Natural Selection and Coalescent Theory

John Wakeley
Harvard University

> We cherish Darwin for we owe to him our enlightened view of the nature of living things; our civilization would be pitifully immature without the intellectual revolution led by Darwin, even if we were equally well off economically without it. H. J. Muller (1960), in celebrating the hundredth anniversary of the publication of *The Origin of Species*, remarked that it can justly be considered as the greatest book ever written by one person.
>
> Motoo Kimura, *The Neutral Theory of Molecular Evolution*, 1983 (p. 2).

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.

DARWIN AND THEORETICAL POPULATION GENETICS

Anyone who has taken a course in population genetics will likely have learned about the origins of mathematical theory in this field (Provine, 1971). Darwin himself became obsessed with the maintenance of variation, especially after Fleeming Jenkin pointed out that if the commonly held notion of blending inheritance were true, variation would erode so quickly that natural selection would be ineffective. In seeking a mechanism for the promotion of variation, Darwin went so far as to advance his own theory of 'gemmules' which were supposed to circulate through the body and cause the inheritance of acquired

characters. The solution to this conundrum lay in the discovery of the particulate nature of inheritance by Mendel (1865) together with the later the mathematical proof that allele frequencies will not change over time, and thus that genetic variation has no inherent tendency to decrease, in an infinite population (Hardy, 1908; Weinberg, 1908). It would have saved Darwin a great deal of anxiety to have known about and appreciated Mendel's work, but unfortunately he was unaware of it.

In fact, Mendel's work was obscure to the great majority of scientists before 1900. Interestingly, the rediscovery of the nature of inheritance at the beginning of the 20th century by Hugo de Vries, Carl Correns, and Erich von Tschermak, although it did lead to the important works of Hardy and Weinberg, sparked further controversy, about the nature of evolutionarily important variation and the importance of natural selection. One camp maintained that evolution proceeded in jumps, as might potentially follow from the particulate inheritance of large mutations, while the other camp asserted that it moved via gradual modification of continuous characters by selection. The latter view was promoted vigorously by the 'biometricians' W. R .F. Weldon and Karl Pearson, while the former was argued equally forcefully by the 'Mendelians' William Bateson, de Vries, and others. Alongside accumulating empirical evidence, Ronald Fisher (later Sir Ronald Fisher) produced a wonderful mathematical resolution to this conflict when he demonstrated that the continuous variation studied by the biometricians follows from a multifactorial model of Mendelian inheritance (Fisher, 1918). Notably, this is also the paper in which Fisher initially proposed the technique called analysis of variance.

Subsequent papers by Fisher, and also by Sewall Wright and J. B. S. Haldane, produced the three great initial works of theoretical population genetics (Fisher, 1930; Wright, 1931; Haldane, 1932). In them, Fisher, Wright, and Haldane established the fundamental dynamics of the evolutionary process, of changes in allele frequencies through the interaction of mutation, selection and random genetic drift. These seminal works are still a crucial part of any advanced education in evolutionary biology. The strange 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.

Aspects of the fundamental dynamics of changes in allele frequencies under selection and genetic drift are reviewed later in this chapter, and considered together with recombination with the goal of modeling the effects of these processes on genomic patterns of variation. One early result deserves to be mentioned first here because it is quite remarkable. Consider the probability of fixation of a new mutant allele under the influence of positive natural selection. In particular, imagine a population of individuals of some species, in which initially every individual has genotype A_1A_1. A mutation produces a new allele, A_2, which gives its carriers an advantage over A_1A_1 individuals. If A_1A_2 individuals have an average of $1+s$ offspring and A_2A_2 individuals have an average of $1+2s$ offspring, relative to A_1A_2 individuals, then the probability that the new mutant allele A_2 goes extinct is approximately equal to $1–2s$. This result holds when s is small relative to 1 and the population size, N, is very large ($Ns>>1$). Following Haldane (1927) and Fisher (1922, 1930), it can be derived using a supercritical branching process model, in which each A_2 allele has a Poisson number of descendants with mean $1+s$ each generation. As we will see later, it can also be obtained using diffusion theory.

The probability of fixation is the probability that eventually the entire population will have genotype A_2A_2, so that every allele is a descendant of the initial mutation. In a finite population it is equal to one minus the probability of extinction. Then under the assumptions stated, the probability of fixation of the mutant A_2 is equal to $2s$, which is small. Although it is not clear whether evolution proceeds in such a simple manner, by the introduction and fixation of single mutations, 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 continues to be 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 (including other organisms)—our modern understanding of genetics is quite detailed. This has led to significant improvements of the models of population genetics, away from the simple A_1, A_2, etc., allelic models, 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 'evolutionary 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. At the same time, the mathematical theory of population genetics developed substantially, but largely in the absence of data about the genetic variation it purported to explain (Lewontin, 1974). It is now common to recognize two additional seminal figures of mathematical population genetics from the mid-twentieth century: Gustave Malécot and Motoo Kimura. Among many important contributions (Nagylaki, 1989; Slatkin and Veuille, 2002), in developing the theory of identity by descent, 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 modeling, which is discussed in detail in the next section. Kimura is probably 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. Among many important results, Kimura obtained the full, time-dependent solution of the stochastic frequency trajectory of an allele subject to selection and genetic drift (Kimura, 1955a,b).

Recognizing Malécot and Kimura is certainly appropriate, but interested readers will find a great deal of ground-breaking and relevant work, by a number of other, equally impressive thinkers in the books by Ewens (1979, 2004) and Moran (1962).

DIFFUSION THEORY

The results of diffusion theory will be used a number of times in this chapter, so a brief review of the setting and basic concepts of the models will be helpful. For an excellent, thorough treatment of diffusion theory in population genetics, see Ewens (1979, 2004). Additional motivation for this brief review comes from the fact that the fundamental assumptions of coalescent theory and diffusion theory are the same.

Diffusion models approximate the dynamics of allele frequencies over time in large populations. To explain, 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 allele A_2 introduced previously) 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 $p = k/(2N)$. Note: the letters p and x are commonly used to denote allele frequencies. In a diffusion model, both time and allele frequency are measured continuously: $p \in [0,1]$, $x \in [0,1]$, 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 or x 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.

Different 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 in the population. The most commonly used one is the Wright-Fisher model (Fisher, 1930; Wright, 1931), although the Moran model is employed often as well (Moran, 1962). Besides tractability, another advantage of using 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 (some) other populations that do not conform to the overly simple Wright-Fisher model.

The qualifier (some) refers to the fact that 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 depend on parameters usually denoted s, u, and r. We saw the definition of s above, and we now define u and r to be the probabilities of a mutation at a single nucleotide site and a recombination event between two adjacent nucleotide sites, respectively, between a parent and its offspring (*i.e.*, per genetic lineage, per generation). Clearly, 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 $4Ns$. Keeping track of symbols and factors of two accounts for approximately 5% of research time in theoretical population genetics.

Kimura's (1955a,b) groundbreaking achievement was to obtain the probability density function of the frequency, x, of an allele at future time, t, given that its initial frequency is p, at a single locus under the influence of natural selection and random genetic drift. We can think of the advantageous allele A_2, whose fixation probability we considered above. The probability density is often written $\phi(x;p,t)$ and is a complicated function which depends on σ in addition to p and t. It may be helpful to note that

$$\int_0^1 \phi(x;p,t)dx = 1$$

in seeing that this is a probabilistic prediction, as it must be since it includes the influence of random genetic drift. Thus, $\phi(x;p,t)dx$ is the probability that the frequency of A_2 is in the interval $(x,x+dx)$ at time t given that it started at frequency p at time zero. 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 using diffusion theory, by himself and others, which is reviewed in Ewens (1979, 2004).

