

Rapid Evolution of the Sex-Determining Gene, *transformer*: Structural Diversity and Rate Heterogeneity Among Sibling Species of *Drosophila*

Rob J. Kulathinal, Lara Skwarek, Richard A. Morton, and Rama S. Singh

Department of Biology, McMaster University, Hamilton, Ontario, Canada

While developmentally regulated genes are generally conserved, *transformer* (*tra*), a key locus involved in the regulation of sexual differentiation, is highly diverged between species of *Drosophila*. With an aim to understand its divergence between sibling species, we investigated *tra* sequence variation among members of the *Drosophila melanogaster* species complex, *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia*. In this species group, *tra* divergence is rapid yet clocklike and exhibits large differences in protein size. *D. melanogaster* contains a 13–amino acid tandem duplication, whereas *D. sechellia* possesses a 72–amino acid tandem duplication representing a 30% increase in total amino acid residues. We also found evidence of a nonrandom distribution of replacement substitutions and heterogeneity in substitution rates using clustering statistics and a codon substitution model. We show that *tra*'s rapid divergence in this species complex is the result of generally lower selective constraints around regions that encode arginine-serine (RS) domains and a significantly higher rate of substitutions around the insertion site of *D. sechellia*'s large duplication. The proximity of rapidly diverged regions to sites of nucleotide insertion suggests that higher local rates of mutation may provide a causal mechanism for *TRA*'s rapid divergence in this subgroup. A comparison of *tra* orthologs across the genus *Drosophila* suggest that *TRA* maintains an assortment of RS domains for proper sex determining function while much of the protein evolves relatively unconstrained.

Introduction

Early developmental processes have long been thought to be more conserved than late-acting processes, even between distantly related organisms. As a result, many classic phylogenetic inferences have been based on shared morphology from early ontogenetic stages (Gould 1977). With the availability of orthologous sequences from diverse taxa, biologists have begun to appreciate even more the extent of homology in development. Examples such as genes involved in embryonic pattern formation and cellular signaling have demonstrated the conservation of structure and function over a broad range of early acting developmental loci (Nusslein-Volhard 1994; Patel 1994; Chan and Jan 1999).

It is therefore surprising when an essential gene, expressed early in development, does not evolve in a typically conserved manner. The sex determining gene, *transformer* (*tra*), has previously been shown to possess the lowest sequence identity among known orthologous proteins between *D. melanogaster* and *D. virilis* (O'Neil and Beloté 1992). The primary function of *tra* is to act as a regulatory switch in the determination of sexual identity in each cell. Depending on the X:autosome ratio, *tra* is expressed as a functional protein (in XX females) or a truncated, nonfunctional protein (in XY males) (see fig. 1) via female-specific alternative splicing of *tra* premRNA controlled by the Sex-lethal protein, SXL. This switch is regulated by SXL binding to a polypyrimidine site in *tra*'s first intron (Sosnowski, Beloté, and McKeown 1989; Handa et al. 1999) (see fig. 1). In females, *TRA* associates with the transformer-2 protein, *TRA-2*, via its arginine-serine (RS) domains, forming part of a protein complex that binds to target *doublesex* (*dsx*) premRNA repeat elements (Inoue et al. 1992). Then *dsx* premRNA is

alternatively spliced in females to produce a female-specific product, whereas in males (without functional *TRA*), default splice sites result in a male-specific product. Each sex-specific protein regulates a host of downstream genes involved in sexual differentiation, including body size, genitalia development, mating preference, and pheromone production (Burtis and Baker 1989; Ferveur et al. 1997). The regulation of *fruitless*, which affects almost all aspects of male courtship behavior (Ryner et al. 1996), also involves *tra*.

So why has this important regulator of sexual differentiation evolved so rapidly? Two different aspects of *transformer* sequence variation—within versus between species—have been studied in the genus *Drosophila* and have suggested two contrary mechanisms of rapid *tra* divergence. In the first study, Walthour and Schaeffer (1994) observed drastically reduced levels of genetic variation in a natural population of *D. melanogaster*. After comparing levels of within species variation to that between species (Hudson, Kreitman, and Aguadé 1987), they rejected a neutral pattern of molecular evolution. Given that *tra* is located in a region of low to moderate recombination (Kliman and Hey 1993; Walthour and Schaeffer 1994), the authors suggested that directional selection via recent selective sweeps acting in or around the *tra* locus may have caused the pattern of low polymorphism in *D. melanogaster* and high divergence between species. In the second study, McAllister and McVean (2000) studied the species pair *D. americana* and *D. virilis* using population genetic and phylogenetic approaches. They concluded that high rates of neutral evolution may be sufficient to explain the high divergence between these species. The authors also tested the presence of a molecular clock on three published sequences of *D. melanogaster*, *D. simulans*, and *D. erecta* and found that a neutral evolutionary model could not be rejected in the three Sophophoran species.

In this paper, we evaluate sequence variation among all four sibling species of the *D. melanogaster* complex—the cosmopolitan species, *D. melanogaster* and *D.*

Key words: Sex determination, neutral evolution, selective constraints, RS domains, tandem duplication, developmental system.

E-mail: rkulathinal@oeb.harvard.edu.

Mol. Biol. Evol. 20(3):441–452. 2003

DOI: 10.1093/molbev/msg053

© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

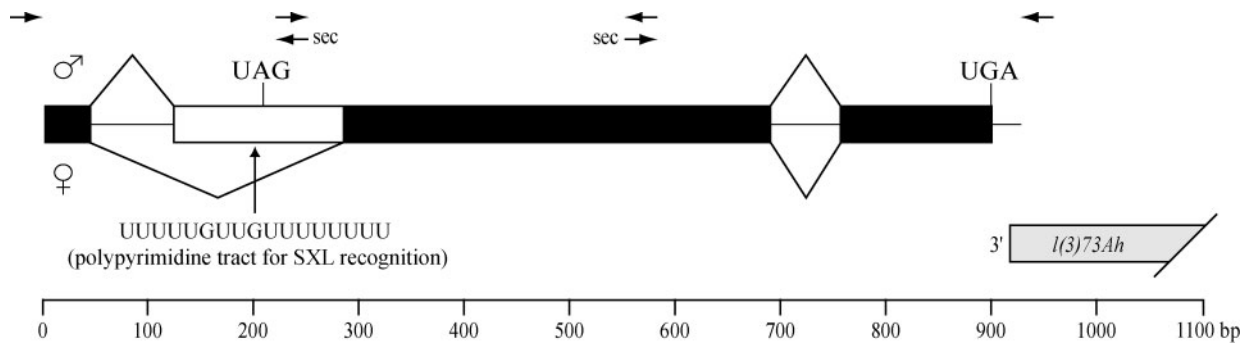


FIG. 1.—Gene structure of the *transformer* locus in *D. melanogaster*. Only regions sequenced in this study are displayed. Solid boxes represent exons and horizontal lines indicate the first and second introns and 3' flanking region. The untranslated region in male-specific exon 2 (with premature stop codon) is shown as an open box. Male-specific mRNA transcripts are shown above, and female-specific transcripts are below. Primers used in amplification and sequencing are indicated by horizontal arrows. Additional sets of primers used in *D. sechellia* are indicated (denoted as sec). Known functional constraints are shown below, including the polypyrimidine SXL recognition domain in females and neighboring 3' gene, which is encoded on opposite strand. Nucleotide numbering (i.e., position 1) commences at the first base pair sequenced, which corresponds to the first nucleotide of the start codon in exon 1.

simulans, and the island endemics, *D. mauritiana* and *D. sechellia*. The latter three species have diverged from each other within the last half a million years (Kliman et al. 2000). In light of its importance in the sex determination pathway, our goal is to understand both the pattern and mechanism of *tra*'s rapid evolution and to gain insight into the functional nature of this essential regulatory protein. Our intent is to advance beyond general arguments of selection versus neutrality in order to better appreciate how various regions of the *tra* locus have evolved in different lineages. We report that this developmentally regulated gene has undergone rapid divergence under varying selective constraints and has evolved drastic changes in protein size in this species clade. We conclude that *tra* evolution in *Drosophila* has been shaped by an overall lack of selective constraints in addition to selective pressures on maintaining an array of RS domains.

