THE ROYAL SOCIETY

Received 5 November 2002
Accepted 9 January 2003
Published online 30 April 2003

The causes of phylogenetic conflict in a classic *Drosophila* species group

Carlos A. Machado* **and Jody Hey**

Department of Genetics, Rutgers University, 604 Allison Road, Nelson Biological Laboratories, Piscataway, NJ 08854-8082, USA (hey@biology.rutgers.edu)

Bifurcating phylogenies are frequently used to describe the evolutionary history of groups of related species. However, simple bifurcating models may poorly represent the evolutionary history of species that have been exchanging genes. Here, we show that the history of three well-known closely related species, *Drosophila pseudoobscura, D. persimilis* and *D. p. bogotana*, is not well represented by a bifurcating phylogenetic tree. The phylogenetic relationships among these species vary widely between different genomic regions. Much of this phylogenetic variation can be explained by the potential of different genomic regions to introgress between species, as measured in laboratory studies. We argue that the utility of multiple markers in species-level phylogenetic studies can be greatly enhanced by knowledge of genomic location and, in the case of hybridizing species, by knowledge of the functional or linkage relationships among the markers and regions of the genome that reduce hybrid fitness.

Keywords: phylogeny; introgression; species; gene trees; multilocus; *Drosophila*

1. INTRODUCTION

The reconstruction of historical relationships between organisms is one of biology's most important endeavours. Obtaining a clear picture of the divergence of closely related species (i.e. species-level phylogeny) is a necessary step for understanding the basic process by which biological diversity is generated (i.e. speciation). However, historical relationships among closely related species are often difficult to reconstruct, even with DNA sequence data. One difficulty is simply that closely related species have low levels of genetic divergence. A more complex issue arises because it takes a long time for gene trees to become reciprocally monophyletic during species divergence. Closely related species often share genetic variation for extensive periods after divergence begins, which is the basis of the well-known 'gene tree versus species tree' problem (Tajima 1983; Pamilo & Nei 1988; Takahata 1989; Wu 1991; Hudson 1992; Maddison 1997; Hudson & Coyne 2002). Furthermore, the occurrence of introgressive hybridization (Anderson & Hubricht 1938), a process more common than previously acknowledged (Arnold 1997), violates the basic bifurcating model of species divergence on which phylogenetic analyses are based.

The complexities and ambiguities found in species-level phylogenetic studies are a simple short-term by-product of the normal processes that underlie the generation of biological diversity. The difficulties that arise from recent divergence, including lineage sorting of ancestral polymorphisms and introgressive hybridization, mean that data from any one locus, or linkage group, may not very well reveal phylogenetic history. Therefore, studies that rely upon one or a few sequences per species from a single locus are likely to result in findings that are either ambiguous or not very representative of the species under investigation. The use of multiple loci and multiple sequences per loci should in principle give a more complete picture of the history of divergence of a group of closely related species, because comparisons can be made across loci to ascertain whether all loci fit a simple model represented by the same bifurcating phylogeny. If the histories are not congruent, one can determine whether lineage sorting of ancestral polymorphisms or introgression are needed to explain the incongruencies.

The classic group of *Drosophila* species *D. pseudoobscura, D. p. bogotana* and *D. persimilis* constitute an interesting study case for such a multilocus phylogenetic approach. Because these species have played a central role in the development of evolutionary theory and in speciation studies (Dobzhansky 1937; Mayr 1942), they are particularly appropriate for in-depth phylogenetic analysis. A recent population genetics multilocus study of these species was conducted using multiple sequences per species from 14 loci scattered across different regions of the genome (Machado *et al.* 2002). The analyses focused on the question of gene flow between species pairs: *D. pseudoobscura* versus *D. p. bogotana* and *D. pseudoobscura* versus *D. persimilis*. A high level of variation across loci was found in patterns of shared polymorphisms and fixed differences between species, indicating the occurrence of gene flow at some loci but not at others between *D. pseudoobscura* and *D. persimilis*. The inference of gene introgression at some loci suggests potential problems with phylogenetic reconstruction and with the use of bifurcating phylogenies in this species group.

Here, we revisit the data of Machado *et al.* (2002), and take an explicit phylogenetic perspective to assess whether different genes or different genomic regions reveal different phylogenetic relationships among these three species of *Drosophila*, and to determine to what degree the history of the divergence of these species can be approximated by a bifurcating phylogeny. We also incorporate new data from two loci selected from non-recombining regions of

*Author and address for correspondence: Department of Ecology and Evolutionary Biology, University of Arizona, 1041 East Lowell Street, Biosciences West Building, Tucson, AZ 85721, USA (cmachado@email.arizona.edu).

the genome: from the dot (fifth) chromosome and from the mitochondria. For genes that undergo recombination and that reveal histories of recombination in their patterns of polymorphism, gene-tree estimates are problematic; thus, it is generally argued that non-recombining genes, or genes that reveal no evidence of recombination, are preferred for phylogenetic studies. Here, we also compare the phylogenetic patterns of non-recombining genes with those of genes that have been undergoing recombination.

(a) *The* Drosophila pseudoobscura *species group*

Drosophila pseudoobscura and *D. persimilis* played a key role in the development of the biological species concept (Dobzhansky 1937; Mayr 1942), and have been the frequent focus of speciation research (Dobzhansky 1936; Dobzhansky & Epling 1944; Orr 1987; Noor 1997; Noor *et al.* 2001*b*). The two species are partially sympatric in the western part of North America (from California to British Columbia) (Dobzhansky & Epling 1944); they hybridize at low frequency in nature (Dobzhansky 1973; Powell 1983) and their hybrid females are fertile, while hybrid males and some hybrid backcross females are sterile (Dobzhansky 1936). In 1963, an isolated population of *D. pseudoobscura* was found at high elevations near Bogotá, Colombia, separated by more than 2000 km from the North American population (Dobzhansky *et al.* 1963). This population, since named *D. pseudoobscura bogotana* (Ayala & Dobzhansky 1974), exhibits unidirectional hybrid male sterility with respect to *D. pseudoobscura* (Prakash 1972). Although *D. p. bogotana* has been classified as a subspecies of *D. pseudoobscura*, for the sake of brevity we refer to it as a species.