NEUTRAL COALESCENT THEORY

Kimura's initial use of diffusion theory flowed out of his desire to explore the dynamics of genetic drift, which Wright had promoted as having a dramatic role in evolution. However, the focus of population genetics soon shifted to explaining 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 change 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 the importance of 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.

Richard Lewontin's indispensable text (Lewontin, 1974) provides a clear analysis of how the 'classical' and 'balance' schools of thought (Dobzhansky, 1955), which had adopted contrasting selectionist theories of variation, gave way to the neutral theory when faced with explaining the extraordinarily high levels of polymorphism and divergence which we accept as given today. Crow (2008) recounts the arguments wonderfully, from the key perspective of someone whose career spanned this and other controversies. Lewontin (1974) argued against an unbridled focus on neutrality. He preferred the term 'neoclassical theory' over 'neutral 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, as 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). Even tempered in this way, the proposal of the neutral theory and the routine gathering of genetic data led to a major proliferation of methods and results from mathematical population genetics.

The basic models of population genetics had been in place since the 1930s, but they had never been tuned to the problem of statistical inference presented by modern genetic data. If we wish to infer the action of natural selection, then neutrality is the appropriate null hypothesis. To establish the null predictions, mathematical models soon commonly included the assumption that *all* genetic variation at a locus was neutral. Notably, they also included increasingly refined assumptions about the mutation-structure of variation, in an effort to be appropriate to the data at hand. 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. However, a major shift in orientation was required, from the earlier prospective view of forward-time population dynamics to the retrospective view of ancestral processes, where time flows from the present back into the past (Ewens, 1990).

Shortly after allozyme data began pouring in, Ewens (1972) described his remarkable sampling formula for selectively neutral alleles under the 'infinite-alleles' mutation model, which assumes that every mutation produces a new allele and that no recombination occurs within the locus under study. At about the same time, Ohta and Kimura (1973) initiated work on the step-wise mutation model, or charge-state model, which may be more suitable for allozyme data, but which does not yield a straightforward sampling formula. The distribution obtained by Ewens (1972) gives the probability of the number of alleles and their frequencies in a sample of size n from the population. Ewens obtained the formula using the results of diffusion theory for the infinite-alleles model. However, the formal proof of the result, by Karlin and McGregor (1972), has embedded in it the fundamental structure of gene genealogies under neutrality. This is nicely illustrated in section 2.3 of the recent paper by Hobolth *et al*. (2008) as well as in section 3 of Kingman (1982a).

The paper by Watterson (1975) is the earliest in which gene genealogies and their relationship to genetic data are easily recognizable. Figure 1 shows a hypothetical gene genealogy of a sample of size $n = 6$. It is drawn as an upside-down binary tree which traces the ancestral lines of the sample back (up) to their most recent common ancestor,

with time measured by vertical distance. Each branch in the tree depicts all of the direct genetic ancestors of particular members of the sample, so that polymorphism in the data must be due to mutation along one or more branches. Watterson's analysis pertained to DNA sequence data, although population samples of DNA sequences were not yet available; see also Ewens (1974). By separately considering mutations on different branches of the tree, Watterson derived the expectation and variance of the number of polymorphic nucleotide sites in a sample, under the assumptions that all variation is at the locus is neutral and that every mutation occurs at a previously unmutated site. This 'infinite-sites' mutation model is still used for DNA sequence data in population genetics, although is it common nowadays to allow intra-locus recombination.

Subsequent developments in the technologies for measuring population genetic variation—from restriction-enzyme digests of mitochondrial DNA (Avise *et al*. 1979; Brown, 1980), to early DNA sequence data (Aquadro and Greenberg, 1983; Kreitman, 1983), to the massive contemporary genome-sequencing effort, such as the 1000 Genomes Project, which is nearly complete, and the Personal Genome Project, which aims to sequence the entire genomes of 100,000 human beings—are too numerous to detail here. Likewise, it is beyond the scope of this chapter to give a detailed review of the mathematical developments in the study of ancestral processes in population genetics. Some of this work is discussed in what follows, but readers may further consult Hein *et al*. (2005) and Wakeley (2008). The following sketch of the standard neutral 'coalescent' (Kingman 1982a,b) and its present-day uses will give the background necessary for understanding coalescent models with selection.

The coalescent process is random genetic drift viewed in reverse, starting with a sample of size n and following the ancestral lines back in time until they reach their most recent common ancestor. Again, Figure 1 shows the result, which is a gene genealogy. Remarkably, under the assumption that all variation is selectively neutral, it is possible to model just the ancestors of the sample, and ignore the other members of the population. Under the same assumptions made in diffusion theory—but without selection, and for the moment restricting ourselves to a non-recombining locus—each pair of ancestral lines coalesces, or reaches its common ancestor, 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. The coalescent and the diffusion are inextricably related as *dual* processes of each other; for mathematical details, see Möhle (1999). Note that 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 are between pairs, and not larger numbers, of alleles.

Because every pair of ancestral lines coalesces with rate equal to one, neutral gene genealogies are random-joining (binary) trees and the time, T_i, during which there exist exactly i lines ancestral to the sample is exponentially distributed with mean equal to $2/(i(i\text{-}1))$, or the inverse of the number of possible pairs of i lines. On average then, neutral gene genealogies have very short branches close to the tips, and much longer branches closer to the root, or the most recent common ancestor of the entire sample. Even for very large samples, T_2 takes up about half of the time to the most recent common ancestor ($T_{\mathrm{MRCA}} = T_n + T_{n\text{-}1} + \ldots + T_2$). The gene genealogy in Figure 1 is drawn with coalescences times, T_i, equal to their expected values. On the coalescent or diffusion time scale, the expected value of T_{MRCA} is given by

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

which can be translated into Wright-Fisher-model generations by multiplying by 2*N*. For large samples this converges to 4*N* generations, which is not unexpected as this is also the expected time to fixation of a neutral mutation obtained using forward-time diffusion theory (Kimura and Ohta, 1969).

One more fact about coalescent modeling deserves mention. Namely, on the coalescent or diffusion time scale, neutral mutations occur along each branch of the gene genealogy with rate equal to $\theta/2$ per site. This, together with the random-joining tree structure and the exponential distributions of coalescence times, T_i, enables us to predict the sampling properties of gene genealogies with mutations, and thus to predict any aspect of neutral genetic variation (although computer simulations may be required). In modeling or simulation, we simply generate the tree and the times, then put mutations randomly with rate $\theta/2$ per site along each branch. Along with a choice of mutation model, this standard coalescent is an efficient mathematical tool for predicting distributions of neutral variation in samples of genetic data.

Kingman (1982a,b) gave the mathematical proof of the coalescent process, and later (Kingman, 2000) cited work on the step-wise mutation model (Moran, 1975; Kingman, 1976; Kesten, 1980) as being instrumental to his realization of the existence and nature of gene genealogies. Independently, biologists were introduced to the theory of gene genealogies by initial the papers of Hudson (1983a) and Tajima (1983), as well as the first review of the topic by Hudson (1990). Hudson's and Tajima's derivations of the theory were motivated by the biological relevance of gene genealogies. Hudson (1983a) described the basic model and, importantly, devised an algorithm for simulating gene genealogies, which expanded into a widely used program (Hudson, 2002). Tajima (1983) derived many now-classical results about the structure of gene genealogies and the concomitant effects on the sampling properties genetic variation. Tavaré's early synthesis and review (Tavaré, 1984) goes some way toward bridging the gap between the biological and the mathematical, and between diffusion theory and coalescent theory.

