Adaptive Loss of an Old Duplicated Gene During Incipient Speciation

Anthony J. Greenberg,* Jennifer R. Moran,† Shu Fang,*1 and Chung-I Wu*†

*Department of Ecology and Evolution, The University of Chicago; and †Committee on Genetics, The University of Chicago

To probe the role of natural selection in species origin, we performed a DNA polymorphism survey of the Drosophila melanogaster desaturase2 (ds2) locus. ds2 is responsible for a cuticular hydrocarbon difference between two behaviorally isolated races—Zimbabwe (Z) and Cosmopolitan (M). The ds2 allele prevalent in the Z populations is functional, while the allele from the M populations harbors a 16-bp deletion upstream of the gene which knocks out its expression. We find a signature of positive selection in the ds2 promoter, but not in the control gene, sus. This signature appears to be confined to the derived M population. We also find that the selection has been recent because the gene retains a signature of a selective sweep evidenced by the departure of Fay and Wu’s H test from neutral expectation. We also find that ds2, as well as its duplicate pair ds1, has been maintained in the Drosophila genus for at least 40 Myr without any sign of adaptive change. Taken together with previous molecular genetic evidence, our results suggest that ds2 is one of the genes responsible for adaptive divergence of the Z and M races of D. melanogaster.

Introduction

One of the crucial unanswered questions in evolutionary biology is the relationship between adaptation and organismal diversity. In order to tackle this complex problem, it is essential to employ a variety of approaches. Studies relying on natural history often implicate ecological adaptation in speciation (Schluter 2000). Likewise, patterns of sequence divergence at loci involved in reproductive isolation show evidence of positive selection (Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003). The third approach involves molecular genetic manipulation of candidate speciation genes and assessment of the resulting phenotypes (Greenberg et al. 2003; Sun, Ting, and Wu 2004; Wu and Ting, 2004). The latter avenue of inquiry has the potential to add mechanistic insights into the forces that generate the patterns observed using the first two methods.

As a model of incipient speciation, we study behavioral isolation between the Zimbabwe (Z) and Cosmopolitan (M) populations of Drosophila melanogaster (Wu et al. 1995; Hollocher et al. 1997a, 1997b). Z females do not readily mate with M males, while the reciprocal cross is possible. Although the genetic architecture of this isolation is complex (Ting, Takahashi, and Wu 2001), one candidate gene, desaturase2 (ds2), has been identified (Takahashi et al. 2001; Fang, Takahashi, and Wu 2002). It encodes a Δ9 fatty acid desaturase that is responsible for a cuticular hydrocarbon (CH) polymorphism (Coyne, Wicker-Thomas, and Jallon 1999; Dallarac et al. 2000; Takahashi et al. 2001). ds2 acts primarily on myristic acid, while the product of its duplicate pair, the essential desaturase1 (ds1) gene (Labeur, Dallarac, and Wicker-Thomas 2002; Marcillach et al. 2005), acts predominantly on palmitic acid (Dallarac et al. 2000). Flies that have high levels of 5,9-heptacosadiene (HD) on their cuticle express ds2 while those with high levels of 7,11-HD do not (Dallarac et al. 2000). The lack of expression of ds2 in adults as well as high levels of 7,11-HD are due to a 16-bp deletion upstream of the ds2 gene (Takahashi et al. 2001; Greenberg et al. 2003). This deletion is derived, while the insertion-type allele is ancestral (Takahashi et al. 2001). The female CHs are mating pheromones (Jallon 1984), and the alleles at ds2 are associated with behavioral isolation between the Z and M populations (Fang, Takahashi, and Wu 2002).

In addition to behavioral isolation, Z and M races exhibit differences in competitive ability when grown in large laboratory cultures (Alipaz et al. 2005a, 2005b). Moreover, precise replacement of the deletion (M) allele of ds2 with the functional (Z) allele in the M background results in reduction in cold tolerance, while simultaneously elevating starvation resistance (Greenberg et al. 2003). These results suggest that ecological forces may have been responsible for differentiation between the two races and that ds2 may be one of the genes responsible. Although this hypothesis is attractive and has intuitive appeal, it is necessary to test whether the observed phenotypes are in fact relevant in nature.

If our hypothesis were correct, we would expect that positive selection acting on the ds2 locus would leave a signature in the patterns of DNA polymorphism in the region. We therefore surveyed DNA polymorphism at the ds2 locus in samples from the Z and M populations. We addressed the following questions. Is there evidence of sequence differentiation at ds2 between the Z and M populations? Is this differentiation confined to the ds2 locus? Is there evidence implicating positive selection in the differentiation? How recent was the selection event? We also studied divergence of ds2 coding sequence in different Drosophila species and compared its evolution to its duplicate pair, ds1. We found that the sequence differentiation is confined to ds2, that the polymorphism pattern is consistent with the action of positive selection in the M population, and that this selection was recent.

