Molecular Population Genetics of Sex Determination Genes: The \textit{transformer} Gene of \textit{Drosophila melanogaster}

C. Scott Walthour and Stephen W. Schaeffer

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received July 16, 1993
Accepted for publication December 17, 1993

\textbf{ABSTRACT}

The \textit{transformer} locus (\textit{tra}) produces an RNA processing protein that alternatively splices the \textit{doublesex} pre-mRNA in the sex determination hierarchy of \textit{Drosophila melanogaster}. Comparisons of the \textit{tra} coding region among \textit{Drosophila} species have revealed an unusually high degree of divergence in synonymous and nonsynonymous sites. In this study, we tested the hypothesis that the \textit{tra} gene will be polymorphic in synonymous and nonsynonymous sites within species by investigating nucleotide sequence variation in eleven \textit{tra} alleles within \textit{D. melanogaster}. Of the 1063 nucleotides examined, two synonymous sites were polymorphic and no amino acid variation was detected. Three statistical tests were used to detect departures from an equilibrium neutral model. Two tests failed to reject a neutral model of molecular evolution because of low statistical power associated with low levels of genetic variation (Tajima/Fu and Li). The Hudson, Kreitman, and Aguade test rejected a neutral model when the \textit{tra} region was compared to the 5'-flanking region of alcohol dehydrogenase (\textit{Adh}). The lack of variability in the \textit{tra} gene is consistent with a recent selective sweep of a beneficial allele in or near the \textit{tra} locus.

\textbf{T}he \textit{transformer} (\textit{tra}) gene in \textit{Drosophila melanogaster} encodes one of the proteins necessary for correct sexual differentiation of somatic cells in females (Baker and Belote 1983; Steinnmann-Zwicky et al. 1990). \textit{tra} regulates the expression of \textit{doublesex}, a gene whose male-specific form represses female-specific differentiation genes and whose female-specific form represses male-specific differentiation genes (Baker 1989). Loss-of-function mutations in \textit{tra} have no effect in chromosomal males (XY), but chromosomal females (XX) fail to express the female form of \textit{doublesex} and are transformed into sterile adult flies with male morphological characters such as sex combs, male-colored abdomen, external genitalia, and reduced gonads (Sturtevant 1945).

The sex-specific regulation of \textit{tra} is mediated by alternative splicing of its pre-mRNA (Boggs et al. 1987). The default splicing pathway of \textit{tra} used by both sexes encodes a truncated protein that has no effect on the \textit{doublesex} gene. A female-specific protein, \textit{Sex-lethal}, alternatively splices the \textit{tra} pre-mRNA to yield a message that encodes an RNA-binding protein (Figure 1) (Nagoshi et al. 1988; Sosnowski et al. 1989; Inoue et al. 1990). The transformer protein and a second RNA-binding protein, transformer-2, alternatively splice the \textit{doublesex} pre-mRNA in females (Amrein et al. 1988; Baker and Wolfer 1988; Burts and Baker 1989; Goralski et al. 1989; Inoue et al. 1992). The default \textit{doublesex} transcript produced in males encodes the protein that represses female development while the alternatively spliced mRNA of females encodes the protein that represses male development.

The importance of the transformer protein in sexual development suggests that its amino acid sequence should be highly conserved throughout the evolution of \textit{Drosophila}. This assumption was shown to be invalid by O’Neil and Belote (1992) who have demonstrated that the transformer protein has an unusually high degree of divergence when the coding regions were compared among five \textit{Drosophila} species. The data of O’Neil and Belote suggest that the transformer protein will be highly polymorphic within species because mutations in populations are the source of diversity between species.

The high level of protein divergence observed among \textit{transformer} sequences may be a result of neutral forces where amino acid substitutions accumulate by mutation and random genetic drift (Kimura 1983). Alternatively, the adaptive fixation of beneficial protein alleles over time may contribute to the accumulation of amino acid differences among species (McDonald and Kreitman 1991; Eanes et al. 1993). These two hypotheses may be distinguished by an examination of diversity in synonymous and nonsynonymous nucleotides within and between species of \textit{Drosophila}. The neutral fixation of new amino acid variants will have no effect on the fixation rate of linked neutral variation such as synonymous sites (Birky and Walsh 1988), while fixations of protein alleles by directional selection will remove linked synonymous variation by genetic hitchhiking (Kaplan et al. 1989; Stephan et al. 1992). We present here a nucleotide

Genetics 136: 1367–1372 (April, 1994)
sequence analysis of 12 tra genes within and between species of the melanogaster subgroup of Drosophila to determine the level of amino acid diversity in the transformer protein and to suggest the evolutionary forces that may explain the rapid divergence of the transformer protein between species.

