Neutral Evolution of the Sex-Determining Gene *transformer* in Drosophila

Bryant F. McAllister* and Gilean A. T. McVean*

*Institute for Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, and
†Department of Biology, University of Texas, Arlington, Texas 76019-0498

Manuscript received September 8, 1999
Accepted for publication December 22, 1999

**ABSTRACT**

The amino acid sequence of the *transformer* (tra) gene exhibits an extremely rapid rate of evolution among Drosophila species, although the gene performs a critical step in sex determination. These changes in amino acid sequence are the result of either natural selection or neutral evolution. To differentiate between selective and neutral causes of this evolutionary change, analyses of both intraspecific and interspecific patterns of molecular evolution of *tra* gene sequences are presented. Sequences of 31 *tra* alleles were obtained from *Drosophila americana*. Many replacement and silent nucleotide variants are present among the alleles; however, the distribution of this sequence variation is consistent with neutral evolution. Sequence evolution was also examined among six species representative of the genus *Drosophila*. For most lineages and most regions of the gene, both silent and replacement substitutions have accumulated in a constant, clock-like manner. In exon 3 of *D. virilis* and *D. americana* we find evidence for an elevated rate of nonsynonymous substitution, but no statistical support for a greater rate of nonsynonymous relative to synonymous substitutions. Both levels of analysis of the *tra* sequence suggest that, although the gene is evolving at a rapid pace, these changes are neutral in function.

The rate of evolutionary change in amino acid sequence varies considerably among proteins. Sequence conservation is most easily explained by the action of purifying selection; changes in an amino acid sequence are deleterious and thus are rapidly eliminated from populations after their occurrence (Hall 1954; Li and Nei 1977; Waxman and Peck 1998). In contrast, evolutionary changes in amino acid sequences present an interesting problem, because two alternative mechanisms are potentially responsible (Kimura 1968, 1983; King and Jukes 1969). The neutral hypothesis proposes that the vast majority of amino acid replacements represent changes that do not affect the function of the gene product. Alternatively, natural selection may lead to the fixation of adaptive amino acid changes. If there are multiple sites in a gene where adaptive changes are possible, such genes may experience an accelerated evolutionary rate. Identifying genes that are evolving rapidly under positive selection is important for understanding organismal evolution, because these genes are potentially significant for adaptation, developmental change, and speciation.

Comparisons among *Drosophila* species reveal that one of the most rapidly evolving genes in this organism’s genome is *transformer* (*tra*), a gene involved in the primary somatic sex-determination pathway (O’Neil and Belote 1992; Cline and Meyer 1996; Civetta and Singh 1998). Production of a functional product from the *tra* gene occurs exclusively in females, which is mediated through differential splicing by the product of the Sex-lethal (Sxl) gene (Boggs et al. 1987; Nagoshi et al. 1988), the primary switch in sexual differentiation (Cline 1993). Regulation of Sxl occurs in response to the chromosomal complement of a developing embryo, where SXL, an RNA binding protein, is produced exclusively in females. In male embryos where SXL is absent, non-sex-specific splicing of *tra* pre-mRNA results in a transcript that contains a stop codon early in the reading frame. In females, SXL protein binds to *tra* pre-mRNA and inhibits the non-sex-specific 3’ splice site of intron 1, generating a message containing a full-length reading frame. The full-length product of *tra* interacts with a gene product of *transformer-2* (*tra2*) to mediate female-specific splicing of doublesex (dsx) pre-mRNA leading to sexual differentiation of somatic cells.

Observation that an essential component of the Drosophila sex-determining pathway has a rapid evolutionary rate presents a case for differentiating between low functional constraints and positive natural selection. Demonstrating that *tra* exhibits an unusually high rate of amino acid change relative to other genes (O’Neil and Belote 1992; Civetta and Singh 1998) does not necessarily indicate that the rapid evolutionary rate is driven by adaptation. A previous study of variability at the *tra* gene in *Drosophila melanogaster* did, however, reveal the presence of low nucleotide sequence variability in this species (Walther and Schaeffer 1994). One
possible cause of low variation is the hitchhiking effect of beneficial substitutions; adaptive amino acid mutations can sweep rapidly through a population, causing loss of variation at linked sites (Maynard Smith and Haigh 1974; Kaplan et al. 1989; Stephan et al. 1992). However, there are several other factors that can also reduce nucleotide variation, but with few polymorphisms there is low resolving power for distinguishing between hypotheses (Simonsen et al. 1995).

