Conditional Gene Genealogies under Strong Purifying Selection

John Wakeley

Department of Organismic and Evolutionary Biology, Harvard University

The ancestral selection graph, conditioned on the allelic types in the sample, is used to obtain a limiting gene genealogical process under strong selection. In an equilibrium, two-allele system with strong selection, neutral gene genealogies are predicted for random samples and for samples containing at most one unfavorable allele. Samples containing more than one unfavorable allele have gene genealogies that differ greatly from neutral predictions. However, they are related to neutral gene genealogies via the well-known Ewens sampling formula. Simulations show rapid convergence to limiting analytical predictions as the strength of selection increases. These results extend the idea of a soft selective sweep to deleterious alleles and have implications for the interpretation of polymorphism among disease-causing alleles in humans.

Introduction

Hermisson and Pennings recently put forward the idea of a "soft selective sweep" and also studied the properties of these interesting events (Hermisson and Pennings 2005; Pennings and Hermisson 2006a, 2006b). A soft sweep is the fixation of a positively selected allele in which multiple independent copies of the allele contribute to the sweep. These may be different copies of a single mutant allele present at a time when selection starts to act (a soft sweep via standing variation), or they may be independently derived copies that appear during the fixation event (a soft sweep via recurrent mutation). In this article, a new kind of soft sweep by recurrent mutation is considered, one in which the allele is "disfavored" and thus does not sweep to fixation. Its appearance in appreciable frequency in the population is a random event in opposition to selection. There is considerable debate over explanations of the diversity of alleles responsible for human diseases (Slatkin and Rannala 1997; Terwilliger and Weiss 1998; Reich and Lander 2001; Pritchard and Cox 2002; Di Rienzo 2006). The results presented here emphasize the importance of stochastic effects in determining allele frequencies.

Strong selection against one allele and in favor of another is modeled here using a coalescent approach (Kingman 1982a, 1982b; Hudson 1983; Tajima 1983). In general, it has proven difficult to incorporate selection into the coalescent and still maintain the analytical tractability and ease of application of the model. This is because different selected alleles have different rates of coalescence—they are not "exchangeable" (Cannings 1974; Kingman 1982c; Aldous 1985)—and because the frequencies of alleles change over time by selection, mutation, and drift. One approach to the coalescent with selection is to condition rates of coalescence on allele frequencies and model changes in allele frequencies explicitly (Kaplan et al. 1988; Barton et al. 2004). A second approach, called the ancestral selection graph, was proposed by Krone and Neuhauser (1997) and provides the framework for the work presented here.

The ancestral selection graph arises via an augmentation of the typical forward-time models of population genetics, the same models that yield the standard Wright–Fisher diffusion (Ewens 2004). The augmented model is built in two layers. For the first layer, an exchangeable forward-time population process (obtained by imagining that the population is composed entirely of the fittest genotype) is run for an infinitely long time to produce a large graph. This graph contains all the ancestor–descendant relationships, birth and death events, etc., that occurred in the population. For the second layer, a small fraction of birth events are marked as being available only to the fittest genotype. Selection is enforced in a second run through the graph, in which allelic states are assigned and their fates are followed forward in time. Less-fit alleles are barred from using the marked birth events. The properties of the original model are preserved, but the inclusion of an exchangeable process in the two-layer model greatly facilitates the study of gene genealogies.

Although the population model behind the ancestral selection graph has been generalized considerably (Neuhauser and Krone 1997; Neuhauser 1999; Fearnhead 2006), for simplicity, the present work is based on the original formulation of Krone and Neuhauser (1997). Two alleles, A_1 and A_2, experience mutations with probability u per generation, and allele A_2 has fitness $1 + s$ relative to A_1. The standard diffusion assumptions are made: time is measured in proportion to N generations and N tends to infinity, whereas u and s tend to zero, such that the dynamics depend on scaled mutation and selection parameters θ and σ. There is no recombination within the locus. Krone and Neuhauser (1997) assumed a haploid Moran model of reproduction (Moran 1958, 1962), in which $\theta = Nu$ and $\sigma = Ns$, but the ancestral selection graph should hold for any model that has the standard diffusion as its limit. It may be assumed, without loss of generality, that A_2 is the fitter of the two alleles ($\sigma > 0$).

One small modification is made, which is to allow for asymmetric mutation in the following way. When a mutation occurs, it has probability α_1 of producing an A_1 allele and probability $\alpha_2 = 1 - \alpha_1$ of producing an A_2 allele. Any asymmetric, two-allele model can be represented in this way, and such "parent-independent" mutation generalizes readily to multiple alleles; for example, see Stephens and Donnelly (2003). Note that this means some mutation events are "empty" (Baake and Bialowons 2008), in the sense that they do not change the allelic type. This further augmentation of the model allows for a simplification (Fearnhead 2002) that will be important in what follows.

When run for a long time, which is from an essentially infinite time in the past to the present, this two-allele

Key words: natural selection, coalescent, ancestral graph, Ewens sampling formula

E-mail: wakeley@fas.harvard.edu.

Mol. Biol. Evol. 25(12):2615–2626. 2008
doi:10.1093/molbev/msn209
Advance Access publication September 17, 2008

population process will reach a statistical equilibrium. Here, the present time is defined as time $t = 0$ and the past as times $t > 0$. Analysis of this model, or its two-layer counterpart, shows that the frequency of A_1 in the population at equilibrium, at time $t = 0$, has distribution

$$h(x) = Bx^{\theta\alpha_1 - 1}(1 - x)^{\theta\alpha_2 - 1}\mathrm{e}^{-\sigma x}, \tag{1}$$

where the constant B is defined such that $\int_0^1 h(x)\mathrm{d}x = 1$ (Wright 1931, 1949; Kimura 1955). The Moran model derivation of equation (1) can be found in Moran (1962, p. 134), where B is expressed in terms of a confluent hypergeometric function; see Slater (1960) or Abramowitz and Stegun (1964, Chapter. 13).

A sample of size n, taken from the population at the present time zero, contains n_1 copies of allele A_1 and $n_2 = n - n_1$ copies of allele A_2 with probability

$$\int_0^1 x^{n_1}(1 - x)^{n_2} h(x)\mathrm{d}x, \tag{2}$$

which can also be expressed in terms of confluent hypergeometric functions. For ease of analysis and explanation below, the sample is assumed to be ordered. One possible ordering is that samples 1 through n_1 are of allelic type A_1 and samples $n_1 + 1$ through n are of type A_2. There are $\binom{n}{n_1}$ possible orderings of such a sample, and every one of these has the same probability; the probability of the corresponding unordered sample is $\binom{n}{n_1}$ times equation (2).

Crucially for the ancestral selection graph, equation (2) also holds at any time t in the past when there are n lineages ancestral to a present-day sample, subject to certain conditions which can be found in Donnelly and Kurtz (1999). Intuitively, this follows from the fact that the population has been evolving for an infinite length of time even before t and because present-day samples are taken at random with respect to genetic variation or any events that have occurred in the population.

The ancestral selection graph is obtained by following a random sample of genetic lineages from the present back into the past under the two-layer population model. Initially, the allelic types of the samples are not specified, and an ancestral graph is obtained by tracing back through the exchangeable population process. This proceeds from time zero back to the ultimate ancestor of the sample, which is reached the first time the entire sample is descended from a single lineage (Krone and Neuhauser 1997). Each ancestral lineage experiences mutations at rate $\theta/2$, each pair of lineages coalesces with rate 1, and each lineage "branches" with rate $\sigma/2$.

Branching events correspond to the marked birth events described above. When a branching event occurs, the lineage that experiences it splits into two lineages. Thus, the number of ancestral lineages can increase as they are followed back in time. Branching events capture the effect of selection in favor of A_2. They must be included in the graph in order to have an ancestral process in which lineages are initially exchangeable. They are resolved in the second run through the graph, with allelic states specified, such that the gene genealogy of the sample is a bifurcating tree and A_2 enjoys a higher fitness than A_1. In order to resolve branching events and retrieve the gene genealogy of the sample, one of the two lineages emanating from each branching event is labeled the "incoming branch" and the other is labeled the "continuing branch" (Krone and Neuhauser 1997). Only one of these will be included in the gene genealogy of the sample. There are four possible values for the allelic states (I, C) of the incoming and continuing branches—(A_1, A_1), (A_1, A_2), (A_2, A_1), and (A_2, A_2)—but these are not specified in the initial construction of the graph.