Right at the start of this new, retrospective age of population genetics, Hudson (1983b) also described the effect of recombination on the coalescent process, and how to simulate gene genealogies with recombination. This complicates the model 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 organisms. Table 4.1 in Lynch (2007) provides a summary. Therefore, a growing number of neutral coalescent approaches to inference take recombination into account; for example, see Becquet and Przeworksi (2009). As we will see in the next section, it is absolutely critical to consider recombination when natural selection is added to the model.

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 can contribute to broad debates about ancient processes and events—the genetic analysis of the peopling of the Americas by Hey (2005) is one of many examples—and the fundamental idea that population genetic data contain information about the past is certainly true. Still, the strict neutrality of all mutations that made its way into these methods, as a logical assumption of the null model, is at odds with some major features of molecular population genetics and evolution. It explains neither the high degree of variation in divergence (the 'over-dispersed molecular clock') nor the weak dependence of levels of polymorphism on population size, which through $\theta = 4Nu$ is predicted to be essentially linear (Lewontin, 1974; Gillespie, 1991, 2001). So, we should be open to the possibility that our framework for inference will need to be modified, perhaps drastically so, as we gather more information about the genetics of populations.

GENOMIC DATA AND THE MODELING RESPONSE

As Lewontin (1974) pointed out, the discovery that some genetic loci have been the targets of selection does not invalidate the neoclassical view. We have already seen that Kimura accepted substantial amounts of selection without rejecting his neutral theory of molecular evolution. However, current genomic data, in particular from *Drosophila* species, suggest staggering amounts of positive selection, such that every aspect of polymorphism and divergence might be affected. We will not review all of the data and associated methods here; interested readers may find details and extensive references in the recent reviews by Thornton *et al*. (2007), Hahn (2008), and Sella *et al*. (2009). 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. Large fractions of positively selected substitutions (~50%) have also been reported for non-coding regions (Begun *et al*., 2007). While Thornton *et al*. (2007) and Jensen (2009) remain agnostic, and caution against drawing strong conclusions based on current methodologies and data, Hahn (2008) argues that we need a new theory in which selection plays the major role, and Sella *et al*. (2009) agree that at least some parts of the neutral theory, or neoclassical theory, are in serious need of an overhaul.

These dramatic conclusions do not seem to hold broadly for 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, which is remarkably similar to the initial estimate that was considered broadly consistent with the neutral theory (Ohta and Kimura, 1971; Kimura, 1983). Also, a major tenet of population genetics, which sits at the base of neutral theory, is not in question: namely that 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. Notably, a principal method 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 when sites may be assumed independent of one another (Sawyer and Hartl, 1992; Sawyer *et al*., 2003)

Positive natural selection for adaptive function has been a primary source of excitement among workers in this field. In addition to estimating the overall prevalence of positive selection, much effort has been put toward the identification of particular loci that have been the recent targets of positive selection. This is possible because the fixation of an advantageous allele at one locus has an effect on other, nearby loci, a phenomenon known as genetic 'hitch-hiking' (Maynard Smith and Haigh, 1974; Kaplan *et al*. 1989). The main effect is a reduction in variation around the site of selection, but the frequencies of alleles at single loci and associations between alleles at two or more loci can also be affected; see the review by Nielsen (2005). The term 'selective sweep' is used loosely to mean the fixation of a positively selected allele or the effects of a fixation event, in particular the reduction in variation. Thornton *et al*. (2007) review recent genomic scans for selective sweeps in *Drosophila*, which have identified large numbers of recently swept loci. In humans, Williamson *et al*. (2007) suggest that recent hitch-hiking affects 10% of sites in the genome. Although there are many unresolved issues—for example, that demographic factors can lead to false positive inferences of selection (Thornton *et al*., 2007) and that 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

In view of (1) the outright complexity of these issues, (2) the fact that even a 10% fraction of positively selected fixations may seem incompatible with neutrality, and (3) because a deeper look at the diffusion of selected alleles is helpful for understanding coalescent models of selective sweeps, this section contains some relevant mathematical and simulation results.

A preliminary question to ask is whether the assumptions of diffusion theory are at all reasonable for loci undergoing positive selective sweeps. As we encountered already, diffusion theory is based on the assumptions 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. In practice, simulations may be used to show that many results of diffusion theory are very accurate even for moderate values of the discrete-model parameters, such as $N = 100$ and $s = 0.01$ (likewise u and r). 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) who estimate $s = 0.0056$ for a recently swept allele affecting coat 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). Thus, the diffusion with large σ appears to be good starting point for modeling selective sweeps.

It is important to note that there is an entirely different diffusion model in population genetics (Norman, 1975), which in fact may be more appropriate for large σ. Unfortunately, few results are available for this "Gaussian" diffusion model. As Ewens (1979, 2004) notes, this Gaussian diffusion and the standard one should overlap for certain parameter values (*i.e*. large values of σ). For some illustrations of the similarities under strong selection and mutation, see Wakeley and Sargsyan (2009).

Armed with our 'standard' diffusion model, for which we can drawn upon Ewens (1979, 2004) for a wealth of results, we can understand the complicated dynamics of allele frequencies under the influence of random genetic drift and strong selection. We define a sweep as the event that a positively selected mutation, which starts in frequency $1/(2N)$ as a new mutation, fixes in the population (*i.e*. reaches frequency 1). We will also assume that all parameters are constant over time; more complicated scenarios will be discussed later. 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 trajectories. For example, the average duration of the sweep is identical to the expected value of the length of the allele-frequency trajectory. Other quantities, such as the probability of coalescence during a sweep, also depend on the characteristics of allele-frequency trajectories.

For the purposes of illustration, and with the emerging estimates from *Drosophila* and humans as a backdrop, let us imagine a locus made up 1000 'deleterious' sites, 1000 'neutral' sites, and a single 'advantageous' site. Estimates of θ for humans are on the order of 0.001 and estimates of the effective population size of humans 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 humans. The total rates of mutation are then $\theta_d = \theta_n = 1.0$ (*i.e*., 1000×0.001) and $\theta_a = 0.001$ for 'deleterious,' 'neutral,' and 'advantageous' sites. Let us also assume fairly strong selection, in particular $\sigma_d = -100$ and $\sigma_a = 100$. With $N = 10^4$, this corresponds to $|s| = 0.0025$ for deleterious and advantageous 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 the sake of illustration and computational efficiency, let us get our "*Drosophila*" parameters simply by multiplying the "human" diffusion-scale parameters by ten, so that $\theta_d = \theta_n = 10.0$, $\theta_a = 0.01$, $\sigma_d = -1000$, and $\sigma_a = 1000$. In the simulations presented below, this is realized by using the same discrete-model, per-generation parameters as for humans, but with $N = 10^5$ instead of $N = 10^4$.

Our model is purposely abstract, but will serve to generate some intuition about selective sweeps and about the relative magnitudes of the processes involved. In relation to the estimates of rates of adaptive substitution in humans and *Drosophila*, with the parameters given, 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 use the $s \rightarrow s/(1+s)$ correction suggested by Bürger and Ewens (1995) in their analysis supporting the applicability of diffusion theory to fixation probabilities of alleles in small copy number (here the mutant is in 1 copy).

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, and in both cases deleterious mutations are exceedingly unlikely to fix.