Materials and Methods

Drosophila Lines

We sequenced 31 lines of D. melanogaster and one line of Drosophila simulans (S132) as the outgroup. The strong Z lines (with discrimination index values in parentheses) were: ZS3 (4.2), Z30 (4.8), ZS6 (3.5), ZS11 (4.4), ZS56 (4.2), from Sengwa, Zimababwe, (Begun and Aquadro 1993; Hollocher et al. 1997b); ZH12 (4.6), ZH16 (3.0),
ZH34 (4.1), ZH32 (4.3), ZH23 (3.3), ZH18 (4.5), ZH21 (5.0) from Harare, Zimbabwe, (Hollocher et al. 1997a); LA2 (2.0), LA34 (3.0) from Luangwa, Zambia, (Hollocher et al. 1997a); and OK87 (4.5) from Okavanga, Botswana, (Hollocher et al. 1997a). The African M lines were: Tunis8 from Tunisia (Takahashi et al. 2001); LA66, LA47 and LA20 from Luangwa, Zambia, (Hollocher et al. 1997a); OK91 and OK17 from Okavanga, Botswana, (Hollocher et al. 1997a); ken8 from Kenya (Takahashi et al. 2001); tâ from Tâi Forest, Côte d’Ivoir, (Hollocher et al. 1997a); and SAFr17 from South Africa (Takahashi et al. 2001). The M lines from outside of Africa were: FRv1-3 from France (Hollocher et al. 1997b); Hg(82)2 from High Grove, Calif., (Hollocher et al. 1997a); and yep2–6 from Yeppoon, Australia, (I. Boussy, personal communication).

Of these lines, yep3 was not included in the analyses of polymorphism at ds2 because of a large deletion eliminating a significant portion of the coding sequence and part of the regulatory region. The LA20 line did not yield any ds2 polymorphism at the PCR product and thus was not sequenced. Likewise, either ZH32 nor Tunis8 yielded useful quantities of sas PCR product and thus were not included in the analyses of this gene.

PCR and Sequencing

DNA from single flies of each iso-female line was amplified with the Takara ExTaq polymerase (obtained through Fisher Scientific) using a standard protocol (see Supplementary Table 3 for PCR conditions and primer sequences, Supplementary Material online). Three kilobases of ds2 and 1.5 kb of sas were amplified. The PCR products were sequenced using multiple primers with an average of twofold coverage. Sequencing reactions were performed with the BigDye v3.0 terminator kit and analyzed on an ABI 3700 machine. We sequenced DNA from two different flies from each line. The bases were called using Phred and sequences assembled separately for each line with Phrap (Ewing and Green 1998; Ewing et al. 1998). We used the mace.pl script from W. Gilliland (personal communication) to align the contigs with ClustalW (Thompson, Higgins, and Gibson 1994). The alignments were then inspected and each contig was checked using Consed (Gordon, Abajian, and Green 1998), paying particular attention to base quality at polymorphic sites. After fixing occasional assembly errors, we realigned the contigs, this time keeping only bases with Phrap quality score of 30 or more (corresponding to expected base calling error rate of 10⁻³ per base). We then inspected the final alignments and corrected the gaps.

Polymorphism Analysis

Frequency Spectrum and Summary Statistics

We used the “compute” program from K. Thornton (available from http://www.molpopgen.org/) to calculate summary statistics (number of segregating sites and \( \theta_w \)) for our polymorphism data sets. We used the same program to calculate site frequency spectrum statistics, such as Tajima’s D (Tajima 1989a), a modified version of Fay and Wu’s H (Fay and Wu 2000) that is scaled by its variance (K. Zeng and Y.-X. Fu, personal communication), and Fu and Li’s D (Fu and Li 1993). \( P \) values were obtained by running 10,000 coalescent simulations under the Wright-Fisher neutral model with infinitely many sites (Hudson 1990) using a version of R. Hudson’s “ms” program (Hudson 2002) implemented within compute (K. Thornton, personal communication). We used the MKtest program from K. Thornton with modifications by J. Shapiro to perform the McDonald-Kreitman test (McDonald and Kreitman 1991). All heterozygous polymorphisms were treated as missing data. To test the sensitivity of our conclusions to this treatment, we recalculated the site frequency spectrum statistics for the ds2 regulatory region after assigning two alleles to each heterozygous single-nucleotide polymorphism (SNP). The values of all statistics calculated this way were identical to those obtained while treating heterozygous SNPs as missing data. Sites with insertion-deletion (indel) polymorphisms and with more than one derived mutation were ignored. Four sites within the ds2 regulatory region (positions 504, 514, 646, and 669 in our alignment) were in a region that could not be confidently aligned with \( D. \ simulans \). We assigned a deletion in \( D. \ simulans \) as the outgroup, but included these sites in the analyses. These sites thus did not contribute to the calculation of \( H \), but influenced Tajima’s \( D \). Although the 16-bp indel polymorphism itself was excluded from the analysis, it is linked to a SNP one nucleotide away that was probably generated by the same mutation event. It is this SNP that is marked by a triangle in figure 1.