When the ultimate ancestor is reached, its type is drawn from the distribution $h(x)$ and the lineages are traced forward in time to the present-day sample, changing type as needed when mutation events are encountered. Branching events are resolved as follows. If $I = A_2$, then the incoming branch replaces the continuing branch, and the allelic state of the descendent lineage is A_2. If $I = A_1$, then the incoming branch does not replace the continuing branch. Instead, the descendent lineage inherits the state and ancestry of the continuing branch. Nonancestral lineages are discarded, and the result is a sample drawn from the joint distribution of allelic states and gene genealogies (Krone and Neuhauser 1997).

The utility of the ancestral selection graph is not that it generates samples with allelic states in proportion to their probabilities, this is known and given by equation (2), but rather that it provides a tool for investigating the properties of gene genealogies under selection. However, the presence of branching events makes analysis and simulation difficult. Fortunately, the ancestral selection graph can also be used to model the ancestry of a sample conditional on allelic types (Slade 2000a, 2000b), and in this case, the problem of multiplying ancestral lineages is not so severe. Following the work of Slade (2000b), Fearnhead (2002), Stephens and Donnelly (2003), and Baake and Bialowons (2008), it is possible to describe a conditional ancestral selection graph in which branching events are minimized and in which superfluous lineages can be discarded upon mutation.

In this work, the conditional ancestral selection graph is used to investigate the analytical properties of gene genealogies of samples of known allelic type in the case when selection is very strong. This case has not yet been considered in the literature, likely due to the explosion of lineages which occurs in the unconditional ancestral selection graph when the selection parameter σ is very large. In addition, simulations show a relatively small effect of selection on gene genealogies of random samples (Golding 1997; Neuhauser and Krone 1997; Przeworski et al. 1999) and of samples of known allelic type when selection is moderate (Slade 2000a, 2000b), so it is of interest to investigate a case where selection should have a dramatic effect on the gene genealogy.

In the limit $\sigma \to \infty$, the ancestry of a random sample or a sample containing at most one deleterious allele (A_1) is shown to be neutral. In contrast, genealogies of samples containing more than one strongly deleterious allele are very different than neutral genealogies. Their structure can be described and is identical to that of a soft selective sweep. That is, the distribution of the number of independently derived, deleterious mutant alleles in the sample is given by the Ewens sampling formula (Ewens 1972). Interestingly, this is identical to the result for a different limiting model, of strong mutation–selection balance (Hartl and Campbell 1982; Sawyer 1983), that Reich and Lander (2001) used in their interpretation of allelic diversity of human disease. Simulations show a rapid approach to limiting analytical predictions, occurring between about $\sigma = 1$ and $\sigma = 100$.

Methods and Results

The notion of "real" and "virtual" lineages (Krone and Neuhauser 1997) is important in describing a conditional ancestral selection graph in which the size of the graph is minimized. Each lineage in the conditional ancestral process carries an allelic type. Thus, in contrast to the unconditional graph, it is possible to resolve branching events when they occur and to know which lineages are ancestral to the sample and which are not. Real lineages are ones that are ancestral to the sample. However, each branching event introduces a virtual lineage, and in general, the gene genealogy of the sample depends on the numbers and types of all real and virtual lineages. The size of the graph can be reduced by recognizing that some virtual lineages, in fact, do not affect the gene genealogy.

The conditional ancestral selection graph is derived from the fundamental recursive equation (Krone and Neuhauser 1997; Slade 2000a; Fearnhead 2002) for the probability that an ordered set of lineages is composed of n_1 real and v_1 virtual lineages of allelic type A_1 and n_2 real and v_2 virtual lineages of allelic type A_2. This state is denoted (n_1, n_2, v_1, v_2), so that the sample itself would be represented as $(n_1, n_2, 0, 0)$. The basic approach is to condition on the first step back in the ancestry of $n = n_1 + n_2 + v_1 + v_2$ lineages in the exchangeable process (the first layer of the ancestral selection graph) and to consider which patterns of ancestral states would produce the configuration (n_1, n_2, v_1, v_2) given each possible event. Equation (A1) in the Appendix gives the basic recursion. It is a straightforward application of ideas that are discussed in detail elsewhere (Krone and Neuhauser 1997; Slade 2000b; Fearnhead 2002; Stephens and Donnelly 2003; Baake and Bialowons 2008), but the Appendix also includes a discussion of each term.

The probabilities in equation (A1) can be computed by a simple extension of equation (2),

$$p(n_1, n_2, v_1, v_2) = \int_0^1 x^{n_1+v_1}(1-x)^{n_2+v_2}h(x)\mathrm{d}x. \quad (3)$$

Thus, although recursive equations like equation (A1) may be used to compute sample probabilities or likelihoods (Griffiths and Tavaré 1994a, 1994b), the interest in equation (A1) here is that it offers a way to study gene genealogies. Equation (A1) is conditioned on allelic types but is derived from the exchangeable ancestral process with total rate equal to $(n + v)\ (\theta + \sigma + n + v - 1)/2$, where $n = n_1 + n_2$ and $v = v_1 + v_2$. A reduced conditional ancestral process is possible because two kinds of events allow a virtual lineage to be discarded and may be filtered out of the process (Slade 2000b; Fearnhead 2002). Baake and Bialowons (2008) provide an illuminating discussion of these simplifications and their interpretations.

Slade (2000b) found that if a branching event occurs in which the incoming branch has type $I = A_2$ and the ancestral lineages not involved in the event all have the correct allelic types to produce the configuration (n_1, n_2, v_1, v_2), then it is unnecessary to create a new virtual lineage with state C. In particular, both $C = A_1$ and $C = A_2$ would yield the correct allelic types of the descendent lineages. Algebraically, these two possibilities can be collected and the simplification $p(n_1, n_2, v_1 + 1, v_2) + p(n_1, n_2, v_1, v_2 + 1) = p(n_1, n_2, v_1, v_2)$ may be applied in lines six and eight of equation (A1).

Fearnhead (2002) showed, similarly, that when a mutation event occurs on a virtual lineage, that lineage may be discarded. This follows from the assumption of parent-independent mutation, in which the parental allele may be of either type. Algebraically, $p(n_1, n_2, v_1, v_2) + p(n_1, n_2, v_1 - 1, v_2 + 1) = p(n_1, n_2, v_1 - 1, v_2)$ and $p(n_1, n_2, v_1 + 1, v_2 - 1) + p(n_1, n_2, v_1, v_2) = p(n_1, n_2, v_1, v_2 - 1)$, which may be applied in lines three and four, respectively, of equation (A1). Note that the corresponding algebraic simplifications will not be used in the first two lines of equation (A1) because the specific objects of study here are the allelic states of, and relationships among, the real lineages ancestral to the sample.

