QP 435-449.

Editors

B. Schierwater
B. Streit
Zoologisches Institut
der Universität Frankfurt
Siesmayerstr. 70
D-60054 Frankfurt

G.P. Wagner
Department of Biology
Yale University
165 Prospect St.
New Haven, CT 06511
USA

R. DeSalle
Department of Entomology
American Museum of Natural History
79th Street at Central Park West
New York, NY 10024
USA

Library of Congress Cataloging-in-Publication Data
Molecular ecology and evolution: approaches and applications / edited by B. Schierwater . . . [et al.].
—(EXS; 69)
Includes bibliographical references and index.
ISBN 3-7643-2942-4 (acid-free):
ISBN 0-8176-2942-4 (U.S.: acid-free)
1. Molecular evolution. 2. Molecular ecology. 3. Population genetics. I. Schierwater, B. (Bernd), 1958– . II. Series.
QH371.M72 1994
575—dc20

Deutsche Bibliothek Cataloging-in-Publication Data
Molecular ecology and evolution: approaches and applications / ed. by B. Schierwater . . . – Basel; Boston; Berlin:
Birkhäuser, 1994
(EXS; 69)
ISBN 3-7643-2942-4 (Basel . . .)
ISBN 0-8176-2942-4 (Boston)
NE: Schierwater, Bernd [Hrsg.]

Printed on acid-free paper produced from chlorine-free pulp
Cover illustration: A. Ender and B. Schierwater
Printed in Germany

ISBN 3-7643-2942-4
ISBN 0-8176-2942-4

9 8 7 6 5 4 3 2 1

ed. by B. Schierwater, B. Streit, G.P. Wagner & R. DeSalle

Bridging phylogenetics and population genetics with gene tree models

J. Hey

Rutgers University, Nelson Labs, Piscataway, NJ 08855-1059, USA

Summary. Current gene tree models, developed and used by population geneticists for research on natural selection, can also be used to ask questions about the formation of species. When these gene tree models are joined with a null model of speciation, a research plan emerges that shows promise of revealing the extent to which genetic variation between populations contributes to the formation of species. The empirical element of this research plan requires that multiple DNA sequences be collected from each of the species investigated, and that these data come from multiple loci. Examples of these models and their application to recent data on the *Drosophila melanogaster* species complex are given.

This report outlines an emergent protocol in evolutionary genetics: the application of genealogical, population genetic models to the inquiry of evolutionary forces (e.g., natural selection, genetic drift, gene flow) associated with the formation of species. On the empirical side the emphasis is on DNA sequence data sets in which multiple sequences have been collected from each of two or more species *and* for multiple loci. The article by Templeton in this volume addresses similar issues.

Simplifying the question – building a model

I will outline a reductionist population genetic approach to the study of speciation that, at its core, draws relatively little from many issues in current speciation debates (e.g., sympatric speciation, genetic architectures, reinforcement; see articles in Otte and Endler, 1989). Typically, population genetics proceeds by statistical assessments of null models. Despite the fact that these models usually employ manifestly extreme assumptions (e.g., strict neutrality of mutations, panmixia, constant population size), they have a diverse record, including being highly explanatory in some circumstances and being strongly rejected in other circumstances. A good example is the neutral theory of molecular evolution (Kimura, 1983). This theory is analytically tractable so that it forms a part of the null hypothesis in a wide array of statistical tests of evolutionary forces (e.g., Hudson et al., 1987; Tajima, 1989; Slatkin, 1989; McDonald and Kreitman, 1991). In patricular, it remains a

standard of current genealogical modeling (Ewens, 1990; Hudson, 1990), most commonly in the form of the infinite sites model (Kimura, 1969). The rejection of the neutral model under some circumstances has lead a number of population geneticists to conclude they have gained significant knowledge of the action of important types of natural selection at or near the loci they study (Hudson et al., 1987; Berry et al., 1991; Begun and Aquadro, 1991; McDonald and Kreitman, 1992; Stephan and Mitchell, 1992; Langley et al., 1993; Eanes et al., 1993).

