Electronic Journal of Probability

Vol. 12 (2007), Paper no. 28, pages 808–847.

Journal URL
http://www.math.washington.edu/~ejpecp/

The Common Ancestor Process for a Wright-Fisher Diffusion

Jesse E. Taylor [*]
Department of Statistics, University of Oxford
1 South Parks Road, Oxford, OX1 3TG, United Kingdom
jtaylor@stats.ox.ac.uk

Abstract

Rates of molecular evolution along phylogenetic trees are influenced by mutation, selection and genetic drift. Provided that the branches of the tree correspond to lineages belonging to genetically isolated populations (e.g., multi-species phylogenies), the interplay between these three processes can be described by analyzing the process of substitutions to the common ancestor of each population. We characterize this process for a class of diffusion models from population genetics theory using the structured coalescent process introduced by Kaplan et al. (1988) and formalized in Barton et al. (2004). For two-allele models, this approach allows both the stationary distribution of the type of the common ancestor and the generator of the common ancestor process to be determined by solving a one-dimensional boundary value problem. In the case of a Wright-Fisher diffusion with genic selection, this solution can be found in closed form, and we show that our results complement those obtained by Fearnhead (2002) using the ancestral selection graph. We also observe that approximations which neglect recurrent mutation can significantly underestimate the exact substitution rates when selection is strong. Furthermore, although we are unable to find closed-form expressions for models with frequency-dependent selection, we can still solve the corresponding boundary value problem numerically and then use this solution to calculate the substitution rates to the common ancestor. We illustrate this approach by studying the effect of dominance on the common ancestor process in a diploid population. Finally, we show that the theory can be formally extended to diffusion models with more than two genetic backgrounds, but that

[*] Supported by EPSRC grant EP/E010989/1.

it leads to systems of singular partial differential equations which we have been unable to solve.

Key words: Common-ancestor process, diffusion process, structured coalescent, substitution rates, selection, genetic drift.

AMS 2000 Subject Classification: Primary 60J70, 92D10, 92D20.

Submitted to EJP on November 14, 2006, final version accepted May 22, 2007.

1 Introduction

One of the key insights to emerge from population genetics theory is that the effectiveness of natural selection is reduced by random variation in individual survival and reproduction. Although the expected frequency of a mutation will either rise or fall according to its effect on fitness, evolution in finite populations also depends on numerous chance events which affect individual life histories in a manner independent of an individual's genotype. Collectively, these events give rise to a process of stochastic fluctuations in genotype frequencies known as genetic drift (Gillespie 2004). For example, a mutation which confers resistance to a lethal infection will still decline in frequency if, by chance, disproportionately many of the individuals carrying that mutation are killed in a severe storm. Moreover, if the mutation is initially carried by only a few individuals, then it may be lost altogether from the population following such a catastrophe. Because it is counterintuitive that populations may evolve to become less fit, there has been much interest in the consequences of stochasticity for other aspects of adaptive evolution, such as the origin of sex (Poon and Chao 2004; Barton and Otto 2005), genome composition (Lynch and Conery 2003), and speciation and extinction (Whitlock 2000; Gavrilets 2003).

Testing these theories requires quantifying genetic drift and selection in natural populations. Although selection and drift can sometimes be inferred from historical changes in the distribution of a trait (Lande 1976) or genotype frequencies (O'Hara 2005), population genetical processes are mainly investigated using sets of contemporaneously sampled DNA sequences. For our purposes, it is useful to distinguish two scenarios. On the one hand, sequences sampled from a single population will usually share a common history shaped by selection and drift, and must be analyzed using models which take that shared history into account. One approach is to reduce the data to a set of summary statistics whose distribution can be predicted using population genetical models (Sawyer and Hartl 1992; Akashi 1995; Bustamante et al. 2001). Alternatively, more powerful analyses can be designed by using coalescent models and Monte Carlo simulations to estimate the joint likelihood of the data and the unobserved genealogy under different assumptions about selection and drift (Stephens and Donnelly 2003; Coop and Griffiths 2004). In both cases, the selection coefficients estimated with these methods will reflect the combined effects of selection and genetic drift in the population from which the sample was collected.

In contrast, when the data consists of sequences sampled from different species, then the time elapsed since any of the ancestors last belonged to a common population may be so great that the genealogy of the sample is essentially unrelated to the population genetical processes of interest. In this case, the genealogy is usually inferred using purely phylogenetic methods, and evolutionary inferences are facilitated by making certain simplifying assumptions about the way in which natural selection influences the substitution process along branches of this tree, i.e., the process of mutations to the ancestral lineages of the members of the sample. It is usually assumed that the substitution process along each branch of the tree is a Markov process, and that substitutions by beneficial or deleterious mutations occur at rates which are either greater than or less than the neutral mutation rate (Yang 1996). While the first assumption is true only when evolution is neutral, i.e., mutations do not affect fitness, the latter assumption reflects the fact that mutations which either increase or decrease the likelihood of a lineage persisting into the future are likely to be over- or under-represented, respectively, on lineages which do in fact persist. For example, it is often possible to identify proteins which are under unusually strong selection simply by comparing the rates of substitutions which change the amino acid composition of the protein with those which do not (Nielsen and Yang 1998).

An important limitation of purely phylogenetic analyses of selection is that the relationship between the phylogenetic rate parameters and population genetical quantities is usually obscure. One exception is when less fit variants are in fact lethal, so that selection is fully efficient and certain substitutions are never observed in live individuals. Alternatively, if the mutation rates are small enough that each new mutation is either rapidly lost or fixed in the population, then under some circumstances the substitution rate can be approximated by the flux of mutations which go to fixation (Kimura 1964). This approach has been used by McVean and Vieira (2001) to estimate the strength of selection on so-called silent mutations (i.e, those which do not change amino acid sequences) in several *Drosophila* species.

The common ancestor process can be used to describe the relationship between phylogenetic substitution rates and population genetical processes when the preceding approximations do not hold. The common ancestor of a population is any individual which is ancestral to the entire population. For the models which will be studied in this paper, such an individual will be guaranteed to exist at some time sufficiently (but finitely) far into the past and will be unique at any time at which it does exist. Denoting the type of the common ancestor alive at time t by z_t, we will define the substitution process to the common ancestor to be the stochastic process $(z_t : t \in \mathcal{R})$ and the common ancestor distribution to be the stationary distribution of z_t. This process will be a good approximation to the substitution process along the branches of a phylogenetic tree provided that the time elapsed along each branch is large in comparison with the coalescent time scales of the populations containing the sampled individuals and their ancestors. In particular, the divergence between the sequences in the sample should be much greater than the polymorphism within the populations from which the sample was collected. As is customary in modeling molecular evolution (Zharkikh 1994), we will assume that these populations are at equilibrium and that evolutionary processes such as mutation and selection do not vary along ancestral lineages. Although common ancestor processes could also be defined for non-equilibrium and time-inhomogeneous models, characterization of such processes will be substantially more difficult than in the idealized cases considered here.

Common ancestor distributions were first described for supercritical multitype branching processes by Jagers (1989, 1992), who showed that the distribution of the type of an individual spawning a branching process which survives forever has a simple representation involving the leading left and right eigenvectors of the first moment generator of the branching process. Because such an individual gives rise to infinitely many lineages which survive forever, but which individually do not give rise to the entire future population, it is not meaningful to speak of the common ancestor process in this setting. Instead, we must study what Georgii and Baake (2003) call the retrospective process, which characterizes the substitution process along lineages which survive forever. This process was also first described by Jagers (1989, 1992), who showed it to be a stationary time-homogeneous Markov process having the common ancestor distribution as its stationary measure. Extensive results concerning the retrospective process and common ancestor distribution can be found in Georgii and Baake (2003) and Baake and Georgii (2007).

Much less is known about the common ancestor process for traditional population genetical models such as the Moran and Wright-Fisher processes in which the population size remains constant. For neutral models, the fact that the substitution process decouples from the genealogy of a sample can be used to deduce that the common ancestor process is simply the neutral mutation process and that the common ancestor distribution is the stationary measure of this process. That this also holds true in the diffusion limit can be shown using the look-down construction of Donnelly and Kurtz (1996), which provides a particle representation for the Wright-Fisher diffusion. The key idea behind this construction is to assign particles to levels and then introduce look-down events which differ from (and replace) the usual neutral two-particle birth-death events of the Moran model in the requirement that it is always the particle occupying the higher level which dies and is then replaced by an offspring of the particle occupying the lower level. In the absence of selection, the common ancestor is the particle occupying the lowest level, as this individual never dies and it can be shown that all particles occupying higher levels have ancestors which coalesce with this lowest level in finite time.

In contrast, when selection is incorporated into the look-down process, particles can jump to higher levels and the common ancestor is no longer confined to the lowest level (Donnelly and Kurtz 1999). Furthermore, because the effect of selection depends on the frequencies of the types segregating in the population, e.g., selection has no effect if the population is monomorphic, we do not expect the non-neutral common ancestor process to be a Markov process. However, the mathematical difficulties which this creates can be overcome with the same technique that is used to characterize the genealogical processes of such models, namely by enlarging the state space of the process of interest until we obtain a higher dimensional process which does satisfy the Markov property. One such enlargement is the ancestral selection graph of Krone and Neuhauser (1997), which augments the ancestral lineages of the genealogy with a random family of 'virtual' lineages which are allowed to both branch and coalesce backwards in time. Fearnhead (2002) uses a related process to identify the common ancestor process for the Wright-Fisher diffusion with genic selection. His treatment relies on the observation that when there is only a single ancestral lineage, certain classes of events can be omitted from the ancestral selection graph so that the accessible particle configurations consist of the common ancestor, which can be of either type, plus a random number of virtual particles, all of the less fit type. This allows the common ancestor process to be embedded within a relatively tractable bivariate Markov process (z_t, n_t), where z_t is the type of the common ancestor and n_t is the number of virtual lineages.

In this article, we will use a different enlargement of the non-neutral coalescent. Our treatment relies on the structured coalescent introduced by Kaplan et al. (1988) and formalized by Barton et al. (2004), which subdivides the population into groups of exchangeable individuals sharing the same genotype and records both the types of the lineages ancestral to a sample from the population and the past frequencies of those types. With this approach, the common ancestor process of a population segregating two alleles can be embedded within a bivariate process (z_t, p_t), where p_t is the frequency at time t of one of the two alleles. We will show that both the stationary distribution and the generator of this process can be expressed in terms of the solution to a simple boundary value problem (Eq. 9) which determines the distribution of the type of the common ancestor conditional on the frequency at which that type occurs within the population. In certain cases we can solve this problem exactly and obtain an analytical characterization of the common ancestor process. However, one advantage of the diffusion-theoretic approach described here is that even when we cannot write down an explicit solution, we can still solve the corresponding boundary problem numerically. This makes it possible to calculate the substitution rates to the common ancestor for a much more general set of population genetical models than can be dealt with using the ancestral selection graph, including models with frequency-dependent selection, which we illustrate in Section 5, as well as fluctuating selection and genetic hitchhiking which will be described elsewhere.

The remainder of the article is structured as follows. In Section 2 we describe the class of diffusion processes to be studied and we briefly recall the construction of the structured coalescent in a fluctuating background as well as its restriction to a single ancestral lineage, which we call the structured retrospective process. Using calculations with generators, we describe the stationary distribution of the structured retrospective process and identify the common ancestor process by reversing the retrospective process with respect to this measure. We also give an alternative probabilistic representation for the conditional distribution of the type of the common ancestor, and in Section 3 we use this to derive asymptotic expressions for the substitution rates to the common ancestor when the mutation rates are vanishingly small. Sections 4 and 5 are concerned with applications of these methods to concrete examples, and we first consider the Wright-Fisher diffusion with genic (frequency-independent) selection. In this case we can write the density of the common ancestor distribution in closed form (Eq. 23), and we show that this quantity is related to the probability generating function of a distribution which arises in the graphical representation of Fearnhead (2002). Notably, these calculations also show that approximations which neglect recurrent mutation (e.g., the weak mutation limits) can underestimate the true substitution rates by an order of magnitude or more when selection is strong. In contrast, few explicit calculations are possible when we incorporate dominance into the model in Section 5, and we instead resort to numerically solving the associated boundary value problem to determine the substitution rates to the common ancestor. In the final section we show that some of these results can be formally extended to diffusion models with more than two genetic backgrounds, but that the usefulness of the theory is limited by the need to solve boundary value problems involving systems of singular PDE's.

2 Diffusions, coalescents and the common ancestor

We begin by recalling the structured coalescent process introduced by Kaplan et al. (1989) and more recently studied by Barton et al. (2004) and Barton and Etheridge (2004). Consider a closed population, of constant size N, and let P and Q be two alleles which can occur at a particular locus. Suppose that the mutation rates from Q to P and from P to Q are μ_1 and μ_2, respectively, where both rates are expressed in units of events per N generations. Suppose, in addition, that the relative fitnesses of P and Q are equal to $1 + \sigma(p)/N$ and 1, respectively, where p is the frequency of P. For technical reasons, we will assume that the selection coefficient $\sigma : [0, 1] \to \infty$ is the restriction of a function which is smooth on a neighborhood of $[0, 1]$, e.g., $\sigma(p)$ could be a polynomial function of the frequency of P. If we let p_t denote the frequency of P at time t and we measure time in units of N generations, then for sufficiently large N the time evolution of p_t can be approximated by a Wright-Fisher diffusion with generator

$$A\phi(p) = \frac{1}{2}p(1-p)\phi''(p) + \big(\mu_1(1-p) - \mu_2 p + \sigma(p)p(1-p)\big)\phi'(p), \tag{1}$$

where $\phi \in \mathcal{C}^2([0, 1])$. If we instead consider a diploid population, then the time evolution of the frequency of P can be modeled by the same diffusion approximation if we replace N by $2N$.

We note that because the drift and variance coefficients are smooth, Theorem 2.1 of Ethier and Kurtz [(1986), Chapter 8] tells us that the set $\mathcal{C}_0^\infty([0, 1])$ of infinitely differentiable functions with support contained in the interior of $(0,1)$ is a core for A. Furthermore, provided that both mutation rates μ_1 and μ_2 are positive, then the diffusion corresponding to (1) has a unique stationary measure $\pi(dp)$ on $[0,1]$, with density (Shiga 1981, Theorem 3.1; Ewens 2004, Section 4.5),

$$\pi(p) = Cp^{2\mu_1-1}(1-p)^{2\mu_2-1}\exp\left(2\int_0^p \sigma(q)dq\right), \tag{2}$$

where C is a normalizing constant. Unless stated otherwise (i.e., when we consider weak mutation limits in Section 3), we will assume throughout this article that both mutation rates are positive.

Although the structured coalescent can be fully characterized for this diffusion model, for our purposes it will suffice to consider only the numbers of ancestral lineages of type P or Q, which we denote $\tilde{n}_1(t)$ and $\tilde{n}_2(t)$, respectively. Here, and throughout the article, we will use the tilde, both on random variables and on generators, to indicate a stochastic process which is running from the present (usually the time of sampling) to the past. Then, as shown in Barton et al. (2004), the generator $\tilde{G}$ of the structured coalescent process $(\tilde{n}_1(t), \tilde{n}_2(t), \tilde{p}_t)$ can be written as

$$\begin{aligned}\tilde{G}\phi(n_1,n_2,p) \quad = \quad & \binom{n_1}{2}\left(\frac{1}{p}\right)[\phi(n_1-1,n_2,p)-\phi(n_1,n_2,p)] + \\ & \binom{n_2}{2}\left(\frac{1}{1-p}\right)[\phi(n_1,n_2-1,p)-\phi(n_1,n_2,p)] + \\ & n_1\mu_1\left(\frac{1-p}{p}\right)[\phi(n_1-1,n_2+1,p)-\phi(n_1,n_2,p)] + \\ & n_2\mu_2\left(\frac{p}{1-p}\right)[\phi(n_1+1,n_2-1,p)-\phi(n_1,n_2,p)] + A\phi(n_1,n_2,p),\end{aligned} \tag{3}$$

where for each $(n_1,n_2) \in \mathcal{N}\times\mathcal{N}$, we have $\phi(n_1,n_2,\cdot) \in \mathcal{C}^2([0,1])$. Barton et al. (2004) prove that a Markov process corresponding to this generator exists and is unique, and moreover that this process is the weak limit of a suitably rescaled sequence of Markov processes describing both the sample genealogy and the allele frequencies in a population of size N evolving according to a Moran model. One particularly convenient property of biallelic diffusion models is that the process $\tilde{p}(t)$ governing the evolution of allele frequencies backwards in time in a stationary population has the same law as the original Wright-Fisher diffusion $p(t)$ corresponding to the generator A. In fact, this property is shared by one-dimensional diffusions in general, which satisfy a detailed balance condition with respect to their stationary distributions (Nelson 1958). This will not be true (in general) of the multidimensional diffusion models considered in Section 6, where we will characterize the common ancestor process at a locus which can occur in more than two genetic backgrounds which can change either by mutation or by recombination.