An ancestral process conditional on the allelic types is obtained by implementing these simplifications in equation (A1), collecting all the terms involving $p(n_1, n_2, v_1, v_2)$ on the left-hand side, then dividing both sides by the result. The ancestral process thus obtained is akin to the one described by Stephens and Donnelly (2003) but minimizes the number of virtual branches that need to be added to the graph. The total rate of events is given by

$$\lambda_{\boldsymbol{n},\boldsymbol{v}} = \binom{n+v}{2} + \frac{\theta}{2}(\alpha_2 n_1 + \alpha_1 n_2 + v_1 + v_2) + \frac{\sigma}{2}(n_1 + v_1), \quad (4)$$

in which $\boldsymbol{n} = (n_1, n_2)$ and $\boldsymbol{v} = (v_1, v_2)$ and again $n = n_1 + n_2$ and $v = v_1 + v_2$. The result of these manipulations to equation (A1) is

$$\begin{aligned} 1 = {} & \frac{\binom{n_1}{2}}{\lambda_{n,v}}\frac{p(n_1-1,n_2,v_1,v_2)}{p(n_1,n_2,v_1,v_2)} + \frac{\binom{n_2}{2}}{\lambda_{n,v}}\frac{p(n_1,n_2-1,v_1,v_2)}{p(n_1,n_2,v_1,v_2)} \\ & + \frac{\frac{\theta}{2}\alpha_1 n_1}{\lambda_{n,v}}\frac{p(n_1-1,n_2+1,v_1,v_2)}{p(n_1,n_2,v_1,v_2)} + \frac{\frac{\theta}{2}\alpha_2 n_2}{\lambda_{n,v}}\frac{p(n_1+1,n_2-1,v_1,v_2)}{p(n_1,n_2,v_1,v_2)} \\ & + \frac{\frac{\sigma}{2}(n_1+n_2+v_1)}{\lambda_{n,v}}\frac{p(n_1,n_2,v_1+1,v_2)}{p(n_1,n_2,v_1,v_2)} \\ & + \frac{\binom{v_1}{2}}{\frac{\theta}{2}\alpha_1 v_1 + n_1 v_1 + \lambda_{n,v}}\frac{p(n_1,n_2,v_1-1,v_2)}{p(n_1,n_2,v_1,v_2)}. \end{aligned} \quad (5)$$

The six terms on the right-hand side above are the probabilities of coalescence between two real A_1 lineages, coalescence between two real A_2 lineages, an $A_1 \rightarrow A_2$ mutation event on a real lineage, an $A_2 \rightarrow A_1$ mutation event on a real lineage, a branching event on any lineage in which the ancestor is always a virtual A_1 lineage, and loss of a virtual A_1 lineage by mutation or coalescence. The conditional ancestral process is a Markov jump chain which remains in state $(\boldsymbol{n}, \boldsymbol{v})$ for an exponentially distributed length of time, with mean $1/\lambda_{\boldsymbol{n},\boldsymbol{v}}$, and then jumps to a new state according to these six probabilities. This is a straightforward generalization of the algorithms in Fearnhead (2002) and Baake and Bialowons (2008) to samples of size larger than one. Note that no virtual A_2 lineages are produced (Slade 2000b). Because the sample also begins without any virtual lineages, v_2 will always be zero. Therefore, it is typically omitted in equations below.

The only case in which equation (5) leads easily to analytical results is when $\sigma = 0$. In this neutral case, no virtual lineages of either type are produced, and the sample may be represented simply by (n_1, n_2). For small samples, the Markov jump chain gives simple systems of equations that can be solved for quantities of interest. For example, the expected times to common ancestry for the three possible samples of size two can be shown to be

$$\begin{aligned} E[T_{(2,0)}] &= 1 - \frac{\theta\alpha_2}{\theta\alpha_1 + 1}\frac{1}{\theta + 1}, \\ E[T_{(1,1)}] &= 1 + \frac{1}{\theta + 1}, \\ E[T_{(0,2)}] &= 1 - \frac{\theta\alpha_1}{\theta\alpha_2 + 1}\frac{1}{\theta + 1}. \end{aligned} \tag{6}$$

The expected time to common ancestry for a sample of one A_1 allele and one A_2 allele, (1, 1), is greater than the standard neutral prediction of one because at least one mutation must occur before the two lineages can coalesce. The expected intraallelic coalescence times are, correspondingly, less than one. It can be checked that a random sample has exactly the neutral expectation of one by averaging the above formulas, weighted by the probabilities of each type of (ordered) sample from equation (2).

Gene Genealogies under Strong Selection

The transition probabilities in the conditional ancestral process described above, and the ones given by equation (A1) in the Appendix, depend on ratios of sampling probabilities. As noted by Stephens and Donnelly (2003), some time may be saved in performing simulations because the constant B in equation (1) cancels in these ratios and, thus, does not need to be calculated. Further, the presence of these ratios changes the probabilities of events substantially when selection is strong because unfit (A_1) alleles are unlikely to be sampled. For example, the rate of branching is reduced in equation (5) because the ratio $p(n_1, n_2, v_1 + 1)/p(n_1, n_2, v_1)$ becomes very small when σ is large. Fairly simple expressions for these ratios are available when σ is large, and this makes it possible to describe a limiting $\sigma \to \infty$ conditional ancestral process.

The analysis follows from a uniform asymptotic (large σ) expansion of Kummer's confluent hypergeometric function, which here is denoted ${}_1F_1[a;b; -\sigma]$. Specifically, if $a > 0$, $b > 0$, and $\sigma > 0$, which will all be true here, then from equations 3.1.2 and 4.1.2 in Slater (1960),

$$\int_0^1 x^{a-1}(1 - x)^{b-a-1}\mathrm{e}^{-\sigma x}\mathrm{d}x = \frac{\Gamma(a)\Gamma(b - a)}{\Gamma(b)}{}_1F_1[a; b; -\sigma] \tag{7}$$

$$= \Gamma(a)\sigma^{-a}\left(\sum_{n=0}^{L-1}\frac{(a)_n(1 + a - b)_n\sigma^{-n}}{n!} + O(\sigma^{-L})\right), \tag{8}$$

in which $(a)_n = a(a + 1) \ldots (a + n - 1)$ denotes the ascending factorial, with $(a)_0 = 1$.

For two sample configurations (n_1', n_2', v_1') and (n_1, n_2, v_1), let $a{=}a_1 + n_1' + v_1'$, $b{=}\theta + n_1' + n_2' + v_1'$, $c = \theta\alpha_1 + n_1 + v_1$, and $d = \theta + n_1 + n_2 + v_1$. Equation (8) gives

$$\frac{p(n_1', n_2', v_1')}{p(n_1, n_2, v_1)} = \frac{\int_0^1 x^{a-1}(1 - x)^{b-a-1}\mathrm{e}^{-\sigma x}\mathrm{d}x}{\int_0^1 x^{c-1}(1 - x)^{d-c-1}\mathrm{e}^{-\sigma x}\mathrm{d}x} \tag{9}$$

$$= \frac{\Gamma(\theta\alpha_1 + n_1' + v_1')}{\Gamma(\theta\alpha_1 + n_1 + v_1)}\sigma^{(n_1 + v_1) - (n_1' + v_1')}(1 + O(\sigma^{-1})). \tag{10}$$

Therefore, each additional A_1 allele, either real or virtual, decreases the sampling probability by a factor of order σ.

The rate $\lambda_{(\boldsymbol{n},\boldsymbol{v})}$ in equation (4) is the total rate of events in the conditional ancestral process, which occurs on a time scale proportional to N generations. Examination of $\lambda_{(\boldsymbol{n},\boldsymbol{v})}$ shows that the limiting conditional ancestral process for the present-day sample (n_1, n_2, 0) must be analyzed separately for $n_1 > 0$ and for $n_1 = 0$. In the first case, when there is at least one deleterious allele in the sample, the total rate of events depends linearly on σ. Then, when σ is large, the waiting time to an event will be of order σ^{-1}. In contrast, when $n_1 = 0$, so that the sample contains only fit alleles, the total rate of events is $\lambda_{\boldsymbol{n},\boldsymbol{v}}{=}n_2(n_2 - 1)/2 + \theta\alpha_1 n_2/2$. In this case, the waiting time to an event does not depend on σ. Instead, it is of order 1 when σ is large. This leads to a "separation of time scales" that is key to the analysis of equation (5) for large σ.

Separation of Times Scales: Fast Processes

In the first case, when $n_1 > 0$ (and $v_1 = 0$), equation (10) allows for the following simplification of equation (5). Taking the limit of equation (5) as $\sigma \to \infty$, or ignoring terms of order σ^{-1} and smaller, and simplifying give

$$1 = \frac{n_1 - 1}{\theta\alpha_1 + n_1 - 1} + \frac{\theta\alpha_1}{\theta\alpha_1 + n_1 - 1}. \tag{11}$$

The two terms on the right are the probability of a coalescent event between two A_1 lineages and the probability of a mutation event from A_1 to A_2. The probabilities of all other events are of order σ^{-1} or smaller because they either lead to the production of an additional unfit allele or simply because $\lambda_{\boldsymbol{n},\boldsymbol{v}}$ is of order σ. Note that in this ($n_1 > 0$) limiting process, all the lineages are real, no virtual lineages are produced, and no events occur among the A_2 lineages, if there are any in the sample.

