

# DISCORDANT DIVERGENCE TIMES AMONG Z-CHROMOSOME REGIONS BETWEEN TWO ECOLOGICALLY DISTINCT SWALLOWTAIL BUTTERFLY SPECIES

Andrea S. Putnam,<sup>1,2</sup> J. Mark Scriber,<sup>3</sup> and Peter Andolfatto<sup>1</sup>

<sup>1</sup>Section of Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093

<sup>2</sup>E-mail: [asputnam@biomail.ucsd.edu](mailto:asputnam@biomail.ucsd.edu)

<sup>3</sup>Department of Entomology, Michigan State University, East Lansing, Michigan 48824

Received July 27, 2006

Accepted December 2, 2006

We investigate multilocus patterns of differentiation between parental populations of two swallowtail butterfly species that differ at a number of ecologically important sex-linked traits. Using a new coalescent-based approach, we show that there is significant heterogeneity in estimated divergence times among five Z-linked markers, rejecting a purely allopatric speciation model. We infer that the Z chromosome is a mosaic of regions that differ in the extent of historical gene flow, potentially due to isolating barriers that prevent the introgression of species-specific traits that result in hybrid incompatibilities. Surprisingly, a candidate region for a strong barrier to introgression, *Ldh*, does not show a significantly deeper divergence time than other markers on the Z chromosome. Our approach can be used to test alternative models of speciation and can potentially assign chronological order to the appearance of factors contributing to reproductive isolation between species.

**KEY WORDS:** Approximate Bayesian computation, divergence, hybrid zones, introgression, Lepidoptera, *Papilio*, speciation.

The understanding of speciation—the evolution of barriers to gene flow between taxa—is central to our current understanding of evolutionary biology (Dobzhansky 1937; Mayr 1942; Coyne and Orr 2004). Evolutionary geneticists are particularly interested in understanding the genetic basis of speciation, namely, how many and which genes are involved, what types of changes to these genes contribute to reproductive isolation, and what population genetic processes led to the fixation of different alleles at these genes (Orr et al. 2004). Historically, the study of the genetic basis of speciation has been hampered by the fact that alleles causing reproductive isolation are not particularly amenable to genetic analysis.

Despite this formidable obstacle, evolutionary geneticists have recently made progress in identifying these barriers to gene

flow using two types of approaches. The first is a series of clever genetic mapping experiments designed to pinpoint genomic regions, and in some cases individual genes, causing reproductive isolation (Wittbrodt et al. 1989; True et al. 1996; Ting et al. 1998; Barbash et al. 2003; Presgraves 2003; Presgraves et al. 2003; Tao et al. 2003; Sawamura et al. 2004; Moehring et al. 2006; Turner et al. 2005). Presgraves (2003) estimated that between *Drosophila melanogaster* and its closest known relative, *D. simulans*, intrinsic hybrid inviability alone involves incompatible alleles at almost 200 genes. Hybrid male sterility factors have also been shown to disproportionately accumulate on the X chromosome (True et al. 1996; Tao et al. 2003), as predicted if the alleles causing them are partly recessive and positively selected (Charlesworth et al. 1987). In addition, several individual genes causing reproductive

isolation have been shown to be targets of recurrent adaptive amino acid substitution (Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003). Interestingly, these latter two observations suggest a link, if only indirect, between adaptive evolution and the evolution of reproductive isolation.

A second approach is based on statistical analysis of hybrid zones between parapatric, incompletely isolated species. The principle of this approach is that hybrid zones, in which hybrids are less fit than parental populations, represent a conflict between selection against unfit hybrids promoting species divergence and gene flow through dispersal preventing divergence (Slatkin 1973; Endler 1977; Mallet and Barton 1989; Harrison 1990; Barton 2001). The degree to which a genomic region can introgress across a hybrid zone can be related to the strength of selection against it in hybrids relative to dispersal (Barton 2001). In this way, genomic regions causing incompatibilities in hybrids can be mapped as those that have higher levels of differentiation between species (Hagen and Scriber 1989; Rieseberg et al. 1999; Payseur and Nachman 2005; Grahame et al. 2006). Though indirect, this approach is amenable to a wide range of species and presumably has the potential to map a broader range of factors contributing to reproductive isolation. Similar to more direct genetic mapping approaches, statistical analyses of hybrid zones suggest that a large number of loci contribute to reproductive isolation (Barton and Gale 1993). For example, Rieseberg et al. (1999) showed that of 26 genome segments showing significantly reduced introgression across a sunflower hybrid zone, 16 were associated with pollen sterility. These results demonstrate the use of hybrid zones in elucidating the genetic architecture of reproductive barriers between species.

Although both approaches above have proven useful, they suffer from caveats that limit how informative they are about the genetics of speciation. One concern is that hybrid zones are probably not stable over long periods of time. Thus, there may be a large historical component to patterns of differentiation between species, complicating estimates of the strength of selection across a cline. The second is that genes currently contributing to reproductive isolation may not have been involved in the initial speciation process (Coyne and Orr 2004). Whereas the evolution of reproduction isolation is a gradual process, the number of loci required to confer almost complete reproductive isolation between species may be small. One example is Presgraves' (2003) estimate that about 200 genes contribute to hybrid inviability alone in *D. melanogaster*/*D. simulans* hybrids. This implies that the total number of loci contribution to reproductive isolation between species (including, among other things, hybrid sterility, ecological differences, premating isolation, etc.) is likely to be much larger. Orr (1995) showed that a rapid accumulation of incompatibility factors (the "snowball effect") is expected after reproductive isolation is complete between two populations. These theoretical con-

siderations imply that a randomly chosen gene that is currently involved in reproductive isolation between completely isolated species is unlikely to have participated in the speciation process itself. The problem thus becomes trying to distinguish between true "speciation" genes from genic incompatibilities that secondarily strengthen reproductive isolation.

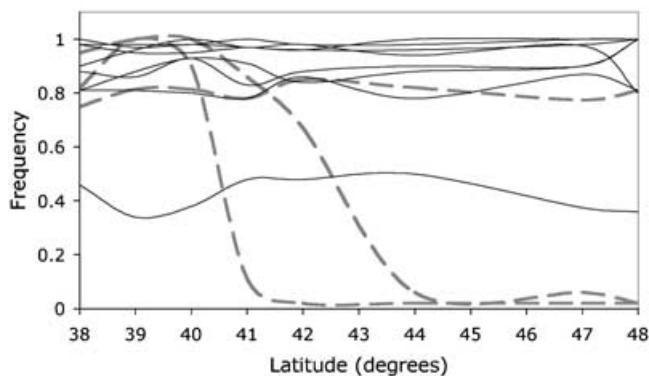
A third, complementary approach is based on coalescent theory. A coalescent approach considers the genealogical properties of samples from parental populations, and can be used to infer population genetic parameters under explicit models of speciation. The simplest of these models considers an allopatric speciation model with no gene flow (Hudson et al. 1987). In the presence of recombination, migration and selection can produce greater heterogeneity in divergence patterns across the genome than expected under a purely allopatric speciation model (Hudson et al. 1987; Palopoli et al. 1996; Wakeley and Hey 1997; Wang et al. 1997; Wu 2001; Machado et al. 2002; Hey and Nielsen 2004; Hey 2005; Bachtrog et al. 2006; Bull et al. 2006). In particular, those parts of the genome that move more freely between species (neutral markers) are expected to diverge more slowly than regions tightly linked to a gene causing reproductive isolation.

By partitioning the genome based on estimated population genetic parameters, such as ancestral and current population size, divergence time and migration rate, we can begin to ask which regions of the genome began to diverge first and/or have the lowest historical migration rates. These regions are more likely to be tightly linked to genes initially causing reproductive isolation rather than neutral parts of the genome or parts of the genome that only recently became associated with reproductive isolation. In addition, we can use estimated population genetic parameters to test explicit models of speciation. Although allopatry is believed to be the dominant mechanism of speciation (Mayr 1942), discordant gene genealogies and divergence time estimates among closely related *Drosophila* species (i.e., *D. pseudoobscura* and relatives; Wang et al. 1997; Machado et al. 2002) and between humans and chimps (Osada and Wu 2005) have rejected models of strict allopatric speciation.

Several Lepidopteran species reveal differential patterns of introgression at multiple loci among ecologically distinct strains or closely related species using various approaches (Lushai et al. 2003; Emelianov et al. 2004; Prowell, et al. 2004; Dopman et al. 2005; Bull et al. 2006; Kronforst et al. 2006). For example, a comparison of three Z-chromosome markers in strains of European corn borer reveals that at one marker haplotypes are not shared (Dopman et al. 2005). This marker is tightly linked to a factor that differentially affects postdiapause developmental time and may contribute to reproductive isolation between strains. Here we develop a new coalescent-based approach and apply it to parental populations of two hybridizing, parapatric species

of Lepidoptera. *Papilio glaucus* and *P. canadensis* are partially reproductively isolated swallowtail butterfly species that form hybrids in a narrow hybrid zone. These species are differentiated by diapause regulation, female-limited mimicry, host-plant preferences, morphological characters, and at least two loci contributing to hybrid inviability (Hagen and Scriber 1989; Hagen et al. 1991; Scriber et al. 1991). Previous surveys of allozymes and mitochondrial DNA (mtDNA) haplotypes revealed a remarkable pattern of differentiation between these two species (Hagen and Scriber 1989; Hagen 1990; Sperling 1993; Bossart and Scriber 1995). In particular, of 21 autosomal allozymes surveyed, most were polymorphic but showed little differentiation between species suggesting high levels of gene flow (Hagen and Scriber 1989). In contrast, mtDNA and three allozymes, including two on the Z chromosome (the Lepidopteran analog of the X in the XY male/XX female system), exhibit strong patterns of differentiation between species, consistent with selection against these markers in hybrids. The two Z-linked allozymes (*Pgd* and *Ldh*) are only loosely linked to each other and show distinct patterns of differentiation across the hybrid zone (Hagen 1990; Fig. 1). These patterns strongly suggest that the genomes of these species are a mosaic of regions that experience differential selection pressures in hybrids, and thus they may also show heterogeneous patterns of differentiation.