Because we are only concerned with substitutions to single lineages, we need only consider sample configurations (n_1,n_2) which are either $(1,0)$ or $(0,1)$, and so we can replace the trivariate process $(\tilde{n}_1(t), \tilde{n}_2(t), \tilde{p}_t)$ with a bivariate process $(\tilde{z}, \tilde{p}_t)$ taking values in the space $E = (\{1\}\times(0,1]) \cup (\{2\}\times[0,1))$, where $\tilde{z}_t = 1$ if the lineage is of type P and $\tilde{z}_t = 2$ if it is of type Q. We will refer to $(\tilde{z}_t, \tilde{p}_t)$ as the structured retrospective process to emphasize the fact that it describes evolution *backwards in time.* (In contrast, Georgii and Baake (2003) define a retrospective process for a multitype branching process which runs forwards in time.) With this notation, the generator of the structured retrospective process can be written as

$$\begin{aligned} \tilde{G}\phi(1,p) &= \mu_1\left(\frac{1-p}{p}\right)[\phi(2,p)-\phi(1,p)] + A\phi(1,p) \\ \tilde{G}\phi(2,p) &= \mu_2\left(\frac{p}{1-p}\right)[\phi(1,p)-\phi(2,p)] + A\phi(2,p), \end{aligned} \tag{4}$$

for functions $\phi \in \mathcal{D}(\tilde{G}) \equiv \mathcal{C}_c^2(E)$ which are twice continuously differentiable on E and have compact support. For future reference we note that $\mathcal{D}(\tilde{G})$ is dense in the space $\hat{C}(E)$ of continuous functions on E vanishing at infinity and that $\mathcal{D}(\tilde{G})$ is an algebra. The key step in proving the existence and uniqueness of a Markov process corresponding to this generator is to show that the ancestral lineage is certain to jump away from a type before the frequency of that type vanishes, e.g., the ancestor will almost surely mutate from P to Q before the diffusion $\tilde{p}_t$ hits 0. This will guarantee that the jump terms appearing in $\tilde{G}$, which diverge at the boundaries of the state space, are in fact bounded along trajectories of the process over any finite time interval $[0,T]$. That the jumps do happen in time is a consequence of Lemma 4.4 of Barton et al. (2004), which we restate below as Lemma 2.1.

We also supply a new proof of this lemma to replace that given in Barton et al. (2004), which contains two errors (Etheridge 2005). One is that the variance $\sigma(W_s)$ appearing in the time change of the Wright-Fisher diffusion needs to be squared, so that the exponent α in the integral displayed in Eq. (16) of that paper is 2 rather than $1+\frac{1}{2(1-2\mu_2)}$. The second is that the divergence of this integral requires $\alpha \geq 2$ rather than $\alpha \geq 1$. Although this condition is (just barely) satisfied, we cannot deduce the divergence of the integral from the Engelbert-Schmidt 0-1 law (Karatzas and Shreve, 1991, Chapter 3, Proposition 6.27; see also Problem 1 of Ethier and Kurtz, 1986, Chapter 6) because this result applies to functionals of a Brownian path integrated for fixed periods of time rather than along sample paths which are stopped at a random time, as is the case in Eq. (16).

Lemma 2.1. *Let p_t be the Wright-Fisher diffusion corresponding to the generator A shown in (1). Then, for any real number $R < \infty$,*

$$\begin{aligned} \lim_{k\to\infty} \mathbf{P}_p\left\{\int_0^{\tau_k}\left(\frac{1}{p_s}\right)ds > R\right\} &= 1 \\ \lim_{k\to\infty} \mathbf{P}_p\left\{\int_0^{\tau_{k'}}\left(\frac{1}{1-p_s}\right)ds > R\right\} &= 1, \end{aligned}$$

where $\tau_k = \inf\{t>0 : p_t = 1/k\}$ and $\tau_{k'} = \inf\{t>0 : p_t = 1-1/k\}$.

Proof. For each positive integer k choose $\phi_k \in \mathcal{C}^{(2)}([0,1])$ such that $\phi_k(p) = -\ln(p)$ on $[1/k,1]$ and observe that on this restricted set,

$$A\phi(p) = \frac{1}{2p} - \frac{1}{2} - \frac{b(p)}{p},$$

where $b(p) = \mu_1(1-p) - \mu_2 p + \sigma(p)p(1-p)$ is the infinitesimal drift coefficient in A. Then, for each $k > p_0^{-1}$, the stopped process

$$\begin{aligned} M_{t\wedge\tau_k} &= \phi_k(p_{t\wedge\tau_k}) - \phi_k(p_0) - \int_0^{t\wedge\tau_k} A\phi_k(p_s)ds \\ &= -\ln(p_{t\wedge\tau_k}) + \ln(p_0) - \frac{1}{2}\int_0^{t\wedge\tau_k} \frac{1-2b(p_s)}{p_s}ds + \frac{1}{2}\left(t\wedge\tau_k\right) \end{aligned}$$

is a continuous martingale with quadratic variation

$$\langle M\rangle_{t\wedge\tau_k} = \int_0^{t\wedge\tau_k} p_s(1-p_s)(\phi_k'(p_s))^2 ds = \int_0^{t\wedge\tau_k} \frac{ds}{p_s} - (t\wedge\tau_k).$$

In particular, on the set $\{\tau_k < \infty\}$, we have

$$\begin{aligned} M_{\tau_k} &= \ln(k) + \ln(p_0) - \frac{1}{2}\int_0^{\tau_k} \frac{1-2b(p_s)}{p_s}ds - \frac{1}{2}\tau_k \\ \langle M\rangle_{\tau_k} &= \int_0^{\tau_k} \frac{ds}{p_s} - \tau_k, \end{aligned}$$

which in turn implies that, for any $R < \infty$, the following three inequalities

$$\begin{aligned} \tau_k &< R \\ \langle M\rangle_{\tau_k} &< R \\ M_{\tau_k} &> \ln(k) + \ln(p_0) - \left(\frac{1}{2} + ||b||_\infty\right) R \end{aligned}$$

are satisfied on the set

$$\Omega_{R,k} = \left\{\int_0^{\tau_k} \frac{ds}{p_s} < R\right\}.$$

Now, because $M_{\cdot\wedge\tau_k}$ is a continuous, one-dimensional martingale, there is an enlargement Ω' of the probability space Ω on which the diffusion p_t is defined and there is also a standard one-dimensional Brownian motion B_t, defined on Ω', such that

$$M_{t\wedge\tau_k} = B_{\langle M\rangle_{t\wedge\tau_k}}.$$

[See Karatzas and Shreve (1991), Chapter 3, Theorem 4.6 and Problem 4.7.] Thus, in view of the conditions holding on $\Omega_{R,k}$, we obtain the following bound

$$\mathbf{P}\{\Omega_{R,k}\} \leq \mathbf{P}\left\{\sup_{t\leq R} B_t > \ln(k) + \ln(p_0) - CR\right\},$$

where $C = \frac{1}{2} + ||b||_\infty$ is independent of k. The first half of the proposition then follows from the fact that the probability on the right-hand side of the preceding inequality goes to 0 as $k \to \infty$ with R fixed. The second half can be proved using a similar argument, with $\phi_k(p) = -\ln(1-p)$ on $[0, 1-1/k]$. □

With Lemma 2.1 established, the next proposition is a special case of the existence and uniqueness results for structured coalescents proved in Barton et al. (2004).

Proposition 2.2. *For any $\nu \in \mathcal{P}(E)$, there exists a Markov process $(\tilde{z}_t, \tilde{p}_t)$, which we call the structured retrospective process, which is the unique solution to the $D_E[0,\infty)$-martingale problem for $(\tilde{G}, \nu)$.*

Proof. Because the operator $\tilde{G}$ is a Feller generator when restricted to twice continuously differentiable functions on each of the sets $E_k = \big(\{1\} \times [k^{-1}, 1]\big) \cup \big(\{2\} \times [0, 1 - k^{-1}]\big)$, we can show that a stopped version of the process exists on each of these sets and that this process is the unique solution of the corresponding stopped martingale problem. Then, using the Lemma 2.1 and noting that the diffusions p_t and $\tilde{p}_t$ are identical in distribution, we can show that the sequence of hitting times of the boundaries of the sets E_k is almost surely unbounded as $k \to \infty$. Consequently, Theorem 4.2 and Proposition 4.3 of Barton et al. (2004) imply the existence of a Markov process $(\tilde{z}_t, \tilde{p}_t)$ defined on all of E which is the unique solution to the $D_E[0,\infty)$-martingale problem for $\tilde{G}$. □

Of course, as the name indicates, the process $(\tilde{z}_t, \tilde{p}_t)$ describes the retrospective behavior of a lineage sampled at random from the population rather than forward-in-time evolution of the common ancestor of the entire population. However, because Kingman's coalescent comes down from infinity (Kingman 1982), we know that, with probability one, all extant lineages, including that ancestral to the sampled individual, will coalesce with the common ancestor within some finite time. That this is still true when we incorporate genetic structure into the coalescent is evident from the fact that the coalescent rates within a background are accelerated by the reciprocal of the frequency of that background; see Eq. (3). Furthermore, because lineages move between genetic backgrounds at rates which are bounded below by the (positive) mutation rates, lineages cannot be permanently trapped in different backgrounds.

These observations lead to the following strategy for identifying the common ancestor process in a stationary population. First, because the asymptotic properties of the retrospective process in the deep evolutionary past coincide with those of the common ancestor process itself, any stationary distribution of $\tilde{G}$ will also be a stationary distribution of the common ancestor process. Indeed, we will call this distribution (assuming uniqueness) the common ancestor distribution. Secondly, given such a distribution, it is clear that we can construct a stationary version of the retrospective process $(\tilde{z}_t, \tilde{p}_t)$ which is defined for all times $t \in \mathcal{R}$. However, because this lineage persists indefinitely, it is necessarily the common ancestor lineage for the whole population. Accordingly, we can characterize the joint law of the stationary process of substitutions to the common ancestor and the forward-in-time evolution of the allele frequencies by determining the law of the time reversal of the retrospective process with respect to its stationary distribution. (Observe that by time reversing the retrospective process, which runs from the present to the past, we obtain a process which runs from the past to the present.)

2.1 The common ancestor distribution

In this section we show that the common ancestor distribution, which we denote $\pi(z, dp)$, can be found by solving a simple boundary value problem. We begin by observing that because $\mathcal{D}(\tilde{G}) = \mathcal{C}_c^2(E)$ is an algebra which is dense in $\hat{C}(E)$ and because the martingale problem for $\tilde{G}$ is well-posed, any distribution $\pi(z, dp)$ which satisfies the condition,

$$\int_0^1 \tilde{G}\phi(1,p)\,\pi(1,dp) + \int_0^1 \tilde{G}\phi(2,p)\,\pi(2,dp) = 0 \tag{5}$$

for all $\phi \in \mathcal{D}(\tilde{G})$, is a stationary distribution for $\tilde{G}$ [Ethier and Kurtz (1986), Chapter 4, Proposition 9.17]. Assuming that we can write $\pi(z, dp) = \pi(z, p)dp$ for $z = 1, 2$, where $\pi(z, \cdot) \in \mathcal{C}^2((0, 1))$, integration-by-parts shows that this condition will be satisfied if

$$\begin{aligned} A^*\pi(1,p) + \mu_2\left(\frac{p}{1-p}\right)\pi(2,p) - \mu_1\left(\frac{1-p}{p}\right)\pi(1,p) &= 0 \\ A^*\pi(2,p) + \mu_1\left(\frac{1-p}{p}\right)\pi(1,p) - \mu_2\left(\frac{p}{1-p}\right)\pi(2,p) &= 0. \end{aligned} \tag{6}$$

Here A^* is the formal adjoint of A with respect to Lebesgue measure on $[0, 1]$ and is defined by the formula

$$A^*\phi(p) = \frac{1}{2}(p(1-p)\phi(p))'' - \big((\mu_1(1-p) - \mu_2 p + \sigma(p)p(1-p))\,\phi(p)\big)'. \tag{7}$$

Because the marginal distribution over $z \in \{1, 2\}$ of the stationary measure $\pi(z, dp)$ is just the stationary measure $\pi(p)dp$ of the diffusion process itself, it is convenient to write $\pi(z, dp)$ in the form

$$\begin{aligned} \pi(1,dp) &= \pi(1,p)dp = h(p)\pi(p)dp \\ \pi(2,dp) &= \pi(2,p)dp = (1-h(p))\pi(p)dp, \end{aligned} \tag{8}$$

where $h(p)$ is the conditional probability that the common ancestor is of type P given that the frequency of P in the population is p. Substituting this expression into (6) leads to the following boundary value problem (BVP) for $h(p)$,

$$\begin{aligned} &Ah(p) - \left(\mu_1\left(\frac{1-p}{p}\right) + \mu_2\left(\frac{p}{1-p}\right)\right)h(p) = -\mu_2\left(\frac{p}{1-p}\right), \\ &h(0) = 0, h(1) = 1. \end{aligned} \tag{9}$$

We show below that the smoothness of the selection coefficient $\sigma(p)$ is sufficient to guarantee the existence of a solution $h(p)$ to (9) which is smooth in $(0, 1)$ and which has a derivative $h'(p)$ that can be continuously extended to $[0, 1]$, and that this implies that the common ancestor distribution can always be represented in the form (8), with $h(p)$ the unique solution to (9). However, we first make two observations concerning equation (9) itself. First, if $\sigma(p) \equiv 0$, i.e., P and Q are selectively neutral, then $h(p) = p$ solves (9) and the distribution of the common ancestor is the same as that of an individual sampled randomly from the population. Of course, this claim can also be deduced directly from the look-down formulation of Donnelly and Kurtz (1996): under neutrality, the common ancestor is the individual occupying the lowest level and, by exchangeability, the distribution of the type of this individual is given by the empirical measure carried by all of the particles, which is just $p\delta_1 + (1 - p)\delta_0$ for a biallelic model.

Secondly, if we write $h(p) = p + \psi(p)$, then a simple calculation shows that $h(p)$ will satisfy the BVP (9) if and only if $\psi(p)$ satisfies the following BVP:

$$\begin{aligned} &A\psi(p) - \left(\mu_1\left(\frac{1-p}{p}\right) + \mu_2\left(\frac{p}{1-p}\right)\right)\psi(p) = -\sigma(p)p(1-p), \\ &\psi(0) = \psi(1) = 0. \end{aligned} \tag{10}$$

This result is useful when numerically calculating $h(p)$ because it replaces the divergent inhomogeneous term on the right-hand side of (9) with a term which is smooth on $[0,1]$. Even so, because the inhomogeneous equation is singular at $p = 0, 1$, the usual shooting method (Press et al. (1992)) used to solve such two-point BVP's must be modified as we discuss briefly in the appendix. More importantly, we can use the BVP (10) to prove the existence and regularity of the conditional probability $h(p)$.

Lemma 2.3. *Suppose that A is the generator of a Wright-Fisher diffusion as in (1). Then there exists a function $\psi(p)$ satisfying the BVP (10) which is holomorphic on* $(0,1)$ *and its first derivative $\psi'(p)$ can be continuously extended to* $[0,1]$. *Furthermore, the function $h(p) = p+\psi(p)$ is the unique solution to the BVP (9) sharing these regularity properties.*

Proof. We begin by noting that $p = 0$ and $p = 1$ are regular singular points for the corresponding homogeneous equation and that the indicial equations have roots $\lambda = 1, -2\mu_1$ at $p = 0$ and $\lambda = 1, -2\mu_2$ at $p = 1$. Because the coefficients are smooth in $(0,1)$, Theorems 7 and 8 of Chapter 9 of Birkhoff and Rota (1989) can be used to deduce the existence of four functions, $u_{0,1}(\cdot)$ and $u_{0,2}(\cdot)$, analytic in a neighborhood of $p = 0$, and $u_{1,1}(\cdot)$, and $u_{1,2}(\cdot)$, analytic in a neighborhood of $p = 1$, as well as two constants C_0 and C_1 such that the following two pairs of functions,

$$\psi_{0,1}(p) = pu_{0,1}(p) \quad \text{and} \quad \psi_{0,2}(p) = p^{-2\mu_1}u_{0,2}(p) + C_0 pu_{0,1}(p)\ln(p)$$

and

$$\psi_{1,1}(p) = (1-p)u_{1,1}(p) \quad \text{and} \quad \psi_{1,2}(p) = (1-p)^{-2\mu_2}u_{1,2}(p) + C_1(1-p)u_{1,1}(p)\ln(1-p),$$

each constitutes a set of linearly independent solutions to the homogeneous equation. Furthermore, because the diffusion operator A is uniformly elliptic on any interval $(\epsilon, 1-\epsilon)$ for any $\epsilon > 0$, these solutions can be analytically continued to $(0,1)$.