**MATERIALS AND METHODS**

**Fly strains and genomic DNA preparation:** Ten strains of *D. melanogaster* were collected from localities near State College, Pennsylvania. Two strains with the designation "P" were collected during August or 1989 by Andrew G. Clark and his colleagues and eight strains with the notation "ST" were obtained in August of 1991. The 10 strains were made homozygous for a black box while sequences that are transcribed but not translated are shown as open boxes.

**Polymerase Chain Reaction (PCR) and nucleotide sequencing:** The nucleotide positions of the tra sequence are those of the previously published *D. melanogaster* sequence (Boggs et al. 1987). A 1,111-base pair fragment was amplified from each of the 10 strains of *D. melanogaster* (Sukai et al. 1988) (Figure 1), with the following oligonucleotide primers: the 5' primer begins at nucleotide -11 (5'GAGTTGGCGGGCACATTGCAAGGTS'); and the 3' primer begins at nucleotide 1100 (5'TAAATCGACCTGTTTGCAATGATG'). Single-stranded templates for nucleotide sequencing were generated with the α exonuclease procedure (Higuchi and Ochman 1989) and sequenced with eight oligonucleotide primers by the chain-termination method of Sanger et al. (1977). The eight oligonucleotide primers were separated by an average of 250 nucleotides and sequenced the tra regions on both strands. The overlapping sequences for each strain were assembled with the SEQMAN and SEQMANED programs (DNASTAR, Madison, Wisconsin).

The 10 tra sequences of *D. melanogaster* determined in this study were deposited in the GenBank/EMBL Data Libraries under the accession numbers L19464–L19470 and L19618–L19620. The sequences of previously published tra sequences from *D. melanogaster* and *Drosophila simulans* (GenBank/EMBL Data Library accession numbers M17468 and X66930, respectively) were also used in this study (Boggs et al. 1987; O'Neill and Belote 1992).

**Nucleotide sequence alignments and statistical analyses:** The 10 tra sequences generated in this study were aligned based on the published comparison of the *D. melanogaster* and *D. simulans* genes (O'Neill and Belote 1992). The aligned sequences were assembled with the Eyeball Sequence Editor [ESEE, version: 1.09; Cabot and Beckenbach (1989)]. Nucleotide sequences within the PCR primers were excluded from all analyses leaving a total of 1063 aligned bases. Any nucleotide site with two nucleotides present was defined as a segregating or polymorphic site. Nucleotide sites found in insertions or deletions assumed in aligned sequences were excluded from further analyses.

We estimated heterozygosity per nucleotide site for tra sequence domains with two methods. Nst (1987; equations 10.6 and 10.9) estimates nucleotide diversity, π, by determining the average number of differences between all pairs of sequences. An alternative measure of nucleotide variability, θ or L, is estimated from the number of segregating sites, S [Watterson (1975); equation 1.4a; Nei (1987) uses the notation, L, for the Watterson estimator θ]. The quantities π and L are both estimates of the neutral mutation parameter 4Np, where N is the effective population size and μ is the neutral mutation per nucleotide per generation. We also estimate a lower (L₁) and upper limit (L₂) on L that is a 95% confidence interval based on the probability distribution of S [Hudson (1990; equations 10, 11, and 12); Kreitman and Hudson (1991; equation 3)].

The tra region was tested for departures from an equilibrium neutral model with three statistical tests. The three tests were designed to determine if the within species nucleotide diversity in tra is consistent with the expectations of a neutral model. The Tajima (1989) test computes a test statistic, D, which is the difference of two estimates of heterozygosity per locus, M(S) and M(k). D should not differ significantly from zero under a strictly neutral model. The Tajima test statistic was estimated for the synonymous variation in the tra region within *D. melanogaster*.