Materials and Methods

Fly Strains

Nine isofemale strains each of *D. melanogaster* and *D. simulans* representing various geographical populations were sequenced in order to sample global *transformer* variation. *D. melanogaster*. Lines from India, Hawaii, and Malaysia originated from the (now defunct) Bowling Green Species Stock Centre. Two lines from State College, Pennsylvania were kindly provided by Brian Lazzaro (Cornell University). One line from Italy was collected by Alberto Civetta (University of Winnipeg). Three homozygous lines from Zimbabwe were provided by the Andrew Clark laboratory (Cornell University; originally from David Begun, University of California, Davis). *D. simulans*: Lines from Colombia, Florida, California and a strain of unknown origin (Solway-Hochman 1088) were also obtained from the Bowling Green Stock Centre. John Roote (Cambridge University) supplied flies from Ethiopia and Madagascar. Lines from Kenya, Italy and a strain of unknown origin (S-148) were obtained from the Umeå Stock Centre, Sweden. *D. mauritiana*: Two lines originat-

ed from the Bowling Green Stock Centre and two lines were obtained from the Umeå Stock Centre. *D. sechellia*: Three lines originated from the Bowling Green Stock Centre and one line (3151) was supplied by Alberto Civetta.

DNA Extraction, PCR Amplification, and Sequencing

A single fly protocol was employed to extract genomic DNA (Gloor and Engels 1992). Briefly, one fly is macerated in 10 mM Tris-Cl (pH 8.2), 1 mM EDTA, 25 mM NaCl, and 200 mg/ml proteinase-K and left to stand at room temperature for 30 min. The preparation is then heated at 95°C for 3 min. A fragment approximately 1 kb long (depending on the species) was amplified by a Perkin-Elmer 480 thermal cycler using the extracted genomic DNA in 1X PCR buffer (MBI Fermentas), 0.2 mM dNTPs, 1.0 pM primers, 2.5 mM MgCl₂, and Taq polymerase (MBI Fermentas). For all four species, the same primers flanking the *tra* locus were used for amplification: 5'GTGCATCATTTAATTTCCAGCA3' and 5'TTTTAATGTACAAAACACACGAATG3'. The amplified product contains the full coding sequence along with *tra*'s two introns, untranslated male exon, and 53 to 54 nucleotides of the 3' flanking region (fig. 1). PCR product was cut out of a 2% agarose gel, DNA extracted, and then purified (Qiagen). Sequencing was performed using an ABI 377 Prism DNA Sequencer using the same primers as above. Two internal primers were also used (5'GAGGTTCGAGAACAGGATCGG3' and 5'CGTTCAGTCTGCGACTTCCG3') so that polymorphisms could be confirmed on both strands. Discrepancies between strands were not observed. Two singletons in *D. melanogaster* (a replacement and silent polymorphism) were reconfirmed through independent DNA extraction, amplification, and sequencing. The large insertion found in *D. sechellia* made necessary the construction and use of a third set of internal primers (5'TAATGCGCAGTTGAGAGTCC3' and 5'GAAGTCGCAGCAGTGAACG3') for amplification and sequencing. The unusually large sequence found in *D. sechellia* was verified separately by three

researchers—each independently extracted, amplified, and sequenced separate lines of this species. Approximate primer locations are shown in figure 1.

Sequence Analysis

In addition to the lines described above, we obtained 11 *D. melanogaster* sequences (O'Neil and Beloté 1992 [accession number M17478]; Walthour and Schaeffer 1994 [accession numbers L19464 to L19470 and L19618 to L19620]) and a *D. erecta* sequence (Walthour and Schaeffer 1994 [accession number X66527]) from GenBank to include in the melanogaster subgroup analyses. Sequences were labelled according to previously published designations. The published sequence of *D. simulans* (O'Neil and Beloté 1992 [accession number X66930]) was excluded from all analyses because six singletons (three of which cause nonsynonymous changes) were not observed in any of our nine *D. simulans* sequences. Although nucleotide diversity is quite high in *D. simulans*, such a high frequency of unique polymorphic sites originating from a single sequence suggests that this sequence may be in error.

Sequences from members of the subgenus *Drosophila* were also used, including *D. virilis* (O'Neil and Beloté 1992 [accession number X66528]) and 31 sequences from *D. americana* (McAllister and McVean 2000 [accession numbers AF208127 to AF208157]). Nucleotide and protein sequences were aligned using ClustalW (Thompson, Higgins, and Gibson 1994). Two estimates of nucleotide diversity, θ , determined from the number of segregating sites in a sample of genes (Watterson 1975), and π , the average pairwise difference between haplotypes (Nei 1987), were calculated using DnaSP v3 (Rozas and Rozas 1997). Both are estimates of $4N_e\mu$ under neutral expectations. HKA tests used the published 5'-flanking region of *Adh* as a reference neutral locus (Kreitman and Hudson 1991). Positional heterogeneity of amino acid variation was assessed using the broken stick model of Goss and Lewontin (1996). Clustering of variable sites is measured by three statistics: L_{\max} , the maximum fractional interval length; $\text{Var}(L)$, the variance in fractional interval length; and Q , a modified variance statistic. (A fourth statistic, L_{\min} or minimum interval length, is not accurate when a large fraction of the sites are variable, as is the case for *tra*, and was not used.) A computer program, Het2, was kindly provided by Peter Goss (Harvard University). In order to specify which region(s) possessed significant differences in rates of nucleotide substitution, we employed the permutation method of Hartmann and Golding (1998). Indels were removed from aligned coding regions and nucleotides were divided into subsets that comprise sites that are more prone to amino acid change (first and second codon positions, nondegenerate sites) as well as sites that serve as internal controls (third codon position, all nucleotide sites). Nondegenerate sites were chosen from a *D. melanogaster* sequence. Topologies with branch lengths were obtained using DNAML, (Phylip package v3.5c) (Felsenstein 1993) for each dataset. Sliding windows of various length then surveyed the sequence for regional rate heterogeneity via maximum likelihood

(code kindly provided by Brian Golding, McMaster University).

To test evolutionary models of constant versus variable rates of substitution, molecular clock versus no clock, and the variability of d_N/d_S among lineages as well as codon sites, we used a maximum likelihood approach implemented by the program codeml in the PAML package v3.0b (Yang 1997). Under each model, estimates of the parameter, $\omega = d_N/d_S$, a measure of a protein's selective constraint, were obtained. Models were evaluated statistically by likelihood ratio tests. Pairwise estimates of d_N and d_S , nonsynonymous and synonymous substitutions per site, were calculated using the method of Nei and Gojobori (1986). To compare TRA divergence with that of other known proteins in the *D. melanogaster* species complex, we compared levels of divergence in the protein coding regions of loci that were found in the NCBI database for all four species of the *D. melanogaster* complex. These loci can be found in online Supplementary Material. Only the portion aligned in all four species was compared when partial sequences were available.