Consequently, by taking suitable linear combinations of the $\psi_{ij}(\cdot)$, we can construct a pair of linearly independent solutions, $\psi_0(\cdot)$ and $\psi_1(\cdot)$, analytic on $(0,1)$, such that for any $\epsilon > 0$,

$$\psi_0(0) = 0,\ \psi_0'(0) = 1,\ \lim_{p\to 1}(1-p)^{2\mu_2+\epsilon}\psi_0(p) = 0,\ \lim_{p\to 1}(1-p)^{2\mu_2+1+\epsilon}\psi_0'(p) = 0$$

$$\psi_1(1) = 0,\ \psi_1'(1) = -1,\ \lim_{p\to 0}p^{2\mu_1+\epsilon}\psi_1(p) = 0,\ \lim_{p\to 0}p^{2\mu_1+1+\epsilon}\psi_1'(p) = 0.$$

A solution to the inhomogeneous equation can then be obtained by the method of variation of parameters, which gives:

$$\psi(p) = \psi_0(p)\int_p^1 \left(\frac{\sigma(q)q(1-q)}{W(q)}\right)\psi_1(q)dq + \psi_1(p)\int_0^p \left(\frac{\sigma(q)q(1-q)}{W(q)}\right)\psi_0(q)dq,$$

where $W(p)$ is the Wronskian of the homogeneous equation

$$W(p) = \exp\left[-2\int_{p_0}^p \frac{\mu_1(1-q) - \mu_2 q + \sigma(q)q(1-q)}{q(1-q)}dq\right],$$

and p_0 is an arbitrary point in $(0,1)$. Furthermore, in light of the boundary behavior of the functions $\psi_1(\cdot)$ and $\psi_0(\cdot)$, it is easy to check that $\psi(\cdot)$ is smooth in $(0,1)$, that $\psi(0) = \psi(1) = 0$, and that the limits

$$\lim_{p\to 0}\psi'(p) = \int_0^1 \left(\frac{\sigma(q)q(1-q)}{W(q)}\right)\psi_1(q)dq \quad \text{and} \quad \lim_{p\to 1}\psi'(p) = -\int_0^1 \left(\frac{\sigma(q)q(1-q)}{W(q)}\right)\psi_0(q)dq$$

exist and are finite. Clearly, these statements also hold for $h(p) = p + \psi(p)$ and a simple calculation verifies that $h(p)$ solves the BVP (9). □

By combining this result with the formal calculations leading from Eq. (5) to (9), as well as Proposition 9.17 of Chapter 4 of Ethier and Kurtz (1986), we can deduce the existence of a stationary distribution for $\tilde{G}$. Our next proposition asserts that this distribution is also unique.

Proposition 2.4. *The retrospective process $(\tilde{z}_t, \tilde{p}_t)$ has a unique stationary distribution $\pi(z, dp)$ of the form (8), where $\pi(p)$ is the density (2) of the stationary distribution for the Wright-Fisher diffusion generated by A and $h(p)$ is the unique solution to the BVP (9).*

Proof. Since we have already demonstrated the existence of a stationary distribution corresponding to Eq. (8)-(9), we need only show that this measure is unique. To do so, we will prove that $\tilde{G}$ is strongly connected (see Donnelly and Kurtz 1999, Section 9): if $P_\nu(t)$ denotes the one-dimensional distribution at time t of the solution to the martingale problem for $(\tilde{G}, \nu)$, then for any pair of distributions $\nu_1, \nu_2 \in \mathcal{P}(E)$ and all times $T > 0$, $\mathcal{P}_{\eta_1}(T)$ and $\mathcal{P}_{\eta_2}(T)$ are not mutually singular. Uniqueness of the stationary distribution will then follow from Lemma 5.3 of Ethier and Kurtz (1993), which implies that if the embedded Markov chain, $((\tilde{z}_{nT}, \tilde{p}_{nT}) : n \geq 1)$, has two distinct stationary distributions (as it will if the continuous time process has two distinct stationary distributions), then it also has two mutually singular stationary distributions.

Let $(\tilde{z}_t^{(1)}, \tilde{p}_t^{(1)})$ and $(\tilde{z}_t^{(2)}, \tilde{p}_t^{(2)})$ be solutions to the $D_E[0, \infty)$-martingale problem for $\tilde{G}$ with initial distributions ν_1 and ν_2, respectively. Because the marginal processes $\tilde{p}_t^{(1)}$ and $\tilde{p}_t^{(2)}$ are Wright-Fisher diffusions corresponding to A, the positivity of the mutation rates μ_1, μ_2 implies that for any $t > 0$, the one-dimensional distributions of $\tilde{p}_t^{(1)}$ and $\tilde{p}_t^{(2)}$ are mutually absolutely continuous with respect to Lebesgue measure on $[0, 1]$. (In particular, these distributions do not have atoms at 0 or 1.) Furthermore, for every $\delta \in (0, 1/2)$ and every $T > 0$, there exists an $\epsilon > 0$ such that the probabilities $\mathbf{P}\{\tilde{p}_t^{(i)} \in [\delta, 1 - \delta] \text{ for all } t \in [T/2, T]\} > \epsilon$ for $i = 1, 2$. Combining this observation with the fact that for fixed $\delta \in (0, 1/2)$, the jump rates of the component $\tilde{z}_t$ are uniformly bounded above 0 and below ∞ whenever the frequency process $\tilde{p}_t$ is in $[\delta, 1 - \delta]$, it follows that $P_{\nu_1}(T)$ and $P_{\nu_2}(T)$ are each mutually absolutely continuous with respect to the product measure $(\delta_1(dz) + \delta_2(dz)) \times m(dp)$ on E, where $m(dp)$ is Lebesgue measure restricted to $(0, 1)$. Since this implies that $P_{\nu_1}(T)$ and $P_{\nu_2}(T)$ are mutually absolutely continuous with respect to one another for every $T > 0$, the proposition follows. □

We can also rewrite the inhomogeneous differential equation in (9) in a form which leads to an alternative probabilistic representation of $h(p)$. Because $h(0) = 0$ and $h(1) = 1$, the solution $h(p)$ to the BVP (9) is a harmonic function for the operator $\hat{A}$, defined as

$$\hat{A}\phi(p) = A\phi(p) + \mu_1 \left(\frac{1-p}{p}\right) (\phi(0) - \phi(p)) + \mu_2 \left(\frac{p}{1-p}\right) (\phi(1) - \phi(p)). \tag{11}$$

Setting

$$\mathcal{D}(\hat{A}) = \{\phi \in \mathcal{C}^2([0,1]) : \lim_{p \to 1} \hat{A}\phi(p) = \lim_{p \to 0} \hat{A}\phi(p) = 0\},$$

we see that $\hat{A}$ is the generator of a jump-diffusion process, $\hat{p}_t$, which diffuses in $(0, 1)$ according to the law of the Wright-Fisher diffusion corresponding to A until it jumps to one of the boundary points $\{0,1\}$ where it is absorbed. It follows from Lemma 2.1 that if the process does reach 0 or 1, then it is certain to have arrived there via a jump rather than by diffusing, even if that boundary is accessible to the pure diffusion process. Indeed, the existence of a unique Markov process $\hat{p}_t$ corresponding to $\hat{A}$ can be deduced from Lemma 2.1 in precisely the same way that the existence and uniqueness of the structured coalescent was obtained, although it is now essential that $\hat{p}_t$ be absorbed once it hits the boundary. Furthermore, because the total rate of jumps to either boundary point from any point in the interior is bounded below by $\mu_1 \wedge \mu_2$, the process is certain to jump to the boundary in finite time. Taken together these observations lead to the following representation for $h(p)$.

Proposition 2.5. *Let $\hat{p}(t)$ be the jump-diffusion process corresponding to the generator $\hat{A}$, and let $\tau = \inf\{t > 0 : \hat{p}_t = 0 \text{ or } 1\}$ be the time of the first (and only) jump to the boundary $\{0,1\}$. Then the solution $h(p)$ to the BVP (9) is the probability that $\hat{p}_t$ is absorbed at 1 when starting from initial value p:*

$$h(p) = \mathbf{P}_p\{\hat{p}_\tau = 1\}. \tag{12}$$

Proof. Because $h(p)$ is in $\mathcal{D}(\hat{A})$ and $\hat{A}h(p) = 0$ for all $p \in [0,1]$, it follows that $h(\hat{p}_t)$ is a bounded martingale with respect to the natural filtration of $\hat{p}_t$. Moreover, $E_p[\tau] < (\mu_1 \wedge \mu_2)^{-1} < \infty$, and so we can use the optional sampling theorem and the fact that $h(0) = 0$ and $h(1) = 1$ to calculate $h(p) = \mathbf{E}_p[h(\hat{p}_\tau)] = \mathbf{P}_p\{\hat{p}_\tau = 1\}$. □

Proposition 2.5 has several interesting consequences. First, by comparing the generator $\hat{A}$ shown in (11) with the generator of the structured coalescent (3) for a sample of size two with one P allele and one Q allele, it is evident that the type of the common ancestor has the same distribution as the type of the sampled lineage which is of the more ancient mutant origin. In other words, the quantity $h(p)$ is the probability that if we sample a P allele and a Q allele from a population in which P occurs at frequency p, then the Q allele has arisen from a mutation to an ancestral P individual more recently than the P allele in the sample has arisen from a mutation to an ancestral Q individual.

Secondly, because the rate at which $\hat{p}_t$ jumps to 1 is a strictly increasing function of p while the rate at which $\hat{p}_t$ jumps to 0 is strictly decreasing, (12) implies that $h(p)$ is a strictly increasing function of p. While we would expect such a relationship to hold if the selection coefficient $\sigma(p)$ is either non-decreasing or non-negative, it is noteworthy that $h(p)$ is increasing even with negative frequency-dependent selection, e.g., under balancing selection, with $\sigma(p) = s \cdot (p_0 - p)$ for $s > 0$ and $p_0 \in (0,1)$.

Another consequence of Proposition 2.5 is that the probability that the common ancestor is of a particular genotype is an increasing function of the fitness of that genotype. To make this precise, suppose that $A^{(1)}$ and $A^{(2)}$ are a pair of Wright-Fisher generators as in (1) with drift coefficients $b_i(p) = \mu_1(1-p) - \mu_2 p + \sigma(p)p(1-p)$ which differ only in their (smooth) fitness functions, $\sigma_1(p)$ and $\sigma_2(p)$, respectively, let $\hat{A}^{(i)}, i = 1,2$ be the generators of the jump-diffusion processes obtained by taking $A = A^{(i)}$ in (11), and let $h_1(p)$ and $h_2(p)$ be the corresponding conditional probabilities that the common ancestor is of type P.

Proposition 2.6. *If $\sigma_1(p) \leq \sigma_2(p)$ for all $p \in [0,1]$, then $h_1(p) \leq h_2(p)$ for all $p \in [0,1]$.*

Proof. In view of the smoothness of the coefficients of the diffusion generators $A^{(i)}, i = 1,2$, there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, a Brownian motion $(W_t, t \geq 0)$, and diffusion processes

$(p_t^{(i)}, t \geq 0), i = 1, 2$ corresponding to these generators such that the stochastic differential equation

$$p_t^{(i)} = p + \int_0^t b_i(p_s^{(i)})ds + \int_0^t \sqrt{p_s^{(i)}(1 - p_s^{(i)})}dW_s \qquad t \geq 0$$

is satisfied a.s. for $i = 1, 2$ and all $p \in [0, 1]$. (For example, such a coupling can be constructed using a sequence of coupled Markov chains which converge weakly to these diffusions.) Furthermore, because the drift coefficients satisfy the inequality $b_1(p) \leq b_2(p)$ for all $p \in [0, 1]$, while the infinitesimal variance satisfies the regularity condition $|\sqrt{p(1-p)} - \sqrt{q(1-q)}| < 2|p - q|^{1/2}$ for all $p, q \in [0, 1]$, we can use Lemma 3.4 in Shiga (1981) to conclude that

$$\mathbf{P}_p\{p_t^{(1)} \leq p_t^{(2)}, \forall t \geq 0\} = 1. \tag{13}$$

To relate this inequality to the jump-diffusion processes generated by the $\hat{A}^{(i)}, i = 1, 2$, observe that because each process jumps exactly once, we can construct coupled versions of these processes, denoted $\hat{p}_t^{(i)}, i = 1, 2$, by taking the motion of $\hat{p}_t^{(i)}$ in the interior $(0, 1)$ to be that of the coupled diffusions $p_t^{(i)}$ and arranging for $\hat{p}_t^{(i)}$ to jump to the boundary points 1 or 0, respectively, at the first time $\tau^{(i)}$ when

$$\mu_1 \int_0^{\tau^{(i)}-} \left(\frac{1 - p_s^{(i)}}{p_s^{(i)}}\right) ds \geq Z_1 \quad \text{or} \quad \mu_2 \int_0^{\tau^{(i)}-} \left(\frac{p_s^{(i)}}{1 - p_s^{(i)}}\right) ds \geq Z_2.$$

Here, Z_1 and Z_2 are unit mean exponential random variables which are independent of each other and of the diffusions $p_t^{(i)}$, but which are shared in common by the two jump-diffusions. (Note that the independence of Z_1 and Z_2, the continuity of the sample paths of the diffusions $p_t^{(i)}$, and the local boundedness of the jump rates in (11) guarantee that $\hat{p}_{\tau^{(i)}}^{(i)}$ is almost surely well-defined.)

Since the rate function governing jumps from $(0, 1)$ to 1 is an increasing function of $p \in [0, 1]$, while that governing jumps from $(0, 1)$ to 0 is a decreasing function of p, the inequality in (13) implies that, with probability one, if the process $\hat{p}_t^{(1)}$ jumps to 1, then the process $\hat{p}_t^{(2)}$ must also have jumped to 1, possibly at an earlier time. Consequently,

$$h_1(p) = \mathbf{P}\{\hat{p}_{\tau^{(1)}}^{(1)} = 1\} \leq \mathbf{P}\{\hat{p}_{\tau^{(2)}}^{(2)} = 1\} = h_2(p),$$

and the proposition follows upon noting that the initial condition $p \in [0, 1]$ is arbitrary. □

In particular, taking $\sigma_2(p) \geq \sigma_1(p) \equiv 0$, we can use Proposition 2.6 to conclude that $h_2(p) \geq h_1(p) = p$. Furthermore, if $\sigma_2(p)$ is strictly positive on $[0, 1]$, then the fact that the diffusions $p_t^{(1)}$ and $p_t^{(2)}$ have continuous sample paths which differ in distribution on every interval $[0, T], T > 0$ can be combined with (13) to deduce that with positive probability the set $\{t \in [0, T] : p_t^{(2)} > p_t^{(1)}\}$ has positive Lebesgue measure whenever the initial condition $p \in (0, 1)$, and therefore that $h_2(p) > h_1(p)$ for every $p \in (0, 1)$. In other words, if P is unconditionally more fit than Q, then the common ancestor will be more likely to be of type P than an individual sampled at random from the population, both on average and when conditioned on the frequency p at which P is segregating in the population. Furthermore, this property implies that the mean fitness of the common ancestor is greater than the mean fitness of an individual chosen at random from the population, and generalizes Theorem 2 of Fearnhead (2002) which applies when P has a fixed (frequency-independent) advantage over Q.

2.2 The common ancestor process

Having found the common ancestor distribution, our next task is to identify the common ancestor process, which we will do by determining the time-reversal of the retrospective process $(\tilde{z}_t, \tilde{p}_t)$ with respect to its stationary distribution. Because Proposition 2.4 asserts that this distribution is unique, the common ancestor process, at least in a stationary population, is also unique. We recall that time reversal preserves the Markov property, and that the generator G of the Markov process obtained by time reversal of the stationary process corresponding to $\tilde{G}$ has the property that it is adjoint to $\tilde{G}$ with respect to the measure $\pi(z, dp)$ (Nelson 1958), i.e.,

$$\sum_{z=1,2} \int_0^1 \big(\tilde{G}\phi(z,p)\big)\psi(z,p)\pi(z,p)dp = \sum_{z=1,2} \int_0^1 \phi(z,p)\big(G\psi(z,p)\big)\pi(z,p)dp, \tag{14}$$

for any $\psi \in \mathcal{D}(\tilde{G})$ and $\phi \in \mathcal{D}(G)$ (which is to be determined). A calculation making repeated use of the product rule and integration-by-parts, along with the characterization of the common ancestor distribution $\pi(z,p)dp$ provided by Proposition 2.4 and the fact that $A^*\pi(p) = 0$, with the density $\pi(p)$ given by (2) and the adjoint operator A^* given by (7), shows that this condition will be satisfied if

$$\begin{aligned} G\psi(1,p) &= A\psi(1,p) + p(1-p)\left(\frac{h'(p)}{h(p)}\right)\psi'(1,p) + \mu_2\left(\frac{p(1-h(p))}{(1-p)h(p)}\right)(\psi(2,p) - \psi(1,p)) \quad (15) \\ G\psi(2,p) &= A\psi(2,p) - p(1-p)\left(\frac{h'(p)}{1-h(p)}\right)\psi'(2,p) + \mu_1\left(\frac{(1-p)h(p)}{p(1-h(p))}\right)(\psi(1,p) - \psi(2,p)), \end{aligned}$$

with $\psi \in \mathcal{D}(G) \equiv \{\psi : \{1,2\} \times [0,1] \to \mathcal{R} \text{ such that } \psi(z,\cdot) \in \mathcal{C}^2([0,1]) \text{ for } z = 1,2\}$.