We examine patterns of divergence between samples of the parental species for five distinct regions of the Z chromosome and the mtDNA (*COI/COII*). One of the Z-linked regions, *Ldh*, is an allozyme locus that shows particularly strong differentiation between species in transects through the hybrid zone and is thus a candidate for tight linkage to a gene causing reproductive isolation. MtDNA haplotypes also show strong differentiation in transects through the hybrid zone, which may be a consequence of its expected linkage to the W chromosome in



**Figure 1.** Frequencies of three Z-linked, *Ldh*, *Pgd*, and *Acp* (grey dashed lines) and eight autosomal (black lines) allozymes across the *P. glaucus*/*P. canadensis* hybrid zone that corresponds to 41–43°. The y-axis plots the frequency of *P. glaucus*-like allozyme variants. Data replotted from Hagen (1990).

Lepidoptera (Andolfatto et al. 2003). This is interesting because female-limited mimicry in *P. glaucus* (a trait that distinguishes species) is partly determined by a W-linked locus (Clarke and Sheppard 1962; Scriber et al. 1996). Here we implement a novel approximate Bayesian approach to estimating speciation time that extends previous approaches (Hudson et al. 1987; Wakeley and Hey 1997; Bachtrog et al. 2006). We use these divergence time estimates to test the strictly allopatric model of speciation, which predicts that each genomic region began to diverge at the same time. Under a model of continuing migration, selection against hybrids, and recombination, we may expect to reject the purely allopatric model. We also relax the strictly allopatric model and estimate locus-specific divergence times. In particular, we expect that our candidate regions, *Ldh* and the mtDNA, should yield deeper divergence time estimates than randomly selected markers. We test this prediction and discuss the implications of our results for mapping speciation genes.

## Materials and Methods

### DNA EXTRACTION AND SEQUENCING

*Papilio glaucus* and *P. canadensis* were collected from a broad geographic sample of their ranges (see online Supplementary Material, Table S1) and identified by hind-wing size and bandwidth. Genomic DNA was isolated using a modified Puregene (Gentra Systems, Minneapolis, MN) protocol. Two legs and the thorax of frozen individuals were ground in 5  $\mu$ L of 20 mg/mL proteinase K. The homogenate was incubated at 55°C for 12 h, followed by a 2-min incubation at 95°C. The standard Purgene protocol for *Drosophila* DNA purification was followed from this point onward using three times the suggested solution volumes.

A  $\lambda$ FIXII genomic DNA library of *P. glaucus* (P. Andolfatto, unpubl. data) was screened for *Lactose Dehydrogenase* (*Ldh*) and *Kettin* (*Ket*) using *D. melanogaster* derived probes and standard library screening protocols (Sambrook and Russell 2001). Positive clones were isolated, sequenced, and intron/exon boundaries were mapped using rapid amplification of cDNA-PCR from total RNA extractions. Primers for *Per* were designed using degenerate primers reported in Regier et al. (1998). Degenerate primers for *Tpi* were designed using multispecies protein sequence alignments (Logsdon et al. 1995). New primers used in this study are listed in Supplementary Material, Table S2. Primers for *COI/COII* were previously reported in Andolfatto et al. (2003). PCR conditions in a thermocycler (Bio-Rad, Hercules, CA) included an initial denaturing step at 95°C for 2 min followed by 40 cycles of 95°C for 30 sec, 52°C for 45 sec, 72°C for 2 min with a final 5 min at 75°C.

The PCR product clean-up was performed using Exo/SAP reagents (Fermentas, Hanover, MD). Templates were directly

sequenced on both strands using primers listed in the Supplementary Materials (Table S2) and the BigDye sequencing kit (ver. 3.1, Roche, Nutley, NJ). Sequence reactions were run on an ABI 3730 sequencer (Applied Biosystems, Foster City, CA). Sequencing each gene in a panel of male and female *P. glaucus* confirmed Z-linkage, as chromatograms showed heterozygous sites in all males, and never in females (data not shown). Nucleotide sequences were edited using Sequencher 4.1 software (Gene Codes, Ann Arbor, MI), aligned using ClustalX (Thompson et al. 1997) and manually adjusted (GenBank accession EF115340–EF115363, and EF126370–EF126497).

### LEVELS OF POLYMORPHISM, DIVERGENCE, AND RECOMBINATION

To quantify and characterize polymorphism and divergence at these Z-linked loci, we considered all silent sites (all synonymous and noncoding sites) in exons and introns. Sites overlapping insertions and/or deletions were excluded. The tRNA separating *COI* and *COII* in the mtDNA was excluded, as were GT/AG splice sites associated with exon/intron boundaries in the case of nuclear genes. Synonymous and nonsynonymous sites were characterized using DnaSP version 3 (Rozas and Rozas 1999).

Levels of intragenic recombination are an important population genetic parameter, particularly in the context of testing population genetic models (Hudson 1983; Wang et al. 1997; Przeworski et al. 2001). We thus jointly estimated the population mutation rate ( $\theta = 4N_e\mu$ , where  $N_e$  is the effective population size and  $\mu$  is the mutation rate per base pair) and levels of intragenic recombination,  $\rho (= 4N_e r$ , where  $r$  is the recombination rate per base pair) in *P. glaucus* and *P. canadensis* using an approximate Bayesian method with rejection sampling (Haddrill et al. 2005; Thornton, unpubl. data). Posterior distributions are based on 5000 acceptances of  $\rho$  and  $\theta$  and were estimated for each individual locus and jointly over all loci. Wide uniform priors were chosen for  $\rho$  (0, 0.9) and  $\theta$  (0.003, 0.02). A fixed tolerance ( $\epsilon$ ) was set to 0.1 for both parameters to reduce computational demand. The analysis took three weeks on two G5 processors.

Levels of silent nucleotide variability within species were summarized using Watterson's estimator,  $\theta_w$  (Watterson 1975), and the average pairwise diversity per nucleotide,  $\pi$  (Tajima 1983). Divergence ( $D_{XY}$ ) between *P. glaucus* and *P. canadensis* was estimated as the average pairwise number of nucleotide substitutions per site between species (Nei 1987). The linkage relationships of our Z-linked markers relative to each other are not known, but throughout this paper we assume that they are sufficiently loosely linked such that they can be treated independently. We performed an exact test for linkage disequilibrium between Z-linked loci as implemented in Arlequin version 2.0 (Schneider et al. 2000). No significant levels of linkage disequilibrium

among loci were detected in either *P. glaucus* or *P. canadensis* ( $P > 0.05$ ).

### NEUTRALITY TESTS

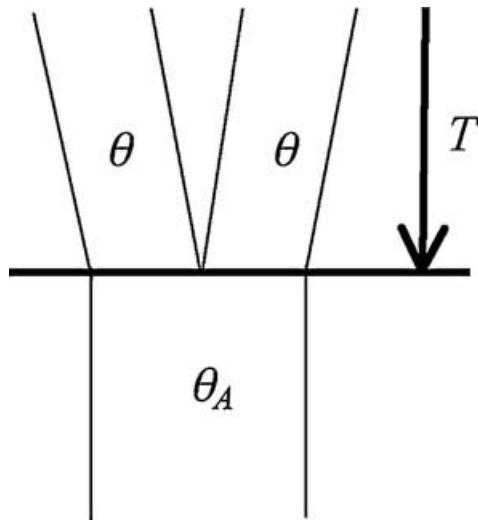
We employ a series of common tests of the neutral equilibrium model in two contexts. First, as the methods employed to estimate divergence times assume neutrality, we used these tests to assess whether there are signs of direct or linked selection acting on the loci used. In a second context, we ask whether our candidates for linkage to hybrid incompatibility loci in the hybrid zone, *Ldh* and the mtDNA, show any signs of recent or ongoing selection, and look any different than other loci we surveyed. In particular, we used two summaries of the distribution of polymorphism frequencies: Tajima's  $D$  (Tajima 1989), a measure of the standardized difference between  $\pi$  and  $\theta_w$ , and Fay and Wu's  $H$  (Fay and Wu 2000), which measures the difference between  $\pi$  and  $\theta_H$ , a summary of  $\theta$  that weights derived variants by the square of their frequencies. For Fay and Wu's  $H$ , we used sequences from *P. rutulus* and *P. multicaudatus* to infer the ancestral state of each nucleotide using standard parsimony criteria with a correction for multiple hits. Under the standard neutral model, both tests are expected to give values close to zero. We also performed three multilocus tests of neutrality as implemented by Haddrill et al. (2005). These tests compared the average Tajima's  $D$  and Fay and Wu's  $H$  across loci to simulated distributions, and the Hudson, Kreitman, Aguadé (HKA) test, which compares levels of polymorphism ( $\theta_w$ ) to levels of interspecific divergence ( $D_{XY}$ ) across loci. Analysis of nucleotide variation and tests of neutrality were implemented using programs available at [www.biology.ucsd.edu/labs/andolfatto/programs.html](http://www.biology.ucsd.edu/labs/andolfatto/programs.html).

### ESTIMATING DIVERGENCE TIMES

Divergence time ( $T$ ) is estimated using a novel approximate Bayesian inference developed from previous methods that use the HKA test (Hudson et al. 1987) as a framework (Wakeley and Hey 1997; Bachtrog et al. 2006). Here we develop a Bayesian extension of the likelihood method of Bachtrog et al. (2006), that incorporates information on the number of shared and fixed polymorphisms. We assume a model of simple allopatric speciation where an ancestral population of size  $\theta_A$  splits into two equal-sized populations of size  $\theta$  with no gene flow (Fig. 2), and  $\theta$  and  $\theta_A$  remain constant over time. For each locus,  $j$ ,  $\theta$  is estimated as

$$\theta_j = S_j / \sum_{i=1}^{n-1} \frac{1}{i}, \quad (1)$$

where  $n$  is the sample size and  $S$  is the number of segregating sites. Using the same  $\theta$  for *P. glaucus* and *P. canadensis* is justified because their estimates of  $\theta$  are not significantly different (results



**Figure 2.** Model of allopatric species divergence. An ancestral population of size  $\theta_A$  splits into two species of size  $\theta$  at time  $T$  with no migration.

not shown); however, our method can be modified to accommodate different population sizes in the two species (see Bachtrog et al. 2006).

