DNA haplotypes in the Tiger Salamander, *Ambystoma tigrinum*. *Genetics* 140: 767–782.
Templeton, A. R., Sing, C. F., Kessling, A. and Humphries, S. (1988) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* 120: 1145–1154.
Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., Neel, J. V., Larsen, M., Smith, D. G., Vullo, C. M. and Wallace, D. C. (1993a) Asian affinities and continental radiation of the four founding native American mtDNAs. *Amer. J. Hum. Genet.* 53: 563–590.
Torroni, A., Sukernik, R. I., Schurr, T. G., Starikovskaya, Y. B., Cabell, M. F., Crawford, M. H., Comuzzie, A. G. and Wallace, D. C. (1993b) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Amer. J. Hum. Genet.* 53: 591–608.
Whittemore, A. T. (1993) Species concepts: a reply to Ernst Mayr. *Taxon* 42: 573–583.
Wiley, E. O. (1981) *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. Wiley, New York.

Testing speciation models with DNA sequence data

John Wakeley[1] *and Jody Hey*

Department of Ecology, Evolution, and Natural Resources, Nelson Biological Labs, Rutgers University, P.O. Box 1059, Piscataway, NJ 08855, USA
[1]Present address:, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

Summary

This chapter reviews an approach to the study of speciation that is based on patterns of genetic variation within and between closely related species. Historically, research on the genetic mechanisms of speciation, and of species divergence, is very difficult – suffering from both practical difficulties in data collection and from theoretical problems. The method outlined in this paper is based on genealogical models of population divergence. We describe a hierarchy of models, and show how these fit into a hypothesis-testing framework that overcomes some of the theoretical problems of studying speciation. The method also advances the empirical study of speciation. Since testing of the models relies only on comparative DNA sequence data from closely related species, it can be applied to existing species regardless of whether it is practical or possible to generate hybrids.

Introduction

Research on the mechanisms of speciation is difficult, and speciation studies have traditionally suffered from two related theoretical uncertainties. One is the "species problem", which is the long-standing debate over the meaning of the word *species* and about the best means of identifying them. Disagreement over the nature of species has contributed to the second problem, which is the present lack of a hypothesis-testing framework for studying speciation.

We do not propose to solve the first of these two problems; rather, we sidestep it in two ways. First, we focus on mathematical models of the genetic divergence of populations. These are simple extensions of well-known population genetic models, and can be applied equally to populations and to species. For example, two separate populations (or species) that were once one will diverge over time unless there is gene flow between them. Under virtually any model that includes mutation, the two will accumulate differences. So long as our focus is on the level of genetic vari-

R. DeSalle, B. Schierwater (eds) Molecular Approaches to Ecology and Evolution

ation within and between populations, we can model divergence without regard to the delineation of the populations into named species. Second, in discussing speciation, we focus mainly on sexually reproducing organisms. Thus, we are able to concentrate on a point of relative agreement among workers in this field. Reproductive isolation, or the inability to interbreed with outsiders, is for most the hallmark of sexual species. In this context, the development of reproductive isolation is synonymous with speciation. The mathematical models that we explore also fall into a hierarchy of complexity, and this hierarchy helps to overcome the second of the two traditional theoretical difficulties in studying speciation (i.e. the lack of a hypothesis-testing framework).

The study of speciation is also difficult from an empirical standpoint, and most studies face one of two common obstacles. The first difficulty occurs for studies of organisms from closely related populations that may be incipient species. These studies face the uncertainty that such present-day examples of population differentiation may not be representative of speciation events in general. A second difficulty can arise when genetic approaches are applied to clearly distinct species. Some methods require crosses and hybrid formation, and these cannot usually be done on clearly distinct species that have separated long ago, for the simple reason that crosses often do not yield fertile progeny. An alternative kind of data, including comparative DNA sequences from multiple loci, can overcome these difficulties. If data are collected from within each of two closely related (but clearly distinct) species, for multiple loci, then the patterns of intraspecific and interspecific polymorphism can be interpreted in light of models of speciation (Hey and Kliman, 1993; Hey, 1994; Hilton et al., 1994).

In this chapter, we describe a conceptual framework for mathematical models of speciation. We begin with a simple model of variation within a single population, then we go on to describe several divergence models and show how they vary in their predictions about patterns of DNA sequence variation within and among populations. Some of these models are simpler than others, and our basic premise is that the hierarchy of model complexity permits a statistical approach to compare models of divergence. The statistical framework that we develop starts with the simplest possible extensions of single-population models to the case of two species. When a simple model can be rejected, another, more complicated one can be proposed and tested. In this way, data can point to successively more precise descriptions of divergence. If two models are both consistent with some data, then we are inclined to work with the simpler of the two until contradictory data appear.

All of the models are genealogical, and they hold a close correspondence to comparative DNA sequence data sets collected from within and between closely related species. The models help to inform our intuition about speciation, and their correspondence to the data permits the actual testing of speciation models.

The standard neutral model

We begin by briefly reviewing a model that makes predictions about levels and patterns of DNA sequence variation within a single population. This model is widely used and forms the basis of most of the statistical tests developed to detect historical patterns of natural selection in DNA sequence data (Hudson et al., 1987; Tajima, 1989b; Fu and Li, 1993). We then show how this model can be extended to study divergence between populations and how these models of divergence are related to theories of speciation.