In this study, we examine the pattern of intraspecific nucleotide variability at tra in natural populations of D. americana and D. a. texana, members of the virilis species group. These two subspecies are very closely related, exhibiting negligible sequence differentiation from each other (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999), and will be referred to collectively as D. americana. Analyses of sequence variability at several genes indicate that D. americana has high levels of sequence diversity in its nuclear genome (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999). An average rate of recombination is expected in the tra gene region of D. americana. The cytological position of tra has been determined at salivary chromosome band 3F in D. virilis (O’Neill and Belote 1992), and the chromosomes of D. americana are colinear with D. virilis in this region (Stalker 1940; Hu 1952). This internal position suggests normal levels of local recombination, which, coupled with the expectation of high sequence diversity, is important for detecting selection pressures acting specifically on tra.

In addition to the tra sequences of D. americana, sequences are available from five other Drosophila species. Analyses of these sequences provide further evidence on the nature of the changes in the tra gene. Neutral evolution predicts a steady, clock-like accumulation of high sequence diversity, is important for detecting selection pressures acting specifically on tra. However, there are several other factors that can also reduce nucleotide variation, but with few polymorphisms there is low resolving power for distinguishing between hypotheses (Simonsen et al. 1995).

In this study, we examine the pattern of intraspecific nucleotide variability at tra in natural populations of D. americana and D. a. texana, members of the virilis species group. These two subspecies are very closely related, exhibiting negligible sequence differentiation from each other (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999), and will be referred to collectively as D. americana. Analyses of sequence variability at several genes indicate that D. americana has high levels of sequence diversity in its nuclear genome (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999). An average rate of recombination is expected in the tra gene region of D. americana. The cytological position of tra has been determined at salivary chromosome band 3F in D. virilis (O’Neill and Belote 1992), and the chromosomes of D. americana are colinear with D. virilis in this region (Stalker 1940; Hu 1952). This internal position suggests normal levels of local recombination, which, coupled with the expectation of high sequence diversity, is important for detecting selection pressures acting specifically on tra.

In addition to the tra sequences of D. americana, sequences are available from five other Drosophila species. Analyses of these sequences provide further evidence on the nature of the changes in the tra gene. Neutral evolution predicts a steady, clock-like accumulation of sequence divergence over time, at both synonymous and nonsynonymous sites. In contrast, adaptive accumulation of replacement changes may deviate from clock-like behavior, due to bursts of substitution in specific lineages. In addition, we test whether specific lineages or regions of the gene show ratios of nonsynonymous to synonymous substitution greater than one, which has been used as evidence for adaptive evolution in several instances (e.g., Hughes et al. 1990; Messier and Stewart 1997).

**MATERIALS AND METHODS**

**DNA sequencing in D. americana:** The published sequences of the tra genes of D. virilis and D. hydei (O’Neill and Belote 1992) were used to identify regions with conserved nucleotide sequence. Oligonucleotides, corresponding to the D. virilis sequence, were synthesized in two relatively conserved regions that encompass most of the tra gene. The 3’ base of the forward primer is located 33 bases upstream of the start codon in the D. virilis sequence and the 3’ base of the reverse primer is located 30 bases upstream of the stop codon (Figure 1). These general primers are designated tra genF (5’att gca agt ggc caa ata gc) and tra genR (5’ctg atc att cct ata gc). Initially, crude sequences were obtained from D. americana and D. virilis using these general primers to amplify and directly sequence the gene from inbred flies. An ~960-bp region was amplified from DNA template of two separate flies from each of three inbred laboratory strains of D. americana and two inbred strains of D. virilis. Standard conditions for the PCR amplification were a 25-μl reaction containing 0.5 μM each dNTP, 0.2 μM each primer, and 2 units Taq polymerase. An MJ Research (Watertown, MA) thermocycler was used to incubate the reaction mixture at 95° for 2 min and then cycle 30 times through the following conditions: 95°, 0.5 min; 58°, 0.5 min; 72°, 1 min. Each of the PCR products was column purified (Qiagen, Valencia, CA) and used directly in cycle sequencing reactions. Approximately 30 ng of each purified PCR product was added to 4 μl of ABI PRISM dye terminator reaction mix (Perkin Elmer, Norwalk, CT) and 3 pg of either the tra genF or tra genR oligonucleotide in a final reaction volume of 10 μl. The reactions were cycled 25 times with the following parameters: 95°, 30 sec; 58°, 30 sec; 60°, 3 min. Products were ethanol precipitated and electrophoresed on an ABI 377 automated DNA sequencer.