Figure 2 shows the trajectories of alleles in simulations of our "human" model (Figure 2A–D) and our "*Drosophila*" model (Figure 2E–H). The upper two panels in each case show the trajectories of advantageous alleles; the lower two panels in each case show the trajectories of deleterious alleles. In the simulations, which are done separately for advantageous and deleterious alleles: (1) 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 (2) the origination times of the mutations were generated using the per-generation population rates of mutation, $\theta_a/2$ for advantageous mutations or $\theta_d/2$ for deleterious mutations. Thus, these simulations are of independent trajectories; they do not take interference between alleles into account. This is reasonable for "human" panels A and B, where successful sweeps are fairly well separated in time (A) and alleles go extinct quickly when sweeps fail (B), 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 described above.

Before looking in detail at Figure 2, note that our model and simulations use 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 $1+2s$ to back 1 at the conclusion of the sweep, and the next mutation again has selection coefficient s. We have also assumed that selection is constant change over time, which might not be realistic even for a single sweep. John 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. These are extremely important

issues, and Gillespie has argued that they are fatal to neutral-theory explanations of the molecular evolution. For us, they are of somewhat less concern because our ultimate focus is the much shorter time scale of single sweeps. Our goal is to understand coalescent models with selection, which importantly also assume two alleles, so it is appropriate that we consider two-allele models here.

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

Panels A and E 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 (over a shorter period of time) as in panels B and F, the trajectories of most alleles are difficult to see. Recall than only 5×10^{-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 allele frequencies.

Relatedly, panels C, D, G, and H show the frequency trajectories of deleterious alleles, which for these values of σ_d (–100 and -1000) will essentially never fix in the population. They enter the population and may drift to appreciable frequencies (*e.g.*, one reaches a frequency of nearly 2% in the "human" panel C) but are lost quickly. Note the different ranges of time for panels C, D, G, and H versus those in A, B, E, and F. The majority spend just one generation in the population, at frequency $1/(2N)$, then are lost (panels D and H). In contrast to the advantageous alleles depicted in panels A, B, E, and F, which are infrequent and whose trajectories mostly don't overlap, the deleterious alleles in panels C, D, G, and H are introduced at a much higher rate (1000×) and several are segregating in the population at any given time. Although, as noted above, we have not dealt properly with the co-segregation of these alleles, for our purposes there is little error in thinking of them as low-frequency variants segregating at the 1000 'deleterious' sites in our locus.

Within our "humans," neither advantageous nor deleterious mutations will greatly affect levels of polymorphism, at least typically. Visual inspection of panels C and D suggests there is a chance of about 1% of observing 1 deleterious polymorphic site in a sample of size $n = 2$ at our locus. These will contribute little to overall levels of polymorphism because at the same time we expect 1 neutral polymorphism ($\theta_n = 1$) in the same sample. Even in the face of the positive selection at this locus, we expect roughly neutral levels of polymorphism because sweeps occur on average only every 400,000

generations while the effective population size is 10,000. However, the same cannot be said of our "*Drosophila*," in which sweeps occur at the locus every 40,000 generations and the effective population size is 100,000.

Methods for and analyses 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, since the *a priori* time back to the last sweep at any given locus 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, but we can gain some better intuition about them by skipping back to the end of the last sweep that occurred in each case, and changing the time scale. 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 models. In contrast to Figure 2, only those trajectories that went to fixation are shown in Figure 3.

Figure 3 illustrates a number of points. First, as is broadly appreciated by those familiar with population genetics, 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. Second, in species like our "*Drosophila*" depicted in Figure 3B, the frequency of selective sweeps might be such that several will have occurred in the recent ancestry of the 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). As we will see shortly, this can have rather drastic consequences for gene genealogies. Third, with time measured in units of $2N$ generations, more strongly favored alleles sweep more quickly through the population ($\sigma_a = 100$ in A versus $\sigma_a = 1000$ in B).

We can use the diffusion model, with large σ, to attain a deeper understanding of these points concerning the time scale of selective sweeps. 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 is that

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

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

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

when σ is large. The symbol γ above is Euler's constant (approximately 0.5772). Note that *s* in Ewens (1979, 2004) is equivalent to our 2*s*, so α = 2*Ns* in Ewens is equivalent to our σ. The equation above gives ~0.2 for the fixation time when σ = 100, and ~0.03 when σ = 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 provides a nice illustration of a fact that has been known for several decades, namely that deterministic equations for allele-frequency trajectories drastically overestimate the amount of time an allele will spend in small frequencies (*e.g*., Ewens (1979) page 149). For example, if an allele, even an advantageous one, is going to sweep to fixation, it must move away from the boundary (*x*=0) very quickly. Fundamentally this is a consequence of the stochastic nature of random genetic drift, and complements the classical finding that a favored allele has probability 1-2*s* of going extinct. Still, it is not uncommon to see deterministic results and methods in the biological literature. For example, 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 2*N* 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(2*N*) is much greater than both –log(2*s*) and γ. For any values of *s* we are likely to consider, the deterministic result will overestimate the diffusion result. For our "human" model, the diffusion result gives 4146 generations and 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 diffusions 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 to fixation that the allele spends in what we might call the 'middle frequencies' becomes negligible; see their Lemma 3.1 and note that their α is our σ/2. Specifically, the time spent going from frequency ε to 1 – ε becomes negligible, for *any* 0 < ε < 1, so that the allele spends half of t_{fix} in the interval (0,ε) and the other half in the interval (1-ε,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).

The results presented above will foster for our investigation of coalescent models with selection. We will focus on selective sweeps, and will consider ancestral processes at other, neutral loci that sit near loci under selection. 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 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, as sweeps will occur only every $20\times 2N$ generations on average. 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 are uncovering, and coalescent models are being developed to aid both in their identification as well as to make estimates of the strength and timing of selection.

COALESCENT MODELS WITH SELECTION

The relative simplicity of the standard coalescent flows from the 'exchangeability' of genetic lineages under neutrality (Kingman, 1982c). Exchangeability means invariant upon permutation, or relabeling, and is the source of the fact all pairs ancestral lineages to have the same rate of coalescence in the standard neutral coalescent model. When selection operates—for example with the two alleles A_1 and A_2 we considered in previous sections—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 in some manner. In this section we will review four different approaches to this problem, guided mostly by our concern for treating selective sweeps.

A second key issue in modeling genetic ancestries with selection is that other loci situated near a locus under selection will also be affected. As alluded to above, the genomic extent of the effects of selection has been the focus of much recent interest. 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 to coalescent approaches to selection. Note that here "linked" and "linkage" refer to physical proximity on a chromosome, rather than to statistical associations that can develop between any pair of loci (as in "linkage disequilibrium").

The Structured Coalescent Approach

In this section we consider the ground-breaking work of Hudson, Kaplan, and colleagues, who extended coalescent models to neutral loci that are linked to loci under selection. It is called the structured coalescent approach to selection by analogy with models of geographic subdivision and migration. Allele frequencies take the place of relative subpopulation sizes, and recombination (or mutation) takes the place of migration between subpopulations. In contrast to migration models, in this case we are concerned with particular kinds of *changes* relative subpopulation sizes, such as when an advantageous allele sweeps through the population, and we know that it may be important to account for the stochasticity in these changes due to random genetic drift.

Hudson and Kaplan (1986) showed how conditioning on the allelic types of the sample alters the coalescent process. 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. 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; and 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.