Analysis of Population Differentiation

We extracted polymorphic sites from our alignments using compute and analyzed them using R (R Development Core Team 2005). We included only biallelic SNPs in the analysis. We calculated per-site \( F_{\text{ST}} \) using the approach of Weir and Cockerham (1984). Simultaneously, we constructed two contingency tables with counts of each of the two alleles in each population and performed Fisher’s exact tests at every site. For cumulative \( F_{\text{ST}} \), we used both the method proposed by Weir and Cockerham (WC) and that developed by Hudson, Boos and Kaplan (HBK) (1992). The latter was calculated using the equation (3e) of Charlesworth (1998) (R code is available from A. J. Greenberg). We ran 9,999 bootstrap replicates of the data by independently sampling the lines from each population with replacement using the bootstrapping package from R (Canty and Ripley 2005). We then estimated the bias and bias-corrected (\( BC_{\text{v}} \)) confidence intervals of \( F_{\text{ST}} \) (WC) and \( F_{\text{ST}} \) (HBK) for ds2 and sas (Davison and Hinkley 1997). To assess the probability that haplotypes at each locus came from the same population, we permuted the data 9,999 times by randomly assigning each line to either Z or M population without replacement (Hudson, Boos, and Kaplan 1992), again using the R bootstrapping package. The \( P \) values were calculated using the formula:

\[
P = \frac{1 + \text{#} F_{\text{ST}} \leq \hat{F}_{\text{ST}}}{1 + R},
\]

where \#\( F_{\text{ST}} \) is the number of permutations with a particular value of \( F_{\text{ST}} \), \( \hat{F}_{\text{ST}} \) is the estimated value of \( F_{\text{ST}} \), and \( R \) is the total number of permutations (Davison and Hinkley 1997).
Between-Species Comparison of Coding Sequences

Sequence Alignment

The *ds2* and *ds1* genomic sequences were obtained from the DroSpeGe database (http://bugbane.bio.indiana.edu). The two genes were on the same contig in all cases. We obtained the coding regions by comparing genomic sequences to *ds2* and *ds1* protein sequences from *D. melanogaster* using GeneWise (http://www.ebi.ac.uk/Wise2/documentation.html, ported to Mac OS X by K. Thornton). The DNA and protein sequences were then aligned using ClustalW. The DNA alignments were inspected by eye and manually edited using the protein alignments as a guide.

Fig. 1.—Per-site \( F_{ST} \) for polymorphic sites in *ds2* and *sas* calculated according to Weir and Cockerham. (A) *Z* versus all *M*; (B) *Z* versus African *M*; and (C) African versus non-African *M*. Triangle marks the 16-bp indel polymorphism. \( P \) values were calculated using Fisher's exact test.
**Tree Construction**

We constructed the tree using the DNA sequences from the ds2 and dsl open reading frames from species indicated in figure 3. We used MrBayes (Ronquist and Huelsenbeck 2003) to infer the phylogeny with separate substitution rates for each codon position, running six Markov chains. After a 20,000 generation burn-in, we ran 200,000 generations, sampling one in ten of them to obtain clade credibility values. The posterior probability $P$ of the tree presented in figure 3 is 0.881. The next most likely tree ($P = 0.060$) has ds2 from Drosophila pseudoobscura in the same clade as ds2 from Drosophila mojavensis and Drosophila virilis. Given our knowledge of the species phylogeny (Russo, Takezaki, and Nei 1995), this tree seems unlikely to us and we do not present it here. We also used the maximum likelihood approach, as implemented in the PHYLIP package version 3.6 (Felsenstein 1989), to construct the phylogeny. This method yielded the same tree as MrBayes.