We also used the statistic suggested by Hudson et al. (1987) (HKA test) to test for departures from neutral expectations. The neutral theory predicts that intraspecific polymorphism will be correlated to interspecific divergence (Kimura 1983). The HKA test rejects a neutral model if the ratio of polymorphism to divergence differs significantly among independent loci (Hudson et al. 1987), provided that one genetic locus evolves according to predictions of the neutral theory. The numbers of variable synonymous sites in the tra region were compared to the 5'-flanking region of alcohol dehydrogenase (Adh) in *D. melanogaster* (Kreitman and Hudson 1991). Tests were performed on synonymous and total nucleotide variation. The between species divergence data is determined from comparisons of each locus with its homolog in *D. simulans*. The comparison of the tra data with the 5'-flanking region of *Adh* probably reflects the most reasonable test of neutrality because this segment has the least functional constraint on its sequence and probably has not been influenced by positive Darwinian selection in its recent history. Comparisons of the Adh or Adh-Dup coding sequences with the tra region are likely to be inappropriate for the HKA test because levels of nucleotide diversity have been altered in each gene by genetic hitchhiking events associated with some form of positive Darwinian selection (Kreitman and Hudson 1991).

Fu and Li (1993, equation 32) have recently derived a test of selective neutrality that examines the number of mutations that occur on external vs. internal branches of a genealogy. The test uses within and between species data to estimate the
RESULTS

Nucleotide variation: Two of the 1963 nucleotide sites sequenced were polymorphic within *D. melanogaster* (Figure 2). Both segregating sites, nucleotide positions 312 and 831 in the DMEL sequence, occurred in synonymous positions. No amino acid polymorphism was observed within *D. melanogaster*. Estimates of $L$ and $\pi$ in each of the sequence domains can be found in Table 1. The lower and upper bounds on $L$ overlap for all the eight sequence domains in the *tra* region. Estimates of nucleotide diversity were quite low for all sequence domains compared to other genes that have been examined in *D. melanogaster* (Table 1) (Aquadro 1993). The nucleotide divergence in the *tra* region was much greater than the *D. melanogaster* and *Drosophila simulans* sequences were compared. Forty-two synonymous substitutions and 16 amino acid replacements have accumulated between the two species (Table 1, Figure 2).

Sequence length variation: One length variant was found among the 11 *D. melanogaster* *tra* sequences studied. The strain ST15 had a 5 base-pair insertion in the 5' flanking region of *tra* relative to the 10 other sequences. The inserted sequence, CATT, is identical to the five bases immediately 5' to the site of the insertion (G. S. Walthour and S. W. Schaeffer, unpublished data). Four deletions of sequence were observed in the *D. simulans* sequence relative to the *D. melanogaster* sequence. Deletions of 39 and 3 base pairs were found in the Female-Specific Exon 2, and Intron 2, respectively. Two deletions of one nucleotide occurred in the 3' flanking sequence (O’Neil and Belote 1992).

Tests of the neutral mutation hypothesis: The nucleotide data in the *tra* region fails to reject a neutral model of molecular evolution with either the Tajima (1989) or the Fu and Li (1993) tests (Tajima: $M(3) = 0.68, M(k) = 0.62, D = -0.29, P > 0.05$; Fu and Li: $\eta_i = 1, D = -0.45, P > 0.05$). The two test statistics in each case above were estimated from the information in two segregating sites observed within species. Thus, these results should be viewed with caution because of the low statistical power of these tests based on so few polymorphic sites. The results of the HKA test (1987) test stand in sharp contrast to the two tests discussed above. The comparison of the *tra* region with the 5' flanking region of *Adh* rejects an equilibrium neutral model (Table 2). The test results are significant because *tra* nucleotide diversity in synonymous and nonsynonymous sites within *D. melanogaster* was significantly lower than that expected given the genetic divergence between *D. melanogaster* and *D. simulans*. The McDonald-Kreitman (1991) test failed to reject a neutral model of protein evolution between species ($P = 0.99$) (Table 3).