For each locus we perform the following set of steps:

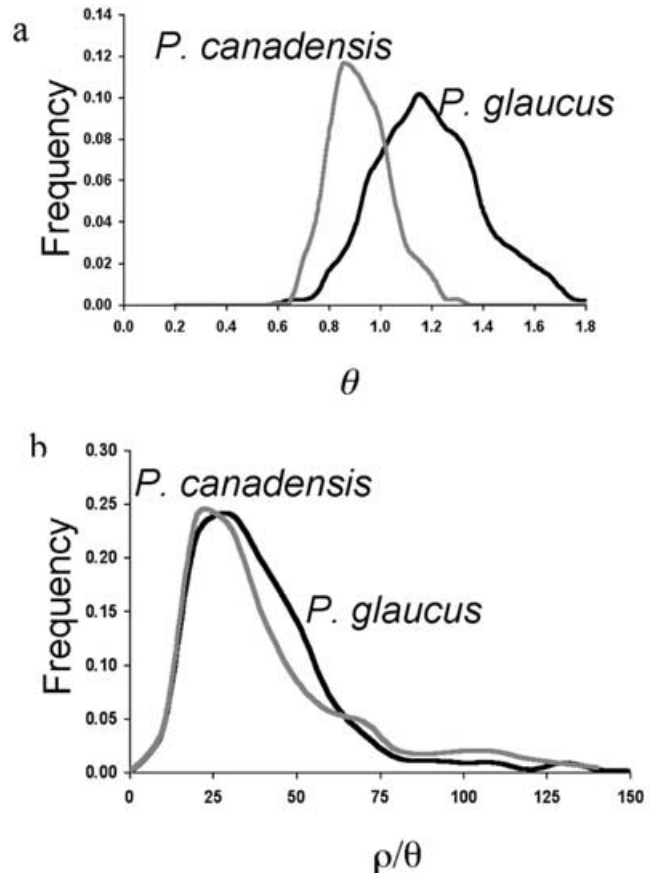
We summarize the observed (obs) data by the sample sizes, locus length in base pairs (representing the total alignment length), levels of variability estimated from the number of segregating sites within a species ( $\theta_j$ ), the number of shared polymorphisms ( $S_{j,obs}$ ), the number of fixed differences between species ( $F_{j,obs}$ ), and the divergence time estimated as

$$T_{j,obs} = (D_{XY,j}/\theta_j) - 1, \quad (2)$$

where  $\theta_j$  is the average of  $\theta$  estimates for *P. glaucus* and *P. canadensis* (Hudson et al. 1987).

Using *ms* (Hudson 2002), simulate the neutral coalescence with recombination of samples drawn from two subdivided populations with no gene flow that diverged at time,  $T$ . This coalescence method is based on the standard Fisher-Wright neutral model, which assumes a large, panmictic population and mutations occurring according to the infinite sites model (Hudson 1983). An infinite sites model is a reasonable assumption in our case (because levels of variability and divergence are low), but may not be appropriate for highly diverged species or genomes with high mutation rates (e.g., some viruses). In these simulations, we use the point estimate of  $\theta_j$  for  $\theta$  and assume  $\rho = 33.5\theta$ , which represents the joint maximum a posteriori (MAP) estimate of  $\rho$  in *P. glaucus* (Fig. 3). The simulated divergence time,  $T$ , is the only free parameter. We use an uninformative (i.e., uniform) prior for  $T$  that is sampled from the interval  $0-16 N_e$  generations.

We summarize the simulated data (sim) in the same way we



**Figure 3.** Approximate Bayesian joint posterior distribution of (a)  $\theta$  and (b)  $\rho/\theta$  for the five Z-linked markers in *Papilio glaucus* (black) and *P. canadensis* (grey).

summarized the observed data. We accept the simulated value of  $T$  if  $S_{j,sim} = S_{j,obs}$ ,  $F_{j,sim} = F_{j,obs}$ , and  $|T_{j,sim} - T_{j,obs}| < \delta$ , where  $\delta$  is a fixed tolerance. A drawback of rejection sampling is that the tolerance parameter affects the efficiency of inference and as a result, acceptance rates may be prohibitively low if a very stringent  $\delta$  is used (Beaumont et al. 2002). Owing to computational constraints,  $\delta$  was set to 0.05, and the average acceptance rate was  $\sim 10^{-3}$  per locus. The simulations took approximately three weeks on four G5 processors. Using a lower tolerance (0.001) had little effect on the posterior distribution (results not shown). To calculate a joint, multilocus estimates of  $T$ , the same priors and tolerance must be used for each locus.

We repeated steps 2 and 3 until 2000 draws of the posterior distribution were collected.

The method produces posterior distributions of  $T$  for each locus. We summarized the posterior distributions and obtained the MAP estimate and 95% confidence interval as implemented in the “locfit” statistical package (Loader 2006) in the library for *R*. To obtain a joint Z-linked MAP estimate of  $T$ , posterior distributions for each locus were binned in increments of  $0.2 N_e$

and probabilities for each bin were multiplied across loci. Likelihood ratio tests were used to test whether assuming unique divergence times for each of the Z-linked markers fit the data significantly better than one (i.e., strictly allopatric) divergence time. We assume the likelihood ratio statistic is chi-squared distributed with degrees of freedom equal to the difference in the number of free parameters. To examine the effect of the assumed ancestral population size ( $\theta_A$ ) on divergence time estimates, we carried out simulations with an ancestral population size that was 1, 2, 5, 8, and 10 times the current population size ( $\theta$ ). In this implementation, we assume that the change in population size is instantaneous, though this assumption can be relaxed (see the *ms* program documentation).

We tested the performance of our method on 100 datasets simulated under the parameters we estimated from the data (i.e., sample size, and our estimates of  $\theta$ ,  $\rho$  and  $T$ ), and assuming that  $\theta_A = \theta$ . The method works well under parameters that closely match our data, with a bias of  $0.2 N_e$  and root mean square error of  $4.8 N_e$  (Supplementary Material, Fig. S1). A library of scripts and programs to implement this procedure, called STE (Speciation Time Estimator), is available from the website [www.biology.ucsd.edu/labs/andolfatto/programs.html](http://www.biology.ucsd.edu/labs/andolfatto/programs.html).

We compared our divergence time estimates among Z-linked markers to results from an alternative method (*WH*) developed by Wakeley and Hey (1997). This program uses the number of exclusive and shared polymorphisms, and the number of fixed differences in the observed data to estimate the population sizes of *P. glaucus* and *P. canadensis* ( $\theta_1$  and  $\theta_2$ , respectively), the size of the ancestral population ( $\theta_A$ ), and the divergence time estimate ( $T$ ) for each locus and jointly across all Z-linked loci. Point estimates of these parameters were then used in neutral coalescent simulations with recombination to test the fit of the data to a strict allopatric speciation model (Wang et al. 1997; Kliman et al. 2000). Whereas both approaches are multilocus methods that use coalescence with recombination to test the fit of observed data to an allopatric speciation model, *WH* differs from our method in that it has an additional free parameter ( $\theta_A$ ), estimates  $\theta$  differently, and is a moment-based method.

#### EVALUATING MODELS USING THE POLYMORPHISM FREQUENCY SPECTRUM

We used two summaries of the frequency spectrum (Tajima's  $D$  and Fay and Wu's  $H$ ) to evaluate the fit of parameters estimated both under our method and the *WH* method to the observed data. We summarized the observed data as the average  $D$  and  $H$  across the five Z-linked loci. For each locus, we used *ms* to simulate 10,000 neutral genealogies with the following parameters:  $\theta_j$ , the joint Z-linked mode for recombination ( $\rho = 33\theta$ ), and  $T$ , drawn from the posterior distribution of  $T$  obtained from our approx-

imate Bayesian analysis and varied  $\theta_A$  to be  $1\times$ ,  $5\times$ ,  $8\times$ , and  $10\times$  the current population size  $\theta$ . For each simulated replicate, we recorded the average Tajima's  $D$  and average Fay and Wu's  $H$  across loci and compared these distributions to the observed averages. Similarly, we evaluated estimates from the *WH* approach by simulating data using point estimates of  $\theta_1$ ,  $\theta_2$ ,  $\theta_A$ ,  $T$  and assumed that  $\rho = 33\theta$ .

## Results

### LEVELS OF DIVERSITY AND RECOMBINATION

Levels of silent variability in *P. glaucus* and *P. canadensis* ( $\sim 1\%$ , see Table 1 and Supplementary Material, Fig. S2) are comparable to *Drosophila* (Moriyama and Powell 1997) and other Lepidoptera surveyed so far (Beltran et al. 2002; Dopman et al. 2005). *Papilio canadensis* has slightly lower levels of variability (0.9%, 95% CI 0.7–1.2) on average than *P. glaucus* (1.1%, 95% CI 0.8–1.7); however, three of the five Z-linked markers are actually more variable in *P. canadensis*. Although suggestive of a smaller effective population size in *P. canadensis*, the lack of a systematic trend implies that we cannot reject the hypothesis that the two species have equal effective population sizes. Levels of diversity on the Z chromosome and mtDNA are similar, but levels of divergence are three-fold higher on the mtDNA (average  $D_{XY}$  is 2.3% for Z-linked and 6.3% for mtDNA). The smaller ratio of polymorphism to divergence for the mtDNA relative to nuclear genes is expected in a neutral equilibrium population with equal numbers of males and females (Kimura 1983; Birky et al. 1989). The higher levels of silent divergence on the mtDNA relative to nuclear genes is typical of arthropods (Moriyama and Powell 1997).

