Natural Selection and Coalescent Theory

John Wakeley
Harvard University

INTRODUCTION

The story of population genetics begins with the publication of Darwin's *Origin of Species* and the tension which followed concerning the nature of inheritance. Today, workers in this field aim to understand the forces that produce and maintain genetic variation within and between species. For this we use the most direct kind of genetic data: DNA sequences, even entire genomes. Our "Great Obsession" with explaining genetic variation (Gillespie, 2004a) can be traced back to Darwin's recognition that natural selection can occur only if individuals of a species vary, and this variation is heritable. Darwin might have been surprised that the importance of natural selection in shaping variation at the molecular level would be de-emphasized, beginning in the late 1960s, by scientists who readily accepted the fact and importance of his theory (Kimura, 1983). The motivation behind this chapter is the possible demise of this Neutral Theory of Molecular Evolution, which a growing number of population geneticists feel must follow recent observations of genetic variation within and between species.

One hundred fifty years after the publication of the *Origin*, we are struggling to fully incorporate natural selection into the modern, genealogical models of population genetics. The main goal of this chapter is to present the mathematical models that have been used to describe the effects of positive selective sweeps on genetic variation, as mediated by gene genealogies, or coalescent trees. Background material, comprised of population genetic theory and simulation results, is provided in order to facilitate an understanding of these models. A strong thread running throughout is the use of population genetic data to draw conclusions broadly about the process of evolution, and the shifting ideas about the causes of evolution that have characterized the field at various times, as our ability to sample genetic data has improved.

THEORETICAL POPULATION GENETICS

Provine (1971) described the birth of theoretical population genetics, which originated with Darwin and culminated in the great works of Fisher, Wright and Haldane. The latter three luminaries established the fundamental dynamics of genetic evolution, of changes in allele frequencies through the interaction of mutation, selection and random genetic drift. The term 'random genetic drift' always requires explanation: it is the stochastic side of evolution, which results from the random transmission of genetic material from one generation to the next in a population due to Mendelian segregation and assortment, as well as the partially unpredictable processes of survival and reproduction. The

founding works of this field (Fisher, 1918, 1930; Wright, 1931; Haldane, 1932) remain a crucial part of any advanced education in evolutionary biology.

Relevant aspects of the genetic evolution are reviewed below, but one early result deserves to be mentioned here. Consider the probability of fixation of a new mutant allele under the influence of positive natural selection. Initially, every individual has genotype A_1A_1. A mutation produces a new allele, A_2, which gives its carriers an advantage. If A_1A_2 individuals have an average of $1+s$ offspring and A_2A_2 individuals an average of $1+2s$ offspring, relative to A_1A_1 individuals, then the probability that the new mutant allele A_2 goes extinct is approximately $1–2s$. This result holds when s is small relative to 1 and the population size, N, is very large ($Ns>>1$). It can be derived using a branching process model, in which each A_2 allele has a Poisson number of descendants with mean $1+s$ each generation (Haldane, 1927; Fisher, 1922, 1930) and it can also be obtained using diffusion theory (see below). The probability of fixation is the probability that eventually the entire population will have genotype A_2A_2. In a finite population this is equal to one minus the probability of extinction, in this case $2s$, which is small. One cannot help but marvel at the possible implications of this result: that the many important adaptations we observe in nature might first have gone extinct several times before they became successful and that many, possibly even better adaptations have not been observed at all because they were lost despite their selective advantage.

It is remarkable that so much of what Fisher, Wright, and Haldane did in the 1920s and 1930s is still relevant today, given that almost nothing was known at that time about the material bases of heredity, development, and ecology. Although our current knowledge of development and ecology is still not sufficient to permit a full evolutionary theory—one that would include the richness between genotype and phenotype, and would extend to interactions between individuals and their environment—our modern understanding of genetics is quite detailed. This has led to improvements of the models of population genetics, away from the simple A_1, A_2, etc., allelic models above, to models which include the structure of DNA, the various kinds of mutations, and, perhaps most importantly, recombination within, as well as between, genetic loci. We may say with some confidence that we know the fundamental components of genetic evolution. As Lynch (2007, p. 366) puts it: “Many embellishments have been added to the theory, and views have changed on the relative power of alternative evolutionary forces, but no keystone principle of population genetics has been overturned by an observation in molecular, cellular, or developmental biology.”

The ‘modern synthesis’ of the mid-twentieth century was initiated in no small part by the early work of Fisher, Wright, and Haldane. It later involved the wide application of ideas from population genetics to explain the patterns of evolution (Dobzhanshy, 1937; Huxley, 1942; Mayr, 1963), although sometimes without the aid of the vital mathematical models of that field. This period also saw the great development of mathematical theory, although largely in the absence of data about the genetic variation the theory purported to explain (Lewontin, 1974). We can recognize two additional seminal figures of mathematical population genetics from the mid-twentieth century: Malécot and Kimura. Among many important contributions (Nagylaki, 1989; Slatkin and Veuille, 2002), Malécot introduced the notion of following a pair of alleles backward in time to their common ancestor (Malécot, 1941, 1948). This is the basic idea behind coalescent theory, which is discussed in detail below. Kimura is best known for the

neutral theory of molecular evolution (Kimura, 1983), but his place in mathematical population genetics derives from his work on the diffusion theory of allele frequencies.

DIFFUSION THEORY

As the results of diffusion theory are used below and the assumptions of coalescent theory and diffusion theory are the same, a brief review of the basic concepts is given here. See Ewens (1979, 2004) for an excellent and thorough treatment.

Diffusion models approximate the dynamics of allele frequencies over time in large populations. The discrete or exact models of population genetics typically imagine a diploid population of constant size N, in which time is measured in discrete units of generations. The number of copies of an allele (*e.g.*, the mutant A_2 above) must, at any one time, be one of $2N+1$ possible values: 0, 1, 2,…, $2N–1$, $2N$. If there are k copies of an allele, then the frequency of that allele is $p = k/(2N)$. In a diffusion model, both time and allele frequency are measured continuously: $p \in [0,1]$, and $t \in [0,\infty)$. This is achieved by taking a limit of the dynamics, as N tends to infinity, with time rescaled so that one unit of time in the diffusion model corresponds to $2N$ generations in the discrete model. Intuitively, when N is large, p may assume very many possible values, so there will be little error in measuring allele frequencies continuously. Similarly, a single generation comprises a very small step when time is viewed on the scale of $2N$ generations. Diffusion models allow the computation many quantities of interest (in order to make predictions, test hypotheses, and estimate parameters), while most discrete models are mathematically intractable.

Discrete models differ in their assumptions about population demography and reproduction, and thus about the dynamics of genetic transmission from one generation to the next. The Wright-Fisher model is the most commonly used (Fisher, 1930; Wright, 1931), although the Moran model is employed often (Moran, 1962). Besides tractability, another advantage of the diffusion approximation is that many different discrete models have the same diffusion limit. Here, "the same" includes the possibility of a constant multiplier of the time scale, so that time is measured in units of $2Nc$ generations. In the Wright-Fisher model, $c = 1$, and in the Moran model, $c = 1/2$, but the mathematical form of the diffusion equations is identical. We say that the *effective population size* is $N_e = cN$ diploid individuals (Ewens, 1982; Sjödin *et al*. 2005). This means that we may use the diffusion approximation of the Wright-Fisher model to illustrate general features of the evolution of populations, knowing that if we replace N with N_e the results will be valid for other populations that do not conform to the overly simple Wright-Fisher model.

A single effective population size may not exist, as in the case of two populations with little or no gene flow, or when the size of the population changes dramatically over time so that the time scale of the diffusion model would also have to change over time. In the interest of brevity and simplicity, populations which deviate so dramatically from the assumptions of the Wright-Fisher model will not be considered here.

On a per-generation basis, the rate of genetic drift, which is the rate at which the frequency of an allele will change, at random due to the vagaries of genetic transmission in a population, is equal to $1/(2N)$ in the Wright-Fisher model. The per-generation effects of selection, mutation and recombination are captured in additional parameters, here denoted s, u, and r. We saw the definition of s above, and we now define u and r to be

the probability of a mutation at a single nucleotide site and the probability of a recombination event between two adjacent nucleotide sites, respectively, between a parent and its offspring. This simple statement of a model leaves our many potentially important things, such as possible variation in these parameters across a genome, among alleles, or through time, and we should add such details to the model later, as needed. In the diffusion limit, where time is rescaled by $2N$, random genetic drift has rate 1 and the strengths of selection, mutation and recombination are given by $2Ns$, $2Nu$, and $2Nr$.

By tradition, the population mutation parameter is defined as $\theta = 4Nu$ or *twice* the population rate of mutation on the diffusion time scale. For consistency, in what follows, the population parameters for selection and recombination will be defined as $\sigma = 4Ns$ and $\rho = 4Nr$. Note, however, that both α and γ are frequently used in place of σ, and are often defined as $2Ns$ rather than $4Ns$.

Kimura's (1955a,b) groundbreaking achievement was to obtain the probability density function of the frequency of an allele at any future time given its current frequency, at a single locus under the influence of natural selection and random genetic drift. Again, it is impossible to make predictions of this sort under most discrete models, in particular the Wright-Fisher model. Kimura's result spurred much further work on diffusion theory, by himself and others, which is reviewed in Ewens (1979, 2004).

NEUTRAL COALESCENT THEORY

Kimura's use of diffusion theory in the 1950s flowed out of his desire to explore the dynamics of genetic drift, which Wright had promoted as having a dramatic role in evolution. In the 1960s, the focus of population genetics shifted to explaining the new observations of protein-sequence divergence between species and allozyme variation within species (Zuckerkandl and Pauling, 1965; Harris, 1966; Lewontin and Hubby, 1966). These and subsequent data caused a dramatic shift in thinking about the role of natural selection, with Kimura and others (Kimura, 1968; King and Jukes, 1969) suggesting a predominant role for neutral mutations in evolution at the molecular level. Later, this concept was greatly expanded by Ohta (1973, 1992) to include weakly selected, or 'nearly neutral' mutations. By emphasizing random genetic drift, the new theories did seem to provide a simple explanation for the observations of the day: that molecular differences between species accumulate surprisingly linearly with time and that natural populations harbor tremendous amounts of genetic variation.

