\frac{r^2}{s^2} \sum_{k=2}^{H} \sum_{l=2}^{k-1} \frac{1}{k} \left(\sum_{d=1}^{n-1} \binom{n-1}{d} q_{l,d,n-1} \right) + O\left(\frac{1}{(\log N)^2}\right). \tag{7.27}$$

Using the probabilistic interpretation of the $q_{l,d,n-1}$ as in Lemma 7.10, we have

$$\sum_{d=1}^{n-1}\binom{n-1}{d}q_{l,d,n-1} = 1 - q_{l,0,n-1} = 1 - \frac{(l-1)((n-1)+l-2)!}{((n-1)+l-1)!}$$
$$= 1 - \frac{l-1}{n+l-2} = \frac{n-1}{n+l-2}.$$

Thus, (7.27) becomes

$$\frac{(n-1)r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{1}{k(n+l-2)} + O\left(\frac{1}{(\log N)^2}\right). \tag{7.28}$$

We get (7.9) by subtracting (7.28) and (7.26) from (7.25). □

LEMMA 7.12. *Let* $\delta_1,\dots,\delta_N \in (0,1)$. *Assume that* $\delta = \delta_1 + \cdots + \delta_n \in (0,1)$. *Then*

$$\delta(1-\delta) \le 1 - \prod_{n=1}^{N}(1-\delta_n) \le \delta.$$

PROOF. The second inequality follows from $|\prod_{n=1}^{N} 1 - \prod_{n=1}^{N}(1-\delta_n)| \le \sum_{n=1}^{N}\delta_n$. To prove the first inequality using the second, note that

$$1 - \prod_{n=1}^{N}(1-\delta_n) = \sum_{m=1}^{N}\left(\prod_{n=1}^{m-1}(1-\delta_n) - \prod_{n=1}^{m}(1-\delta_n)\right)$$
$$= \sum_{m=1}^{N}\left(\prod_{n=1}^{m-1}(1-\delta_n)\right)\delta_m$$
$$\ge \sum_{m=1}^{N}(1-\delta)\delta_m = \delta(1-\delta).$$

□

PROOF OF LEMMA 7.8. Let $B_1^k = \{Z_i \le k \text{ for } i = 1,\dots,n\}$. Let $B_2^k = \{Z_i = k \text{ for } 1 \le i \le a \text{ and } Z_j < k \text{ for } a+1 \le j \le n\}$. Let $B_3^k = \{Z_i = 1 \text{ for } a+1 \le i \le n\}$. We have

$$P(B_1^k \cap B_2^k \cap B_3^k \text{ for some } k \ge 2)$$
$$= \sum_{k=2}^{H} P(B_1^k)P(B_2^k|B_1^k)P(B_3^k|B_1^k \cap B_2^k) \tag{7.29}$$
$$= \sum_{k=2}^{H}\left(\prod_{l=k+1}^{H} E[(1-V_l)^n]\right)\left(\frac{r}{s}q_{k,a,n}\right)\left(\prod_{l=2}^{k-1} E[(1-V_l)^{n-a}]\right).$$

Using Lemma 7.9,

$$\begin{aligned}E[(1-V_l)^m] &= \left(1-\frac{r}{s}\right)+\frac{r}{s}q_{l,0,m} = \left(1-\frac{r}{s}\right)+\frac{r}{s}\left(\frac{(l-1)(m+l-2)!}{(m+l-1)!}\right)\\ &= 1-\frac{rm}{s(m+l-1)}.\end{aligned}$$

Therefore, the expression on the right-hand side of (7.29) is

$$\frac{r}{s}\sum_{k=2}^{H}\left[\prod_{l=k+1}^{H}\left(1-\frac{nr}{s(n+l-1)}\right)\right]\left[\prod_{l=2}^{k-1}\left(1-\frac{(n-a)r}{s(n-a+l-1)}\right)\right]q_{k,a,n}.$$

Let $\delta=\frac{r}{s}\sum_{l=k+1}^{H}\frac{n}{n+l-1}+\frac{r}{s}\sum_{l=2}^{k-1}\frac{n-a}{n-a+l-1}$. Then

$$\begin{aligned}\delta^2 &= \frac{r^2}{s^2}\left(\sum_{l=k+1}^{H}\frac{n}{n+l-1}+\sum_{l=2}^{k-1}\frac{n-a}{n-a+l-1}\right)^2\\ &\le \frac{r^2}{s^2}\left(n\sum_{l=1}^{H}\frac{1}{l}\right)^2 \le Cr^2(\log H)^2.\end{aligned}$$

Since $q_{k,a,n}\le C/k$ by (7.14), we have $\frac{r}{s}\sum_{k=2}^{H}\delta^2 q_{k,a,n}\le Cr^3(\log H)^2\sum_{k=2}^{H}\frac{1}{k}\le Cr^3(\log H)^3$. Using Lemma 7.12, the right-hand side of (7.29) can be written as

$$\begin{aligned}&\frac{r}{s}\sum_{k=2}^{H}\left(1-\sum_{l=k+1}^{H}\frac{nr}{s(n+l-1)}-\sum_{l=2}^{k-1}\frac{(n-a)r}{s(n-a+l-1)}\right)q_{k,a,n}\\ &\qquad+O\left(\frac{(\log\log N)^3}{(\log N)^3}\right).\end{aligned}\tag{7.30}$$

