17

Genealogical Portraits of Speciation in the *Drosophila melanogaster* Species Complex

Jody Hey and Richard M. Kliman

Introduction

A particular class of DNA sequence data set, one that is increasingly being generated for purposes of exploring natural selection, can also be used to study speciation. Minimally, these analyses require, for each of multiple loci, several DNA sequences from each of the species under investigation. The analysis of a recently generated data set for the four species of the *Drosophila melanogaster* complex is reviewed within the context of speciation. For three loci, a consistent pattern of variation emerges, in which it appears that ancestral *D. simulans* gave rise to both *D. mauritiana* and *D. sechellia*. Surprisingly, modern *D. simulans* appears to retain much of the genetic variation that existed prior to those speciation events.

This report describes recent progress in the application of modern population genetic methods to speciation problems. The basic methodology draws heavily from advances in genealogical modeling and from the use of DNA sequence data for studying natural selection. In recent years, the theoretical arm of population genetics has seized onto genealogical models (also called coalescent models). The paradigm has shifted away from a focus on allele frequencies and toward a focus on the structure of gene trees. Much of this shift has been driven by the actual data, simply because homologous DNA sequences are readily interpreted in terms of estimates of historical gene trees. As the nature of the data has changed, so have the models.[1,2]

The synergism of theoretical and empirical research has been especially effective in the investigation of natural selection acting on genetic loci, and several recent reports present statistical evidence for adaptive natural selection at or near the loci studied. The cases include fairly recent occurrences of balancing selection[3] and strong directional selection.[49] The methods used in these studies of natural

selection draw on two types of contrasts: (1) comparison of variation within species to that between species; and (2) comparison of variation among genomic regions (sometimes different loci and sometimes different portions of a single locus). These contrasts take place within the framework of a null model, in which each species has a constant population size and is not subject to natural selection. The model permits variation in the neutral mutation rate among loci (but not among species for a given locus), and also permits variation in population size among species (but not among loci for a given species). In short, the model assumes that the ratio of within-species variation to between-species variation will be the same for all of the loci. The resulting statistical test can take the form of a simple chi-square design,[6] or it can require more complicated strategies.[3]

The approach of using multiple genomic regions and multiple species lends itself to the study of speciation. Because the null model assumes that the relationship between intraspecific and interspecific variation is the same for all regions of the genome, the model can be used to test whether or not all loci are consistent in terms of divergence during speciation. In other words, these methods can be used to statistically test whether the pattern of variation among a set of loci from different species is consistent with a model in which the timing of the cessation of gene flow during species formation is the same for all of the loci.

A Case Study

The remainder of this review summarizes recent findings from a study of the four species of the *Drosophila melanogaster* species complex and of three X-linked loci[10,11]: approximately 1900 base pairs of the *period* locus (*per*), 1000 base pairs of the *zeste* locus, and 1100 base pairs of the *yolk protein* 2 locus (*yp*2). Two of the species, *D. melanogaster* and *D. simulans,* are cosmopolitan, while the other two, *D. mauritiana* and *D. sechellia,* are endemic to oceanic islands.[12] Numerous phylogenetic and species-hybridization studies have revealed only that *D. melanogaster* is a sister-taxon to the other species. *Drosophila simulans, D. sechellia* and *D. mauritiana* are morphologically very similar to each other and, despite considerable effort, a bifurcating phylogeny for these species has not been clearly resolved.[13–16] Collectively, these four species represent three speciation events, two of which appear to have occurred very recently and at similar times.

X-linked loci were chosen for the simple convenience that DNA preparations from individual male flies avoids heterozygosity in PCR-generated DNA sequences. Each species and locus was represented by six isofemale lines (*i.e.,* the laboratory strain was started with a single wild caught female) collected from multiple locations in the species range (excepting *D. mauritiana,* which is found only on the island of Mauritius). One DNA sequence was generated from each line for each locus. With the exception of five of the *zeste* gene sequences, all three genes were sequenced from the same DNA preparations, meaning that the three sequences for each strain actually came from the same chromosome. In the case of the *zeste* exceptions, the sequence still came from the same strains.