**Analysis of Evolutionary Rates**

We used the “codeml” program from the PAML package (Yang 1997) to analyze patterns of divergence at silent and replacement sites in the coding sequences from ds2 and dsl. Analyses presented here are based on the tree shown in figure 3. We repeated the analyses using the next most likely tree (see above) and obtained the same results (not shown). We used a likelihood ratio test to reject the molecular clock for our data set (see table 4). Models with local clock for the tree ($P = 0.060$) has ds2 from Drosophila pseudoobscura in the same clade as ds2 from Drosophila mojavensis and Drosophila virilis. Given our knowledge of the species phylogeny (Russo, Takezaki, and Nei 1995), this tree seems unlikely to us and we do not present it here. We also used the maximum likelihood approach, as implemented in the PHYLIP package version 3.6 (Felsenstein 1989), to construct the phylogeny. This method yielded the same tree as MrBayes.

As a measure of sequence differentiation between populations, we calculated the commonly used $F_{ST}$ statistic (Wright 1951). We computed $F_{ST}$ for each biallelic SNP according to the formula proposed by Weir and Cockerham (1984). We simultaneously assessed the significance of the population subdivision at each SNP by constructing $2 \times 2$ contingency tables (number of lines from each population carrying either of the two segregating alleles) and performing a Fisher’s exact test. The results of this analysis are shown in figure 1. It is clear that the 16-bp indel (marked by a triangle in fig. 1) polymorphism is the highest $F_{ST}$ site and is significantly differentiated between the two populations (Fisher’s exact test $P = 5.4 \times 10^{-5}$). The differentiation remains significant even after a Bonferroni correction for multiple tests. It is also noteworthy that $F_{ST}$ at this site remains significant when we compare the Z and African M populations (fig. 1B). In contrast, there are no significantly differentiated sites between African and non-African M populations. The population structure at ds2 is thus due to racial rather than geographic subdivision.

To further verify that the population structure at ds2 is significant, we calculated cumulative $F_{ST}$ at this locus as well as at sas. We used two estimators of $F_{ST}$: one proposed by Weir and Cockerham (Weir and Cockerham 1984) and another developed by Hudson, Boos, and Kaplan (Hudson, Boos, and Kaplan 1992; Charlesworth 1998). Unfortunately, neither the distributions nor variances of these estimators are known. Moreover, they are guaranteed to be unbiased only if the effective sizes of the populations being compared are equal (Weir and Cockerham 1984; Hudson, Boos, and Kaplan 1992), a situation that is unlikely to be true for Z and M populations. We therefore turned to nonparametric bootstrapping (Efron 1979; Davison and Hinkley 1997) to estimate the confidence intervals, bias, and $P$ values of the statistics (table 1; see Materials and Methods for details). We indeed find that both estimators of $F_{ST}$ are somewhat biased when applied to our sample. Despite this, they show clear evidence of significant population structure at ds2 but not at sas.

Recent Positive Selection on the ds2 Upstream Region in the M Population

Having established significant sequence differentiation at ds2, we next wanted to know if it arose due to positive selection. We therefore analyzed patterns of polymorphism at ds2 and sas in each population separately. Consistent with previous reports (Begun and Aquadro 1993; Haddrill et al. 2005; Ometto et al. 2005), we see a reduction of nucleotide diversity ($\theta_{N}$) in the M population sample compared to Z (table 2). This pattern is evident at both ds2 and sas loci. Although $\theta_{N}$ in the ds2 coding region from Zimbabwe is lower than that in the M lines, this is due to lack of amino acid replacement polymorphism in the Z population. If only silent and intron sites are considered, $\theta_{S}$ in the Z population is greater than that in the M population for this region, too.

To test whether our data are consistent with neutral evolution, we calculated various frequency spectrum statistics (table 2). The ds2 regulatory region displays a significant departure from neutrality in the M population, but not
in Z. Tajima’s $D$ test is significant at both $ds2$ and the neighboring $sas$ locus, and thus may reflect demographic rather than selective forces. In contrast, Fay and Wu’s $H$ is significant only in the $ds2$ upstream region. Moreover, the signal does not come solely from the 16-bp deletion itself. When it is removed from the analysis, the $H$ statistic still deviates from the neutral expectation ($H = -2.73$, $P = 0.0119$).

To further pinpoint the region of the $ds2$ locus responsible for the departure from neutrality, we performed a sliding window analysis of three estimators of $h$ along the $ds2$ sequence. Each window covers five polymorphic sites and is moved one polymorphic site at a time. Differences between two of the $h$ estimators ($h_w$ and $h_1$) and $h_2$ form the basis for Tajima’s $D$ and Fay and Wu’s $H$ tests, respectively. Plots of these measures of nucleotide diversity show that the departure from neutrality is due to a 750-bp region of $ds2$ that encompasses the 16-bp indel polymorphism (fig. 2).