In the proof of the next proposition we show that the existence of a Markov process corresponding to G is a consequence of the continuity of $h(p)$ and $h'(p)$ (Lemma 2.3). Recall that the state space of the retrospective process is $E = (\{1\} \times (0,1]) \cup (\{2\} \times [0,1))$.

Proposition 2.7. *For any $\nu \in \mathcal{P}(\{1,2\} \times [0,1])$, there exists a Markov process (z_t, p_t), which we call the common ancestor process, which is the unique solution to the martingale problem for (G, ν). Furthermore, $(z_t, p_t) \in E$ for all $t > 0$.*

Proof. Since $h(0) = 0$ and $h(1) = 1$, the continuity of $h'(p)$ on $[0,1]$ implies the existence of constants $C_1 < C_2$ such that $C_1 p < h(p) < C_2 p$ and $C_1(1-p) < 1 - h(p) < C_2(1-p)$ for all $p \in [0,1]$. Consequently, all of the terms appearing on the right-hand side of (15) can be continuously extended (as functions of p) to $[0,1]$, and we define $G\psi(z,p)$ accordingly if $(z,p) = (1,0)$ or $(z,p) = (2,1)$. In particular, the operators $A_1\psi(p) = A\psi(p) + p(1-p)(h'(p)/h(p))\psi'(p)$ and $A_2\psi(p) = A\psi(p) - p(1-p)(h'(p)/(1-h(p)))\psi'(p)$, with $\psi \in \mathcal{C}^2([0,1])$, are the generators of a pair of diffusion processes on $[0,1]$ (Ethier and Kurtz 1986, Chapter 8, Theorem 1.1), and so there exists a diffusion process on the space $\{1,2\} \times [0,1]$ corresponding to the generator $\mathcal{A}\psi(z,p) \equiv A_z\psi(z,p)$ for $\psi \in \mathcal{D}(G)$. The existence of a process (z_t, p_t) corresponding to G then follows from Theorem 7.1 of Ethier and Kurtz (1986, Chapter 1) and the fact that G is a bounded perturbation of $\mathcal{A}$, i.e., the jump rates are bounded. Furthermore, the uniqueness of solutions to the martingale problem for (G, ν) is then a consequence of Theorem 4.1 of Ethier and Kurtz (1986, Chapter 4).

To prove that $(z_t, p_t) \in E$ for all $t > 0$, we observe that boundary points inconsistent with the type of the common ancestor are entrance boundaries for the frequency component of the common ancestor process. For example, if the type of the common ancestor is P, i.e., if $z = 1$, then because $h(p)$ is continuously differentiable on $[0,1]$ (Lemma 2.3) and because $p/h(p) \approx 1/h'(0)$ when $p \approx 0$ we can write the drift coefficient of A_1 as

$$b_1(p) \equiv b(p) + p(1-p)\left(\frac{h'(p)}{h(p)}\right) = (\mu_1 + 1) + O(p),$$

where $b(p)$ is the drift coefficient of A. That $p = 0$ is an entrance boundary for the diffusion corresponding to A_1 can then be shown using Feller's boundary classification (Ewens 2004, Section 4.7) and the fact that $\mu_1 + 1 > 1/2$. Similar remarks apply to the boundary $p = 1$ and the diffusion corresponding to A_2. □

If A is the generator of a neutral Wright-Fisher diffusion (i.e., $\sigma(p) \equiv 0$), then $h(p) = p$ and the generator of the common ancestor process is just

$$\begin{aligned} G\psi(1,p) &= A\psi(1,p) + (1-p)\psi'(1,p) + \mu_2(\psi(2,p) - \psi(1,p)) \\ G\psi(2,p) &= A\psi(2,p) - p\psi'(2,p) + \mu_1(\psi(1,p) - \psi(2,p)). \end{aligned}$$

As expected, the process governing the change of type of the common ancestor decouples from the frequency process and coincides with the mutation process itself. The only novel feature of the neutral common ancestor process is the presence of the additional drift terms in the diffusion which reflect the fact that because the common ancestor contributes more offspring to the population than an individual chosen at random, the population has a tendency to evolve towards the type of the common ancestor. Indeed, these extra births can be made explicit by formulating a finite population model $(z^{(N)}, p^{(N)})$ which combines the usual Moran resampling with a neutral look-down process that operates only on the lowest level (i.e., birth-death events involving the lowest level always assign the birth to the lowest level, but all other birth-death events are resolved by choosing the parent at random from the two participating individuals). It is then straightforward to show that as $N \to \infty$, suitably rescaled versions of $(z^{(N)}, p^{(N)})$ converge weakly to the jump-diffusion process generated by G.

When there are fitness differences between the two alleles, in general $h(p) \neq p$ and the substitution rates to the common ancestor depend on the allele frequency p, i.e., the substitution process to the common ancestor z_t is not a Markov process. In this case, the substitution rates will differ from the corresponding mutation rates for most values of p. Moreover, because Proposition 2.6 shows that $h(p) > p$ for all $p \in (0,1)$ whenever P is unconditionally more fit than Q (i.e., $\sigma(p) > 0$ for all $p \in [0,1]$), Eq. (15) shows that the rate of substitutions from the less fit allele to the more fit allele is greater than the corresponding mutation rate, and *vice versa*. A less intuitive property of the generator of the common ancestor process is that for each value of p the geometric mean of the two substitution rates is the same as that of the two neutral mutation rates. While it is unclear what biological interpretation this invariant might have, one mathematical consequence is that for each fixed value of p only one of the two substitution rates can exceed the corresponding neutral mutation rate, while the other is necessarily less than it. However, the direction of these two inequalities may differ according to the frequency p if selection is frequency-dependent or fluctuates in time.

3 Weak mutation limits

Because single nucleotide mutation rates in DNA sequences are typically on the order of 10^{-8} mutations per site per generation, while most effective population size estimates are less than 10^7 (Lynch and Connery 2003), the asymptotic properties of the common ancestor process in the limit of vanishing mutation rates are of special interest. (Here we temporarily relax our earlier assumption that the mutation rates are positive.) We first observe that if μ_1 and μ_2 are both zero, then the BVP (9) simplifies to the equation $Ah(p) = 0$, with $h(0) = 0$ and $h(1) = 1$, and the solution,

$$h_0(p) = \frac{\int_0^p e^{-2S(q)}dq}{\int_0^1 e^{-2S(q)}dq}, \tag{16}$$

is just the fixation probability of P when its initial frequency is p (Ewens 2004, Section 4.3). Furthermore, if we substitute this expression into the generator of the common ancestor process (15), then because both jump rates vanish, we are left with a pair of operators,

$$\begin{aligned} G\psi(1,p) &= A\psi(1,p) + p(1-p)\left(\frac{h_0'(p)}{h_0(p)}\right)\psi'(1,p) \\ G\psi(2,p) &= A\psi(2,p) - p(1-p)\left(\frac{h_0'(p)}{1-h_0(p)}\right)\psi'(2,p), \end{aligned} \tag{17}$$

which we recognize to be the generators of the diffusion process corresponding to A conditioned to absorb either at 1 (top line) or at 0 (lower line) (Ewens 2004, Section 4.6). That the limiting generator takes this form reflects the fact that in the absence of mutation, any population which is descended from the common ancestor will also be fixed for the type of that individual.

A more useful observation is that if the mutation rates are small enough that mutations occur rarely on the coalescent time scale, then we can approximate the non-Markovian substitution process to the common ancestor by a continuous time two state Markov chain. Although approximate, such a process would greatly simplify the numerical or Monte Carlo computations needed to infer selection coefficients and other model parameters from a set of DNA sequences. One possibility is to define the transition rates of the Markov chain to be equal to the mean substitution rates obtained by averaging the frequency-dependent substitution rates of the bivariate process (15) over the conditional distribution of the allele frequencies given the type of the common ancestor,

$$\begin{aligned} \mu_2^{CA} &\equiv \mu_2 \int_0^1 \left(\frac{p(1-h(p))}{(1-p)h(p)}\right) h(p)\pi(p)dp \Big/ \int_0^1 h(p)\pi(p)dp \\ \mu_1^{CA} &\equiv \mu_1 \int_0^1 \left(\frac{(1-p)h(p)}{p(1-h(p))}\right)(1-h(p))\pi(p)dp \Big/ \int_0^1 (1-h(p))\pi(p)dp, \end{aligned} \tag{18}$$

e.g., μ_2^{CA} is the mean substitution rate to the common ancestor given that the type of that individual is P (which mutates to Q). Indeed, the ergodic properties of Wright-Fisher diffusions (Norman 1977) offer some justification for this approximation. Provided that the mutation rates are sufficiently small, the time elapsed between successive substitutions to the common ancestor will with very high probability be large enough for the allele frequencies to have relaxed to their stationary distribution well in advance of the next mutation. Moreover, because Lemma 2.3 guarantees that the jump rates appearing in the generator G in (15) are continuous functions of the frequency p, the time averages of the jump rates along paths of the diffusion will be approximately equal to the product of the time elapsed and the mean substitution rates shown above. Of course, for this approximation to be relevant to data, we will also need the phylogenetic tree describing the relationships among the sequences to be deep enough that the ergodic averages of the substitution rates are approached along each branch of the tree.

When the mutation rates are very small, the average substitution rates shown in (18) can be replaced by simpler expressions which depend only on the mutation rates and the fixation probabilities. Suppose that $\mu_i = \theta\nu_i$, $i = 1, 2$, and write $h_\theta(p)$, $\pi_\theta(p)$, and $\mu^{CA}_{i,\theta}$ to indicate the dependence of these quantities on θ. In Proposition 3.2 we evaluate the scaled, weak mutation limits $\mu^{CA}_{i,low} \equiv \lim_{\theta\to 0} \theta^{-1}\mu^{CA}_{i,\theta}$. However, we first state a technical lemma which will be needed in the proof of that proposition. Recall that we assume that the selection coefficient $\sigma(p)$ is holomorphic on some neighborhood of $[0, 1]$.

Lemma 3.1. *The functions $h_\theta(p)$ and $h'_\theta(p)$ converge uniformly on $[0, 1]$ to $h_0(p)$ and $h'_0(p)$, respectively, as $\theta \to 0$.*

Proof. We begin by using the probabilistic representation of $h_\theta(p)$ given in Proposition 2.5 to prove that $h_\theta(p)$ converges pointwise on $[0, 1]$ to $h_0(p)$. For each $\theta \geq 0$, let $p_\theta(t)$ be the diffusion process corresponding to the generator $A_\theta\phi(p) = \frac{1}{2}p(1-p)\phi''(p) + (\theta\nu_1(1-p) - \theta\nu_2 p + \sigma(p)p(1-p))\,\phi'(p)$, and if $\theta > 0$, let $\hat{p}_\theta(t)$ be the jump-diffusion process corresponding to the generator

$$\hat{A}_\theta\phi(p) = A_\theta\phi(p) + \theta\nu_1\left(\frac{1-p}{p}\right)(\phi(0) - \phi(p)) + \theta\nu_2\left(\frac{p}{1-p}\right)(\phi(1) - \phi(p)).$$

As in the proof of Proposition 2.6, we can construct coupled versions of the two processes $p_\theta(t)$ and $\hat{p}_\theta(t)$ in the following way. Let $Z_{0,\theta}$ and $Z_{1,\theta}$ be a pair of independent, unit mean exponentially distributed random variables, define

$$J_{0,\theta}(t) = \theta\nu_1\int_0^t \left(\frac{1-p_\theta(s)}{p_\theta(s)}\right) ds, \quad \text{and} \quad J_{1,\theta}(t) = \theta\nu_2\int_0^t \left(\frac{p_\theta(s)}{1-p_\theta(s)}\right) ds,$$

and let $\tau_\theta = \inf\{t > 0 : J_{0,\theta}(t) \geq Z_{0,\theta} \text{ or } J_{1,\theta}(t) \geq Z_{1,\theta}\}$. For $t < \tau_\theta$, define $\hat{p}_\theta(t) = p_\theta(t)$, while for $t \geq \tau_\theta$, define $\hat{p}_\theta(t) = 0$ if $J_{0,\theta}(\tau_\theta) \geq Z_0$, and $\hat{p}_\theta(t) = 1$ otherwise. Then, according to Proposition 2.5, $h_\theta(p) = \mathbf{P}\{\hat{p}_\theta(\tau_\theta) = 1\}$. Morally, we expect the probabilities $h_\theta(p)$ to converge to the fixation probability $h_0(p)$ as $\theta \to 0$ because the diffusions $p_\theta(t)$ converge in distribution to $p_0(t)$, and because when θ is small but positive the process $\hat{p}_\theta(t)$ is likely to jump to the boundary at 1 only along those sample paths of $p_\theta(t)$ which hit 1 before they hit 0.

To make this argument precise, let us introduce the hitting times $T_{q,\theta} = \inf\{t > 0 : p_\theta(t) = q\}$ for $q \in [0, 1]$, with $T_{q,\theta} = \infty$ if $p_\theta(t) \neq q$ for all $t > 0$, and recall that $\mathbf{P}_p\{T_{b,\theta} < T_{a,\theta}\} = (s_\theta(p) - s_\theta(a))/(s_\theta(b) - s_\theta(a))$ for any $0 < a < b < 1$, where the scale function $s_\theta(p)$ for the Wright-Fisher diffusion $p_\theta(t)$ is

$$s_\theta(p) = \int_c^p \left(\frac{q}{c}\right)^{-2\theta\nu_1}\left(\frac{1-q}{1-c}\right)^{-2\theta\nu_2} e^{-2(S(q)-S(c))}dq,$$

with $S(p) \equiv \int_0^p \sigma(q)dq$ and c some arbitrary point in $(0, 1)$ (Ewens 2004, Section 4.3). Furthermore, if $2\theta\nu_1$ and $2\theta\nu_2$ are both less than 1, then the scale function is finite on $[0, 1]$ and we can also allow $a = 0$ and $b = 1$ in the previous expression for the hitting probability. Consequently, for every $p \in [0,1]$, $s_\theta(p)$ converges pointwise (in fact, uniformly on $[0,1]$) to $s_0(p)$ as $\theta \to 0$, and for any fixed $0 \le a < b \le 1$, the probabilities $\mathbf{P}_p\{T_{b,\theta} < T_{a,\theta}\}$ converge to $\mathbf{P}_p\{T_{b,0} < T_{a,0}\}$. In particular, if we define $u_\theta(p) = \mathbf{P}_p\{T_{1,\theta} < T_{0,\theta}\}$ to be the probability that the diffusion $p_\theta(t)$ hits 1 before hitting 0, then $u_\theta(p)$ converges uniformly to $h_0(p)$ on $[0,1]$. We also observe that if we let $T_\theta \equiv T_{0,\theta} \wedge T_{1,\theta}$ denote the first hitting time of 0 or 1 by the diffusion $p_\theta(t)$, then the expectation $\mathbf{E}_p[T_\theta]$ is finite whenever $2\theta\nu_1$ and $2\theta\nu_2$ are both less than 1, in which case

$$\begin{aligned}\mathbf{E}_p[T_\theta] \;=\;& 2(1-u_\theta(p))\int_0^p q^{2\theta\nu_1-1}(1-q)^{2\theta\nu_2-1}e^{2S(q)}dq\int_0^q z^{-2\theta\nu_1}(1-z)^{-2\theta\nu_2}e^{-2S(z)}dz\;+\\ & 2u_\theta(p)\int_p^1 q^{2\theta\nu_1-1}(1-q)^{2\theta\nu_2-1}e^{2S(q)}dq\int_q^1 z^{-2\theta\nu_1}(1-z)^{-2\theta\nu_2}e^{-2S(z)}dz,\end{aligned}$$

(Ewens 2004, Section 4.4). A simple calculation shows that $\mathbf{E}_p[T_\theta]$ converges to $\mathbf{E}_p[T_0]$ for every $p \in [0,1]$ as $\theta \to 0$.

Now fix $p \in [0,1]$, let $\epsilon > 0$, and use the continuity of the sample paths of $p_0(t)$ to choose $\delta > 0$ small enough that

$$\mathbf{P}_p\{T_{1,0} < T_{\delta,0}\} > \mathbf{P}_p\{T_{1,0} < T_{0,0}\} - \epsilon = h_0(p) - \epsilon.$$

Because the scale functions $s_\theta(p)$ converge to $s_0(p)$ as $\theta \to 0$, we can choose $\theta_1 > 0$ such that

$$\mathbf{P}_p\{T_{1,\theta} < T_{\delta,\theta}\} > \mathbf{P}_p\{T_{1,0} < T_{\delta,0}\} - \epsilon,$$

for all $\theta \in [0,\theta_1]$. Furthermore, by combining Markov's inequality with the convergence of $\mathbf{E}_p[T_\theta]$ to $\mathbf{E}_p[T_0]$, we can also choose $\theta_2 > 0$ and $T < \infty$ such that

$$\mathbf{P}_p(T_\theta < T) > 1 - \epsilon,$$

for all $\theta \in [0,\theta_2]$. Consequently, if we let

$$\Omega_{T,\delta,\theta} \equiv \{T_{1,\theta} < (T_{\delta,\theta} \wedge T)\}$$

denote the event that the diffusion $p_\theta(t)$ hits 1 before time T without first hitting δ and we set $\theta_3 = \theta_1 \wedge \theta_2$, then

$$\mathbf{P}_p(\Omega_{T,\delta,\theta}) > h_0(p) - 3\epsilon,$$

for all $\theta \in [0,\theta_3]$.