Upon comparison of the sequences that were obtained, many nucleotide differences between the D. americana and D. virilis lines were evident, thus representing potential fixed differences between these species. Two regions containing nucleotide differences were identified and oligonucleotides were synthesized corresponding to the sequence present in the three D. americana lines. These species-specific primers were designated tra amF2 (5’cgc aat tct cg cag ctg) and tra amR (5’gca gct gga tga ggt ctg a). Each of these primers, used with the appropriate general primer, amplifies ~600 bp of either the upstream or downstream portions of the tra gene from D. americana (Figure 1). To obtain sequences of single alleles of tra, the two regions of the gene were amplified in separate reactions from DNA template of single F1 hybrids between D. americana (both D. a. americana and D. a. texana) and the V.46 strain of D. virilis. Sequences of tra were obtained from three different samples (McAllister and Charlesworth 1999): laboratory strains of D. a. americana maintained by the Drosophila Species Center, a sample (G96) of D. a. americana collected near Gary, IN, and a sample (LP97) of D. a. texana collected near Lone Star, TX. The same conditions as given above were used for PCR and direct sequencing. Editing of sequences was performed using SeqEd (ABI, Norwalk, CT). In the absence of a balancer-chromosome system for extracting wild chromosomes into homozygous lines, this procedure provides a quick and reliable method for generating an effectively haploid nuclear genome for obtaining a DNA sequence of a single allele.

**Sequence analysis:** To reveal patterns that may be limited to individual regions of tra, the gene was subdivided into its functional domains for many of the analyses (Figure 1). The tra gene has standard eukaryotic gene regions: exon 1, intron 1, exon 2, intron 2, and exon 3. However, the region of tra between the end of intron 1 and beginning of exon 2 is alternately spliced in males and females. This region is subdivided into two regions: the non-sex-specific protein (NSS) is the alternatively spliced region of the transcript that upon translation ends at a premature stop codon, whereas the untranslated region (UTR) begins after the stop codon of the non-sex-specific protein and ends at the junction with exon 2. These two regions were treated independently when possible and otherwise combined as the female-specific intron (FSI). The published sequence of D. virilis (X66528; O’Neill and Belote 1992) was used in many of the analyses; however, it was modified from the published version by reversing the
Alignments of the complete sequence data set of D. a. americana and D. a. texana was performed using SeqPup (D. Gilbert, Indiana University). Comparisons of the aligned data set and some of the analyses were performed using SITES (Hey and Wakeley 1997). For the entire sequence region and each of the functional domains (NSS and UTR treated separately), standard estimates of nucleotide diversity, $\pi$ and $\theta$, were calculated according to Tajima (1993). The estimate of $\pi$ is based on the average number of nucleotide differences from pair-wise comparisons among the sequences in each sample, and $\theta$ is based on the number of sites that exhibit variable nucleotides among the sequences in each sample. To determine whether the distribution of nucleotide variation within the samples was consistent with the neutral model, D statistics were calculated (Tajima 1989; Fu and Li 1993). Both statistics were calculated separately for synonymous, replacement, and intron changes and for all types combined. The sequence of D. virilis was used as the outgroup to determine the ancestral state for the Fu and Li (1993) test. To determine the level of sequence divergence that was present between D. a. americana and D. a. texana, the net number of pairwise differences ($d_{s}$) was estimated (Nei 1987, p. 276). This measure was estimated individually at changes at synonymous sites, replacement sites, and intron sites (including FSI) and for all of these sites combined. In the analysis of the polymorphism to divergence ratio for different types of changes (McDonald and Kreitman 1991), this ratio was determined for synonymous, replacement, intron (including FSI), and insertion/deletion (indel) changes among all of the D. a. americana and D. a. texana sequences combined into one "population" to identify polymorphic and fixed differences relative to the sequence of D. virilis.

A single sequence from each of six Drosophila species was used in the interspecific analyses (GenBank accession numbers in parentheses): D. melanogaster (M17478), D. simulans (X66930), and D. erecta (X66527) from the Sophophoran subgenus and D. hydei (X66931), D. virilis (X66528), and D. americana (G96.03, AF208140) from the Drosophila subgenus (see Figure 2). Each subgenus was analyzed independently as the high degree of divergence between the two makes for poor sequence alignment and reduces the number of conserved sites. DNA sequences were aligned so as to be consistent with the amino acid sequence alignment of O’Neill and Belote (1992). Likelihood-ratio tests were used to test different models of evolution for tra by application of the program codeml within the PAML software package (Yang 1997). We consider the difference in log-likelihood between models requiring different numbers of parameters to test the hypotheses of (1) a molecular clock for silent and replacement changes and (2) homogeneity of the nonsynonymous to synonymous substitution rate across each evolutionary lineage, both for the coding region as a whole and for each exon separately. Twice the difference in log-likelihood between models is $\chi^2$ distributed with the degrees of freedom equal to the difference in number of parameters. For the best model in each case (that for which the addition of no extra parameters provides a significant increase in likelihood) we also estimated the value of $d_{s}/d_{a}$ (the parameter $\omega$ of Goldman and Yang 1994) and its 2-unit support interval.

