Molecular Evolution of the Testis TAFs of *Drosophila*

Victor C. Li,† Jerel C. Davis,† Kapa Lenkov,† Benjamin Bolival,‡ Margaret T. Fuller,‡ and Dmitri A. Petrov†*

*Harvard Medical School, Biological and Biomedical Sciences Program; †Department of Biology, Stanford University; and ‡Department of Developmental Biology, Stanford University School of Medicine

The basal transcription machinery is responsible for initiating transcription at core promoters. During metazoan evolution, its components have expanded in number and diversified to increase the complexity of transcriptional regulation in tissues and developmental stages. To explore the evolutionary events and forces underlying this diversification, we analyzed the evolution of the Drosophila testis TAFs (TBP-associated factors), paralogs of TAFs from the basal transcription factor TFIID that are essential for normal transcription during spermatogenesis of a large set of specific genes involved in terminal differentiation of male gametes. There are five testis-specific TAFs in Drosophila, each expressed only in primary spermatocytes and each a paralog of a different generally expressed TFIID subunit. An examination of the presence of paralogs across taxa as well as molecular clock dating indicates that all five testis TAFs likely arose within a span of ∼38 My 63–250 Ma by independent duplication events from their generally expressed paralogs. Furthermore, the evolution of the testis TAFs has been rapid, with apparent further accelerations in multiple Drosophila lineages. Analysis of between-species divergence and intraspecies polymorphism indicates that the major forces of evolution on these genes have been reduced purifying selection, pervasive positive selection, and coevolution. Other genes that exhibit similar patterns of evolution in the Drosophila lineages are also characterized by enriched expression in the testis, suggesting that the pervasive positive selection acting on the tTAFs is likely to be related to their expression in the testis.

Introduction

The basal transcription machinery of metazoans (also called the "core promoter recognition machinery") is composed of conserved, generally expressed initiation factor protein complexes, including TFIIA–TFIIH and RNA polymerase II (Thomas and Chiang 2006). A number of alternative homologs of components of the basal apparatus have been identified in metazoans, including many paralogs of the TATA binding protein (TBP) and its associated factors, called TBP-Associated Factors (TAFs). For example, the Drosophila genome encodes four related, but distinct proteins homologous to TBP in addition to TBP itself (Crowley et al. 1993; Rabenstein et al. 1999; Levine and Tjian 2003), as well as paralogs of six of the TAFs, five of which are expressed only in primary spermatocytes.

The alternative variants of TFIID subunits play roles in a range of biological processes. For example, Drosophila TRF1, a paralog of TBP, plays a major role in transcription of targets of Pol III (Isogai et al. 2007), whereas the mammalian TAF4 homolog TAF4b is required for folliculogenesis (Freiman et al. 2002) in the ovary, human TAF9L may be involved in apoptosis (Frontini et al. 2005), and the Drosophila testis TAFs are required for proper spermatogenesis (Hiller et al. 2004). Specialized variant forms of TFIID components have been proposed to provide selective activation of certain Pol II promoters (Verrijzer 2001; Hochheimer and Tjian 2003) for the coordinated regulation of cell type–specific development and differentiation. Indeed, emerging evidence suggests that alternate forms of core transcription machinery components are used during metazoan development to switch between transcription programs as cells differentiate. Incorporation of TAF4b in place of one of the two subunits of the generally expressed TAF4 allows TFIID to stimulate binding of certain DNA binding transcription factors to specific target genes, turning on a cell type–specific transcription program (Liu et al. 2008). Strikingly, differentiation of skeletal muscle involves destruction of TFIID and its replacement by a novel complex of TBP and TAF homologs required for expression of differentiation genes (Deato and Tjian 2007).

A number of new variants of basal transcription factors have been identified (Thomas and Chiang 2006), and their importance for differential and cell type–specific control of gene expression programs in metazoan development is becoming clear. However, little is known about the evolutionary events and forces that have accompanied the genesis of these variants. Here, we examine the evolution of the testis TAFs (also called tTAFs) of Drosophila. Each of the five testis TAFs is a paralog of a different one of the TAFs, designated TAF1-15, which together with TBP comprise the general transcription complex TFIID (Thomas and Chiang 2006). A number of functions have been ascribed to the generally expressed TAFs, including specific promoter binding (e.g., to the Initiator [Int] and Downstream Promoter Element [DPE]), physical contact with mediators and enhancers, and positioning of the preinitiation complex on DNA (Albright and Tjian 2000; Shao et al. 2005). In addition, several TAFs seem to be associated with other large protein complexes, including the histone acetylase complexes and the Polycomb transcriptional repression complex in Drosophila embryos (Struhl and Moqtaderi 1998; Saurin et al. 2001).

The tTAFs of Drosophila are paralogs of TAFs 4, 5, 6, 8, and 12, encoded by the genes *no hitter* (*nht*/*TAF4L*), *cannonball* (*can*/*TAF5L*), *meiosis I arrest* (*mia*/*TAF6L*), *spermatocyte arrest* (*sa*/*TAF8L*), and *ryan express* (*rye*/*TAF12L*), respectively. Expression of the tTAFs turns on in primary spermatocytes, midway through differentiation of male gametes. The tTAFs are required for progression of the meiotic cell cycle through the G2/M1 transition and for robust transcription in primary spermatocytes of a number of genes involved in spermatid differentiation (Lin et al. 1996; White-Cooper et al. 1998; Hiller et al.

Key words: general transcription machinery, evolutionary diversification, gene duplication, testis TAF, transcription complex origin.

E-mail: victorli@fas.harvard.edu.

Mol. Biol. Evol. 26(5):1103–1116. 2009
doi:10.1093/molbev/msp030
Advance Access publication February 25, 2009

2001, 2004). This requirement is gene specific, as many other transcripts are normally expressed in spermatocytes from tTAF null mutant males (White-Cooper et al. 1998). The tTAFs may allow robust expression of terminal differentiation genes in part by counteracting the repression of target genes by Polycomb (Chen et al. 2005). Additionally, Hiller et al. (2004), Chen et al. (2005), and Metcalf and Wassarman (2007) have provided genetic, biochemical, and microscopic evidence that the tTAFs may function together as a physical complex in vivo.

Here, we explored the origin of the tTAFs and the forces that shaped their evolution using the newly available genomic sequences for 12 Drosophila species (Drosophila 12 genomes consortium 2007). Our results suggest that the tTAFs likely arose within a relatively short span of evolutionary time through multiple independent duplications of the generally expressed TAFs and that evolution of the tTAFs since duplication has been driven by coevolution, positive selection, and relaxed purifying selection. These findings constitute a case study of how the basal transcriptional control system has diversified and evolved in metazoans, facilitating cell type and stage-specific regulation of gene expression programs during development.

Materials and Methods

Genomic Search Analysis

Blast searches to all Drosophila genomes were conducted using the TBlastN program at FlyBase under default, unfiltered parameters. All subsequent queries back to the *Drosophila melanogaster* genome or otherwise to *Anopheles gambiae* were conducted under the same parameters. Hits with *E*-values below or near 10^{-20} were filtered and downloaded for further analysis. All queries to *A. gambiae* were made to the AgamP3 assembly (released July 31, 2006).

Synteny Verification

Synteny analysis was conducted using the annotations in FlyBase originally from the Drosophila 12 genomes consortium (2007) or Richards et al. (2005). One hundred kilobases both 5′ and 3′ around a predicted ortholog was checked for neighbors. See main text for the definitions of synteny conservation and relaxed conservation. We restricted cases of gene movement to where synteny was not conserved or to where very few neighbors were present.

Sequencing and Testis Expression of *Drosophila pseudoobscura* Orthologs

5′ and 3′ rapid amplification of cDNA ends (RACE) (Invitrogen, Carlsbad, CA) was used to obtain the *D. pseudoobscura* sequences of *nht* and *rye*. First, a cDNA strand was generated by ligating a known oligomer to the 5′ end of the mRNA message and then performing reverse transcription with an OligodT primer. Next, separate DNA segments representing the 5′ and 3′ ends were polymerase chain reaction (PCR) amplified using nested primers designed to generate overlapping products. These 5′ and 3′ segments were then cloned into a pCR4-TOPO vector, sequenced, and combined to obtain the full-length sequence.

To check the testis expression of *nht* and *rye* in *D. pseudoobscura*, tissue samples were prepared from either whole *D. pseudoobscura* flies (male or female) or dissected testes and remaining residual male carcasses. Ambion's MicroPoly(A)Purist kit (Ambion, Foster City, CA) was used to isolate mRNA from these samples. The reverse-transcription reaction was performed using Ready-to-Go reverse transcription (RT)-PCR beads (Amersham, Piscataway, NJ). Genomic DNA was extracted from samples of male and female flies using the DNAeasy kit (Qiagen, Valencia, CA). PCR was performed using two gene-internal primers in each case.