A single population

The standard, single-population, neutral model is a combination of the well-known Wright-Fisher model (Fisher, 1930; Wright, 1931; Ewens, 1979; Hudson, 1990) and the assumption that mutations have negligible consequences on fitness (Kimura, 1983. In brief, mating is random, the population has had a constant size for a very long period of time, and mutations are neutral and occur in such a way that individual base positions are segregating at most only a single mutation at any point in time.

Under this model, variation is lost through genetic drift at a rate that is proportional to the inverse of the population size (i.e. $1/N$), and variation is input to the population at a rate that is proportional to the mutation rate, u. In practice, the compound parameter $\theta = 4Nu$ has proven useful for describing the amount of genetic variation that is expected under the model. Typically, estimating θ for a population or species is the first step in a population genetic study.

Assume that we have taken a sample of homologous DNA sequences from a diploid population which conforms to this standard, neutral model. If the mutation model is at least approximately true, then most base positions will not be variable in the sample, and it is not difficult to align the DNA sequences. When the sequences are aligned in rows, for example, some base positions may be revealed as polymorphic, and these appear simply as columns in which not all of the base values are identical. If we take a sample of n DNA sequences, then under our simple model the expected number of these polymorphic, or segregating, sites is

$$E(S) = \theta \sum_{i=1}^{n-1} \frac{1}{i} . \qquad (1)$$

Suppose that we have a sample of two sequences ($n = 2$). Then it is easy to see from (1) that the expected number of polymorphic sites is equal to θ. In fact, the average number of polymorphic sites among pairs of sequences in a larger sample, com-

monly called pairwise differences, has this same expectation. If we add a third sequence, then expression (1) means that we expect to add half again as many polymorphic sites as were observed with just two sequences. As more sequences are added, the expected number of polymorphic sites in the entire sample rises but at a slower and slower rate.

If we count up all of the polymorphic sites S, we can use (1) to estimate θ:

$$\hat{\theta} = \frac{S}{a_n}, \tag{2}$$

$$\text{where} \quad a_n = \sum_{i=1}^{n-1} \frac{1}{i} .$$

Expression (2) is commonly referred to as Watterson's estimator of θ. Watterson (1975) also derived the variance of S that will arise under the combination of Fisher-Wright and neutral mutation models. An understanding of the variance can be especially useful when data have been collected from multiple loci from the same population. Even though the data are from the same organisms, if they are from independently segregating loci, then the variance among them for S is expected to follow Watterson's expressions. We will return to this idea of using variation among loci when we discuss ways to evaluate models of population divergence.

Expression (1) is one way to connect the model parameter θ to an observable quantity, in this case S. However, it does not make use of all of the information regarding polymorphisms that is available from a set of aligned DNA sequences. In particular, each polymorphic site has an associated frequency, because each site divides the sample into two groups. There are, for instance, sites at which a single sequence is different from all the others, and sites where two sequences bear one nucleotide and the other $n-2$ bear another, and so on. Under our simple model, the expected numbers of polymorphic sites in each frequency class are known. Thus, if we could distinguish the ancestral from the mutant base at a particular site, and if we let ξ_i represents the number of sites at which the mutant nucleotide has frequency i/n in the sample, then

$$E(\xi_i) = \frac{\theta}{i} \tag{3}$$

(Tajima, 1989b; Fu and Li, 1993;Fu, 1995). Without an outgroup sequence, it is impossible to distinguish mutants in frequency i/n from those in frequency $(n-i)/n$. Thus, we are limited to measuring $\eta_i = (\xi_i + \xi_{n-i})/\delta$. In this expression δ is just an adjustment for the special case when n is even and $i = n/2$. Thus δ is equal to 1 if $i \neq n-i$ and equal to 2 if $i = n-i$. Then η_i has expectation

$$E(\eta_i) = \left(\frac{\theta}{i} + \frac{\theta}{n-i}\right) / \delta \tag{4}$$

(Fu, 1995) also gives the variances and covariances of site frequencies.

The genealogies of nucleotide sites

In the absence of recombination, a sample of homologous DNA sequences will have a single gene tree history, or genealogy. The single-population model described above predicts a particular probability distribution of genealogies. A typical one of these is pictured in Figure 1. Under the assumption typically made that the number of sequences sampled is much smaller than the total number of individuals in the population ($n \ll N$), the genealogy of a sample is a random bifurcating tree and is expected to have characteristic branch lengths; namely, the times between successive common ancestor events (the nodes in the genealogy) are expected to be shorter when there are many ancestral lineages and longer when there are fewer. Thus, for the genealogy in Figure 1, we expect that $t_6 < t_5 < t_4 < t_3 < t_2$. More specifically, under the Wright-Fisher model, t_i is approximately exponentially distributed with parameter $i(i-1)/(4N)$, so that the expectation of t_i is equal to $4N/[i(i-1)]$.