Previously, with little data available, population geneticists had formed two opposing selectionist camps: the 'classical' and 'balance' schools (Dobzhansky, 1955). Lewontin (1974) provides a clear analysis of how these gave way to the neutral theory when faced with explaining high observed levels of polymorphism and divergence, and Crow (2008) recounts the arguments from the key perspective of someone whose career spanned this and other controversies. Lewontin (1974) argued against an unbridled focus on neutrality. He suggested the term 'neoclassical theory' because, although a shockingly large fraction of the functional differences at the molecular level might be invisible to selection, still most mutations are disadvantageous and some or all adaptations must be driven by natural selection. Kimura recognized these points in his concept of the neutral theory: he was ready to accept that approximately 10% of amino acid substitutions between species could be driven by positive selection (see Ohta and

Kimura, 1971), and that 85-95% of non-synonymous, or amino-acid-changing, mutations are substantially deleterious (see pp. 206-210 in Kimura, 1983).

If we wish to infer the action of natural selection, then neutrality is the appropriate null hypothesis. Mathematical models soon began to include the assumption that *all* genetic variation was neutral. Importantly, they also included increasingly refined assumptions about the mutation-structure of variation, in an effort to be appropriate to the data at hand (Ewens, 1972; Ohta and Kimura, 1973, Moran, 1975; Watterson, 1975). It seems inevitable in hindsight that this would lead to the consideration of the mathematical structure of ancestral relationships among sampled alleles, or gene genealogies. For example, the famous Ewens sampling formula (Ewens, 1972; Karlin and McGregor, 1972) clearly has the fundamental structure of gene genealogies under neutrality embedded in it; see Hobolth *et al*. (2008) and section 3 of Kingman (1982a). However, a major shift in orientation was required: from the prospective view of classical population genetics to the retrospective view of coalescent theory (Ewens, 1990).

The paper by Watterson (1975) is the earliest in which gene genealogies and their relationship to genetic data are easily recognizable. Remarkably, if all variation is selectively neutral, it is possible to model just the ancestors of the sample, and ignore the other members of the population. Figure 1 shows a hypothetical gene genealogy, or coalescent tree, for a sample of size $n = 6$. It is a binary tree which traces the ancestral lines of the sample back (up) to their most recent common ancestor. Time is measured by vertical distance. The nodes in the tree represent coalescent events, where a pair of ancestral lines reaches a common ancestor. Each branch in the tree depicts all of the genetic ancestors of particular members of the sample. Therefore, any polymorphisms in the data must be due to mutations along the branches. Watterson used this idea to derive the expectation and variance of the number of polymorphic nucleotide sites in a sample at a locus that does not undergo recombination.

The coalescent and the diffusion are inextricably related as *dual* processes; for mathematical details, see Möhle (1999). Under identical assumptions to those made in diffusion theory, but for the moment without selection or recombination, each pair of ancestral lines coalesces independently with rate equal to one. In the Wright-Fisher model, this corresponds to their being a probability $1/(2N)$, in each generation looking back, that two alleles descend from a common ancestral allele. In considering the limit $N \to \infty$, the sample size n is treated as a (finite) constant, and this is the reason that all coalescent events occur between pairs of alleles rather than larger numbers.

Because every pair of ancestral lines coalesces with rate equal to one, neutral gene genealogies are random-joining trees and the time, T_i, during which there exist exactly i lines ancestral to the sample is exponentially distributed with mean $2/(i(i-1))$, which is the inverse of the number of possible pairs of i lines. As a result, T_2 tends to be the longest coalescent interval, comprising about half of the time to the most recent common ancestor ($T_{\mathrm{MRCA}} = T_n + T_{n-1} + \ldots + T_2$). The gene genealogy in Figure 1 is drawn with the times, T_i, equal to their expected values. On the coalescent or diffusion time scale, the expected value of T_{MRCA} is equal to

$$E(T_{MRCA}) = 2(1 - 1/n).$$

This can be translated into Wright-Fisher-model generations by multiplying by 2*N*. For large samples it converges to 4*N* generations, which is not unexpected because this is also the expected fixation time for an allele from forward-time diffusion theory without selection (Kimura and Ohta, 1969). With a choice of mutation model, this standard coalescent is an efficient way of predicting patterns of neutral variation. In modeling or simulation, we simply generate the tree and the times, then place mutations randomly along each branch with rate $\theta/2$ per site.

Kingman (1982a,b) gave the mathematical proof of the coalescent process. Independently, biologists were introduced to the theory of gene genealogies and their biological relevance by Hudson (1983a, 1990) and Tajima (1983). Tavaré (1984) helped bridge the gap between the biological and the mathematical, and between diffusion theory and coalescent theory. Hudson also studied the effects of recombination, and described how to simulate gene genealogies both with and without recombination (Hudson, 1983a, 1983b, 2002). Recombination complicates the coalescent process substantially, but is difficult to brush aside because the per-site rates of mutation and recombination (θ and ρ) appear to be of the same order of magnitude in many species. Table 4.1 in Lynch (2007) gives examples. Thus, a growing number of neutral coalescent approaches to inference take recombination into account; for example, see Becquet and Przeworksi (2009). When natural selection is added to the coalescent, it becomes absolutely critical to include recombination.

Coalescent theory is best known today for having produced a repertoire of tools for statistical inference under the assumption that genetic 'markers' (*i.e*. polymorphisms) are neutral. In the 1980s, this was fueled by the remarkable utility of uni-parentally inherited, non-recombining animal mitochondrial DNA for uncovering plausible histories of population expansions and contractions, and complex patterns of geographic subdivision, in many different species (Avise *et al*., 1987). Appropriate statistical machinery was developed, and work flourished after the introduction of Markov chain Monte Carlo, importance sampling, and Bayesian approaches in computational methods of coalescent-based inference. Stephens and Donnelly (2000), Marjoram and Tavavré (2006), and Felsenstein (2007) together give a comprehensive review of methods. Estimates made using these tools are sensible enough that they have contributed to broad debates about ancient processes and events; for example, see Hey (2005).

GENOMIC DATA AND THE MODELING RESPONSE

The continued development of technologies for measuring genetic variation after the 1960's—from restriction-enzyme digests of mitochondrial DNA (Avise *et al*. 1979; Brown, 1980), to early DNA sequence data (Aquadro and Greenberg, 1983; Kreitman, 1983)—has led us to the massive contemporary genome-sequencing efforts, such as the 1000 Genomes Project and the Personal Genome Project. As new data are gathered, paradigms are questioned. At present, genomic polymorphism and divergence data from a growing number of taxa suggest staggering amounts of positive selection. For example, Hahn (2008) cites estimates from *Drosophila melanogaster* and *D. simulans* that 30% to 94% of amino acid substitutions between species have been driven by positive selection. Halligan *et al.* (2010) put this figure at 57% for substitutions between the mouse *Mus musculus castaneus* and the rat *Mus famulus*. Large fractions of positively selected

substitutions (~50%) have also been reported for non-coding regions in *Drosophila* (Begun *et al*., 2007). These dramatic observations do not seem to extend to some other well studied species, in particular *Arabidopsis* (Bustamante *et al*., 2002) and humans. Sella *et al*. (2009) summarize a number of studies of humans, and conclude that ~10% of amino acid substitutions have been driven by positive selection.

The latter figure (~10%) is remarkably similar to the initial estimate that was considered broadly consistent with the neutral theory (Ohta and Kimura, 1971; Kimura, 1983). As Lewontin (1974) pointed out, the discovery that some genetic loci have been the targets of selection does not invalidate the neoclassical view. In addition, Thornton *et al*. (2007) and Jensen (2009) caution against drawing strong conclusions based on current methodologies and data. Still, Hahn (2008) argues that a new theory is required, one in which selection plays the major role, and Sella *et al*. (2009) agree that at least some parts of the neutral, or neolcassical, theory are in dire need of an overhaul.

One major tenet of population genetics, which sits at the base of neutral theory, is clearly not in question: a large fraction of mutations alter function so ruinously that they are extremely unlikely to be observed, either as substitutions between species or as polymorphisms within species. In fact, one of the principal methods of estimating the fraction of positively selected amino acid changes (Smith and Eyre-Walker, 2002) is to use the ratio of non-synonymous to synonymous polymorphisms within species to set a low baseline expectation for the ratio of non-synonymous to synonymous substitutions between species. Positively selected amino acid changes may then be uncovered, by an "excess" of non-synonymous substitutions, even if the number of non-synonymous substitutions per site is much smaller than the number of synonymous substitutions per site. Such considerations lead to sophisticated yet tractable statistical approaches to estimating selection in the case where sites may be assumed independent of one another (Sawyer and Hartl, 1992; Sawyer *et al*., 2003).

Positive natural selection for adaptive traits has been the primary source of excitement among workers studying the genomic effects of selection. In addition to estimating the overall prevalence of positive selection, effort has focused on identifying recently selected loci. This is possible because the fixation of an advantageous allele at one locus affects loci nearby, in a phenomenon is known as genetic 'hitch-hiking' (Maynard Smith and Haigh, 1974; Kaplan *et al*. 1989). The primary signal of this is a reduction in variation around the site of selection, but a number of subtler effects occur as well (Nielsen, 2005). The term 'selective sweep' is used loosely to mean the fixation of a positively selected allele or the attendant reduction in variation. Thornton *et al*. (2007) review genomic scans for recent selective sweeps in *Drosophila*, which have identified large numbers loci. In humans, Williamson *et al*. (2007) suggest that recent hitch-hiking affects 10% of sites in the genome. Although demographic factors can lead to false positive inferences of selection (Thornton *et al*., 2007), and divergence data and polymorphism data give rather different estimates of the prevalence of selection (Jensen, 2009), these recent findings motivate the development of coalescent approaches to modeling selective sweeps.