The 16-bp deletion upstream of $ds2$ results in the loss of expression of the gene (Dallerac et al. 2000), at least in the adults. It is thus possible that $ds2$ is on its way to becoming a pseudogene in the M populations. We indeed see evidence of this in four of the M lines we sequenced (see the alignments in the supplementary material, Supplementary Material online). These lines, one from Africa (LA66, also observed by Takahashi et al. [2001]) and three from Australia (yep3, yep4, and yep6), harbor deletions and insertions in the $ds2$ coding region that result in frameshifts. We had to eliminate the yep3 line from our polymorphism insertions in the $ds2$ coding region because it carries a 350-bp deletion that eliminates a pseudogene in the M populations. We indeed see evidence that a recent, single selection event gave rise to the patterns we observe.

$ds2$ is a Constrained Ancient Duplicate

The $ds2$ gene forms a tandem duplicate pair with $ds1$ (Dallerac et al. 2000). Unlike $ds2$, $ds1$ is an essential gene (Labeur, Dallerac, and Wicker-Thomas 2002; Marcillac et al. 2005). It is thus plausible that $ds2$ has been historically less constrained than its counterpart. To see if this is true, we constructed a phylogeny of the $ds2$ and $ds1$ open reading frames from seven $Drosophila$ species (fig. 3; see Materials and Methods for details). We found both genes in all species examined, suggesting that the duplication event happened before the divergence between Sophophora and Drosophila clades, an event that happened around 40 MYA (Russo, Takezaki, and Nei 1995). Each gene forms its own well-supported clade (fig. 3), and both branches follow the species trees (Russo, Takezaki, and Nei 1995).

Using the PAML package (Yang 1997), we performed likelihood ratio tests to look for evidence of positive selection or relaxation of constraint on either $ds2$ or $ds1$. Although the tests suggest that the ratio of substitution rates at amino acid replacement and silent sites ($K_a/K_s$) in the $ds2$ gene is significantly bigger than in $ds1$ (middle portion of table 4 and fig. 4), the difference is not large, and

### Table 2

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of Sites</th>
<th>$\theta_w 	imes 100$</th>
<th>Tajima $D$</th>
<th>Fay and Wu $H$</th>
<th>Fu and Li $D$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cosmopolitan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sas$ gene</td>
<td>20</td>
<td>0.2131</td>
<td>$-1.86 (0.016)$</td>
<td>$-1.57 (0.064)$</td>
<td>$-1.42 (0.133)$</td>
</tr>
<tr>
<td>$ds2$ regulatory</td>
<td>11</td>
<td>0.1669</td>
<td>$-2.00 (0.010)$</td>
<td>$-3.09 (0.011)$</td>
<td>$0.21 (0.678)$</td>
</tr>
<tr>
<td>$ds2$ gene</td>
<td>5</td>
<td>0.0859</td>
<td>$-0.85 (0.224)$</td>
<td>$0.69 (0.736)$</td>
<td>$-1.17 (0.243)$</td>
</tr>
<tr>
<td><strong>Zimbabwe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sas$ gene</td>
<td>19</td>
<td>0.3719</td>
<td>$-0.21 (0.462)$</td>
<td>$-1.39 (0.072)$</td>
<td>$0.29 (0.679)$</td>
</tr>
<tr>
<td>$ds2$ regulatory</td>
<td>12</td>
<td>0.3058</td>
<td>$-0.63 (0.202)$</td>
<td>$-0.79 (0.142)$</td>
<td>$0.31 (0.709)$</td>
</tr>
<tr>
<td>$ds2$ gene</td>
<td>5</td>
<td>0.0699</td>
<td>$-1.20 (0.109)$</td>
<td>$0.70 (0.740)$</td>
<td>$-1.21 (0.232)$</td>
</tr>
<tr>
<td><strong>Zimbabwe and Cosmopolitan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sas$ gene</td>
<td>29</td>
<td>0.2834</td>
<td>$-1.49 (0.047)$</td>
<td>$-1.19 (0.093)$</td>
<td>$-1.74 (0.080)$</td>
</tr>
<tr>
<td>$ds2$ regulatory</td>
<td>16</td>
<td>0.2773</td>
<td>$-1.05 (0.151)$</td>
<td>$-1.92 (0.042)$</td>
<td>$0.03 (0.598)$</td>
</tr>
<tr>
<td>$ds2$ gene</td>
<td>9</td>
<td>0.0834</td>
<td>$-1.56 (0.044)$</td>
<td>$0.60 (0.685)$</td>
<td>$-1.71 (0.111)$</td>
</tr>
</tbody>
</table>

 NOTE.—Significant $P$ values (less than 5%) are in bold.