Estimation of Evolutionary Rates, Linear Regressions, and Statistical Tests

Maximum likelihood (ML) estimates of TAF and tTAF branch lengths were calculated using PAML (Yang 1997) under the amino acid Poisson model (AAML). For the linear regressions, K_A values were estimated using DnaSP 4.0 (Rozas et al. 2003) after a multiple ClustalW alignment. All regression line intercepts were forced to zero. Branch-length tests were performed with the help of the LINTREE program developed by N. Takezaki (downloaded from the Indiana University ftp site). For these tests, Neighbor-Joining (NJ) trees were initially constructed for each tTAF using the 12 Drosophila species and its paralogous TAF in *D. melanogaster* as the outgroup. Subsequently, iterative branch-length tests were performed if the results of a previous test determined the overall rates to be significantly inhomogeneous ($P < 0.05$). After each iteration, a significantly deviated sequence was removed and the NJ tree reconstructed. Iteration was stopped once the overall hypothesis of rate constancy could not be rejected at the 95% level. Tajima one-tailed relative rate tests were performed using MEGA version 3.1 (Kumar et al. 2004) with *D. melanogaster* as one of the sequences and *Drosophila virilis* as the outgroup.

Duplication Date Estimates

BEAST (v.1.4.8)(Drummond and Rambaut 2007) was used to date duplications. The molecular clock model used was the relaxed, uncorrelated lognormal clock. Calculations were performed using the 24 Drosophilid sequences from each tTAF and TAF paralog pair. To calibrate the divergence dates, we set constraints on three different nodes: 1) the divergence of the Drosophila and Sophophora subgroups, 2) the divergence of *D. melanogaster* and *Drosophila ananassae*, and 3) the divergence of *D. melanogaster* and *Drosophila simulans*/*Drosophila sechellia*. The dates for these divergences were set with a uniform distribution and a range determined by the date plus or minus one standard error given in Tamura et al. (2004). These divergences and the root node (the duplication node) were the only constraints that were forced on phylogenetic topology. A constant speciation rate per lineage (the Yule process) was used for the speciation model. The Markov Chain Monte Carlo chain length was set at 2,000,000.

Assessment of Blast Sensitivity for tTAF-Like Singletons

This analysis involved four main steps. First, the *D. melanogaster* proteome was downloaded from the Genbank ftp Blast database and then a reciprocal BlastP search was conducted to identify singletons. These genes were conservatively defined as those that produced no significant hits below an *E*-value of 0.1. Next, the tTAF protein sequences and these singletons were queried to the *Drosophila yakuba* genome using TBlastN. The length (in amino acids), %gap, and percent identities of the top hits for the tTAFs were then determined. The overall maximum and minimum value of these properties within the group of tTAFs were then used to define the range within which singletons would be tTAF-like. This range was then used to filter the set of singletons after they had been similarly queried to *D. yakuba*. Finally, this set of 80 singletons was then queried through TBlastN to the *A. gambiae* genome under the same parameters as in the original genomic search analysis. Throughout this analysis, retrieval of the lengths, %gap, and % identities from the results was conducted using a self-developed script. Blast queries were all performed locally using NCBI's Blast program (under default, unfiltered parameters) and using genomes downloaded from the Genbank ftp Blast database.

Coevolution Tests

Distance profiles were generated for the tTAFs and a general set of 330 REGs (see section below on identifying genes with tTAF properties; for the coevolution test, we narrowed down an original set of 370 REGs with rates within the range of the tTAFs or greater to only those that had rates within the range of the tTAFs) by PAML. A random number generator was then used to select 10,000 unique combinations of five genes. We calculated the average profile for each group/combination by taking the mean of the values of the group members in each species. To calculate the weighted residuals (WR), we then subtracted from each data point (Y) its species average (Y_{av}) and divided it by the same value. That is, WR = $(Y - Y_{\mathrm{av}})/Y_{\mathrm{av}}$. The sum of squares for the group was then calculated as the sum of all WR^2 values. We counted the number of groups that had sum of squares below or equal to that of the tTAFs. For all our tests, *D. yakuba* was excluded intentionally because the range of the REG values there was originally set to be equal to the tTAFs (as a means of identifying their rapid evolution). In the *D. pseudoobscura*, *Drosophila persimilis*, and *Drosophila willistoni* exclusion test, all three species were excluded together. For the robustness tests, species were excluded individually. The coevolution test for the tTAFs using the 38 genes highly enriched in the testis was performed in a similar manner, except that all 12 Drosophila species were included.

Sliding Window Analysis and Tests for Positive Selection

K_S and K_A values were calculated using the method described in Comeron (1995). The sliding window analysis used a window size of 100 bp and step size of 10 bp, which were determined to be optimal based on 1) minimizing the number of windows with $K_S = 0$ and 2) keeping the size of a window to a fraction of the smallest gene. There were several windows where $K_S = 0$ and others where *K*-estimator found K_A or K_S were not applicable (NA). These windows were left blank in figure 5. Confidence intervals for K_A, K_S, and K_A/K_S were determined from Monte Carlo simulations performed in *K*-estimator (Comeron 1999). The Mcdonald–Kreitman test (1991) for selection was done with the help of DnaSP 4.0 (Rozas et al. 2003). The M1a, M2a, M7, and M8 site models were run in PAML v.3.14 (Yang 1997) using the nucleotide sequences of the tTAFs from all 12 Drosophila species. A chi-square distribution with two degrees of freedom was used to calculate the probability of the likelihood-ratio tests between site models.

Collection of Polymorphisms

Polymorphism data were obtained by PCR from the genomic DNA of eight North American *D. melanogaster* strains (We25, We47, We60, Wi15, Wi415, Wi98, Wi45, and Wi18). Products were amplified using gene-specific primers with predicted melting temperatures of 60 ± 3 °C. Sequencing was performed bidirectionally by Genaissance. Testis TAF and general TAF full-length DNA sequences from strains of *D. simulans* were obtained from the *D. simulans* Washington University Genome Sequencing Center database. For our analysis, we only used the protein-coding regions of these sequences.

Identification of Genes Sharing Testis TAF Properties

Predicted homologous gene clusters and their coding sequences were first downloaded from the AAA annotation data sets (Drosophila 12 genomes consortium 2007). Then, genes with 1:1 orthology calls in all 12 Drosophila species were selected. A subset of these genes (6,279) contained easily identifiable tissue-expression data within the FlyAtlas microarray database. This database contains values of mRNA levels for a large number of *D. melanogaster* genes in different adult (and some larval) tissues, which were obtained using four independent replicates of Affymetrix Drosophila Genome 2 expression arrays. As a quality control for this data set, several known tissue-specific genes have been confirmed using these data (Chintapalli et al. 2007). Additionally, approximately 50 genes have been verified using quantitative PCR (http://www.flyatlas.org).

Of this subset of 6,279, genes with accelerations in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* were defined by two criteria: 1) whether they obeyed the linearity of Tamura, Subramanian, and Kumar's molecular clock as well as the tTAFs when the amino acid distances were plotted against speciation time and 2) whether their *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* distances deviated as much as the tTAFs from the expected value predicted through the linear relationship determined in 1). Genes that satisfied both criteria were then further filtered according to evolutionary rate, which we defined as the raw proportion of amino acid differences with their

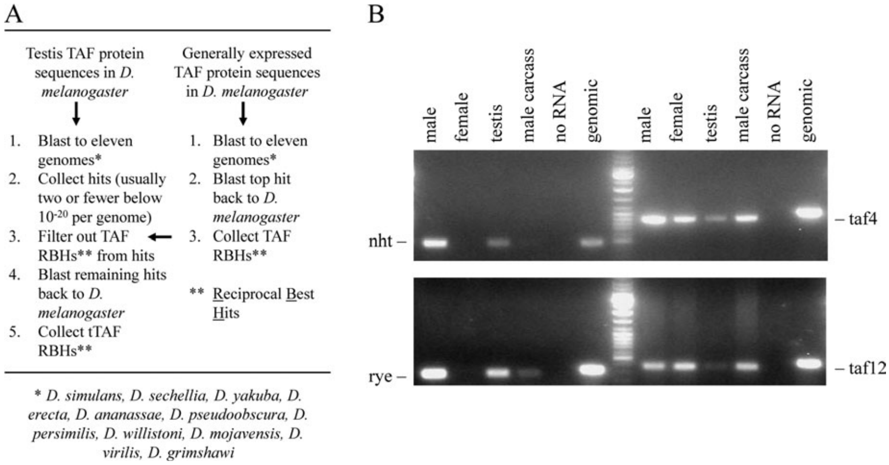
A
Testis TAF protein sequences in D. melanogaster
Generally expressed TAF protein sequences in D. melanogaster
1. Blast to eleven genomes*
2. Collect hits (usually two or fewer below 10-20 per genome)
3. Filter out TAF RBHs** from hits
4. Blast remaining hits back to D. melanogaster
5. Collect tTAF RBHs**
1. Blast to eleven genomes*
2. Blast top hit back to D. melanogaster
3. Collect TAF RBHs**
** Reciprocal Best Hits
* D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi
B
male
female
testis
male carcass
no RNA
genomic
male
female
testis
male carcass
no RNA
genomic
nht –
– taf4
rye –
– taf12