Null speciation models

To extend a population genetics approach to the study of speciation, we need to consider the effect of speciation on genetic variation between species. Specifically, we need a simple, or null model of speciation with as few parameters as possible. From the viewpoint of DNA sequence variation at genetic loci, the simplest model is one in which no variation accumulates between species, either as a cause or an effect of speciation. In the face of ubiquitous evidence of DNA sequence variation among species, a more general model is required. We will begin with the assumption that genetic variation at the loci under investigation has had nothing to do with speciation. In other words, assume that genetic variation at the loci under investigation has made no contribution to the defining characteristics of the species (whatever they may be, see below). Secondly, allow the possibility that genetic variation between species may have arisen at these loci as a result of an absence of gene flow between species. Thus, we permit the possibility that for all of the loci under investigation, there was a point in time when gene flow ceased between species. This more general view includes the possibility that, for the loci under study, gene exchange has not ceased between the species. An even more general model with reduced, but non-zero gene flow could also be examined, however, this is more difficult and will not be considered in this report.

Note that this strict focus on gene flow need not imply a particular idea of the nature of species. For the most part, a particular species concept enters the discussion at the point in which species are identified for study. However, it is possible that the identification of species will complicate this null model. For example, if species are viewed under the biological species concept, then the nature of species and species formation are defined by an actual or hypothetical test of gene flow failure. It is preferable that the biological species concept be avoided in identifying species, since this is tantamount to assuming non-zero values of a key parameter of the null model (i.e., the time of cessation of gene flow). The null model is most useful for cases where species are identified without implications on the form or magnitude of gene flow. Typically in practice, a group of organisms is called a species if the individuals share a number of characteristics *and* if those characteristics differentiate them from other organisms. While it is often known or assumed that the defining character state differences reflect gene sequence differences (sometimes character states are gene sequences), these differences may be limited to only a small portion of the genome.

Population genetics

So far, the model does not yet include any ideas about evolution within species. To begin, we assume that mutations are neutral and follow the infinite sites model of Kimura (1969). The simplification of assuming an infinite number of sites for mutation is often quite justifiable for recently diverged gene copies where few mutations have occured and multiple hits are not likely. Furthermore, by assuming neutrality and no natural selection, the process of genetic drift occurs independently from the process of mutation. In a gene tree view (see below), neutrality means that the processes that determine branch lengths and the shape of the gene trees (meaning the actual pattern of historical relationships, *not* estimates of gene trees from data) are not affected by the mutations that have occured. With natural selection out of the way, the remaining fundamental population issue is genetic drift, which depends on population size, the distribution of family sizes, and population structure. For this, the most widely used model is the Fisher-Wright model, which assumes a constant population size, with a Poisson distribution of family sizes, and no population structure (Ewens, 1979). For the case of a model of the divergence of two recently formed species, there are three population sizes to be considered: those of the two species *and* that of their common ancestor.

Gene trees

At this point it is helpful to belabor a review of the widely used figure that forms a graphical basis for much of the theory that has been developed for the study of DNA sequence differences. Figure 1 shows a rooted bifurcating network, or tree, representing the history of a sample of three gene copies. Note that Figure 1 is given as a hypothetical depiction of a true history, and is not to be confused with a tree that has been estimated from comparative data. The principle features include: the directionality of time, from the past to the present; straight lines, also called branches; a series of points that mark the ends of lines at the present moment; and nodes, the junction of branches. In a genealogical context, the tips of branches at the present refer to different copies of

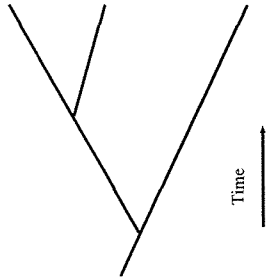

Figure 1. A rooted bifurcating network.

homologous DNA sequences, and the rest of the diagram is a description of the history. The branches refer precisely to the persistence of a DNA sequence through time. This persistence means at times the physical persistence, but also includes numerous cases of replication, wherein it is the information in the sequence that persists. The tree for a single gene copy would simply be a straight line extending from the past to the present. The nodes of the tree refer precisely to those cases of DNA replication in which both copies of the sequence that arose from a replication event were ancestors of sequences represented as tips of branches at the present moment. Thus, although the persistence of sequences through many replication events is represented with branches, nodes are used to represent the minority of replication events for which both copies are ancestors of sampled gene copies. It is not possible, given current knowledge of the action of DNA and RNA polymerases, to have three branches descendent from a node, because that would represent three sequences emerging from a single replication event. Thus, the graphical model of a genealogy flows from well established knowledge of DNA replication and, with one important exception, will not be tested. The exception is recombination. If recombination occured among the ancestral sequences of a sample, then a bifurcating diagram cannot match the historical topology. This realization has actually been used to develop tests of recombination using gene tree models and DNA sequence data (Hudson and Kaplan, 1985). Again for simplicity, assume recombination has not taken place within the region to be considered.