SELECTION AND GENETIC DRIFT FORWARD IN TIME

The models of population genetics are often based on diffusion approximations, but are the assumptions of diffusion theory reasonable for loci undergoing positive selective sweeps? Diffusion theory assumes that s, u, and r are very small and N is very large. Formally, the limit $N\to\infty$ is taken with $\sigma = 4Ns$, $\theta = 4Nu$, and $\rho = 4Nr$ held constant. However, simulations show that many results of diffusion theory are very accurate for moderate values of the discrete-model parameters, such as $N = 100$ and $s = 0.01$. The occurrence of a sweep implies that selection is strong in some sense, so we ask more specifically whether it is reasonable to use a model in which s is assumed to be much less than one. Estimates from recent selective sweeps suggest that the answer is yes. One example, not from *Drosophila* but from deer mice, was reported recently by Linnen *et al*. (2009). They estimated $s = 0.0056$ for a recently swept allele affecting pelage color of mice in the Nebraska Sand Hills. This is similar to the larger estimates for swept loci in *Drosophila* (Thornton *et al*., 2007; Sella *et al*., 2009), so assuming small s appears safe. Still, a sweep certainly indicates that selection has overwhelmed random genetic drift. In the diffusion model, this occurs when σ is large. Estimates of σ for swept loci in *Drosophila* range from values in the tens to values in the thousands (Thornton *et al*., 2007; Sella *et al*., 2009). In sum, the diffusion with large σ appears to be good starting point for modeling selective sweeps.

Note that there is an entirely different diffusion model in population genetics (Norman, 1975), which may be more appropriate for large σ. Unfortunately, few results are available for this "Gaussian" diffusion model, and we will not pursue it further. Ewens (1979, 2004) points out that the Gaussian diffusion and the standard one should overlap for certain parameter values (*i.e*. large values of σ), and this is illustrated for strong balancing selection and mutation in Wakeley and Sargsyan (2009).

We will define a sweep as the event that a positively selected allele, which starts in frequency $1/(2N)$ as a new mutation, reaches frequency 1, or fixes in the population. For simplicity, we will also assume that all parameters are constant over time (but see below). Diffusion theory can tell us about the distribution of trajectories the allele will take on its way to fixation. Knowing this distribution is helpful because many things we are interested in are functions of the allele-frequency trajectory. For example, the average duration of the sweep is identical to the expected value of the length of the trajectory. Other quantities, such as the probability of coalescence during a sweep or the chance of observing a sweep in a sample of genetic data, also depend on the characteristics of allele-frequency trajectories.

To illustrate the simplest dynamics of genetic evolution, and with the emerging estimates from *Drosophila* and humans as a backdrop, let us imagine a locus comprised of a single 'advantageous' site, 1000 'deleterious' sites, and 1000 'neutral' sites. For humans, estimates of θ are on the order of 0.001 and estimates of the effective population size are on the order of 10000. Thus, we will use a Wright-Fisher model with $N = 10^4$ and $u = 2.5\times10^{-8}$ for our "humans." Then, the total rates of mutation are $\theta_a = 0.001$ for 'advantageous' sites and $\theta_d = \theta_n = 1.0$ (*i.e*., 1000×0.001) for 'deleterious' and 'neutral' sites. Let us also assume fairly strong selection, in particular $\sigma_a = 100$ and $\sigma_d = -100$. With $N = 10^4$, this corresponds to $|s| = 0.0025$ for advantageous and deleterious mutants. Of course, we have $\sigma_n = 0$. Estimates of θ for *Drosophila* are somewhat more than an order of magnitude greater than those for humans. For computational efficiency, let us get our "*Drosophila*" parameters simply by multiplying the "human" diffusion-

scale parameters by ten, so that $\theta_a = 0.01$, $\theta_d = \theta_n = 10.0$, $\sigma_a = 1000$, and $\sigma_d = -1000$. In the simulations presented below, this is realized by using the same per-generation parameters as for humans, but with $N = 10^5$ instead of $N = 10^4$.

This idealized model will serve to generate intuition about selective sweeps and the relative magnitudes of the processes involved. In relation to the estimates of rates of adaptive substitution in humans and *Drosophila*, with these parameters, our model predicts that ~9% of substitutions will be driven by positive selection in "humans" and ~50% of substitutions will be driven by positive selection in "*Drosophila*." These percentages are derived, in the usual way, by multiplying the per-generation rates of introduction each type of mutation ($\theta_d/2$, $\theta_n/2$, $\theta_a/2$) by their probabilities of fixation from diffusion theory,

$$P(\text{fix}) = \begin{cases} \dfrac{1}{2N} & \text{if } \sigma = 0, \\ \dfrac{1 - e^{-\sigma/(2N)}}{1 - e^{-\sigma}} & \text{if } \sigma \neq 0. \end{cases}$$

Note that this is the standard result, which is sufficient for our purposes, and does not include the $s \rightarrow s/(1+s)$ correction suggested by Bürger and Ewens (1995).

For our "humans," we have $P(\text{fix})$ approximately equal to 5×10^{-3}, 5×10^{-5}, and 1.9×10^{-46} for advantageous, neutral, and deleterious mutations, respectively. For our "*Drosophila*" model, the corresponding values are 5×10^{-3}, 5×10^{-6}, and 2.5×10^{-437}. Note that when $\sigma = 4Ns$ is large and s is small, as is true here, the second case in the equation above gives $P(\text{fix}) \approx 2s$, which is the classical population genetic result we saw earlier. The probabilities of fixation of advantageous mutants are thus the same for our "humans" and our "*Drosophila*," while the probabilities for neutral mutants differ by a factor of ten due to the difference in population size. In both cases deleterious mutations are exceedingly unlikely to fix.

Figure 2 shows the trajectories of advantageous alleles in simulations of our "humans" (Figure 2A,B) and "*Drosophila*" (Figure 2C,D). First, a large number of trajectories was simulated, from the introduction of a mutant in a single copy until the mutant either fixed or went extinct. Then, the origination times of the mutations were generated using the per-generation population rates of advantageous mutation, $\theta_a/2$. Thus, these simulations are of independent trajectories; they do not take interference between alleles into account. This is reasonable for "humans" because successful sweeps are fairly well separated in time (2A) and alleles go extinct quickly when sweeps fail (2B), and does not invalidate the qualitative points we will draw from the figure as a whole. Simulations of each trajectory were done according to the discrete Wright-Fisher model, with the parameters above.

Before looking in detail at Figure 2, note that our model and simulations follow the fairly common convention of using one-locus, two-allele dynamics to portray a situation which is probably much more complicated. For example, we have assumed that *every* mutation at the 'advantageous' site has selection parameter σ_a. Recalling our classical A_1/A_2 model, this would be realized if the average number of offspring of A_2A_2 is

mysteriously reset from 1+2*s* to back 1 at the conclusion of the sweep, and the next mutation again has selection coefficient *s*. We have also assumed that the strength of selection is constant over time, which might not be realistic even within a single sweep. Gillespie has shown repeatedly (*e.g.*, Gillespie, 1991, 2004b) that key features of the dynamics of fully specified, multi-allele models, in which selection parameters differ among alleles and may change over time, are simply not captured using two-allele approaches, and has further argued that these shortcomings are fatal to neutral-theory explanations of the molecular evolution. For us they are of somewhat less concern because our focus is the much shorter time scale of single sweeps.

Figure 2 shows the frequency trajectories of advantageous alleles over a period of time during which we expect 1000 advantageous mutations to occur. Sweeps appear as nearly vertical curves, in which the frequency (x) of an allele rises quickly from $1/(2N)$ to 1. Because the probability of fixation is $\sim 5\times10^{-3}$ in both "humans" and "*Drosophila*", we expect about five selective sweeps in each. This is exactly what was observed in these particular simulations, but just by chance: in both cases the number of sweeps is Poisson distributed with mean ~5. The time it takes to observe 1000 advantageous mutations is ten times shorter in "*Drosophila*" than in "humans" because the rate of introduction of advantageous mutations ($\theta_a/2$) is ten times greater. If we ran our "Drosophila" simulations over 2×10^6 generations, as we did for "humans," we would expect to see 50 sweeps. Time in Figure 2 is measured in generations, and accordingly the panels for "*Drosophila*" (C,D) are one tenth the length of the panels for "humans" (A,B).

Panels A and C display the entire range of frequencies, and on this scale only a handful of the trajectories are visible. At our hypothetical locus, with its one positively selected site, we expect one advantageous mutation to occur about every 2000 generations in "humans" and one about every 200 generations in "*Drosophila*." Even focusing on much smaller frequencies, as in panels B and D, the trajectories of most alleles are difficult to see. Recall than only $\sim 5\times10^{-3}$ of advantageous mutations will sweep to fixation. The other ~99.5% of them go extinct, and they do so very quickly, without ever reaching substantial frequencies. The trajectories of the deleterious alleles at our loci are not shown. For our values of σ_d (–100 and –1000), deleterious alleles will essentially never fix in the population. They enter the population and may drift to frequencies of 1% or so, but then are lost.

Within our "humans," advantageous and deleterious mutations will not typically have much effect on levels of neutral polymorphism. Either they will never reach appreciable frequencies or they will sweep quickly through the population and only rarely be observed. Sweeps in our "humans" occur on average only every 400,000 generations while the effective population size, which sets the average time for neutral variation to reach equilibrium levels, is only 10,000. The situation is rather different for our "*Drosophila*," in which sweeps occur at the locus every 40,000 generations and the effective population size is 100,000. In this case, we expect sweeps to greatly affect levels and patterns of neutral polymorphism.