To relate these observations to the behavior of the jump-diffusions $\hat{p}_\theta(t)$, note that Lemma 2.1 implies that on $\Omega_{T,\delta,\theta}$,

$$\lim_{t\to T} J_{1,\theta}(t) = \infty,$$

almost surely, and therefore $\hat{p}_\theta(T) = 1$ whenever the sample path of $p_\theta(t)$ is in this set, unless $\hat{p}_\theta(t)$ jumps to 0 before the diffusion $p_\theta(t)$ hits 1. However, because a jump to 0 can only occur if $J_{0,\theta}(T) > Z_0$ and because

$$J_{0,\theta}(t) \le \frac{\theta\nu_1 t}{\delta},$$

on $\Omega_{T,\delta,\theta}$, we can bound the probability of this exceptional event by

$$\mathbf{P}_p\left(\{\hat{p}_\theta(\tau_\theta)=0\}\cap\Omega_{T,\delta,\theta}\right)<\mathbf{P}_p\left\{Z_0\leq\frac{\theta\nu_1 T}{\delta}\right\}<\frac{\theta\nu_1 T}{\delta},$$

and therefore

$$h_\theta(p)>\mathbf{P}_p(\Omega_{T,\delta,\theta})-\frac{\theta\nu_1 T}{\delta}$$

Thus, if we set $\theta_4=\theta_3\wedge\left(\frac{\delta\epsilon}{\nu_1 T}\right)$, it follows that for all $\theta\in[0,\theta_4]$,

$$h_\theta(p)\geq h_0(p)-4\epsilon,$$

and since $\epsilon>0$ can be taken arbitrarily small, we see that $\liminf_{\theta\to 0}h_\theta(p)\geq h_0(p)$. Since we can also show that $\liminf_{\theta\to 0}(1-h_\theta(p))\geq(1-h_0(p))$ by considering trajectories of $p_\theta(t)$ which hit 0 before hitting 1, it follows that $\lim_{\theta\to 0}h_\theta(p)=h_0(p)$ for every $p\in[0,1]$. Furthermore, because $h_\theta(p)$ is continuous and non-decreasing for every $\theta\geq 0$ (by Lemma 2.3 and Proposition 2.5, respectively), this result implies that $h_\theta(p)$ converges uniformly to $h_0(p)$ on $[0,1]$ as $\theta\to 0$.

To prove that the first derivatives converge uniformly on $[0,1]$, recall that according to Lemma 2.3 each function $h_\theta(p)$ is holomorphic on $(0,1)$. Since $h_\theta(p)$ converges uniformly to $h_0(p)$ on $[0,1]$, it follows that $h'_\theta(p)$ converges uniformly to $h'_0(p)$ on the interval $[\epsilon,1-\epsilon]$ for any $\epsilon>0$. To extend this result to all of $[0,1]$, rewrite the differential equation (9) satisfied by $h_\theta(p)$ in the form:

$$h''_\theta(p)=-2\sigma(p)h'_\theta(p)-2\theta\nu_1\left(\frac{h'_\theta(p)}{p}\right)+2\theta\nu_2\left(\frac{h'_\theta(p)}{1-p}\right)+2\theta\nu_1\left(\frac{h_\theta(p)}{p^2}\right)-2\theta\nu_2\left(\frac{1-h_\theta(p)}{(1-p)^2}\right).$$

If we fix $c\in(0,1)$ and integrate this equation over $[c,p]$ for $p\in(0,1)$, we obtain the equation

$$\begin{aligned}h'_\theta(p)-h'_\theta(c)\quad=\quad&-2\int_c^p\sigma(q)h'_\theta(q)dq-2\theta\nu_1\int_c^p\left(\frac{h'_\theta(q)}{q}\right)dq+2\theta\nu_2\int_c^p\left(\frac{h'_\theta(q)}{1-q}\right)dq+\\&2\theta\nu_1\int_c^p\left(\frac{h_\theta(q)}{q^2}\right)dq-2\theta\nu_2\int_c^p\left(\frac{1-h_\theta(q)}{(1-q)^2}\right)dq,\end{aligned}$$

which, by integration-by-parts in the first, fourth and fifth integrals, we can rewrite as

$$\begin{aligned}h'_\theta(p)\quad=\quad&h'_\theta(c)+2\sigma(c)h_\theta(c)-2\sigma(p)h_\theta(p)+2\int_c^p\sigma'(q)h_\theta(q)dq+\\&2\theta\nu_2\left(\frac{1-h_\theta(c)}{1-c}\right)-2\theta\nu_2\left(\frac{1-h_\theta(p)}{1-p}\right)+2\theta\nu_1\left(\frac{h_\theta(c)}{c}\right)-2\theta\nu_1\left(\frac{h_\theta(p)}{p}\right).\end{aligned}\tag{19}$$

Since c is fixed and since we know that all of the terms on the right-hand side except possibly the last one converge uniformly on $[0,c]$ as $\theta\to 0$, it follows that the quantity

$$h'_\theta(p)+2\theta\nu_1\left(\frac{h_\theta(p)}{p}\right),\tag{20}$$

which we extend continuously to $p=0$, also converges uniformly on $[0,c]$. In particular, because $h_\theta(p)$ is continuously differentiable on $[0,1]$ and $h_\theta(0)=0$, we obtain the identity

$$h'_\theta(0)\quad=\quad\frac{1}{1+2\theta\nu_1}\left[h'_\theta(c)+2\sigma(c)h_\theta(c)-2\int_0^c\sigma'(q)h_\theta(q)dq+2\theta\nu_2\left(\frac{c-h_\theta(c)}{1-c}\right)+2\theta\nu_1\left(\frac{h_\theta(c)}{c}\right)\right],$$

and the convergence of the right-hand side as $\theta \to 0$ to an expression equal to $h_0'(0)$ implies that $h_\theta'(0)$ converges to this quantity as well. Thus $h_\theta'(p)$ is pointwise convergent to $h_0'(p)$ on $[0,c]$. To strengthen this to uniform convergence, we need to show that the second term in Eq. (20) converges uniformly to 0 as $\theta \to 0$, and because this term is non-negative and contains θ as a factor, it suffices to show that there is a constant $M < \infty$ such that

$$sup\left\{\frac{h_\theta(p)}{p} : p \in (0,c], \theta \in [0,1]\right\} < M.$$

Observe that we can write Eq. (19) as

$$h_\theta'(p) = D(\theta,p) - 2\theta\nu_1\left(\frac{h_\theta(p)}{p}\right),$$

where the quantity $|D(\theta,p)|$ is bounded by a constant C for all $p \in [0,c]$ and all $\theta \in [0,1]$. Since $h_\theta(p) \geq 0$ and $h_\theta(0) = 0$, this implies that $h_\theta'(p) \leq C$ and therefore $h_\theta(p) \leq Cp$ for all $p \in [0,c]$ and $\theta \in [0,1]$, and so we can take $M = C$. Thus, $h_\theta'(p)$ converges uniformly to $h_0'(p)$ on $[0,c]$ as $\theta \to 0$, and a similar argument shows that this is true on $[c,1]$ as well. □

Proposition 3.2. *Let $S(p) = \int_0^p \sigma(q)dq$. Then the weak mutation limits of the scaled, stationary substitution rates defined in (18) are*

$$\begin{aligned} \mu_{2,low}^{CA} &\equiv \lim_{\theta\to 0}\theta^{-1}\mu_{2,\theta}^{CA} = \frac{\nu_2\, e^{-2S(1)}}{\int_0^1 e^{-2S(q)}dq} \\ \mu_{1,low}^{CA} &\equiv \lim_{\theta\to 0}\theta^{-1}\mu_{1,\theta}^{CA} = \frac{\nu_1}{\int_0^1 e^{-2S(q)}dq}. \end{aligned} \tag{21}$$

Proof. We first observe that if $\phi(\cdot)$ is any continuous function, then by splitting the domain of integration into the three regions $[0,\epsilon)$, $[\epsilon, 1-\epsilon]$ and $(1-\epsilon, 1]$ and taking $\epsilon > 0$ arbitrarily small, we can show that

$$\lim_{\theta\to 0}\int_0^1 \phi(p)\pi_\theta(p)dp = \pi\phi(1) + (1-\pi)\phi(0),$$

where $\pi_\theta(p)$ is determined from (2) and

$$\pi \equiv \frac{\nu_1 e^{2S(1)}}{\nu_2 + \nu_1 e^{2S(1)}},$$

i.e., the measures $\pi_\theta(dp)$ converge weakly to $\pi\delta_1(dp) + (1-\pi)\delta_0(dp)$ as $\theta \to 0$. Since $h_\theta(0) = h_0(0) = 0$ and $h_\theta(1) = h_0(1) = 1$ for all $\theta > 0$ and since $h_\theta(p)$ converges uniformly to $h_0(p)$, it follows that

$$\lim_{\theta\to 0}\int_0^1 h_\theta(p)\pi_\theta(p)dp = \lim_{\theta\to 0}\int_0^1 (h_\theta(p) - h_0(p))\pi_\theta(p)dp + \lim_{\theta\to 0}\int_0^1 h_0(p)\pi_\theta(p)dp = \pi.$$

Likewise, if we define $\phi_\theta(p) = (1-h_\theta(p))/(1-p)$ and $\phi_0(p) = (1-h_0(p))/(1-p)$ for $p \in [0,1)$ and $\phi_\theta(1) = h_\theta'(1)$ and $\phi_0(1) = h_0'(1)$, then Lemma 2.3 guarantees that all of these functions are continuous on $[0,1]$, while it follows from Lemma 3.1 that $\phi_\theta(p)$ converges uniformly to $\phi_0(p)$ on this interval. Consequently,

$$\lim_{\theta\to 0}\int_0^1 p\phi_\theta(p)\pi_\theta(p)dp = \lim_{\theta\to 0}\int_0^1 p(\phi_\theta(p) - \phi_0(p))\pi_\theta(p)dp + \lim_{\theta\to 0}\int_0^1 p\phi_0(p)\pi_\theta(p)dp = \pi h_0'(1).$$

It follows that $\theta^{-1}\mu_{2,\theta}^{CA}$ converges to $\nu_2 h'(1)$ as $\theta \to 0$ and a similar argument shows that $\theta^{-1}\mu_{1,\theta}^{CA}$ converges to $\nu_1 h'(0)$ in this same limit. The limiting expressions shown in (21) can then be derived from the formula for the fixation probability $h_0(p)$ shown in (16). □

The proof of Proposition 3.2 shows that the weak mutation limits $\mu_{i,low}^{CA}$ are closely related to an approximation commonly used to describe the 'flux of selected alleles' (Kimura 1964; Otto and Whitlock 1997) and incorporated into a phylogenetic framework by McVean and Vieira (2001). Writing $\mu_{1,low}^{CA} = \nu_1 h_0'(0) = N\nu_1 h_0(N^{-1}) + O(\nu_1^2)$, we see that the limiting substitution rate of P is approximately equal to the product of the number, $N\nu_1$, of new P mutants produced per generation and the fixation probability, $h_0(N^{-1})$, of a single such mutant in a population otherwise fixed for Q. In contrast, if we let $u_\theta(p)$ denote the fixation probability of P when the mutation rates are $\theta\nu_1$ and $\theta\nu_2$, then it is not true that $\nu_1 u_\theta'(0)$ converges to $\nu_1 h_0'(0)$ as $\theta \to 0$ since $u_\theta'(0) = \infty$ whenever $\theta > 0$ (see the scale function $s_\theta(p)$ introduced in the proof of Lemma 3.1). Thus the additional regularization of $h_\theta'(p)$ afforded by recurrent mutation to the common ancestor lineage (represented by the 'jump terms' in the BVP (9)) appears to be essential to the existence of the low mutation rate limit. We shall see in the next two sections that the approximation given by Proposition 3.2 is generally very good when selection and mutation are both weak, but tends to underestimate the substitution rates if either selection is strong or the mutation rates are high.

4 Purifying selection in a haploid population

We next show how the theory developed in the preceding section can be used to characterize the common ancestor process of a haploid population evolving according to a Wright-Fisher diffusion (1) with frequency-independent fitness differences between the alleles, i.e., $\sigma(p) \equiv s \neq 0$. Because we know from (2) that the density of the stationary distribution of this diffusion is $\pi(p) = Cp^{2\mu_1-1}(1-p)^{2\mu_2-1}e^{2sp}$, our description of the common ancestor distribution will be complete if we can solve (9) for the conditional probability $h(p)$.

To do so, we begin by supposing that we can expand $h(p)$ in a power series in s

$$h(p) = p + \sum_{n=1}^{\infty} h_n(p)s^n.$$

Substituting this expansion into (9) and collecting all terms multiplying s^n leads to a recursive series of BVP's for the functions $h_n(\cdot)$, $n \geq 1$,

$$\frac{1}{2}p(1-p)h_n''(p) + (\mu_1(1-p) - \mu_2 p)h_n'(p) - \left(\mu_2\left(\frac{p}{1-p}\right) + \mu_1\left(\frac{1-p}{p}\right)\right)h_n(p) \\ = -p(1-p)h_{n-1}'(p), \qquad (22)$$

subject to the conditions $h_n(0) = h_n(1) = 0$. To solve these inhomogeneous equations, we first need to determine the general solution to the corresponding homogeneous equation. Some guesswork leads to one solution

$$\psi_1(p) = p^{-2\mu_1}(1-p)^{-2\mu_2},$$

and a reduction of order calculation leads to a second linearly independent solution

$$\psi_2(p) = \psi_1(p)\beta(p) \equiv \psi_1(p) \int_0^p q^{2\mu_1}(1-q)^{2\mu_2} dq.$$

With these in hand, integration-by-parts and the method of variation of parameters can be used to find a recursive solution to the boundary value problem given in (22)

$$h_n(p) = \frac{2}{\beta'(p)} \left[\frac{\beta(p)}{\beta(1)} \int_0^1 \beta'(q) h_{n-1}(q) dq - \int_0^p \beta'(q) h_{n-1}(q) dq \right].$$

Defining $H_n(p) = \beta'(p)h_n(p)$ and $H(p) = \beta'(p)h(p)$, we can rewrite this recursion as

$$H_n(p) = 2\left(\frac{\beta(p)}{\beta(1)}\right) \int_0^1 H_{n-1}(q)dq - 2\int_0^p H_{n-1}(q)dq,$$

which, upon differentiating with respect to p, gives

$$H_n'(p) = 2\left(\frac{\beta'(p)}{\beta(1)}\right) \int_0^1 H_{n-1}(q)dq - 2H_{n-1}(p).$$

Term-by-term differentiation of the series expansion of $H(p)$ itself leads to the following first-order differential equation

$$H'(p) = \beta'(p) + p\beta''(p) + 2s\left(\frac{\beta'(p)}{\beta(1)}\right) \int_0^1 H(q)dq - 2sH(p),$$

and to find $h(p)$, we must divide the general solution to this equation by $\beta'(p)$ and impose the original boundary conditions $h(0) = 0$ and $h(1) = 1$ (which can be jointly satisfied). These calculations lead to the following expression for the conditional probability that the genotype of the common ancestor is P,

$$h(p) = p + 2s \int_0^p (\tilde{p} - q) e^{2s(q-p)} \left(\frac{q}{p}\right)^{2\mu_1} \left(\frac{1-q}{1-p}\right)^{2\mu_2} dq, \tag{23}$$

where the constant $\tilde{p}$ is the expectation of the allele frequency p with respect to the variance-biased stationary distribution $\tilde{\pi}(p)dp \equiv Csp(1-p)\pi(p)dp$ (where C is a normalizing constant):

$$\tilde{p} \equiv \int_0^1 p\tilde{\pi}(p)dp = \frac{\int_0^1 e^{2sq} q^{2\mu_1+1}(1-q)^{2\mu_2} dq}{\int_0^1 e^{2sq} q^{2\mu_1}(1-q)^{2\mu_2} dq}. \tag{24}$$

(Observe that $\tilde{p}$ is also the probability that a sample of three individuals from a stationary population contains two P and one Q individual conditional on it containing at least one individual of each genotype.)