The meaning intended for Figure 1 is similar in many ways to that for tree diagrams that represent the history of species. In these cases, the tips of the branches represent extant species; the branches represent the persistence of species through time, and the nodes represent cases of speciation. Species trees face at least two elements of uncertainty. First, it is often unclear (because the species concept is often not articulated) what is meant by the persistence of species through time. Second, the

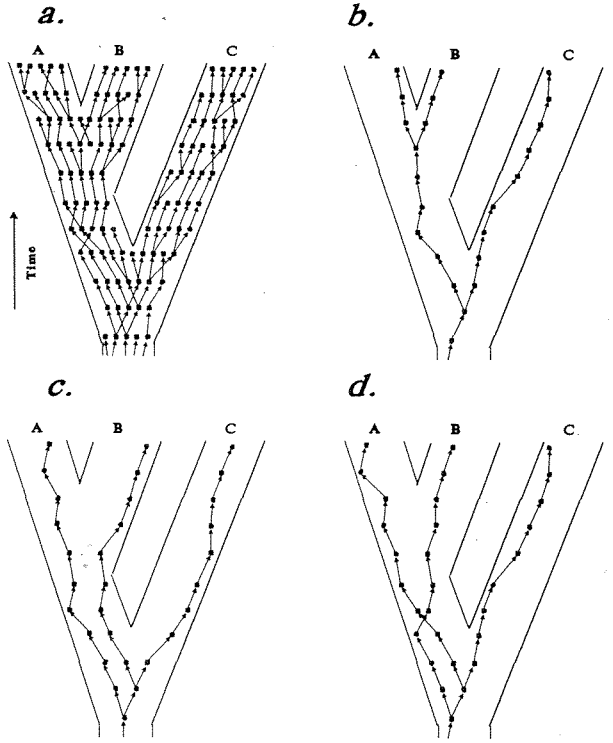

Figure 2. Gene trees within population trees. See text for explanation.

nodes represent speciation as an instant in time, when in fact the duration of the speciation process is generally not known. Also, there is generally no biological reason (akin to the constraints of DNA replication) to limit speciation to the formation of just two species, and thus no reason to exclude from consideration nodes with multiple descendent branches.

To include genealogies within a discussion of species divergence, many authors have enclosed gene trees within wider "population" trees (sometimes referred to hereafter as skinny trees and fat trees, respectively). Figure 2 depicts a fat tree in which widely spaced parallel lines are branches representing the persistence of population of gene copies. In practice, populations are identified with different species, but with a focus on the history of specific loci and given the null speciation model, it is more precise to refer to populations of gene copies. The junctions of wide branches are nodes representing time points when two popula-

tions of gene copies ceased exchange. The time point may coincide with the time of speciation depending on the species concept, but to be general and precise, it will be referred to as the point of gene flow cessation. Dots represent gene copies and arrows represent the flow of information (i.e., the DNA sequence) from one cell generation to the next. Figure 2a shows a particular realization for all copies of a gene within three recently diverged populations. Each of the populations has a very small size with only a few gene copies each. Figures 2b, c, and d show just a subset of the dots and arrows of Figure 2a, with different gene copies sampled in each case. The varying topologies of the skinny trees in these figures are intended to drive home the point that gene trees for sequences drawn from different populations need not match the topology, or branching order of the fat free for the populations from which the sequences came. It should also be clear that for sample sizes larger than one, no single branching pattern for the populations (let alone the correct one) may be evident from the gene tree. For example, consider a hypothetical data set for a locus at which five sequences were sampled from each of three populations, and suppose that from the sequence variation found within the data, the gene tree in Figure 2a – minus the embracing fat tree and minus those branches that end before the present – was found to be the best estimate of the historical branching pattern. Clearly (as shown in Figs 2b, c, and d), the gene tree within the fat tree of Figure 2a does not simply correspond to any one single population tree.