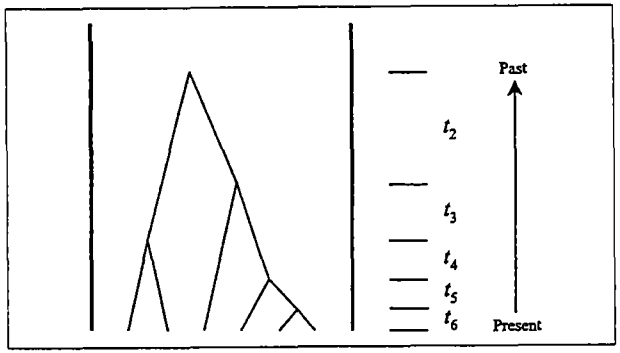

Figure 1. One possible genealogy of a sample of six sequences from a Wright-Fisher population. The thick lines represent the population's boundaries, emphasizing its finite size. The thin lines trace the ancestral lineages of the sample back (up) in time. The times, t_i, are the periods during which there were i ancestral lineages of the sequences in the sample.

The quantities S and η_i (which we can calculate from data) contain information about the genealogy of a sample, because the mutations that cause the variation observable in a sample must occur on the ancestral lineages like those depicted in Figure 1. For example, η_i is the number of mutations which occurred on lineages that left i or $n-i$ descendants in the sample. The total number of polymorphic sites, S, is simply the total number of mutations that occurred on the entire genealogy. Deviations from the standard model can be understood easily if we keep these relations between observable quantities and underlying genealogies in mind. For example, η_1 will increase as t_6 increases in Figure 1 because the greater t_6 is, the longer are the lineages leading to single descendent sequences. Thus, a larger than expected value of η_1 might indicate the failure of the standard model and suggest a different history that would cause this kind of genealogy.

In general, we cannot directly observe the genealogy of a sample, but even if we could, the particular one we saw would be just a single point in the universe of all possible genealogies. Looking at Figure 1, it is easy to see that this universe is incredibly large. Not only are there a very great number of different possible patterns of branching, but for each one of these there are an infinite number of possible t_i.

Multiple loci and recombination

When a number of unlinked loci are sampled, but within each of them no recombination occurs, then each one represents a single, independent draw from the sample space of genealogies. Thus, equations (1) and (4) would apply to each locus separately, but they would also apply to the total S and η_i for all the loci combined. The variances of S and η_i among the loci should follow the expressions given by Watterson (1975) and Fu (1995). Suppose that the data from several independent small loci are pooled, and S and η_i calculated for all the loci together, as if they constituted one large composite locus. In this case the variances of the total S and η_i would not follow those same expressions. Rather the variances for the composite locus would be considerably reduced. The reason is that the composite locus would be made up of several segments, each providing information independent from the others, and would not be just a single realization of the genealogical process.

Recombination is another component which can be included in the model. Like mutation, it is important in shaping the pattern of genetic variation, but in a very different way. The effect of recombination within a locus is similar to that of sampling multiple loci. The expected values of S and of η_i in expressions (1) and (4) are unaffected by recombination, but the variances are smaller. When recombination occurs in the history of some sequences, its effect is to break the sequence up into pieces that segregate more or less independently of each other. Thus, different segments within a locus that undergoes recombination may have different genealogies, just as different loci may. This means that observed single-locus values of S and η_i will tend

to be closer to their expectations when the recombination rate is high than when it is low. Sampling multiple loci and the occurrence of recombination within loci both act to increase the number of independent observations, thus decreasing the variance. In general, lower variances permit more accurate estimates of parameters like θ, and will also give us more power to test the other assumptions of the model.

It is important to note that, because of recombination and independent segregation, some loci or even small regions within a single locus may be subject to selection and thus differ markedly from the standard neutral model, while most loci or most sites within a single locus conform well to the model. Sites that are under selection will influence the histories of adjacent regions, but the magnitude of this effect will decrease with the recombinational distance (Hudson and Kaplan, 1988). Genetic "hitchhiking", the phenomenon that neutral or deleterious mutations can be swept to fixation if they are tightly linked to a positively selected variant (Hudson et al., 1987), is one well-known example of this. Thus, we need to draw a distinction between processes which act on single loci or small regions of the genome and ones which affect all loci identically.

For instance, if we use Tajima's (1989b) or Fu and Li's (1993) tests on a single locus, and find that the data are not compatible with the standard neutral model, without sampling more loci, we cannot know whether the process we have uncovered is locus-specific or occurs at the population level. A rapid growth in population size and a selective sweep will have nearly identical effects on the character of genetic variation (Tajima, 1989a; Slatkin and Hudson, 1991), as will population subdivision and balancing selection (Kaplan et al., 1988; Simonsen et al., 1995). However, selection acts on specific, limited regions, whereas population growth and subdivision act on the entire genome. This represents an advantage rather than a problem. By sampling many different loci, we can potentially distinguish between these two kinds of processes.

Neutral two-population models of divergence

The simple model of a single population can be extended to include two (or more) populations. To do so, we must introduce new model parameters that describe the historical relationship between the two populations. We must also consider the data, which will have a component that is not apparent when just one population has been sampled, i.e. divergence between the two populations. When data come from just one population, all polymorphisms appear as variations within that population, but with two populations we must also consider DNA sequence variation that distinguishes the two populations.