RESULTS

Intraspecific variability: Sequences of an 871-bp region of the transformer gene were obtained for single alleles from 4 standard laboratory strains of D. americana, 18 strains of the G96 sample of D. americana, and 9 strains of the LP97 sample of D. a. texana. All nucleotide positions that are variable among the 31 sequences are presented in Figure 3. Nucleotide position 1 corresponds to the 5' base of the start codon of the tra gene. A total of 64 segregating sites were identified, and two of these sites (261 and 716) have three different nucleotide variants present among the alleles. Four different indel polymorphisms are present among the sequences of D. americana. Although these indels occur in coding regions of the tra gene, none generate shifts in the reading frame. Nucleotide variants and indels are distributed very evenly among the sequences, and each allele represents a unique sequence (or haplo-
1714 B. F. McAllister and G. A. T. McVean

Figure 3.—Nucleotide and insertion/deletion polymorphisms among transformer sequences of D. americana. The functional regions of the transformer gene are indicated, and the positions of polymorphisms are based on the aligned sequences and relative to the first position of the first codon being position one. Each polymorphism is classified by type: intron (I), synonymous (S), replacement (R), or insertion/deletion (D). Alleles of D. a. americana from the laboratory lines are identified as D_am.01, alleles of D. a. americana collected from Gary, Indiana, are identified as G96.99, and alleles of D. a. texana collected from Lone Star, Texas, are identified as LP97.99. Dashes represent identity to D_am.01, and asterisks (*) represent deletions. Complete sequences have been deposited in GenBank under accession nos. AF208127-AF208157.

Belote (1992) demonstrated the sequence conservation of the region among different species of Drosophila, and we observe no nucleotide variation in this region among the 31 different alleles of tra in D. americana. Of the 4Nr, the scaled recombination rate, gives a value of 0.12. This is an order of magnitude higher than has been observed for gene regions in D. melanogaster and is indicative of a high rate of recombination in this region.

Measures of nucleotide variability for the entire sequenced region and each functional domain are summarized in Table 1. No divergence was present among the samples (see below), so the combined data for the three samples are presented. The numbers of nucleotide sites that are fixed in D. americana and different from the available sequence of D. virilis (O’Neil and Belote 1992) are also presented in Table 1. Very few nucleotide variants are segregating in exon 1 and intron 1 of the tra gene sequence, and these two domains contain only one nucleotide difference relative to the D. virilis sequence. The region of intron 1 near the non-sex-specific splice site is the target for binding of SXL and hence is necessary for alternative splicing (Sosnowski et al. 1989, 1994; Inoue et al. 1990). O’Neil and Belote (1992) demonstrated the sequence conservation of the region among different species of Drosophila, and we observe no nucleotide variation in this region among the 31 different alleles of tra in D. americana.
also observed in comparison with D. virilis, as a 16-bp sequence is present in D. virilis, but absent in the D. americana sequences. This indel apparently causes a frameshift in the resulting amino acid sequence and results in a larger polypeptide in D. virilis.

Many segregating nucleotide variants are present in the UTR, exon 2, intron 2, and exon 3 (Figure 3). In exons 2 and 3, both replacement and synonymous nucleotide substitutions are present, but the level of nucleotide variation per site is higher at synonymous positions, indicating some selective constraint. Overall nucleotide diversity is relatively homogeneous throughout these coding and noncoding regions (Table 1). This is unexpected given that selective constraint on the coding regions should maintain levels of nucleotide diversity that are lower in exons 2 and 3 than in the UTR and intron 2. All of these regions also exhibit a large number of fixed differences in comparison to the D. virilis sequence (Table 1).

One method for revealing the influence of selection on a coding sequence is by examining the distribution of segregating variants at that locus in a population and testing this distribution against the neutral model. Two measures of the deviation of segregating variation from neutral expectations are the D statistics of Tajima (1989) and Fu and Li (1993). Upon analysis of segregating variation over the entire tra gene, neither statistic provides evidence for a departure from neutral evolution (Table 1). Values of the D statistics are presented for all types of segregating variants combined. When synonymous, replacement, and intron sites were analyzed independently, no significant results were obtained either. The D statistics were also applied separately to the nucleotide variants present in each functional domain of tra, and the only significant result is the negative D statistics for the NSS domain. Lack of fit to the neutral model may be a spurious result due to few variants being present in the region; only four variants are present in this domain and all are present in single sequences. Alternatively, due to the involvement of this region in the sex-specific splicing reaction (O’Neill and Belote 1992), these may represent mildly deleterious mutations that are maintained at low frequency (Tajima 1989). Overall, the pattern of intrapopulation distribution of variability in the tra gene region is consistent with the neutral expectation; hence there is no evidence that positive Darwinian selection has recently influenced, or is currently influencing, nucleotide variation in this region of the genome.