Joining the models

Together, the neutral mutation model, the Fisher-Wright model, and the null speciation model can be used to generate quantitative probabilistic descriptions of gene tree lengths (Takahata and Nei, 1985; Hudson et al., 1987; Hey, 1991). We also have the fat trees and skinny trees that form an accessible graphical footing to aid our analytical models and intuition.

With these models in hand, an empirical research plan emerges. If speciation has in fact happened in a way roughly like the null model, and genetic variation at the loci under study has not played a role in speciation, then the cessation of gene flow has happened at the same time for all of the loci under study. This means that studies on multiple loci should all reveal the same underlying population level processes. In other words, the model predicts that the interspecific divergence that is encountered in the data for each locus should be consistent with a single time point at which gene exchange ceased *and* that this time point should be the same for all loci. Alternatively, if speciation has not occurred in this way and, in particular, if one or more of the loci were "involved" in speciation, then the gene trees for samples from these loci may reflect a different history than for loci that did not contribute to the speciation.

The idea of studying multiple loci to address speciation questions has much in parallel with genealogical studies on natural selection. Because of linkage, strong directional or balancing selection on a very small part of the genome (e.g., a single nucleotide) is expected to affect the structure of the gene tree (and thus levels of variation) over a larger region of the genome. By extending a research program on natural selection (or the role of natural selection in speciation) to multiple loci, one can discriminate between forces that are expected to affect all loci similarly and forces that act on smaller portions of the genome. The first category consists of forces that act on populations, such as genetic drift and population subdivision. In contrast, natural selection acting on functional variation at individual loci is not expected to affect variation at effectively unlinked loci. In short, loci with a recent history of natural selection may have different patterns of variation within *and* between species than do other loci.

A widely used test of natural selection is the HKA test (Hudson et al., 1987) which is intended for data sets that have multiple DNA sequences (or RFLP data) from within each of two species for each of two or more loci. The test proceeds by fitting a neutral model (that includes a speciation component essentially identical to that described here) to all of the data. The procedure generates expected quantities of intra-and interspecific variation for each of the loci, as well as an overall measure of the goodness-of-fit. Thus, although originally intended and routinely applied to questions concerning natural selection, the test can also be used to see whether all of the loci are consistent with the null model of speciation.

Examples of models and data

The probability of an exclusive node as a function of divergence time

The depth of a genealogy (i.e., the times of the nodes) for a sample of gene copies from a single species is a function of the number of gene copies sampled *and* the population size. For a population of effective size N_e under Fisher-Wright assumptions, two sequences chosen at random will have had a common ancestor $2N_e$ generations ago, on average. The common ancestor for all gene copies will have occured $4N_e$ generations ago on average. Thus, for a data set with sequences from within and among multiple species, discordance between the topology of the genealogy and the topology of the species tree (or population tree, see *Gene trees* above) is only expected to occur when the time between

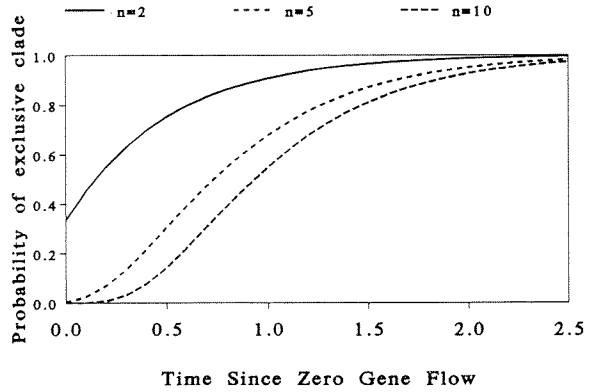

Figure 3. The probability of obtaining a sample with a genealogical history of an exclusive clade in one or both species. The calculations were done using expressions (7) and (19) of Hey (1991). These calculations assume that both species as well as the ancestral species have the same effective population size. The time since the cessation of gene flow is in units of $2N_e$ generations. The sample size, n, refers to the number of sequences randomly drawn from each species.

speciation events is of the order of N_e generations or less. One way to consider the issue of topological correspondence between genealogies and species trees, is to consider samples from two species and inquire about the time that must pass following the cessation of gene flow for the gene trees within species to be exclusive of lineages of the other species. Figure 3 shows, for two species of identical N_e, the relationship between sample size, the time since the cessation of gene flow, and the probability that one of the samples has a genealogy exclusive of the other. For cases where gene flow ceased very recently, there is a large chance that DNA sequences will have an intermingled genealogy, especially for large samples. Furthermore, different independently segregating loci will have different genealogies, simply by chance. For cases of recent speciation some loci may reveal a genealogy in which the sequences from the different species are separate on the tree while others do not.