Methods of inference of selective sweeps often assume that the population has been sampled just at the end of the sweep. In applications this needs to be justified, because, *a priori*, the time back to the last sweep is unknown. In our model it would be roughly exponentially distributed with mean $(P(\text{fix})\theta_a/2)^{-1}$. Sweeps appear to go to completion almost instantaneously on the time scale in Figure 2. However, by traveling

back to the end of the last sweep that occurred in each case, and changing the time scale, we can see the shape of sweeps. Figure 3 shows this for "humans" (panel A) and "*Drosophila*" (panel B), with the time scale given in the coalescent or diffusion units of $2N$ generations. In both cases, the total range is 0.5 on the new time scale, which is equivalent to N generations (10,000 for "humans" and 100,000 for "*Drosophila*"). Also, time now flows from the moment the population is sampled back into the past, as is the custom in coalescent modeling. In contrast to Figure 2, only those trajectories that went to fixation are shown in Figure 3.

Figure 3 illustrates that sweeps tend to follow sigmoidal trajectories, with allele frequencies changing relatively slowly when $x(t)$ is close to 0 or 1, but moving rapidly through the middle frequencies. With time measured in units of $2N$ generations, more strongly favored alleles will sweep more quickly through the population ($\sigma_a = 100$ in A versus $\sigma_a = 1000$ in B). In addition, in species like our "*Drosophila*" shown in Figure 3B, the rate of occurrence of selective sweeps might be such that several will have happened in the recent ancestry of a locus under study. We can recall the results of neutral coalescent theory, that the average time back to the common ancestor for a sample of size $n = 2$ is equal to 1 (*i.e.* $2N$ generations) and the average time to the most recent common ancestor of all members of a large sample is ~2 (*i.e.* ~$4N$ generations).

The standard diffusion model, with large σ, allows us to quantify these observations. A fundamental result of diffusion theory in population genetics, due to Ewens (1963, 1964), concerns the average time that an allele, which begins in frequency p and sweeps through the population, spends at each frequency x on its way to fixation. The function, called $t^*(x;p)$ and given as equation 5.52 in Ewens (1979) or 5.53 in Ewens (2004), has the interpretation that

$$\int_{x_1}^{x_2} t^*(x;p)dx$$

is the average amount of time, on the diffusion time scale, that the allele frequency spends in the interval (x_1,x_2) before the allele fixes in the population. Integrating over the entire frequency range gives the expected total sweep time of a new mutant. When σ is large, this may be approximated as

$$t_{\text{fix}} = \int_0^1 t^*(x;p)dx \approx \frac{4(\log(\sigma)+\gamma)}{\sigma}.$$

The symbol γ above is Euler's constant (approximately 0.5772). Note that s in Ewens (1979, 2004) is equivalent to our $2s$, so $\alpha = 2Ns$ in Ewens is equivalent to our σ. The equation above gives ~0.2 for the fixation time when $\sigma = 100$, and ~0.03 when $\sigma = 1000$, and these match the simulation results very well (*e.g.* see Figure 3).

Although this particular result for the fixation time of an strongly advantageous allele starting from a single copy seems to have appeared in the literature only recently (Hermisson and Pennings 2005; Etheridge *et al*. 2006; Hermisson and Pfaffelhuber 2008), it illustrates something that has been known for several decades. That is, deterministic equations for allele-frequency trajectories can drastically overestimate the

amount of time an allele will spend in small frequencies (*e.g.*, Ewens (1979) page 149). If an advantageous alleleis going to sweep to fixation, it must move away from the boundary (x=0) faster than the deterministic equations predict. Still, it is not uncommon to see deterministic results and methods used in this context in the biological literature. The deterministic model (*e.g.*, 1.28 in Ewens (1979) but with our s) gives

$$\int_{\frac{1}{2N}}^{1-\frac{1}{2N}} (sx(1-x))^{-1} dx = \frac{2\log(2N-1)}{s} \approx \frac{2\log(2N)}{s}$$

for the fixation time, in generations. This appears in many publications. It may be compared to the diffusion result above by multiplying t_{fix} by $2N$ and rearranging:

$$2Nt_{\text{fix}} \approx \frac{2(\log(2N) + \log(2s) + \gamma)}{s}.$$

Recall that $\log(a) < 0$ when $0 < a < 1$, and tends to negative infinity as a tends to zero. Even if σ is large, it might not be reasonable to assume that $\log(2N)$ is much greater than both $-\log(2s)$ and γ. For any values of s we are likely to consider, the deterministic result will overestimate the diffusion result. In our "humans," the diffusion result gives 4146 generations, while the deterministic result gives 7923 generations. For "Drosophila," the corresponding numbers are 5988 generations and 9765 generations.

The fact that allele-frequency trajectories are sigmoidal is key to understanding coalescent models of selective sweeps because the rates of events in the ancestry of a sample depend on the allele frequencies. As a final point about diffusion models of sweeps before turning to coalescent models, Etheridge *et al*. (2006) have recently obtained the very interesting result that as σ grows, the fraction of the time that the allele spends the 'middle frequencies' becomes negligible; see their Lemma 3.1 and note that their α is our $\sigma/2$. Specifically, the time spent going from frequency ε to $1-\varepsilon$ becomes negligible, for *any* $0 < \varepsilon < 1$, so that the allele ultimately spends half of t_{fix} in the interval $(0,\varepsilon)$ and the other half in the interval $(1-\varepsilon,1)$. Because of this, it is possible to make some detailed calculations concerning the approximate behavior of coalescent process during selective sweeps when σ is large (Etheridge *et al*., 2006).

Our investigation of coalescent models with selection will be drawn upon these results from diffusion theory. We will focus on selective sweeps, in particular the effect these have on ancestral processes at nearby neutral loci. It seems increasingly clear that the hitch-hiking effect studied by Maynard Smith and Haigh (1974), which reduces polymorphism levels around the site of a sweep, has affected many loci. For example, based on data from humans, Sabeti *et al*. (2007) listed 22 regions in the human genome where selection appears to have decreased polymorphism over spans of 0.2 to 3.5 Mb, at least in some populations. Kimura (1983) did not cite Maynard Smith and Haigh (1974) and yet he accepted the estimate that roughly 10% of substitutions might be driven by positive selection (Ohta and Kimura, 1971). This is roughly what our "human" model predicts. On the one hand, it is true that polymorphism levels at our "human" locus should not typically deviate from neutral predictions, because sweeps will occur only about once every $20\times2N$ generations. Using the diffusion result for t_{fix}, with $\sigma = 100$, the

chance of catching a sweep in progress is only about 1%. On the other hand, if a large number of loci are surveyed, we should not be surprised to find several that have recently been affected by sweep. It is these loci that current genome-wide scans for selection may uncover, and coalescent models are being developed to aid both in their identification and to make estimates of the strength and timing of selection.

COALESCENT MODELS WITH SELECTION

When selection operates, for example with two alleles A_1 and A_2, then the population is structured by allelic type such that the average number of offspring of genetic lineages labeled A_1 differs from that of lineages labeled A_2. In order to model the genetic ancestry of a sample, we need to keep track of these labels. Here we will review three different approaches to this problem, guided by our concern for selective sweeps. A fourth coalescent approach to selection, the ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997), will not be reviewed because it has not been extended to apply to selective sweeps. Hitch-hiking will be a key phenomenon in our investigations: neutral genetic markers contain information about past histories of selection just as they do about other demographic processes and events. The fact that strongly selected adaptive substitutions have probably occurred at a small minority of sites in the genome means that the bulk of signals of selection will be in patterns linked variation. Recombination is the process that modulates the effect of linkage, so recombination is fundamental in what follows.

The Structured Coalescent Approach

Hudson and Kaplan (1986) showed how conditioning on the allelic types of a sample alters the coalescent process, in a way that is similar to the effect of geographic stucture and migration. Two lineages with the same allelic type may coalesce, but two lineages with different types must wait for mutation to change the type of one or the other. Kaplan *et al*. (1988) applied this idea to a locus under selection, showing that rates of coalescence and mutation in the ancestral process depend on the frequencies of the two alleles. Hudson and Kaplan (1988) extended the model to describe the coalescent process at a linked neutral locus, conditional on the frequency trajectory at the selected locus. Darden et al. (1989) described the joint process of coalescence at the linked neutral locus and changes in allele frequencies at the selected locus by the standard diffusion. Barton *et al*. (2004) investigated this model more rigorously, and found boundary conditions necessary to allow analytical work. Kaplan *et al*. (1989) considered the specific application of this approach to a strong selective sweep and the effect this has on variation at the linked neutral locus.

This structured coalescent approach has led to a number of useful simulation methods (Slatkin, 2001; Kim and Stephan, 2002; Przeworski, 2003; Coop and Griffiths, 2004)., in which an allele-frequency trajectory is generated, then the structured coalescent process is run conditional on the trajectory. A main goal in developing these simulations is to devise methods of estimating the characteristics of sweeps, such as the selection parameter σ and the time the last sweep began. Kim and Wiehe (2009) review the issues and available software.

A key feature of the structured coalescent approach to selection is that the rate of coalescence within an allelic class depends inversely on the allele frequency. Consider our selectively favored allele A_2, whose frequency is $x(t)$ at time t in the past, measured in units of $2N$ generations. If there are i ancestral lineages of type A_2, then the rate of coalescence between any pair of them is $1/x(t)$, and the total rate is

$$\frac{\binom{i}{2}}{x(t)}.$$

If $x(t) = 1$, the rate is, rightly, the same as in the standard neutral coalescent. However, if $x(t) < 1$, then the rate is *greater* than in the standard neutral coalescent. The reason for this is that, when $x(t)$ is smaller, there are fewer possible parents of the i lineages, so the probability of a common ancestor in a single generation is larger. The same notion applies to lineages that possess the A_1 label, but with $1–x(t)$ instead of $x(t)$.