Estimates based on DNA sequences from two nuclear genes indicate that this group of species diverged less than 0.5 Myr ago (Schaeffer & Miller 1991; Wang *et al.* 1997). The traditional phylogeny of the species ((*pseudoobscura, bogotana*), *persimilis*) reflects several factors, including degrees of reproductive isolation (Dobzhansky & Epling 1944; Prakash 1972), levels of genetic divergence at random allozyme loci (Ayala & Powell 1972; Singh 1983) and the presence of fixed inversion differences on the X and second chromosomes between *D. persimilis* and the others (Dobzhansky & Epling 1944). Genes that cause sterility in male *D. pseudoobscura*/*D. persimilis* hybrids map to chromosomal inversions located on the X and second chromosomes (Orr 1987; Noor *et al.* 2001*b*). Introgression in the laboratory can occur across most of the autosomal chromosomes, including the polymorphic inversions of the third chromosome (Noor *et al.* 2001*a*,*b*).

2. MATERIAL AND METHODS

(a) *Data collection*

Nucleotide sequences were collected for 16 loci located on all five chromosomes and in the mitochondrial genome of these species. Each locus was sequenced in 10–20 inbred lines of each of the three species, as well as in one to four lines of an outgroup species, *D. miranda*. Seven loci are protein-coding regions (including exons and introns) (*period, Hsp82, rh1, bicoid, Adh, ey, mtDNA*) and nine are non-coding regions that flank microsatellites (*X008, X009, X010, 2001, 2002, 2003, 3002, 4002, 4003*) (Machado *et al.* 2002). The average sequence length per locus was 1242 nucleotides (table 1), and 43 lines were sequenced on average per locus. Information on collection sites, names of inbred lines and standard molecular methods used to collect the sequence data can be found elsewhere (Wang *et al.* 1997; Machado *et al.* 2002).

The sequences from 14 of the loci have been previously published (Wang *et al.* 1997; Machado *et al.* 2002). Sequences from the two new loci reported here (*ey* and *mtDNA*) have been deposited in GenBank (accession numbers AF451009–AF451152). The locus *ey* (*eyeless*) is located in the fifth chromosome of these species. A total of 1753 base pairs (bp) from the 3′-end of the large second intron of *ey* were amplified and sequenced using the primers FW1ey (5′-AAAATGCCAAATGCCTCT-3′), RV1ey (5′-TTCTGTCAGTTTCGCACTACAC-3′), FW4ey (5′-ACAAGAAAGGCTCTCGGATT-3′) and RV7ey (5′-AGTAAATGCATGGCATAGCTG-3′). The FW1ey primer is located in a conserved region in the middle of the intron. This region was detected by aligning the *ey* sequences of *D. melanogaster* (GenBank accession number AJ131630) and *D. virilis* (AF098329). RV1ey was designed using partial sequences of the 3′-end of the intron from *D. pseudoobscura* and *D. persimilis* (Noor *et al.* 2001*b*). FW4ey and RV7ey are internal primers used to reamplify shorter overlapping fragments.

The *mtDNA* data consist of sequences from two mitochondrial regions separated by about 4.3 Kb: the 3′-end of the COI gene (829 bp), and a region of 997 bp that includes the last 111 bases of *ND4*, the complete $tRNA_{His}$ and the first 820 bases of *ND5*. The 3′-end of COI was amplified and sequenced using primers Jerry (5′-CAACATTTATTTTGATTTTTTGG-3′) and Pat (5′-TCCAATGCACTAATCTGCCATATTA-3′) (Simon *et al.* 1994). The *ND4*–$tRNA_{His}$–*ND5* region was amplified and sequenced using primers ND4-1F (5′-AGCATGGTAAATTATTTTCTGG-3′) and ND5-3R (5′-TGTCTAAGAGTTGACAAAGCAA-3′). The sequences from *ND4*, $tRNA_{His}$, *ND5* and COI were concatenated into a single dataset, referred to as *mtDNA*.

(b) *Phylogenetic analyses*

We used MEGA v. 2.0 (Kumar *et al.* 2002) to reconstruct gene trees using the neighbour-joining (NJ) algorithm with Tamura–Nei distances (Tamura & Nei 1993). The observation of recombination in most of the loci included here (Wang *et al.* 1997; Machado *et al.* 2002) complicates interpretations of the phylogenetic history. Recombination has a large effect on estimated gene trees within species as it causes the different portions of a locus to have different histories. Further, it shuffles the variation among sequences within species, generating the appearance of large amounts of homoplasy, preventing precise assessments of homoplasy resulting from recurrent mutation and raising the concern that any observed variation among gene trees estimated for different loci could be a result of undetected recurrent mutation at individual loci. However, for the present purpose, these concerns are reduced for two main reasons. First, for the purpose of understanding species relationships, recombination may not be a difficulty simply because sequences from separate species will not have recombined with one another. By this argument, gene trees from recombining nuclear loci may still be useful for considering species phylogenetic relationships. Second, three considerations suggest a low rate of homoplasy by mutation in the data.

(i) The overall level of mutation accumulation is low because of recent common ancestry. Out of a total of 18 999 bases

Table 1. Patterns of monophyly at 16 loci.
(Shown are the bootstrap support values larger than 50% for monophyletic clades formed by sequences of each species or by sequences from two species. The *D. persimilis–D. p. bogotana* clade was never observed.)

			bootstrap value (> 50%)				
locus	length (bp)	chromosome	*pseudo*	*per*	*bog*	(*pseudo*, *bog*)	(*pseudo*, *per*)
period	1475	X			100	82	
X008	1026	X		100	79	76	
X009	700	X			63	63	
Hsp82	1953	X		97	85	66	
X010	877	X		99		99[a]	
2003	520	second					
rh1	1382	second			80		
bicoid	1371	second					
2002	918	second		99			
2001	685	second					
3002	614	third					
4003	623	fourth					
Adh	3451	fourth					
4002	825	fourth					
ey	1626	fifth			84		
mtDNA	1826	mtDNA			99		99

[a] No outgroup sequences available; therefore this value is the same as the value for *D. persimilis*.

sequenced in each of an average of 43 inbred lines, only 1481 positions (7.8%) were variable.

(ii) These loci have a high average GC content (46.7%), and the apparent transition–transversion ratio from counts of two-base polymorphisms (847/729 = 1.16) suggests that mutations tend to occur between all four base types at rates that are not greatly dissimilar.

(iii) Out of a total of 1481 variable positions (866 phylogenetically informative sites) only 76 (5.1%) revealed three or four base types. By comparison, under an infinite-alleles model and zero homoplasy by mutation, we would expect from the Poisson distribution that a region with 18 999 bases, of which 17 518 were invariant, would show 1422 sites with two alleles and 59 sites with more than two alleles.