FIG. 1.—Identification and confirmation of testis TAF orthologs. (*A*) Overview of the genomic search method that was used to identify testis TAF orthologs. Two searches are essentially carried out in parallel, one using the tTAF amino acid sequences in *Drosophila melanogaster* and another using the generally expressed TAFs. During the course of the search, the RBHs of the TAFs are filtered out from the Blast results to help identify the best hits of the tTAFs. This method could identify potential orthologs for the tTAFs in the Drosophila species but not in the mosquito genome. (*B*) The *nht* and *rye* orthologs in *Drosophila pseudoobscura* are testis specific. mRNAs from whole males, whole females, dissected testes, and residual male carcasses were isolated and amplified with RT-PCR. A genomic DNA sample from combined male and female flies is also shown.

ortholog in *D. yakuba*. Genes that had rates greater or equal to the tTAFs we called Rapidly Evolving Genes (REGs; 370 identified). As a positive control, it was confirmed that the tTAFs were not filtered out through any step in the analysis. The final set of 54 genes that were both rapidly evolving and accelerated in the three lineages was then scored for enrichment patterns in adult tissues. Here, we counted a gene to be "enriched" only if the statistical call from the FlyAtlas microarray analysis determined the gene to be upregulated relative to the whole fly (this call was made by the Affymetrix MAS 5.0 software based on a two-tailed *t* test with a *P* value of 0.05). Hence, it is possible for a gene to be enriched in more than one tissue.

Results

Identification of Testis TAF Orthologs

The testis TAFs had previously only been identified and studied in *D. melanogaster* (Hiller et al. 2001, 2004). We identified and collected the sequences of the tTAFs in other Drosophila species by using Blast to compare the protein sequences of each *D. melanogaster* TAF and tTAF paralog to the sequences of the 11 other fully sequenced Drosophila genomes (Drosophila 12 genomes consortium 2007) (fig. 1*A*). Each genome generally contained two TAF-like sequences with *E*-values below 10^{-20} (supplementary table 1, Supplementary Material online). One of these was always the reciprocal best hit (RBH) of the *D. melanogaster* generally expressed TAF sequence. We checked to make sure that the remaining hit was an RBH of the *D. melanogaster* tTAF after excluding the *D. melanogaster* TAF sequence. The identified tTAF orthologs in the 12 Drosophila species are listed in table 1. For two species, *D. pseudoobscura* and *D. persimilis*, we did not find any *rye* (TAF12L) orthologs but found more than two TAF5L potential orthologs. We later identified the *D. pseudoobscura* and *D. persimilis* TAF12L orthologs by synteny (see below) and studied the *D. pseudoobscura* and *D. persimilis* TAF5L sequences that are most similar in the amino acid sequence to the TAF5L sequence in *D. melanogaster* (see Materials and Methods).

In almost all cases, we were also able to verify that the identified RBHs of the tTAFs are located in syntenic regions. Neighboring genes flanking the candidates in the studied Drosophila genomes were compared with genes flanking the *D. melanogaster* tTAFs using either Flybase or GLEANR high-confidence annotations (see Materials and Methods). Conservation of synteny was defined as orthology of the genes located in the vicinity of the presumed tTAF orthologs in *D. melanogaster* and the species in question (see Materials and Methods for details). This analysis identified conservation of synteny among the Drosophila species in 49 of 55 cases (table 1). In many cases, all neighbors were conserved in the vicinity of the tTAFs. However, there were also instances in which only one or a few neighbors were conserved (supplementary table 2, Supplementary Material online). As mentioned above, we were also able to identify *rye* (TAF12L) in *D. pseudoobscura* and *D. persimilis* by first finding orthologs of the flanking protein-coding sequences to the left and right of *D. melanogaster rye* (TAF12L), then searching between them to discover a predicted protein having sequence similarity to *rye* in the syntenic location. The *D. pseudoobscura* and the *D. persimilis* TAF12L homologs found in this way both had 27% amino acid identity to *D. melanogaster rye*.

To further test the biological similarity of the predicted tTAF orthologs, we examined the expression pattern of two of the orthologs most highly diverged from *D. melanogaster*, the orthologs of *nht* (TAF4L) and *rye* (TAF12L) in *D. pseudoobscura*. The 5′ and 3′ untranslated regions of the transcripts were identified by 5′ and 3′ RACE from whole *D. pseudoobscura* flies, and this information was used to isolate cDNAs representing the full protein-coding regions by RT-PCR. Sequencing of the transcripts indicated open reading frames consistent with the bioinformatic

Table 1
Ortholog Sequences Collected

	TAF12L (*rye*)		TAF4L (*nht*)		TAF6L (*mia*)		TAF8L (*sa*)		TAF5L (*can*)		
	RBH	Syn	RBH	Syn	RBH	Syn	RBH	Syn	RBH	Syn	T_s (Ma)
Drosophila simulans	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*5.4*
Drosophila sechellia	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*5.4*
Drosophila yakuba	Y	Y	Y	Y	Y	***Y***	Y	Y	Y	Y	*12.8*
Drosophila erecta	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	*12.6*
Drosophila ananassae	Y	Y	Y	Y	Y	***Y***	Y	Y	Y	Y	*44.2*
Drosophila pseudoobscura	N	Y	Y	Y	Y	***N***	Y	Y	Y*	***N***	*54.9*
Drosophila persimilis	N	Y	Y	Y	Y	***N***	Y	Y	Y*	***N***	*54.9*
Drosophila willistoni	Y	(Y)	Y	Y	Y	***N***	Y	Y	Y	Y	*62.2*
Drosophila mojavensis	Y	***Y***	Y	Y	Y	***(Y)***	Y	Y	Y	Y	*62.9*
Drosophila virilis	Y	***Y***	Y	Y	Y	***(Y)***	Y	Y	Y	Y	*62.9*
Drosophila grimshawi	Y	***Y***	Y	Y	Y	***N***	Y	Y	Y	Y	*62.9*
Anopheles gambiae	N	—	N	—	N	—	N	—	N	—	*250*

RBH: reciprocal best hits; Syn: synteny conserved. Bold, italicized sequences indicate instances of gene movement. Speciation dates with reference to *Drosophila melanogaster* (T_s) are taken from Tamura et al. (2004) and Bolshakov et al. (2002). (*) indicates the genome contains several additional hits (E-value below 10^{-20}) for TAF5L. None have conserved synteny with the original gene. (Y) indicates orthologs fit a more relaxed definition of synteny where a neighboring gene one over is conserved. See supplementary figure 1, Supplementary Material online, for an alignment of these sequences.

predictions of the gene structure for *D. pseudoobscura nht* (TAF4L) and *rye* (TAF12L). mRNA samples from whole *D. pseudoobscura* males, females, dissected testes, and the residual male carcasses were reverse transcribed and amplified by PCR. Transcripts from the presumed *D. pseudoobscura nht* and *rye* orthologs were detected in males but not females (fig. 1*B*). In addition, the levels of the *nht* and *rye* ortholog transcripts detected were much higher in the testis than in the remaining carcasses. In contrast, the orthologs of the generally expressed homologs TAF4 and TAF12 showed similar levels of expression in males and females and robust transcript levels in the male carcass, consistent with general expression of the TAF4 and TAF12 orthologs in *D. pseudoobscura* (fig. 1*B*). Thus, the identified orthologs of *nht* and *rye* in *D. pseudoobscura* appear to recapitulate the testis-specific or testis-enriched expression of the *D. melanogaster* tTAFs.

Rapid Rates of Evolution and Acceleration in Multiple *Drosophila* Lineages

The tTAFs appear to be evolving at faster rates than their paralogous generally expressed TAFs. Estimates of the rate of protein evolution for each TAF–tTAF pair using PAML (Yang 1997) revealed that the branch lengths of the tTAF subtrees are longer than the branch lengths of the generally expressed TAFs. Figure 2 shows the dramatic difference in evolutionary rates for *nht* (TAF4L) and its paralog TAF4. The same pattern was observed in the trees of the other tTAFs and their generally expressed paralogs (supplementary fig. 2, Supplementary Material online). Notably, the branch lengths of the tTAFs were longer in most or all of the branches, even those between recently diverged species, such as *D. simulans* and *D. melanogaster*, indicating consistently rapid rates of evolution, even in more recent times and not just shortly after duplication (the number of branches that were longer for each tTAF was 19[*nht*], 16[*rye*], 21[*mia*], 18[*sa*], and 22[*can*] out of a total 22 branches; sign test P values all < 0.0005).