DISCUSSION

The prediction that the *tra* gene would be polymorphic at the amino acid level within species was not supported by our data. The high level of conservation in amino acid sequence observed within species could re-
sult from strong functional constraints on the tra protein product or from a recent directional selection event in or near the tra locus. An important developmental protein such as tra might be expected to be highly conserved at the amino acid level because of its funda-
mental role in sex determination. If strong constraint on amino acid sequence was the main force restricting variation in
the protein, then amino acid sequence should be highly conserved between species and diversity in synonymous sites should be unaffected by changes in nonsynonymous positions (Birky and Walsh 1988). Sixteen amino acid replacements have accumulated in the tra protein between D. melanogaster and D. simulans. This is much less constrained than ADH, which has only two amino acid differences between the two species (Bodmer and Ashburner 1984). The most striking feature of the tra

\[ \text{TABLE 1} \]

<table>
<thead>
<tr>
<th>Region</th>
<th>m</th>
<th>S</th>
<th>D</th>
<th>( L_x )</th>
<th>L (SE)</th>
<th>( L_C )</th>
<th>( \pi ) (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'-Flanking</td>
<td>58</td>
<td>0</td>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.029</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Exon 1*</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.170</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Intron 1</td>
<td>78</td>
<td>0</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.023</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Exon 2 (m and f)</td>
<td>175</td>
<td>10</td>
<td>10</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Exon 2 (f)</td>
<td>101</td>
<td>13</td>
<td>13</td>
<td>0.000</td>
<td>0.000</td>
<td>0.017</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Intron 2</td>
<td>57</td>
<td>0</td>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.030</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Exon 3*</td>
<td>37</td>
<td>1</td>
<td>8</td>
<td>0.000</td>
<td>0.009</td>
<td>0.074</td>
<td>0.012 (0.013)</td>
</tr>
<tr>
<td>5'-Flanking</td>
<td>109</td>
<td>0</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.016</td>
<td>0.000 (0.000)</td>
</tr>
<tr>
<td>Summary</td>
<td>148</td>
<td>22</td>
<td>22</td>
<td>0.000</td>
<td>0.002</td>
<td>0.018</td>
<td>0.003 (0.005)</td>
</tr>
<tr>
<td>Noncoding</td>
<td>472</td>
<td>20</td>
<td>20</td>
<td>0.000</td>
<td>0.001</td>
<td>0.006</td>
<td>0.0004 (0.001)</td>
</tr>
</tbody>
</table>

\[ m, \text{number of effectively synonymous sites; } S, \text{ number of segregating sites; } D, \text{ number of divergent sites between the } D. \text{ melanogaster and } D. \text{ simulans sequences, DMEL and DSIM; } L_x \text{ and } L_C \text{, lower and upper bounds on } L \text{ for which there is a 2.5}\% \text{ probability of observing the same number or more extreme values of } S \text{; } \pi, \text{ nucleotide diversity determined from the number of pairwise differences (Nei, 1987) and the number of segregating sites (Watterson 1975), respectively; and SE, standard error. Exon 2 is shown in two parts, the non-sex-specific portion (m/f) and the female-specific coding region (f).} \]

\[ \text{TABLE 2} \]

<table>
<thead>
<tr>
<th>Region</th>
<th>Synonymous nucleotides</th>
<th>Total nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Polymorphism</td>
<td>Divergence</td>
</tr>
<tr>
<td>transformer</td>
<td>2 (9.3)</td>
<td>42 (34.7)</td>
</tr>
<tr>
<td>Adh 5'-flanking</td>
<td>30 (22.7)</td>
<td>78 (85.3)</td>
</tr>
<tr>
<td>Test</td>
<td>4.13</td>
<td>P &lt; 0.05</td>
</tr>
</tbody>
</table>

\[ \text{Note: Synonymous refers to synonymous substitutions in coding and noncoding regions. Total includes both synonymous and nonsynonymous substitutions. Numbers in parentheses are expected values for the Hudson, Kreitman and Aquadro (HKA) (1987) test. The row designated Test gives the results of the HKA (1987) test whose statistic approximates a chi-square distribution with one degree of freedom and the probability of the chi-square value, } P. \]

\[ \text{TABLE 3} \]

<table>
<thead>
<tr>
<th>Sites</th>
<th>Fixed differences</th>
<th>Polymorphisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsynonymous</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Synonymous</td>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>
variation over a large region (MAYNARD SMITH and HAIGH 1974; KAPLAN et al. 1989), which would make it difficult to know what gene was being selected. Alternatively, selective sweeps in regions with free recombination would have more localized reductions in heterozygosity.