The presence of intragenic recombination particularly affects phylogenetic inference using maximum parsimony (MP) or maximum likelihood (ML), because the apparent homoplasy generated by recombination greatly slows heuristic search algorithms by generating large numbers of equally optimal trees. To avoid these methodological problems we used the NJ method (Saitou & Nei 1987), which is faster, renders a single bifurcating tree and permits easily determined bootstrap-support values. Recombination also affects NJ by reducing intraspecific phylogenetic resolution but, as described above, this is less of a concern for understanding species relationships. Analyses of the two non-recombining loci with MP and ML produced almost identical trees to the ones estimated by NJ (not shown).

3. RESULTS

(a) *Phylogenetic relationships vary across loci and genomic location*

All loci had many phylogenetically informative sites (55 per locus on average, 44 on average excluding the outgroup) and probably have had little recurrent mutation (see § 2b), yet, even if we consider just those parts of the genealogies with strong bootstrap support, we find considerable variation among loci (figure 1). Table 1 lists, for each locus, the species and species pairs that are monophyletic in the gene-tree estimates. At nearly half of the loci the sequences from *D. p. bogotana* form a monophyletic group, while *D. persimilis* sequences are monophyletic at just four loci. No locus shows exclusive monophyly of *D. pseudoobscura* sequences. Further, *D. pseudoobscura* appears in different monophyletic groups with each of the other species, depending on the locus. Five loci show *D. pseudoobscura* forming a monophyletic group with *D. p. bogotana*, and one locus shows *D. pseudoobscura* in a monophyletic group with *D. persimilis*. The gene-tree estimates also reveal several cases where *D. pseudoobscura* and *D. persimilis* share DNA sequence variation, as revealed by nodes of common ancestry with descendant sequences in both species.

Genealogies of the five X-linked loci show strong support for the traditional phylogeny ((*pseudoobscura*, *bogotana*), *persimilis*) (figure 1*a–e*). The tree for locus *2002* (figure 1*i*), a locus found within the region spanned by the second chromosome fixed inversion, is also consistent with the traditional phylogeny, although this is suggested by the strong support for the monophyly of *D. persimilis* sequences and not by the monophyly of the *D. pseudoobscura–D. p. bogotana* clade. The eight additional loci located in recombining regions of the genome (chromosomes 2, 3 and 4) show one of two phylogenetic patterns: (i) no monophyly for any species or pair of species (*2003, bicoid, 2001, 3002, 4003, Adh, 4002*), or (ii) monophyly of *D. p. bogotana* sequences and paraphyly of *D. pseudoobscura* and *D. persimilis* sequences (*rh1*; figure 1*g*).

Finally, the genealogies of the two loci located in regions of no recombination (*ey*, *mtDNA*; figure 1*o*,*p*) show that *D. pseudoobscura* and *D. persimilis* share multiple haplotypes, suggesting that they have experienced considerable recent gene flow. The genealogies of both non-recombining loci are not compatible with the traditional phylogeny