General analytical results are difficult to obtain, but the 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 these, an allele-frequency trajectory is generated, then the structured coalescent process is run, conditional on the trajectory. A main goal in developing these simulations has been 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 both the issues involved and the 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

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

Then 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, meaning that coalescence happens faster. The essential 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, so that the total scaled rate of recombination is

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

where ρ is the per-site rate of recombination we defined before. In particular, recall that $\rho = 4Nr$, and that ρ/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 need to specify allelic types at this locus. Rather, we can use the convenient technique of separately modeling the genealogical and mutational processes which, as discussed above, is common practice in standard neutral coalescent models.

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* 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 following them back into the past. 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. If this occurs, it means that more than one A_2 mutant contributed to the sweep (Hermisson and Pennings, 2005). The rate of this type of switching depends on the mutation rate (θ_a) at the *A* locus, and is not modulated by the distance *m* between the loci. The probability of escape by mutation 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 (and only one mutation gave rise to A_2), 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 considering how to model the ancestral process depicted in Figure 4A, it is clear that the frequency of A_2 will decrease from 1 down to $1/(2N)$ as we follow it back through the sweep, and then it will disappear by mutation. For the B-locus alleles that are linked to A_2, the rates of coalescence and escape by recombination, which are given by the formulas above, 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, as $1 - x(t)$. The combination of these time-inhomogeneous 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 $t_{fix} \approx 0.03$). Further, the result of Etheridge *et al*. (2006) quoted above implies that a strong sweep will be divided fairly neatly into two halves. Then, 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, this corresponds to the first half of the sweep, during the convex part of the allele-frequency trajectory (*e.g*., see Figure 3).

Beginning with Kaplan *et al*. (1989), a number of workers have considered approximations to $x(t)$, based on different models of how the allele A_2 moves away from its initial frequency of $1/(2N)$, then races toward the middle frequencies. 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, followed by the deterministic model for the middle frequencies, and finally a subcritical branching process for the final part of the trajectory to fixation. They chose critical frequencies 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 virtually guaranteed to fix.

Wiehe and Stephan (1993) used a deterministic model for the entire trajectory, and were able to obtain an analytical result for the decrease is neutral heterozygosity, *i.e*. for a sample of size $n = 2$. Following Kaplan *et al*. (1989), the formula of Wiehe and Stephan (1993) captures the effects of 'recurrent' selective sweeps, which result from linkage 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 well used (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 substituting the average time that the advantageous allele spends in each frequency class on its way to fixation for the actual deterministic predictions. This is equivalent to using the function called $t^*(x;p)$ above, which nicely captures the fact that A_2 moves quickly through the small frequencies. Although Eriksson *et al*. (2008) did not justify their approach mathematically, this might turn out to be a useful approximate method for simulations (Kim and Wiehe, 2009), and perhaps it could be justified using the results of Nagylaki (1974). Eriksson *et al*. (2008) used a Moran population model, but were apparently unaware that some of their results were previously known (Ewens, 1963) and of the connection to $t^*(x;p)$ from the diffusion model.

The Yule Process Approximation

Heterozygosity and probabilities of identity by descent were major concerns of classical population genetics. Because of this, the idea of following a *pair* of alleles back to their common ancestor, due to Malécot, is not a novel concept of the coalescent. The power of coalescent approach comes from its ability to handle samples larger than two. The object of many of the computational methods of inference for neutral coalescent models discussed above is to compute the likelihood of a sample, in all its intricate detail. In principle, this likelihood captures all of the information in the data. Similarly, the goal of simulation-based methods like the one of Coop and Griffiths (2004) mentioned above is to apply the power of the (structured) coalescent approach to inferences about selection. However, methods based on the structured coalescent model, in which it is necessary to account for the unknown allele frequency in the population as it changes through time, are computationally costly and mathematically difficult. 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. In this context, 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. For a single sweep and for recurrent sweeps, respectively, Kim and Stephan (2002) and Kim (2006) derived the expected allele-frequency spectrum at a single neutral polymorhic in a sample of size n. Kim and Stephan (2002) and Kim (2006) used a forward-time analysis, but in a coalescent model, the expected allele-frequency spectrum would depend explicitly on the distribution of family sizes. The Yule process approximation provides a way to generate the numbers and sizes of families that descend from the sweep as well as the times of events in the ancestry of the sample.

Durrett and Schweinsberg (2004, 2005) introduced this approximation in an analysis of selective sweeps in a Moran population model. Etheridge *et al*. (2006) approached the same problem starting with the standard diffusion, which shows that the

Yule process approximation applies to a variety of models, in the limit as $N \to \infty$. In a presentation that is more accessible to biologists, Pfaffelhuber *et al*. (2006) described 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 structured coalescent simulations are quite accurate for many purposes, especially for small samples, but Pfaffelhuber *et al*. (2006) showed that the Yule process approximation is better than the deterministic model for predicting the distribution of family sizes in larger samples.

The Yule approximation is derived under the assumption that σ is large, for the same model we considered above, *i.e*. a neutral locus *B* sitting near the selected locus *A*. As noted above, when σ is very large, the period in the first half of the sweep, when allele A_2 is increasing in frequency rapidly, becomes particularly important. The Yule process approximation is obtained by transforming the diffusion time scale during the sweep by the frequency $1 - x(t)$ of allele A_1, with the effect that the second half of the sweep becomes greatly compressed; this is depicted clearly in Figure 1 of Pfaffelhuber *et al*. (2006). On the 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 original sample can be modeled as a random subsample of a larger random tree, called the Yule tree. We can imagine (descending from the first A_2 mutant) the entire population gene genealogy of all 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 locus-*B* 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.

The process that generates the Yule tree is simple, although 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 the process by which the sample lineages percolate up from the $\lfloor \sigma \rfloor$ tips of the Yule tree towards the root, possibly coalescing each time two lines in the Yule tree join in a common ancestor. Because of this, simulations of the Yule approximation become slower when σ is larger

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

A lot of mathematical details go into demonstrating the validity of the Yule process approximation as well as 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 there will be at most two non-singleton families: one that escapes (possibly along with some number of singleton families) and one that descends from the original A_2 allele.

The fact that the families that escape tend to be singletons has led to approximations 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). While these approximations are accurate for some purposes but not others, they do serve illustrate an important feature of any strong selective sweep. Heuristically, we can argue as follows. On the coalescent time scale, the first half of the sweep is all that is relevant, and this has length roughly $2\log(\sigma)/\sigma$ when is very large. 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. We can now see that, 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. For example, for given values of *m*, *r*, and *s*, increasing *N* will *increase* the probability of escape. It may seem counterintuitive that the genomic extent of swept regions in our "*Drosophila*" model above will tend to be smaller than in our "human" model, but the reason is that by making *N* ten times larger, we also made ρ tens times larger, but we did not make the length of sweeps ten times shorter because of the $\log(\sigma)$ term in the numerator of t_{fix}.

The Ancestral Selection Graph

The structured coalescent approach to a sweep and the Yule process approximation are clearly related, in that both begin by imagining a single adaptive substitution that has occurred at a locus in a population. The model we will encounter in this section is quite different, in that it begins by imagining a population at equilibrium, or stationarity, with respect to the processes of mutation, selection, and random genetic drift at a locus. It shares something with the two previous approaches, because it is also based on results from diffusion theory but the models are so different that this is just a technical point. This third model, called the Ancestral Selection Graph, or ASG for short, has so far not

been used to study selective sweeps, but we will include it in our survey of coalescent models of selection because it presents an ingenious solution to the problem of non-exchangeability. However, the large number of technical details and definitions required to understand the ASG would be an unjustifiable distraction from our focus on selective sweeps, so we will have to be content with a relatively brief verbal description.