Results

Nucleotide Polymorphism in *D. melanogaster* and Its Sibling Species

The *transformer* locus was sequenced from the first position of its start codon to its 3' flanking region (fig. 1). In a previously published study on *tra* variation, Walthour and Schaeffer (1994) found two segregating sites, both silent, out of 1063 nucleotides spanning the *tra* locus in 10 sequences from a single North American population (Pennsylvania). One of these polymorphisms was a singleton, whereas the other was found at an intermediate frequency. In order to survey this species' allelic diversity, we sampled nine sequences from global populations of *D. melanogaster*. Among 926 nucleotide sites, we observe two polymorphic silent sites and one replacement polymorphism, which is only present in an Asian population. Thus, genetic variation at the *tra* locus is reduced not only in the Pennsylvanian population but also throughout the species range. In total, only four variable sites constituting six unique haplotypes, three of which vary in exons, were detected from a collection of 20 *D. melanogaster* sequences. Haplotype diversity was quite low among sequences from the Pennsylvanian population study ($H = 0.15$) while slightly higher among the worldwide sample ($H = 0.25$).

Two measures of nucleotide diversity, θ and π , were calculated in order to compare the amount of variation found at the *tra* locus in *D. melanogaster* with that found in other species, as well as to published values for other loci (table 1). In *D. melanogaster*, nucleotide diversities from *tra* are smaller than the average value calculated for a sample of 24 loci reported by Moriyama and Powell (1996) and almost an order of magnitude lower than its ortholog in *D. simulans*. These differences are observed for both synonymous and replacement sites, as well as noncoding regions of the *tra* locus. Nucleotide variation is absent in both introns and 3' flanking region and nearly

Table 1
Comparison of transformer Nucleotide Diversity Among Species of the *Drosophila melanogaster* Complex

	Coding Regions										Noncoding Regions															
	Sequence Summary					Segregation Sites		Total Sites			Nonsynonymous Sites			Synonymous Sites			Intron 1 and Untranslated Male			Intron 2			3' Flanking Region			
	N	H ^a	L ^b	S ^c	S ^c	r ^d	s ^e	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ	
<i>D. melanogaster</i>																										
<i>tra</i> (1 pop.) ^f	10	3	926	2	1	0	1	0.79	0.60	0.00	0.00	0.00	3.15	2.39	0.81	1.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>tra</i> (global)	9	5	926	3	2	1	2	1.92	1.79	0.45	0.79	6.32	4.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
<i>tra</i> (combined) ^g	20	6	926	4	1	2	2	1.36	1.42	0.22	0.63	4.81	3.81	0.40	1.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average (24 loci) ^h	—	—	—	—	—	—	—	4.02	4.03	NA	NA	13.5	13.5	—	—	—	—	—	—	—	—	—	—	—	—	—
<i>D. simulans tra</i>	9	8	883	29	4	12	9.51	10.6	3.50	3.50	NA	28.2	32.8	11.0	11.9	29.8	40.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average (16 loci) ^h	—	—	—	—	—	—	7.83	7.99	NA	NA	30.4	31.1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
<i>D. mauritiana tra</i>	4	4	884	13	4	4	7.55	7.91	5.26	5.26	0.63	14.8	16.1	8.06	8.80	9.43	10.3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average (4 loci) ⁱ	—	—	—	—	—	—	4.51	4.37	0.61	0.61	20.1	19.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—
<i>D. sechellia tra</i>	4	2	1106	1	0	1	0.64	0.70	0.85	0.92	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average (4 loci) ⁱ	—	—	—	—	—	—	0.26	0.30	0.23	0.23	0.55	0.55	—	—	—	—	—	—	—	—	—	—	—	—	—	—

NOTE.—π is calculated as in Nei (1987) and θ is calculated as in Watterson (1975); both are calculated per base pair. Corrections for multiple hits were not performed. Nucleotide diversities are multiplied by 10⁻³. Noncoding regions sequenced in *tra* include only intronic, untranslated male exon and 3' regions comprising 328 to 332 nucleotides in the sibling species. Values in square brackets denote pooled diversity measures in introns and 5' and 3' ends as measured by Moriyama and Powell (1996). NA indicates data not available.

^a Number of distinct haplotypes.

^b Length of sequence excluding sites with alignment gaps.

^c Total number of segregating sites.

^d Number of replacement polymorphic sites.

^e Number of synonymous polymorphic sites.

^f Data from previously published population sample (Walthour and Schaeffer 1994).

^g Includes sequences from this study, the aforementioned population study, and a previously published sequence (O'Neil and Beloté 1992).

^h Values determined by Moriyama and Powell (1996).

ⁱ Calculated from the coding region of the following loci available from GenBank: *nillo*, *period*, *yr-2*, and *zeste*.

absent in the untranslated region of the male-specific exon (table 1).

Nucleotide variation among nine global strains of *D. simulans* is almost an order of magnitude greater than *D. melanogaster* at the *tra* locus and is similar to that reported for other loci from *D. simulans* (table 1). Haplotype diversity is also high ($H = 0.89$) compared with *D. melanogaster*. Four replacement and 12 synonymous polymorphic sites were found in the coding region. Singletons made up almost half of the total polymorphic sites.

The nucleotide diversity of *tra* is also relatively high in *D. mauritiana* (table 1). A total of 13 variable sites were distributed among four sequences and each sequence sampled represented a unique haplotype ($H = 1.0$). Variability is comparable between exonic versus intronic regions, except for the complete absence of substitutions in the 3' flanking region, which was common to all species sampled. In the coding regions, four replacement and four synonymous polymorphisms were observed. Compared with the overall level of coding region variation observed from four other loci sequenced in *D. mauritiana*, *tra* locus diversity is significantly higher ($P < 0.01$) than the species average. This is entirely due to the excess of replacement polymorphisms—nonsynonymous diversity is almost an order of magnitude greater in *tra* compared with other *D. mauritiana* loci. In contrast, among four lines of *D. sechellia*, only one polymorphism was observed. This replacement singleton is situated in one of the tandem repeats unique to *D. sechellia*. The low level of polymorphism found at this locus, however, was statistically no different ($P > 0.10$) from levels found at other sequenced loci in *D. sechellia* such as *asense*, *period*, *yp2*, and *zeste* (table 1).

We further examined whether within and between species variation adheres to a neutral model. According to neutral expectations, similar d_N/d_S ratios should be found within and between species (McDonald and Kreitman 1991). In all cases, we did not detect any deviation from this expectation among species of the *D. melanogaster* complex (data not shown). Neutrality was not rejected even in pairwise species comparisons involving *D. mauritiana* which possessed a higher within-species d_N/d_S ratio (four replacement versus four synonymous polymorphisms, table 1). HKA tests, which simply compare within versus between species variation (Hudson, Kreitman, and Aguadé 1987), also failed to reject a neutral model with one exception. Only *D. melanogaster* revealed significant differences between expectations based on polymorphism and divergence. Using the complete *D. melanogaster* data set ($N = 20$ sequences) against a consensus *D. simulans* sequence, HKA tests revealed significant differences in both silent site comparisons ($\chi^2 = 4.36$, $P = 0.04$) and total site comparisons ($\chi^2 = 5.36$, $P = 0.02$).

Divergence and Constraints Among Species of the *melanogaster* Subgroup

Out of 16 loci which have been sequenced in all four species of the *D. melanogaster* complex (Supplementary

Material), *tra* ranked second (after *Acp26Aa*) in average amino acid divergence, d_N , between *D. melanogaster* and its three sibling species. Synonymous divergence, d_S , on the other hand, ranked fifth. In terms of species pair divergence between the three sibling species of the *D. simulans* clade, *D. sechellia*, and *D. mauritiana* possessed the greatest differences (six fixed silent and six fixed replacement differences), particularly in nonsynonymous substitutions. The *D. simulans*–*D. mauritiana* comparison reveals a relative deficiency of fixed differences (one fixed silent and one fixed replacement difference) as many of the polymorphisms remain shared between these species. Due to the presence of these shared ancestral polymorphism between *D. simulans* and *D. mauritiana*, monophyletic lineages were not statistically supported (data not shown).