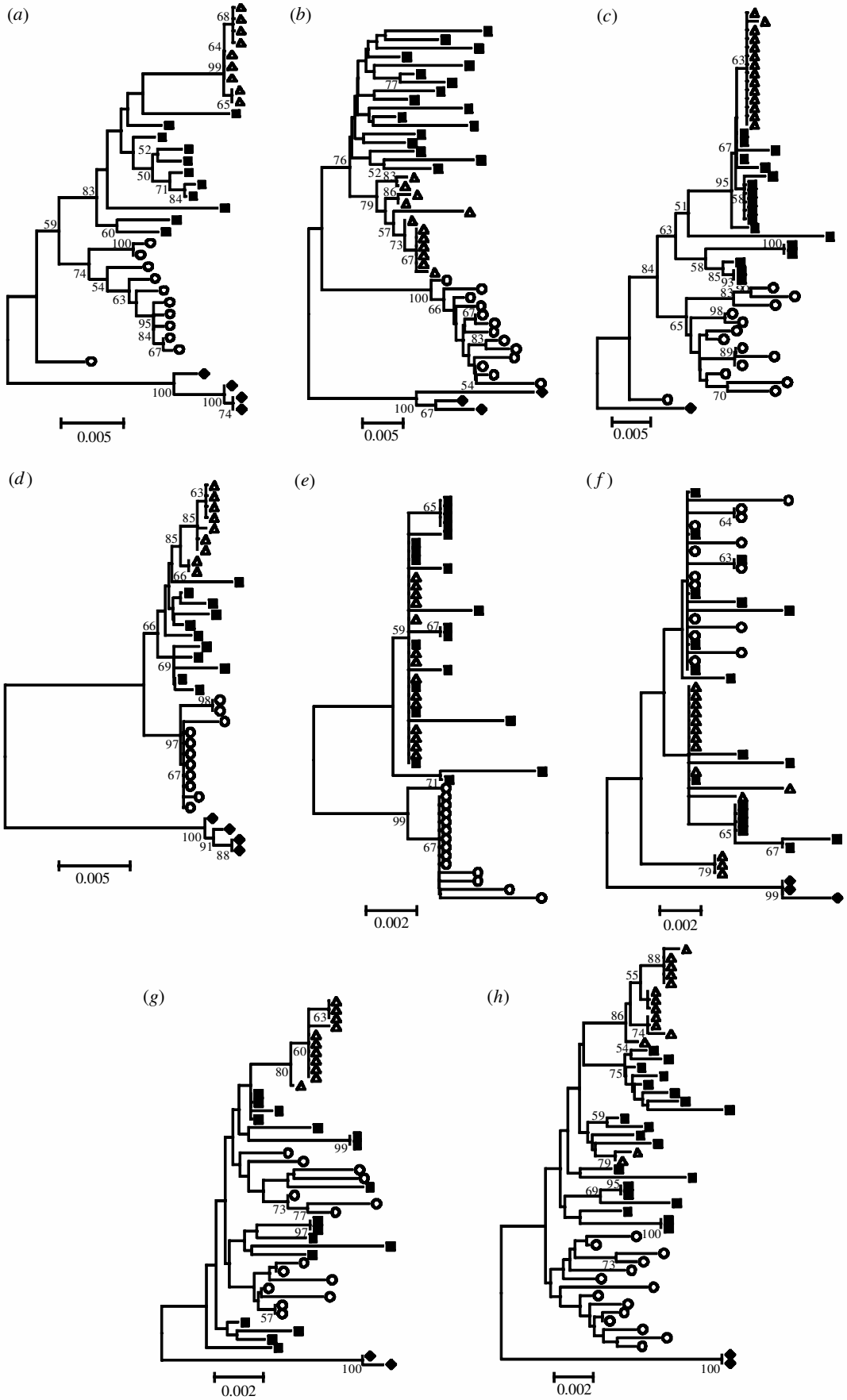

Figure 1. Gene genealogies from the 16 loci used in this study. Numbers on the branches are bootstrap support values based on 500 pseudoreplications. Bootstrap values lower than 50% are not shown. Scale bar numbers represent nucleotide divergence per base pair. Species codes: *D. pseudoobscura* (filled squares), *D. persimilis* (open circles), *D. p. bogotana* (open triangles), *D. miranda* (filled diamonds). Loci: (*a*) *period*, (*b*) *X008*, (*c*) *X009*, (*d*) *Hsp82*, (*e*) *X010*, (*f*) *2003*, (*g*) *rh1*, (*h*) *bicoid*, (*i*) *2002*, (*j*) *2001*, (*k*) *3002*, (*l*) *4003*, (*m*) *Adh*, (*n*) *4002*, (*o*) *ey* and (*p*) *mtDNA*.

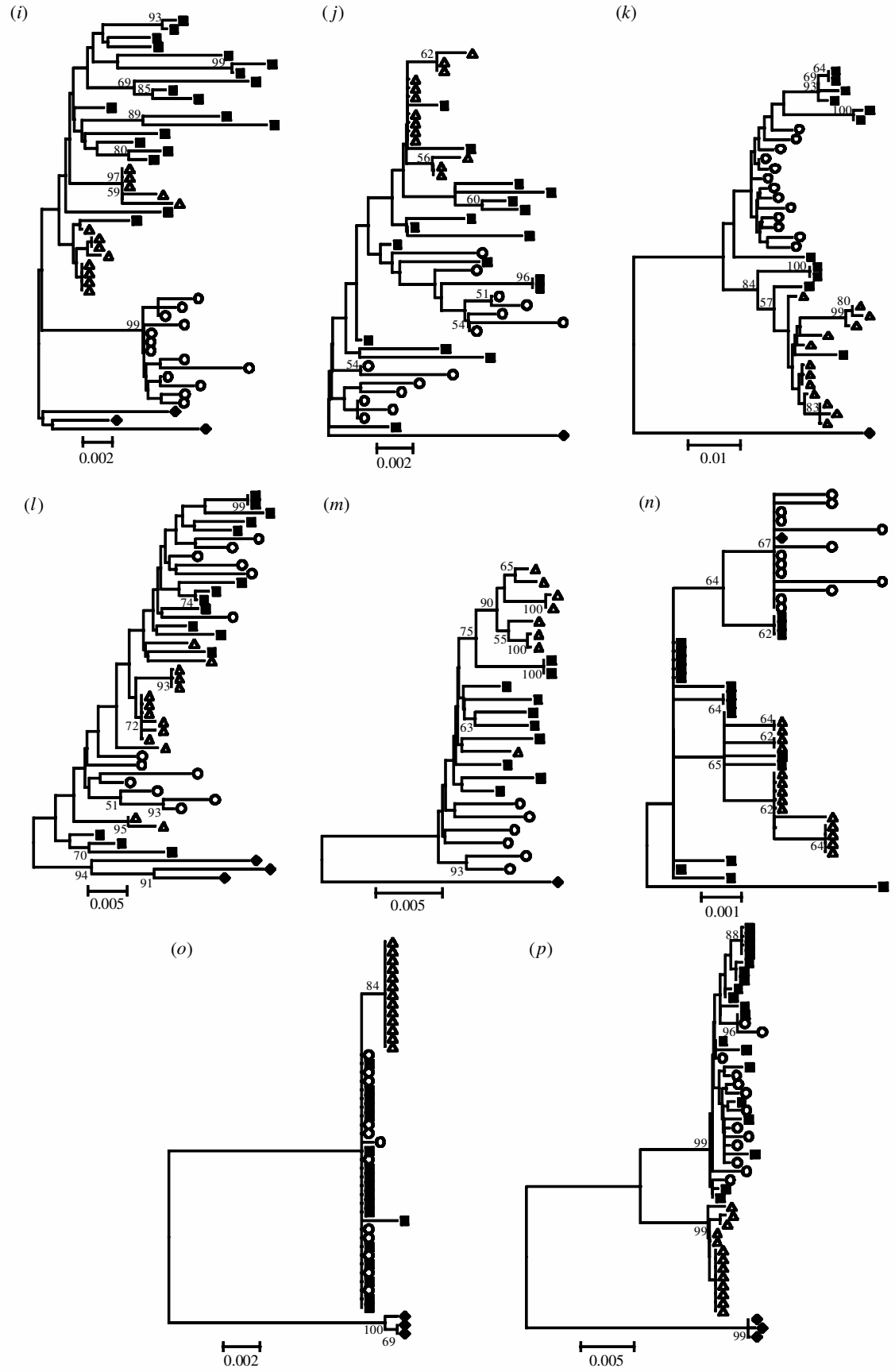

Figure 1. (*Continued.*)

of the species, because sequences from *D. p. bogotana* form a distinct monophyletic cluster that is well separated from the mixture of *D. pseudoobscura* and *D. persimilis* sequences.

The observed variation in phylogenetic patterns across loci can be readily summarized by comparing levels of interspecific sequence divergence across loci (figure 2). We defined a quantity d as the mean sequence divergence