In considering the effects of linkage, we imagine a site or locus B that sits at a distance m from the selected locus A. Let m be in units of base pairs and the total scaled rate of recombination be

$$\rho^* = m\rho,$$

where ρ is the per-site rate of recombination we defined before. Recall that $\rho = 4Nr$, and that $\rho/2$ is the rate of recombination between two adjacent base pairs on the coalescent time scale. In defining ρ^* as the product $m\rho$, we have implicitly assumed that m is small enough that we can ignore interference between cross-over events. Note also that, since (by assumption) variation at locus B is neutral, we do not yet need to specify allelic types at this locus. Rather, we can use the convenient coalescent technique, discussed above, of modeling the genealogical and mutational processes separately.

Each of the members of a sample of size n taken at the B locus will be linked either to an A_1 allele or to an A_2 allele at the selected locus, and the same is true of the ancestral lineages of the sample. It is this linkage that makes the ancestry at the B locus differ from the predictions of the standard neutral coalescent. Thus, in modeling the ancestry of the B-locus sample, the appropriate label for each B-locus lineage is the allelic type at the A locus, to which it is linked.

If i B-locus lineages are linked to A_2 alleles, then the rate of coalescence between each pair is $1/x(t)$ and the total rate is identical to the total rate for the A locus given above. This will be true as long as m is not too large, as it neglects the possibility that both recombination and coalescence occur in a single generation. Crucially, B-locus lineages can switch labels as we follow them back in time. This occurs when a lineage ancestral to the sample was the product of a recombination event in an individual who was heterozygous at the A-locus. If i B-locus lineages are linked to A_2 alleles, then the total rate of this type of event at time t in the ancestral process is

$$i\rho^*(1 - x(t))/2$$

with ρ^* as defined above. If an event of this type occurs, one of the *B*-locus lineages switches types at the *A* locus (from A_2 to A_1). To explain the equation above, each of the i lineages hits a recombination event between *A* and *B* with rate $\rho^*/2$, but only $1–x(t)$ of these events occur in heterozygous individuals. There is no additional 2 in the formula, as might be expected given the Hardy-Weinberg proportions implicitly assumed, because we have conditioned on the type of one allele. For *B*-locus lineages that are linked to A_1 alleles, the rate of label switching depends on $x(t)$ instead on $1–x(t)$.

As suggested above, *B*-locus lineages can also escape the sweep due to mutations at the *A* locus. The probability of this depends on the mutation rate (θ_a) at the *A* locus but not on distance *m* between the loci, and over the entire sweep is of order θ_a (Hermisson and Pennings, 2005). For simplicity, we will ignore this possibility.

Figure 4A shows a hypothetical gene genealogy of a sample of size $n = 6$ at the *B* locus under this structured coalescent model, for a population that has experienced a recent sweep at the *A* locus. The allele-frequency trajectory, shown in pink, is from the simulations described above. Blue boxes mark recombination events by which two *B*-locus lineages were able to 'escape' the sweep by switching labels. As a result, these two members of the sample may carry mutations at the *B*-locus that occurred in the ancestral population before the sweep. If there is no recombination, then the entire sample will coalesce during the sweep, and any variation in the sample must be due to mutations that occurred since the sweep. The hypothetical time scale in Figure 4A can be compared to the standard neutral one in Figure 1. The lineages that predate the sweep travel up out of the figure because their expected time to common ancestry is much greater than the range given in Figure 4A. Among these we would expect to see neutral levels of polymorphism. Thus, polymorphism will be reduced at the *B* locus only to the extent that extra coalescent events occur during the sweep. Due to the dependence on $\rho^* = m\rho$, larger reductions will occur when locus *B* is close to locus *A*.

In the ancestral process depicted in Figure 4A, the frequency of A_2 decreases from 1 down to $1/(2N)$ as we follow it back through the sweep, then the single A_2 is converted into an A_1 allele by mutation. For the *B*-locus alleles that are linked to A_2, the rates of coalescence and escape by recombination will *increase* as $x(t)$ decreases. The rate of coalescence increases very dramatically because it depends on $1/x(t)$, while the rate of escape by recombination increases mildly, like $1–x(t)$. These rates, and the changes in $x(t)$, make coalescent analyses of selective sweeps complicated. However, we can see from the large-σ diffusion approximation for t_{fix} that the duration of a sweep will be small on the coalescent time scale, and will become negligible if σ is very large (recall that for $\sigma = 1000$, we have $t_{fix} \approx 0.03$). The results of Etheridge *et al*. (2006) imply that a strong sweep will be divided fairly neatly into two halves. Because of the way the rates of coalescence and escape by recombination depend on the frequency of A_2, we expect most events to occur when $x(t)$ is small. Considered forward in time, small $x(t)$ corresponds to the first half of the sweep, during the convex part of the trajectory.

Beginning with Kaplan *et al*. (1989), a number of workers have considered approximations to $x(t)$, based on different models of how the frequency of allele A_2 increases from its initial frequency of $1/(2N)$. Kaplan *et al*. (1989) used the supercritical branching process, that gave the classical population genetic result $P(\text{fix}) \approx 2s$, to model first part of the trajectory, then followed it with the deterministic model for the middle frequencies, and finally a subcritical branching process for part just before fixation. They

chose frequency cuttoffs of $10/\sigma$ and $1–10/\sigma$ for the boundaries between the three phases, based on the fact that once allele A_2 reaches frequency $10/\sigma$, it is essentially sure to fix.

Later, using the deterministic model over the entire trajectory, Wiehe and Stephan (1993) were able to obtain an analytical result for the decrease is neutral heterozygosity, *i.e*. for a sample of size $n = 2$. Following considerations of Kaplan *et al*. (1989), the formula of Wiehe and Stephan (1993) captures the effects of 'recurrent' selective sweeps, that is where a neutral locus is linked to several selected loci that undergo adaptive fixation events at some rate. There has been a great deal of interest in recurrent selective sweeps, and the formula of Wiehe and Stephan (1993) has been used extensively (Jensen, 2009; Sella *et al*., 2009).

Barton (1998) inserted a fourth phase into the trajectory, between the initial branching process and the deterministic model, based on the finding by Otto and Barton (1997) of an acceleration above deterministic increase over a range of small frequencies of A_2. Barton (1998) obtained a number of new analytical results, also for samples of size $n = 2$, in particular probabilities of identity by descent, and by extension, distributions of pairwise coalescence times.

Eriksson *et al*. (2008) recently suggested modeling sweeps deterministically, but using the average time that the advantageous allele spends in each frequency class in place of the actual deterministic predictions. The authors used a Moran model, but were apparently unaware that some of their results were previously known (Ewens, 1963). In the context of the standard diffusion, their method is equivalent to assuming that the fixation event follows the 'expected trajectory' $t^*(x;p)$ exactly. Although Eriksson *et al*. (2008) offered no mathematical justification for their approach, it has some appeal because $t^*(x;p)$ captures the fact that A_2 moves quickly through the small frequencies.

The Yule Process Approximation

Results for the decrease of heterozygosity are simple and useful, but greater power should be possible in coalescent approaches to samples larger than $n = 2$. For this reason, the computational methods of inference discussed above for neutral models aim to compute likelihoods for any sample. The likelihood captures all of the information in the data. The goal of simulation-based methods like the one of Coop and Griffiths (2004) is to apply this power of the (structured) coalescent approach to inferences about selection. However, because it is necessary to account for the unknown allele frequency in the population as it changes through time, structured coalescent methods for computing likelihoods are computationally costly. In this section, we consider a promising new model called the Yule process approximation.

In addition to pairwise measures, Barton (1998) investigated the distribution of 'family sizes' descending from a sweep, using simulations. Here, families are the descendants (in the sample) of each lineage that emerges from the sweep, looking backward in time. For example, in Figure 4A there are three families, and these have sizes 4, 1, and 1. If we knew the distribution of family sizes, we could derive key quantities, such as the distribution of allele frequencies after a sweep, using coalescent methods rather than forward-time analyses (Kim and Stephan, 2002; Kim, 2006). The Yule process approximation provides a way to generate the numbers and sizes of families that descend from the sweep, and the times of events in the ancestry of the sample.

Durrett and Schweinsberg (2004, 2005) introduced this approximation through an analysis of selective sweeps in a Moran population model. Etheridge *et al*. (2006) approached the same problem starting with the standard diffusion, showing that the Yule process approximation applies to a variety of models, in the limit as $N \to \infty$. In a presentation more accessible to biologists, Pfaffelhuber *et al*. (2006) describe a simulation algorithm for sampling gene genealogies at a neutral locus that is linked to a selected locus, based on a modified version of the Yule process approximation.

Durrett and Schweinsberg (2004, 2005) and Pfaffelhuber *et al*. (2006) assess the accuracy of the Yule process approximation compared to simulations of the discrete Wright-Fisher model and of the structured coalescent model with a deterministic trajectory. A number of authors, including Braverman *et al*. (1995), Simonsen *et al*. (1995), and Przeworski (2002) have used the deterministic model in simulations. In fact, these deterministic simulations are quite accurate for many purposes, especially for small samples, but Pfaffelhuber *et al*. (2006) showed that the Yule process approximation gives better predictions for the distribution of family sizes in larger samples.

The Yule process approximation is derived under the assumption that σ is large, in the model above, of a neutral locus *B* sitting near the selected locus *A*. For very large σ, recall that events in the ancestry of the sample are concentrated in the first half of the sweep (forward in time), when the frequency of A_2 is increasing rapidly. The Yule process approximation is obtained by transforming the diffusion time scale during the sweep by the frequency of allele A_1, $1–x(t)$. The result is that the second half of the sweep becomes greatly compressed; see Figure 1 of Pfaffelhuber *et al*. (2006). On this new time scale, the rate of escape by recombination becomes constant along each ancestral lineage. The process of coalescence between *B*-locus lineages that are linked to A_2 alleles follows from the fact that the sample at the *A* locus can be modeled as a random subsample of a larger random tree, called the Yule tree. If we imagine the whole gene genealogy of all the A_2 alleles that do not go extinct, then roughly speaking, the Yule tree is the portion of this genealogy corresponding to the first half of the sweep.