The two probabilities in equation (11) are identical in form to those of a fundamental stochastic process in population genetics, namely the process of tracing the ancestry of a sample under the infinite-alleles mutation model that gives the Ewens sampling formula (Ewens 1972). Note that whichever event occurs above, the number of A_1 lineages decreases by one. Then equation (11) may be reapplied with $n_1 \to n_1 - 1$, and so on, continuing until no A_1 lineages remain. Counting the number, K, of ($A_1 \rightarrow A_2$) mutations leads to

$$P(K = k|n_1) = \frac{S_k^{(n_1)}(\theta\alpha_1)^k}{(\theta\alpha_1)_{n_1}} \tag{12}$$

for the probability that the n_1, A_1 alleles in the sample are descended from k A_2 alleles, where $S_k^{(n_1)}$ is an unsigned Stirling number of the first kind. Equation (12) is identical to the probability function for the number of alleles in the Ewens sampling formula (Ewens 1972). The full Ewens sampling formula, with mutation parameter $\theta\alpha_1$, gives the probability function for the numbers of A_1 descendants in the sample of each of these k ancestral A_2 alleles. As mentioned above, this result is identical to the result for soft selective sweeps; for example, see Pennings and Hermisson (2006a).

Again, the amount of time this takes will be of order σ^{-1}, which is negligible on the coalescent time scale of the ancestral selection graph, where one unit of time is proportional to N generations ($N^2/2$ steps in the discrete Moran model). Thus, a sample containing n_1 copies of A_1, where $n_1 > 0$ and n_2 copies of A_2 will quickly be converted into an ancestral sample of $K + n_2$ real A_2 lineages, where K is a random variable with $1 < K < n_1$ and the probability function $P(K = k|n_1)$ above.

Separation of Times Scales: Slow Processes

Of course, the above is only part of the ancestry of the sample. It is still necessary to trace the ancestry of the resulting $k + n_2$ real A_2 lineages back to their most recent common ancestor. For an ancestral sample of this sort, or for a present-day sample containing only fit alleles, a different limiting process arises. Without loss of generality, k may be omitted for simplicity, and the sample may be represented as ($n_1 = 0$, $n_2 > 0$, $v_1 = 0$). In this case, using equations (10) and (5) and taking the limit, or ignoring terms of order σ^{-1} and smaller, gives

$$1 = \frac{n_2 - 1}{\theta\alpha_1 + n_2 - 1} + \frac{\theta\alpha_1}{\theta\alpha_1 + n_2 - 1}.$$

The first term on the right is the probability of a coalescent event between two A_2 lineages, but now the second term corresponds to the fifth term on the right-hand side of equation (5) and is the probability of a branching event, in particular the production of a single virtual A_1 lineage.

The creation of this virtual A_1 lineage induces a third case, similar to the case $n_1 > 0$ above, but in which a different type of "fast" event is possible: the annihilation of the virtual lineage by mutation. When the ancestral configuration is ($n_1 = 0$, n_2, $v_1 = 1$), the total rate of events is again of order σ. An analysis like those above shows that the last term in equation (5) is $1 + O(\sigma^{-1})$, and all other terms are $O(\sigma^{-1})$. In the limit $\sigma \rightarrow \infty$, the virtual lineage is annihilated with probability equal to one, and this happens in a negligible amount of time. The sample thus reverts immediately to state ($n_1 = 0$, n_2, $v_1 = 0$) and the above "slow" process resumes.

This shows that a present-day sample or ancestral configuration comprised only of n_2 copies of the fit allele, A_2, undergoes a filtered ancestral process in which branching events occur, but the resulting virtual A_1 lineages are instantly removed. Eventually, with probability equal to one in the limit $\sigma \rightarrow \infty$, a coalescent event will occur between two of the n_2 A_2 alleles. The waiting time to this event is exponentially distributed with rate equal to the total rate of events times the probability that the event is a coalescent event or

$$\left(\binom{n_2}{2} + \frac{\theta\alpha_1}{2}n_2\right)\frac{n_2 - 1}{\theta\alpha_1 + n_2 - 1} = \binom{n_2}{2}.$$

Therefore, the ancestry of a sample containing only fit alleles is given by the standard neutral coalescent. The ancestry of a random sample should also be neutral because the probability that a random sample contains any A_1 alleles is of order σ^{-1}.

Comparing Analytical Predictions to Simulations

In the limiting ancestral process, analytical predictions can be made for any quantity of interest simply by conditioning on K, the number of A_2 ancestors of the n_1 copies of allele A_1 in the sample. For example, the total length of the gene genealogy of the sample (n_1, n_2) is given by

$$E[T_{\text{total}}] = 2\sum_{k=1}^{n_1} P(K = k|n_1)\sum_{j=1}^{n_2+k-1}\frac{1}{j}, \tag{13}$$

which can be recognized as Watterson's (1975) expected value averaged over all possible ancestral samples. Because $1 \leq K \leq n_1$, the expected length of the gene genealogy above is less than or equal to the neutral expectation for a sample of size $n = n_1 + n_2$. Due to equation (12), it will be greatest when θ is large, so that K is equal to n_1 with high probability. When θ is small, $E[T_{\text{total}}]$ will be close to the neutral expected value for a sample of size $n = 1 + n_2$.

Further, let T_i be the time during which there are i lineages ancestral to the sample. Under the standard neutral coalescent, T_i has expectation $2/(i(i-1))$, and i ranges from 2 to n. Under the limiting conditional ancestral selection graph, the expected value is

$$E[T_i] = \frac{2}{i(i-1)}\sum_{k=\max(1,i-n_2)}^{n_1} P(K = k|n_1), \tag{14}$$

given that the starting sample is (n_1, n_2). Equation (14) is the usual expectation, multiplied by the probability that the sample includes a period during which there are i lineages. Note that for i such that $2 \leq i \leq n_2$, the expected value of T_i for any sample is given exactly by the neutral expected value. For samples in which $n_1 > 0$, the expected value of T_{n_2+1} is also given by the neutral expected value because there must be at least one A_2 ancestor of the n_1 A_1 alleles. On the other hand, the expected value of T_i may be much smaller than the neutral expected value for values of i such that $n_2 + 1 < i \leq n_1 + n_2$ because the fast coalescence/mutation process described above may cause some coalescent

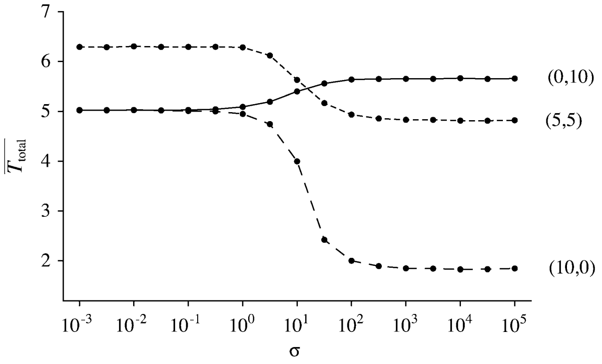

FIG. 1.—The average total length of the gene genealogy of a sample of size 10, for three different allelic configurations, (n_1, n_2), and over a broad range of the scaled selection parameter σ. In all cases, θ = 1.0, α_1 = α_2 = 0.5, and for each point, the average is taken over 200,000 simulation replicates.