The population recombination rate,  $\rho$ , is inversely proportional to levels of intragenic linkage disequilibrium. We estimate  $\rho$  per site to be lower (but not significantly so) in *P. canadensis* (mode 0.15, 95% CI 0.06–1.5) than in *P. glaucus* (mode 0.35, 95% CI 0.11–1.0). This difference may in part be attributed to a smaller population size, as reflected by lower levels of diversity on average in *P. canadensis* (Fig. 2A). The ratio of the recombination rate,  $\rho$ , relative to the mutation rate,  $\theta$ , however, is expected to be similar in two equilibrium populations of different size (Hudson et al. 1987; Andolfatto and Przeworski 2000). Interestingly, the mode of  $\rho/\theta$  in *P. glaucus* (33.5, 95% CI 9.0–90.7) that is very close to that for *P. canadensis* (33.0, 95% CI 9.1–110.2; see Fig. 2B). This estimate is considerably higher than recent estimates from *D. melanogaster* ( $\rho/\theta = 10$ ; Thornton and Andolfatto 2006), which is consistent with *Papilio* having at least a three-fold longer genetic map (Ashburner 1989; Winter and Porter, pers. comm.) but roughly only two times more genomic DNA (Celniker and Rubin 2003; Gregory and Herbert 2003).

Per locus posterior distributions of  $\rho/\theta$  are broad (Supplementary Material, Fig. S3) and MAP estimates of  $\rho/\theta$  appear to

**Table 1.** Polymorphism and divergence statistics for *Papilio glaucus* and *P. canadensis*

Gene	Species	Sample size	Total length	Silent sites	S <sup>1</sup>	$\pi^2$ (%)	$\theta^3$ (%)	D <sub>XY</sub> <sup>4</sup> (%)	Sh <sup>5</sup>	F <sup>6</sup>	E(ss) <sup>7</sup>
Kettin	<i>P. glaucus</i>	12	1206	251	4	.4	.5	3.8	0	7	.08
	<i>P. canadensis</i>	10			5	.6	.7				
<i>Ldh</i>	<i>P. glaucus</i>	11	439	317	16	1.9	1.7	2.9	0	3	.25
	<i>P. Canadensis</i>	9			5	.7	.6				
<i>Period</i>	<i>P. glaucus</i>	12	212	162	3	.5	.6	2.5	0	3	.56
	<i>P. canadensis</i>	7			3	.7	.8				
<i>Titin</i>	<i>P. glaucus</i>	12	611	410	13	1.2	1.1	1.7	10	1	.41
	<i>P. canadensis</i>	8			13	1.4	1.2				
<i>Tpi</i>	<i>P. glaucus</i>	12	296	211	15	2.4	2.4	4.4	3	0	.64
	<i>P. canadensis</i>	10			9	1.2	1.3				
Z-linked	<i>P. glaucus</i>	12 <sup>8</sup>	2764	1351	51	1.3	1.3	2.8 <sup>8</sup>	13	14	1.4
	<i>P. canadensis</i>	9 <sup>8</sup>			36	.9	.9				
<i>COI/COII</i>	<i>P. glaucus</i>	29	2289	494	24	.7	1.3	6.3	1	19	.49
	<i>P. canadensis</i>	13			10	.3	.6				

<sup>1</sup>Total number of polymorphisms observed.

<sup>2</sup>Average pairwise diversity per site.

<sup>3</sup>Estimate of  $\theta = 3N_e\mu$  ( $N_e\mu$  for mtDNA) per site using the number of polymorphic sites.

<sup>4</sup>Average pairwise divergence per silent site.

<sup>5</sup>The number of shared polymorphisms at silent sites.

<sup>6</sup>The number of fixed differences at silent sites.

<sup>7</sup>The expected number of shared mutations from recurrent mutation (Clark 1997; Kliman et al 2000).

<sup>8</sup>Averages.

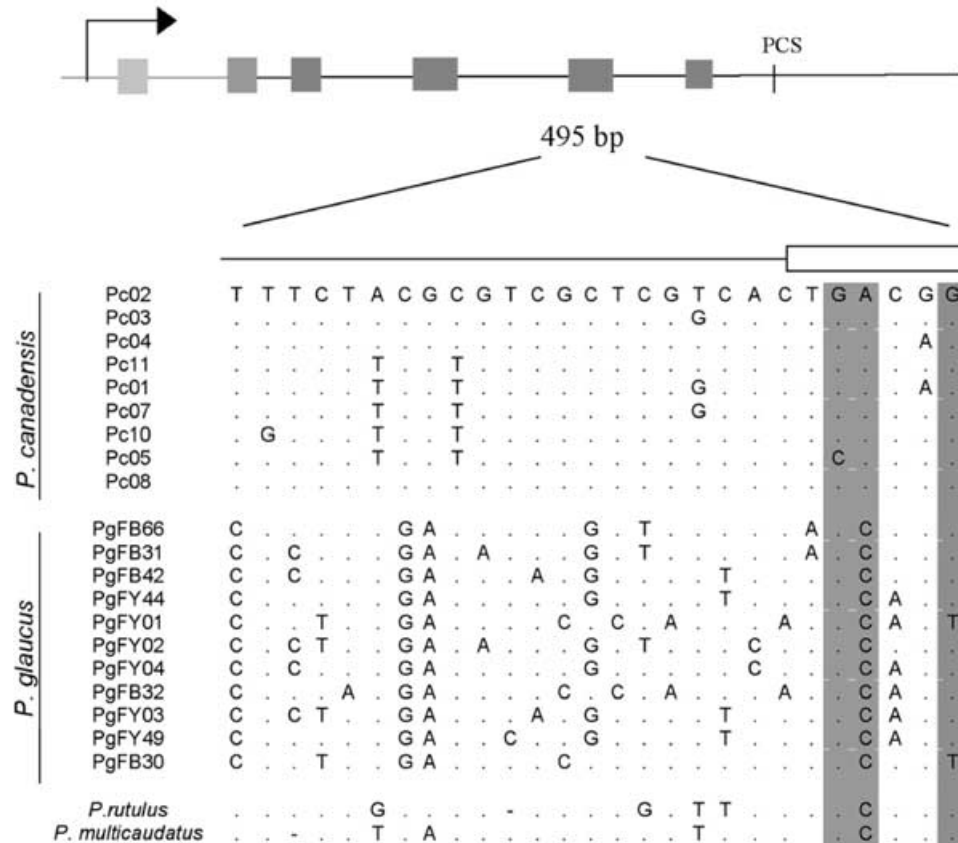
vary somewhat among loci. We implemented a likelihood ratio test to evaluate whether this reflected significant heterogeneity in recombination rates among loci. In particular, we tested whether assuming one  $\rho/\theta$  for all loci is a better fit to the data than assuming five different  $\rho/\theta$  (i.e., one for each locus). Using this approach, we found no significant evidence for recombination rate heterogeneity in either species (*P. glaucus*:  $P = 0.66$ , *P. canadensis*:  $P = 0.21$ ). Given lack of evidence for heterogeneity in recombination rates and the worry that per locus MAP estimates may be biased due to the sample size and number of segregating sites (Andolfatto and Wall 2003), we prefer to use the joint multilocus estimate of  $\rho/\theta$  in lieu of locus-specific estimates.

#### PATTERNS OF VARIATION AT CANDIDATE LOCI *Ldh* AND mtDNA

Transects through the hybrid zone between *P. glaucus* and *P. canadensis* have revealed that *Ldh* harbors a single allozyme variant that shows strong clinal differentiation between species (Hagen 1990). Because most polymorphic allozymes show little differentiation, our working hypothesis is that *Ldh* is tightly linked to a factor causing reproductive isolation between the species. Our sequenced *Ldh* clone is 6.4 kilobases long and contains five exons, which comprises most of the *Ldh* protein (we are missing 40 amino acids on the 5' end). We sequenced all exons in both species and found three nonsynonymous differences in the fourth exon of our clone (Fig. 4). Two of the mutations are conservative

amino acid changes. One mutation is from a serine to threonine substitution that is only polymorphic in one *P. canadensis*, and the other is a glycine to valine substitution that is polymorphic in *P. glaucus*. The third nonsynonymous mutation is a nonconservative change from an uncharged glutamine in *P. glaucus* to a negatively charged lysine in *P. canadensis* that appears to be fixed between species. This likely represents the electrophoretic difference underlying the *Ldh* allozyme variants that distinguish these species. Using *P. rutulus* and *P. multicaudatus* as outgroup species, the fixed Gln to Lys change appears to be derived in the *P. canadensis* lineage. We surveyed a 495 base pair region that included this exon and 306 base pairs of the upstream intron in population samples from both species. Surprisingly, despite being a candidate for selection in the hybrid zone, *Ldh* does not stand out as unusual when compared to other loci. Levels of variability, Tajima's  $D$ , and Fay and Wu's  $H$  for *Ldh* in both species were close to the average across Z-linked markers, and did not significantly differ from neutral expectations (Supplementary Material Table S3). A possible explanation for this pattern is that if selection is weak compared to levels of recombination in *Ldh* a signature of historical divergent selection at linked sites might be obscured.

The mitochondrion is another candidate target of selection opposing introgression owing to its expected linkage to the W chromosome. Interestingly, the mtDNA marker, *COI/COII*, is the only marker surveyed in this study that significantly departs from



**Figure 4.** Diagram of the *Ldh* gene and nucleotide variants in *Papilio glaucus* and *P. canadensis*. In the gene diagram the grey box represents an exon that was not sequenced in this study. Shaded nucleotides indicate nonsynonymous substitutions. Outgroup sequences of *P. rutulus* and *P. multicaudatus* are included. Dots indicate identity to the reference sequence; hyphens indicate a gap. The PolyA cleavage site is indicated as PCS.

neutral expectations (Supplementary Material Table S3). A significantly negative Tajima's  $D$  ( $P = 0.02$ , without correcting for multiple tests) and somewhat reduced polymorphism in *P. canadensis* is suggestive of a possible selective sweep involving the mitochondrion or the W chromosome. Additional markers will be necessary to distinguish between demographic and selective causes for this departure from the neutral equilibrium model.