In the following sections, we review some simple extensions of single-species models to models of population divergence. Three kinds of models are discussed: isolation without gene flow; limited migration; and mixtures of migration and isolation.

Isolation

In the isolation model, a single ancestral species splits into two descendants. The split is assumed to happen instantaneously at some time in the past. After that time, the two descendent populations are assumed to be completely isolated from each other – there is no genetic exchange between them at any time thereafter. All three populations, the ancestor and the two descendants, are Wright-Fisher populations (see section "A single population"). Figure 2a gives a graphic depiction of the isolation model.

The model has three Wright-Fisher populations. Each may have its own unique effective size, N, so that there may be three characteristic parameters, θ. We propose that this general version of the isolation model is a good starting point for the study of speciation. It is fairly simple, and it corresponds roughly to the case of population divergence under complete allopatry. Other isolation models impose restrictions on the relative sizes of the three populations: Takahata and Nei (1985) and others assumed that all three populations or species are of the same size, and Hudson et al. (1987) assumed that the ancestral population is equal in size to the average of the two descendant population sizes. The salient feature of the isolation model is the complete absence of genetic exchange. In designing statistical tests, the more general version of the isolation model is preferable; we would not want to reject the model simply because of differences in population size among species.

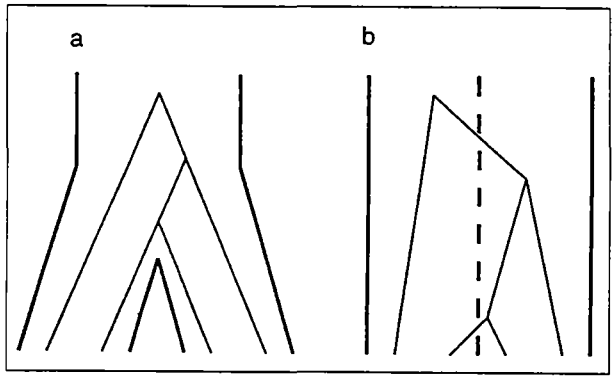

Figure 2. The strict isolation model, (a), and the equilibrium migration model, (b). As in Figure 1, the thick lines represent population boundaries, now with dashed lines to indicate that migration can occur, and thin lines trace the genealogy of each sample back (up) in time.

Wakeley and Hey (1997) considered this general isolation model with two descendant and one ancestral population size: N_1, N_2, and N_A. Four parameters then characterize the model, and these are θ_1, θ_2, θ_A, and T, where T is the time of separation measured in units of $2N_1$ generations. Correspondingly, four categories of segregating sites characterize variation within and between the two species. The first two of these comprise sites that are polymorphic in one of the species but monomorphic in the other. The numbers of these are called S_{x1} and S_{x2}, for the counts in species 1 and 2, respectively. Next are sites which show fixed differences, that is, which are monomorphic in both species but with different nucleotides. These are called S_f and were previously studied by Hey (1991). Last, there are sites at which the same polymorphism segregates in both species. The number of these shared polymorphisms is referred to as S_s. Wakeley and Hey (1997) derived the expectations of each of these four mutually exclusive categories of segregating sites in a sample and used them to estimate the four parameters of the isolation model. The expectations depend on all four parameters, θ_1, θ_2, θ_A and T, so the parameters of the model are estimated by solving numerically for values that most closely equate the expectations to observations. Extensive simulations showed that the numbers of exclusive, fixed and shared polymorphisms do contain the information necessary to estimate ancestral population parameters. Wakeley and Hey (1997) also derived the expectations of the joint site frequencies in a two-species sample.

Figure 3 shows two examples of plots of the four types of polymorphisms as a function of the time of isolation. The only difference in the parameters for Figures 3a and 3b is that in 3a the common ancestral population was equal to one of the descendants, whereas in 3b, the ancestor was much larger than either descendant. Note the large difference in the shapes of the curves, especially for times less than $T = 1$. Clearly the size of the common ancestor can have a very large impact on the levels of variation in each of the four classes.

Migration

Restricted, but nonzero genetic exchange between two populations is another possible cause of differentiation. Migration models stand in contrast to isolation models which permit no gene flow, but in practice the two give similar predictions about many aspects of genetic variation (Slatkin and Maddison, 1989; Takahata and Slatkin, 1990).

The most-studied migration model assumes that both populations are of the same size and have been exchanging migrants at a constant rate for an essentially infinite length of time into the past. This simple equilibrium migration model, pictured in Figure 2b, has just two parameters, the population migration rate, M, and a single θ. It is easily compared with the simple isolation model studied by Takahata