The constancy of *tra* divergence rate between members of this species complex was supported using a codon-based maximum likelihood approach (Yang 1997). Evolutionary models which utilized a molecular clock were not significantly different from models employing free rate parameters between species lineages (table 2). In other words, models which allowed rates to vary between branches did not offer a significant improvement to constant rate models. This was the case whether we used *D. erecta* or *D. melanogaster* as the outgroup, or whether we employed representative sequences ($N = 4$) or all sequences ($N = 37$) with estimated substitution rate parameters specific (or “local”) to each species (table 2). Different topologies were also compared and showed similar results.

The parameter $\omega = d_N/d_S$ was estimated using a codon-based substitution model and provides evidence that *tra*'s evolution is relatively unconstrained among species of this complex. Only three loci (*Acp26Aa*, *Acp70*, and *asense*) of 16 genes sequenced in all four species of the complex, revealed similar or lower levels of selective constraints (Supplementary Material). Over all *tra* coding regions, ω was estimated as 0.20 (table 2). The inclusion of *D. erecta* increased ω in all evolutionary models, including ones which factored in site and lineage variation. However, these models did not produce significantly better likelihoods. A null constant rate model maximized the likelihood at $\omega = 0.32$. To localize the region(s) of the gene with the least selective constraints, we subdivided the coding region into its three exons. Using the four sibling species with or without the inclusion of *D. erecta*, we found that all exons had moderate functional constraints, with exons 2 and 3 bearing slightly less constraints (table 2).

Duplications and Protein Size Evolution

TRA has dramatically diverged in size between the *D. melanogaster* sibling species. These differences in amino acid sequence length—a fixed 13–amino acid duplication in *D. melanogaster* as well as a 74–amino acid duplication in *D. sechellia*—are shown in figure 2. One interesting feature is that both species-specific sets of repeats are tandemly arranged, suggesting common mechanisms of origin. The 222-bp duplication in *D. sechellia* corresponds to a region near the N-terminus, and both repeats lie

Table 2
Likelihood Ratio Tests and d_N/d_S Estimates for *transformer* Using Codon Substitution Models

Species Sequence Subset	$2\Delta l$					Estimate of $\omega = d_N/d_S$ Under Null Model (Model of Best Fit)
	Codon Length	Site Variation		Lineage Variation		
		Constant vs. Variable Rate of Substitution	Constant vs. Variable d_N/d_S	Molecular Clock. vs No Clock	Constant vs. Variable d_N/d_S	
mel/sim/sec/mau (N = 4)						
Coding region	183	10.6**	2.11	0.81	2.06	0.20 (0.21)
Exon 1	13	1.27	0	2.22	1.06	0.12
Exon 2	120	7.28*	4.91	2.14	6.24	0.24 (0.23)
Exon 3	49	1.97	0	2.41	3.28	0.24
mel/sim/sec/mau (N = 20/9/4/4)						
Coding region	183	37.7**	12.0*	5.44 ^a	2.37 ^a	0.21 (0.19)
ere/mel/sim/sec/mau (N = 5)						
Coding region	176	2.75	5.41	0.52	4.50	0.32
Exon 1	13	1.49	2.07	3.97	2.03	0.24
Exon 2	114	2.08	3.91	0.26	6.40	0.34
Exon 3	49	3.89	0.62	1.90	4.71	0.33

NOTE.—A representative sequence from each species was randomly selected (see *Supplementary Information*) except in the one data set where all available sequences were used ($N_{\text{mel}} = 20$, $N_{\text{sim}} = 9$, $N_{\text{sec}} = 4$, $N_{\text{mau}} = 4$). A *D. mauritiana*–*D. sechellia* split from *D. simulans* was modeled (a trichotomy corroborated all results). Codon frequencies were directly calculated from nucleotide frequencies, and the transition/transversion ratio was estimated from the data. Each model is tested against its subset null model. $2\Delta l$ is twice the difference of the log-likelihood value and used in the likelihood ratio test. * $P < 0.05$. ** $P < 0.01$. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*; ere, *D. erecta*.

^a In this case, “local” substitution rates were assigned by estimating a single rate parameter for all sequences from a given species.

adjacent to each other. Immediately 3' to *D. sechellia*'s second repeat are *D. melanogaster*'s two tandemly arranged repeats (fig. 2). Another feature of interest is that each of the inserts contains regions that code for arginine-serine (RS) domains. One such domain is found in the *D. melanogaster* duplication, whereas in *D. sechellia*, two large RS-rich regions are found (fig. 3). In both cases, basic amino acid regions are also present (fig. 3).

Distribution of Variable Sites and Substitution Rates

Amino acid variation among species of the *D. melanogaster* complex appear to be clustered in regions near the RS domains and in a C-terminal segment of the coding region (fig. 2). To test if replacement substitutions are randomly distributed, we used the broken stick model of Goss and Lewontin (1996). The distribution of amino acid sites that are variable within the *D. melanogaster* complex is significantly nonrandom using two of the statistics recommended by Goss and Lewontin (table 3). On the other hand, when *D. erecta* is included with the four sibling species, the distribution of variable amino acid sites is not significantly different from random expectations (table 3). In contrast to replacement sites, variable synonymous sites are not distributed significantly different from random expectations, with or without the inclusion of *D. erecta* sequence.

While the broken stick model simply tests the observed distribution of variable sites against a random distribution, a sliding window permutation analysis identifies the regions that significantly differ in substitution rate (Hartmann and Golding 1998). Using either non-degenerate nucleotide sites ($n = 333$ bp) or a data set with first and second codon positions ($n = 364$ bp), only one region possessed a significantly different substitution rate

($N = 37$ sequences, $P > 0.05$). This region included several windows of moderate size (<20 nucleotides), all of which spanned the cluster of substitutions surrounding the putative *D. sechellia* insertion site (open box in fig. 2). Similar results were obtained either using topologies with interspersed sequences of *D. mauritiana* and *D. simulans* or by constraining each species to have a monophyletic origin. When data sets with only third positions sites ($n = 183$ bp) or all coding sites ($n = 549$ bp) were applied, there appeared no regions with significantly different substitution rates.

Rate heterogeneity among sites was also supported using maximum likelihood and a codon-based substitution model (Yang 1997). These analyses differ from the previous broken stick and permutation tests as they test for differences in substitution rate between codon sites as opposed to the regional clustering of substitutions. We observed that a model which incorporated rate heterogeneity between sites (allowing for three separate categories of substitution rates for the discrete-gamma distribution) fit the data significantly better than a model that used a single category of substitution rate when tested on these four species ($2\Delta l = 10.6$, $P < 0.01$). When *tra* was subdivided into its exons, the second exon alone gave significant results (table 2). Using all available *tra* sequences ($N = 37$) from the melanogaster subgroup, it was found that a variable d_N/d_S between sites model produced a significantly better likelihood ($2\Delta l = 37.7$, $P < 0.01$) (table 3).