A more thorough description of the ASG can be found in Wakeley (2008) and in the review by Baake *et al*. (2008). A more mathematical description can be found in Stephens and Donnelly (2003) and a significant generalization of the ideas behind the ASG is available in Donnelly and Kurtz (1999).

The ASG was introduced in a pair of papers by Krone and Neuhauser (1997) and Neuhauser and Krone (1997). Consider our locus A, at which positive selection favors one allele, but now allow mutation in both directions between the two alleles. In this case, if the population has been evolving for a some time, with the parameters constant over time, it will approach a well-known stationary distribution for the frequencies of the two alleles (Wright, 1931, 1949). A defining feature of the ASG is that it applies to a random sample of size n taken from this population, at stationarity, without conditioning on any particular event and without even knowing the allelic types of the sample. The aim of the ASG is to model the joint distribution of the gene genealogy and the allelic types of the sample, so these can be simulated or studied analytically.

In order to do this, it is necessary to know the stationary distribution of allele frequencies. Given this, a wonderful trick by Krone and Neuhauser does away, at least temporarily, with the unpleasant problem of non-exchangeability of lineages in the coalescent with selection. We can imagine our population, of constant size $2N$ alleles. running from forever in the past, down to the present time. A sample taken now would be a sample from the population at stationarity, but so would a sample taken $2N$ generations ago, or even $100\times2N$ generations ago. We can see all the A_1 and A_2 alleles as they have been passed down, as well as all the births and deaths in the population.

The trick works as follows. Toss the first population aside for the moment, and in its place make a new population, running from forever in the past, etc., but this time get rid of the allelic types. Be generous, and let every allele in the population have the advantage of allele A_2. Relative to our first population, this new one has some extra birth events, and since the population size is constant, there must also be an equal number of extra deaths. Therefore, on average, lives are shorter now than they were in the population that contained both A_1 and A_2. Here, "births," "deaths," and "lives" belong to alleles, not to diploid individuals. Now, mark a fraction s of the birth events in this new population with the label "2" for later reference.

Note a few things about this new population. First, and quite importantly, alleles in the new population, with its additions, are exchangeable since they all have the same distribution of the number of offspring. Second, if we removed all of the marked events from the population, the rate of birth would be that of allele A_1 and life-spans would increase accordingly. Third, if we went very far back in time and assigned types to the $2N$ alleles, then we could reproduce the behavior of the first population we constructed. All we would need to do is forbid A_1 from reproducing if it happens to hit a birth events that is marked "2." This would cause the correct fitness difference between A_1 and A_2, so that if the population were followed forward in time, it would reach the same stationary distribution as the original population.

Krone and Neuhauser showed that these properties of the marked population allow gene genealogies to be constructed together with the allelic types of the sample. We start with a sample of size n at the present time, without specifying the allelic types. We then trace the genetic lineages back in time. A complication arises from the fact that we do not know yet if the events marked "2" really happened. When a lineage hits one of these events, we temporarily follow *both* the allele that might have died and the newborn allele that might have replaced it. This is a nuisance, since we have lost the relatively compact tree neutral coalescent theory, but is straightforward to model because the lineage are exchangeable. We must follow all of these lineages, as they split and coalesce, until there is just one left. The graph obtained in this way contains the gene genealogy of the sample, plus a lot more branches we'll need to discard. As in the neutral coalescent, any polymorphism in the sample results from mutation along the branches of this graph. So, we pick the allelic type of that one ultimate ancestral allele from the stationary distribution, then follow its descendants, allowing mutations, back down the tree. We keep events marked "2" only if the parent of the newborn allele has allelic type A_2. With some further rules described by Krone and Neuhauser, we can identify and discard all the non-ancestral branches in the graph and be left with the gene genealogy of the sample, with allelic types at the tips.

Figure 4C shows a hypothetical realization of this process, in which the ultimate ancestor was not reached within the range time in Figure 4C. Following 4A and 4B, the parts of the ancestral graph that did not end up in the gene genealogy of the sample are shown in pink while the gene genealogy is drawn in black. The vertical axis in Figure 4C is identical to the one in Figure 1, reflecting the fact the gene genealogies generated under the ASG tend not to differ very dramatically from neutral gene genealogies; for example, see the simulation study of Przeworski *et al.* (1999).

It is possible to use the ASG formulation to generate gene genealogies conditional on knowing the types of the sample (Slade, 2000a,b), and Slade (2001) has further described how to model a linked neutral locus within the conditional ASG. However, this idea has not been applied to hitch-hiking due to a single selective sweep. It would be difficult or impossible to do so because the stationary distribution of allele frequencies required in the ASG does not exist under a single (non-equilibrium) sweep.

The Coalescent with Multiple Mergers

The extension of the recurrent hitch-hiking model referred to above bring us to our last coalescent model of selection. Although the basic idea behind what follows can be found in many papers in the literature on hitch-hiking, the extreme version which justifies this separate section is due to Gillespie (2000, 2001). The notion, which is heretical under the neutral theory of molecular evolution, is that essentially all of what we attribute to random genetic drift might instead be due to strong recurrent sweeps. Gillespie used the term 'genetic draft' to emphasize the difference. Nielsen (2005), Hahn (2008), and Sella *et al*. (2009) promote this idea as a possible, or even likely, explanation for genomic patterns of variation.

Recall the preponderance of sweeps in our "*Drosophila*" simulations. Figure 2E shows, for example, shows five sweeps in $2N$ "Drosophila" generations, which again is the expected number for the parameter values we used. When sweeps occur so

frequently, polymorphism levels will be dramatically reduced relative to neutral predictions. If the rate of sweeps per $2N$ generations is very high, then might make sense to dispense with genetic drift altogether and replace it with genetic draft. Gillespie (2000, 2001) illustrates this using heterozygosity, *i.e*. for $n = 2$, but this case does not capture the full dynamics of the resulting coalescent models. These are called coalescents with multiple mergers, or multiple collision, and they differ very dramatically from the standard neutral coalescent, because several genetic lineages may come together at exactly the same time.

Most of the work on these models has been in fact been done under the assumption of selective neutrality (Pitman, 1999; Sagitov, 1999; Schweinsberg, 2000; Möhle and Sagitov, 2001; Birkner *et al*., 2005). Coalescents with multiple mergers arise in neutral models when the variance in the number of offspring among individuals in the population is very large. Briefly, Kingman (1982a,b) assumed that this variance is finite, while coalescents with multiple mergers occur when this variance approaches infinity in the diffusion limit. Because of this, the time scale for coalescents with multiple mergers is not $2N$ generations, but rather will depend on the demography of the population, through the variance in offspring numbers. These models offer a wide variety of possible behaviors, and consequently require us to follow and interpret a great many mathematical details in order to assess their biological plausibility. An introductory look at the possibilities under one particularly simple model of a population may be found in Eldon and Wakeley (2006).

We will consider these models heuristically, in the context of genetic hitch-hiking under strong selective sweeps. Durrett and Schweinsberg (2005) proved that the gene genealogy at a neutral locus which is imbedded in a genomic region where sweeps occur at some rate and at random locations will follow a coalescent with simultaneous multiple mergers. The additional word 'simultaneous' distinguishes models in which several multiple coalescent events may occur at the same time from model in with multiple mergers occur at different times. We can understand this result with reference to our analyses of selective sweep. If sweeps hit the population frequently, then the time scale of the ancestral process will depend on their rate of introduction and not necessarily on the population size. In addition, if selection is very strong (large σ) the duration of sweeps may be much shorter than average time between them. The result of Durrett and Schweinsberg (2005) requires that sweeps do not overlap. The focal, neutral locus will be affected by sweeps in the region around it, subject to conditions about the relative strengths of recombination and selection like the ones discussed above under the Yule process approximation.