Figure 4B depicts the model, with a hypothetical Yule tree drawn in the background, in pink, and the lineages that are ancestral to the sample drawn in black. In this representation, the time-change $(1–x(t))$ used in the Yule process approximation has been undone, and the figure is drawn to correspond to the sweep in Figure 4A. Blue boxes again show recombination events by which two *B*-locus lineages escape the sweep. As the range of time on the vertical axis in 4B is the same as in 4A, the three lineages that emerge from the sweep again continue up out of the graph, where they are expected to accrue standard neutral levels of polymorphism. Notice, with respect to Figure 4A, we have dispensed with the allele-frequency trajectory itself.

The process that generates the Yule tree is simple enough, but there is no point to describing it here. We note only that it is a binary tree with $\lfloor\sigma\rfloor$ tips, where $\lfloor\sigma\rfloor$ is the largest integer less than or equal to σ. Importantly, it is not necessary to actually generate the Yule tree, so this approximation relieves us of the detailed, explicit conditioning inherent in the structured coalescent approach, even though we're modeling the same process. However, in generating coalescent times during the sweep using the algorithms in the Appendix Pfaffelhuber *et al*. (2006), it is necessary to model events in the Yule tree and consider whether these events occur in the ancesrtry of the sample. Because of this, simulations of the Yule approximation become slower when σ is larger (Pfaffelhuber *et*

al., 2006), and it might be that other methods are more efficient than the Yule process approximation when σ is very large.

Many mathematical details go into demonstrating the validity of the Yule process approximation and in using it to compute quantities of interest analytically. Readers are referred to Etheridge *et al*. (2006). Note that if none of the *B*-locus lineages escape the sweep, there will be a single family of size *n*. Further let a 'singleton family' be a family with just one member, like the two families descending from the blue boxes in Figures 4A and 4B. Etheridge *et al*. (2006) were able to prove that the probability there will be more than two non-singleton families—one that escapes the sweep (along with some number of singleton families) and one that descends from the original A_2 allele—is of order $1/\log(\sigma)^2$, which tends to zero, albeit slowly, as $N\to\infty$.

Because family sizes are biased toward singletons, approximations have been proposed in which the number of singleton families is a binomial random variable and all remaining lineages descend from the original A_2 allele (Barton 1998; Kim and Nielsen, 2004; Pennings and Hermisson, 2006; Schweinsberg and Durrett, 2005). The heuristic argument for this, which illustrates some important features of strong selective sweeps, is as follows. First, only the first half of the sweep is relevant, and this has length $\sim 2\log(\sigma)/\sigma$ when is very large. Next, the rate of recombination per lineage during this period is effectively $\rho^*/2$ (= $m\rho/2$) per unit of time, because $x(t)$ is very small and $1-x(t)$ is close to one. Thus, in order for there to be an appreciable effect of recombination during a sweep, the product $\rho^*\log(\sigma)/\sigma$ must also be appreciable. This product, which is equal to $mr\log(\sigma)/s$ is the total rate of escape by recombination for a single *B*-locus lineage. The probability that a single lineage escapes the sweep is given by

$$1-\exp\left(-\frac{mr\log(\sigma)}{s}\right)=1-\sigma^{-mr/s}=1-(4Ns)^{-mr/s}.$$

The final expression is written in terms of the discrete model parameters, so that the effects of each can be seen. One possibly counterintuive result is that, for a given selection coefficient, *s*, and for a locus at a fixed recombination distance from the selected site, as measured by *mr*, the probability of escape is *greater* when *N* is greater. Increasing *N* increases ρ proportionally, but does not decrease the duration of sweeps proportionally, due to the $\log(\sigma)$ term in the numerator of t_{fix}.

The Coalescent with Multiple Mergers

We can discern three characteristic times in the processes presented above. The first is the time between new mutations that will fix in the population. The second is the duration of a selective sweep. The third is the neutral coalescence time, for a sample to reach its most recent common ancestor. Only the last of these is simple: the neutral coalescence time does not depend on the rates of mutation and recombination or on the selection coefficient. Again, for a large sample it is close to 2 when time is measured in units of $2N$ generations. In this section, we consider the possibility that the time between sweeps is much smaller than this, such that many sweeps may have occurred within 2 units of time. However, we will require that the duration of a sweep is much smaller than the time between sweeps, so that sweeps are non-overlapping. In this case, the ancestry

of a sample follows a process which is conceptually similar to the neutral coalescent process, but which is very different in detail.

In other words, we consider ancestries of samples of size n under the recurrent hitch-hiking model mentioned above. Coalescence will be driven by the occurrence of selective sweeps rather than by a neutral process of reproduction, by 'genetic draft' rather than by genetic drift (Gillespie, 2000). Gillespie (2000, 2001) illustrated this idea with analyses of samples of size $n = 2$, and Nielsen (2005), Hahn (2008), and Sella *et al.* (2009) promote it as a possible explanation for genomic patterns of variation.

Durrett and Schweinsberg (2005) proved that the gene genealogy at a neutral locus which is embedded in a genomic region where sweeps occur at some rate and at random locations will follow a process known as a coalescent with simultaneous multiple mergers. As the name implies, such processes are distinguished from the standard neutral coalesent because more than two lineages may coalesce at the same time. Multiple-mergers coalescent processes were in fact discovered first in neutral population models, in cases where the variance of reproductive success among individuals is very large (Pitman, 1999; Sagitov, 1999; Schweinsberg, 2000; Möhle and Sagitov, 2001; Birkner *et al*., 2005). An introductory look at the wide range of possible behaviors under one particularly simple neutral model of a population, motivated by organisms that repoduce by broadcast spawning, can be found in Eldon and Wakeley (2006).

Without going into the considerable mathematical details of multiple-mergers coalescent processes for recurrent selective sweeps, we can understand the basic idea from Figure 4C. Genetic lineages at the neutral locus travel backward in time, undergoing a possible burst of coalescent events (with associated family sizes) each time a selective sweep happens at a locus in the vicinity of the neutral locus. In Figure 4C, four lineages coalesce in the first sweep and two escape by recombination, then the remaining three lineages coalesce during the next sweep. Note the much shorter range of time depicted in Figure 4C than in Figures 4A and 4B. Sweeps hit the population very frequently compared the rate of neutral genetic drift or coalescence. The rate at which they occur may be difficult to describe. In a simple model, the rate will depend on the rate of advantageous mutations, but in reality this may in turn depend on changing environments and selection pressures (Gillespie, 1991, 2000, 2001, 2004b). As a consequence, the rate of sweeps may have little to do with the size of the population, so polymorphism levels will not necessarily be predicted to depend linearly on population size, as they are under standard neutral models. A long-standing observation, and conundrum with respect to the neutral theory, is that levels of polymorphism do not in fact increase linearly with estimates of population size (Lewontin, 1974; Gillespie, 1991; Meiklejohn *et al.*, 2007). Thus, while work on these coalescent models with multiple mergers for selection is still in its infancy, they could prove useful in long-standing debates about the origin and maintenance of genetic variation.

HOPES FOR THE FUTURE

Darwin could scarcely have imagined the world we live in today. Only the most active imagination could find passages in the *Origin* relevant to coalescent models of natural selection. However, Darwin's fundamental insights are as relevant today as they ever have been. In the field of population genetics, in particular, attention to natural selection

as a factor shaping variation at the molecular level has been boosted greatly by recent analyses of genomic data.

This shift comes after a perhaps overly long focus on neutral models of genetic variation, received from Kimura (1983). Given the lack of force of theoretical arguments for the neutral theory (Ewens, 1979), the empirical evidence against it and the fact the selective models can both provide a better fit to the observations and mimic neutrality itself (Gillespie, 1991), the longevity of the neutral theory may be surprising. However, as Crow (2008) points out, the very simplicity of the neutral theory accounts for a lot of its appeal. In large part, this is probably due to the wonderful paper of Kimura and Ohta (1971), which seemingly explained both molecular evolution among species and molecular variation within species as different facets of one relatively simple process.

Neutrality, as a logical assumption of a null model with respect to selection, has also been difficult to reject statistically. Perhaps we have been lulled into accepting a false null model. Gillespie (1994) has shown that there is low power to detect crucial deviations from neutrality and to distinguish among some selective alternatives to the neutral theory, at least using simple statistical tests. A similar conclusion applies to multiple-mergers coalescent processes (Sargsyan and Wakeley, 2008). Another issue, which we only glanced at above, is that selective neutrality is just one of several assumptions of a "neutral" model, so rejecting such a model does not necessarily identify selection as the reason. Although choosing among alternatives to neutrality will be a major challenge—few will "wish to slog through 100 pages of mathematics," as Gillespie puts it, in Chapter 4 of *The Causes of Molecular Evolution* (Gillespie, 1991)—we can have some hope that the current rapid pace of research at the interface of theoretical and empirical population genetics will allow more precise inferences to be made.

In closing, we can ask what is to become of the sophisticated coalescent machinery for making inferences about the demographic history of populations. To what extent will we be able to rely on genetic makers containing information about changes in population size over time or patterns of population structure? Will inferences of migration rates or divergence times have to be reinterpreted in terms of selection? We might hope that our inferential tools can be developed in ways that are robust to the presence of natural selection, even for species in which selection is a dominant force.

LITERATURE CITED

Aquadro, C. F. and B. D. Greenberg. 1983. Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* 103: 287-312.

Avise, J. C., C. Gilbin-Davidson, J. Laerm, J. C. Patton, and R. A. Lansman. 1979. Mitochondrial DNA clones and matriarchal phylogeny within and among geographic populations of the pocket gopher, *Geomys pinetis*. *Proc. Natl. Acad. Sci. USA* 76: 6694-6698.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst*. 18: 489-522.