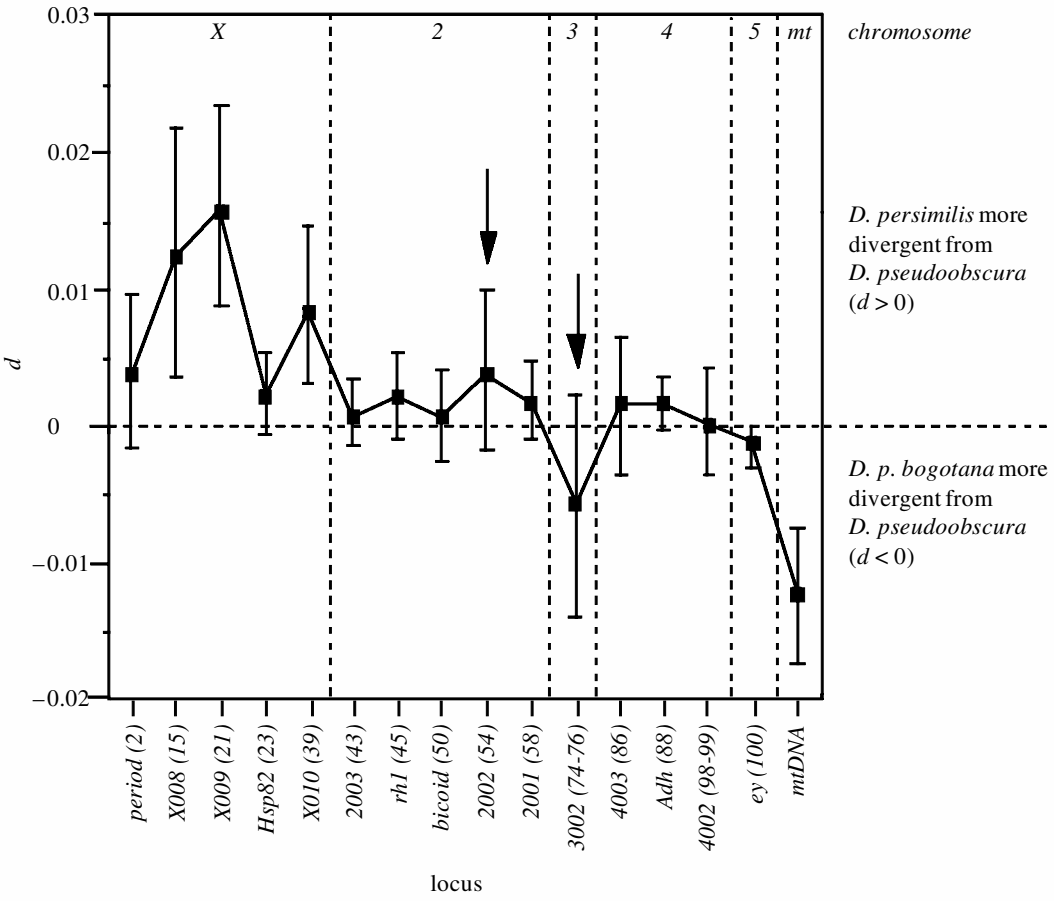

Figure 2. Differences in interspecific sequence divergence between *D. pseudoobscura*, *D. persimilis* and *D. p. bogotana*. Each point represents the value of absolute divergence d for the locus shown on the x-axis: $d = d_{ps/per} - d_{ps/bog}$, where $d_{ps/per}$ is the mean interspecific sequence divergence between *D. pseudoobscura* and *D. persimilis*, and $d_{ps/bog}$ is the mean interspecific sequence divergence between *D. pseudoobscura* and *D. p. bogotana*. Under a strict isolation model consistent with the traditional phylogeny the value of d is expected to be greater than zero for most loci. Chromosomal locations are shown for each locus, and loci are arranged in order based on their cytological-band locations (shown in parentheses) (Machado *et al.* 2002). Arrows indicate data from loci located in chromosomal inversions (*2002*, *3002*). Error bars correspond to 95% confidence intervals for d estimated with 10 000 bootstrap pseudoreplicates of the data for each locus.

between *D. pseudoobscura* and *D. persimilis* minus the mean sequence divergence between *D. pseudoobscura* and *D. p. bogotana*. For loci that have evolved according to the traditional phylogeny ((*pseudoobscura*, *bogotana*), *persimilis*) without gene flow, d should be positive. On the other hand, under a model in which *D. pseudoobscura* and *D. persimilis* have exchanged genes since the separation of *D. p. bogotana*, negative values of d are expected for those loci that have experienced gene flow. Indeed, the sign and magnitude of d varies greatly across different regions of the genome (figure 2), suggesting the occurrence of gene flow at some loci (see § 4a).

(b) *Phylogenies of non-recombining loci*

The phylogeny estimates from *ey* and *mtDNA* are not consistent with the traditional phylogeny (figure 1). At both loci the sequences from *D. pseudoobscura* and *D. persimilis* form an intermingled tree, much as if they had come from a single species—one that is separate from *D. p. bogotana*. Unlike the other loci, *ey* and *mtDNA* reveal complete shared haplotypes between *D. pseudoobscura* and *D. persimilis* (all but two haplotypes are identical in *ey*; there is a range of different *mtDNA* haplotypes of which two are shared). These shared haplotypes strongly suggest recent gene flow, which is consistent with the absence of linkage between these loci and the loci that contribute to hybrid sterility (Hutter & Rand 1995; Noor *et al.* 2001*b*). Levels of gene flow estimated from the *mtDNA* data are large ($Nm = 4.172$) (lack of variation in *ey* does not permit an estimate of Nm using an F_{ST}-based estimator) (Hudson *et al.* 1992). Neither locus exhibits fixed differences between *D. pseudoobscura* and *D. persimilis*, but both loci reveal fixed differences between these two species and *D. p. bogotana*. At *ey* there are two fixed differences between *D. p. bogotana* and both *D. pseudoobscura* and *D. persimilis*, and in *mtDNA D. p. bogotana* has 18 fixed differences with respect to *D. pseudoobscura* and 21 fixed differences with respect to *D. persimilis*.

The *ey* locus also differs from the other loci in having little variation within species (figure 1*o*; table 2) (Machado *et al.* 2002). Low variation at this locus is not the result of sequence conservation arising from selective constraints, because divergence from the outgroup *D. miranda* is similar to that observed at other loci (Machado *et al.* 2002). The observed reduction in variation is consistent with the action of natural selection (table 3), making *ey* the only locus out of the 16 for which the neutral model is rejected. Because *ey* is located in a region of no or very little recom-

Table 2. Summary of data for *ey* and *mtDNA*.
(Dashes indicate that values could not be obtained for small samples or for groups of sequences with few informative sites.)

locus	species	N[b]	L[c]	S[d]	$\hat{\theta}$[e]	π[f]	D[g]	Fu–Li D[h]	div.[i]
mtDNA	*pseudoobscura*	19	1826	30	0.004 70	0.003 25	–1.2260	–1.6644	0.0389
	bogotana	13	1826	7	0.001 24	0.000 93	–0.9473	–0.4763	0.0382
	persimilis	13	1826	27	0.004 86	0.003 31	–1.3280	–0.9346	0.0390
	miranda	3	1826	1	0.000 37	0.000 37	—	—	
ey	*pseudoobscura*	18	1607	3	0.000 54	0.000 21	–1.7130[a]	–2.5115[a]	0.0226
	bogotana	13	1609	0	0.000 00	0.000 00	—	—	0.0238
	persimilis	12	1607	2	0.000 41	0.000 21	–1.4514	–1.9122[a]	0.0226
	miranda	3	1680	2	0.000 79	0.000 79	—	—	

[a] Significant at $p < 0.05$.
[b] Number of lines sequenced.
[c] Average length (bp) of the sequences from each species.
[d] Number of polymorphic sites.
[e] Estimate of $2N_f\mu$ (*mtDNA*) (N_f is the effective number of females) or $4N\mu$ (*ey*) per base pair using the number of polymorphic sites (Watterson 1975).
[f] Estimate of $2N_f\mu$ (*mtDNA*) or $4N\mu$ (*ey*) using the average number of nucleotide differences per site (Nei 1987).
[g] Tajima's statistic (Tajima 1989). Significance was determined by 1000 coalescent simulations.
[h] Fu and Li's D statistic (Fu & Li 1993). Significance was determined by 1000 coalescent simulations.
[i] Average divergence (div.) per base pair between sequences from each taxon and the sequences of *D. miranda*.