Figure 1 shows estimates of the amount of variation within each species for each locus. Figure 2 shows the divergence between *D. melanogaster* and *D. simulans*, per base pair, for each locus. Both *per* and *zeste* appear more variable than does *yp*2, and this pattern appears both in the intraspecific comparisons (Figure 1) and the interspecific comparisons (Figure 2). *Drosophila sechellia* has less variation than the other species, and this pattern is consistent among loci. Despite considerable variation among loci and among species, a casual appraisal suggests that this variation is consistent with the neutral model. The fit of the null model can be tested more formally using the procedure of Hudson, Kreitman and Aguadé[3] (the procedure is now widely used and frequently referred to as the HKA test). When this method was applied to the data, the overall fit of the model was quite good. The method also returns scalars of proportionality for population sizes and mutation rates. Relative to *D. melanogaster* (1.0), the population sizes for the other species are 1.600 for *D. simulans,* 1.296 for *D. mauritiana,* and 0.111 for *D. sechellia*. Similarly, for mutation rates, the scalars of proportionality relative to *zeste* are 2.372 for *per* and 0.722 for *yp2*. In sum, there is no evidence for recent directional or balancing selection; and the data are consistent with a neutral model in which population sizes vary across species, but not across loci, and in which neutral mutation rates vary across loci, but not species.

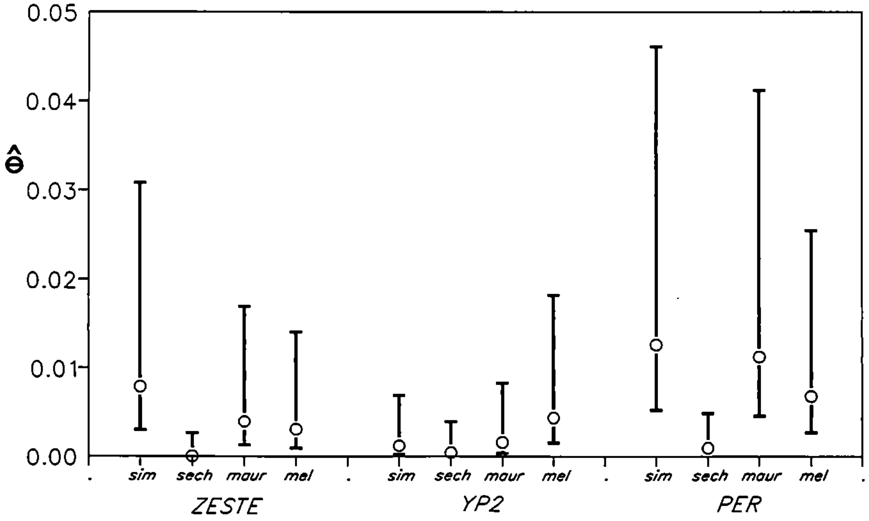

Figure 1. Estimates of $4Nu$ ($\hat{\theta}$) per base pair, where N is the effective population size of a species and u is the neutral mutation rate. $\hat{\theta}$, and the 95% confidence intervals were calculated from the number of polymorphic sites observed in each species and locus according to the procedure of Kreitman and Hudson.[27]