If the tra gene is under selection at the incipient stage of speciation (Civetta and Singh 1998), more rapid divergence between incipient species may occur for nucleotide changes at replacement sites relative to changes at synonymous or intron sites. No fixed nucleotide differences between D. a. americana and D. a. texana were observed throughout the tra gene region, consistent with findings from other genes (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999). The net number of nucleotide changes (d) at synonymous, replacement, and intron sites was determined between the G96 sample of D. a. americana and the LP97 sample of D. a. texana (Table 2). These two subspecies are extremely closely related and may be actively in the process of incipient speciation (Throckmorton 1982). Replacement sites did show nearly twice as many net differences than did synonymous sites (0.00097 and 0.00047 net differences per site, respectively); however, the error associated with these small numbers is very large (Takahata and Nei 1985). This pattern of divergence between D. a. americana and D. a. texana is not consistent with tra playing a significant role in driving speciation.

Another method of detecting selection on a coding sequence is to examine the ratio of segregating variation relative to fixed differences between species at different types of sites (McDonald and Kreitman 1991). This method is based on the prediction of the neutral theory that the level of polymorphism and rate of divergence are both proportional to the neutral mutation rate and hence should be proportional to each other. The ratio of polymorphism to divergence should thus be equal

**TABLE 1**

<table>
<thead>
<tr>
<th>Differences</th>
<th>Segregating sites</th>
<th>Total</th>
<th>Silent</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Silent Replc.</td>
<td>Silent Replc.</td>
<td>( \pi / \text{site} )</td>
<td>( \theta / \text{site} )</td>
</tr>
<tr>
<td>Total, 871 bp</td>
<td>26</td>
<td>15</td>
<td>38</td>
<td>26</td>
</tr>
<tr>
<td>Exon 1, 34 bp</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Intron 1, 69 bp</td>
<td>1</td>
<td>−</td>
<td>1</td>
<td>−</td>
</tr>
<tr>
<td>NSS, 62 bp</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>UTR, 111 bp</td>
<td>8</td>
<td>−</td>
<td>9</td>
<td>−</td>
</tr>
<tr>
<td>Exon 2, 383 bp</td>
<td>7</td>
<td>10</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Intron 2, 59 bp</td>
<td>9</td>
<td>−</td>
<td>3</td>
<td>−</td>
</tr>
<tr>
<td>Exon 3, 153 bp</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

*P < 0.05; **P < 0.01.

† Differences are relative to the D. virilis sequence.
TABLE 2

Average pairwise differences within samples of \textit{D. a. americana} and \textit{D. a. texana} and net nucleotide difference between samples

<table>
<thead>
<tr>
<th>Type of change (sites)</th>
<th>Pairwise difference: \textit{amer-G96} (\textit{n} = 18)</th>
<th>\textit{tex-LP97} (\textit{n} = 9)</th>
<th>Net difference: \textit{G96}\textit{/LP97}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous (149)</td>
<td>4.45</td>
<td>5.53</td>
<td>0.07</td>
</tr>
<tr>
<td>Replacement (424)</td>
<td>5.05</td>
<td>3.50</td>
<td>0.41</td>
</tr>
<tr>
<td>Intron (322)</td>
<td>3.44</td>
<td>4.33</td>
<td>2.00</td>
</tr>
<tr>
<td>Total (895)</td>
<td>12.73</td>
<td>13.36</td>
<td>0.34</td>
</tr>
</tbody>
</table>

for different classes of site (e.g., silent and replacement). In contrast, unique selection pressures on certain classes of site will lead to heterogeneity in this ratio. All of the sequences from the different samples of \textit{D. americana} were pooled to determine the number of changes that are polymorphic or fixed relative to a single sequence from \textit{D. virilis}. In Table 3, the number of polymorphic and fixed changes is presented for synonymous, replacement, and intron sites, along with indels. There is significant heterogeneity in the ratio of polymorphic to fixed changes at these different types of sites (\(\chi^2 = 8.21, P < 0.05, 3\) d.f.). At synonymous sites, three times as many substitutions are polymorphic than are fixed, whereas this ratio is 1.5 at replacement sites, 0.94 at intron sites, and 0.5 for indels (Table 3). However, tests of heterogeneity between replacement sites and either synonymous sites or silent (synonymous plus intron) sites did not yield significant results. Heterogeneity appears to be a general trend of decreasing the polymorphism/fixed ratio across these four types of changes and is not attributable to one class of site, thus providing no evidence for positive selection acting specifically to fix amino acid replacements in TRA.