#### MULTILOCUS PATTERNS OF DIFFERENTIATION

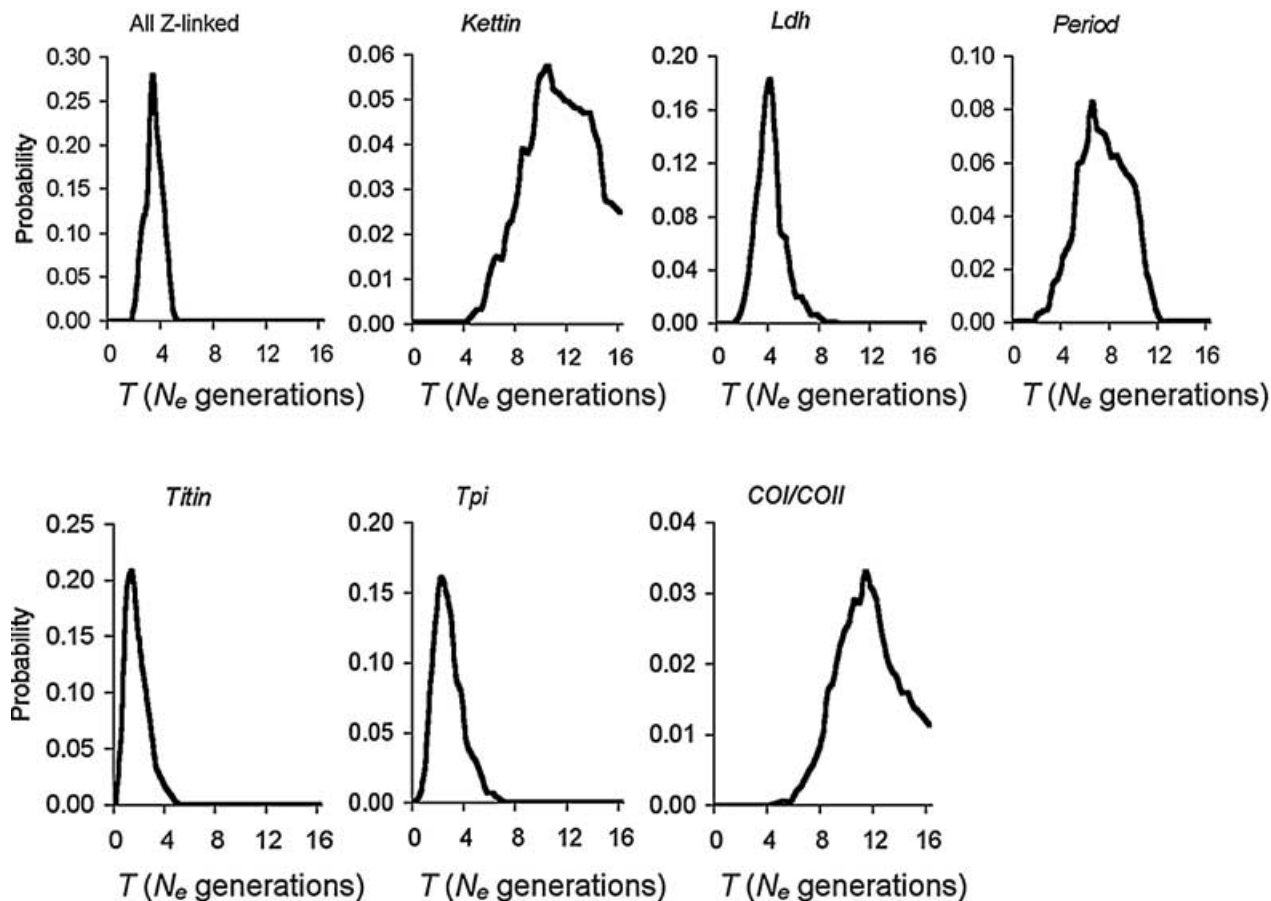
Shared polymorphisms can result from the persistence of ancestral polymorphism, introgression, or recurrent mutation. The distribution of shared and fixed, and exclusive polymorphisms between the species varies considerably among loci (Table 1). Three of the five Z-linked loci exhibited fixed differences between species and no shared polymorphisms. *Titin* has 10 shared polymorphisms (of 16 total) with one fixed difference and *Tpi*, has three shared polymorphisms (of 21 total) and no fixed differences. The mtDNA marker, *COI/COII*, contained 19 fixed differences and 0 shared polymorphism (of 33 total).

The source of shared polymorphisms of nuclear genes and the mtDNA marker may have different causes. The probability

of shared polymorphisms by parallel mutation was determined following the expectation of a hypergeometric distribution in the two species (Clark 1997). In *P. glaucus* and *P. canadensis* 13 of the 74 Z-linked polymorphisms are shared but only 1.4 are expected to be due to recurrent mutation ( $P < 0.001$ , Table 1). This rules out recurrent mutation as the cause of shared polymorphism and leaves ancestral polymorphism or introgression as possible causes. However, recurrent mutation cannot be ruled out as the cause of the one shared of 33 polymorphisms observed at *COI/COII* ( $P = 0.31$ ). This shared difference was excluded from our estimation of divergence time for *COI/COII* (see below).

#### DIVERGENCE TIME ESTIMATES

To estimate the time that *P. glaucus* and *P. canadensis* began to diverge,  $T$ , we implemented a novel approximate Bayesian method that is an extension of previous approaches (Hudson et al. 1987; Wakeley and Hey 1997; Bachtrog et al. 2006). These estimates are based on average pairwise divergence between species,  $D_{XY}$ , and within species polymorphism,  $\theta$ , and the number of shared and fixed differences observed in samples from both species. Posterior distributions of  $T$  for each locus, and a joint posterior distri-



**Figure 5.** Posterior distributions of  $T$  (in units of  $N_e$  generations) for *Papilio glaucus* and *P. canadensis*, assuming  $\theta_A = \theta$ .

bution for all Z-linked loci are shown in Figure 5. The joint maximum a posteriori (MAP) estimate for the Z-linked loci is  $3.2 N_e$  generations ago. The mutation rate in *Papilio* is not known, and the generation time of *P. glaucus* and *P. canadensis* are likely to differ (Hagen et al. 1991). However, if we crudely assume that the mutation rate per generation is similar to that of *Drosophila* (i.e.,  $1.5 \times 10^{-8}$  per year; Li 1997) and assume one generation per year, this implies that these species began to diverge around 0.6 million years ago.

In this study, we were interested in comparing divergence time estimates for two candidate loci—*Ldh* and the mitochondrion—to other Z-linked loci, with the expectation that these loci should yield deeper divergence times due to their putative linkage to reproductive isolation factors. Surprisingly, the divergence time estimate for *Ldh* is close to the joint estimate based on all Z-linked loci. The divergence time estimate for the mtDNA ( $11.7 N_e$  generations ago) is difficult to compare to nuclear genes because the appropriate coalescent scaling factor relative to the Z chromosome is not known. However, if we assume a neutral equilibrium population with equal numbers of males and females, the appropriate scaling factor would be 3, suggesting that the

scaled divergence time for the mtDNA ( $3.9 N_e$  generations ago) also agrees well with the joint estimate of  $T$  for the Z chromosome. Thus there is no evidence that these candidate loci have unusual patterns of divergence compared to other Z-linked loci.

#### TESTING THE ALLOPATRIC SPECIATION MODEL

We address the validity of the purely allopatric speciation model using only Z-linked genes. MAP estimates of  $T$  varied considerably among Z-linked loci ranging from 1.2 (*Titin*) to 10.5 (*Kettin*)  $N_e$  generations ago. A likelihood ratio test was implemented to ask whether a model positing unique divergence times for each locus fit the data significantly better than a single divergence time equal to the joint MAP estimate. We exclude the mtDNA because of uncertainty about appropriate scaling factor and possible evidence for selection (Supplementary Material Table S3). Using this test, we can reject the simple allopatric speciation model depicted in Figure 2 ( $P < 0.001$  assuming  $\theta_A = \theta$ ; Table 2). Because  $\theta_A$  is not a free parameter in our approach, we performed our test of allopatric model assuming  $\theta_A$  is  $2\times$ ,  $5\times$ ,  $8\times$ , and  $10\times \theta$ . We find that allopatry can be rejected when the assumed size of the ancestral population was 2, 5, and 8 times the current popula-

**Table 2.** Divergence time,  $T$ , in units of  $N_e$ , estimated under a simple speciation model with no migration.  $T$  is estimated from the raw data (using  $D_{XY}/\theta - 1$ ) when  $N_A = 1N_e$ , and as the MAP estimate when the ancestral population size is increased 2, 5, 8, and 10 times.

Gene	Estimated divergence time ( $T$ )				
	$N_A = 1N_e$	$N_A = 2N_e$	$N_A = 5N_e$	$N_A = 8N_e$	$N_A = 10N_e$
Kettin	10.5 (5.9–16.0) <sup>1</sup>	10.4 (5.6–15.2)	5.7 (3.0–8.0)	4.8 (2.0–6.9)	4.0 (1.6–6.6)
Ldh	4.0 (1.8–6.8)	3.6 (1.6–6.4)	2.1 (1.4–4.2)	2.0 (1.4–4.1)	1.6 (1.0–3.4)
Period	6.3 (3.4–11.0)	4.5 (2.8–8.8)	2.8 (1.2–3.6)	2.8 (1.2–3.6)	2.8 (1.2–3.4)
Titin	1.2 (.4–3.6)	.9 (.4–2.4)	.8 (.4–2.0)	.6 (.2–1.5)	.4 (.1–1.2)
Tpi	2.8 (.7–5.1)	2.4 (.6–4.9)	1.6 (.6–4.2)	1.6 (.5–4.0)	1.6 (.4–4.0)
All Z-linked	3.2 (1.8–4.2)	2.8 (1.3–3.6)	2.4 (1.1–2.2)	2.4 (1.0–2.2)	2.4 (1.0–2.0)
mt_COI/COII <sup>2</sup>	3.9 (1.3–5.3)	3.6 (1.2–4.6)	1.9 (1.1–3.6)	1.6 (1.0–3.5)	1.4 (.6–3.4)

<sup>1</sup>Confidence interval of  $T$  estimated as two log-likelihood units around the maximum.