We can calculate the marginal probability, π_1, that the common ancestor is of type P by integrating the density of the joint probability $\pi(1,p) = h(p)\pi(p)$ over $[0,1]$. Because this integral cannot be evaluated analytically, π_1 must be calculated by numerical integration, which can be done accurately using the method described in the appendix. Furthermore, by interchanging the order of integration in the resulting double integral, we arrive at the following intriguing expression for π_1

$$\pi_1 \equiv \mathbf{P}\{z=1\} = \mathbf{E}_\pi[p] + \mathbf{E}_\pi[2sp(1-p)]Cov_{\tilde{\pi}}\left(p, \ln\left(\frac{p}{1-p}\right)\right), \tag{25}$$

where $Cov_{\tilde{\pi}}(\cdot,\cdot)$ denotes the covariance with respect to the variance-biased stationary measure defined above. Although this expression is reminiscent of Price's equation (Price 1970), which states that the change in a trait caused by selection is equal to the covariance between that trait and fitness, it is not clear how to interpret the terms appearing within the covariance in a way that would make this correspondence precise.

When $s > 0$, i.e., P is fitter than Q, it is clear that the integral on the right-hand side of (23) vanishes at $p = 0, 1$ and is strictly positive when $p \in (0,1)$. Consequently, $h(p) \geq p$, as follows from Proposition 2.6, and thus the common ancestor is more likely to be of the fitter type than an individual chosen at random. Plots of $h(p)$ for different values of the (symmetric) mutation rates and selection coefficients are shown in Figure 1A. For fixed values of the mutation rates, we see that $h(p)$ is an increasing function of the selection coefficient s, which also follows from Proposition 2.6. On the other hand, for fixed positive values of s, $h(p)$ is a decreasing function of the mutation rates, probably because mutation reduces the correlation between the type of an extant lineage and its probability of surviving into the future.

Although expressions (15) and (23) fully determine the generator G of the common ancestor process, none of the terms containing $h(p)$ simplify and so we do not reproduce these here. Less cumbersome, approximate expressions for the substitution rates can be derived with the help of Proposition 3.2, which shows that the weak mutation limits are

$$\mu_{2,low}^{CA} = \mu_2\left(\frac{2s}{e^{2s}-1}\right) \text{ and } \mu_{1,low}^{CA} = \mu_1\left(\frac{2se^{2s}}{e^{2s}-1}\right).$$

These are also derived in Corollary 3 of Fearnhead (2002) and have been used by McVean and Vieira (2001) to estimate the strength of selection on codon usage in several *Drosophila* species. Of course, we can also use expressions (15) and (23) to calculate the exact common ancestor substitution rates. Figure 1B shows how the relative deleterious substitution rate, $\mu_2^{CA}(p)/\mu_2 = \frac{p(1-h(p))}{(1-p)h(p)}$, varies as a function of the frequency p of P. (As can be seen in (15), the relative beneficial substitution rate, $\mu_1^{CA}(p)/\mu_1$, is always the reciprocal of this quantity and so is not shown.) Note that the mutation rates are symmetric in parts A-C of Figure 1, i.e., $\mu_1 = \mu_2 \equiv \mu$, and that the substitution rates are scaled by the mutation rate. As expected, the relative deleterious substitution rate is always less than 1, i.e., the absolute substitution rate is less than the mutation rate, and this rate decreases as the selective advantage of P increases, but increases as the mutation rate μ increases. For comparison, we have also plotted the average deleterious substitution rates, μ_2^{CA}/μ_2, calculated using (18) and scaled by μ, as bold horizontal line segments on the right side of Figure 1B. Examining this figure reveals that for each fixed pair of values of μ and s, the average deleterious substitution rate is nearly as small as the smallest frequency-dependent rate (i.e., the bold horizontal lines lie beneath the corresponding curve for most values of p). Presumably this is because the conditional distribution of p given that the common ancestor is of type P is concentrated in a small region abutting the boundary $p = 1$ whenever P is selectively advantageous and the mutation rate is not too large.

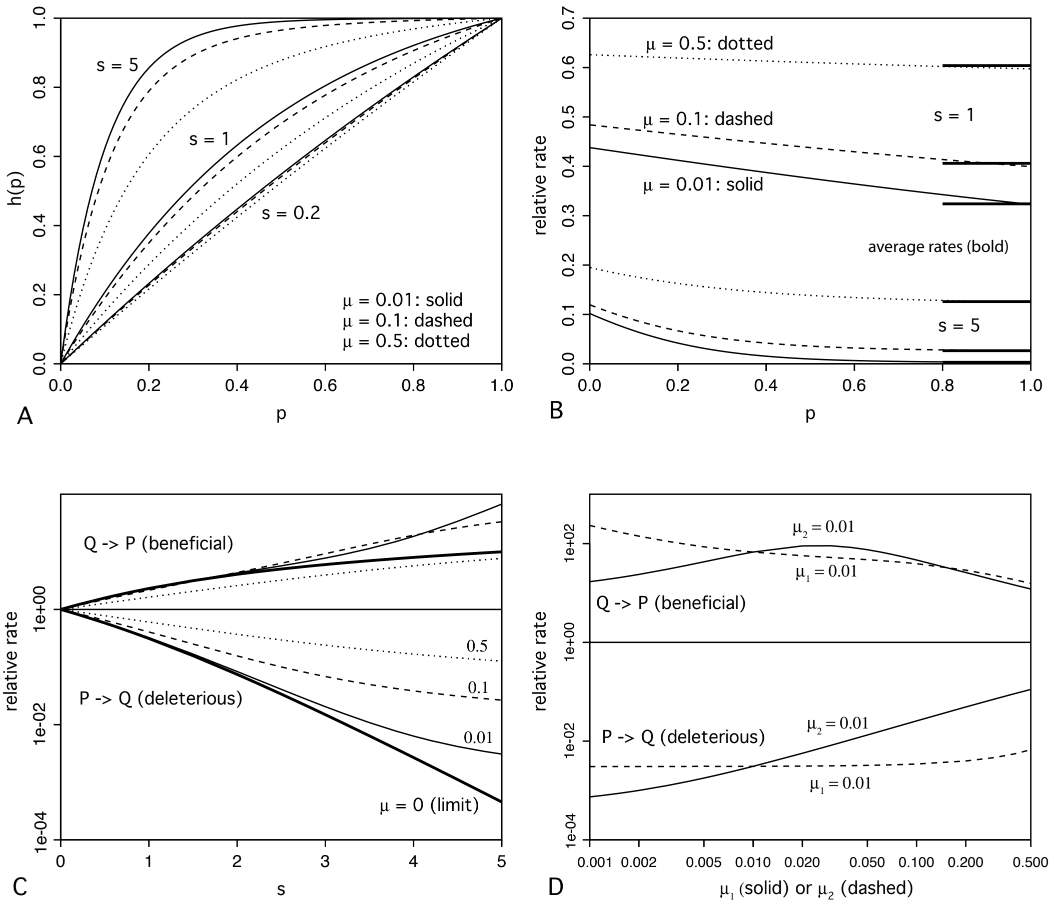

Figure 1: Stationary distribution and substitution rates of the common ancestor of a haploid population subject to purifying selection ($\sigma(p) \equiv s$) and either symmetric ($\mu_1 = \mu_2 \equiv \mu$) (A-C) or asymmetric (D) mutation. (A) shows the conditional probability $h(p)$ (Eq. 23) that the common ancestor is of type P, while (B) shows the frequency-dependent deleterious substitution rates (from Eq. 15) and their stationary averages (bold line segments; Eq. 18). (C) and (D) show the stationary averages of the beneficial ($Q \to P$) and deleterious ($P \to Q$) substitution rates, and C also shows the weak mutation limits (Eq. 21) (bold lines; $\mu = 0$). Note that solid, dashed, and dotted curves in (A-C) correspond to the symmetric mutation rates $\mu = 0.01$, 0.1, and 0.5, respectively, while in (D), $\mu_1 = 0.01$ is fixed and μ_2 varies along dashed lines, and $\mu_2 = 0.01$ is fixed and μ_1 varies along solid lines. Note that all substitution rates have been scaled by the corresponding mutation rate.

Figures 1C and D give a more detailed picture of the variation in substitution rates under different regimes of mutation and selection. In Figure 1C, we plot the average deleterious and favorable substitution rates, again scaled by the corresponding (symmetric) mutation rates, as functions of the selective advantage of P. (Because the conditional distributions used in (18) to define these two rates differ, the reciprocity property noted for the frequency-dependent rates no longer holds.) Also shown are the weak mutation limits, calculated using (21), which are shown as bold curves. It is notable that except when the mutation rate is very large ($\mu = 0.5$), both the beneficial and the deleterious substitution rates are greater than their weak mutation limits, although the discrepancy is greatest for deleterious substitutions and grows as the selection coefficient s increases. For example, when $s = 5$ and either $\mu = 0.01$ or $\mu = 0.1$, the average deleterious substitution rates are approximately 7 and 59-times greater than the corresponding weak mutation limit, respectively, while the average beneficial substitution rates for these two cases are approximately 7 and 3 times as large. Furthermore, because Figure 1B shows that the average substitution rates are nearly as small as the minimum frequency-dependent rates, we can conclude that the discrepancy between the average substitution rates and their weak mutation limits arises primarily because the limiting values underestimate the true substitution rates and not because the average substitution rates are too large. While this difference is small when the mutation rates are small, it could be as much as an order of magnitude or more in organisms having either very large effective population sizes or high mutation rates, e.g., in HIV-1 the mutation rate is approximately 3×10^{-5} mutations per nucleotide per viral generation, while the effective viral population size within infected hosts is usually estimated to be between $10^3 - 10^4$ (Koyous et al. 2006), giving $\mu \approx 3 \times 10^{-2} - 3 \times 10^{-1}$. In such cases, the stationary averages given by (18) will be much more accurate numerical summaries of the true, frequency-dependent substitution rates than approximations which neglect recurrent mutation.

To disentangle the effects of the two mutation rates, Figure 1D shows how the two substitution rates change when one of the mutation rates is held fixed and the other is varied. The behavior of the deleterious substitution rate is easiest to understand: increasing either mutation rate increases this substitution rate, although the effect is greatest when the mutation rate to the favorable allele is increased, presumably because a Q lineage is then more likely to mutate to P before going extinct. In contrast, the beneficial substitution rate eventually decreases when either mutation rate is increased, but is initially an increasing function of μ_1, possibly because of mutation-drift interactions, i.e., a larger mutation rate helps drive a rare favorable allele up to frequencies where selection can be effective compared with genetic drift.

4.1 Complementarity of the diffusion and graphical representations

The common ancestor process for a Wright-Fisher diffusion with frequency-independent selection has also been characterized by Fearnhead (2002) using the ancestral selection graph. This approach embeds the common ancestor process within a pure jump Markov process (z_t, n_t) taking values in the state space $E = \{0,1\} \times \mathcal{N}$, where n_t denotes the number of virtual lineages, all of the less fit type. Fearnhead (2002) shows that the stationary distribution of this process is given by

$$\begin{aligned} \pi(1;n) &= \left(\prod_{i=1}^{n} \lambda_i\right) \mathbf{E}_\pi[p(1-p)^n] \\ \pi(0;n) &= (1-\lambda_{n+1}) \left(\prod_{i=1}^{n} \lambda_i\right) \mathbf{E}_\pi[(1-p)^{n+1}], \end{aligned} \tag{26}$$

where $\mathbf{E}_\pi[\cdot]$ denotes an expectation with respect to the stationary measure $\pi(dp)$ and $\lambda_n = \lim_{k\to\infty} \lambda_n^{(k)}$ with

$$\lambda_{n-1}^{(k)} = \frac{2s}{n + 2(\mu_2 + \mu_1) + 2s - (n + 2\mu_2)\lambda_n^{(k)}} \tag{27}$$

and $\lambda_{k+1}^{(k)} = 0$, and we interpret empty brackets ($n = 0$) as being equal to 1. (This formula assumes that $s \geq 0$; if P is less fit than Q, then we simply exchange indices. Also note that (26) and (27) have been rewritten to reflect the scalings of μ_i and s used in this article rather than those in Fearnhead (2002).) The transition rates of the common ancestor process can be calculated by reversing the modified ancestral selection graph with respect to $\pi(z, dn)$ and also depend on the λ_n's; we refer the reader to Corollary 2 of Fearnhead (2002) for their values.

Because the marginal laws of the genealogical processes embedded within the structured coalescent and the ancestral selection graph are identical, it is clear that this will also be true of the common ancestor processes identified by these two methods. However, deducing this equality directly from the generators of the bivariate processes appears to be difficult and here we merely show that the marginal stationary distributions of the type of the common ancestor are the same.

Lemma 4.1. *Let $h(p)$ be the conditional probability defined in (23) and let $(\lambda_n)_{n\geq 1}$ be the sequence defined by (27). Then,*

$$h(p) = p + p\sum_{n\geq 1}(1-p)^n \left(\prod_{i=1}^n \lambda_i\right)$$

Proof. Using recursion (27) and its initial condition, we can show inductively that, for every $n \geq 1$ and every $k \geq 1$, $\lambda_n^{(k)} < s/(s+\mu_1)$, and therefore that $\lambda_n \leq s/(s+\mu_1)$ and $a_n \equiv \prod_{i=1}^n \lambda_i \leq (s/(s+\mu_1))^n$ as well. It follows that the function $g(p) = p\sum_{n\geq 1} a_n(1-p)^n$ is holomorphic in the open disk $D = D(1; 1+\epsilon)$ for some $\epsilon > 0$ and thus can be differentiated term-by-term on D. Substituting $g(p)$ into the left-hand side of equation (10), we obtain the following expression

$$\begin{aligned} &p(1-p)\left[(1+\mu_2)a_2 - (1+\mu_2+\mu_1+s)a_1\right] + \\ &\quad p\sum_{n\geq 2} n\left[\left(\frac{1}{2}(n+1)+\mu_2\right)a_{n+1} - \left(\frac{1}{2}(n+1)+\mu_1+\mu_2+s\right)a_n + sa_{n-1}\right](1-p)^n, \end{aligned}$$

and, using recursion (24), which holds with λ_n in place of $\lambda_n^{(k)}$, we can show that this is equal to $-sp(1-p)$ for all $p \in [0,1]$. Since $g(0) = g(1) = 0$, the uniqueness of solutions to second-order boundary value problems with smooth coefficients implies that $h(p) = p + g(p)$. □

The equality of the marginal stationary distributions follows upon integrating both sides of the identity asserted by the lemma with respect to the stationary measure $\pi(dp)$,

$$\pi_1 = \int_0^1 h(p)\pi(p)dp = \mathbf{E}_\pi[p] + \mathbf{E}_\pi[\sum_{n\geq 1} p(1-p)^n]\left(\prod_{i=1}^n \lambda_i\right) = \sum_{n\geq 0}\pi(1,n).$$

Another consequence of the lemma is that it provides an explicit formula for the constants λ_n,

$$\lambda_n = -\frac{v^{(n)}(1)}{v^{(n-1)}(1)}, \tag{28}$$

where $v(p) \equiv (h(p) - p)/p$. Of course, the algebraic operations needed to analytically evaluate successive derivatives of $v(p)$ essentially replicate the recursion satisfied by the λ_n. However, the one advantage that (28) does have over (27) is that it gives an explicit formula for λ_1,

$$\lambda_1 = \left(\frac{2s}{1 + 2\mu_2}\right)(1 - \tilde{p}),$$

which allows the recursion to be solved from the bottom-up, starting with the calculated value of λ_1, rather than from the top-down, with the approximation $\lambda_n \approx 0$ for some large value of n.

5 Selection and dominance in a diploid population

To illustrate the generality of the diffusion theoretic methods, in this section we will consider a diploid population and explore what effect the degree of dominance of fitness has on the common ancestor process. Several observations suggest that dominance plays an important role in molecular evolution. For example, it has long been known that deleterious mutations in coding sequences are usually recessive, possibly because of complementation by the fully functional allele or because of structural features of metabolic pathways (Kondrashov and Koonin 2004). Furthermore, even when the different alleles segregating at a locus are not individually advantageous, non-additive interactions between alleles can cause heterozygous genotypes to have higher or lower fitness than any of the possible homozygotes, leading to balancing or disruptive selection (Richman 2000). From classical population genetics theory we know that dominance relations affecting fitness can profoundly alter fixation probabilities and rates (Ewens 2004; Williamson et al. 2004) and so we would expect the same to be true of the substitution process of the common ancestor.

We formulate our model by considering a diploid population of effective population size $2N_e$ in which the relative fitnesses of the genotypes PP, PQ, and QQ are $1 + 2s : 1 + 2ds : 1$, respectively, and d is a constant which quantifies the dominance ($d > 0.5$) or recessiveness ($d < 0.5$) of P relative to Q. Note that when $d > 1$, heterozygotes have higher fitness than either homozygote and are said to be overdominant, whereas when $d < 0$, heterozygotes have lower fitness and then are said to be underdominant. By rescaling both the selection coefficient s and the mutation rates by a factor of $1/2N_e$ and speeding up time by a factor of $2N$, we can again approximate the changes in the frequency of P by a Wright-Fisher diffusion with generator (1) where $\sigma(p) = 2s(d - (2d - 1)p)$. When $d = 0.5$, $\sigma(p) \equiv s$ is constant and the common ancestor process can be characterized using either the results in the preceding section or those of Fearnhead (2002). However, for any other value of d, $\sigma(p)$ is frequency-dependent and neither set of results applies.