The observed heterogeneity of substitution rate may be specific to members of the *D. melanogaster* complex since when *D. erecta* sequence is included, rate heterogeneity no longer significantly improves the likelihood (table 2). It also appears evident from the transformer protein sequence alignment of these five species, that regions which appear well conserved in the *D. melanogaster* complex have been substituted in the *D. erecta* protein

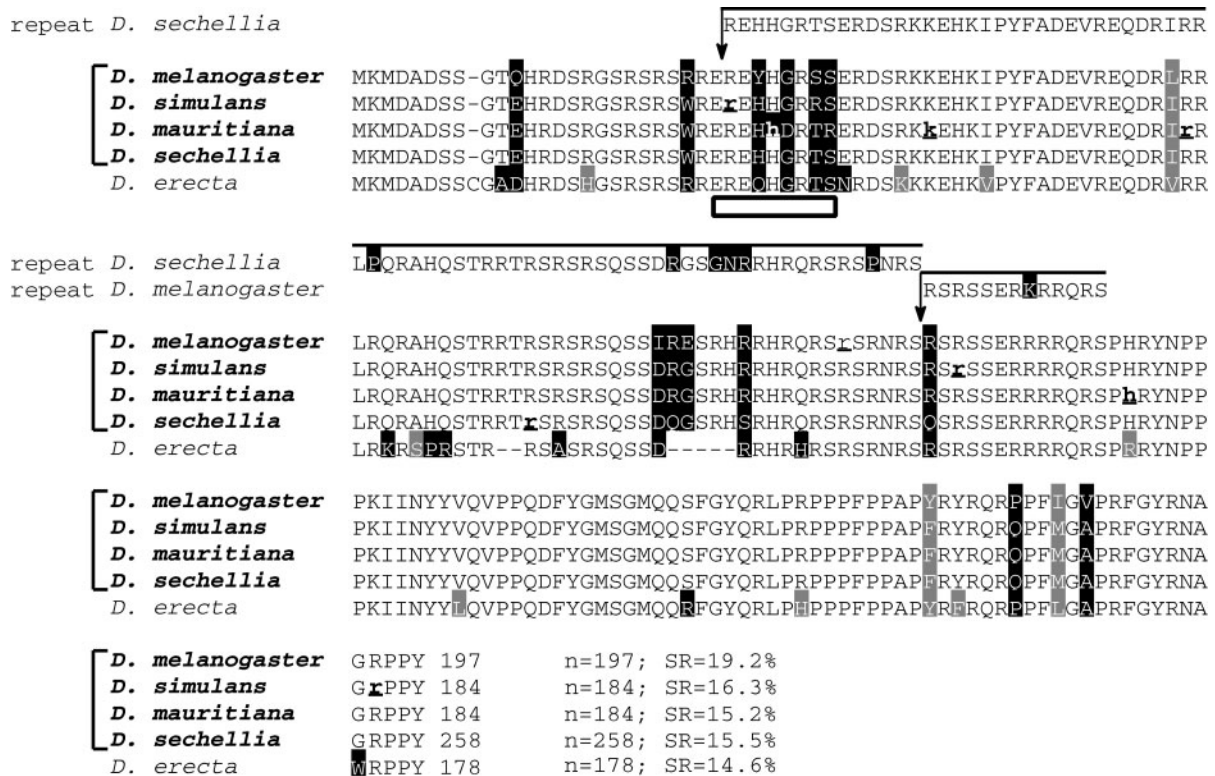


Fig. 2.—Protein alignment of the *transformer* protein in the *D. melanogaster* subgroup. Variable sites in the *D. melanogaster* complex are shaded in all sequences. Variable sites found solely in *D. erecta* are highlighted only in its sequence. Lightly shaded sites denote similar amino acid substitutions. A consensus sequence from each species was obtained by choosing at each site amino acids that were represented in the majority of lines. Replacement polymorphisms are highlighted, underlined, and denoted in lowercase boldface. Indels are indicated by dashes. Both sets of repeats are tandemly arranged and shown above consensus sequences. Insertion sites are indicated. Differences between repeats are shaded accordingly. Region showing a significant higher substitution rate by permutation test (see text) is shown below alignment as an open box. n = number of amino acids in TRA. SR = percentage of TRA consisting of serine-arginine (or arginine-serine) dipeptides.

(fig. 2). This does not mean, however, that these regions are not constrained. In fact, substitutions in the C-terminal segment for which amino acid variability was not found within the *D. melanogaster* species clade are relatively conservative amino acid replacements in *D. erecta* (see fig. 2), suggesting that it may represent a functionally important domain.

Discussion

Evolution of *tra* in the melanogaster Subgroup

Compared with other loci sequenced among species of the *D. melanogaster* complex (Moriyama and Powell 1996; Kliman et al. 2000; Supplementary Material), *transformer* is highly variable. Previous studies on *tra* variation have suggested both selective (Walthour and Schaeffer 1994) and neutral (McAllister and McVean 2000) modes of evolution. Our findings generally support a high rate of neutral evolution among these sibling species and furthermore direct our attention to regions within the *tra* locus, which may be under higher rates of neutral mutation.

Sibling species of the *D. melanogaster* species complex offer an attractive model system, as many loci have been surveyed for both polymorphism and divergence and each species' effective population size has

been satisfactorily estimated (Kliman et al. 2000). Sequence variation at the *tra* locus in the two sibling species *D. simulans* and *D. mauritiana* revealed relatively high levels of polymorphism consistent with other conspecific loci. The low diversity found in *D. sechellia* supports previous studies, indicating that it has a small effective population size (Hey and Kliman 1993; Kliman et al. 2000). This finding is consistent with the discovery of a large nucleotide insertion in the coding region of *D. sechellia*, which suggests that *tra* can accommodate major changes in protein structure. The insertion in *D. sechellia* may be mildly deleterious in its effect on fitness, as such mutations are more likely to be fixed due to increased drift in species with small population sizes (Ohta 1973).

But while the divergence of *tra* in this species subgroup reveal features consistent with a neutral model and low functional constraint, how do we explain the reduced variation in *D. melanogaster*? Our results confirm that *tra* has low levels of polymorphism, relative to other conspecific loci, across a global sample of *D. melanogaster*, and HKA tests performed on species-wide variation suggest a nonneutral model of evolution, consistent with Walthour and Schaeffer's (1994) previous results. In fact, the recombinational landscape in *D. melanogaster*, *D. simulans*, and *D. mauritiana* correlates rather well to the observed levels of species diversity at the *tra* locus,