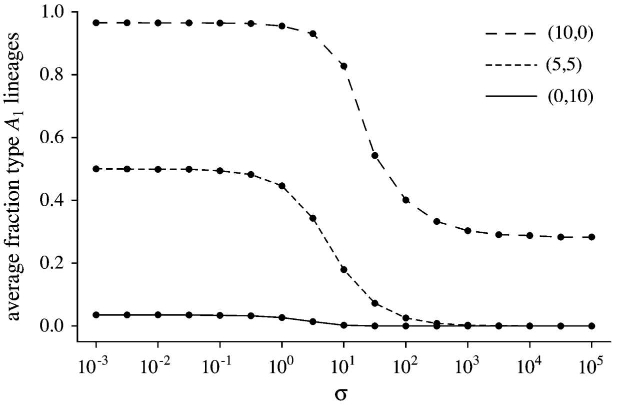

FIG. 2.—The average value of the fraction of the gene genealogy comprised of A_1 allelic lineages for a sample of size 10, for three different allelic configurations, (n_1, n_2), as a function of the scaled selection parameter σ. In all cases, θ = 1.0, α_1 = α_2 = 0.5, and results are averaged over 200,000 simulation replicates.

intervals to be skipped with high probability. That is, given a value of $K = k$, T_i will be equal to zero for $n_2 + k < i \leq n_1 + n_2$.

The process described by equation (5) is straightforward to simulate. The simulations presented below were done in Mathematica (Wolfram 1999), version 5.2. The Mathematica notebook used to generate the results is available from the author upon request. A single simulation run begins with a sample (n_1, n_2, 0) and ends in a random configuration, either (1, 0, ν_1) or (0, 1, ν_1), in which there is only one real lineage, which is the most recent common ancestor of the sample. Exponential waiting times with means $1/\lambda_{(\boldsymbol{n},\boldsymbol{\nu})}$ are generated conditional upon each configuration of ancestral lineages encountered, and transitions are implemented stochastically according to the probabilities in equation (5). This algorithm is very similar to that of Stephens and Donnelly (2003), the only difference being that the simplifications due to Slade (2000b) and Fearnhead (2002) have been implemented here but were not used in Stephens and Donnelly (2003).

The two main aims of the simulations are to assess the convergence of various quantities to the limiting $\sigma \rightarrow \infty$ predictions, such as equations (13) and (14) above, and to illustrate how conditional gene genealogies depend on σ. The program was also tested against available analytical results. In particular, with σ = 0, the average pairwise coalescence times become closer and closer to the predictions of equation (6) as the number of replicates increases (results not shown). Further, when σ is very large, the simulation conforms to limiting ($\sigma \rightarrow \infty$) analytical predictions. Simulation results change monotonically between these two extremes, but no analytical predictions are available for arbitrary σ.

Results are presented for samples of size 10, for three different sampling configurations: a sample containing only A_1 alleles (10, 0), a sample split evenly between A_1 and A_2 alleles (5, 5), and a sample containing only A_2 alleles (0, 10). In all cases, θ = 1 and α_1 = α_2 = 0.5, and 200,000 replicates were done to produce each result. The exception to this is figure 4*C*, which required a larger number of replicates. In this case, 1 million replicates were done for each combination of parameters. The average values of four quantities were computed for 17 values of σ, from 10^{-3} to 10^{5}, evenly spaced on a log scale. The speed of these simulations is greatly improved by tabling values of equation (7) for each value of σ (and θ, α_1, α_2). The 17 million replicates that produced figure 4*C* took 22 h on a Macintosh 1.5 GHz PowerPC G4.

Simulation Results

Figure 1 shows the average total length of the gene genealogy of the sample, that is, the sum of the lengths of all the real branches in the ancestry, back to the most recent common ancestor of the sample. On the left, as σ decreases, the values converge on neutral expectations. In this case, because of the dependence on mutation and consistent with equation (6), the value for the sample (5, 5) is larger than the values for the samples (10, 0) and (0, 10). On the right, the values fit the limiting $\sigma \rightarrow \infty$ predictions well, which in this case are given by 1.82, 4.82, and 5.66 for samples (10, 0), (5, 5), and (0, 10), respectively. The most rapid change in values occurs between σ = 1 and σ = 100. When $\sigma < 0.1$, the behavior is very close to that of the neutral model, and when $\sigma > 1000$, the behavior is very close to that of the limiting $\sigma \rightarrow \infty$ model. Interestingly, these "cutoffs" of σ appear insensitive to the value of θ (results not shown).

Figure 2 shows the average fraction of the gene genealogy made up of lineages of type A_1. For each simulation replicate, the total length of real A_1 lineages was computed and divided by the total length of all real ($A_1 + A_2$) lineages for that replicate. These values were averaged over all simulation replicates. On the left, when σ is small, the value for the sample (5, 5) is close to one-half because mutation is symmetric and the sample configuration is also symmetric. The unbalanced samples (10, 0) and (0, 10) have values close to 1 and 0, respectively, when σ is small. Note that, when θ becomes large, the effect of the sample

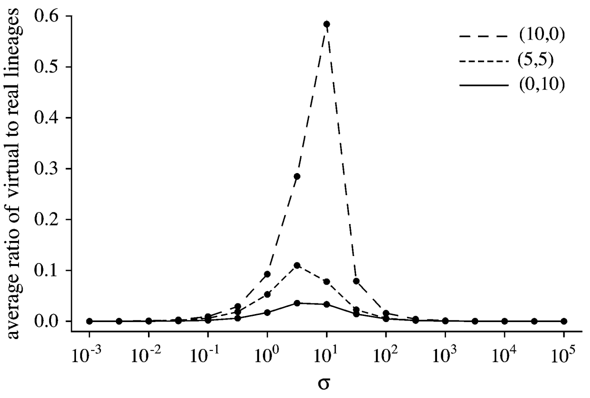

FIG. 3.—The average value of the ratio of the total length of virtual branches to the total length of the gene genealogy (i.e., real branches) for a sample of size 10, for three different allelic configurations, (n_1, n_2), and over a broad range of the scaled selection parameter σ. In all cases, θ = 1.0, $\alpha_1 = \alpha_2 = 0.5$, and the average is taken over 200,000 simulation replicates.

configuration disappears and the values for all three samples converge on one-half or on α_1 if mutation is asymmetric (results not shown).

On the right in Figure 2, when σ is large, the average fraction of A_1 lineages decreases to zero for any sample that contains at least one A_2 allele, in this case samples (5, 5) and (0, 10). This illustrates that any A_1 alleles in the sample will disappear rapidly as a result of the fast process described above as they coalesce or get converted to A_2 alleles by mutation. In the special case that only A_1 alleles are sampled, there is a chance, equal to $P(K = 1|n_1)$ in the limit, that all n_1 copies will coalesce before the first $A_1 \rightarrow A_2$ mutation event. If this occurs, then the entire gene genealogy will be composed of A_1 lineages, and the fraction will be equal to one, even though the total length of the tree may be very small. If $K > 1$, the fraction of A_1 lineages will be negligible for the reason just discussed. Thus, the average fraction for the sample (10, 0) converges on $P(K = 1|n_1 = 10) \approx 0.28$ as σ increases in Figure 2.

Figure 3 shows the average ratio of virtual branches to real branches in the ancestry of the sample back to the most recent common ancestor. As in Figure 2, the ratio is taken for each replicate and then averaged across replicates. For any sample, the largest numbers of virtual branches are generated when σ is slightly less than 10. The total time of virtual branches is very small when σ is either small or large. The intuition behind this is clear when σ is small: selection is weak and few branching events occur. On the other hand, when σ is large, virtual A_1 branches will be created but then will be annihilated quickly by mutation. There is also a strong effect of sample type on the total length of virtual branches, with samples containing more copies of A_2 having smaller numbers of virtual branches. Figure 3 demonstrates that the conditional ancestral selection graph, with transitions given by equation (5), does not suffer from the explosion of virtual lineages that plagues the unconditional ancestral selection graph. For these samples and parameter values, virtual branches never outnumber real branches, at least on average.