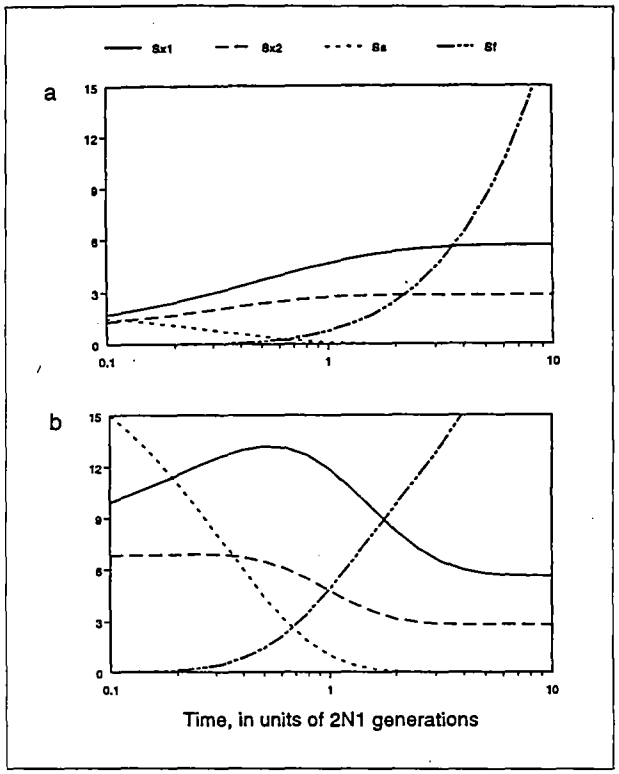

Figure 3. Expected values for S_{x1}, S_{x2}, S_s and S_f under the general isolation model as a function of T (time in units of $2N_1$ generations). For both figures, $n_1 = n_2 = 10$, $\theta_1 = 20$ and $\theta_2 = 1.0$ (a), $\theta_A = 1.0$ (b) $\theta_A = 10$.

and Nei (1985) and others, which is characterized by the time of separation, T, measured in units of $2N$ generations, and a single θ.

In order to compare these two models, we need a measure of divergence for which the expectation and the variance is known under both. At present, the only measure that is sufficiently well understood in both contexts is the average number of pairwise differences. For two populations, there are three measures of pairwise difference: the averages within each population, d_1 and d_2, and the average between populations, d_{12}. These are easy to calculate from sequence data. For instance, d_{12} is calculated simply by comparing each sequence from one population with each from the other, determining the number of differences between the two sequences, and taking the average. It is a well-known result that when $T = 1/(2\ M)$ these two models give identical predictions about the expected values of d_1, d_2 and d_{11} (Li, 1976; Gillespie and Langley, 1979). However, the two models make different predictions about the variances. Wakeley (1996a) derived expressions for the variances of pairwise differences in the two-population equilibrium migration model and compared them with those found under isolation by Takahata and Nei (1985). The results showed that, when the expectations of the average numbers of pairwise differences are the same in both models, the variances are larger under migration than under isolation. This can, again, be understood by considering the genealogy of a sample. Figure 2 compares the isolation model and the migration model, and shows the genealogy of a sample of two gene copies from each population. Looking first at the variance of between-population pairwise differences, it is easy to see that under isolation the occurrence of interpopulation common ancestors is restricted in time to the ancestral population. Under migration, however, there may be both very recent and very ancient interpopulation events. The genealogies shown in Figure 2 illustrate this. Because interpopulation common ancestor events occur over a broader range under migration than under isolation, the variance is larger. The variance of intrapopulation pairwise differences is inflated also, but does not depend on there being more than one interpopulation common ancestor.

Wakeley (1996b) used this result to devise a test of the simple isolation model. The test is formulated such that the isolation model is rejected, for given values of average pairwise difference, when the variances are too large. The usefulness of the variance of pairwise differences in testing the simple isolation model suggests a test of the more general isolation model studied by Wakeley and Hey (1997) that could also detect migration. Other things being equal, the variances of the numbers of fixed, shared and exclusive polymorphisms among loci will be higher in models that include migration than in ones that assume strict isolation. Thus, a test could be made that rejects the four-parameter isolation model in favor of some sort of limited migration. We expect that such a test will not be as sensitive to changes in population size as Wakeley's (1996b) test, because fewer restrictions on population sizes are imposed.

Mixed models

Consider a situation in which the isolation model has been fitted to a data set and the fit is so poor that the model is rejected. This would occur if we found variances among loci that were considerably larger than expected under the isolation model. The next step is to consider realistic alternatives that included limited migration. If we are studying species, the equilibrium migration model, described above, is not appropriate, because it predicts a standing level of differentiation between the two populations which does not increase over time. It is a model of equilibrium divergence but not speciation. Realistic models of speciation must include some element of isolation, such that gene flow is ultimately prevented. Of course, natural selection may be a part of this, and we address this below. Focusing for the moment on neu-

tral models of divergence, the alternatives to a strict isolation model are mixed models that include some migration and some isolation.

We consider two mixed isolation-migration models that appear to be reasonable alternatives and that may aid in our attempts to understand the divergence of species. These two models are pictured in Figure 4, and while no theoretical work has been done on either of them, we can predict some of their characteristics from what is known about strict isolation and equilibrium migration. The first, shown in Figure 4a is a hybrid of the isolation model and the model considered by Wakeley (1996c). Originally there was a single Wright-Fisher population, then a period of migration between two nascent species and, finally, complete isolation. The second model, shown in Figure 4b, is one in which the original Wright-Fisher population gradually splits into two which exchange migrants. The migration rate is initially a very high value (i.e. as if there were just one panmictic population) and then decays until isolation is complete. By adjusting the parameters of these two models, we can