We have $\frac{1}{l}-\frac{1}{n+l-1}=\frac{n-1}{l(n+l-1)}\le\frac{n}{l^2}$. Since $q_{k,a,n}\le C/k$, it follows that

$$\frac{nr^2}{s^2}\sum_{k=2}^{H}\sum_{l=k+1}^{H}\left|\frac{1}{(n+l-1)}-\frac{1}{l}\right|q_{k,a,n}\le Cr^2\sum_{k=2}^{H}\sum_{l=k+1}^{H}\frac{1}{kl^2}\le\frac{C}{(\log N)^2}.\tag{7.31}$$

Since $q_{k,a,n}\le C/k^a$ by (7.14), when $a\ge 2$ we have

$$\frac{r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{n-a}{n-a+l-1}q_{k,a,n}\le Cr^2\sum_{l=2}^{H}\sum_{k=l+1}^{H}\frac{1}{lk^2}=O\left(\frac{1}{(\log N)^2}\right).\tag{7.32}$$

By combining (7.30), (7.31) and (7.32), we get (7.10) when $a\ge 2$. When $a=1$, note that

$$\frac{n-a}{n-a+l-1}q_{k,a,n}=\frac{(n-1)(k-1)}{(n+l-2)(n+k-1)(n+k-2)}.$$

Also, note that $|\frac{k-1}{(n+k-1)(n+k-2)} - \frac{1}{k}| \le \frac{C}{k^2}$. It follows that, when $a = 1$, we have

$$\begin{aligned}&\frac{r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{n-a}{n-a+l-1}q_{k,a,n}\\ &\qquad = \frac{(n-1)r^2}{s^2}\sum_{k=2}^{H}\sum_{l=2}^{k-1}\frac{1}{k(n+l-2)} + O\biggl(\frac{1}{(\log N)^2}\biggr).\end{aligned} \tag{7.33}$$

Equations (7.30), (7.31) and (7.33) establish (7.11) when $a = 1$. □

REFERENCES

[1] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, New York.

[2] BARTON, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetic Research Cambridge* **72** 123–133.

[3] BARTON, N. H., ETHERIDGE, A. M. and STURM, A. K. (2004). Coalescence in a random background. *Ann. Appl. Probab.* **14** 754–785.

[4] DONNELLY, P. and KURTZ, T. G. (1999). Genealogical processes for Fleming–Viot models with selection and recombination. *Ann. Appl. Probab.* **9** 1091–1148.

[5] DURRETT, R. (1996). *Probability*: *Theory and Examples*, 2nd ed. Duxbury, Belmont, CA.

[6] DURRETT, R. (2002). *Probability Models for DNA Sequence Evolution*. Springer, New York.

[7] DURRETT, R. and SCHWEINSBERG, J. (2004). Approximating selective sweeps. *Theoret. Population Biol.* **66** 129–138.

[8] DURRETT, R. and SCHWEINSBERG, J. (2004). A coalescent model for the effect of advantageous mutations on the genealogy of a population. Preprint. Available at http://front.math.ucdavis.edu/ math.PR/0411071.

[9] ETHERIDGE, A. M., PFAFFELHUBER, P. and WAKOLBINGER, A. (2005). An approximate sampling formula under genetic hitchhiking. Preprint. Available at http://front.math.ucdavis.edu/math.PR/0503485.

[10] GADAG, V. G. and RAJARSHI, M. B. (1987). On multitype processes based on progeny length particles of a supercritical Galton–Watson process. *J. Appl. Probab.* **24** 14–24.

[11] GADAG, V. G. and RAJARSHI, M. B. (1992). On processes associated with a super-critical Markov branching process. *Serdica* **18** 173–178.

[12] GILLESPIE, J. H. (2000). Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155** 909–919.

[13] JOHNSON, N. L., KOTZ, S. and KEMP, A. W. (1992). *Univariate Discrete Distributions*, 2nd ed. Wiley, New York.

[14] JOYCE, P. and TAVARÉ, S. (1987). Cycles, permutations and the structure of the Yule process with immigration. *Stochastic Process. Appl.* **25** 309–314.

[15] KAPLAN, N. L., HUDSON, R. R. and LANGLEY, C. H. (1989). The “hitchhiking effect” revisited. *Genetics* **123** 887–899.

[16] KINGMAN, J. F. C. (1978). The representation of partition structures. *J. London Math. Soc.* **18** 374–380.

[17] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248.

[18] MAYNARD SMITH, J. and HAIGH, J. (1974). The hitch-hiking effect of a favorable gene. *Genetic Research* **23** 23–35.

[19] MORAN, P. A. P. (1958). Random processes in genetics. *Proceedings of the Cambridge Philosophical Society* **54** 60–71.

[20] MÖHLE, M. and SAGITOV, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29** 1547–1562.

[21] O'CONNELL, N. (1993). Yule process approximation for the skeleton of a branching process. *J. Appl. Probab.* **30** 725–729.

[22] PITMAN, J. (1999). Coalescents with multiple collisions. *Ann. Probab.* **27** 1870–1902.

[23] PRZEWORSKI, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics* **160** 1179–1189.

[24] SAGITOV, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** 1116–1125.

[25] SCHWEINSBERG, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5** 1–50.

[26] SIMONSEN, K. L., CHURCHILL, G. A. and AQUADRO, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141** 413–429.

[27] STEPHAN, W., WIEHE, T. and LENZ, M. W. (1992). The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theoret. Population Biol.* **41** 237–254.

DEPARTMENT OF MATHEMATICS, 0112
UNIVERSITY OF CALIFORNIA AT SAN DIEGO
9500 GILMAN DRIVE
LA JOLLA, CALIFORNIA 92093-0112
USA
E-MAIL: jschwein@math.ucsd.edu

DEPARTMENT OF MATHEMATICS
MALOTT HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-4201
USA
E-MAIL: rtd1@cornell.edu