Data

One of the best studied groups of recently formed species is the *Drosophila melanogaster* species complex. Two of the species, *D. melanogaster* and *D. simulans*, are cosmopolitian, while the other two, *D. mauritiana* and *D. sechellia*, are endemic to oceanic islands. Individuals are identified to species on the basis of the morphology of external male genitalia (see, e.g., Ashburner, 1989). Numerous phylogenetic studies have shown only that *D. melanogaster* is a sister taxon to the other species (Lachaise et al., 1988). *Drosophila simulans, D. sechellia, and D. mauritiana* (sometimes referred to collectively as the *simulans* complex) are similar to one another and, despite considerable effort, a bifurcating species tree has not been unambiguously determined (Bodmer and Ashburner, 1984; Cohn, Thompson and Moore, 1984; Coyne and Kreitman, 1986; Lachaise et al., 1988; Caccone, Amato and Powell, 1988).

Recent reports out of my laboratory describe a DNA sequence data set of three X-linked loci; for each locus six gene copies were sequenced from each of the four species (Kliman and Hey, 1993; Hey and Kliman, 1993). Figure 4 shows an estimated genealogy for the *zeste* locus. The trees for the *period* and *yolk protein 2* loci differ in detail, but are fully supportive of the points to be made in this review.

Like previous studies, Figure 4 supports a historical view of *D. melanogaster* having a relatively ancient separation from other species. Also like previous studies, our data do not resolve a clear bifurcating species tree for the *simulans* complex. Figure 4 shows, as did the trees for *yolk protein 2* and the *period* locus, that some of the *D. simulans*

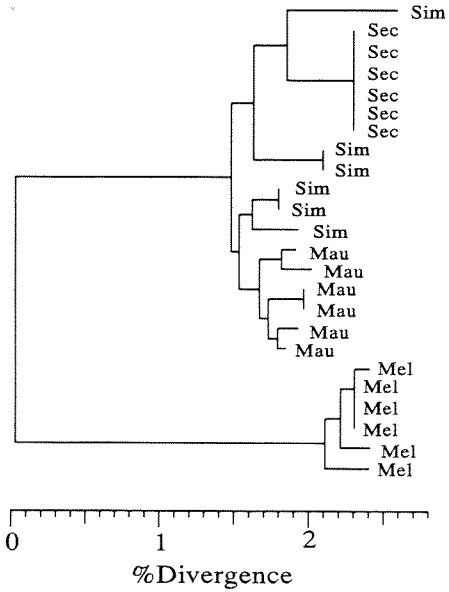

Figure 4. A neighbor-joining tree (Saitou and Nei, 1987) redrawn from Hey and Kliman (1993). The species designations are: *D. melanogaster* (Mel); *D. simulans* (Sim); *D. mauritiana* (Mau); and *D. sechellia* (Sec).

sequences shared a common ancestor with modern *D. mauritiana* sequences more recently than with other *D. simulans* sequences. Similarly, some *D. simulans* are more closely related to *D. sechellia*. Thus, it appears that the loci in present day *D. simulans* are still segregating lineages that have persisted since before the origin of the island species.

The *zeste* data were also interesting for what they revealed about current effective population sizes of the species. In particular, the *D. sechellia* sequences were all identical and it appears that this species has very little variation, which is suggestive of small population size. If this were the only locus, we would not be able to rule out natural selection or accidental sampling as the cause of this low variation. However, the same pattern was found at the other loci. Interestingly, the other island endemic, *D. mauritiana*, revealed as much or more intraspecific variation as *D. simulans* and *D. melanogaster*.

One of the most interesting findings was that at the period locus, *D. simulans* and *D. mauritiana* share several polymorphisms. At present, it is not clear whether these shared polymorphisms reflect a large population size for *D. mauritiana* during and since formation, or some pattern of limited gene flow after divergence began.