* African and non-African M.
the *ds2* gene is clearly quite constrained (*K_\alpha/K_s* = 0.08 (0.06, 0.10)). We were unable to find any evidence for historic positive selection (bottom portion of table 4). We also note that the essential elements that define fatty acid desaturases are present and conserved in both *ds2* and *ds1* (Supplementary Figure 1, Supplementary Material online). The only exception can be found in the *ds1* gene of *D. mojavensis* which lacks the third and fourth transmembrane domains because of an in-frame deletion. Although this deletion probably affects function in some way, the predicted topology of this protein (Knipple et al., 2002) should still allow for a functional active site.

To further investigate the degree of constraint on the *ds2* gene, we assayed its expression in male and female adults in species of *D. melanogaster* subgroup (*D. melanogaster*, *D. simulans*, *Drosophila mauritiana*, and *Drosophila sechellia*; fig. 4). We find that this gene is expressed in males and females in all of these species, with the exception of *M D. melanogaster*. We also find that *ds2* is expressed in African male *D. melanogaster*, despite a previous report (Dallerac et al. 2000) that its expression is restricted to females. We do not know the reason for this discrepancy, but we note that A. Llopart (personal communication) independently found that *ds2* is indeed expressed in male *D. melanogaster* lines that carry the insertion allele.

We thus conclude that although *ds2* has been evolving slightly faster than *ds1*, it was highly constrained for tens of millions of years and its loss in the *M* population of *D. melanogaster* could not have been predicted based on its prior evolutionary history.

**Discussion**

**Positive Selection Accompanies Sequence Differentiation at *ds2***

In this study, we examined DNA sequence polymorphism patterns at the *ds2* locus in the *Z* and *M* behavioral races of *D. melanogaster*. We find that this locus is significantly different between these populations at the nucleotide level. This difference is not attributable to geographic subdivision because it is not observed between African and non-African *M* populations. Furthermore, we detect a signature of positive selection on the *ds2* regulatory sequence within the *M* population. A widespread polymorphic inversion, *In(3R)K*, breaks close to our locus. This could have an effect on the population recombination levels in the region and thus influence our conclusions. To control for this, we sequenced *sas*, a gene 7 kb closer to the breakpoint. Unlike *ds2*, this gene shows no sequence differentiation between *Z* and *M* populations. Neither does it show a significant skew toward the high-frequency–derived variants present in the *ds2* regulatory region. Moreover, *In(3R)K* segregates within most of our *Z* lines (M.-L. Wu, personal communication).

<table>
<thead>
<tr>
<th>Table 3</th>
<th>No Excess of Divergence at the <em>ds2</em> Locus. <em>P</em> Values Were Calculated Using Fisher’s Exact Test on 3 × 2 Contingency Tables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regulator</td>
</tr>
<tr>
<td><strong>Cosmopolitan</strong></td>
<td></td>
</tr>
<tr>
<td>Polymorphism</td>
<td>11</td>
</tr>
<tr>
<td>Divergence</td>
<td>61</td>
</tr>
<tr>
<td><strong>Zimbabwe</strong></td>
<td></td>
</tr>
<tr>
<td>Polymorphism</td>
<td>12</td>
</tr>
<tr>
<td>Divergence</td>
<td>61</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
</tr>
<tr>
<td>Polymorphism</td>
<td>16</td>
</tr>
<tr>
<td>Divergence</td>
<td>60</td>
</tr>
</tbody>
</table>

* African and non-African *M*. 

**FIG. 2.—** Sliding window analysis of polymorphism at *ds2*. Window size five polymorphic sites, step size one site. Each window starts at the nucleotide after the previous polymorphic site and ends on the nucleotide after the last included site. Windows with significant Tajima’s *D* or Fay and Wu’s *H* are marked at the top: ‘+’ indicates 0.01 < *P* ≤ 0.05 and ‘++’ indicates *P* ≤ 0.01. (A) Analysis of Cosmopolitan population (African and non-African lines pooled). (B) Analysis of the *Z* population. Empty boxes represent the exons; shaded box represent the region of problematic alignment with *Drosophila simulans*. *θ_n*, *θ_w*, and *θ_l* are plotted as indicated in the legend.
even though these lines are monomorphic for the insertion-type allele of ds2. It is thus unlikely that \textit{In(3R)K} influences our conclusions. The essential \textit{ds1} gene is located on the other side of ds2, with a small hypothetical gene (CG17207) in between the two desaturases. Like ds2, ds1 is involved in CH production (Marcillac, Grosjean, and Ferveur 2005), and therefore could potentially account for the association of ds2 with Z/M behavioral incompatibility observed by Fang, Takahashi, and Wu (2002). However, this is highly unlikely because no sequence differentiation between Z and M races can be seen at ds1 (J. Shapiro, personal communication).