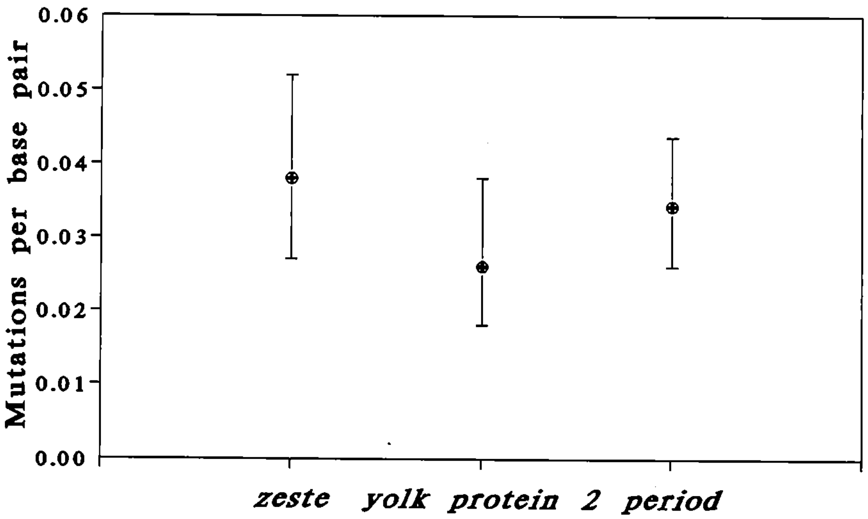

Figure 2. The average number of differences per base pair separating *D. melanogaster* sequences from *D. simulans* sequences. The 95% confidence intervals where generated assuming a Poisson distribution of differences.

Figure 3 shows tree diagrams generated with the neighbor-joining procedure.[17] The diagrams can be considered in two different ways. They are first, and most literally, the output of a particular clustering algorithm that joins sequences and groups of sequences by a complicated assessment of sequence similarity and tree length. Secondly, these diagrams may be considered as estimates of the historical genealogy of the gene copies in the sample. In this light, the tips of branches refer to the actual DNA sequences, and the branches refer to the persistence of ancestral DNA sequence lineages through time. The nodes of the tree refer precisely to DNA replication events in which both copies produced are ancestors of sequences included in the sample.

One way a branching tree diagram may not fit a genealogy is if the history includes recombination among the ancestral sequences of a sample.[18] In fact, the *per* locus data suggest a history with considerable recombination.[10] The patterns of variation, especially within *D. simulans* and *D. mauritiana,* suggest that a bifurcating diagram like that in Figure 3 is not a good model of the genealogy. Because of recombination, some parts of the *per* gene in these species have different genealogical histories than do other parts of the gene. Other evidence of recombination at *per* in these species is the observation that 11 polymorphisms are shared between them. Even though the neighbor-joining algorithm separated all the *D. mauritiana* sequences from the *D. simulans* se-