Because the ancestral selection graph has been identified for this diffusion model (Neuhauser 1999), one might try to identify the common ancestor process by generalizing the methods used by Fearnhead (2002). The main obstacle to implementing this approach is that trinary branchings are required to account for the frequency-dependence of fitness and it is unclear that

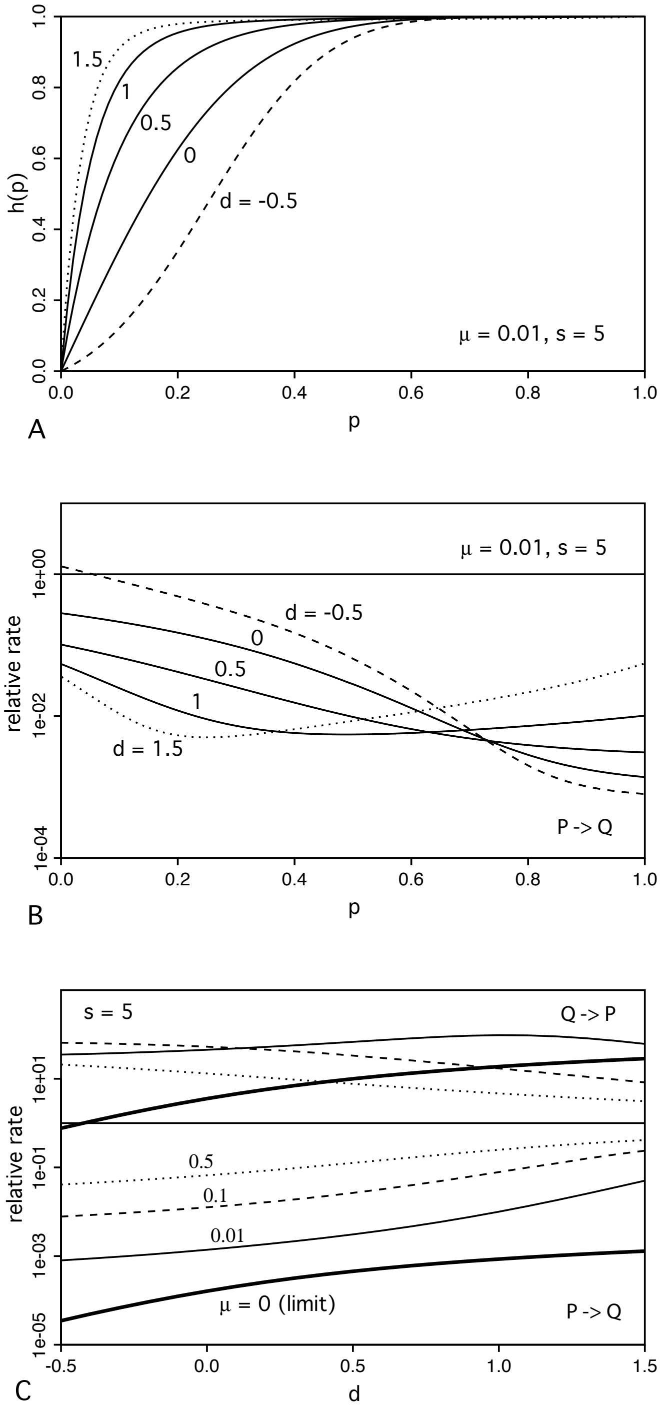

Figure 2: Stationary distribution and substitution rates of the common ancestor in a diploid population with relative genotypic fitnesses $1 + 2s$, $1 + 2ds$, and 1 for PP, PQ, and QQ, and symmetric mutation rates ($\mu_1 = \mu_2 \equiv \mu$). (A) shows the conditional probability $h(p)$ that the common ancestor is of type P, determined numerically by solving BVP (9), while (B) shows the frequency-dependent $P \rightarrow Q$ substitution rates (from Eq. 15). (C) shows the stationary averages (Eq. 18) and weak mutation limits (Eq. 21; bold lines) of the substitution rates, with $P \rightarrow Q$ substitution rates falling below 1 and $Q \rightarrow P$ substitution rates lying (mainly) above 1. Note that all substitution rates have been scaled by the corresponding mutation rate, and that $P \rightarrow Q$ substitutions are deleterious and $Q \rightarrow P$ substitutions are beneficial at all frequencies p

there is a pruned version of the ancestral selection graph of the kind found in Fearnhead (2002). However, without such a simplification, the common ancestor process will be embedded within a trivariate process $(z_t, n_1(t), n_2(t))$, where $n_i(t)$ is the number of virtual lineages of type A_i, and the stationary distribution will have to be determined by solving a multi-dimensional recursion.

In contrast, the diffusion theory developed in the first part of this article can be applied without modification to the new model. The price we pay for the added complexity of frequency-dependent selection is that we can no longer write down an explicit formula for $h(p)$: although we can still solve the BVP (9) recursively as in the previous section, we no longer obtain an integrating factor for the original equation. On the other hand, it is relatively easy to solve this problem numerically using the shooting method, even for large values of $|s|$. (See the appendix for a description of our numerical methods.) Furthermore, because the density $\pi(p)$ of the stationary distribution is known explicitly,

$$\pi(p) = Cp^{2\mu_1-1}(1-p)^{2\mu_2-1}e^{2sp[d-(d-1/2)p]},$$

we can use our numerical estimates of $h(p)$ to evaluate both the density of the common ancestor distribution and the substitution rates to the common ancestor. Of course, we can also use Proposition 3.2 to calculate the weak mutation limits to the substitution rates.

Figure 2 shows the results of these calculations. In Figure 2A we have plotted the numerical solutions themselves to show how $h(p)$ varies as a function of the dominance coefficient d when selection is moderately strong ($s = 5$). Qualitatively similar results are obtained both with weaker ($s = 1$) and stronger ($s = 10$) selection and so are not shown. We then substituted these numerical values into (15) to obtain the frequency-dependent 'deleterious' substitution rates ($P \to Q$) shown in Figure 2B. (Note that these substitutions are unconditionally deleterious only when the dominance coefficient d lies between 0 and 1.) The patterns evident in both figures can be interpreted by considering how dominance affects the marginal fitnesses of the two alleles at high and low frequencies of P. With increasing dominance of P over Q, the difference between the marginal fitnesses of the two alleles is reduced at high frequencies of P, rendering selection less effective and causing both a small decline in $h(p)$ but also a marked increase in the deleterious substitution rate at frequencies p close to 1. In contrast, because higher levels of dominance expose heterozygotes to stronger selection in populations which are nearly fixed for Q, these relationships are reversed when p is close to 0.

Similar considerations apply when heterozygotes are over- or under-dominant. One interesting feature of Figure 2B is that with disruptive selection ($d = -0.5$) the relative substitution rate of $P \to Q$ increases above 1 when the frequency of P is sufficiently close to 0. This is because the deterministic dynamics corresponding to this fitness scheme have an unstable internal equilibrium below which the marginal fitness of P is less than that of Q and selection favors Q substitutions. In contrast, when heterozygotes are over-dominant, the corresponding deterministic dynamics have a stable internal equilibrium and the marginal fitness of Q is an increasing function of the frequency of P, leading to the convex substitution rates seen in Figure 2B when $d = 1.5$.

An overview of how dominance affects the substitution process can be gleaned from Figure 2C, which shows plots of the (relative) average substitution rates for different values of d and μ, as well as the low mutation rate limits. We again see that the weak mutation rate limits (21) generally underestimate the average substitution rates (18), except when the mutation rate is so large that the long-term fitness of a lineage is partially decoupled from its current type. Also, whereas the deleterious ($P \to Q$) substitution rates are increasing functions of the dominance coefficient d, the beneficial ($Q \to P$) substitution rates are either unimodal or decreasing. In fact, even when μ is as small as 0.01, the beneficial substitution rate is seen to decrease slightly as d exceeds 1.3, probably because heterozygote advantage favors Q when the frequency of P is very high.

6 Multiple genetic backgrounds: prospects and problems

In this article we have used the structured coalescent in a fluctuating background to characterize the common ancestor process associated with a class of diffusion models important in population genetics theory. In addition to the classical Wright-Fisher diffusion, which can model general forms of frequency-dependent selection in panmictic populations, one-dimensional diffusions arise as scaling limits of models incorporating population structure (Cherry and Wakeley 2003), group selection (Roze and Rousset 2004) and environmental variation (Gillespie 1991). Although a closed-form solution was found only for the model with genic selection, the theory can also be used to quantify the influence of selection and genetic drift on the rate of molecular evolution under more complicated scenarios by first solving the BVP (9) numerically and then substituting the results into the expressions for the substitution rates which were derived as part of (15). Furthermore, as the section on the weak mutation limits illustrates, we can also use the theory to obtain analytical approximations for the substitution rates when selection, mutation or genetic drift are either very strong or very weak.

The most serious limitation of the diffusion-theoretic approach is that it leads to a much less tractable description of the common ancestor process when there are more than two genetic backgrounds. To illustrate both the difficulties and the potential interest of this approach, consider a locus which we will call the *focal locus* and which can occur in m different genetic backgrounds, $P_1, \cdots, P_m$, present at frequencies $p_1, \cdots, p_m$, respectively. Because the frequencies sum to one, we can describe the genetic composition of the population using any $m-1$ of these and so we will consider diffusion processes which take values in the $(m-1)$-dimensional simplex $K_{m-1} = \{(p_1, \cdots, p_{m-1}) : p_1, \cdots, p_{m-1} \geq 0, p_1 + \cdots + p_{m-1} \leq 1\}$. As before, let N_e denote the effective population size, let $1 + \sigma_i(p)/N_e$ denote the relative fitness of background P_i, and suppose that mutations from P_i to P_k occur at rate μ_{ik}/N_e, and that recombinations (or gene conversion events) involving individuals of type P_j change the background of the focal locus from type P_i to type P_k at rate $\rho(i,j|k)/N_e$. It will also be convenient to define $\mu_{ii} = \rho(i,j|i) = 0$ for all $i, j = 1, \cdots, m$. By rescaling the parameters and time in the usual manner and writing $\bar{\sigma}(p) = \sum_{i=1}^m p_i \sigma_i(p)$ for the mean fitness of the population, we obtain a Wright-Fisher diffusion with generator

$$\begin{aligned} A\psi(p) = {} & \frac{1}{2} \sum_{i,j=1}^{m-1} p_i(\delta_{ij} - p_j)\partial_i \partial_j \psi(p) \qquad (29) \\ & + \sum_{i=1}^{m-1} \left(\sum_{k=1}^{m} (p_k \mu_{ki} - p_i \mu_{ik}) + \sum_{j,k=1}^{m} (p_k p_j \rho(k,j|i) - p_i p_j \rho(i,j|k)) + p_i(\sigma_i(p) - \bar{\sigma}(p)) \right) \partial_i \psi(p), \end{aligned}$$

for $\psi \in \mathcal{C}^2(K_{m-1})$. Although we allow for recombination in this model, we emphasize that we are considering the common ancestor process at the focal locus only, and that $\rho(i,j|k)$ is the rate at which recombinations involving a non-ancestral lineage in background P_j change the background of the ancestral lineage from P_i to P_k. We could also define a structured ancestral recombination graph (Griffiths and Marjoram 1997) and use this to characterize the type of the common ancestor at several recombining loci, but this process would be even more complicated than the one we do consider here.

Unfortunately, when we try to write down the generator for the coalescent process of a sample of n genes from a population evolving according to this model, we encounter several complications that do not occur in biallelic models. One is that because the diffusion corresponding to generator (29) need not be not time-reversible with respect to its stationary distribution, i.e., the detailed balanced conditions need not hold, the generator $\tilde{A}$ of the time-reversed process may differ from A. If we denote the density of the stationary distribution of A by $\pi(p)$ (for which we will assume both existence and uniqueness), then we can use the adjoint condition (Nelson 1958)

$$\int_{K_{m-1}} \big(A\psi(p)\big)\phi(p)\pi(p)dp = \int_{K_{m-1}} \psi(p)\big(\tilde{A}\phi(p)\big)\pi(p)dp,$$

for $\phi, \psi \in \mathcal{C}^2(K_{m-1})$, to arrive at the following formal expression

$$\tilde{A}\phi(p) = \frac{1}{2}\sum_{i,j=1}^{m-1} p_i(\delta_{ij} - p_j)\partial_i\partial_j\phi(p) + \sum_{i=1}^{m-1}\left(\frac{1}{\pi(p)}\sum_{k=1}^{m-1}\partial_k\left(a_{ik}(p)\pi(p)\right) - b_i(p)\right)\partial_i\phi(p), \qquad (30)$$

where $b_i(\cdot)$ denotes the drift coefficient associated with $\partial_i\psi(\cdot)$ in (29). Although $\pi(p)$ is known explicitly only in those special cases where the diffusion is in fact reversible (see Li et al. 1999), (30) does at least provide us with a semi-explicit formula for the generator of the time-reversed diffusion process which we can use to study the common ancestor process.

When there is only one ancestral lineage, the structured coalescent corresponding to (29) can be denoted $(\tilde{z}_t, \tilde{p}_t) \in E \equiv \{1, \cdots, m\} \times K_{m-1}$, where $\tilde{z}_t = i$ if the type of the ancestral lineage at time t in the past was A_i, and $\tilde{p}_t = (\tilde{p}_1(t), \cdots, \tilde{p}_{m-1}(t))$ is the time-reversal of the frequency process. (Recall that we use the tilde to denote processes and generators which run backwards in time.) The generator of the structured retrospective process can then be written as

$$\begin{aligned}\tilde{G}\phi(z,p) \;\;=\;\; & \sum_{k=1}^{m}\mu_{kz}\left(\frac{p_k}{p_z}\right)(\phi(k,p) - \phi(z,p)) + \sum_{j,k=1}^{m}\rho(k,j|z)\left(\frac{p_kp_j}{p_z}\right)(\phi(k,p) - \phi(z,p)) + \\ & \tilde{A}\psi(z,p), \qquad (31)\end{aligned}$$

provided that $\phi(z,\cdot) \in \mathcal{C}^2(K_{m-1})$ for $z = 1, \cdots, m$. While the first and third terms correspond to terms appearing in the generator of the biallelic structured coalescent, the second term is novel and accounts for changes in the type of the common ancestor caused by recombination.

We again define the common ancestor distribution to be the stationary distribution of the retrospective process and we denote this quantity by $\pi(z,p)dp$, assuming the existence of a density with respect to Lebesgue measure on each copy of the simplex K_{m-1}. As in the biallelic case, $\pi(z,p)$ can be formally characterized by the condition

$$\sum_{z=1}^{m}\int_{K_{m-1}} \tilde{G}\phi(z,p)\,\pi(z,p)dp = 0,$$

valid for all $\phi(z,p)$ in the domain of $\tilde{G}$, which leads to the following system of coupled partial differential equations

$$\tilde{A}^*\pi(z,p) + \sum_{k=1}^{m}\left[\mu_{zk}\left(\frac{p_z}{p_k}\right)\pi(k,p) - \mu_{kz}\left(\frac{p_k}{p_z}\right)\pi(z,p)\right] + \\ \sum_{j,k=1}^{m}\left[\rho(z,j|k)\left(\frac{p_z p_j}{p_k}\right)\pi(k,p) - \rho(k,j|z)\left(\frac{p_k p_j}{p_z}\right)\pi(z,p)\right] = 0, \qquad z = 1,\cdots,m; \tag{32}$$

here $\tilde{A}^*$ denotes the formal adjoint of $\tilde{A}$ with respect to Lebesgue measure on K_{m-1}. We have, of course, little hope of being able to solve these equations, even numerically: not only are we confronted with a system of singular PDE's, but to make matters worse, we do not have a fully explicit expression for $\tilde{A}^*$.

On the other hand, if we write $\pi(z,p) = h_z(p)\pi(p)$, where $h_z(p)$ denotes the conditional probability that the common ancestor is of type A_z given that the backgrounds are segregating at frequencies p, then we can at least overcome the latter problem. Substituting this expression into (32) and noting that $h_1(p)+\cdots+h_m(p) = 1$ for every $p \in K_{m-1}$, we find that the functions $h_z(p)$ also satisfy a system of coupled equations,

$$Ah_z(p) + \sum_{k=1}^{m}\left[\mu_{zk}\left(\frac{p_z}{p_k}\right)h_k(p) - \mu_{kz}\left(\frac{p_k}{p_z}\right)h_z(p)\right] + \\ \sum_{j,k=1}^{m}\left[\rho(z,j|k)\left(\frac{p_z p_j}{p_k}\right)h_k(p) - \rho(k,j|z)\left(\frac{p_k p_j}{p_z}\right)h_z(p)\right] = 0, \qquad z = 1,\cdots,m-1, \tag{33}$$

but that now the partial differential operator is known explicitly. The boundary conditions for this system are given by $h_z(e_z) = 1$, where e_z is the vertex of K_{m-1} with z'th coordinate equal to 1 and all other coordinates equal to 0, and $h_z(p) = 0$ for all $p \in K_{m-1}$ such that $p_z = 0$.