Estimating times of gene flow cessation with gene tree models

In the case of a data set of aligned DNA sequences, collected from two populations or species, a commonly used descriptor of divergence is the simple average of the number of differences observed between all possible interspecific comparisons. For example, with a data set of three sequences from one species and four from another, gross divergence is calculated as the average of 12 different pairwise comparisons. However, this quantity is expected to include variation that was present within the ancestral species prior to speciation, as well as variation that has accumulated since speciation. To estimate the time since speciation, or more strictly, the amount of divergence that has occured since gene flow stopped, we would like to have a measure of just that component of interspecific variation that has accumulated since speciation. By far, the most common approach is to ignore the ancestral intraspecific component. For speciation events that are very old relative to the time scale of the persistence of intraspecific variation, this works just fine. However, for recent speciation events, the ancestral intraspecific component may make up much or most of the divergence. An adjustment of this component can be made if we assume that the amount of variation in the ancestral population was similar to that observed within the two descendent species. A common descriptor of intraspecific variation is the average number of pairwise differences. Commonly called π (Nei and Tajima, 1981), this quantity is calculated much like gross interspecific divergence: for n sequences, the average is taken among all $n\cdot(n-1)/2$ pairwise comparisons. This measure of nucleotide heterozygosity is often used as an estimate of $4N_e u$ (a widely used parameter in population genetic models), where u is the neutral mutation rate. If we let D_{ij} refer to the observed gross divergence between species i and j, then net divergence (δ_{ij}) is equal to $D_{ij}-(\pi_i+\pi_j)/2$ (Nei, 1987, page 276). In other words, net divergence is equal to gross divergence less the average of the two species's intraspecific variation. Put another way, if the size of the ancestral population prior to speciation was equal to the average population size of the decendent species, then net divergence is equal to twice the average number of mutations that has occured on a lineage since the speciation event. If gene flow has not yet stopped, then net divergence has an expectation of zero.

The HKA test (Hudson et al., 1987) employs the assumption that ancestral population size was the average of the descendent populations. In a sense, the method uses current population sizes as a way to guess about ancestral population sizes. Contingent on this assumption (and others), the HKA test can account for that portion of interspecific variation due to ancestral polymorphism, and returns estimates of the time since gene flow cessation. Thus, although the branching order of the species is not clear from Figure 4, we may be able to estimate the times when gene flow stopped. Also, since the HKA test is essentially a goodness-of-fit procedure, we can assess the overall fit of the model.

Table 1 shows the outcomes of two HKA tests; in each case *D. simulans* was paired with one of the island endemic species for the data sets in Kliman and Hey (1993) and Hey and Kliman (1993). The first thing to notice is that the goodness-of-fit statistic, X^2, does not approach statistical significance when compared with the appropriate chi-square distribution (4 degrees of freedom in this case). This means that at least for these loci, the data are consistent with the null population-genetic/speciation model. Secondly, we can compare the estimated times since zero gene flow. From Table 1 the times of the splits involving *D. mauritiana* and *D. sechllia* are 0.52 and 1.03, respec-

Table 1. HKA tests for three loci (*zeste*, *period*, and *yolk protein 2*) and two species (Hudson et al., 1987; Kliman and Hey, 1994; Hey and Kliman, 1993)

Species 1	Species 2	$\hat{T}$	95% Limits	X^2	p
simulans	*mauritiana*	0.52	0.0–3.13	0.95	0.918
simulans	*sechellia*	1.03	0.58–4.29	2.82	0.588

$\hat{T}$ is an estimate of the time since the cessation of gene flow between species in units of 3/2 N_e generations, where N_e is effective population size for *D. simulans*. The 95% confidence intervals were determined by simulation of 1000 replicates, and then taking those values in the 97.5% and 2.5% positions in the ranked values (see text). X^2 is the goodness-of-fit statistic. p is the probability of observing an X^2 value greater than or equal to the actual value, assuming a X^2 distribution with 4 degrees of freedom.

tively. These times are in units of $3/2N_e$ generations (3/2 rather than 2 because the loci are sex-linked), where N_e, in this case, is the effective population size of *D. simulans*. If we assume that the time of speciation corresponded roughly with the times that gene flow ceased for these loci, then it appears the speciation event giving rise to *D. sechellia* occurred prior to (i.e., longer ago) than that for *D. mauritiana*. Table 1 also shows the 95% confidence intervals, generated by simulation, of the estimated time since zero gene flow. The overlapping confidence intervals suggest that we cannot reject either branching order, though this comparison is not strictly a test of this. The wide confidence intervals also reflect the difficulty in discerning species trees for recent and closely spaced speciation events. These simulations were carried out in a coalescent fashion (Hudson, 1990) in strict accord with the assumptions of the HKA test (Hudson et al., 1987) using the parameter estimates generated using the HKA test on the original data.