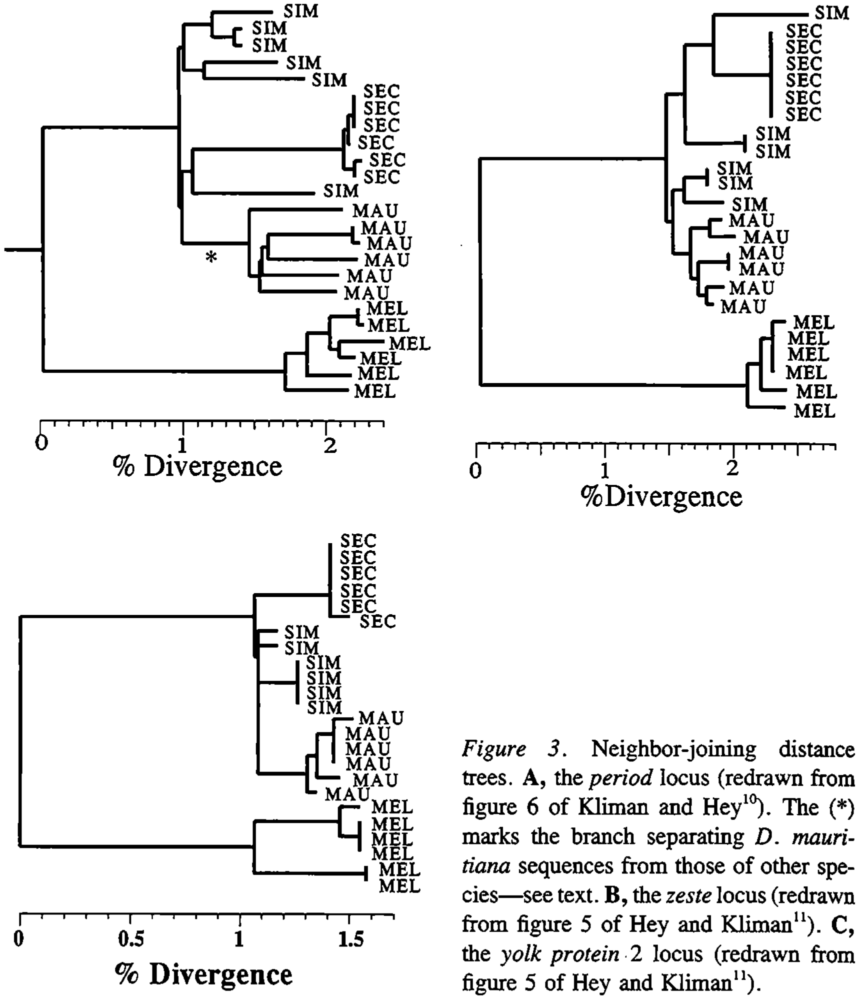

Figure 3. Neighbor-joining distance trees. **A,** the *period* locus (redrawn from figure 6 of Kliman and Hey[10]). The (*) marks the branch separating *D. mauritiana* sequences from those of other species—see text. **B,** the *zeste* locus (redrawn from figure 5 of Hey and Kliman[11]). **C,** the *yolk protein* 2 locus (redrawn from figure 5 of Hey and Kliman[11]).

quences, this part of the diagram cannot be interpreted as a genealogy. The branch of the tree marked with an asterisk (Figure 3A) cannot represent a historical DNA sequence, for the simple reason that polymorphisms cannot be carried by a single DNA sequence.

Recombination at *per* notwithstanding, the three trees share several characteristics: *D. melanogaster* lineages are clearly separated from those of the other species, consistent with all other phylogenetic studies on this group[12]; all of the lineages within *D. schellia* and *D. mauritiana* form discrete clusters; and the earliest nodes of the tree, excluding the split between *D. melanogaster* and the other species, separate lineages of *D. simulans*.

Considering Speciation

Because no evidence was found for recent balancing or directional selection at these loci, and since they have similar genealogical histories, the data can be interpreted in terms of evolutionary forces that are expected to affect all loci in the same way. In other words, with no evidence that locus-specific forces have been at work, the data are more useful for considering evolutionary forces that affect all loci in the same way.

The shapes of the gene trees show that present day *D. simulans* still has genetic variation that predates the origin of the island endemic species. All three of the gene trees show *D. simulans* lineages extending to nodes that occurred prior to the base nodes of the *D. mauritiana* and *D. sechellia* clusters. It appears that *D. mauritiana* and *D. schellia* arose independently from ancestral *D. simulans,* and that modern *D. simulans* has changed relatively little since that time. Since modern day *D. mauritiana* and *D. sechellia* are endemic to different oceanic islands, a plausible model for their formation is that they diverged from ancestral *D. simulans* following the isolation of a small number of individuals that successfully crossed the oceanic barrier. It has been suggested that this kind of "founder" event, involving only a small number of individuals, describes a possibly frequent scenario for speciation.[19–21] However, this does not appear to be the case for *D. mauritiana,* which has nearly as much variation as does *D. simulans,* indicating a large effective population size. Also, 11 of the *per* polymorphisms are shared between *D. mauritiana* and *D. simulans;* it is unlikely that such a number of presumably ancient polymorphisms would persist if *D. mauritiana* ever experienced a population bottleneck. Thus, a very small effective population size for *D. mauritiana* at any time in its history seems unlikely. The repeated observation of very little variation in *D. sechellia* indicates a small effective population size for this species, which is consistent with speciation via "founder" event. However, one can only say that *D. sechellia* has had a small population size recently, and it remains possible that the population size was larger in the past.