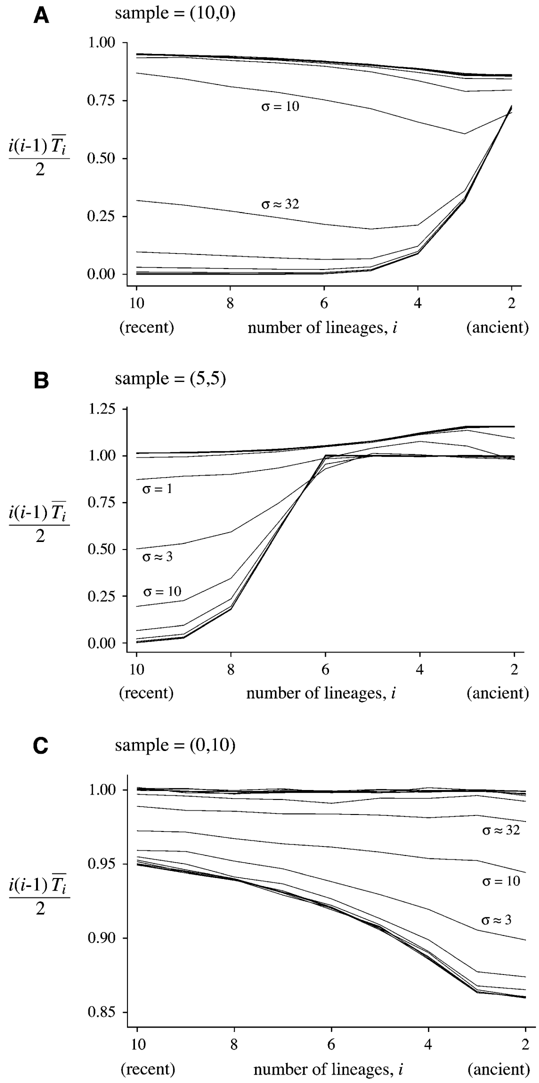

FIG. 4.—The ratio of the average time during which there are i ancestral lineages to the expected value of the same quantity under the standard neutral coalescent. The parameters are the same as in figures 1–3, but the results for the three different allelic configurations, (n_1, n_2), are presented separately in panels (*A*), (*B*), and (*C*). To obtain smooth curves in (*C*), 1 million simulation replicates were performed for each point.

Figure 4 shows how the average time during which there are i lineages ancestral to the sample ($2 \leq i \leq n$) compares to the neutral expectation for a random sample, $E[T_i] = 2/(i(i-1))$. Thus, the plots are similar to skyline plots (Strimmer and Pybus 2001), except that the horizontal axis here is the number of ancestral lineages (i) rather than time before the present. Values close to 1 indicate coalescence times close to neutral expectations. The samples and parameter values are the same as in figures 1–3. Each line shows the results for a single value of σ, and again, there are 17 of these ranging from 10^{-3} to 10^{5}. The different curves for small σ are difficult to distinguish, as are those for large σ. Thus, figure 4 displays the same sharp transition between the behaviors for small and large σ seen in figures 1–3. For reference, curves that fall in the steepest part of the transition are labeled by their σ values.

The bundle of lines at the top in figure 4*A* displays the behavior under neutrality but conditional on the sample being of one allelic type (A_1) and with $\theta = 1$. Thus, the bundle of lines sits below one. As might be expected, a sample of all A_2 displays the same behavior when σ is small. This is shown in figure 4*C*, but in this case, the bundle of lines is at the bottom of the plot rather than at the top. Thus, increasing σ has the opposite effect on relative coalescence times for the sample (0, 10) as it does for the sample (10, 0). In the case of (0, 10), shown in figure 4*C*, increasing σ leads to more and more neutral looking gene genealogies, for the reasons discussed above. In the case of (10, 0), shown in figure 4*A*, increasing σ leads to very short times for first $10 - K$ coalescent intervals looking back ($i = 10, \ldots, K + 1$), where K is the random variable with distribution given by equation (12). The bundle of lines at the bottom of figure 4*A* sits right on top of the limiting predictions obtained from equation (14).

Figure 4*B* shows the behavior for an evenly split sample. In this case, the bundle of lines for small σ tends to lie above one, especially for the more ancient coalescent intervals (small i), because the sample requires at least one mutation before the most recent common ancestor can be reached. As σ increases, the average coalescence times converge on the limiting predictions of equation (14). In this case, the $n_1 = 5$ copies of A_1 in the sample coalesce (and mutate) rapidly into K lineages of type A_2, then the subsequent ancestry of the remaining $n_2 + K$ lineages is neutral. As $K \geq 1$, there is at least one A_2 ancestor of the five A_1 alleles, so the times T_i are given by the neutral model for $i = 2, 3, 4, 5$, and 6.

Discussion

The ancestral selection graph is a mathematical tool for studying gene genealogies of alleles under selection. Due to the complicated nature of ancestral processes with selection, few analytical results are available. Krone and Neuhauser (1997) proved that the ancestral selection graph collapses to the neutral coalescent when $\theta = 0$ or $\sigma = 0$, and Neuhauser and Krone (1997) obtained the same result when $\theta \to \infty$ for a given σ. Here, using the conditional ancestral selection graph, it was shown that neutral gene genealogies also dominate when $\sigma \to \infty$. However, gene genealogies of samples which happen to contain some number of deleterious alleles are very different than neutral genealogies. Their ancestries consist of a two-phase process, in which the deleterious (A_1) alleles in the sample quickly coalesce and mutate into a random number of advantageous (A_2) alleles, and then the ancestry of those alleles and the rest of the sample is given by the neutral coalescent process. Interestingly, the Ewens sampling formula describes the result of the fast process.

As mentioned above, the results presented here are fundamentally similar to those for soft selective sweeps by recurrent mutation. Pennings and Hermisson (2006a) describe how the Ewens sampling formula gives the probability function for the number of independent ancestral alleles of a sample of size n taken at the end of the sweep. Pennings and Hermisson (2006a) assumed that each copy of the advantageous allele arises uniquely via one-way mutation from the background allele and that the allele-frequency trajectory of the advantageous allele follows the standard prediction for strong positive selection. However, they argue that the result should not depend on the shape of the allele-frequency trajectory as long as the sweep occurs quickly.

It is not surprising that the same result should be found, as it was here, for a strongly deleterious allele that happens to reach a large enough frequency to be observed in a sample. Such an allele would have arrived at high frequency by a sweep-like process (see further discussion below). Given the starting and ending frequency of an allele, some properties of the trajectory do not depend on whether the allele is advantageous or deleterious, but only on the absolute value of selection parameter, see Sections 4.6 and 5.4 in Ewens (2004). The comparison of the present results to those of Pennings and Hermisson (2006a) is of value because it shows that deleterious alleles which are observed in a sample are no more or less likely to be derived from independent mutations than advantageous alleles which have undergone a sweep.

The results presented here also have implications for the interpretation of allelic diversity at human disease loci and make a contribution to ongoing modeling efforts in that area (Di Rienzo 2006). Hartl and Campbell (1982) studied a model of classical mutation–selection balance, in which the frequency of alleles that cause a simple Mendelian disorder is held constant over time by strong mutation and selection and found an identical role for the Ewens sampling formula as that discovered here. Their model exists in the limit as $N \to \infty$ with $\theta \to \infty$ and $\sigma \to \infty$ but θ/σ constant (Sawyer 1983), meaning that mutation is a strong force that keeps the deleterious allele at an appreciable frequency in the population despite strong deleterious selection.

Pritchard (2001) considered similar ideas in the context of complex diseases, where selection on particular alleles that cause susceptibility is expected to be weaker. He discussed the importance of the mutation–selection–drift equilibrium (eq.1) in interpreting the diversity of alleles at each locus that contributes to a complex disease. Again, equation (1) holds in the limit as $N \to \infty$ with θ and σ constant. Pritchard (2001) used simulations to study patterns of allelic diversity and estimated that the scaled rate of mutation to deleterious alleles—denoted $\theta\alpha_1$ here and β_S in Pritchard (2001)—is between about 0.1 and 5 for a typical locus.

Using a completely different approach and set of assumptions, Slatkin and Rannala (1997) also showed that the diversity of alleles at a disease locus should follow the Ewens sampling formula. In particular, Slatkin and Rannala (1997) used a birth–death process, forward in time, in which every copy of the allele reproduced independently. Their mathematical analysis did not require the alleles to be deleterious, but they focused on this case because they were interested in applications to disease. The method and results of Slatkin and Rannala (1997) should be valid over a fairly broad range of parameter values, provided that the overall frequency of the alleles is small.