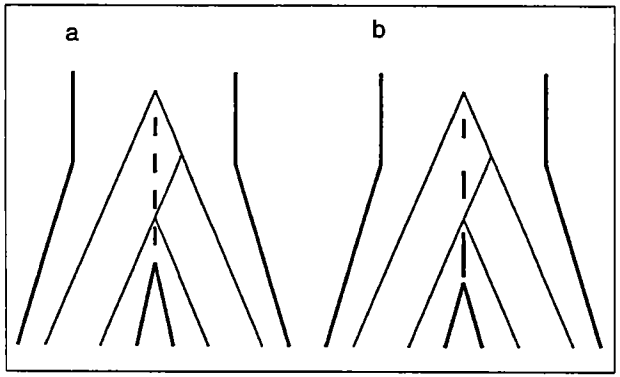

Figure 4. The two mixed neutral models discussed in the text: (a) the hybrid of migration and isolation, and (b) the decay-of-gene-flow model. The dashed lines in (b) are drawn to indicate that gene flow occurs readily at first, then becomes less likely as time passes.

mimic a great number of different scenarios for speciation, ranging from strict isolation to very recent divergence with a long history of migration.

We can expect that both of these mixed models will have a greater variance among loci than a strict isolation model. They will be able to explain more phenomena than strict isolation, but at the price of greater complexity, i.e. more parameters. The hybrid model requires at least two more parameters: a time of onset for migration, and a rate of migration. The decay-of-gene-flow model will require one extra parameter: a rate of decay in migration rate. These two models may be extremely similar in their predictions about genetic variation and divergence, and so may be very difficult to distinguish. All other things being equal, the decay-of-gene-flow model may be preferable, since it requires fewer parameters.

Adding natural selection

The models discussed so far have all assumed that natural selection has not shaped the pattern of variation. If speciation could occur simply as the by-product of divergence via genetic drift between populations – with no natural selection – then the isolation model and the mixed models could be considered neutral models of speciation. However, it is well known that natural selection can have a large impact, and most theories of speciation include selection. As in the traditional within-species neutralist-selectionist debate, we can adopt the neutral models of divergence as null models of speciation, and ask whether observed patterns of genetic variation require that natural selection be invoked in addition to, or together with, the processes of migration and isolation.

Variation among loci is even more important here than in the context of single populations. Some loci may be subject to selection, and this may contribute to species differences, while other, unlinked loci may conform perfectly to the neutral isolation model or to one of the mixed models. The neutral models of section 3 do not specify the causes of isolation and migration. Under a neutral isolation or mixed model, we generally imagine some sort of geographic barrier. However, when natural selection enters, the possibility exists that selection itself is the cause of isolation and divergence.

Divergence models and speciation theories

As mentioned above, the isolation model of divergence corresponds roughly to (neutral) allopatric speciation. One of the most interesting ways in which natural selection might contribute to divergence under allopatry is if one population is very small and isolated from a larger primary population. Such an isolated population may be in new circumstances, both environmental and genetical. For instance, if the population is very small, individuals may become quite inbred. A variety of scenarios have been envisioned whereby such a founder population might undergo considerable and novel adaptations (Mayr, 1963; Carson, 1978; Templeton, 1981), and these could increase the rate of divergence for a period of time just after the formation of the population. These adaptations may collectively have a pleiotropic effect

such that the new population is reproductively isolated from the primary population, should they again come into contact.

These kinds of allopatric speciation models that require a small founder population yield specific predictions of patterns of variation. Small founder population models predict that data should show evidence of a population bottleneck, i.e. reduced variation for a period of time after the bottleneck. Even if the population rebounds in size and variation accumulates, Tajima's (1989b) or Fu and Li's (1993) tests could detect residual effects on the genealogies of nucleotide sites. Also, a population bottleneck will affect all loci in the same way, so every locus sampled should give similar evidence.

When natural selection is considered within the context of a model that includes migration, a fairly general situation emerges in which natural selection contributes to speciation. In brief, natural selection acting to reduce migration is tantamount to selection for reproductive isolation, and thus will further the process of speciation. This situation can arise whenever there are two groups of organisms that can exchange genes, but in which the hybrids have lower fitness than the parents. There seem to be at least two categories of natural selection that can arise in this context. One kind acts on loci that are the sites of differential adaptation in the two populations or that are epistatic with loci associated with differential adaptation. Individuals that are heterozygous at these loci, for one allele from each population, may have poor fitness. Natural selection is then manifest as poor hybrid fitness. The second kind of locus is one that can contribute to mate choice and where alleles that lead to preferential mating within a population (i.e. avoidance of hybrid mating) are favored by natural selection. This kind of adaptation could arise if some individuals are involved in relatively unsuccessful hybrid matings, and again, the source of selection is poor hybrid fitness.

Models that include these types of selection have traditionally been divided into several geographic categories: sympatric speciation wherein incipient species exist at least partly within overlapping geographic ranges; parapatric speciation where the populations abut one another; and models in which divergence begins to accrue under allopatry, but reproductive isolation is not complete before the populations come back into contact.