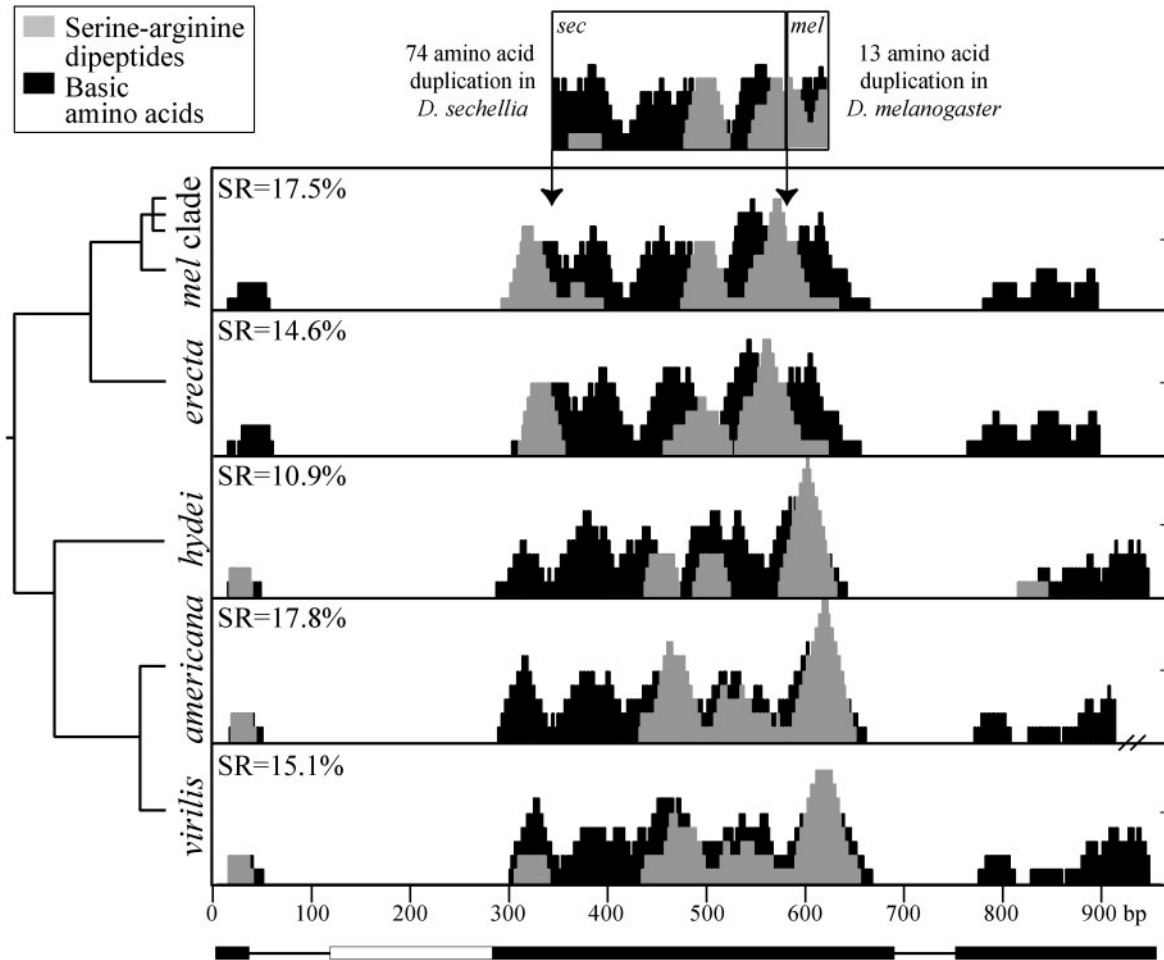


FIG. 3.—TRA amino acid composition in different species of *Drosophila*. A consensus sequence of *D. melanogaster* and its three sibling species was compared with *D. erecta*, *D. hydei*, *D. americana*, and *D. virilis*. Duplicated regions of *D. sechellia* and *D. melanogaster* aligned above. Their phylogenetic relationship is indicated on the left. *tra* gene structure is indicated below and as described in figure 1. Full coding sequence was not available for *D. americana*. Graphs indicate the fractional score of arginine-serine (RS) dipeptides (shown in gray) or basic amino acids (arginine, histidine, or lysine [shown in black]) along the *transformer* locus. Scores are calculated as the fraction of RS dipeptides or basic amino acids within a sliding window of 11 amino acids and are indicated at each window's midpoint. Values range from 0 to 1 and scores equal or greater than 0.5 represent a region of the protein with RS dipeptides or basic amino acids constituting a majority. SR = percentage of TRA consisting of serine-arginine (or arginine-serine) dipeptides.

consistent with the presence of a recent selective sweep (or alternatively background selection). True, Weir, and Laurie (1996) compared rates of recombination between these three species and found high and almost equivalent coefficient of exchanges in *D. mauritiana* and *D. simulans* (≈ 0.1). *D. melanogaster* exhibited a so-called "centromere effect," whereby a suppression of crossovers takes place around the centromere and decreases with cytological distance. In the region of the third chromosome where *tra* resides, recombination rate in *D. melanogaster* was shown to be an order of magnitude lower than its siblings, *D. simulans* and *D. mauritiana* (True, Weir, and Laurie 1996). Any effect of selection (either positive or negative) will be more evident in *D. melanogaster*, where recombination is lower (Hill and Robertson 1966; Charlesworth, Morgan, and Charlesworth 1993; Hudson and Kaplan 1995).

While *tra* variation within *D. melanogaster* is unusually low, phylogenetic tests using all four species

of the *D. melanogaster* complex suggest a neutral mechanism of change. In the *D. virilis/D. americana* subgroup, McAllister and McVean (2000) demonstrated rapid yet neutral evolution of the *tra* locus. In addition, using three representative sequences of *D. melanogaster*, *D. simulans*, and *D. erecta*, they found their divergence to be consistent with a molecular clock. With the inclusion of the two sibling species *D. mauritiana* and *D. sechellia*, as well as a *D. simulans* sequence that is noticeably different from the previously published sequence (see *Materials and Methods*), we too show that species of this subgroup evolve in a clocklike manner. These sibling species diverged from *D. melanogaster* within the last 3 Myr and from a *D. simulans* common ancestor within the last 0.5 Myr (Kliman et al. 2000). However, we note that while it has been argued that violations of the molecular clock assumption may provide evidence for lineage-specific selection (Yang and Nielsen 1998), its corollary may not necessarily be correct (i.e., a clocklike substitution rate

Table 3
Clustering Statistics for Variable Codon Sites Among Sophophora and Drosophila Subgenera in *transformer*

Subgenus	Codon Sites	Variable	Statistic (Probability)		
			Var(L)	Q	L _{max}
Sophophora					
<i>mell/sim/mau/sec</i>					
Replacement	184	21	0.00313* (0.03)	0.0250 (0.09)	0.2584* (0.03)
Synonymous	184	33	0.00062 (0.54)	0.0085 (0.60)	0.1067 (0.54)
<i>ere/mell/sim/mau/sec</i>					
Replacement	177	37	0.00050 (0.52)	0.0079 (0.18)	0.0899 (0.69)
Synonymous	177	49	0.00018 (0.93)	0.0040 (0.47)	0.0569 (0.97)
Drosophila					
<i>americana</i>					
Replacement	187	19	0.00318 (0.09)	0.0258 (0.39)	0.2287 (0.12)
Synonymous	187	24	0.00136 (0.38)	0.0181 (0.19)	0.1596 (0.30)
<i>virilis/americana</i>					
Replacement	185	30	0.00134* (0.05)	0.0102 (0.59)	0.1774 (0.06)
Synonymous	185	28	0.00081 (0.63)	0.0112 (0.74)	0.1183 (0.59)

NOTE.—Protein sequence from each of the two subgenera were separately aligned and variable amino acid sites linearly mapped. Indels were ignored. A consensus sequence was utilized for each species. Var(L) indicates variance of interval lengths statistic. Q indicates modified variance statistic. L_{max} indicates fractional length of longest interval statistic. Significance (in parentheses) was assessed using the computer program of Goss and Lewontin (1996). The probabilities are the fraction of replicates for which the replicate statistic was equal to or greater than the sample value. Statistically significant values ($P < 0.05$) are indicated with an asterisk. mel, *D. melanogaster*; sim, *D. simulans*; sec, *D. sechellia*; mau, *D. mauritiana*; ere, *D. erecta*.

may not necessarily indicate neutral evolution). For example, *Acp26Aa*, exhibiting the highest nonsynonymous divergence rate among genes that were available from all four species of the *D. melanogaster* complex (Supplementary Material), also reveals a clocklike evolutionary rate. Yet positive selection has been found to act on this locus using standard tests of neutrality (Aguadé 1998; Tsauro, Ting, and Wu 1998). Furthermore, significant heterogeneity between sites also suggests positive selection among regions of the *Acp26Aa* locus (Supplementary Material). Therefore, *tra*'s rapid and clocklike behavior in both the melanogaster and the virilis subgroups (McAllister and McVean 2000) may indicate the constant fixation of neutral substitutions or, alternatively, the constant fixation of advantageous alleles acting in each of these lineages. McDonald-Kreitman and HKA tests of selective neutrality generally corroborate the former hypothesis. These results emphasize the need to utilize both within and between species studies of variation when inferring the action of natural selection or genetic drift.