Under this model, genetic lineages at the neutral locus may travel backward in time, undergoing coalescent events whenever a selective sweep happens at some other locus in the vicinity of the neutral locus. Figure 4D shows a hypothetical gene genealogy under this scenario. Four lineages coalesce in the first sweep and two escape by recombination, then the remaining three lineages coalesce during the next sweep. For comparison with the other panels in Figure 4 and with Figure 1, time in Figure 4D is given in the standard coalescent units of $2N$ generations. The range of times in Figure 4D is much smaller than in the other panels to emphasize how sweeps may drastically shorten the time scale of the ancestral process. Recurrent sweeps will similarly affect the process of genetic drift/draft forward in time in the population at the neutral locus.

Notice that there are really three time scales here: the usual coalescent time scale, the time scale at which sweeps occur, and the time scale (duration) of a single sweep. The usual coalescent time scale is always lurking in the background, and would become the primary driver of coalescence if the rate of sweeps was decreased somehow. It is possible also that multiple-merger events and regular coalescent events occur on the same time scale. If so, then the ancestral process would not be the standard coalescent, but it would occur on the same time scale as the standard coalescent. Further, even if the time scale of a coalescent with multiple mergers is much shorter than the time scale of the standard coalescent, it is possible for the distribution of gene genealogies (in terms of tree structure and relative coalescence times) to be identical to the standard neutral coalescent. To visualize this last point, we can imagine a neutral locus sitting in a genomic region that is hit by sweeps extremely frequently but where the recombination rate is high enough that at most two lineages get caught by each sweep.

Therefore, it may be difficult to distinguish between multiple-mergers coalescent processes and the standard neutral coalescent, at least based on patterns of variation at a single neutral locus. For examples and a discussion of this in the context of neutral population models, see Wakeley and Sargsyan (2009). For multi-locus data within a single species, it may be possible to make inferences based on the variation and correlations between loci, as in Wiehe and Stephan (1993), among others, have done. In addition we might appeal to comparisons among species to make inferences more broadly. For example, genetic draft can explain the problem, which has attracted a lot of debate over the years, that polymorphism levels do not depend linearly on population size, as predicted by the neutral theory (Lewontin, 1974; Gillespie, 1991; Meiklejohn *et al.*, 2007). These coalescent models with multiple mergers could prove useful in long-standing debates about the neutral theory of molecular evolution.

HOPES FOR THE FUTURE

The world we live in today, in many ways, could scarcely have been imaginable to Darwin, and yet Darwin's ideas are as relevant now as they ever have been. Clearly, in the intervening 150 years since the publication of the *Origin of Species*, a lot has happened to make biology what it is today. In this chapter, we have focused on developments in mathematical population genetics, a field that one might rightly say has at times developed inordinately compared to applications of the theory. Today, on the other hand, while the mathematical tools of population genetics and evolutionary biology are impressive, they pale in comparison to the genetic data, which is accumulating at a truly mind-boggling rate. It gives us hope to see the current rapid pace of research at the interface of theoretical and empirical population genetics.

Although levels and patterns of genetic variation continue to surprise us today, as they first did in the 1960's, our modern view is much more focused on natural selection as an important force in shaping variation than it was in the years after Kimura (1983) gave his summary of the field. 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 (Gillespie, 1991), the longevity of the neutral theory may be surprising. Crow (2008) points out that the very simplicity of the neutral theory accounts for a lot of its appeal. At once, in a

modern classic of science, Kimura and Ohta (1971) explained both molecular evolution among species and molecular variation within species as different facets of one relatively simple process.

In focusing on data sampled from populations, the retrospective approach of standard coalescent theory was built on the assumption of strictly neutrality. This is certainly justified from the standpoint of statistical theory, assuming we are interested making inferences about natural selection, because strict neutrality is the logical null model. Due to the fundamental directionality of statistical tests, however, our belief in the null model should not increase when we fail to reject it. Gillespie (1994) has shown that it there is low power to detect fundamental deviations from neutrality and to distinguish among some selective alternatives to the neutral theory, at least using simple statistical tests. Choosing a new "null model" among alternatives to neutrality might prove difficult. Although it is clearly short-sighted to be discouraged by difficult mathematics, it is also no small task to "slog through 100 pages," as Gillespie's puts it, of Chapter 4 in *The Causes of Molecular Evolution* (Gillespie, 1991). Again, we can draw some hope from the flurry of current activity in empirical and theoretical population genetics, and the application of the models presented in this chapter, as well as others, to genomic data from many species.

In closing, we can also be hopeful that the sophisticated coalescent approaches for making inferences about the demographic history of populations—for example changes in population size over time and patterns of population structure by subdivision and migration or by the isolation of populations without migration—can be further 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.

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

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.

Donnelly P., and T. G. Kurtz. 1999. Genealogical models for FlemingViot models with selection and recombination. *Ann. Appl. Probab*. 9: 1091-1148.

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. 1974. A note on the sampling theory for infinite alleles and infinite sites models. *Theor. Popul. Biol*. 6: 143-148.

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.

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

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.

Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science*, 18: 49-50.

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

Hein, J., Schierup, M. H., and C. Wiuf 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

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.

Kesten, H. 1980. The number of distinguishable alleles according to the Ohta–Kimura model of neutral mutation. *J. Math. Biol*. 10: 167-187.

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. 1976. Coherent random walks arising in some genetical models. *J. Roy. Stat. Soc. B* 351: 19-31.

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.

Kingman, J. F. C. 1982c. Exchangeability and the evolution of large populations. In G. Koch and F. Spizzichino (eds.), *Exchangeability in Probability and Statistics*, pp. 97-112. North-Holland, Amsterdam.

Kingman, J. F. C. 2000. Origins of the coalescent: 1974–1982, *Genetics* 156: 1461-1463.

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.

Mendel, J. G. 1866. *Versuche über Plflanzenhybriden* Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, 1865. Abhandlungen:3–47.

Translated in: Druery, C. T., and Bateson, W. 1901. *Experiments in plant hybridization*. Journal of the Royal Horticultural Society 26: 1–32.

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. 1974. The moments of stochastic integrals and the distribution of sojourn times. *Proc. Nat. Acad. Sci. USA* 71: 746-749.

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

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

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.

Przeworski, M., B. Charlesworth, and J. D. Wall. 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol*. 16: 264-1252.

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.

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.

Slade, P. F. 2000a. Simulation of selected genealogies. *Theoret. Popul. Biol*. 57:35–49.

Slade, P. F. 2000b. Most recent common ancestor distributions in genealogies under selection. *Theoret. Popul. Biol*. 58: 291-305.

Slade, P. F. 2001. Simulation of 'hitch-hiking' genealogies. *J. Math. Biol*. 42: 41-70.

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.

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

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. 2008. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, Colorado.

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

Weinberg, W. 1908. Über Vererbungsgesetze beim Menschen. *Jaresh. Verein. f. vaterl. Naturk. Würtem*. 64: 368—382. Translations: Pages 4–15 in *Papers on Human Genetics* (S. H.. Boyer, ed.), Prentice Hall, Englewood Cliffs, NJ (1963) and pages 115–125 in *Evolutionary Genetics* (D. L. Jameson, ed.), Dowden, Hutchinson, and Ross, Stroudsburg, PA (1977).

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.