All these situations have the potential to lead to the evolution of premating barriers to gene flow and thus to speciation. Whether or not speciation occurs depends on the details of available genetic variation, levels of hybridization and gene flow, and the magnitude of selection coefficients against hybridization. Hereafter we will consider these models collectively. We may call them gene flow-selection models, as they all have in common the feature that gene flow can occur during speciation, but that it is most restricted for those loci that are the cause of low hybrid fitness.

Gene flow, selection and hypothesis testing

The gene flow-selection models could be quite difficult to evaluate directly, as has been done for the neutral models. Formal models would need additional, possibly many more, parameters than the neutral models. However, the gene flow-selection models exhibit two basic differences from neutral models: gene flow-selection models have different predictions, particularly regarding variation among loci; and they are more complex. Thus gene flow-selection models are ideal alternative hypotheses, to be contrasted with simpler null speciation models in statistical tests. While it may be difficult to specify the best gene flow-selection model for a particular data set, it may be possible to reject neutral models and show that, as a class, gene flow-selection models must be considered.

For some models, and some patterns of variation, these tests may be difficult, and it may be difficult to distinguish between a gene flow-selection model and a neutral model. For example, a model of sympatric or parapatric speciation with a period of migration before complete reproductive isolation would be impossible to distinguish from the neutral hybrid model of the section "Mixed models" if the loci under study are not linked to selected loci. A gene flow-selection model might be indicated if the variation found among loci is too great to be explained by neutral mixed models. This will occur only when some of the loci sampled are either under selection or are linked to selected loci and others are not. Loci that are not affected by selection for divergence or speciation will reflect the underlying, population-level process of migration. Despite the inherently large variances expected under migration, selected loci may appear as outliers.

This framework – isolation or mixed models as null and gene flow-selection models as alternative – can only be informative to the extent that the two classes of models differ in the predictions about divergence. At the level of genealogies, gene flow-selection models have a fundamental difference from neutral isolation or mixed models. Under neutral models, all loci are subject to common factors of shared effective population size and a presence or absence of gene flow between populations. Under gene flow-selection models, different loci may be subject to different levels of selection against gene flow. Thus, tests which examine variation among loci will be very useful. However, tests which focus on the pattern of variation at a single locus in the way Tajima's (1989b) and Fu and Li's (1993) do in the context of single populations may also be helpful in rejecting mixed neutral models in favor of speciation via selection.

Consider a comparative DNA data set collected from two divergent populations, and suppose the data include information on variation within and between the populations from each of several loci. The isolation and mixed models predict that all loci will have experienced common effective population sizes, and that divergence began at a single time point for all loci. In contrast, the gene flow-selection models predict that some portions of the genome may have relatively high levels of gene

flow and little divergence, while loci linked to sites under selection (because of low hybrid fitness) will experience low or zero gene flow. If the data set includes loci of both types (or with sufficient linkage to both types), then some loci may show very little divergence, while others may show considerable divergence. This variation among loci may be much higher than is expected under the isolation or mixed models.

When neutral models are rejected in favor of gene flow-selection models, the process will also necessarily generate hypotheses of natural selection that may be amenable to additional tests. For example, a locus that does not reveal evidence of gene flow, when contrasted with others that do, may be a candidate for linkage to a site that is under selection against gene flow. If either an isolation or mixed model of divergence were true in this case, then the locus is just an outlier of the neutral model distribution. Thus if the locus could be subjected to an independent test of gene flow, or natural selection, in a controlled setting, it might become possible to further distinguish the gene flow-selection models from the neutral models.

Examples from *Drosophila*

Table 1 shows a three-locus data set collected from two species of the *D. melanogaster* species complex. The values shown are the numbers of S_{x1}, S_{x2}, S_f and S_s (see section "Isolation") observed for *D. mauritiana* and *D. simulans*. These numbers lead to the following parameter estimates: $\theta_1 = 30.2$, $\theta_2 = 23.0$, $\theta_A = 28.6$ and $T = 0.6$. Thus the isolation model fit indicates a common ancestor population intermediate in size to both descendants. It turns out that a statistical test of a specific isolation model that assumes the ancestor had intermediate population size has been in wide use for some years. The HKA test, though primarily used to test for the effect of natural selection on patterns of variation, is also a test of this specific isolation model (Hudson et al., 1987). When the data in Table 1 were put to this test, the fit between the data and model expectations was very good (Hey and Kliman, 1993).

Table 2 shows another three-locus data set, this time from *D. pseudoobscura* and *D. persimilis*. Note that the data in Table 1 show some variation among loci, particularly in S_s, but that the variation among loci is much greater in the data in Table 2. Indeed, one locus (Adh) shows no fixed differences and a very large number of shared polymorphisms. When the general isolation model was fit to these data, the model parameter estimates were: $\theta_1 = 28.7$, $\theta_2 = 24.9$, $\theta_A = 102.9$ and $T = 0.48$. Thus, taken together the data suggest a model in which the ancestral population size was far larger than that of either descendant. When a statistical test of the quality of fit between the data and this isolation model was made (via computer simulation) the fit was quite poor, and the model was rejected (Wang et al., 1997). In the paper describing these findings, we conclude that an isolation speciation

Table 1. Segregating sites in D. simulans *and* D. mauritiana

	n_1	n_1	S_{x1}	S_{x2}	S_s	S_f
Period	6	6	43	37	11	3
Zeste	6	6	18	9	0	1
Yp2	6	6	3	4	0	2

Note. Species 1 is *D. simulans* and species 2 is *D. mauritiana*; n_1 and n_2 are the number of sequences. These data are from Kliman and Hey (1993) and Hey and Kliman (1993).