Rapid Evolution and the Nature of Constraints on RS Domains

TRA is highly diverged not only between sibling species of the *D. melanogaster* complex but also throughout the genus *Drosophila*. So, why does this key developmental gene appear to lack the strong selective constraints typical among other developmental loci? And what is the significance of insertions within the melanogaster subgroup that appear to maintain the total proportion of RS domains (fig. 3)? One explanation is that only certain regions of this locus are required for proper functioning, and all other regions evolve under reduced selective constraints. Our results suggest that

much of this protein is evolving relatively unhampered while maintaining a certain fraction of RS domains.

TRA is part of a protein family containing arginine-rich and serine-rich (RS) domains. These SR proteins are involved in spliceosome assembly and the regulation of alternative splicing (Fu 1995). Specifically, RS domains are thought to help bridge 5' and 3' splice sites in *pre-mRNA* transcripts by interacting with other SR proteins (Fu 1995). In *Drosophila* sex determination, TRA interacts with another SR protein, TRA-2, to form part of the spliceosome complex (Hoshijima et al. 1991). If RS domains (i.e., protein regions with a high concentration of arginine-serine dipeptides) used for spliceosome recruitment represent the major functional part of the protein, it may not be surprising that TRA has undergone high rates of neutral evolution. Domain-swap experiments have demonstrated the exchangeability of loosely conserved TRA RS domains between different SR proteins as well as between distantly related orthologs. For example, RS domains from the suppressor-of-white-apricot protein can be substituted by TRA RS domains to yield partial function (Li and Bingham 1991). In another experiment, SXL, which acts as a splicing suppressor in *Drosophila* sex determination, was transformed into a splicing activator by the addition of RS domains from another SR protein, U2AF (Valcárcel et al. 1993). Finally, in a transgenic experiment with *tra*, O'Neil and Beloté (1992) transferred the wild type *tra* gene of *D. virilis* to *D. melanogaster* by P-element mediated germ line transformation. The much diverged *D. virilis* gene was capable of shifting male structures in *D. melanogaster* flies, which were chromosomally female but homozygous for a *tra* deletion, towards femaleness.

While most of the *tra* locus is practically unalignable between both *Drosophila* and Sophophoran subgenera (O'Neil and Beloté 1992; McAllister and McVean 2000), the presence of RS domains remains conserved. We have

compared the numbers and concentration of RS amino acid dipeptides between species of *Drosophila* (fig. 3) and show that a fraction of the protein has been maintained to consist of domains of RS dipeptides. The fraction of RS dipeptides ranges from 10.9% in *D. hydei* (fig. 3) to 19.2% in *D. melanogaster* (fig. 2). Even with the large insertions in the *D. melanogaster* species complex, the proportion of RS domains remains relatively similar and ranges between 14.6% and 19.2%. Thus, the basic functional requirements of TRA may be fulfilled if at least 10% to 20% of its protein remain RS dipeptides, allowing a large remainder to be functionally unconstrained and amenable to rapid evolutionary change. This, of course, may be an oversimplification as the location of these RS domains, as well as basic amino acid domains, may also prove to be important. Rooney, Zhang, and Nei (2000) demonstrated a similar phenomenon in primate protamines, whereby the proportion of arginine residues, important for DNA binding, remains conserved across distant taxa.

Different species lineages appear to exhibit different selective constraints. In other words, regions of the *tra* locus possess varying levels of purifying selection that are lineage-specific. When *D. erecta*, a species that diverged from the *D. melanogaster* lineage between 10 and 15 MYA, is included in either the broken stick or maximum likelihood heterogeneity tests, heterogeneity in the distribution of replacement substitutions and substitution rate, respectively, is not statistically supported (tables 2 and 3). Although many of the amino acid substitutions between *D. erecta* and its sister species are conservative in nature, they nevertheless suggest that different selective constraints may be acting in both lineages. The inclusion of *D. erecta* in analyses that test for the random distribution of replacement sites and rate heterogeneity between sites contrasts similar analyses in the subgenus *Drosophila*—*D. virilis*' inclusion produces an even more significant nonrandom effect (table 3). This contrast does not appear to be the result of a difference in statistical power since both subgenera possess similar amounts of replacement changes (table 3). Other evidence which may indicate that functional constraints have evolved in a lineage-specific manner concern the particular regions of conserved versus low constraints. In the subgenus *Drosophila*, McAllister and McVean (2000) found a very high d_N/d_S ratio in the third exon of *D. americana* and *D. virilis* lineages ($d_N/d_S = 1.19$) but a very low ratio in *D. hydei* lineages ($d_N/d_S = 0.01$) and a more moderate ratio (similar to the *melanogaster* subgroup) in the second exon ($d_N/d_S = 0.31$). Interestingly, RS domains are not found in the third exon and may indicate differences in purifying selection among the basic amino acid domains (fig. 3). Differences in selective constraint in exon 1 were also evident between these two lineages. In contrast, our results show that among species of the *D. melanogaster* complex, selective constraints were equally moderate in the second and third exons ($d_N/d_S = 0.24$). The presence of varying selective constraints in *tra* orthologs may ultimately affect the affinity of their RS domains to regulate splicing. This affinity may be affected by conformational changes in the protein structure or by direct changes to the RS domains themselves. Differences in the size of these RS domains

are also apparent and may provide a mechanism for the net loss or gain of other such domains in the protein. For example, *D. hydei* (the species with the lowest fraction of RS dipeptides) possesses a relatively large consecutive run of RS dipeptides (fig. 3) possibly reducing the selective constraints in other RS-rich regions of the protein.

We must also remark that the size and nature of the repeats found at the *tra* locus are unprecedented in this species clade. Not only are the lengths of these tandem duplications relatively large (39 and 222 bp in *D. melanogaster* and *D. sechellia*, respectively) compared with the overall size of the protein but also that these two tandem duplications are situated directly adjacent to each other (fig. 2). Another interesting property of these repeats is the apparent correlation between repeat structure and rapid divergence. For example, the only region which exhibited significantly higher rates of amino acid substitution is found around the putative *D. sechellia* insertion site (fig. 2). This region may designate a mutational hotspot for both nonsynonymous substitutions and duplications in these closely related species, thus providing a nonselective causal mechanism for rapid evolutionary change. A similar correlation between repeat structure and rapid divergence is also observed at the *tra* locus in the subgenus *Drosophila*—a series of indel polymorphisms are found among *D. americana* sequences within its largest RS domain (data not shown). This correlation between repeated sequences and faster rates of evolution may be part of a general pattern of divergence (Huntley and Golding 2000).

Rapid Evolution of a Key Developmental Regulator of Sexual Dimorphism

While the relationship between development and evolution is a central theme in evolutionary biology, there exists an unfortunate lack of variational studies on early acting ontogenetic genes (Richter et al. 1997). This study attempts to address this concern by examining the rapid evolution of a gene from a well-characterized early acting developmental pathway that controls sexual identity. TRA is involved in the regulation of somatic sexual differentiation in females by binding (with TRA2 and other spliceosome proteins) to regulatory elements in *doublesex* and *fruitless* (Hoshijima et al. 1991; Steinmann-Zwicky 1994), and it indirectly controls aspects of female sexual differentiation, including all somatic and behavioral components (i.e., Ferveur et al. 1997; Arthur et al. 1998; Gatti, Ferveur, and Martin 2000). Consequently, changes to the transformer protein, particularly in the RS domain's binding affinity, may affect targeted processes of sexual differentiation such as pheromone production, mating behavior, abdominal pigmentation, and bristle number. For example, slight stoichiometric changes in the amount of downstream target transcripts, initiated by amino acid changes in TRA, may affect sex-specific phenotypes. This is consistent with the growing evidence that many reproductive traits are rapidly evolving during the early stages of species divergence (Civetta and Singh 1998; Singh and Kulathinal 2000).