Natural selection and speciation models

Suppose that the null speciation model is not correct, and that some of the genetic variation at the loci under study has contributed to the maintenance of the criteria used for identifying species. We may envision a scenario of differential adaptation, with different functional alleles suited to different environments; alternatively, there may be underdominance whereby heterozgotes (i.e., hybirds) are less fit. At any rate, there are numerous speciation models in which, because of natural selection, some loci experience less gene flow between the populations than others. If speciation occurs while there is some gene flow, then those loci that are not included in the gene flow will have different gene trees than those that are included. Put another way, the time of the node on the fat tree will be different for the two classes of loci. A comparison of gene trees, one for a locus with a history of contributing to species formation and one for a locus not involved in limiting gene flow, is shown in Figure 5. Stephan and Mitchell (1992) describe for two Asian populations of *D. ananassae* a pattern of variation that is in some ways consistent with the model depicited in Figure 5. Whether or not this pattern reflects an early stage in speciation (i.e., whether these two populations will become species) is not clear.

Discussion

The purpose of this contribution has been to outline a research program whereby current models and methods of population genetics can be extended to the study of recent speciation events. Much of what has been written is implicit (and at times explicit) within the large body of literature on population level and species level variation in mitochondrial genomes (see Avise, 1991). The research program outlined here contrasts with much of the mitochondrial literature in two ways. The principal one is that here I emphasize (as have many others) the need for data from multiple loci, while organelle genomes segregate effectively as a single locus. Certainly, data from organelle genomes can inform on population genetic processes and can be included with studies of nuclear loci. A second distinction arises from the fact that a population genetic approach to speciation questions will be most informative for those cases in which some of the intraspecific variation predates the species divergence. This means that the species must be very closely related or they must have large population sizes so that they have maintained variation for a long period of time. It also means that genealogies of organelle genomes, which are expected to segregate under an N_e roughly one-quarter that of diploid nuclear loci, will carry old variation less often.

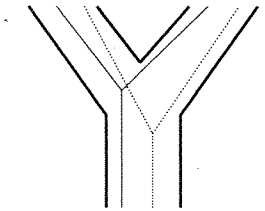

Figure 5. Contrasting gene trees for selected and non-selected loci, for samples with one gene copy drawn from each species. The fat tree represents the separation of the species as seen for a locus not under selection and not involved in speciation. The thin line represents a typical non-selected genealogy that might be expected for the locus. The dotted line represents an expected genealogy for a different selected locus (see Text).

The recent inquiries on speciation in the *D. melanogaster* species complex are surprising for what they do reveal as well as for what they do not. On the one hand, the conclusions about the age of variation within *D. simulans* relative to the origins of *D. mauritiana* and *D. sechellia* are exceptional. So, too, is the finding that *D. simulans* and *D. mauritiana* share a number of polymorphisms, thus providing unique evidence against a role for small effective population size in the formation of *D. mauritiana*. On the other hand, these data do not help much in addressing many long standing questions about speciation. For instance, we do not know whether the variation shared by *D. simulans* and *D. mauritiana* came about because of gene flow during speciation or whether it predates speciation and both species have been large since their isolation. The overall pattern of variation for *zeste*, *per*, and *yp2* was consistent with the null speciation model when examined with the

method of Hudson et al. (1987), and supporting the view that there has not been genetic contribution to the formation of barriers to gene flow. However, this support is weak because the possibility of gene flow after divergence began cannot be ruled out (especially at the *period* locus), and because the study included only three loci.

The paradox of new questions raised and old questions unresolved is especially clear for those interested in the shape of the species tree. We can say some interesting and novel things about the formation of *D. simulans*, *D. sechellia*, *and D. mauritiana*, but we still do not have a strong conclusion on the order of the speciation events.