Linking a Gene Tree Approach with a Gene Mapping Approach

The *zeste, yp*2, and *per* data reveal a unique view of population sizes during and since species formation. However, by themselves, these data do not inform on the role of genetic variation in the process of speciation. Consider loci that were in some fashion affected by natural selection as a component of the speciation process. One can imagine a history of differential adaptation, with different functional alleles suited to different environments; alternatively, there may be underdominance at some loci whereby heterozygotes (*i.e.*, hybrids) are less fit. Regardless, there are a variety of speciation models in which, because of natural selection, some loci experience different patterns of gene flow between incipient species than others. If there is opportunity for some gene flow while reproductive isolation develops, those loci that are not included in the gene flow will have different gene trees than those that are included.

The *zeste, yp*2 and *per* data provide an excellent opportunity for testing whether specific loci have played an active role in species formation. If candidates for such "speciation genes" are found, and if sequence data is obtained from within and between the several species, then the contrasting patterns between loci may well inform on the historical role of such "speciation genes." In short *per, zeste* and *yp*2 may serve as reference loci.

One way to include loci associated with speciation is to map, at a very fine scale, those loci responsible for fitness loss in hybrids. Once identified, a locus can be examined for intraspecific and interspecific variation in a manner parallel to that for *zeste, yp*2 and *per*. Not all of these loci, and possibly none, have necessarily contributed to speciation, as hybrid fitness loss is expected to accumulate after speciation. However, for very recent speciation events, there is the chance that loci identified in this way may inform on speciation. Significant progress with this approach has recently been made using *D. simulans-D. mauritiana* hybrids and *D. simulans-D. sechellia* hybrids.[22,23] In particular, a small gene region, containing a putative locus named *Odysseus,* has been located that appears to contribute to reproductive isolation between *D. simulans* and *D. mauritiana*.[23] Interestingly, this gene does not seem to contribute to reproductive isolation between *D. simulans* and *D. sechellia,* consistent with independent evolution of reproductive isolation from *D. simulans* in the two island species.

In some cases, it may also be possible to map loci responsible for premating isolation, though quantitative analysis of behavioral traits can be difficult. In the *D. simulans* complex, at least two loci appear to contribute to premating isolation between *D. sechellia* females and *D. simulans* males, while at least three loci (including two on the second chromosome) contribute to sexual isolation between *D. mauritiana* females and *D. simulans* males.[24,25] It is unlikely that this represents a single, shared evolutionary process (*i.e.,* that the two island species formed from a single species that evolved premating isolation from *D. simulans*). First, *D. sechellia* and *D. mauritiana* are, themselves, reproductively isolated. Second, the quantitative effects on sexual isolation of loci linked to specific autosomes differs in the two species, as does the degree of dominance. Third, the behavioral bases to sexual isolation differ, with *D. simulans* males actively courting *D. mauritiana* females, while avoiding *D. sechellia* females.[24] It should be added that, although courtship between *D. mauritiana* and *D. simulans* is often successful, the duration of copulation is short, resulting in the actual reproductive isolation.[25] These results are consistent with independent speciation events for the two island endemics, as also indicated by the molecular data.

The *zeste, yp*2 and *per* studies have laid a baseline that can be used to test the role of other loci in speciation. As mapping studies proceed, and other methods[26] for identifying candidate loci for involvement in speciation are developed, we may find a wider diversity of genealogical histories among loci.

Acknowledgments

This work was supported by National Science Foundation grant BSR 8918164.