Slatkin and Rannala (1997) also pointed out that the standard homozygosity test (Watterson 1978; Slatkin 1994, 1996) could be used to detect deviations from the model, including

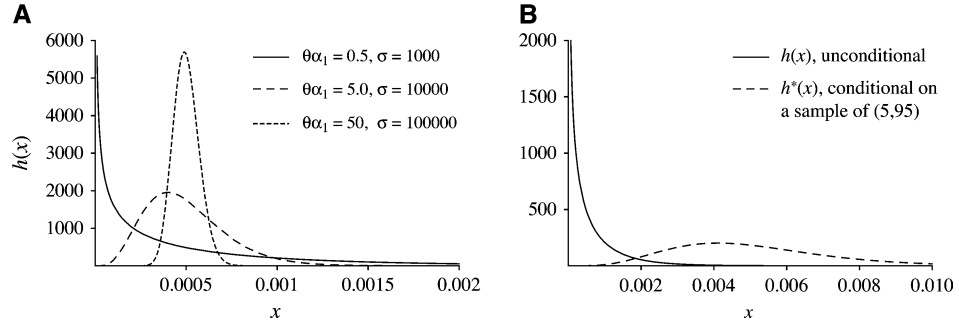

FIG. 5.—(*A*) Plots of $h(x)$, given in equation (1), for three different sets of parameters. In all three cases, the average frequency of the deleterious allele is equal to 1/2000. (*B*) Plots of $h(x)$ and $h^*(x)$, given in equation (15) and conditional on observing $n_1 = 5$ and $n_2 = 95$ in a sample of size 100, when $\theta = 1$, $\sigma = 1000$, and $\alpha_1 = \alpha_2 = 0.5$; these are the same parameter values as for the solid curve in (*A*).

different rates of mutation to different alleles, differential selection, differential penetrance, and changes in population size (in particular, growth). They rejected the null model for the *BRCA*1 locus, which is implicated in early-onset breast cancer, and the factor VIII locus, which is associated with hemophilia A. They concluded that population growth could explain the deviations at both loci. Beginning instead with the model of Hartl and Campbell (1982), Reich and Lander (2001) extended this conclusion about growth to a general observation among many loci: that diseases in higher frequencies tend to have a simpler pattern of diversity, with one most common allele. For further discussion, see Pritchard and Cox (2002) and Di Rienzo (2006).

Taken together, the results of Hartl and Campbell (1982), Slatkin and Rannala (1997), Pennings and Hermisson (2006a), and those presented here support the broad applicability of the Ewens sampling formula as a model of diversity among selectively equivalent alleles at a locus without recombination. To illustrate some of the above ideas and to get a sense of the domain of application (to deleterious/disease alleles) of the present model compared with the model Hartl and Campbell (1982), consider figure 5. The model of Slatkin and Rannala (1997) will not be considered in this context because its assumptions are implicit, about the age and frequencies of alleles, rather than explicit, about the parameters. Figure 5*A* shows three distributions of x, the frequency of the deleterious allele A_1, all of which correspond to a disease whose average frequency in the population is 1/2000, but among which there are very different levels of variation around this average. Note that $h(x)$, given in equation (1), may be interpreted as the relative amount of time the population spends with the frequency of A_1 equal to x. A different but related way to think of $h(x)$ is that it represents the relative chance that the current frequency of A_1 in the population is equal to x.

The shape of $h(x)$ depends on $\theta\alpha_1$, $\theta\alpha_2$, and σ. It is L shaped when both $\theta\alpha_1$ and $\theta\alpha_2$ are less than or equal to one and has a nonzero mode if the mutation rate to the less-frequent allele (here $\theta\alpha_1$) is greater than one. The solid curve in Figure 5*A* corresponds to one set of parameters used in the simulations presented above. Recall that when $\sigma = 1000$, as for the solid curve, the simulation results were very close the limiting predictions for strong selection, here meaning $\sigma \to \infty$ for a given θ. It is clear that, for this solid curve, the population is not particularly likely to have an x close to its expected value of 1/2000 and that much of the time the frequency of A_1 is close to zero. On the other hand, as θ and σ increase, but θ/σ remains constant, the distribution becomes more and more concentrated on the expected value. The finely dashed curve in figure 5*A* represents a case in which the model of Hartl and Campbell (1982) would be appropriate. Amazingly, despite quite different assumptions, both models predict the same sampling distribution of alleles.

Figure 5*B* displays the same solid curve that appears in figure 5*A*, but over a wider range of x. Again, in this case, there is only a 1/2000 chance of sampling a deleterious allele. Also included in figure 5*B* is the distribution of x, conditional on observing five deleterious alleles in a random sample of size 100. The general formula for this posterior distribution of allele frequencies is

$$h^*(x) = \frac{x^{n_1}(1-x)^{n_2}h(x)}{p(n_1,n_2)}. \qquad (15)$$

The dashed curve in figure 5*B* shows that if such a sample is observed, in which the frequency of A_1 is 5% rather than the expected 0.05%, then the population is likely to be in a very uncommon state, where x is far off in the tail of its predicted distribution $h(x)$. Because the population must spend most of its time with values of x in the bulk of $h(x)$, it follows that the frequency of A_1 only recently rose to such a high level. In keeping with the suggestion of Pennings and Hermisson (2006a)—that the Ewens sampling formula for soft sweeps should hold regardless of the sign of the selection coefficient and of the exact trajectory of the allele frequency—the conditional gene genealogy of a sample containing some number of deleterious alleles resembles closely that of a positive selective sweep that has gone only part way to completion. This highlights the possibility that some human diseases may be at higher frequencies than expected based on mutation rates and selection coefficients, simply stochastically.

Acknowledgments

I thank Joachim Hermisson, Tee Muirhead, Pleuni Pennings, and Monty Slatkin for helpful discussions of this work.

Appendix

In the following equation, the probability of the ordered sample (n_1, n_2, v_1, v_2) is computed by conditioning on the first step in the exchangeable ancestral process which underlies the ancestral selection graph (see text). The total rate of events is $(n + v)$ $(\theta + \sigma + n + v - 1)/2$. Fourteen different kinds of events are distinguished, based on whether they are mutation, coalescent, or branching events and on which lineages they affect. This is a special case of the general, multiallele processes described in Fearnhead (2002) and Stephens and Donnelly (2003). Similar equations, for unordered samples, can be found in Krone and Neuhauser (1997) and Slade (2000a).

$$\begin{aligned}
p\Big(n_1,n_2,v_1,v_2\Big) &= \frac{n_1\theta\alpha_1}{(n+v)(\theta+\sigma+n+v-1)}(p(n_1,n_2,v_1,v_2)+p(n_1-1,n_2+1,v_1,v_2))\\
&+\frac{n_2\theta\alpha_2}{(n+v)(\theta+\sigma+n+v-1)}(p(n_1+1,n_2-1,v_1,v_2)+p(n_1,n_2,v_1,v_2))\\
&+\frac{v_1\theta\alpha_1}{(n+v)(\theta+\sigma+n+v-1)}(p(n_1,n_2,v_1,v_2)+p(n_1,n_2,v_1-1,v_2+1))\\
&+\frac{v_2\theta\alpha_2}{(n+v)(\theta+\sigma+n+v-1)}(p(n_1,n_2,v_1+1,v_2-1)+p(n_1,n_2,v_1,v_2))\\
&+\frac{n_1\sigma}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1+1,v_2)\\
&+\frac{n_2\sigma}{(n+v)(\theta+\sigma+n+v-1)}(2p(n_1,n_2,v_1+1,v_2)+p(n_1,n_2,v_1,v_2+1))\\
&+\frac{v_1\sigma}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1+1,v_2)\\
&+\frac{v_2\sigma}{(n+v)(\theta+\sigma+n+v-1)}(2p(n_1,n_2,v_1+1,v_2)+p(n_1,n_2,v_1,v_2+1))\\
&+\frac{n_1(n_1-1)}{(n+v)(\theta+\sigma+n+v-1)}p(n_1-1,n_2,v_1,v_2)\\
&+\frac{n_2(n_2-1)}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2-1,v_1,v_2)\\
&+\frac{n_1v_1}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1-1,v_2)\\
&+\frac{n_2v_2}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1,v_2-1)\\
&+\frac{v_1(v_1-1)}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1-1,v_2)\\
&+\frac{v_2(v_2-1)}{(n+v)(\theta+\sigma+n+v-1)}p(n_1,n_2,v_1,v_2-1)
\end{aligned} \qquad . \qquad \text{(A1)}$$

Each term on the right-hand side of equation (A1) has the form $P\{\text{Event}\}P\{\text{Data|Event}\}$, where Data means the ordered sample (n_1, n_2, v_1, v_2), and the sum is taken only over Events that have a nonzero probability of producing the data. In particular, coalescent events between lineages whose allelic types in the data are different are omitted from equation (A1), as are mutation events in which the descendent lineage is not of the allelic type required by the data. There are four kinds of events among the 14 terms on the right-hand side of equation (A1): mutation events in which the descendant lineage is of the correct allelic type (terms 1–4), branching events in which the descendent lineage must be of allelic type A_1 (terms 5 and 7), branching events in which the descendent lineage must be of allelic type A_2 (terms 6 and 8), and coalescent events in which both descendent lineages must be of the same allelic type (terms 9–14).