**Interspecific analysis:** Within the genus \textit{Drosophila}, the degree of sequence divergence at the \textit{tra} gene between the \textit{Drosophila} and \textit{Sophophora} subgenera is such that for several regions of the peptide there are few conserved sites, and sequence alignment is highly ambiguous. For this reason we have analyzed each subgenus separately. By using maximum likelihood methods of phylogenetic inference (Yang 1997) we can test alternative models for the evolution of the \textit{tra} gene. When three species are used, such tests are similar to the relative-rate test for detecting heterogeneity in the rate of evolution between species pairs by comparison to an outgroup species. The one difference is that by using both synonymous and nonsynonymous substitutions in the same analysis, we can detect heterogeneity between all lineages. In the tests presented we consider whether relaxing the assumptions of (1) clock-like accumulation of synonymous and nonsynonymous substitutions and (2) homogeneity in the ratio of nonsynonymous to synonymous substitution rates across lineages provides a significantly better fit to the data. Relaxing the first assumption is equivalent to allowing between-lineage variation in the mutation rate but maintaining selective constraint; relaxing the second assumption allows for differences in selective constraint, but assumes a constant rate of synonymous substitution. Tests are considered both for the coding region as a whole and for each functional exon separately. We do not consider intron sequences or the NSS exon; however, these are discussed elsewhere in the article.

Table 4 shows the increase in log-likelihood from the base model (molecular clock and single \(d_0/d_s\) ratio for all lineages) for models incorporating increased numbers of parameters. For the \textit{Sophophoran} subgenus we find no evidence for either deviation from clock-like accumulation of silent and replacement changes or heterogeneity in the \(d_0/d_s\) ratio between lineages. Differences in the pattern of substitution between \textit{D. melanogaster} and \textit{D. simulans} are not evident. In addition, the estimated value of the \(d_0/d_s\) ratio is very similar for all exons when considered separately (though branch lengths cannot be constrained to be identical between exons).

In contrast, within the \textit{Drosophila} subgenus we find evidence for differences between lineages in the \(d_0/d_s\) ratio, but no evidence for differences in the synonymous substitution rate between species. When the exons are considered separately, it is clear that exon 3 is almost exclusively responsible for this heterogeneity, and the estimated value of \(d_0/d_s\) is greater than one for the
TABLE 4
Likelihood-ratio tests of variation in the rate of molecular evolution and the $d_{ns}/d_s$ ratio in the transformer gene

<table>
<thead>
<tr>
<th>Subgenus</th>
<th>Region of gene</th>
<th>Homologous codons</th>
<th>Difference in log-likelihood from base model</th>
<th>$d_{ns}/d_s$ ratio for lineages under best model (2-unit support interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No clock</td>
<td>Lineage-specific $d_{ns}/d_s$</td>
</tr>
<tr>
<td>Sophophora</td>
<td>CDS</td>
<td>177</td>
<td>0.28</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>Exon 1</td>
<td>14</td>
<td>1.24</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Exon 2 (SS)</td>
<td>114</td>
<td>0.68</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>Exon 3</td>
<td>49</td>
<td>0.15</td>
<td>0.27</td>
</tr>
<tr>
<td>Drosophila</td>
<td>CDS</td>
<td>180</td>
<td>0.13</td>
<td>6.17*</td>
</tr>
<tr>
<td></td>
<td>Exon 1</td>
<td>12</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Exon 2 (SS)</td>
<td>117</td>
<td>0.06</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td>Exon 3</td>
<td>51</td>
<td>0.10</td>
<td>10.51**</td>
</tr>
</tbody>
</table>

* $P < 0.01$; ** $P < 0.0005$.

The base model assumes a molecular clock and a single $d_{ns}/d_s$ ratio for all lineages.

DISCUSSION

Rapid evolution of sex-determination mechanisms:

Developmental programs that direct sexual differentiation exhibit an unexpectedly high level of variability among organisms (Bull 1983; Marin and Baker 1998). Detailed analyses of the genetic mechanisms governing sex determination in flies, nematodes, and mammals have revealed patterns of hierarchical control that are superficially similar, but in which there is almost no conservation of specific genes and mechanisms. Given that sexual reproduction (and sexual differentiation) is present in all eukaryotic groups, such evolutionary flexibility is paradoxical and suggests that there exist evolutionary forces that drive turnover in the genetic mechanisms of sex determination.