<sup>2</sup>Divergence time is scaled to 1/3  $N_e$  of Z-linked markers.

tions size ( $P = 0.001$ ,  $P = 0.007$ ,  $P = 0.02$ , respectively) but not 10 times ( $P = 0.09$ ). These results suggest that shared ancestral polymorphism cannot account for the variation in divergence time estimates across loci unless the ancestral population was 10 times the size of the current population size.

We compared the results of our approximate Bayesian method to the *WH* method of Wakeley and Hey (1997). Estimated parameter values from the *WH* analysis were  $\theta_1 = 6.77$ ,  $\theta_2 = 5.48$ ,  $\theta_A = 49.05$ , and  $T = 0.7 N_e$ , and this approach failed to reject the allopatric speciation model using the *WH* test statistic ( $P = 0.5$ ). In the *WH* method,  $\theta_1$ ,  $\theta_2$ , and  $\theta_A$  are free parameters, and  $\theta_A$  is estimated to be about eight times larger than  $\theta_1$  and  $\theta_2$ . Based on coalescent simulations with recombination, we found the estimates of  $\theta_1$  and  $\theta_2$  are not significantly different (results not shown). However, using our approximate Bayesian approach we reject allopatry with a likelihood ratio test when  $\theta_A = 8\times\theta$  (see above;  $P = 0.02$ ). It is unclear whether it is the *WH* statistic or the more recent divergence time inferred by the *WH* method compared to our approximate Bayesian approach that explains the discrepancy between the two methods.

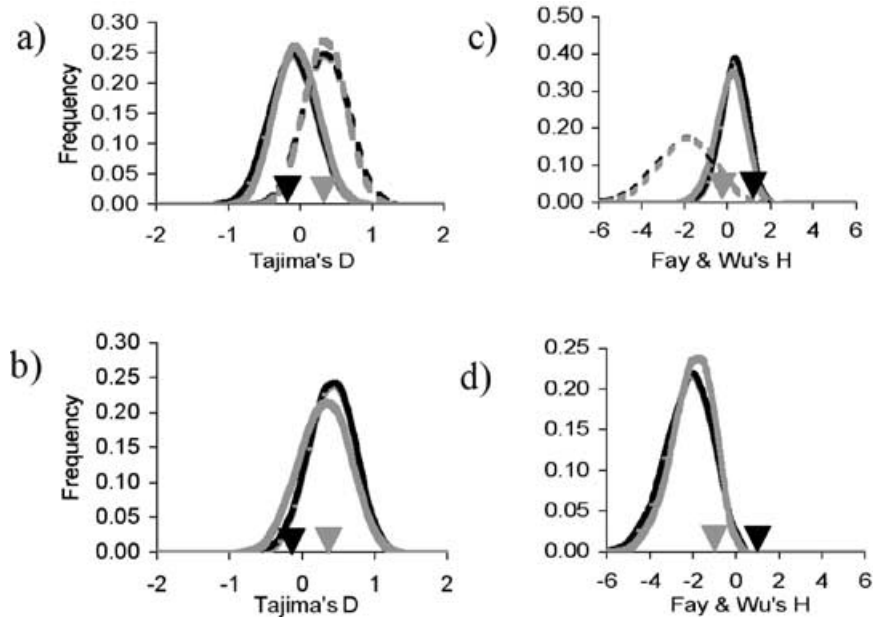
A large variance among loci in the number of shared and fixed polymorphisms (and thus estimates of divergence time in models with no migration) is expected when the ancestral population size is large compared to the current population size (Wakeley and Hey 1997). Both tests of the allopatric model suggest that the data can be reconciled with an allopatric model only if that size of the ancestral population was large relative to the current size (i.e., 10 $\times$  by our approach and 8 $\times$  by the *WH* approach). However, drastic changes in population size are expected to leave characteristic signatures in the frequency spectrum of polymorphisms (Tajima 1989; Fay and Wu 1999; Haddrill et al. 2005). We thus used information from the frequency spectrum of polymorphisms (summarized as the averages of Tajima's  $D$  and Fay and Wu's  $H$  across loci) to evaluate the fit of the data to different assumptions

about  $\theta_A$ . We compared the average Tajima's  $D$  and Fay and Wu's  $H$  for our observed data to simulations using point estimates of parameters from the *WH* method and the posterior distributions of parameters from our approximate Bayesian approach (Fig. 6). The observed average values for Tajima's  $D$  and Fay and Wu's  $H$  in *P. glaucus* are significantly different from the distributions of average  $D$  and average  $H$  obtained under the *WH* parameters ( $P = 0.03$ ). Similarly, the average Fay and Wu's  $H$  is significantly different from the observed average for *P. glaucus* using parameters estimated by our approximate Bayesian method assuming that  $\theta_A$  is 5 $\times$  and 10 $\times$  larger than  $\theta_1$  ( $P = 0.02$ ). In contrast, both the observed  $D$  and  $H$  are compatible with the distributions obtained assuming  $\theta_A = \theta$  ( $P = 0.34$  and  $P = 0.10$ ). These results suggest that the ancestral population size is unlikely to be larger than 5 $\times$  the current size as suggested by the *WH* method, and allows us to reject the allopatric model for this data. Incorporating information from the frequency spectrum may be a valuable addition to methods for estimating current and ancestral population sizes under speciation models.

## Discussion

### TESTING ALTERNATIVE MODES OF SPECIATION

Multilocus coalescent methods provide a potentially powerful means to piece together the history of speciation for recently diverged species. The role of continuing gene flow in the speciation process is much debated (reviewed in Coyne and Orr 2004). Two types of models have been developed to address different speciation scenarios. The first are simple isolation models with no gene flow that mimic strictly allopatric speciation (Takahata and Nei 1985; Hudson et al. 1987; Hey 1994; Wakeley and Hey 1997; Wang et al. 1997; Bachtrog et al. 2006). A second class of models allow for continuing migration between populations (Nath and Griffiths 1996; Beerli and Felsenstein 1999; Nielsen and Wakeley



**Figure 6.** The polymorphism frequency spectrum under alternative allopatric speciation models. For all figures, *Papilio glaucus* is represented in black and *P. canadensis* is represented in gray. Observed averages are indicated by triangles. Each graph plots 10,000 simulated replicates of the average  $D$  and  $H$  for five loci. The distributions of Tajima's  $D$  and Fay and Wu's  $H$  in (a) and (c) are based on neutral simulations using parameters estimated under the approximate Bayesian method. Solid lines represent distributions where  $\theta_A = \theta$ , and dashed lines where  $\theta_A = 5\theta$ . Distributions shown in (b) and (d) are based on neutral simulations using parameters estimated by the *WH* method (Wakeley and Hey 1997). See Methods for simulation parameters.

2001) as expected under parapatric or sympatric speciation. By assessing the fit of data from multiple unlinked regions of the genome to these models, we can begin to assess the prevalence of one mode of speciation relative to another, and the general importance of continuing gene flow in particular.

**Table 3.** Goodness-of-fit tests for one versus five divergence times among Z-linked markers.

$N_A^1$	Divergence times <sup>2</sup>	ln L	Likelihood-ratio statistic (df = 4)	$P$
1×	1	-19.78	18.36	.001
	5	-10.60		
2×	1	-19.27	18.23	.001
	5	-10.16		
5×	1	-15.01	14.14	.007
	5	-7.93		
8×	1	-13.95	12.15	.016
	5	-7.88		
10×	1	-11.74	8.13	.087
	5	-7.67		

<sup>1</sup>The size of the ancestral population relative to the current population size of *P. glaucus* and *P. canadensis*.

<sup>2</sup>The two models tested posit one divergence time for all loci versus a unique divergence time for each of the five Z-linked loci.

In the absence of recombination and recurrent mutations, genealogies will show either fixed differences or shared polymorphisms, but not both (Wakeley and Hey 1997). The presence of both shared and fixed polymorphisms among loci on the Z chromosome of *P. glaucus* and *P. canadensis* can thus be best explained by different evolutionary histories for different parts of the Z chromosome. Here we have combined a likelihood ratio test of the allopatric model with information from the frequency spectrum to show that a strictly allopatric model can be rejected for these two *Papilio* species (Table 3, Fig. 6). The wide range of divergence time estimates among loci strongly suggests that historical introgression has occurred for some parts of the Z chromosome but not in others. This pattern is consistent with the observation of differential introgression of molecular markers through the hybrid zone between these two species (Hagen 1990). Given our inference of continuing gene flow between species, we conclude that our joint speciation time estimate across loci ( $3.2 N_e$  generations ago) is most likely an underestimate of the true time the species began to diverge (Wakeley and Hey 1997; Osada and Wu 2005).

#### LOCUS-SPECIFIC DIVERGENCE TIME ESTIMATES AND IMPLICATIONS

An additional use of the coalescent-based approach is to test hypotheses about the evolutionary history of specific genes or regions

of the genome. In particular, by treating loci separately, we can assign each with an estimated divergence time. In the case of two hybridizing incipient species, we expect that regions of the genome linked to hybrid incompatibility alleles will begin to diverge as soon as these incompatibilities arise. It is also possible that some regions of the genome began to differentiate between species prior to the evolution of reproductive isolation between these species. Identifying both types of regions can potentially yield information about the genetic basis of speciation.