Table 3. Results of HKA tests.
(*D. miranda* was the outgroup, and all autosomal loci were used (except *X010*, for which no outgroup sequence is available): χ^2 is the HKA test statistic (Hudson *et al.* 1987); and p is the probability of a value of χ^2 higher than observed, estimated with 1000 coalescent simulations. The value of the HKA statistic for *D. p. bogotana* without *ey* is close to significance owing to the pattern of polymorphism in the *period* locus (Wang & Hey 1996).)

	without *ey*		including *ey*	
species	χ^2	p	χ^2	p
pseudoobscura	4.472	0.816	16.976	0.049
bogotana	19.420	0.054	28.216	0.003
persimilis	11.184	0.311	21.101	0.025

bination (fifth chromosome), the reduction in variation within species could have been the result of background selection (Charlesworth *et al.* 1993) or a selective sweep (Maynard Smith & Haigh 1974) at or near the locus. However, the lack of variation between species is more consistent with a selective sweep across species boundaries. Similar patterns have been observed in other *Drosophila* species for loci in regions of low recombination and have been interpreted as 'trans-species' selective sweeps (Hilton *et al.* 1994; Stephan *et al.* 1998). Significantly negative values (table 2) of Fu and Li's D (Fu & Li 1993) in *D. pseudoobscura* and *D. persimilis* ($D = -2.5115$, $p = 0.005$, and $D = -1.9122$, $p = 0.040$, respectively), and of Tajima's D (Tajima 1989) in *D. pseudoobscura* ($D = -1.7130$, $p = 0.029$) also suggest that the pattern is consistent with a recent selective sweep, although other loci also show strongly negative values for these statistics (Machado *et al.* 2002). An alternative explanation for the pattern at this locus is the occurrence of background selection coupled with large levels of gene flow between species.

The *mtDNA* data are also unusual as regards the estimated time of divergence of *D. p. bogotana* from *D. pseudoobscura*. Based on sequences from *Adh* and *Hsp82*, the proposed time of divergence of the two species is 0.15–0.23 Myr (Schaeffer & Miller 1991; Wang *et al.* 1997), and the estimated divergence time from the outgroup *D. miranda*, using *Hsp82*, is 2.63 Myr (Wang *et al.* 1997). Simple calibrations of the *mtDNA* molecular clock using a standard 2% divergence per Myr (Brower 1994) show that, while the divergence from *D. miranda* (2.02 Myr) is roughly similar to the previous estimate, the estimated divergence time of *D. p. bogotana* from *D. pseudoobscura* (0.80 Myr) is four times greater than previous estimates.

4. DISCUSSION

(a) *The causes of phylogenetic variation among loci*

Three basic types of gene tree were observed after analyses of the 16 datasets:

(i) those that agree with the traditional phylogeny ((*pseudoobscura*, *bogotana*), *persimilis*; six loci),
(ii) those that are incongruent with the traditional phylogeny and support a different one (two loci), and
(iii) those in which there is a lack of resolution of the relationships among the *D. pseudoobscura* and *D. persimilis* sequences (eight loci).

Under the assumption that a single bifurcating phylogeny (the traditional phylogeny) correctly represents the history of divergence of these species, patterns that are not consistent with that phylogeny or that show a lack of reciprocal monophyly for *D. pseudoobscura* and *D. persimilis* can result from two non-exclusive processes: lineage sorting of ancestral polymorphisms and introgression.

Some of the observed phylogenetic variation among loci is an expected consequence of recent speciation events, because following divergence some loci will develop monophyletic patterns more quickly than others simply by chance (lineage sorting). Similarly, the absence of monophyly for *D. pseudoobscura* sequences and the frequent monophyly of *D. p. bogotana* sequences can be seen as a consequence of large and small population sizes, respectively (Machado *et al.* 2002). However, such arguments cannot explain the finding that *D. pseudoobscura* appears in different monophyletic groups with each of the other species, depending on the locus. Further, those arguments cannot explain why reciprocal monophyly for the sequences of *D. pseudoobscura* and *D. persimilis* is correlated with genomic location, and specifically why it is observed only in regions of the genome where hybrid-sterility loci have been mapped (i.e. only in X-linked and second inversion-linked loci) (Noor *et al.* 2001*b*) (see below). If that pattern resulted by chance, no obvious correlation with genomic location should have been observed. It can be argued that X-linked loci should show more phylogenetic resolution (i.e. more reciprocal monophyly) than autosomal loci because of the reduced effective population size for the X chromosome. However, X-linked loci revealed levels of variation that were virtually identical to those of the autosomal loci in these species: average values of $\hat{\theta}$ for autosomal and X-linked loci are 0.0148 and 0.0149, respectively, for *D. pseudoobscura,* and 0.0097 and 0.0090, respectively, for *D. persimilis* (Machado *et al.* 2002). Additionally, the phylogenies of the non-recombining genes can hardly be explained by lineage sorting, despite a previous suggestion based on a different mitochondrial dataset (Powell 1991).