In addition to the influence of polymorphic inversions, several circumstances could confound our evaluation of the extent of nucleotide divergence between the Z and M populations. The estimators of \( F_{ST} \) we used are likely biased because our two populations are almost certainly of unequal effective size. Indeed, using a bootstrap procedure we calculate that both measures (Weir and Cockerham 1984; Hudson, Boos, and Kaplan 1992) overestimate the true value. However, even the bias-corrected confidence intervals for \( F_{ST} \) at the ds2 (unlike \textit{sas}) locus do not include zero. Another potential problem with inferring the extent of population structure at ds2 is due to unequal nucleotide diversity in our Z and M samples. Charlesworth (1998) pointed out that paucity of variants in one population could lead to high \( F_{ST} \) that is due to disparities in levels of polymorphism rather than high levels of differentiation. However, the ratio \( \theta_0(Z)/\theta_0(M) \) is the same for ds2 and \textit{sas}, while estimates of \( F_{ST} \) are different. Variations in levels of nucleotide diversity are thus not sufficient to explain our observations.

Although the differentiation between the Z and M populations at ds2 is clearly significant, it is not very high due to low linkage disequilibrium (LD) between the 16-bp allele that probably drives the pattern and the rest of the polymorphic sites. Fang, Takahashi, and Wu (2002) found that the insertion allele of ds2 is associated with strength of behavioral isolation between the Z and M populations. The low level of LD we observed in the region makes it likely that this association is indeed due to ds2 and not some untyped polymorphism at a linked locus.

Patterns of nucleotide variability in the region are clearly not compatible with an equilibrium neutral model. However, one must not be too hasty in inferring the action of adaptive forces from polymorphism data. Effects of demography and selection on the mutation frequency

### Table 4

<table>
<thead>
<tr>
<th>Clock</th>
<th>Number of ( K_e/K_s )</th>
<th>Site Model</th>
<th>( \chi^2 ) (df)</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
<td>NA</td>
<td>(-5680.55)</td>
<td>21.60 (12)</td>
</tr>
<tr>
<td>No</td>
<td>1</td>
<td>NA</td>
<td>(-5660.99)</td>
<td>4.15 (1)</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
<td>NA</td>
<td>(-5658.91)</td>
<td>2.21 (2)</td>
</tr>
<tr>
<td>No</td>
<td>1</td>
<td>M7</td>
<td>(-5530.56)</td>
<td>0.0245</td>
</tr>
<tr>
<td>No</td>
<td>1</td>
<td>M8</td>
<td>(-5529.46)</td>
<td>0.0245</td>
</tr>
</tbody>
</table>

**NOTE.**—df, degrees of freedom; NA, not applicable.

---

**FIG. 3.**—A Bayesian phylogeny of ds2 and ds1. Numbers on branches represent clade credibility values. Branch lengths are proportional to \( K_e \) as calculated using PAML (model with two \( K_e/K_s \) ratios). Confidence intervals and the \( P \) value were calculated using 999 bootstrap permutations of codons.

spectrum are often similar. For example, if a population either has been increasing in size or has gone through a bottleneck, as is likely for the M population (David and Capy 1988), one expects an excess of low-frequency polymorphism and thus negative values of Tajima’s $D$ statistic (Tajima 1989b; Slatkin and Hudson 1991). Indeed, the reduction in levels of nucleotide diversity and the significant excess of low-frequency variants in the M population is evident at both $ds2$ and $sas$. These patterns are consistently seen in samples from this population (Begun and Aquadro 1993; Haddrill et al. 2005; Ometto et al. 2005) and are therefore likely due to demographic rather than adaptive processes. In contrast, the significant skew toward high-frequency–derived variants (as measured by Fay and Wu’s $H$ statistic) is confined to the $ds2$ regulatory region. This statistic should be less sensitive to demographic events such as population expansion (Fay and Wu 2000), although it may falsely reject neutrality when the population is subdivided (Przeworski 2002). However, we see no sign of structure in the M population. Thus, a significant $H$ test probably reflects the action of positive selection in our case. Even so, one can never guarantee that some plausible demographic model with no selection will not fit the data (Przeworski 2002; Haddrill et al. 2005). If we are to be reasonably confident that selection has generated the pattern we observe, experimental evidence that variants in the sample influence fitness is necessary. Fortunately, such experimental data are available for $ds2$ (Greenberg et al. 2003). Precise substitution of the deletion allele for the insertion one leads to reduction in cold tolerance while elevating starvation tolerance. Taken together with the polymorphism patterns we see, these observations strongly suggest that adaptation has played a role in generating sequence differentiation at $ds2$.