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

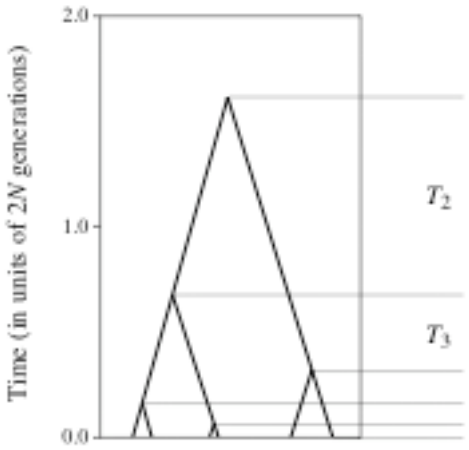
Zuckerkandl, E. and L. Pauling 1965. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*. Academic Press, New York.

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.

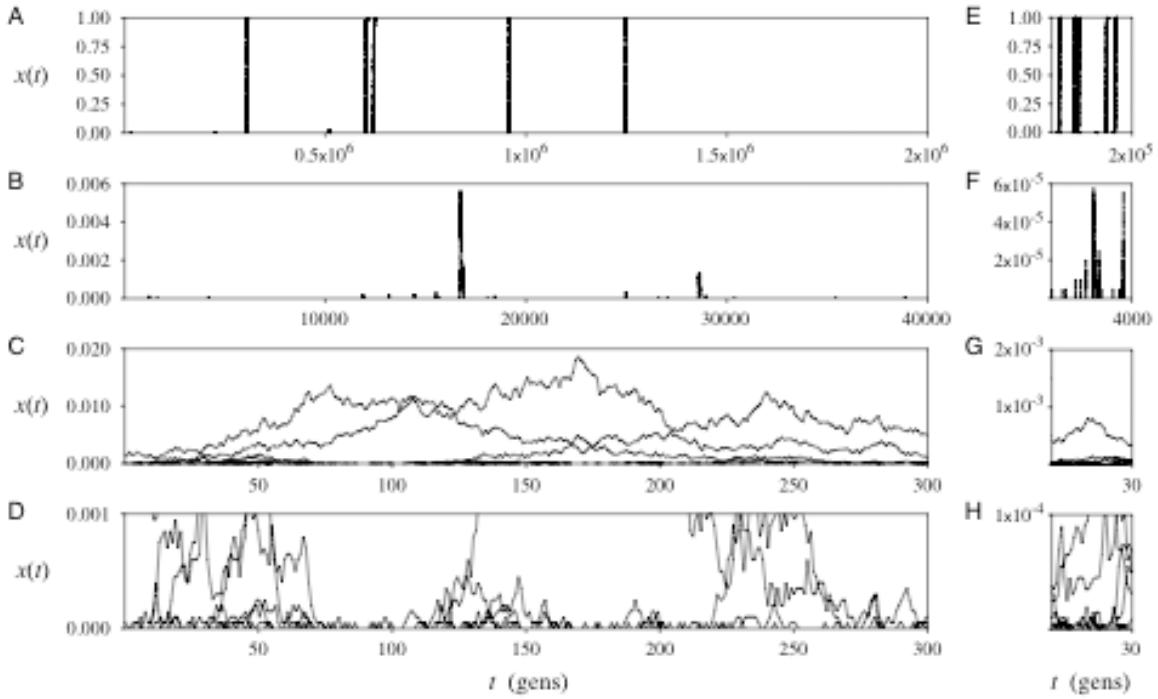
FIGURE 2—Simulated allele-frequency trajectories for the evolution of the hypothetical "human" and "*Drosophila*" loci described in the text. Panels A through D show results for "humans" and panels E through H show results for "*Drosophila*." A & E show Advantageous mutations that reached high frequencies. B & F show advantageous mutations that went extinct. C & G show deleterious mutations that reached appreciable frequencies then went extinct. D & H show deleterious mutations that went extinct without reaching appreciable frequencies. 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 2E.

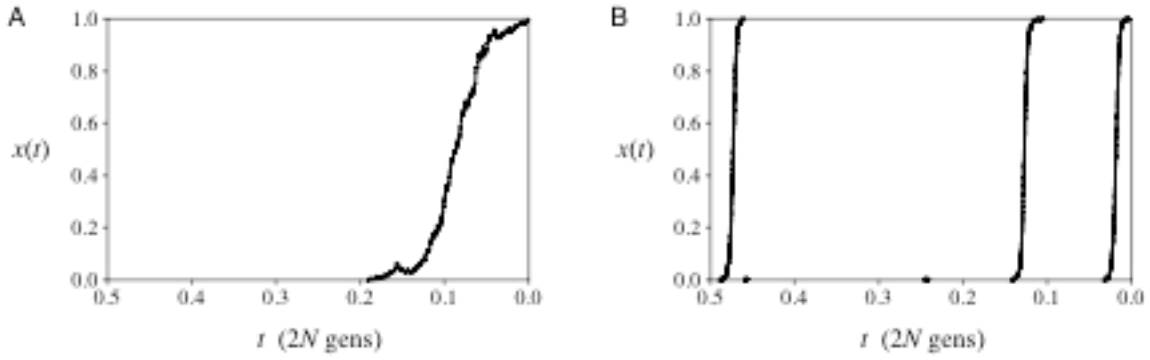
FIGURE 4—Hypothetical gene genelogies for a sample of size $n = 6$ at a neural locus linked to a selected locus, showing the four coalescent approaches to modeling natural selection described in the text. Panel A depicts the structured coalescent approach, B depicts the Yule process approximation, C depicts the Ancestral Selection Graph, and D 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 in A, B, and D mark recombination events that allow linked neutral lineages to "escape" a sweep.



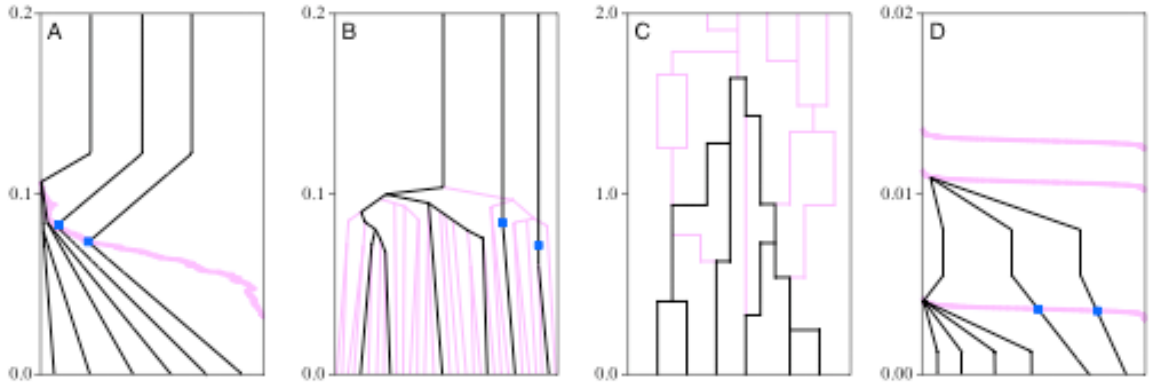
(figure 1)


A
B
C
D
E
F
G
H
x(t)
1.00
0.75
0.50
0.25
0.00
0.5x10^6
1x10^6
1.5x10^6
2x10^6
0.006
0.004
0.002
0.000
10000
20000
30000
40000
0.020
0.010
50
100
150
200
250
300
0.001
2x10^5
6x10^-5
4x10^-5
2x10^-5
4000
2x10^-3
1x10^-3
30
1x10^-4
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)