Given that lineage sorting cannot explain all the patterns of phylogenetic variation, we argue that insights from another line of research, on the mapping of genes that cause reduced fitness in hybrids, do help identify an alternative cause for the different patterns observed across loci: that is, the capacity of loci to introgress between species. Recent studies of patterns of DNA sequence variation at multiple loci have provided evidence of gene flow between *D. pseudoobscura* and *D. persimilis* (Wang *et al.* 1997; Machado *et al.* 2002). Those analyses have revealed a high level of variation across loci in patterns of shared polymorphisms and fixed differences between species, indicating the occurrence of gene flow at some loci but not at others. Those observations suggest a divergence-with-gene-flow model of speciation (Maynard Smith 1966; Rice & Hostert 1993) for *D. pseudoobscura* and *D. persimilis*, under which natural selection acts to impede gene introgression at regions of the genome involved in the adaptive or reproductive divergence of the species. Other genes, not linked to such regions, will introgress more readily. Genes that cause sterility in male *D. pseudoobscura–D. persimilis* hybrids map to chromosomal inversions located on the X and second chromosomes (Orr 1987; Noor *et al.* 2001*b*), and introgression in the laboratory can occur across most of the autosomal chromosomes, including the polymorphic inversions of the third chromosome (Noor *et al.* 2001*a*,*b*) and the mitochondrial genome (Hutter & Rand 1995). Thus, it is expected that loci located in or linked to the X and second chromosome inversions should have experienced less gene flow and diverged more than loci located in other genomic regions. Further, gene-tree estimates for those loci are expected to be consistent with the traditional phylogeny, while gene-tree estimates for loci located in regions that can introgress are expected to be less consistent with that phylogeny (i.e. show lack of reciprocal monophyly) and in some cases, depending on the level and timing of gene flow, may even suggest an alternative species phylogeny ((*pseudoobscura, persimilis*), *bogotana*).

The observed phylogenetic patterns agree well with these expectations. The predicted lack of introgression between *D. pseudoobscura* and *D. persimilis* at loci linked to regions of the genome involved in hybrid sterility (X and second chromosome inversions) is consistent with the observation of reciprocal monophyly only at those loci. The remaining 10 loci, which are in regions of the genome that can be introgressed in the laboratory, either show lack of resolution for the monophyly of *D. pseudoobscura* and *D. persimilis* sequences, or strongly support a monophyletic *D. pseudoobscura–D. persimilis* clade. Those patterns are also consistent with predictions based on introgression capacity, as gene flow will reduce the rate of or stop differentiation at a locus, thus reducing the rate at which reciprocal monophyly could be attained. Out of those 10 loci, the two non-recombining ones (*ey* and *mtDNA*) show strong evidence of recent gene flow, as *D. pseudoobscura* and *D. persimilis* share several identical haplotypes, and their genealogical patterns stand in stark contrast to patterns from other loci that show considerable divergence between the species (e.g. X-linked loci). Unfortunately, it is difficult to obtain meaningful estimates of gene-flow levels for any of the remaining eight loci, owing to recombination. For those loci it is difficult to assess the precise cause of shared variation across species because, when recombination has occurred, sequences that have moved between species become shuffled with those that have not (Slatkin & Maddison 1989), thus making it hard to assess whether shared variation is the result of introgression or ancestral polymorphism. However, the rejection of a strict model of isolation for the 14 recombining loci based on the high variance in the number of shared polymorphisms, fixed differences across loci and a linkage disequilibrium test (Machado *et al.* 2002) suggest that gene flow is responsible for some of the shared variation at those loci.

Levels of interspecific sequence divergence across loci are also consistent with the expected potential of loci to introgress (figure 2). *Drosophila pseudoobscura* and *D. persimilis* are more differentiated at loci linked to regions of the genome that do not introgress, as shown by the high positive value of *d* for most of the X-linked loci and for the locus located in the fixed inversion of the second chromosome (*2002*). The values of *d* in regions of the autosomes not associated with reproductive isolation are either close to zero or negative, suggesting the presence of ancestral shared variation and/or introgression recently

or in the past, consistent with the observation of few fixed differences and many shared polymorphisms across species at those loci (Machado *et al.* 2002). Both species are more differentiated from *D. p. bogotana* at the two loci from the third and fifth chromosomes (*3002, ey*) and at *mtDNA*, which all show evidence of recent gene flow.

In conclusion, the variation in phylogenetic patterns among loci is consistent with the predicted capacity of the loci to introgress between *D. pseudoobscura* and *D. persimilis*, based on their genomic location. Although some of the observed phylogenetic variation is likely to be the result of lineage sorting, that process alone cannot explain all the variation among loci. For instance, it cannot explain patterns in the non-recombining loci or the observed correlation of reciprocal monophyly with genomic location, observations that are fully consistent with the predicted capacity of loci to introgress depending on genomic location. Although recent introgression is inferred from the two non-recombining loci, the lack of shared full haplotypes at any of the recombining loci that can introgress in the laboratory suggests that introgression at those loci has not occurred recently.

(b) *The use of non-recombining loci for reconstructing phylogenies*

Because of the analytical difficulties presented by recombining loci, most phylogenetic analyses of DNA sequence variation within and between closely related species have focused on non-recombining molecules, particularly in the mitochondrial genome (see Avise (2000) for references). A major reason for that choice is that a bifurcating gene tree constructed from non-recombining sequences is more meaningful than one constructed from a recombining locus, because in the latter any given pair of sequences has more than one common ancestor. A major drawback, however, is that patterns of variation in non-recombining loci are greatly dependent on selective forces acting at linked sites. Selective sweeps and background selection can obliterate any sequence variation present in non-recombining loci, thus decreasing the utility of inferences about demographic or phylogenetic history.

Our data from *ey* and *mtDNA* are clear examples of these difficulties. The *ey* data are unique among the 16 loci in showing evidence of a reduction in polymorphism caused by natural selection. The *ey* data also suggest a genealogical history with recent introgression between *D. pseudoobscura* and *D. persimilis* and with the occurrence of selective sweeps across the species barrier. The *mtDNA* data also suggest introgression, but are particularly striking because the large divergence of *D. p. bogotana* sequences is difficult to explain under any of the current models of divergence of these species. The data suggest either the presence of an ancestral population structure in the ancestor of these species or an old balanced polymorphism that has been resolved with separate alleles to separate species. Uninformed use of the *mtDNA* or *ey* data without information from other loci would have led us to infer a history of divergence that greatly differs from that suggested by the data from other genomic regions and that would simply correspond to a piece in the puzzle of an already complex history. Incongruencies between mtDNA and nuclear genealogies caused by introgression have also been recently observed in other groups of insects (Sota & Vogler 2001; Shaw 2002). These combined results illustrate the dangers of using data from a single locus, and from non-recombining loci, to infer the demographic or evolutionary history of recently diverged species.