The rapid divergence of this essential developmental

gene among sibling species may present another facet in our understanding of early species divergence. We suggest that random genetic drift serves to quickly change much of *tra*'s protein structure and over time, *accumulated* change may induce various nonneutral effects on downstream targets. Therefore, rapid TRA divergence through neutral drift may contribute to the considerable variation found among *Drosophila* mating systems. The transgenic experiment of O'Neil and Beloté (1992), which restored only *partial* sexual function, is consistent with the idea that slight changes in TRA structure may affect a cascade of sexual traits. It remains an intriguing possibility that the rapid evolution of *transformer* may indirectly contribute to important species differences in *Drosophila*.

Supplementary Material

All new *transformer* species/strain sequences presented in this paper are found in GenBank under accession numbers, AY183382 to AY183407. A table summarizing a set of evolutionary analyses on genes sequenced among all four species of the *D. melanogaster* complex is provided on this journal's Web site.

Acknowledgments

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

- Aguadé, M. 1998. Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. *Genetics* **150**:1079–1089.
- Arthur, B. I., J. M. Jallon, B. Caflisch, Y. Choffat, and R. Nothiger. 1998. Sexual behaviour in *Drosophila* is irreversibly programmed during a critical period. *Curr. Biol.* **8**:1187–1190.
- Burtis, K. C., and B. S. Baker. 1989. *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNA's encoding related sex-specific polypeptides. *Cell* **56**:997–1010.
- Chan, Y. M., and Y. N. Jan. 1999. Conservation of neurogenic genes and mechanisms. *Curr. Opin. Neurobiol.* **9**:582–588.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**:1289–1303.
- Civetta, A., and R. S. Singh. 1998. Sex and speciation: genetic architecture and evolutionary potential of sexual versus non-sexual traits in the sibling species of the *Drosophila melanogaster* complex. *Evolution* **52**:1080–1092.
- Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- Ferveur, J.-F., F. Savarit, C. J. O'kane, G. Sureau, R. J. Greenspan, and J.-M. Jallon. 1997. Genetic feminization of pheromones and its behavioural consequences in *Drosophila* males. *Science* **276**:1555–1558.
- Fu, X.-D. 1995. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**:663–680.
- Gatti, S., J.-F. Ferveur, and J.-R. Martin. 2000. Genetic identification of neurons controlling a sexually dimorphic behaviour. *Curr. Biol.* **10**:667–670.
- Gloor, G., and W. Engels. 1992. Single-fly preps for PCR. *D. I. S.* **71**:148–149.
- Goss, P. J. E., and R. C. Lewontin. 1996. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* **143**:589–602.
- Gould, S. J. 1977. *Ontogeny and phylogeny*. Harvard University Press, Cambridge, Mass.
- Handa, N., O. Nureki, K. Kurimoto, I. Kim, H. Sakamoto, Y. Shimura, Y. Muto, and S. Yokoyama. 1999. Structural basis for recognition of the *tra* mRNA precursor by the sex-lethal protein. *Nature* **398**:579–585.
- Hartmann, M., and G. B. Golding. 1998. Searching for substitution rate heterogeneity. *Mol. Phylog. Evol.* **9**:64–71.
- Hey, J., and R. M. Kliman. 1993. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol. Biol. Evol.* **10**:804–822.
- Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**:269–294.
- Hoshijima, K., K. Inoue, I. Higuchi, H. Sakamoto, and Y. Shimura. 1991. Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in *Drosophila*. *Science* **252**:8333–8336.
- Hudson, R. R., and N. L. Kaplan. 1995. Deleterious background selection with recombination. *Genetics* **141**:1605–1617.
- Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**:153–159.
- Huntley, M., and G. B. Golding. 2000. Evolution of simple sequence in proteins. *J. Mol. Evol.* **51**:131–140.
- Inoue K., K. Hoshijima, I. Higuchi, H. Sakamoto, and Y. Shimura. 1992. Binding of the *Drosophila transformer* and *transformer-2* proteins to the regulatory elements of *doublesex* primary transcript for sex-specific RNA processing. *Proc. Natl. Acad. Sci. USA* **89**:8092–8096.
- Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**:1913–1931.
- Kliman, R. M., and J. Hey. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**:1239–1258.
- Kreitman, M., and R. R. Hudson. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* **127**:565–582.
- Li, H., and P. M. Bingham. 1991. Arginine/serine-rich domains of the *su(wa)* and *tra* RNA processing regulators target proteins to a subnuclear compartment implicated in splicing. *Cell* **67**:335–342.
- McAllister, B. F., and G. A. McVean. 2000. Neutral evolution of the sex-determining gene *transformer* in *Drosophila*. *Genetics* **154**:1711–1720.
- Mcdonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**:652–654.
- Moriyama, E. N., and J. R. Powell. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**:261–277.
- Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
- Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- Nusslein-Volhard, C. 1994. Of flies and fishes. *Science* **266**:572–574.
- Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246**:96–98.

- O'Neil, M. T., and J. M. Beloté. 1992. Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence. *Genetics* **131**:113–128.
- Patel, N. H. 1994. Developmental evolution: insights from studies of insect segmentation. *Science* **266**:581–590.
- Richter, B., M. Long, R. C. Lewontin, and E. Nitasaka. 1997. Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in *Drosophila*. *Genetics* **145**:311–323.
- Rooney, A. P., J. Zhang, and M. Nei. 2000. An unusual form of purifying selection in a sperm protein. *Mol. Biol. Evol.* **17**: 278–283.
- Rozas, J., and R. Rozas. 1997. dnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.
- Ryner, L. C., S. F. Goodwin, D. H. Castrillon, A. Anand, A. Vilella, B. S. Baker, J. C. Hall, B. J. Taylor, and S. A. Wasserman. 1996. Control of male sexual behavior and sexual orientation in *Drosophila* by the *fruitless* gene. *Cell* **87**: 1079–1089.
- Singh, R. S., and R. J. Kulathinal. 2000. Sex gene pool evolution and speciation: a new paradigm. *Genes Genet. Syst.* **75**: 119–130.
- Sosnowski, B. A., J. M. Beloté, and M. Mckeown. 1989. Sex-specific alternative splicing of RNA from the *transformer* gene results from sequence-dependent splice site blockage. *Cell* **58**:449–459.
- Steinmann-Zwicky, M. 1994. Sex determination of the *Drosophila* germ line: *tra* and *dsx* control somatic inductive signals. *Development* **120**:707–716.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- True, J. R., B. S. Weir, and C. C. Laurie. 1996. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. *Genetics* **142**:819–837.
- Tsaur, S.-C., Ting, C.-T., and C.-I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of *Drosophila*: II. Divergence versus polymorphism. *Mol. Biol. Evol.* **15**:1040–1046.
- Valcárcel, J., R. Singh, P. D. Zamore, and M. R. Green. 1993. The protein sex-lethal antagonizes the splicing factor U2AF to regulate alternative splicing of *transformer* pre-mRNA. *Nature* **362**:171–175.
- Walthour, C. S., and S. W. Schaeffer. 1994. Molecular population genetics of sex determination genes: the *transformer* gene of *Drosophila melanogaster*. *Genetics* **136**:1367–1372.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13**:555–556.
- Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**:409–418.

Pierre Capy, Associate Editor

Accepted November 22, 2002