The accumulation of fixed differences and lack of shared polymorphism at three of five of the Z-linked loci surveyed (*Kettin*, *Ldh*, and *Period*) suggests there has been little introgression at these markers. Of particular interest is *Ldh*, for which we had prior information that alternative allozyme alleles show strong clinal differentiation between species in transects through the hybrid zone, suggesting either selection on *Ldh* itself or a linked character such as diapause, Batesian mimicry, or hybrid inviability factors. We have identified a single nonsynonymous fixed difference between *P. glaucus* and *P. canadensis* that results in a change in amino acid charge (Fig. 4) and may account for the diagnostic alleles observed in a previous allozyme study of transects through the hybrid zone (Hagen 1990). Surprisingly, however, our divergence time estimate for *Ldh* (4.0  $N_e$  generations ago) is intermediate compared to other loci, and very close to the average for Z-linked genes (3.2  $N_e$  generations ago). *Kettin* and *Period*, for which we have no prior hypotheses, have similar patterns of polymorphism to *Ldh* but deeper divergence estimates and thus are more likely to be linked to loci causing reproductive isolation. Such hypotheses could be confirmed by reciprocally introgressing these regions from one species into the other and testing for effects on hybrid inviability, or surveying patterns of introgression through the hybrid zone at these loci relative to the rest of the genome. Such regions are predicted to have a larger effect on hybrid viability than *Titin* and *Tpi*, for example, where many shared polymorphisms and few fixed differences are observed. Thus, coalescent methods may prove to be a useful tool in partitioning the Z chromosome into candidate regions of species-specific functional and/or ecological importance.

The mtDNA is a second candidate locus for selection in hybrids due to its expected linkage to the W chromosome in Lepidoptera (both are maternally transmitted), which in *P. glaucus* and *P. canadensis* carries alternative alleles for a diagnostic mimicry difference between species (Clarke and Sheppard 1962; Scriber et al. 1996). A relatively deep divergence time estimate might be expected for the mitochondrion if mimicry contributes to reproductive isolation between these species. At first glance, the divergence time for the mtDNA does appear to be deeper than for most Z-linked markers. However, given the lower expected effective population size for the mtDNA relative to the Z chromosome, it is probably more appropriate to scale the di-

vergence time by a factor of three. This correction results in a divergence time estimate for the mtDNA that is similar to joint estimate for the Z chromosome (3.9 and 3.2  $N_e$  generations, respectively). It should be noted that possible recombination between the mtDNA and the W chromosome in *Papilio* (Andolfatto et al. 2003) may uncouple their evolutionary histories to some extent, weakening the association between the mtDNA and the W chromosome.

#### HITCHHIKING EFFECTS AND ESTIMATING $T$

Inference of speciation times from population genetic data assume that the markers used are neutral and are not closely linked to other loci experiencing purifying or positive selection. In general, because speciation times are estimated in units of the effective population size, any form of selection that influences the effective population size (and thus neutral variation) in a genomic region—such as recurrent selective sweeps (Kaplan et al. 1989), background selection (Charlesworth et al. 1993), or balancing selection (Kaplan et al. 1988)—can lead to incorrect time estimates. In particular, selection that reduces variation will lead to overestimates of the divergence time whereas selection that enhances variation will lead to underestimates. In principle, hitchhiking effects could account for the greater than expected variance in estimated divergence times among loci under a strictly allopatric model that has now been described in several species (Wang et al. 1997; Machado et al. 2002; Llopart et al. 2005).

In our study, we can rule out the possibility that recent species-specific selective sweeps in either species contributes to deep divergence times estimated at *Kettin* and *Period* by considering intraspecific estimates of  $\theta$  (Supplementary Material Fig. S2). At these two markers, levels of variability are very similar in the two species, whereas a recent selective sweep in either species would be apparent as a species-specific reduction in levels of variability. More difficult to rule out is the possibility of reduced variability in both species. In *Drosophila*, regions of reduced crossing-over harbor reduced levels of variability consistent with the effects of recurrent hitchhiking and/or background selection (Begun and Aquadro 1992; Andolfatto 2001). Currently, we cannot rule out the possibility that *Kettin* and *Period* are simply located in regions of reduced recombination on the Z chromosome, and as a result are more prone to linked variation-reducing selection. Within *P. glaucus* and *P. canadensis*, we did not detect significant variation in the recombination rate among loci; however, our test is weak and clearly more loci will be necessary to determine if there is a negative correlation between recombination rates and divergence time estimates. Uncertainty about recombination rates, and the effects of linked selection, are a general problem for using coalescent-based approaches to estimate speciation times in organisms that lack genetic and physical maps. Hey and Nielsen (2004) address

this issue by adding an inheritance scalar to each locus as an additional free parameter in their model (noting that they also assume no recombination within loci). Unfortunately, this approach would be computationally prohibitive using our method.

### FUTURE PROSPECTS

Our coalescent-based approach to divergence time estimation provides the means to test models of speciation as well as identify regions of a genome potentially contributing to phenotypic species differences and factors underlying reproductive isolation between them. Here we have employed a simple allopatric model with no migration to show that such a model can be rejected based on the greater than expected variance in divergence time estimates among loci under this model. Although computationally intensive, the framework can be extended to add ancestral and current population sizes and continuing migration as free parameters within the rejection-sampling method. Here we have also shown the frequency spectrum distribution can yield useful information about the relative sizes of current and ancestral populations, and thus may be useful in estimating speciation parameters. A disadvantage of our rejection-sampling method is that it becomes less computationally feasible as the number of free parameters increases. In contrast, Markov Chain Monte Carlo-based approaches (i.e., Nielsen and Wakeley 2001; Hey and Nielsen 2004) do not suffer from this limitation. A clear advantage of our approach over these methods is the inclusion of intragenic recombination, which appears to be very strong relative to mutation in *Papilio*.

Our method can also be used to assign a temporal sequence to the appearance of reproductive isolation factors. Because reproductive isolation is not complete between *P. glaucus* and *P. canadensis*, each reproductive isolation factor that is found can be thought of as actively contributing to the on-going speciation process. Such is not the case for completely reproductively isolated species that may harbor hundreds of reproductive isolation factors, most of which were likely not involved in the speciation process (Coyne and Orr 2004). In principle, our approach could be used to distinguish between true speciation genes (i.e., those genes that were actively involved in restricting gene flow between species during speciation) and secondary reproductive isolation factors that have accumulated long after the speciation process has been completed. It will be of considerable interest to determine if genomic regions underlying locally adapted traits in *P. glaucus* and *P. canadensis*, such as diapause, host-plant preferences, and mimicry differences, show evidence for reduced gene flow between species and higher levels of differentiation than the genomic background.

Our approach, like all methods so far, makes a number of simplifying assumptions that are necessary given computational constraints. Although clearly we can look forward to more advances in coalescent-based methods, these approaches should

probably at best be used as guides to identify potentially interesting regions of the genome. Coalescent approaches should be complementary to examining patterns of introgression of candidate regions through the hybrid zone or introgressing these regions from one species to another so we can potentially verify their effect on phenotypic differences and/or reproductive isolation between species.

### ACKNOWLEDGMENTS

We thank K. Thornton and J. Huelsenbeck for helpful discussions, and D. Bachtrog for comments on the manuscript. JMS was supported by MAES Project no. 01644 at Michigan State University. PA was supported by an Alfred P. Sloan Research Fellowship in Molecular and Computational Biology.

### LITERATURE CITED

- Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* 18:279–290.
- Andolfatto, P., and M. Przeworski. 2000. A genome wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* 156:257–268.
- Andolfatto, P., and J. D. Wall. 2003. Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics* 165:1289–1305.
- Andolfatto, P. J., Scriber, and B. Charlesworth. 2003. No association between mitochondrial DNA haplotypes and a female-limited mimicry phenotype in *Papilio glaucus*. *Evolution* 57:305–316.
- Ashburner, M. 1989. *Drosophila: a laboratory handbook*. Cold Spring Harbor Laboratory Press, New York.
- Bachtrog, D., K. Thornton, A. Clark, and P. Andolfatto. 2006. Extensive introgression of mitochondrial DNA in the absence of nuclear gene flow in *Drosophila yakuba* species group. *Evolution* 60:292–302.
- Barbash, D., D. Sinno, A. Tarone, and J. Roote. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 100:5302–5307.
- Barton, N., and K. Gale. 1993. Genetic analysis of hybrid zones. Pp. 13–45 in R. Harrison, ed. *Hybrid zones and the evolutionary process*. Oxford Univ. Press, Oxford, U.K.
- . 2001. The role of hybridization in evolution. *Mol. Ecol.* 10:551–568.
- Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.
- Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rate in *D. melanogaster*. *Nature* 365:519–520.
- Beerli, P., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763–773.
- Beltran, M., C. Jiggins, V. Bull, M. Linares, J. Mallet, W. McMillan, and E. Bermingham. 2002. Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Mol. Biol. Evol.* 19:2176–2190.
- Birky, C., P. Fuerst, and T. Maruyama. 1989. Organelle gene diversity under migration, mutation, and drift: equilibrium expectation, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* 121:613–627.
- Bossart, J. L., and J. M. Scriber. 1995. Maintenance of ecologically significant