5. CONCLUSIONS

The genomes of these species are mosaics, with different regions varying greatly in their genealogical histories and divergences across species. Whether or not *D. pseudoobscura* and *D. persimilis* first began to diverge under allopatry remains an open question; however, we have no evidence that they were ever completely separate species or that they were ever parapatric or allopatric. Notwithstanding their ongoing role as a classic model system for the study of speciation, these species join a growing list of species complexes that appear to have exchanged genes at some loci and not others (see Hey (2001, pp. 100–101) for references). Collectively, these findings offer a cautionary tale for the phylogenetic study of recent speciation events. A traditional phylogenetic approach that assumes a simple bifurcating history will necessarily be inaccurate for many of the genes of these species. Our findings highlight a pitfall that can arise either by using just a small number of loci, or by combining data from different loci, to reconstruct the history of closely related species. For cases like the present one, these approaches will lead to misrepresentations of the history of divergence by imposing a simple bifurcating model upon species that have undergone complex divergence.

We thank K. Shallop for his assistance in the laboratory. Research supported by NIH grant GM58060.

REFERENCES

Anderson, E. & Hubricht, L. 1938 The evidence for introgressive hybridization. *Am. J. Bot.* **25**, 396–402.

Arnold, M. L. 1997 *Natural hybridization and evolution*. New York: Oxford University Press.

Avise, J. C. 2000 *Phylogeography*. Cambridge, MA: Harvard University Press.

Ayala, F. J. & Dobzhansky, T. 1974 A new subspecies of *Drosophila pseudoobscura* (Diptera: Drosophilidae). *Pan-Pacific Entomol.* **50**, 211–219.

Ayala, F. J. & Powell, J. R. 1972 Allozymes as diagnostic characters of sibling species in Drosophila. *Proc. Natl Acad. Sci. USA* **69**, 1094–1096.

Brower, A. V. 1994 Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proc. Natl Acad. Sci. USA* **91**, 6491–6495.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Dobzhansky, T. 1936 Studies of hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* **21**, 113–135.

Dobzhansky, T. 1937 *Genetics and the origin of species*. New York: Columbia University Press.

Dobzhansky, T. 1973 Is there gene exchange between *Drosophila pseudoobscura* and *Drosophila persimilis* in their natural habitats? *Am. Nat.* **107**, 312–314.

Dobzhansky, T. & Epling, T. 1944 *Contributions to the genetics, taxonomy, and ecology of* Drosophila pseudoobscura *and its relatives*. Washington, DC: Carnegie Institute of Washington.

Dobzhansky, T., Hunter, A. S., Pavlovsky, O., Spassky, B. & Wallace, B. 1963 Genetics of an isolated marginal population of *Drosophila pseudoobscura*. *Genetics* **48**, 91–103.

Fu, Y. X. & Li, W. H. 1993 Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709.

Hey, J. 2001 *Genes, categories and species*. New York: Oxford University Press.

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

Hudson, R. R. 1992 Gene trees, species trees and the segregation of ancestral alleles. *Genetics* **131**, 509–512.

Hudson, R. R. & Coyne, J. A. 2002 Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557–1565.

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

Hudson, R. R., Slatkin, M. & Maddison, W. P. 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589.

Hutter, C. M. & Rand, D. M. 1995 Competition between mitochondrial haplotypes in distinct nuclear genetic environments: *Drosophila pseudoobscura* vs. *D. persimilis*. *Genetics* **140**, 537–548.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. 2002 MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Machado, C. A., Kliman, R. M., Markert, J. A. & Hey, J. 2002 Inferring the history of speciation using multilocus sequence data: the case of *Drosophila pseudoobscura* and its close relatives. *Mol. Biol. Evol.* **19**, 472–488.

Maddison, W. P. 1997 Gene trees in species trees. *Syst. Biol.* **46**, 523–536.

Maynard Smith, J. 1966 Sympatric speciation. *Am. Nat.* **100**, 637–650.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.

Mayr, E. 1942 *Systematics and the origin of species*. New York: Columbia University Press.

Nei, M. 1987 *Molecular evolutionary genetics*. New York: Columbia University Press.

Noor, M. A. F. 1997 Genetics of sexual isolation and courtship dysfunction in male hybrids of *Drosophila pseudoobscura* and *D. persimilis*. *Evolution* **51**, 809–815.

Noor, M. A., Grams, K. L., Bertucci, L. A. & Reiland, J. 2001*a* Chromosomal inversions and the reproductive isolation of species. *Proc. Natl Acad. Sci. USA* **98**, 12 084–12 088.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., Almendarez, Y., Reiland, J. & Smith, K. R. 2001*b* The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution* **55**, 512–521.

Orr, H. A. 1987 Genetics of male and female sterility in hybrids of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **116**, 555–563.

Pamilo, P. & Nei, M. 1988 Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583.

Powell, J. R. 1983 Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from *Drosophila*. *Proc. Natl Acad. Sci. USA* **80**, 492–495.

Powell, J. R. 1991 Monophyly/paraphyly/polyphyly and gene/species trees: an example from *Drosophila*. *Mol. Biol. Evol.* **8**, 892–896.

Prakash, S. 1972 Origin of reproductive isolation in the absence of apparent genetic differentiation in a geographic isolate of *Drosophila pseudoobscura*. *Genetics* **72**, 143–155.

Rice, W. R. & Hostert, E. E. 1993 Laboratory experiments on speciation: what have we learned in forty years? *Evolution* **47**, 1637–1653.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for constructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Schaeffer, S. W. & Miller, E. L. 1991 Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proc. Natl Acad. Sci. USA* **88**, 6097–6101.

Shaw, K. L. 2002 Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proc. Natl Acad. Sci. USA* **99**, 16 122–16 127.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. & Flook, P. 1994 Evolution, weighting and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann. Entomol. Soc. Am.* **87**, 651–701.

Singh, R. S. 1983 Genetic differentiation for allozyme and fitness characters between mainland and Bogota populations of *Drosophila pseudoobscura*. *Can. J. Genet. Cytol.* **25**, 590–604.

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

Sota, T. & Vogler, A. P. 2001 Incongruence of mitochondrial and nuclear gene trees in the Carabid beetles *Ohomopterus*. *Syst. Biol.* **50**, 39–59.

Stephan, W., Xing, L., Kirby, D. A. & Braverman, J. M. 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. *Proc. Natl Acad. Sci. USA* **95**, 5649–5654.

Tajima, F. 1983 Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105**, 437–460.

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

Takahata, N. 1989 Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966.

Tamura, K. & Nei, M. 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526.

Wang, R. L. & Hey, J. 1996 The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the period locus. *Genetics* **144**, 1113–1126.

Wang, R. L., Wakeley, J. & 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. Popul. Biol.* **7**, 256–276.

Wu, C. I. 1991 Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* **127**, 429–435.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.