Barton, N. H., A. M. Etheridge, and A. K. Sturm. 2004. Coalescence in a random background. *Ann. Appl. Prob*. 14: 754-785.

Becquet, C., and M. Przeworski. 2009. Learning about modes of speciation by computational approaches. *Evolution* 63: 2547-2562.

Begun, D. J., A. K. Holloway, K. Stephens, L. W. Hillier, Y.-P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. Dewey, L. Pachter, E. Myers, and C. H. Langley. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol*. 5: e310.

Birkner, M., J. Blath, M. Capaldo, A. Etheridge, M. Möhle, J. Schweinsberg, and A. Wakolbinger. 2005. Alpha-stable branching processes and beta-coalescents. *Electron. J. Probab*. 10: 303-325.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140: 783-796.

Brown, W. M. 1980. Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci. USA* 77: 3605-3609.

Bustamante C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416: 531-534.

Coop, G., and R. C. Griffiths. 2004. Ancestral inference on gene trees under selection. *Theoret. Pop. Biol*. 66: 219-232.

Crow, J. F. 2008. Mid-century controversies in population genetics. *Annu. Rev. Genet*. **42**:1-16.

Darden, T., N. L. Kaplan, and R. R. Hudson. 1989. A numerical method for calculating moments of coalescent times in finite populations with selection. *J. Math. Biol*. 27: 355-368.

Darwin, C. 1859. *On the Origin of Species*. Murray, London.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*, 1st ed. Columbia University Press, New York.

Dobzhansky, T. 1955. A review of some fundamental problems of and concepts of population genetics. *Cold Spring Harb. Symp. Quant. Biol*. 20: 1-15.

Durrett, R. and J. Schweinsberg. 2004. Approximating selective sweeps. *Theoret. Pop. Biol*. 66: 129-138.

Durrett, R. and J. Schweinsberg. 2005. A coalescent model for the effect of advantangeous mutations on the genealogy of a population. *Stochast. Proc. Appl*. 115: 1628-1657.

Eldon, B., and J. Wakeley 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172: 2621-2633.

Etheridge, A., P. Pfaffelhuber, and A. Wakolbinger. 2006. An approximate sampling formula under genetic hitchhiking. *Annals of Applied Probability* 16: 685-729.

Ewens, W. J. 1979. *Mathematical Population Genetics*, Springer-Verlag, Berlin. Note: see also the revised and update version, Ewens (2004).

Ewens, W. J. 1982. On the concept of effective size. *Theoret. Pop. Biol*. 21: 373-378.

Ewens, W. J. 1990. Population genetics theory—the past and the future. In S. Lessard (ed.), *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177-227. Kluwer Academic Publishers, Amsterdam.

Ewens, W. J. 2004. *Mathematical Population Genetics*, *Volume I: Theoretical Foundations*, Springer-Verlag, Berlin.

Felsenstein, J. 2007. Trees of genes in populations. In O. Gascuel and M. Steel (eds.), *Reconstructing Evolution: New Mathematical and Computational Advances*, pp. 3-29. Oxford University Press, Oxford.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Royal Soc. Edin*. 52: 399-433.

Fisher, R. A. 1922. On the dominance ratio. *Proc. Royal Soc. Edin*. 42: 321-341.

Fisher, R. A. 1930. *The Genetical Theory of Natural Selection*. Clarendon, Oxford.

Gillespie, J. H. 1991. *The Causes of Molecular Evolution*. Oxford University Press, New York.

Gillespie, J. H. 1994. Alternatives to the neutral theory. In B. Golding (ed.), *Non-Neutral Evolution: Theories and Molecular Data*, pp. 1-17. Chapman & Hall, New York.

Gillespie, J. H. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155: 909-919.

Gillespie, J. H. 2001. Is the population size of a species relevant to its evolution? *Evolution* 55: 2161–2169.

Gillespie, J. H. 2004a. *Population Genetics: A Concise Guide*. 2nd ed. Johns Hopkins University Press, Baltimore, Maryland.

Gillespie, J. H. 2004b. Why $k = 4N_e su$ is silly. In R. Singh and M. Uyenoyama (eds.), *The Evolution of Population Biology—Modern Synthesis*, pp. 181-192. Cambridge University Press, Cambridge.

Hahn, M. W. 2008. Toward a selection theory of molecular evolution. *Evolution* 62: 255-265.

Haldane, J. B. S., 1927. A mathematical theory of natural and artificial selection, Part V Selection and mutation. *Proc. Camb. Philos. Soc.* 23: 838-844.

Haldane, J. B. S. 1932. *The Causes of Natural Selection*. Longmans Green & Co., London.

Halligan, D. L., F. Oliver, A. Eyre-Walker, B. Harr, and P. D. Keightley. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics* 6: e1000825.

Harris, H. 1966. Enzyme polymorphism in man. *Proc. Royal Soc. London, Ser. B* 164: 298-310.

Hey, J. 2005. On the number of New World founders: A population genetic portrait of the peopling of the Americas. *PLoS Biol* 3(6): e193.

Hermisson, J., and P. S. Pennings. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* 169:2335–2352.

Hermisson, J., and P. Pfaffelhuber. 2008. The pattern of genetic hitchhiking under recurrent mutation. *Electronic Journal of Probability* 13: 2069-2106.

Hobolth, A., M. K. Uyenoyama, and C. Wuif. 2007. Importance sampling for the infinite sites model. *Statistical Applications in Genetics and Molecular Biology* Vol. 7, Iss. 1, Art. 32.

Hudson, R. R. 1983a. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37: 203-217.

Hudson, R. R. 1983b. Properties of a neutral allele model with intragenic recombination. *Theoret. Pop. Biol*. 23: 183-201.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. In D. J. Futuyma and J. Antonovics (eds.), *Oxford Surveys in Evolutionary Biology*, Volume 7, pp. 1-44. Oxford University Press, Oxford.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337-338.

Hudson, R. R., and N. L. Kaplan. 1986. On the divergence of alleles in nested subsamples from finite populations. *Genetics* 113: 1057-1076.

Hudson, R. R., and N. L. Kaplan. 1988. The coalescent process in models with selection and recombination. *Genetics* 120: 831-840.

Huxley, J. S. 1942. *Evolution: The Modern Synthesis*. Allen and Unwin, London.

Jensen, J. D. 2009. On reconciling single and recurrent hitchhiking models. *Genome Biol. Evol*. 1: 320-324.

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

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

Karlin, S., and J. McGregor. 1972. Addendum to a paper of W. Ewens. *Theoret. Pop. Biol*. 3: 113-116.

Kim, Y. 2006. Allele frequency distribution under recurrent selective sweeps. *Genetics* 172: 1967-1978.

Kim, Y., and R. Nielsen. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513-1524.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765-777.

Kim, Y., and T. Wiehe. 2009. Simulation of DNA sequence evolution under models of recent directional selection. *Briefings in Bioinformatics* 10: 84-96.

Kimura, M. 1955a. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci., USA* 41: 144-150.

Kimura, M. 1955b. Stochastic processes and the distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* 20: 33-53.

Kimura, M. 1968. Evoutionary rate at the molecular level. *Nature* 217:624-626.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M., and T. Ohta. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61: 763-771.

Kimura, M., and T. Ohta. 1971. Protein polymorphism as a phase of molecular evolution. *Nature* 229: 467-469.

King, J. L. and T. H. Jukes. 1969. Non-Darwinian evolution. *Science* 164: 788-798.

Kingman, J. F. C. 1982a. On the genealogy of large populations. *J. Appl. Prob*. 19A: 27-43.

Kingman, J. F. C. 1982b. The coalescent. *Stochastic Process. Appl*. 13: 235-248.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304: 412-417.

Krone, S. M., and C. Neuhauser. 1997. Ancestral processes with selection. *Theoret. Popul. Biol*. 51: 210-237.

Lewontin, R. C. 1974. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R. C. and J. L. Hubby. 1966. A molecular aproach to the study of genic diversity in natural populations II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54: 595-609.

Linnen, C.R., E.P. Kingsley, J.D. Jensen and H.E. Hoekstra. 2009. On the origin and spread of an adaptive allele in deer mice. *Science* 325: 1095-1098.

Lynch, M. 2007. *The Origins of Genome Architecture*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts.

Malécot, G. 1941. La consaguinité dans une population limitée. *C. R. Acad. Sci., Paris* 222: 841-843.

Malécot, G. 1948. *Les Mathématiques de l'Hérédité*. Masson, Paris. Extended translation: *The Mathematics of Heredity*. W. H. Freeman, San Francisco (1969).

Marjoram, P., and S. Tavaré. 2006. Modern computational approaches for analyzing molecular genetic variation data. *Nature Reviews Genetics* 7: 759-770.

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

Mayr, E. 1963. *Animal Species and Evolution*. Belknap Press, Cambridge, Massachusetts.

Meiklejohn, C. D., K. L. Montooth, and D. M. Rand. 2007. Positive and negative selection on the mitochondrial genome. *Trends in Genetics* 23: 259-263.

Möhle, M. 1999. The concept of duality and applications to Markov processes arising in neutral population genetics models. *Bernoulli* 5: 761-777.

Möhle, M. and S. Sagitov. 2001. A classification of coalescent processes for haploid exchangeable population models. *Ann. Appl. Probab*. 29: 1547-1562.

Moran, P. A. P. 1962. *Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Moran, P. A. P. 1975. Wandering distributions and the electrophoretic profile. *Theoret. Pop. Biol*. 8: 318-330.

Nagylaki, T. 1989. Gustave Malécot and the transition from classical to modern population genetics. *Genetics* 122: 253-268.

Neuhauser, C., and S. M. Krone. 1997. The genealogy of samples in models with selection. *Genetics* 145: 519-534

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet*. 39: 197-218.