The evolution of the tTAFs in the *Drosophilids* appeared to obey a molecular clock, suggesting the tTAFs were evolving fast at a consistent rate. Plots of the nonsynonymous distances (K_A) from the *D. melanogaster* ortholog of each tTAF to its ortholog in the 11 other species against the species divergence dates proposed by Tamura et al. (2004) (fig. 3) revealed that the divergence of the majority of tTAF orthologs closely followed a linear relationship (r^2 values > 0.95), with the exception of three outliers (*D. pseudoobscura*, *D. persimilis*, and *D. willistoni*), which showed apparent acceleration of the rate of evolution. For most of the tTAF data, therefore, the assumption of a

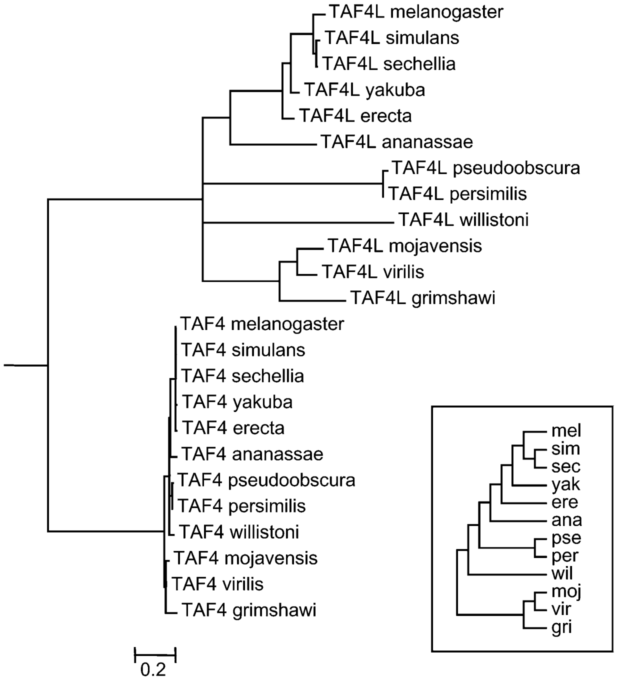

FIG. 2.—Rapid evolution of the testis TAFs. Tree showing the typical rapid evolution of the testis TAFs. Branch lengths were estimated using an ML approach implemented in PAML. Inset, the topology of the species tree used for the analysis.

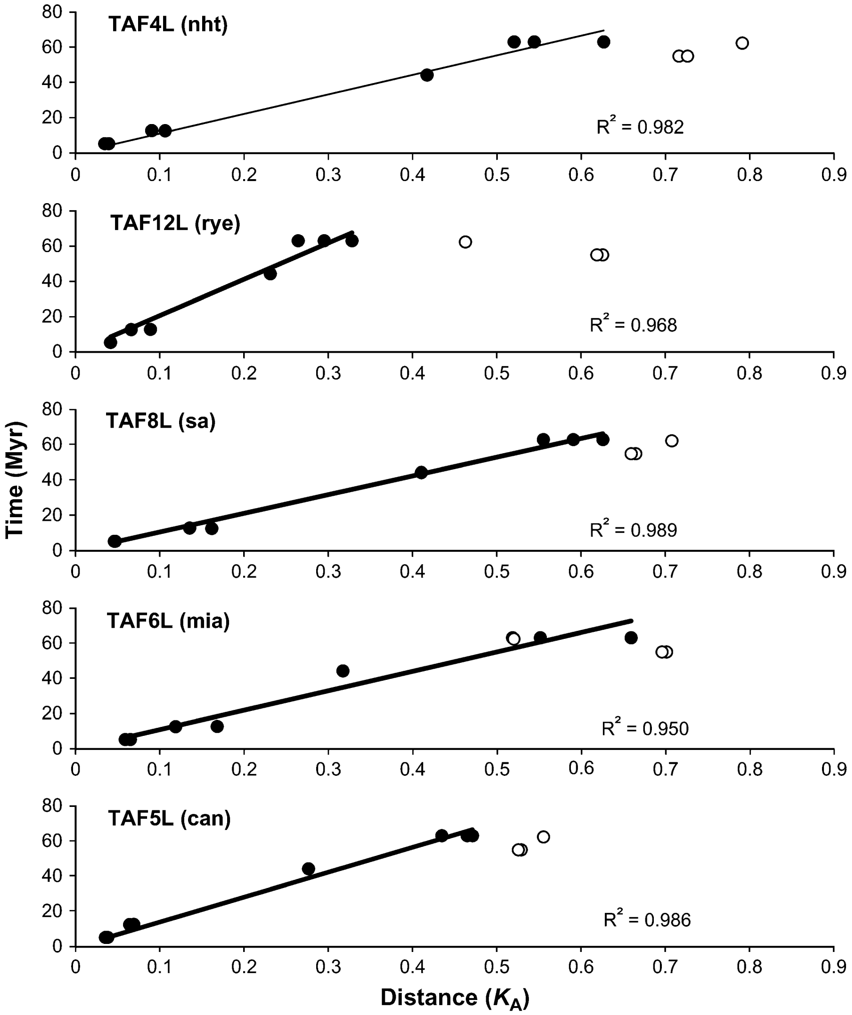

FIG. 3.—Amino acid changes in the testis TAFs approximately follow a molecular clock. The numbers of amino acid substitutions per nonsynonymous site (K_A) are plotted for sequences of *Drosophila melanogaster* and other Drosophila species. The unfilled circles represent the outlier points of *Drosophila pseudoobscura*, *Drosophila persimilis*, and *Drosophila willistoni* in each plot. Regression lines are meant to illustrate the linearity of the data without these outliers.

constant molecular clock is reasonable. Iterative branch-length tests (Takezaki et al. 1995) indicated that the evolution of tTAFs was indeed accelerated significantly for all five tTAFs in the common lineages leading to *D. pseudoobscura* and *D. persimilis*, and for four of the five tTAFs in the lineage leading to *D. willistoni*. In all cases, the changes in amino acid sequence were more extensive than expected, suggesting accelerated evolution of the tTAFs in these lineages. Tajima's relative rate test further substantiated the accelerated evolutionary rate within these three outlier species: *Drosophila pseudoobscura* and *D. persimilis* were accelerated for all five tTAFs, whereas *D. willistoni* showed significantly faster evolution for three of the five tTAFs (table 2). Two other tTAF orthologs, TAF4L and TAF6L in *Drosophila grimshawi*, showed minor deviations in evolutionary rate in the iterative branch-length tests (table 2). When these outlier sequences were removed, the remaining data, which contained the majority of the points, showed no overall significant deviation from rate constancy in branch-length tests ($P > 0.05$).

Molecular Clock Estimates of Duplication Dates

To gain insight into the origins of the tTAFs, we dated their duplication events using a relaxed molecular clock, which allows for rate differences within a phylogeny (Drummond et al. 2006). Because we already identified the tTAFs within the 12 Drosophila genomes, we knew their duplications from the generally expressed TAFs must have occurred before the divergence of the Drosophila and Sophophora subgroups (~62.9 Ma; Tamura et al. 2004). We used this divergence date, the divergence date of the *melanogaster* subgroup from the *ananassae* subgroup, and the divergence date of *D. melanogaster* from *D. simulans* to calibrate our calculations (for details see Materials and Methods). Our analysis estimates the duplications to have occurred ~73.3(*nht*), 89.8(*mia*), 78.7(*can*), 110.9(*sa*), and 80.5(*rye*) Ma. Surprisingly, the duplication dates of the tTAFs fall within a narrow time range of 37.6 My, from 73.3 to 110.9 Ma. This indicates that the tTAF duplications took place after the split of the mosquito *A. gambiae* from

Table 2
Statistical Tests of Acceleration in the *Drosophila pseudoobscura*, *Drosophila persimilis*, and *Drosophila willistoni* Lineages

	Branch-length test				
Gene	TAF5L	TAF12L	TAF4L	TAF6L	TAF8L
Deviated sequences	3	3	4	3	3
Sequence species	*D. pseudoobscura*, *D. persimilis*, *D. willistoni*	*D. pseudoobscura*, *D. persimilis*, *D. willistoni*	*D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *Drosophila grimshawi*	*D. pseudoobscura*, *D. persimilis*, *D. grimshawi*	*D. pseudoobscura*, *D. persimilis*, *D. willistoni*
Direction	(+,+,+)	(+,+,+)	(+,+,+,+)	(+,+,+)	(+,+,+)
	Tajima relative rates test				

P value	0.02(pse)
	0.01(per)
	<0.001(wil)
		<0.001(pse)
		<0.001(per)
		n.s.(wil)
			<0.001(pse)
			<0.001(per)
			<0.001(wil)
				0.002(pse)
				0.003(per)
				n.s.(wil)
					0.04(pse)
					0.03(per)
					<0.001(wil)

For the branch-length test, all sequences identified to have significantly deviated root-to-tip branch lengths ($P < 0.05$) compared with the average are shown. (+) indicates a faster than average rate, whereas (−) indicates slower than average. *P* values of one-tailed Tajima relative rate tests for each tTAF are also given, using the sequences of *Drosophila melanogaster* and the species in parentheses. *Drosophila virilis* serves as the outgroup.