The probabilities of each event are given by the fractions in each term and are computed in the usual way as the rate of each particular event divided by the total rate of events in the unconditional, exchangeable process. At the time of the first event, the lineages in the graph are a random sample from the equilibrium population (Donnelly and Kurtz 1999). The probabilities of the data given each event are computed by considering what type of ancestral sample would yield the data. For example, the lineages not involved in the event must simply have the same allelic state that they do in the data.

For mutation events, the ancestral sample would yield the data if it were identical to the data (in which case it is an empty mutation event) or if it contained one fewer allele of the type required for the mutant lineage by the data and one more of the other allelic type (in which case the mutation converts the lineage to the correct allelic type). In the case of branching events, all four possibilities for the allelic states of the incoming and continuing branches must be considered. For branching events in which the descendent lineage must be of allelic type A_1, only an ancestral sample in which the incoming (virtual) branch has allelic type A_1 (and the remaining lineages have the types required by the data) would yield the data. For branching events in which the descendent lineage must be of allelic type A_2, three cases of (I, C) could produce the data: (A_1, A_2), (A_2, A_1), and (A_2, A_2). In the first two cases, the ancestral sample possesses one additional (virtual) lineage of type A_1 relative to the data; hence, these are grouped together in equation (A1), whereas in the third case the ancestral sample possesses one additional (virtual) lineage of type A_2. Finally, coalescent events in which both descendent lineages are of the same allelic type will yield the data if the coalesced ancestral lineage is of the correct allelic type, so that ancestral sample contains one fewer of that allelic type than the descendent sample does.

Literature Cited

Abramowitz M, Stegun IA. 1964. Handbook of mathematical functions. New York: Dover Publications.

Aldous DJ. 1985. Exchangeability and related topics. In: Dold A, Eckmann B, editors. École d'Été de Probabilités de Saint-Flour XII—1983. Berlin (Germany): Springer-Verlag. p. 1–198 (Lecture notes in mathematics; vol. 1117).

Baake E, Bialowons R. Forthcoming. 2008. Ancestral processes with selection: branching and Moran models. In: Miekisz J, editor. Banach center publications. Warsaw (Poland): Institute of Mathematics, Polish Academy of Sciences.

Barton NH, Etheridge AM, Sturm AK. 2004. Coalescence in a random background. Ann Appl Probab. 14:754–785.

Cannings C. 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv Appl Probab. 6:260–290.

Di Rienzo A. 2006. Population genetics models on common diseases. Curr Opin Genet Dev. 16:630–636.

Donnelly P, Kurtz TG. 1999. Genealogical models for Fleming-Viot models with selection and recombination. Ann Appl Probab. 9:1091–1148.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. Theoret Popul Biol. 3:87–112.

Ewens WJ. 2004. Mathematical population genetics. Volume I: theoretical foundations. Berlin (Germany): Springer-Verlag.

Fearnhead P. 2002. The common ancestor at a nonneutral locus. J Appl Probab. 39:38–54.

Fearnhead P. 2006. Perfect simulation from nonneutral population genetic models: variable population size and population subdivision. Genetics. 174:1397–1406.

Golding GB. 1997. The effect of purifying selection on genealogies. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag. p. 271–285 (IMA volumes in mathematics and its applications; vol. 87).

Griffiths RC, Tavaré S. 1994a. Simulating probability distributions in the coalescent. Theoret Popul Biol. 46:131–159.

Griffiths RC, Tavaré S. 1994b. Ancestral inference in population genetics. Stat Sci. 9:307–319.

Hartl DL, Campbell RB. 1982. Allelic multiplicity in simple Mendelian disorders. Am J Hum Genet. 34:866–873.

Hermisson J, Pennings PS. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 169:2335–2352.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 37:203–217.

Kaplan NL, Darden T, Hudson RR. 1988. Coalescent process in models with selection. Genetics. 120:819–829.

Kimura M. 1955. Stochastic processes and the distribution of gene frequencies under natural selection. Cold Spring Harb Symp Quant Biol. 20:33–53.

Kingman JFC. 1982a. The coalescent. Stochastic Process Appl. 13:235–248.

Kingman JFC. 1982b. On the genealogy of large populations. J Appl Probab. 19A:27–43.

Kingman JFC. 1982c. Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F, editors. Exchangeability in probability and statistics. Amsterdam: North-Holland. p. 97–112.

Krone SM, Neuhauser C. 1997. Ancestral processes with selection. Theoret Popul Biol. 51:210–237.

Moran PAP. 1958. Random processes in genetics. Proc Camb Philos Soc. 54:60–71.

Moran PAP. 1962. Statistical processes of evolutionary theory. Oxford: Clarendon Press.

Neuhauser C. 1999. The ancestral graph and gene genealogy under frequency-dependent selection. Theoret Popul Biol. 56:203–214.

Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. Genetics. 145:519–534.

Pennings PS, Hermisson J. 2006a. Soft sweeps II: molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 23:1076–1084.

Pennings PS, Hermisson J. 2006b. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2:e186.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases?, Am J Hum Genet. 69:124–137.

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common-disease–common variant ... or not? Hum Mol Genet. 11:2417–2423.

Przeworski M, Charlesworth B, Wall JD. 1999. Genealogies and weak purifying selection. Mol Biol Evol. 16:264–1252.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet. 17:502–510.

Sawyer SA. 1983. A stability property of the Ewens sampling formula. J Appl Probab. 20:449–459.

Slade PF. 2000a. Simulation of selected genealogies. Theoret Popul Biol. 57:35–49.

Slade PF. 2000b. Most recent common ancestor distributions in genealogies under selection. Theoret Popul Biol. 58:291–305.

Slater LJ. 1960. Confluent hypergeometric functions. Cambridge: Cambridge University Press.

Slatkin M. 1994. An exact test for neutrality based on the Ewens sampling distribution. Genet Res. 64:71–74.

Slatkin M. 1996. A correction to the exact test based on the Ewens sampling distribution. Genet Res. 68:259–260.

Slatkin M, Rannala B. 1997. The sampling distribution of disease-associated alleles. Genetics. 147:1855–1861.

Stephens M, Donnelly P. 2003. Ancestral inference in population genetics models with selection. Aust N Z J Stat. 45:395–430.

Strimmer K, Pybus OG. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol Biol Evol. 18:2298–2305.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.

Terwilliger JD, Weiss KM. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotech. 9:578–594.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theoret Popul Biol. 7:256–276.

Watterson GA. 1978. The homozygosity test of neutrality. Genetics. 88:405–417.

Wolfram S. 1999. The mathematica book. 4th edition. Cambridge (UK): Wolfram Media/Cambridge University Press.

Wright S. 1931. Evolution in Mendelian populations. Genetics. 16:97–159.

Wright S. 1949. Population structure in evolution. Proc Am Philos Soc. 93:471–478.

Marcy Uyenoyama, Associate Editor

Accepted September 9, 2008