E. C. KINDAHL and C. F. AQUADRO have estimated relative rates of recombination or coefficients of exchange (COE) on the third chromosome of D. melanogaster with methods described previously (BEGUN and AQUADRO 1992). The COE estimates they have obtained range from 0.001 to approximately 0.070, where a region with a COE of zero experiences little or no recombination while a region with a COE greater than zero experiences higher levels of recombination. They have estimated the COE for the tra gene region to be 0.015 (E. C. KINDAHL and C. F. AQUADRO, Cornell University, personal communication). The estimate of COE for tra suggests that recombination rates are low to moderate in the region. Thus, a selective sweep in or near the tra locus might be expected to effect a large to moderate sized chromosomal segment. We sequenced approximately 1 kilobase of the tra gene, but we do not know to what extent variation is reduced in the local region around the gene. Additional sequence from the 5' and 3' regions may delimit the region affected by the selective sweep in the tra region.

Are multiple adaptive fixations of beneficial alleles responsible for the protein evolution observed between species? The MCDONALD-KREITMAN (1991) test failed to reject a neutral model of molecular evolution, which suggests that the between species divergence from mutation and random genetic drift. The MCDONALD-KREITMAN test lacked sufficient power to reject the neutral hypothesis because we had one polymorphic site among the D. melanogaster tra sequences. It is possible that multiple selective sweeps have been responsible for the rapid evolution of the transformer protein, however, the recent directional selection event in or near tra prevent a strong inference from being drawn.

CHARLESWORTH et al. (1993) have considered the effect that slightly deleterious alleles have on levels of variation in the genome, the so called background selection model. They have shown that selection against slightly deleterious alleles maintained by mutation can lower genetic diversity at linked neutral sites. The effect of background selection is most pronounced in regions where recombination rates are low such as the centromeric or proximal regions of autosomes. Background selection is a less potent force in regions with free recombination because the few deleterious genes contribute to reductions in fitness levels. Background selection may contribute to reduced levels of heterozygosity in tra because the estimate of recombination rate in the region is low to moderate.

The HRA test assumes that the sampled populations are at equilibrium. If the equilibrium assumption is not valid, then the significant test results could suggest that the our local D. melanogaster population is not at equilibrium. BEGUN and AQUADRO (1993) found that populations of D. melanogaster collected from the United States and Zimbabwe differ significantly in genetic composition in a four-cutter survey of eight genes. Their data suggest that U.S. populations have lost variation compared to the Zimbabwe population, which is consistent with various population bottleneck scenarios. The reduced heterozygosity in tra may have resulted from a recent bottleneck that established U.S. populations. A bottleneck would reduce diversity in all loci within the genome, but other loci in D. melanogaster have not had the severe loss of variation as that observed in tra (BERRY and KREITMAN 1993). An examination of nucleotide diversity in African populations may clarify whether the tra gene has low diversity because of a selective sweep or a recent population bottleneck.

Why would the two other statistical tests (TAJIMA 1989; Fu and Li 1993) of selective neutrality fail to reject a neutral model? These two tests have little power to detect departures from a neutral model when the number of segregating sites is small. Without interspecies data, we would have concluded that transformer protein lacks diversity because of high selective constraint on the tra sequence. This highlights the importance of collecting between species data to control for differences in selective constraints among genes.

The data presented in this report suggest that a selective fixation of a beneficial allele occurred in or near the tra locus. Further data are necessary to show that tra was the source of the selected variation. These data are intriguing because they suggest that an important developmental gene may have been the target of adaptive evolution.

We would like to thank ELLEN L. MILLER for her technical assistance and ANDREW G. CLARK for two genomic DNAs and balancer stocks. We would also like to thank CHARLES F. AQUADRO, DAVID BEGUN, MARTHA HAMBLIN, MARTIN KREITMAN and an anonymous reviewer for helpful comments and discussions used in the preparation of this manuscript. This work was supported in part by grant GM42472 to S.W.S. from the National Institutes of Health.

LITERATURE CITED


Higuchi, R. G., and H. Ochman, 1989 Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Research 17: 5865.


Communicating editor: W. F. Eanes.