Norman, M. F. 1975. Approximation of stochastic processes by Gaussian diffusions, and applications to Wright–Fisher genetic models. *SIAM J. Appl. Math*. 29: 225-242.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96-98.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst*. 23: 263-286.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res., Camb*. 22: 201-204.

Otto, S. P., and N. H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. *Genetics* 147: 879-906.

Pennings, P. S., and J. Hermisson. 2006. Soft sweeps III: The signature of positive selection from recurrent mutation. *PLoS Genet*. 2(12): e186

Pfaffelhuber, P., B. Haubold, and A. Wakolbinger. 2006. Approximate genealogies under genetic hitchhiking. *Genetics* 174: 1995-2008.

Pitman, J. 1999. Coalescents with multiple collisions. *Ann. Probab*. 27: 1870-1902.

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics*, University of Chicago Press, Chicago.

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

Przeworski, M., 2003. Estimating the time since the fixation of a beneficial allele. Genetics 164: 1667-1676.

Sabeti, P.C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander and the International HapMap Consortium. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913-918.

Sagitov, S. 1999. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab*. 36: 1116-1125.

Sargsyan, O. and J. Wakeley. 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoret. Pop. Biol.* 74: 104-114.

Sawyer, S. A., and D. L. Hartl 1992. Population genetics of polymorphism and divergence. *Genetics* 132: 1161-1176.

Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol*. 57 Suppl 1: S154-164.

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

Slatkin, M. 2001. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res., Camb*. 78: 49-57.

Slatkin, M., and M. Veuille. 2002. *Modern Developments in Theoretical Population Genetics: The legacy of Gustave Malécot*. Oxford University Press, Oxford.

Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. 2005. On the meaning and existence of an effective population size. *Genetics* 169: 1061-1070.

Smith, N. G., and A. Eyre-Walker. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415: 1022-1024.

Stephens, M. and P. Donnelly. 2000. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B* 62: 605-655.

Schweinsberg, J., and R. Durrett. 2005. Random partitions approximating the the coalescence of lineages during a selective sweep. *The Annals of Applied Probability* 15: 1591-1651.

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

Thornton, K. R., J. D. Jensen, C. Becquet, and P. Andolfatto. 2007. Progress and prospects in mapping recent selection in the genome. *Heredity* 98: 340-348.

Wakeley, J., and O. Sargsyan. 2009. Extensions of the coalescent effective population size. *Genetics* 181: 341-345.

Wakeley, J. and O. Sargsyan. 2009. The conditional ancestral selection graph with strong balancing selection. *Theoret. Pop. Biol.* 75: 355-364.

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

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97-159.

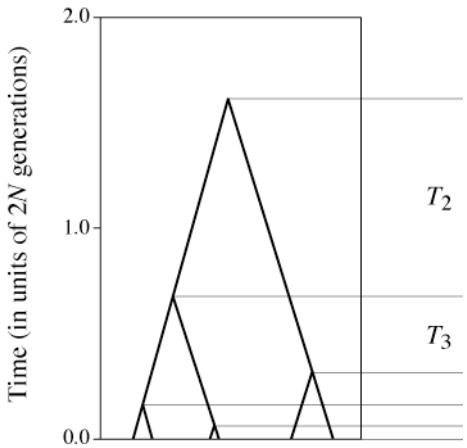
Zuckerkandl, E. and L. Pauling 1965. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*. Academic Press, New York.

FIGURE 1—Example gene genealogy of a sample of size $n = 6$, with coalescence times (T_i on the right) drawn to match expectations from the standard neutral coalescent.

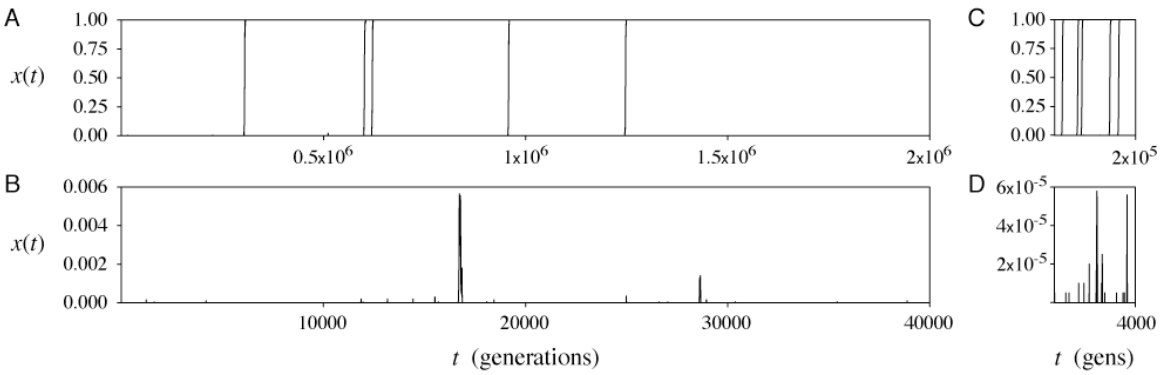
FIGURE 2—Simulated allele-frequency trajectories of advantageous mutants at the hypothetical "human" and "*Drosophila*" loci described in the text. Panels A and B show results for "humans" and panels C and D show results for "*Drosophila*." A & C show advantageous mutations that reached high frequencies. B & D show advantageous mutations that went extinct. Parameter values are described in the text.

FIGURE 3—Example population ancestries, in which a selective sweep has just reached completion. Panel A is from the simulations depicted in Figure 2A, and panel B is from the simulations depicted in Figure 2C.

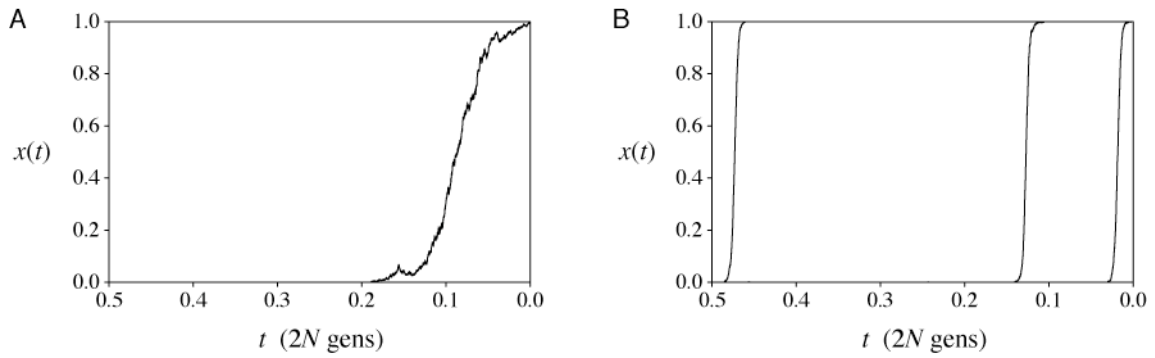
FIGURE 4—Hypothetical gene genelogies for a sample of size $n = 6$ at a neural locus linked to a selected locus, showing the three coalescent approaches to modeling natural selection described in the text. Panel A depicts the structured coalescent approach, B depicts the Yule process approximation, and C depicts the mutliple-merges coalescent for recurrent selective sweeps. Possible characteristic ranges of time (measured in units of $2N$ generations) for each model are displayed on the veritical axes. Black lines show the ancestry of the sample, while unobserved allele-frequency trajectories and genetic lineages not directly ancestral to the sample are shown in pink. Blue boxes mark recombination events that allow linked neutral lineages to "escape" a sweep.



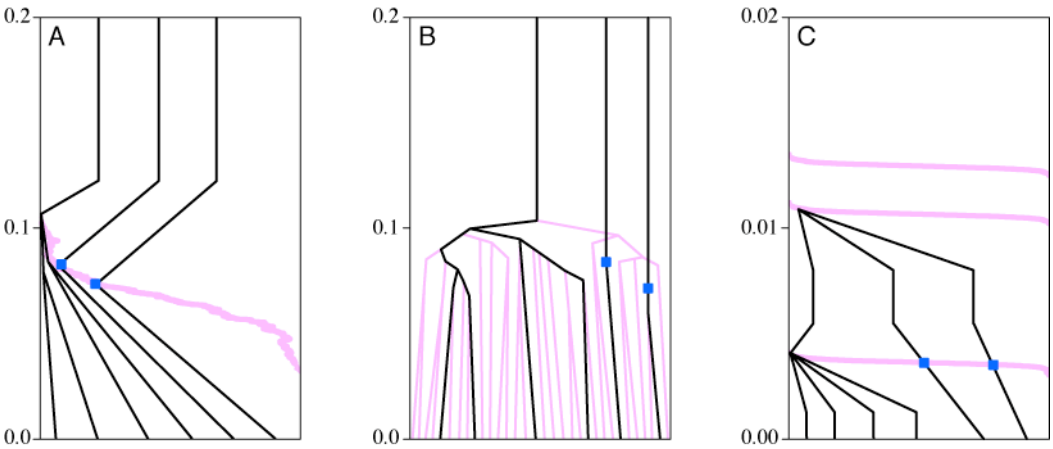
(figure 1)


A
1.00
0.75
0.50
0.25
0.00
x(t)
0.5x10^6
1x10^6
1.5x10^6
2x10^6
C
1.00
0.75
0.50
0.25
0.00
2x10^5
B
0.006
0.004
0.002
0.000
x(t)
10000
20000
30000
40000
t (generations)
D
6x10^-5
4x10^-5
2x10^-5
4000
t (gens)

(figure 2)


A
1.0
0.8
0.6
0.4
0.2
0.0
x(t)
0.5
0.4
0.3
0.2
0.1
0.0
t (2N gens)
B
1.0
0.8
0.6
0.4
0.2
0.0
x(t)
0.5
0.4
0.3
0.2
0.1
0.0
t (2N gens)

(figure 3)



(figure 4)