This system of equations can also be used to obtain a semi-explicit expression for the generator G of the common ancestor process associated with the diffusion (29). The adjoint condition on G and $\tilde{G}$ with respect to $\pi(z,p)dp$ is now

$$\sum_{z=1}^{m}\int_{K_{m-1}}\big(\tilde{G}\phi(z,p)\big)\psi(z,p)\pi(z,p)dp = \sum_{z=1}^{m}\int_{K_{m-1}}\phi(z,p)\big(G\psi(z,p)\big)\pi(z,p)dp,$$

for any $\phi \in \mathcal{D}(\tilde{G})$ and $\psi \in \mathcal{D}(G)$, and this in combination with (33) formally implies that the generator of the common ancestor process is

$$G\psi(z,p) = A\psi(z,p) + \sum_{i=1}^{m-1}\left(\sum_{k=1}^{m-1}p_i(\delta_{ik}-p_k)\left(\frac{\partial_k h_z(p)}{h_z(p)}\right)\right)\partial_i\psi(z,p) + \\ \sum_{k=1}^{m}\mu_{zk}\left(\frac{p_z h_k(p)}{p_k h_z(p)}\right)(\psi(k,p)-\psi(z,p)) + \sum_{j,k=1}^{m}\rho(z,j|k)\,p_j\left(\frac{p_z h_k(p)}{p_k h_z(p)}\right)(\psi(k,p)-\psi(z,p)) \tag{34}$$

As in the biallelic case, the forward diffusion is modified by a drift term which reflects the excess offspring produced by the common ancestor, while the substitution rates are modified by factors

$\left(\frac{p_z h_k(p)}{p_k h_z(p)}\right)$ which account for the effects of selection. One new feature of (34) is that the rate at which recombination changes the type of the common ancestor is also influenced by selection, although the effect depends only on the types of the backgrounds A_z and A_k of the common ancestor before and after the recombination event, and not on the type A_j of the individual with which the common ancestor recombines. This suggests that attempts to quantify recombination using phylogenetic methods (e.g., Patterson et al. 2006) could be confounded by selection.

There are a few situations in which the multidimensional Wright-Fisher diffusion is reversible with respect to its stationary distribution, allowing analytical expressions for the conditional probabilities $h_k(p)$ and the generator G to be found (Li et al. 1999). Under complete neutrality and parent-independent mutation, we have $\sigma_k(p) \equiv 0$ and $\mu_{ik} \equiv \mu_k$ for all $i, k = 1, \cdots, m$, and direct substitution into equation (33) shows that $h_z(p) = p_z$ as expected. In this case, the stationary distribution of background frequencies is known to be the Dirichlet distribution with parameters $(2\mu_1, \cdots, 2\mu_m)$. Moreover, under complete neutrality, equation (33) shows that it is true that $h_z(p) = p_z$ even when the mutation rates are parent-dependent, although we are then unable to write down an explicit formula for the density $\pi(p)$.

If the genetic backgrounds can be partitioned into two fitness classes, say $\mathcal{F}$ and $\mathcal{U}$, with fitnesses $1+s$ and 1, respectively, and mutation is parent-independent, then as in Fearnhead (2002) we can use the solution from the corresponding biallelic model to determine the stationary distribution and generator of the multi-allelic common ancestor process. Suppose that $\mathcal{F} = \{P_1, \cdots, P_l\}$ and $\mathcal{U} = \{P_{l+1}, \cdots, P_m\}$, and let $\mu_{\mathcal{F}} = \mu_1 + \cdots + \mu_l$, $\mu_{\mathcal{U}} = \mu_{l+1} + \cdots + \mu_m$, and $p_{\mathcal{F}} = p_1 + \cdots + p_l$. Then $p_{\mathcal{F}}(t)$ evolves according to a Wright-Fisher diffusion with parameters $\mu_{\mathcal{F}}$, $\mu_{\mathcal{U}}$, and s, and so the probability that the common ancestor belongs to the fitness class $\mathcal{F}$ given that the frequency of that class is p is given by equation (23), where we set $\mu_1 = \mu_{\mathcal{F}}$ and $\mu_2 = \mu_{\mathcal{U}}$. Furthermore, using equation (33), we can show that the multi-allelic conditional distribution of the type of the common ancestor is given by

$$h_z(p) = \left(\frac{p_z}{p_{\mathcal{F}}}\right) h(p_{\mathcal{F}}) \text{ if } z = 1, \cdots, l \quad \text{and} \quad h_z(p) = \left(\frac{p_z}{1 - p_{\mathcal{F}}}\right) (1 - h(p_{\mathcal{F}})) \text{ if } z = l+1, \cdots, m. \tag{35}$$

Since the density of the stationary distribution is given by Wright's formula,

$$\pi(p) = C e^{2sp_{\mathcal{F}}} \prod_{k=1}^{m} p_k^{2\mu_k - 1},$$

it follows that (35) determines both the common ancestor distribution and the generator of the common ancestor process.

Unfortunately, analytical solutions such as these are rarely available, and thus the difficulty of numerically solving the system of singular PDE's in (33) limits the usefulness of this theory. Extensions to models based on multidimensional diffusions are important for several reasons. On the one hand, while neutral substitutions at different sites will occur independently of one another (assuming that the mutation rates are not context-dependent), selection will lead to correlated substitution processes whenever fitness is determined epistatically or when there is genetic linkage between polymorphic loci. It is important to understand and to quantify these correlations not only because they may alter the marginal substitution rates, but also because of the significant role which they might play in processes such as speciation and the evolution of recombination. Furthermore, even if we could assume that the substitution processes at different sites or different codons were independent, we would still need to consider multidimensional diffusions in order to correctly describe single nucleotide substitution processes which can involve any one of four different DNA or RNA bases and which exhibit parent-dependent mutation. Indeed, as McVean and Vieira (2001) found in their study of codon bias, mutation rates can exhibit pronounced asymmetries which, if ignored, could be incorrectly interpreted as evidence of selective constraints. For these reasons, the development of efficient numerical methods to solve equations such as (33) would greatly enhance the value of this theory.

Appendix

Singular boundary value problems such as (9) and (10) can be solved using the shooting method, as described in Barton and Etheridge (2004) and Press et al. (1992). Our approach follows that described in the former paper, but with one modification which is needed to solve problems with either large mutation rates or large selection coefficients.

The difficulty posed by singular equations is that the gradients of highest order generally diverge wherever their coefficients vanish. In our case, this means that $h''(p)$ may diverge as p approaches 0 or 1. Accordingly, it is not possible to integrate directly from the boundary points and instead shooting proceeds from interior points which are offset from the boundaries by some small quantity, say ϵ. To do so, the boundary conditions must also be transferred to these interior points. Often this is done simply by shifting the boundary conditions, unchanged, to the interior points, which in our case would mean setting $h(\epsilon) = 0$ and $h(1-\epsilon) = 1$. However, we found that with large values of μ_1, μ_2 or s, this approach did not produce accurate solutions, as evidenced by numerical solutions exceeding 1 in the interior ($h(p)$ is a probability and so must always take values between 0 and 1), and by unacceptably large discrepancies between the numerical solution and the exact solution (23) for the BVP corresponding to the Wright-Fisher diffusion with genic selection.

To resolve these problems, we found it necessary to modify the shifted boundary conditions to account for the displacement of the initial points of the integration, which we were able to do by expanding $h(p)$ in a Taylor series. This leads to the first-order approximations $h(\epsilon) \approx \epsilon h'(\epsilon)$ and $h'(1-\epsilon) \approx 1 - \epsilon h'(1-\epsilon)$, where the boundary gradients $h'(\epsilon)$ and $h'(1-\epsilon)$ are determined by the shooting algorithm. With these corrections, the numerical solutions satisfied the necessary upper and lower bounds and in the case of genic selection agreed with the known solution to four or more decimal places.

Accurate numerical evaluation of the integrals in (18) can also be delicate when the mutation rates are small ($\mu_i < 0.5$) because then the density $\pi(p)$ of the stationary measure diverges at the boundaries. As with the BVP (9), one cannot numerically integrate all the way to the boundary, and we found that truncating the domain of integration to $(\epsilon, 1-\epsilon)$ led to poor approximations whenever the mutation rates were very small ($\mu_i \leq 0.01$) and the selection coefficient was large ($s \geq 5$). Furthermore, we were unable to resolve this simply by taking ϵ to be very small without exceeding the tolerance of the numerical integration algorithms implemented in *Mathematica*.

This problem can be resolved by splitting the singular integrals into three parts,

$$\begin{aligned}\int_0^1 F(p)\pi(p)dp &\equiv \int_0^1 F(p)G(p)p^{2\mu_1-1}(1-p)^{2\mu_2-1}dp \\ &= \int_0^\epsilon F(p)G(p)p^{2\mu_1-1}(1-p)^{2\mu_2-1}dp + \int_\epsilon^{1-\epsilon} F(p)G(p)p^{2\mu_1-1}(1-p)^{2\mu_2-1}dp \\ &\quad + \int_{1-\epsilon}^1 F(p)G(p)p^{2\mu_1-1}(1-p)^{2\mu_2-1}dp,\end{aligned}$$

where $\epsilon > 0$ is chosen small enough that the locally smooth functions $F(p)G(p)(1-p)^{2\mu_2-1}$ and $F(p)G(p)p^{2\mu_1-1}$ can be approximated by $F(0)G(0)$ and $F(1)G(1)$ in the first and third integral, respectively. With this approximation, the boundary integrals can be evaluated analytically, while the non-singular integral over $(\epsilon, 1-\epsilon)$ can be evaluated numerically. The accuracy of this scheme was tested by comparing the expected substitution rates for the Wright-Fisher model with genic selection obtained using the diffusion characterization with those reported in Fearnhead (2002) using an independent characterization, and the two sets of rates were seen to agree to within the number of digits reported in the latter paper. A *Mathematica* program implementing both the shooting and the integration methods described here is available from the author upon request.

Acknowledgements I would like to thank Ellen Baake for bringing this problem to my attention, Alison Etheridge for helpful discussions of the mathematics, and Nick Barton, Gil McVean and two anonymous reviewers for comments on the paper.

References

[1] H. Akashi. Inferring Weak Selection from Patterns of Polymorphism and Divergence at Silent Sites in Drosophila DNA. *Genetics*, 139:1067–1076, 1995.

[2] E. Baake and H.-O. Georgii. Mutation, selection, and ancestry in branching models: a variational approach. *J. Math. Biol.*, 54:257–303, 2007. MR2284067

[3] N. H. Barton and A. M. Etheridge. The Effect of Selection on Genealogies. *Genetics*, 166:1115–1131, 2004.

[4] N. H. Barton, A. M. Etheridge, and A. K. Sturm. Coalescence in a Random Background. *Ann. Appl. Prob.*, 14:754–785, 2004. MR2052901

[5] N.H. Barton and S. P. Otto. Evolution of Recombination Due to Random Drift. *Genetics*, 169:2353–2370, 2005.

[6] G. Birkhoff and G.-C. Rota. *Ordinary differential equations.* Wiley, New York, 1989. MR0972977

[7] C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl. Directional Selection and the Site-Frequency Spectrum. *Genetics*, 159:1779–1788, 2001.

[8] J. L. Cherry and J. Wakeley. A Diffusion Approximation for Selection and Drift in a Subdivided Population. *Genetics*, 163:421–428, 2003.

[9] G. Coop and R.C. Griffiths. Ancestral inference on gene trees under selection. *Theor. Pop. Biol.*, 66:219–232, 2004.

[10] P. Donnelly and T. G. Kurtz. A Countable Representation of the Fleming-Viot Measure-Valued Diffusion. *Ann. Prob.*, 24:698–742, 1996. MR1404525

[11] P. Donnelly and T. G. Kurtz. Genealogical Processes for Fleming-Viot Models with Selection and Recombination. *Ann. Appl. Prob.*, 9:1091–1148, 1999. MR1728556

[12] A. M. Etheridge. Evolution in Fluctuating Populations. In A. Bovier, F. Dunlop, F. den Hollander, A. van Enter, and J. Dalibard, editors, *Mathematical Statistical Physics*, volume 83, pages 489–545. Les Houches, Elsevier, 2005.

[13] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence.* John Wiley & Sons, New York, N.Y., 1986. MR0838085

[14] W. J. Ewens. *Mathematical Population Genetics. Biomathematics, Vol 9.* Springer, New York, 2 edition, 2004. MR2026891

[15] P. Fearnhead. The Common Ancestor at a Nonneutral Locus. *J. Appl. Prob.*, 39:38–54, 2002. MR1895142

[16] S. Gavrilets. Perspective: Models of speciation: What have we learned in 40 years? *Evolution*, 57:2197–2215, 2003.

[17] H.-O. Georgii and E. Baake. Supercritical Multitype Branching Processes: The Ancestral Types of Typical Individuals. *Adv. Appl. Prob.*, 35:1090–1110, 2003. MR2014271

[18] J. H. Gillespie. *The Causes of Molecular Evolution.* Oxford University Press, Oxford, 1991.

[19] J. H. Gillespie. *Population Genetics: A Concise Guide.* Johns Hopkins University Press, Baltimore, 2004.

[20] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavare, editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer-Verlag, 1997. MR1493031

[21] P. Jagers. General branching processes as Markov fields. *Stoch. Proc. Appl.*, 32:183–242, 1989. MR1014449

[22] P. Jagers. Stabilities and instabilities in population dynamics. *J. Appl. Prob.*, 29:770–780, 1992. MR1188534

[23] N. L. Kaplan, T. Darden, and R. R. Hudson. The coalescent process in models with selection. *Genetics*, 120:819–829, 1988.

[24] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus.* Springer, New York, 1991. MR1121940

[25] M. Kimura. Diffusion Models in Population Genetics. *J. Appl. Prob.*, 1:177–232, 1964. MR0172727

[26] F. A. Kondrashov and E. V. Koonin. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.*, 20:287–291, 2004.

[27] R. D. Koyous, C. L. Althaus, and S. Bonhoeffer. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends Microbiol.*, 14:507–511, 2006.

[28] S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theor. Pop. Biol.*, 51:210–237, 1997.

[29] R. Lande. Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, 30:314–334, 1976.

[30] Z. Li, T. Shiga, and L. Yao. A Reversibility Problem for Fleming-Viot Processes. *Elect. Comm. in Probab.*, 4:65–76, 1999. MR1711591

[31] M. Lynch and J. S. Conery. The Origins of Genome Complexity. *Science*, 302:1401–1404, 2003.

[32] G. McVean and J. Vieira. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics*, 157:245–257, 2001.

[33] E. Nelson. The Adjoint Markoff Process. *Duke Math. J.*, 25:671–690, 1958. MR0101555

[34] C. Neuhauser. The ancestral graph and gene genealogy under frequency-dependent selection. *Theor. Pop. Biol.*, 56:203–214, 1999.

[35] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936, 1998.

[36] M. Frank Norman. Ergodicity of Diffusion and Temporal Uniformity of Diffusion Approximation. *J. Appl. Prob.*, 14:399–404, 1977. MR0436355

[37] R. B. O'Hara. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. *Proc. Roy. Soc. Lond. B*, 272:211–217, 2005.

[38] S. P. Otto and M. C. Whitlock. The Probability of Fixation in Populations of Changing Size. *Genetics*, 146:723–733, 1997.

[39] N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441:1103–1108, 2006.

[40] A. Poon and L. Chao. Drift Increases the Advantage of Sex in RNA Bacteriophage Φ6. *Genetics*, 166, 2004.

[41] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C. The Art of Scientific Computing.* Cambridge University Press, Cambridge, 1992.

[42] G. R. Price. Selection and Covariance. *Nature*, 227:520–521, 1970.

[43] A. Richman. Evolution of balanced genetic polymorphism. *Mol. Ecol.*, 9:1953–963, 2000.

[44] D. Roze and F. Rousset. Selection and Drift in Subdidived Populations: A Straightforward Method for Deriving Diffusion Approximations and Applications Involving Dominance, Selfing and Local Extinctions. *Genetics*, 165:2153–2166, 2004.

[45] S. A. Sawyer and D. L. Hartl. Population Genetics of Polymorphism and Divergence. *Genetics*, 132:1161–1176, 1992.

[46] T. Shiga. Diffusion processes in population genetics. *J. Math. Kyoto Univ.*, 21:133–151, 1981. MR0606316

[47] M. Stephens and P. Donnelly. Ancestral inference in population genetics models with selection. *Austral. & New Zealand J. Stat.*, 45:395–423, 2003. MR2018460

[48] M. C. Whitlock. Fixation of New Alleles and the Extinction of Small Populations: Drift Load, Beneficial Alleles, and Sexual Selection. *Evol.*, 54:1855–1861, 2000.

[49] S. Williamson, A. Fledel-Alon, and C. D. Bustamante. Population Genetics of Polymorphism and Divergence for Diploid Selection Models With Arbitrary Dominance. *Genetics*, 168:463–475, 2004.

[50] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, 11:367–372, 1996.

[51] A. Zharkikh. Estimation of Evolutionary Distances Between Nucleotide Sequences. *J. Mol. Evol.*, 39:315–329, 1994.