It is of key interest to know the time scale for the evolutionary change in the mechanisms and components of sex determination and to understand the processes that influence this change. Genetic analysis of sex determination in model organisms shows the occurrence of large changes on the time scale (>544 million years) of the divisions of metazoan phyla (Cline and Meyer 1996; Marin and Baker 1998). There are suggestions, however, that the functions of the Drosophila sex-determining genes tra2 and dsx may be conserved throughout higher eukaryotes (Dauwalder et al. 1996; Raymond et al. 1998; Shearman and Frommer 1998). At the opposite end of the time scale (60 million years) within the Drosophilidae, the genetic components form-
ing the cascade leading to sexual differentiation are conserved, implying that the pathway was in place prior to the diversification of these species (O’Neil and Belote 1992; Bopp et al. 1996; Hertel et al. 1996; Chandler et al. 1997; Erickson and Cline 1998). Sex-determination mechanisms are, however, variable within the order Diptera (MRCA = 225 mya), and the function of Sxl is not conserved among other flies (Bull 1983; Müller-Holtkamp 1995; Marin and Baker 1998; Meise et al. 1998; Saccone et al. 1998). Therefore, the complete pathway leading to sex determination in Drosophila has been stable for a relatively short period of time.

Evolution of transformer: Despite the role of the tra gene product in somatic sexual differentiation in Drosophila, it demonstrates rapid evolution (in terms of amino acid divergence and insertions and deletions) to an extent that homology inferences between the Drosophila subgenera are highly ambiguous for much of the amino acid sequence. Transgenic analyses indicate that a highly divergent D. virilis gene sequence can rescue at least some of the function of the native D. melanogaster gene; hence its role in sex determination is at least partially maintained (O’Neil and Belote 1992). The combination of functional conservation yet rapid evolution suggests that understanding the evolution of tra may provide insight into the more general phenomenon of turnover of sex-determination mechanisms. A link between rapid evolution of sex-determination genes and turnover of sex-determination mechanism is also suggested by the high rates of evolution seen in key sex-determining genes in other model organisms, notably the Sry gene in mammals (Tucker and Lundrigan 1993; Whiffin et al. 1993) and the tra-2 gene of nematodes (Kuwabara 1996).

There are two possible explanations for high rates of substitution in amino acid sequences: multiple events of adaptive evolution or accumulation of neutral, or nearly neutral, mutations in proteins of low functional constraint. By means of both interspecific and intraspecific comparisons of patterns of evolution and polymorphism at the tra gene in Drosophila, we conclude that the rapid evolution is best explained by low functional constraints. Analysis of substitution rates in a phylogenetic context reveals no evidence for a ratio of nonsynonymous to synonymous divergence significantly greater than one in any lineage or region of the gene. Nor do we find evidence for widespread heterogeneity in the $d_{NS}/d_S$ ratio between lineages or regions of the gene. While this does not rule out certain scenarios of adaptive evolution, such patterns are expected from neutrality. The only exception is the evolution of exon 3 in the Drosophila subgenus. Here we find evidence for a relatively low $d_{NS}/d_S$ ratio in the lineages leading to D. hyde and the MRCA of D. americana and D. virilis and a relatively high ratio in the lineages leading to D. americana and D. virilis following species divergence.

Heterogeneity in the $d_{NS}/d_S$ ratio between lineages, but where the value of the ratio does not exceed 1, is predicted by models of accumulation of weakly deleterious mutations (Ohta 1992) and models of adaptive evolution in which a minority of sites are under positive selection and others are highly constrained. These models can be distinguished by analysis of within-species polymorphism.

Under the hypothesis that rapid evolution of tra is influenced by positive Darwinian selection, replacement polymorphisms are being driven to fixation in the population. Neutral replacement variants are also transient, but their frequency in the population is only influenced by genetic drift. Consistent with a nonadaptive model for the high rates of sequence divergence between species, we have observed a large number of amino acid changes as intraspecific polymorphisms within D. americana. We can also test this assertion by statistical means through comparison of the distribution of replacement and linked silent variants in the samples from D. americana, both to each other and to the distribution expected under neutrality. The distribution of amino acid variants and linked silent variants among sequences is consistent with the pattern expected for neutral and mildly deleterious variants. When patterns of polymorphism and divergence at different types of sites are compared we find evidence for heterogeneity between these classes (replacement, synonymous, noncoding, and indels), but no evidence that this heterogeneity is the result of selection acting on amino acid replacements. In short, analyses of patterns of polymorphism within D. americana provide no evidence that natural selection is driving the rapid evolution of the tra gene sequence.