Recent Loss of a Historically Constrained Duplicate Gene

Because the 16-bp deletion leads to loss of expression of $ds2$ (Dallerac et al. 2000; Takahashi et al. 2001; Greenberg et al. 2003), it is logical to look for signs that this gene has become nonfunctional in the M population. We indeed see hints that this is happening. Four out of 15 M lines we sequenced harbor insertions or deletions in the $ds2$ open reading frame that result in elimination of large portions of the protein. However, the signs of relaxation of constraint are weak, probably because the deletion arose too recently for a significant number of changes to have accumulated. The notion that selection at $ds2$ was relatively recent is further supported by an unremarkable extent of divergence in the region, as compared to levels of polymorphism. Simulations show that the signature of a selective sweep that can be captured by the $H$ test is short lived (Przeworski 2002). However, it is possible that the 16-bp deletion allele is kept from fixation in the M population by gene flow. Thus, in the case of $ds2$, the $H$ test could be detecting incomplete fixation of the positively selected allele rather than incomplete hitchhiking. If this is the case, it is hard to estimate the time since the selective sweep took place. Nevertheless, taking all the data together, we are confident that the selection event happened approximately at the time the M and the Z populations split.

The $ds2$ gene forms a duplicate pair with $ds1$, an essential gene also responsible for CH biosynthesis (Labeur, Dallerac, and Wicker-Thomas 2002; Marcillac et al. 2004, Marcillac, Grosjean, and Ferveur 2005; Ueyama et al. 2005). The two genes are the only representatives of the KPSE group of desaturases in D. melanogaster (Knipple et al. 2002). It is reasonable to suppose that $ds2$ would be under less constraint than $ds1$, and perhaps has experienced a history of adaptive evolution. However, despite the signs that $ds2$ is becoming a pseudogene in the M population of D. melanogaster, we found that this gene has been highly constrained for at least 40 Myr. We saw no signs of positive selection during the evolution of $ds2$ in the Drosophila genus.

It is interesting to note that this is the second example of a duplicated gene being involved in speciation. The other such gene, $Ods$ (Ting et al. 1998, 2004), has had a longer history of rapid evolution and has accumulated enough changes to show a significant excess of amino acid substitutions. Although the number of speciation genes is still too limited to draw any firm conclusions, this pattern does seem to bear out the suggestion (Lynch and Force 2000) that duplicated genes may play a major role in generation of reproductive incompatibility. However, neither $Ods$ nor $ds2$ seem to evolve according to the complementary degeneration model (Force et al. 1999; Lynch and Force 2000). Indeed, although $ds2$ is being lost, its demise appears to be driven by adaptation rather than through a chance accumulation of degenerative mutations, while its counterpart, $ds1$, remains conserved.

Conclusions

We find that a recent selective sweep likely drove the evolution of the $ds2$ regulatory region. It is thus becoming clear that the $ds2$ locus is involved in differential adaptation of the Z and M populations, probably to disparate ecological circumstances (Greenberg et al. 2003). It also appears to contribute to reproductive isolation between the two populations (Fang, Takahashi, and Wu 2002; J. R. Moran, unpublished data). Our results provide further support to the hypothesis that selection leads to speciation, this time at the early stages of the process.

Supplementary Materials

Supplementary Table 3 and Figure 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank K. Thornton for software and discussion, J. Shapiro for providing his version of MKtest and for useful discussion, M.-L. Wu for unpublished data on inversion polymorphism, A. Llopart for unpublished data on $ds2$ expression, J. Felsenstein for information on bootstrapping of codons in PHYLIP, and S. Moores for help with sequencing. The study was supported by the National Institutes of Health (NIH) grant to C.-I.W., an NIH Ruth L. Kirschstein National Research Service Award fellowship to A.J.G., and a Graduate Assistance in Areas of National Need grant number P-P200A030043 from the Department of Education to J.R.M.
Literature Cited


Takahashi, A., S. C. Tsaur, J. A. Coyne, and C.-I Wu. 2001. The nucleotide changes governing cuticular hydrocarbon variation


Marta Wayne, Associate Editor

Accepted October 19, 2005