D. melanogaster, which occurred approximately 250 Ma (Bolshakov et al. 2002).

Drosophila Testis TAF Orthologs Are Absent in the *A. gambiae* Genome

Consistent with the relatively recent duplication dates suggested from the molecular clock calculations, we failed to detect tTAF orthologs in the genome of the mosquito, *A. gambiae*. Although TBlastN identified close homologs of all the generally expressed TAFs, we were not able to identify orthologs of the tTAFs even using a permissive *E*-value cutoff of 0.01 (table 1). In two of the cases (TAF4/*nht* and TAF6/*mia*), both the generally expressed and testis-specific TAF sequences in *D. melanogaster* had the same single TBlastN hit in *A. gambiae*. The TAF8/*sa* pair retrieved two hits—one of these was the RBH of the generally expressed TAF8, and the other was a RBH of another gene in *D melanogaster*, bip2. The TAF12/*rye* pair, which is the shortest protein of the five TAFs, initially retrieved five hits, but retrieved only one hit when a low complexity filter was employed (see Materials and Methods). Finally, TAF5 and *can* both retrieved several hits from the *A. gambiae* genome due to their conserved WD40 repeats. Once we removed the WD40 repeats from the query sequences, TAF5/*can* retrieved only a single hit corresponding to the general TAF ortholog (*E*-value = 1.4e−82 from TAF5 query).

Failure to find orthologs of the tTAFs in *A. gambiae* might either be because the tTAF protein sequences are so rapidly evolving that the orthologs can no longer be detected by TBlastN of the *A. gambiae* genome or because the genes arose after the evolutionary divergence of Drosophila and *A. gambiae*. To test the first possibility, we identified 80 single copy *D. melanogaster* genes that encoded predicted proteins with similar lengths (in amino acids) and rates of evolution (% amino acid identity and % sequence gaps between the *D. melanogaster* and *D. yakuba* orthologs) as the tTAFs, then determined the frequency that a homolog of these 80 genes could be identified in the genome of *A. gambiae* by TBlastN. For the 0.01 *E*-value threshold that was used in the tTAF search above, 87.5% of the genes in the test set identified homologs in the *A. gambiae* genome. This implies that our failure to identify orthologs of all five tTAFs in *A. gambiae* by TBlastN was unlikely to happen by chance ($P \sim 3 \times 10^{-5}$). Rather, consistent with the estimates of duplication dates from the molecular clock analysis, the tTAFs probably arose since the split of Drosophila and Anopheles approximately 250 Ma (Bolshakov et al. 2002).

The Testis TAFs Are Coevolving

The high rates of evolution and accelerations in the *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* lineages shared by the tTAFs suggested that the tTAFs may be coevolving. The patterns of evolution of the five tTAFs was compared with that for a set of 330 genes (REG set) picked because they had clear 1:1 orthologs in each of the 12 Drosophila species and evolve at rates similar to those of the tTAFs between *D. melanogaster* and *D. yakuba* (see Materials and Methods for details). tTAFs appear to evolve in Drosophila at more similar rates compared with REG genes overall (fig. 4). To estimate whether the homogeneity of the rates of evolution exhibited by tTAFs is unusual, we generated 10,000 random combinations of five genes from our set of 330 REGs and compared the variability in the rates of evolution of these sets of five genes with that exhibited by the five tTAFs. Heterogeneity was measured by calculating the sum of squares of the deviations of the rates of evolution for individual genes compared with the average rate of evolution for the five genes in the set (Materials and Methods). We found that the tTAFs indeed evolve at unusually similar rates, with only 195 combinations of five REGs out of 10,000 displaying as much or more similarity in their rates of evolution as the tTAFs ($P = 0.0195$). The significance of this result was not due to the common accelerations in the *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* lineages. When we excluded these lineages from the counting criteria, the test still obtained a significant *P* value of 0.0493. To estimate the robustness of the test to influence from any single species, we removed each species one at a time (before excluding *D. pseudoobscura*, *D. persimilis*, and *D. willistoni*); the *P* value in all cases was less than

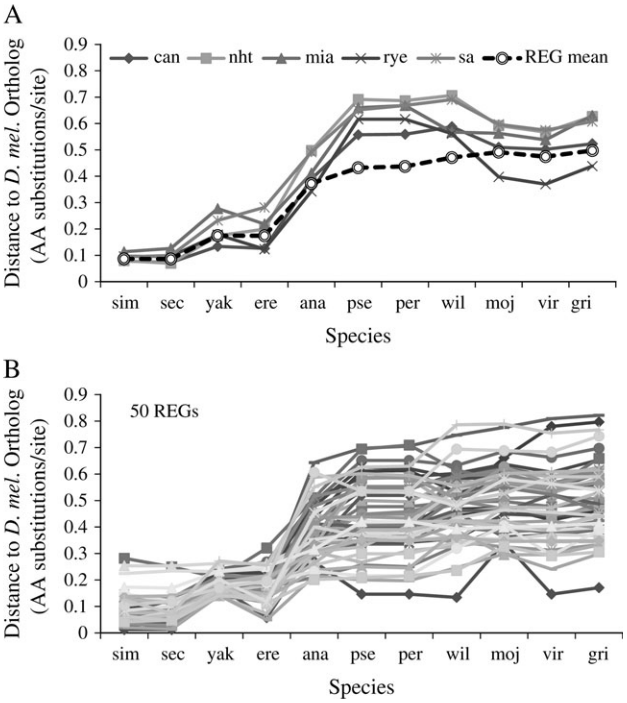

FIG. 4.—The Drosophila tTAFs are coevolving. Evolutionary distance profiles of the tTAFs showing qualitatively the tTAFs share a common pattern of sequence divergence across species. The mean profile for a large sampling of 502 tTAF rate-matched REGs is shown as a dotted line. (*B*) Evolutionary distance profiles of tTAF rate-matched REGs (Here only 50 are displayed). Note the reduced spread of distances in *A* compared with *B*. This difference is significant ($P = 0.0195$). *Drosophila yakuba* should not be compared because the range in this species was specifically set to be similar to the tTAFs (as a means of selecting REGs).

0.05, indicating that the coevolution of the tTAFs has been pervasive in Drosophila.

Positive Selection and Relaxed Purifying Selection Drive Testis TAF Evolution

The relatively rapid evolution of the tTAFs since their duplication from their generally expressed paralogs could result from an elevation in mutation rate, relaxation of purifying selection, or pervasive positive selection. Analysis of the rate of synonymous (K_S) and nonsynonymous (K_A) substitutions, as well as their ratio K_A/K_S, for both the generally expressed TAFs and the tTAFs indicated relaxation of purifying selection or positive selection acting on the tTAFs. The K_S values for the paired tTAFs and TAFs were approximately the same (table 3). Although Monte Carlo simulations found three tTAFs to have significantly elevated K_S values, the absolute changes were small. Thus, an elevated mutation rate did not appear to be responsible for the rapid evolution of the tTAFs. Both K_A and K_A/K_S values, however, were dramatically increased in the tTAFs. Three of the tTAFs (*nht*, *can*, and *mia*) had K_A values ten or more times the value for their corresponding general TAF, whereas the other two (*rye* and *sa*) had K_A values approximately three to four times larger.

Table 3
K_A/K_S Test for the Testis TAFs

General	TAF4	TAF5	TAF6	TAF8	TAF12
K_A	0.0000	0.0058	0.0077	0.0141	0.0134
K_S	0.1002	0.1013	0.0865	0.1321	0.0898
K_A/K_S	0.0000	0.0573	0.0892	0.1069	0.1494
Testis	TAF4L (*nht*)	TAF5L (*can*)	TAF6L (*mia*)	TAF8L (*sa*)	TAF12L (*rye*)
K_A	0.047^{***}	0.059^{***}	0.078^{***}	0.047^{***}	0.035^{***}
K_S	0.093 (ns)	0.141^{**}	0.132^{**}	0.117 (ns)	0.143^{*}
K_A/K_S	0.505^{***}	0.421^{***}	0.589^{***}	0.399^{***}	0.245(ns)

Different rates of evolution (nonsynonymous, synonymous, and ratio) are presented between orthologs of *Drosophila melanogaster* and *Drosophila simulans*. Confidence intervals were estimated for K_A, K_S, and K_A/K_S in the general TAFs using Monte Carlo simulations (see Materials and Methods). Significances represent degrees of departure for testis TAF rates from that of their general paralogs.***$P \ll 0.001$, **$P < 0.01$, and *$P < 0.05$.

Analysis of synonymous versus nonsynonymous polymorphisms among strains of *D. melanogaster* and *D. simulans* by McDonald–Kreitman tests (Mcdonald and Kreitman 1991) indicated that some of the tTAFs have been evolving under global positive selection. Polymorphism data for the tTAF genes in North American *D. melanogaster* strains were collected by sequencing PCR products covering each gene obtained with gene-specific primers (Materials and Methods). In addition, the sequences of the TAFs and tTAFs were obtained from genomic sequences of several *D. simulans* strains deposited in a public data set. The McDonald–Kreitman test results indicated significant positive selection for *nht*, *can*, and *sa* (table 4). In these three cases (*nht*, *can*, and *sa*), the ratios of nonsynonymous-to-synonymous divergence (D_n:D_s) were significantly higher than the ratios of nonsynonymous-to-synonymous polymorphism (P_n:P_s), indicating the action of positive selection.