Table 2. Segregating sites in D. pseudoobscura *and* D. persimilis

	n_1	n_1	S_{x1}	S_{x2}	S_s	S_f
Period	11	11	42	30	6	2
Hsp82	11	11	33	9	1	8
Adh	99	6	333	27	67	0

Note. Species 1 is *D. pseudoobscura* and species 2 is *D. persimilis*; n_1 and n_2 are the number of sequences. These data are from Wang et al. (1997), Wang and Hey (1996) and Schaeffer and Miller (1991, 1992).

model does not fit the *D. pseudoobscura/D. persimilis* data, and other models that include migration must be considered.

Conclusion

Traditional approaches to the study of speciation have faced practical difficulties (i.e. it has not always been clear what kind of data should be collected) and epistemological uncertainties (due to the species problem and a lack of a hypothesis-testing framework). This chapter has outlined an approach for the study of speciation that overcomes some of these shortcomings. We have described several formal population genetic models of speciation that generate specific predictions of patterns of genetic variation. These predictions bear a very close correspondence to the kinds of observations that are made using multilocus comparative DNA sequence data sets, so it is possible to fit the speciation models to data. Finally, we show how some speciation models are more complex than others, and how this complexity permits a hypothesis testing hierarchy.

Acknowledgements

This work was supported by NIH grant GM 17745 to JW and NSF grant DEB-9306625 to JH.

References

Carson, H. L. (1978) Speciation and sexual selection in Hawaiian *Drosophila*. *In*: Brussard, P. F. (ed.) *Ecological Genetics: The Interface*, Springer-Verlag, New York, pp. 93–107.

Ewens, W. J. (1979) *Mathematical Population Genetics*, Springer-Verlag, Berlin.

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

Fu, X.-Y. (1995) Statistical properties of segregating sites. *Theoret. Pop. Biol.* 48: 172–197.

Fu, X.-Y. and Li, W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.

Gillespie, J. H. and C. H. Langley (1979) Are evolutionary rates really variable? *J. Mol. Evol.* 13: 27–34.

Hey, J. (1991) The structure of genealogies and the distribution of fixed differences between DNA sequences from natural populations. *Genetics* 128 831–840.

Hey, J. (1994) Bridging phylogenetics and population genetics with gene tree models. *In*: Schierwater, B., Streit, B., Wagner, G, P. and DeSalle, R. (eds) *Molecular Ecology and Evolution: Approaches and Applications*, Birkhäuser, Basel, pp. 435–449.

Hey, J. and Kliman, R. M. (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Molec. Biol. Evol.* 10: 804–822.

Hilton, H., Kliman, R. M. and Hey, J. (1994) Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* complex. *Evolution* 48: 1900–1913.

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

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

Hudson, R. R., Kreitman, M. and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.

Kaplan, N. L., Darden, T. and Hudson, R. R. (1988) Coalescent process in models with selection. *Genetics* 120: 819–829.

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

Kliman, R. M. and Hey, J. (1993) DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. *Genetics* 133: 375–387.

Li, W.-H. (1976) Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. *Theor. Pop. Biol.* 10: 303–308.

Mayr, E. (1963) *Animal Species and Evolution*, Belknap Press, Cambridge, MA.

Schaeffer, S. W. and Miller, E. L. (1991) Nuceotide sequence analysis of *adh* genes estimates the time of geographic isolation of the B ogota population of *Drosophila pseudoobscura*.

Proc. Natl. Acad. Sci. USA 88: 6097–6101.

Schaeffer, S. W. and Miller, E. L. (1992) Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. *Genetics* 132: 471–480.

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

Slatkin, M. and Hudson, R. R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555–562.

Slatkin, M. and Maddison, W. P. (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* 123: 603–613.

Tajima, F. (1989a) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-population model. *Genetics* 123: 229–240.

Tajima, F. (1989b) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

Takahata, N. and Nei, M. (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110: 325–344.

Takahata, N. and M. Slatkin (1990) Genealogy of neutral genes in two partially isolated populations. *Theor. Pop. Biol.* 38: 331–350.

Templeton, A. R. (1981) Mechanisms of speciation – a population genetic approach. *Annu. Rev. Ecol. Syst.* 12: 23–48.

Wakeley, J. (1996a) The variance of pairwise nucleotide differences in two populations with migration. *Theor. Pop. Biol.* 49: 39–57.

Wakeley, J. (1996b) Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Pop. Biol.* 49: 369–386.

Wakeley, J. (1996c) Pairwise differences under a general model of population subdivision. *J. Genetics* 75: 81–89.

Wakeley, J. and Hey, J. (1997) Estimating ancestral population sizes. *Genetics* 145: 847–855.

Wang, R.-L and Hey, J. (1996) Speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. *Genetics* 114: 1113–1126.

Wang, R.-L., Wakeley, J. and Hey, J. (1997) Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* 147: 1091–1106.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7: 256–276.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* 16: 97–159.