Differences between D. melanogaster and D. americana: Presence of a large amount of nucleotide variability at the tra gene of D. americana is clearly different from the near absence of variation that was previously reported at this gene in D. melanogaster (Walsh and Schaeffer 1994). Nucleotide variants were identified at 64 sites in the tra gene region in D. americana, whereas only 2 nucleotide sites were variable in a sample of D. melanogaster. The level of nucleotide variability present at synonymous and intron sites in the tra gene of D. americana is similar to the variability that has been observed at other genes in this species (Hilton and Hey 1996, 1997; McAllister and Charlesworth 1999), which is consistent with the expectation of homogenous levels of neutral variation throughout freely recombining regions of the genome (Hudson et al. 1987). Nucleotide variability at tra in D. melanogaster is significantly lower than other genes (Walsh and Schaeffer 1994). Differences in variability at tra between these two species may be due to either selection in D. melanogaster or differences in local rates of recombination.

Location of the tra gene in the D. americana and D. melanogaster genomes indicates that there is a difference in the local recombination rate in these two species. In
D. virilis, the tra gene is located 15.5 map units proximal to the telomere of Muller's element D (chromosome 3 of the virilis group), which has a total length of 145 map units (Gubenko and Evgen'ev 1984; O'Neil and Belote 1992). This suggests that the genomic region containing tra experiences a normal rate of recombination. Observation of 31 haplotypes among 31 tra alleles supports this inference, because intragenic recombination is the source of these haplotypes (Hudson 1987; Hey and Wakeley 1997; Depaulis and Veuille 1998).

The rate of recombination in the tra gene region in D. melanogaster is apparently much lower, due to the gene being located two map units from the centromere of element D (chromosome 3L). This places tra in a region of below-average recombination within the genome of D. melanogaster (Kliman and Hey 1993). Furthermore, the overall chromosomal recombination rate in D. melanogaster would lead to a lower expected level of nucleotide variation in this region, relative to regions of normal recombination, through the action of background selection (Charlesworth et al. 1995; Hudson and Kaplan 1995). Recent selective sweeps at other genes in the tra gene region of D. melanogaster would also reduce sequence variation at tra (Begun and Aquadro 1992; Walthour and Scharff 1994).

**Low functional constraint in the tra gene:** Rapid non-synonymous evolution in the tra gene raises an important question. How can it be that such an important developmental gene has such low functional constraint? Some features within the amino acid sequence of the tra gene are clearly constrained, because synonymous sites exhibit a higher rate of change than replacement sites at both the intraspecific and interspecific levels. It is, however, notable that a highly conserved region of the tra gene is effectively noncoding, intron 1 and the NSS. In D. americana, few polymorphisms are present in the region and there is evidence of purifying selection. Furthermore, a stretch of 21 nucleotides at the 3’ end of intron 1 (total length 69 bases) is conserved among all Drosophila species that have been examined (O'Neil and Belote 1992). This region of the gene serves as the binding site of the SXL polypeptide (Sosnowski et al. 1994) and it presumably directs preferential splicing in the absence of SXL (O'Neil and Belote 1992).

This picture is very similar to that seen in the Sry gene of mammals and the tra-2 gene of nematodes (Tucker and Lundrigan 1993; Whitfield et al. 1993; Kuwabara 1996). In the SRY protein, the DNA-binding region known as the HMG box is both the only region that demonstrates conservation within mammals and also the location of almost all known Sry-related sex-reversing mutations in humans (Hawkins 1994). Likewise, two of the three known disruptive mutations in the TRA-2 protein, required for female development in nematodes, occur in conserved regions of the gene. In contrast, the average identity for TRA-2 between Caenorhabditis elegans and C. briggsae is only 43%, the most divergent protein identified in this species pair (Kuwabara 1996).

While there are many genes involved in the sex-determination process in each of these species, several of which are highly conserved, the similarities between these genes suggest that low functional constraint may be a general property of certain aspects of the sex-determination process. One possibility is that such low constraint arises naturally from the relatively rapid turnover of sex-determination mechanisms. Changes in the sex-determination mechanism, for whatever reason, should rely on genes being co-opted into the sex-determination pathway. If some of these arise by divergence from a duplicated gene, as has been suggested for Sry (Stevanovic et al. 1993), then one would expect limited pleiotropy and, as a consequence, lower functional constraint than average (Waxman and Peck 1998).

The authors are grateful for the input and support of Brian Charlesworth. We thank Jody Hey and the anonymous reviewers for comments on the manuscript. This work was supported by a National Science Foundation/Alfred P. Sloan Foundation Postdoctoral Fellowship in Molecular Evolution to B. McAllister. G. McVean is funded by the Natural Environment Research Council.

**LITERATURE CITED**


Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide


Communicating editor: J. Hey