In addition to demonstrating strong signals of positive selection in three of five tTAF proteins, the *D. melanogaster* and *D. simulans* polymorphism data also indicated relaxation of constraints in the tTAFs compared with their generally expressed paralogs. P_n:P_s ratios for each of the tTAFs were 2- to 4-fold higher than the P_n:P_s ratios for their generally expressed TAF paralogs ($P < 0.05$ for *can*, *mia*, and *nht*; $P = 0.051$ for *sa*; $P = 0.142$ for *rye*; 2 × 2 one-tailed G-test).

In addition to evidence of global positive selection, several of the tTAFs also had regions that showed particularly high rates of protein evolution in a sliding window analysis using the orthologous sequences from *D. melanogaster* and *D. erecta* (fig. 5). These species were chosen for calculating K_A/K_S values because of their appropriate overall evolutionary distance for the tTAFs (mean K_S 0.353), which provided a substantial number of substitution events while keeping distances low enough such that multiple hits are unlikely to obscure the patterns of evolution. In general, K_A/K_S was <1 for the majority of windows, confirming our previous conclusion that the tTAFs have been subject to purifying selection. However, for each tTAF, K_A/K_S approached or exceeded one in certain regions. Monte Carlo simulations (Comeron 1995) identified windows with significantly elevated K_A/K_S in three of the five tTAFs (*can*, *mia*, and *rye*, $P < 0.01$). The signal in *mia* was particularly

Table 4
Mcdonald–Kreitman Test for Global Positive Selection

Gene	*rye*	TAF12	*nht*[*]	TAF4	*mia*	TAF6	*can*[*]	TAF5	*sa*[*]	TAF8
Mel. strains	8	1	8	1	8	1	6	1	8	1
Sim. strains	1	3	4	4	3	2	2	3	4	2
D_n[a]	9	5	24	0	89	8	120	6	24	6
D_s[a]	9	15	7	12	48	28	72	16	19	24
P_n[a]	2	1	9	1	15	1	16	10	4	1
P_s[a]	2	3	19	10	12	13	34	40	19	13
P_n/P_s	1.0000	0.3333	0.4737	0.1000	1.2500	0.0769	0.4706	0.2500	0.2105	0.0769
χ^2 P value	1.0000	1.0000	0.0066	0.7673	0.8350	0.6701	0.0018	0.9260	0.0285	0.7579
G-test P value	1.0000	1.0000	0.0004	0.4783[c]	0.3584	0.1803	0.0002	0.4997	0.0019	0.2489

The number of *Drosophila melanogaster* and *Drosophila simulans* strains used to study each gene are provided. P_n/P_s is also shown to reveal relaxation of constraints (compare between paralogs). The P values of two different tests (chi-square and G-test) are given, which test for differences between D_n/D_s ratios with P_n/P_s ratios. Genes with significant P values are indicated by *.

[a] Nonsynonymous (n) and synonymous (s) polymorphisms (P) and fixed differences (D).

[b] Parentheses enclose counts of polymorphisms from (*D. melanogaster*, *D. simulans*).

[c] G-test cannot be computed for zero; Fisher's exact test used instead.

strong, with a region lying near the C-terminal end of the mia-TAF6 conserved domain having a K_A/K_S ratio of 9.32. In *rye*, a strong positive signal ($K_A/K_S = 2.45$) was located near the N-terminus of the protein, before the histone fold domain (HFD). In *can,* the window of high K_A/K_S indicating positive selection was located near the N-terminus of the protein, in a region that is not obviously homologous to the generally expressed homolog dTAF5 (Hiller et al. 2001). The remaining tTAFs (*nht* and *sa*) also contained windows with K_A/K_S marginally exceeding 1, but the Monte Carlo simulation-based statistical test in K-estimator did not confirm these as significant. K_S values for these regions with $K_A/K_S > 1$ exceeded the lowest value for the gene, or contained several windows which exceeded the lowest value, indicating that $K_A/K_S > 1$ in those locations was not due to K_S being particularly low. In contrast, the generally expressed TAF homologs did not possess any windows with $K_A/K_S > 1$.

Codon-level analysis performed using PAML did not reveal strong evidence of recurrent evolution under positive selection at any sites. The two likelihood-ratio tests M1a versus M2a and M7 versus M8 were run on each of the tTAFs (using all 12 Drosophila ortholog sequences), but significant ($P < 0.05$) results were not obtained for either test on any gene. The Naïve and Bayes Empirical Bayes calculations associated with models M2a and M8 also did not reveal any positively selected sites with probability of positive selection >95%, although for several tTAFs, a small number of sites with probability of positive selection >50% but less than 95% were identified. There were eight such sites for *mia* (124V, 126E, 127A, 132K, 133K, 200D, 424T, and 431T—sites are numbered based on the amino acid sequence in *D. melanogaster*), four for *sa* (45R, 91L, 95N, and 172Q), two for *can* (858F, 884E), and one for *nht* (78W). There were no such sites identified for *rye*. Several of these sites corresponded to the regions of positive selection identified using the sliding window analysis (200D, 424T, and 431T in *mia* and 858F in *can*). Taken together, the results of the McDonald–Kreitman tests and the regional high scores in the sliding window analysis support positive selection for change in specific regions of the tTAF proteins, whereas the results of the PAML analysis suggest that few individual amino acids have undergone substantial recurrent changes in Drosophila. We posit that this evolutionary behavior may reflect adaptation of tTAF proteins to new functions in the testis, perhaps due to interaction with altered partners or participation in tissue-specific complexes different from those in which the generally expressed TAFs participate.

Testis TAF Evolutionary Properties Correlate with Their Expression Pattern

To elucidate possible sources for the selective pressures affecting the tTAFs, we searched for other Drosophila genes that possess the same rapid rates of evolution and lineage-specific accelerations as the tTAFs. For all *D. melanogaster* genes that have clear 1:1 orthologs in the 12 Drosophila species (Drosophila 12 genomes consortium 2007), we translated and aligned the sequences with their orthologs and calculated the proportion of amino acid differences between the *D. melanogaster* protein and the ortholog in each species. Selecting sequences that displayed the same or greater rate of evolution between *D. melanogaster* and *D. yakuba* and possessed the same or greater acceleration within *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* as the tTAFs (see Materials and Methods for details) provided a final list of 54 genes. Gene ontology annotations for these 54 genes did not reveal any obvious commonalities, although a few were known to be involved in spermatogenesis (table 5). However, analysis of the mRNA expression patterns of this gene set in *D. melanogaster* revealed that all of these 54 genes had higher expression in the testes than in the other tissues (fig. 6). This was not the case if we used only one of the two criteria to select the genes. Neither the set of genes rapidly evolving in Drosophila that did not experience an even higher rate in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* nor the set of genes evolving unusually rapidly in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* but not evolving fast overall in the other Drosophila species was strongly biased for genes upregulated in the testis.

To see if the tTAFs' high expression levels in the testis versus other tissues, rapid evolutionary rates, and lineage-specific accelerations are determining factors of tTAF