- genetic variation in the tiger swallowtail butterfly through differential selection and gene flow. *Evolution* 49:1163–1171.
- Bull, V., M. Beltran, C. D. Jiggins, W. O. McMillan, E. Bermingham, and J. Mallet. 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biol.* 4:11.
- Celniker, S. E., and G. M. Rubin. 2003. The *Drosophila melanogaster* genome. *Ann. Rev. Genom. Hum. Genet.* 4:89–117.
- Charlesworth, B., J. Coyne, and N. Barton. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* 130:113–146.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–1303.
- Clark, A. 1997. Neutral behavior of shared polymorphism. *Proc. Natl. Acad. Sci. USA* 94:7730–7734.
- Clarke, C. A., and P. M. Sheppard. 1962. The genetics of the mimetic butterfly, *Papilio glaucus*. *Ecology* 43:159–161>.
- Coyne, J., and H. Orr. 2004. *Speciation*. Sinauer Associates, Sunderland, MA.
- Dobzhansky, T. 1937. *Genetics and the origin of species*. Columbia Univ. Press, New York.
- Dopman, E., L. Perez, S. Bogdanowicz, and R. Harrison. 2005. Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proc. Natl. Acad. Sci. USA* 102:14706–14711.
- Endler, J. 1977. *Geographic variation, speciation and gene flow*. Princeton Univ. Press, Princeton, NJ.
- Emelianov, I., F. Marec, and J. Mallet. 2004. Genomic evidence for divergence with gene flow in host races of the larch budmoth. *Proc. R. Soc. Lond. B, Bio. Sci.* 271:97–105.
- Fay, J. C., and C.-I. Wu. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol.* 16:1003–1005.
- Fay, J., and C. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Grahame, J. W., C. S. Wilding, and R. K. Butlin. 2006. Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution* 60:268–278.
- Gregory, T. R., and P. D. N. Hebert. 2003. Genome size variation in lepidopteran insects. *Can. J. Zool.* 81:1399–1405.
- Hadrill, P., K. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* 15:790–799.
- Hagen, R., and M. Scriber. 1989. Sex-linked diapause, color and allozyme loci in *Papilio glaucus*: linkage analysis and significance in a hybrid zone. *J. Hered.* 80:179–185.
- . 1990. Population structure and host use in hybridizing subspecies of *Papilio glaucus*. *Evolution* 44:1914–1930.
- Hagen, R., R. Lederhouse, J. Bossart, and M. Scriber. 1991. *Papilio canadensis* and *P. glaucus* are distinct species. *J. Lepidopterist Soc.* 45:245–258.
- Harrison, R. 1990. Hybrid zones: windows on evolutionary processes. *Oxf. Surv. Evol. Biol.* 7:69–128.
- Hey, J. 1994. Bridging phylogenetics and population genetics with gene tree models. Pp. 435–447 in B. Schierwater, B. Streit, G. Wagner and R. DeSalle, eds. *Molecular ecology and evolution: approaches and applications*. Birkhauser Verlag, Basel, Switzerland.
- Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.
- . 2005. On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol.* 3:e193.
- Hudson, R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23:183–201.
- Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* 50:245–250.
- Hudson, R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
- . 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
- Kaplan, N. L., T. Darden, and R. R. Hudson. 1989. The coalescent process in models with selection. *Genetics* 120:819–829.
- Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1988b. The “hitchhiking effect” revisited. *Genetics* 123:887–899.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge Univ. Press, Cambridge, U.K.
- Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, and J. Hey. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* 156:1913–1931.
- Kronforst, M. R., L. G. Young, L. M. Blume, and L. E. Gilbert. 2006. Multilocus analyses of admixture and introgression among hybridizing *Heliconius* butterflies. *Evolution* 60:1254–1268.
- Lushai, G., D. A. S. Smith, I. J. Gordon, D. Goulson, J. A. Allen, N. Maclean. 2003. Incomplete sexual isolation in sympatry between subspecies of the butterfly *Danaus chrysippus* (L.) and the creation of a hybrid zone. *Heredity* 90:236–246.
- Li, W. 1997. *Molecular evolution*. Sinauer Associates, Sunderland, MA.
- Llopart, A., D. Lachaise, J. A. Coyne. 2005. Multilocus analysis of introgression between two sympatric sister species of *Drosophila*: *Drosophila yakuba* and *D. santomea*. *Genetics* 171:197–210.
- Loader, S. 2006. locfit: local regression, likelihood and density estimation. R package ver. 5–3. Available from <http://www.locfit.info/>.
- Logsdon, J., M. Tyshenko, C. Dixon, J. -Jafari, V. Walker, and J. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. USA* 92:8507–8511.
- Machado, C., R. Kliman, J. Markert, and J. Hey. 2002. Inferring the history of speciation from multilocus DNA sequence data. *Mol. Biol. Evol.* 19:472–488.
- Mallet, J., and N. Barton. 1989. Strong natural selection in a warning-color hybrid zone. *Evolution* 43:421–431.
- Mayr, E. 1942. *Systematics and the origins of species*. Columbia Univ. Press, New York.
- Moehring, A., A. L. Llopart, S. Elwyn, J. Coyne, and T. Mackay. 2006. The genetic basis of postzygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to hybrid male sterility. *Genetics* 173:225–233.
- Moriyama, E., and J. Powell. 1997. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J. Mol. Evol.* 45:378–391.
- Nath, H., and R. Griffiths. 1996. Estimation in an island model using simulation. *Theor. Popul. Biol.* 50:227–253.
- Nei, M. 1987. *Molecular evolutionary genetics*. Columbia Univ. Press, New York.
- Nielsen, R., and J. Wakeley. 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. *Genetics* 158:885–896.
- Orr, H. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139:1805–1813.
- Orr, H., J. Masly, and D. Presgraves. 2004. Speciation genes. *Curr. Opin. Genet. Dev.* 14:675–679.
- Osada, N., and C.-I. Wu. 2005. Testing the mode of speciation with genomic data—examples from the great apes. *Genetics* 169:259–264.

- Palopoli, M., A. Davis, and C.-I. Wu. 1996. Discord between the phylogenies inferred from molecular versus functional data: uneven rates of functional evolution or low levels of gene flow? *Genetics* 144:1321–1328.
- Payseur, B., and M. Nachman. 2005. The genomics of speciation: investigation the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. *Biol. J. Linn. Soc.* 84:523–534.
- Presgraves, D. 2003. A fine-scale genetic analysis of hybrid incompatibilities in *Drosophila*. *Genetics* 163:955–972.
- Presgraves, D., L. Balagopalan, S. Abmayr, and H. Orr. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* 243:715–719.
- Powell, D. P., M. McMichael, J-F. Silvain. 2004. Multilocus genetic analysis of host use, introgression, and speciation in host strains of fall armyworm (Lepidoptera: Noctuidae). *Ann. Entomol. Soc. Am.* 97:1034–1044.
- Przeworski, M., J. Wall, and P. Andolfatto. 2001. Recombination and the frequency spectrum in *Drosophila melanogaster* and *D. simulans*. *Mol. Biol. Evol.* 18:291–298.
- Regier, J., Q. Fang, C. Mitter, R. Peigler, T. Friedlander, and M. Solis. 1998. Evolution and phylogenetic utility of the *period* gene in Lepidoptera. *Mol. Biol. Evol.* 15:1172–1182.
- Rieseberg, L., J. Whitton, and K. Gardner. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two wild sunflower species. *Genetics* 152:713–727.
- Rozas, J., and R. Rozas. 1999. DnaSP ver. 3: an integrated program for molecular population genetics and molecular evolution. *Bioinformatics*. 15:174–175.
- Sambrook, J., and D. Russell. 2001. *Molecular cloning, a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sawamura, K., J. Roote, C.-I. Wu, and M. Yamamoto. 2004. Genetic complexity underlying hybrid male sterility. *Genetics* 166:789–796.
- Schneider, S., D. Roessli, and L. Excoffier. 2000. Arlequin ver. 2000 a software for genetic data analysis. *Genetics and Biometry Laboratory, University of Geneva, Switzerland*.
- Scriber, J., B. Giebink, and D. Snider. 1991. Reciprocal latitudinal clines in oviposition behaviour of *Papilio glaucus* and *P. canadensis* across the Great Lakes hybrid zone. *Oecologia* 87:360–368.
- Scriber, J. M., R. H. Hagen, and R. C. Lederhouse. 1996. Genetics of mimicry in the tiger swallowtail butterflies, *Papilio glaucus* and *P. canadensis*. (Lepidoptera: Papilionidae). *Evolution* 50:222–236.
- Slatkin, M. 1973. Gene flow and selection in a cline. *Genetics* 75:733–756.
- Sperling, F. A. H. 1993. Mitochondrial DNA variation's rule in *Papilio glaucus* and *Papilio troilus* groups. *Heredity* 71:227–233.
- Tajima, F. 1983. Evolutionary relationship 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., and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344.
- Tao, Y., S. Chen, D. Hartl, and C. Laurie. 2003. Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritana*. I. Differential accumulation of hybrid male sterility factors on the X and autosomes. *Genetics* 164:1383–1397.
- Thompson, J., T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24:4876–4882.
- Thornton, K. R., and P. Andolfatto. 2006. Approximate bayesian inference reveals evidence for a recent, severe, bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* 172:1607–1619.
- Ting, C., S.-C. Tsauro, and C.-I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282:1501–1504.
- True, J., B. Weir, and C. Laurie. 1996. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. *Genetics* 142:819–837.
- Turner, T. L., M. W. Hahn, S. V. Nuzhdin. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3:1572–1578.
- Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. *Genetics* 145:847–855.
- Wang, R., J. Wakeley, and J. Hey. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* 147:1091–1106.
- Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:256–276.
- Wittbrodt, J., D. Adam, B. Malitschek, W. Mauerer, and F. Raulf. 1989. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing *Tu* locus in *Xiphorhorus*. *Nature* 341:415–421.
- Wu, C.-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* 14:851–865.

Associate Editor: M. Noor

## Supplementary Material

The following supplementary material is available for this article:

**Table S1.** Sampling locations of *P. glaucus* and *P. canadensis*.

**Table S2.** Primer sequences for Z-linked loci used in this study. Genbank accession numbers are listed.

**Table S3.** Summary of the frequency distribution of polymorphisms.

**Figure S1.** Performance of divergence time estimation by our approximate Bayesian approach. Posterior distributions of 100 datasets simulated using parameters estimated from the data (see Materials and Methods) are shown. The vertical line shows the true divergence time of 3.2  $N_e$  generations.

**Figure S2.** Approximate Bayesian posterior distributions of  $\theta$  for each Z-linked locus for *P. glaucus* (black) and *P. canadensis* (gray).

**Figure S3.** Approximate Bayesian posterior distributions of  $\rho/\theta$  for each Z-linked locus for *P. glaucus* (black) and *P. canadensis* (gray).

This material is available as part of the online article from:

<http://www.blackwell-synergy.com/doi/abs/10.1111/j.1558-5646.2007.00076.x>

(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.