Future perspectives

The research program outlined here is likely to prove especially informative when loci are included for the express purpose of testing whether they have contributed to the differentiation of species. At least two recent reports outline approaches that may provide candidate loci, within the near future, for the *simulans* complex. First, is the fine scale mapping of loci responsible for fitness loss of species hybrids (Wu et al., 1993). Second, is the large scale search, using two-dimensional electrophoresis, of loci encoding proteins with large interspecific differences in mobility or quantity (Zeng and Singh, 1993). Both approaches must overcome significant technical hurdles before specific genetic loci are identified; however, if and when they are found, it will be very interesting to see how their gene trees resemble those for *zeste*, *period*, and *yolk protein 2*.

Acknowledgements
Thanks to Holly Hilton for comments on the manuscript. This work was supported by National Science Foundation grant BSR 8918164.

References

Ashburner, M. (1989) *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, New York.
Avise, J.C. (1991) Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* 43: 1192–1208.
Begun, D. and Aquadro, C.F. (1991) Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* 129: 1147–1158.
Berry, A.J., Ajioka, J.W. and Kreitman, M. (1991) Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* 129: 1111–1117.
Bodmer, M. and Ashburner, M. (1984) Conservative and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. *Nature* 309: 425–430.
Caccone, A., Amato, G.D. and Powell, J.R. (1988) Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. *Genetics* 118: 671–683.
Cohn, V.H., Thompson, M.A. and Moore, G.P. (1984) Nucleotide sequence comparison of the Adh gene in three Drosophilids. *J. Mol. Evol.* 20: 31–37.
Coyne, J.A. and Kreitman, M. (1986) Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. *Evolution* 40: 673–691.
Eanes, W.F., Krichner, M. and Yoon, J. (1993) Evidence for adaptive evolution of the G6pd gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. *Proc. Natl. Acad. Sci. USA* 90: 7475–7479.
Ewens, W.J. (1979) *Mathematical Population Genetics*. Springer Verlag, New York.
Ewens, W.J. (1990) Population Genetics Theory – the past and the future. *In*: S. Lessard (ed.): *Mathematical and Statistical Development of Evolutionary Theory*. Kluwer Academic Publishers, Drodrecht, pp. 177–227.
Hey, J. (1991) The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics* 128: 831–840.
Hey, J. and Kliman, R.M. (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* complex. *Mol. Biol. Evol.* 10: 804–822.
Hudson, R.R. and Kaplan, N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164.
Hudson, R.R., Kreitman, M. and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.
Hudson, R.R. (1990) Gene genealogies and the coalescent process. *In*: P. H. Harvey and L. Partridge (eds): *Oxford Surveys in Evolutionary Biology*, Vol. 7, Oxford University Press, New York, pp. 1–44.
Kimura, M. (1969) The number of heterozgous nucleotide sites maintained in a finite population due to a steady flux of mutations. *Genetics* 61: 893–903.
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.
Lachaise, D., Cariou, M.-L., David, J.R., Lemeunier, F., Tsacas, L., and Ashburner, M. (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* 22: 159–225.
Langley, C., MacDonald, J.M., Miyashita, N., Aguadé, N. and M. (1993) Lack of correlation between interspecific divergence and intraspecific polymorphism at the *suppressor of forked* region in *Drosophila melanogaster* and *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* 90: 1800–1803.
Martin-Campos, J.M., Comeron, J.M., Miyashita, N. and Aguadé, M. (1992) Intraspecific and Interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics* 130: 805–816.
Nei, M. and Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* 97: 145–163.
Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, N.Y.
Otte, D. and Endler, J.A. (1989) *Speciation and its Consequences*. Sinauer Associates Inc., Sunderland, Massachusetts.
Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406–425.
Slatkin, M. (1989) Detecting small amounts of gene flow from phylogenies of alleles. *Genetics* 121: 609–612.
Stephan, W. and Mitchell, S.J. (1992) Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. *Genetics* 132: 1039–1045.
Tajima, F. (1989) 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.
Wu, C.-I., Perez, D.E., Davis, A.W., Johnson, N.A., Cabot, E.L., Palopolis, M.F. and Wu, M.-L. (1993) Molecular genetic studies of postmating reproductive isolation. *In*: N. Takahata and A.G. Clark (eds): *Mechanisms of molecular Evolution*. Sinauer Associates Inc., Sunderland, Massachussetts, pp. 191–212.
Zeng, L.-W. and Singh, R.S. (1993) A combined classical genetic and high resolution two-dimensional electrophoretic approach to the assessment of the number of genes affecting hybrid male sterility in *Drosophila simulans* and *Drosophila sechellia*. *Genetics* 135: 135–147.