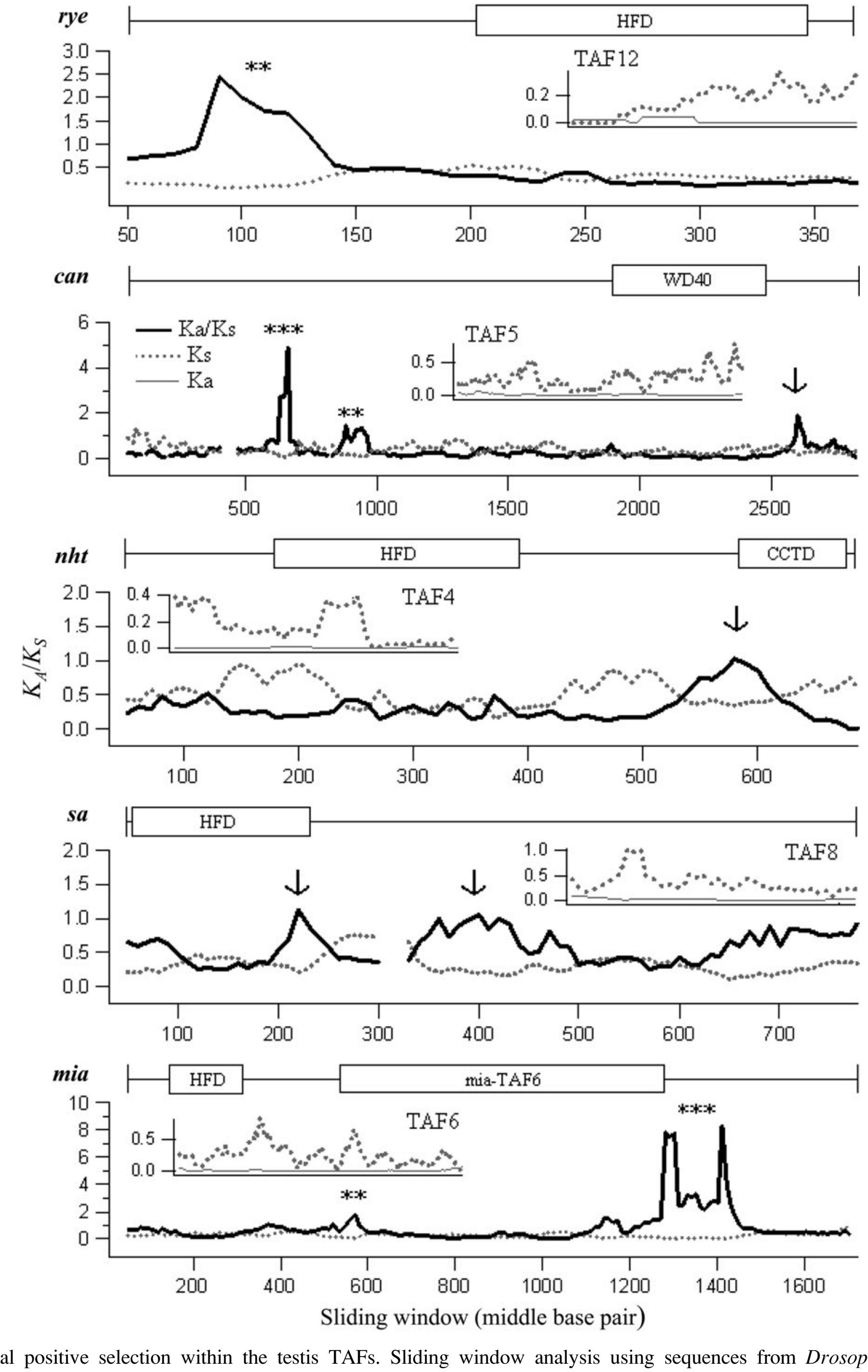

FIG. 5.—Regional positive selection within the testis TAFs. Sliding window analysis using sequences from *Drosophila melanogaster* and *Drosophila erecta* showing positive selection in regions of the tTAFs. Regions in which $K_A/K_S > 1$ are indicated with arrows, and if significant, with stars. The sliding window profile of the homologous general TAFs are shown in the inset. Conserved domains are HFD, WD40 domains, Conserved C-terminal domain, and a *mia*-TAF6 domain.

coevolution, we generated 10,000 random combinations of five genes from 38 testis-enriched REGs with accelerations in *D. pseudoobscura*, *D. persimilis*, and *D. willistoni* and compared the homogeneity in their rates of evolution with the tTAFs. Out of this set, we found 445 combinations that had heterogeneities (sum of squares) less than or equal to the tTAFs ($P = 0.0445$), indicating that the degree of homogeneity within the tTAFs was unusual even for genes that had their evolutionary properties.

Discussion

Our results indicate that the tTAFs arose by duplication most likely between 63 and 250 Ma (within a short time span ~73.3–110.9 Ma), after the divergence of mosquitoes and true flies. In support of this date range, we found the absence of tTAF orthologs in the *A. gambiae* genome and duplication dates that fell within the predicted range when calculated with a molecular clock. The alternative scenario

Table 5
Genes Sharing the Same Evolutionary Properties as the Testis TAFs (Partial List of 15)

Name/CG Number	GO Biological Process/Molecular Function	Expression[a] (Flyatlas)
CG2075	aly; male meiosis I, spermatogenesis, establishment and/or maintenance of chromatin architecture, spermatid development, transcription initiation	Testis enriched
CG9929	Art9; protein–arginine *N*-methyltransferase activity	Testis specific
CG10694	Base-excision repair, proteasomal ubiquitin-dependent protein catabolic process, damaged DNA binding	Testis specific
CG13493	Comr; positive regulation of transcription from RNA polymerase II promoter, spermatogenesis, chromatin	Testis specific
CG16940	Cytoplasmic exosome (RNase complex), nuclear exosome (RNase complex), exoribonuclease II activity	Testis and ovary enriched
CG6539	Dhh1; RNA helicase activity, ATP-dependent RNA helicase activity, ATP binding	Testis and ovary enriched
CG31835	Intracellular zinc ion binding	Testis specific
CG3219	Klp59C; mitotic sister chromatid segregation, kinesin complex, microtubule motor activity	Testis specific
CG14660	Laf; embryonic development via the syncytial bastoderm	Testis specific
CG3162	Nuclear mRNA splicing, via spliceosome, snRNP U2, mRNA binding	Testis specific
CG12857	Protein binding	Testis specific
CG10254	Regulation of protein metabolic process, posttranslational protein modification, ubiquitin–protein ligase activity	Testis and ovary enriched
CG4711	Squash; dorsal appendage formation, oogenesis	Testis and ovary enriched
CG15262	Transcription regulator activity	Testis specific
CG30156	Unfolded protein binding, heat shock protein binding	Testis specific

For the full list, see supplementary table 3, Supplementary Material online.

Testis specific: greater than 25-fold higher mRNA signal in testis than elsewhere and little or no expression detected in other tissues assayed.

Testis enriched: mRNA signal high in testis. Also expressed in other tissues but greater than 4-fold higher in testis than elsewhere.

Testis and ovary enriched: mRNA signal 3-fold higher in testis and ovary than in other tissues assayed. Also expressed in other tissues but greater than 3-fold higher in testis and ovary than elsewhere.

[a] Expression data in adult tissues based on Chintapalli et al. (2007).

that all five genes duplicated prior to the split of Drosophila from Anopheles is difficult to explain in light of these observations. First, this would imply that all five duplicated copies have been lost in the mosquito lineage. Second, it would suggest that duplication date estimates under the molecular clock are dramatic underestimates in all five cases, which we believe is unlikely. With the exception of *D. pseudoobscura*, *D. persimilis*, and *D. willistoni*, analysis of evolutionary distances of tTAF orthologs in Drosophila closely followed the assumption of a consistent molecular clock. Moreover, there is some reason to suspect that our estimates are overestimates, as evolutionary rates are expected to be higher than the calculated rates because a theoretical period of functional redundancy following duplication can lead to increased rates of evolution (Hurles 2004). If a deviation from rate constancy occurred early after gene duplication, we suspect the true duplication dates would only be nearer to the 63 Ma mark than our estimates

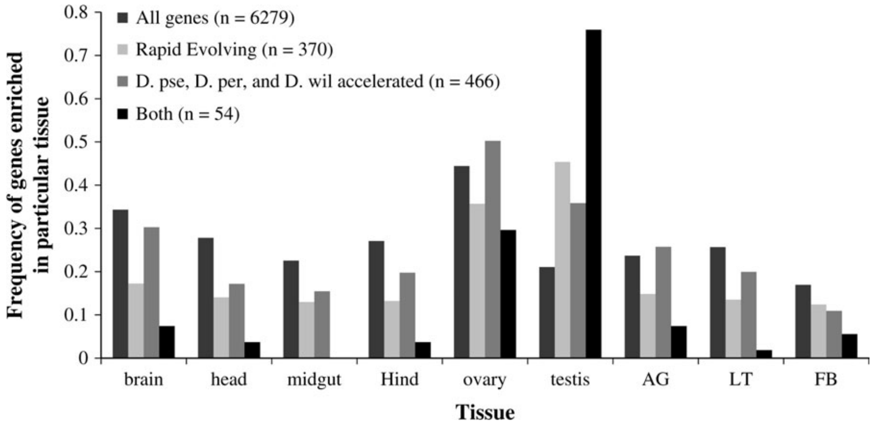

FIG. 6.—Tissue-expression patterns for genes that share the same evolutionary properties as the testis TAFs. Tissue-expression data for these different sets of genes were obtained from a database of *Drosophila melanogaster* gene mRNA levels called FlyAtlas (Chintapalli et al. 2007). These mRNA levels were determined through a microarray analysis of different dissected tissues and whole adult flies (see Materials and Methods for details). "Enrichment" refers to the statistical call of upregulation in the mRNA level of a gene in a particular tissue relative to the whole fly (as determined by the FlyAtlas microarray analysis). Each gene can be enriched in more than one tissue. Frequency refers to the percentage of genes in a particular set enriched in that tissue. AG = accessory gland; LT = larval tubule; and FB = larval fat body.

suggest. It will be possible to provide additional evidence for this date range as more genomic data from species that diverged after the Drosophila–Anopheles split but before the most recent common ancestor of Drosophila (e.g., housefly) becomes available.

Because the genes encoding the TAFs are not collocated with each other in the genome and there is no known mechanism that could have coordinated the simultaneous duplication of all five tTAFs, the most likely scenario is that each of the five tTAFs arose from a separate duplication event. If so, it is surprising that the molecular clock duplication dates of all tTAFs fall within such a short time period (~38 My, 43% the average age of Drosophila tTAFs estimated by molecular clock and 15% of the divergence time between mosquitoes and other flies). Caution should be taken, however, in interpreting these dates, as there is usually some error in the calibration times used (taken from Tamura et al. 2004). Nevertheless, the proximity among tTAF origin dates suggests that although they may have arisen from separate events, their duplications may not have been selectively independent. One possibility is that the initial duplication of a single tTAF modified the selective environment to favor maintenance of a duplication of a second tTAF, a third, and so on. Because the tTAFs appear to work together in a common process, possibly in a common complex (Hiller et al. 2004; Chen X, Fuller MT, personal communication), the appearance and evolution of one subunit might influence the appearance and evolution of the others.