References

1. W. J. Ewens. (1990). in *Mathematical and Statistical Developments of Evolutionary Theory* (S. Lessard, ed.) pp. 177–227, Kluwer Academic Publishers.
2. R. R. Hudson. (1990). In *Oxford Surveys in Evolutionary Biology* (*Vol. 7*) (P. H. Harvey and L. Partridge, ed) pp. 1–44 Oxford University Press.
3. R. R. Hudson, M. Kreitman, and M. Aguadé. (1987). *Genetics* 116, 153–159.
4. A. J. Berry, J. W. Ajioka, and M. Kreitman. (1991). *Genetics* 129, 1111–1117.
5. D. Begun and C. F. Aquadro. (1991). *Genetics* 129, 1147–1158.
6. J. H. McDonald and M. Kreitman. (1991). *Nature* 351, 652–654.
7. W. Stephan and S. J. Mitchell. (1992). *Genetics* 132, 1039–1045.
8. C. Langley, J. M. MacDonald, N. Miyashita, and M. Aguadé. (1993). *Proc Nat. Acad. Sci. USA* 90, 1800–1803.
9. W. F. Eanes, M. Krichner, J. Yoon. (1993). *Proc. Natl. Acad. Sci. USA* 90, 7475–7479.
10. R. M. Kliman, and J. Hey. (1993). *Genetics* 133, 375–387.
11. J. Hey and R. M. Kliman. (1993). *Mol. Biol. Evol.* 10, 804–822.
12. D. Lachaise, M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. *Evolutionary Biology* 22, 159–225.
13. M. Bodmer and M. Ashburner. (1984). *Nature* 309, 425–430.
14. V. H. Cohn, M. A. Thompson, and G. P. Moore. (1984). *J. Mol. Evol.* 20, 31–37.
15. J. A. Coyne and M. Kreitman. (1986). *Evolution* 40, 673–691.
16. A. Caccone, G. D. Amato, and J. R. Powell. (1988). *Genetics* 118, 671–683.
17. N. Saitou and M. Nei. (1987). *Mol. Biol. Evol.* 4, 406–425.
18. R. R. Hudson and N. L. Kaplan. (1988). *Genetics* 120, 831–840.
19. E. Mayr. (1954). In *Evolution as a Process* (J. Huxley, C. Hardy, and E. B. Ford, eds.) pp. 157–180, Allen & Unwin.
20. H. L. Carson. (1975). *Am. Nat.* 109, 83–92.
21. A. R. Templeton. (1980). *Genetics* 94, 1011–1038.
22. C.-I. Wu, D. E. Perez, A. W. Davis, N. A. Johnson, E. L. Cabot, M. F. Palopolis,

and M.-L. Wu. (1993). In *Mechanisms of Molecular Evolution* (N. Takahata and A. G. Clark, ed.) pp. 191–212, Sinaur.

23. D. E. Perez, C.-I. Wu, N. A. Johnson, and M.-L. Wu. (1993). *Genetics* 134, 261–275.
24. J. A. Coyne. (1992). *Genet. Res., Cambr.* 60, 25–31.
25. J. A. Coyne. (1993). *Evolution* 47, 778–788.
26. L.-W. Zeng. and R. S. Singh. (1993). *Genetics* 135, 135–147.
27. M. Kreitman and R. R. Hudson. 1991. *Genetics* 127, 565–582.

18

Genetic Divergence, Reproductive Isolation and Speciation

Rama S. Singh and Ling-Wen Zeng

> *Discontinuities observed between species must have owed its origin to discontinuities occurring in the evolution of each. . . .*
>
> Bateson (1894)
>
> *Species and higher categories originate in single macroevolutionary steps as completely new genetic systems. The general process which is involved consists of a repatterning of the chromosomes, which results in new genetic system.*
>
> Goldschmidt (1940)

Introduction

Species are the most readily recognizable units in the diversity of life and the mechanism of speciation has always been and still is a central problem in evolutionary biology.[33] Almost all species with which we come in contact in daily life and certainly a large proportion of the formally described species show large gaps of qualitative or quantitative nature and this large gap has been a stumbling block to the study of speciation. This is for two reasons. First, presence of large gaps between species meant that species-specific traits could not be subjected to Mendelian genetic analysis, and second, the large gaps observed between species were used to propose theories of speciation which went against the neo-Darwinian mechanisms of gradual evolution (e.g., see Bateson 1894, Goldschmidt 1940). It is therefore not surprising that most genetic theories of speciation advocating macroevolutionary mechanisms of speciation, prior to the advent of molecular techniques in 1960s, were based on large changes in the genome such as chromosomal changes or macromutations.[23,59] On the other hand, the neo-Darwinian theories were based on a collection of genetic and ecological factors, with strong emphasis on natural selection and geographic isolation which was required to complete the job of reproductive isolation.[20,27,32,55]

The recognition that evolutionary processes of gradual evolution initially must produce species that are far from having complete reproductive isolation led to