One question that remains unresolved is the mechanism of tTAF duplication. Hiller et al. (2004) previously mapped the TAFs and their paralogous tTAFs to different chromosomal loci (supplementary table 1, Supplementary Material online), indicating that tandem duplication alone could not be a sufficient explanation. Retroposition is not entirely consistent with the data either because several of the tTAFs contain introns (Hiller et al. 2001, 2004). Although *nht* is intronless, *can, rye*, *sa*, and *mia* have introns, some of which appear to be in similar locations to introns of their paralogous general TAFs. In addition, *can* appears to have gained additional four introns not present in TAF5 (Hiller et al. 2001). Retroposition that involved incompletely processed mRNA is one possible scenario. This would be consistent with the generally frequent recruitment of retroposed genes into the male germline (Vinckenbosch et al. 2006; Bai et al. 2007). Another possible explanation is that the tTAFs initially duplicated in tandem to their generally expressed homologs but subsequently broke up by interchromosomal transpositions. However, we did not find any evidence for this in terms of neighboring genes that might have coduplicated with the tTAFs. It will be interesting to see in the future if chromosomal translocations can be mapped that might explain the locations of the tTAFs.

The Rapid Evolution of the tTAFs

The high rate of evolution displayed by the tTAFs is rare among components of the core transcription machinery. We do not know of any components of the generally expressed basal transcription apparatus that share comparable rates of sequence evolution. For example, TBP and the general TAFs evolve very slowly across eukaryotes (Hernandez 1993; our data). Instead, the evolutionary rates of the tTAFs are most similar to those of the fastest-evolving genes in metazoa (e.g., accessory gland proteins; Swanson et al. 2001). It is likely that the forces responsible for these rapid rates have some relationship to the role that tTAFs play in the testis function, because we demonstrated that genes evolving similarly to the tTAFs are almost exclusively expressed in the testis (fig. 6). The finding that many male-biased genes in *D. melanogaster* evolve at accelerated rates is consistent with this hypothesis (Swanson et al. 2001). Functionally, the rapid evolution of the tTAFs also suggests that the tTAFs may have a role in germline functional divergence and speciation. Because tTAFs interact physically and coevolve, high rates of evolution of tTAFs could lead to Dobzhansky–Muller incompatibilities where the tTAFs can no longer interact and function together.

Sources of Positive Selection and Coevolution

The tTAFs of Drosophila appear to be evolving through rapid, positive selection both globally and regionally. They also appear to be coevolving and subject to weakened purifying selection compared with their generally expressed paralogs. Each of the five tTAFs is encoded by a paralog of a different generally expressed TAF that is a component of the general transcription factor TFIID (Hiller et al. 2004). The generally expressed TAFs 4, 5, 6, 8, 9, 10, and 12 assemble into a stable complex that appears to form the core structure of TFIID (Leurent et al. 2004). Two molecules of TAF5 lie with their N termini in proximity and their WD40 domain–containing C termini separate. The TAF5 dimer appears to form a binding platform for the remaining six TAFs, which interact via their histone fold domains in heterodimer pairs (TAF4-TAF12; TAF6-TAF9; and TAF8-TAF10) (Gangloff et al. 2000, 2001). The result is a trilobed structure, which is decorated with the other TAFs and the TBP to form TFIID. This core structure also appears to participate in other protein complexes that lack TBP, such as TFTC/SAGA (Leurent et al. 2004). Because the tTAFs are the paralogs of TAF4, TAF5, TAF6, TAF8, and TAF12, it may be that the tTAFs form a structure similar to the core complex formed by their generally expressed paralogs.

The requirement for action of the tTAFs is gene specific. Although wild type function of the tTAFs is required for robust expression in spermatocytes of a large number of target genes implicated in terminal differentiation, many genes are expressed normally in tTAF mutant spermatocytes (White-Cooper et al. 1998). Consistent with gene-selective function, analysis by chromatin immunoprecipitation revealed tTAF binding at promoters of three representative target genes but not at two representative nontarget genes that are also expressed in spermatocytes (Chen et al. 2005). Positive selection on the tTAFs could arise as a result of a rapid selection for better interaction of a tTAF-containing complex with rapid adaptive evolution of the promoter–enhancer sequences or transcription activators associated with tTAF target genes. Strikingly, at least two of the regions in the tTAF proteins that we identified to be under positive selection are homologous to or lie next to regions implicated in DNA binding in the generally expressed TAFs. These are amino acids 363–500 of *mia* and 177–210 of *nht*. The region containing amino acids 363–500 of *mia* lies directly C-terminal to a region with known DNA-binding activity in human TAF6. This DNA-binding region (amino acids 300–400) is evolutionarily conserved, resides C-terminal to a HFD, and binds DNA cellulose (Shao et al. 2005). In *nht,* the 177–210 region is homologous to the 311–350 region of yeast TAF4, which lies within a spacer region linking the TAF4 HFD to a Conserved C-Terminal Domain. This 311–350 region has highly conserved DNA-binding activity in the TAF4 orthologs of human, Drosophila, and yeast (Shao et al. 2005).

It is not yet known whether the tTAFs participate in a testis-specific TFIID-like complex or a TFTC–SAGA-like complex. However, it is tempting to speculate that they may confer special gene-selective action on a tissue-specific form of TFIID at work in primary spermatocytes to control expression of genes required for subsequent spermatid differentiation. Recent work has shown that incorporation of one subunit of the variant mammalian isoform TAF4b into TFIID strongly influences transcriptional activation at selected promoters and potentiates the binding and action of specific transcriptional activators compared with the canonical TFIID (Liu et al. 2008). Thus, tissue-specific TAFs may help direct gene-selective action of more generally expressed transcriptional activators to turn on expression of banks of tissue-specific target genes. If similar mechanisms are at play in Drosophila spermatocytes, the rapid evolution of the tTAFs may have allowed and been driven by the formation of a novel TFIID-like structure that regulates a particular subset of target genes in the testis that evolve fast under positive selection. tTAFs would then be both tracking and allowing adaptive and fast evolution of the target genes (Swanson et al. 2001).

In addition to promoter selection and/or ability to interact with specific transcription factors, positive selection within the testis TAFs may have in part been driven by compensatory mutations that maintain ability of specific tTAFs to interact with each other, either as heterodimer partners or within the core complex. This, coupled with a rapid rate of evolution within the testis TAFs due to the forces described above, may have driven changes in amino acid sequence in a back-and-forth game of "catch up." So far, only the biochemical interaction between the tTAF subunits *nht* and *rye* has been tested and confirmed. However, our detection of coevolution supports the notion that the testis TAFs are subject to shared evolutionary pressures.

Supplementary Material

Supplementary tables 1–3 and supplementary figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

V.L. was supported in part by NIH Cell and Developmental Biology Training Grant GM0722 at Harvard. This work was supported by NIH 3 RO1 GM061986 to M.T.F, NIH R01 GM077368 to D.P., and NSF DEB-0317171 to D.P.

Literature Cited

Albright SR, Tjian R. 2000. TAFs revisited: more data reveal new twists and confirm old ideas. Gene. 242:1–13.

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8:R11.

Bolshakov VN, Topalis P, Blass C, Kokoza E, Della Torre A, Kafatos FC, Louis C. 2002. A comparative genomic analysis of two distant Diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. Genome Res. 12:57–66.

Chen X, Hiller M, Sancak Y, Fuller MT. 2005. Tissue-specific TAF counteract polycomb to turn on terminal differentiation. Science. 310:869–872.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Gen. 39:715–720.

Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol. 41:1152–1159.

Comeron JM. 1999. *K*-estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics. 15:763–764.

Crowley TE, Hoey T, Liu JK, Jan YN, Jan LY, Tjian R. 1993. A new factor related to TATA-binding protein has highly restricted expression patterns in *Drosophila*. Nature. 361: 557–561.

Deato MD, Tjian R. 2007. Switching of the core transcription machinery during myogenesis. Genes Dev. 21:2137–2149.

Drosophila 12 genomes consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Freiman RN, Albright SR, Zheng S, Sha WC, Hammer RE, Tjian R. 2002. Redundant role of tissue-selective $\text{TAF}_{\text{II}}105$ in B lymphocytes. Mol Cell Biol. 22:6564–6572.

Frontini M, Soutoglou E, Argentini M, Bole-Feysot C, Jost B, Scheer E, Tora L. 2005. TAF9b (formerly TAF9L) is a bona fide TAF that has unique and overlapping roles with TAF9. Mol Cell Biol. 25:4638–4649.

Gangloff YG, Sanders SL, Romier C, Kirschner D, Weil PA, Tora L, Davidson I. 2001. Histone folds mediate selective heterodimerization of yeast TAF(II)25 with TFIID components yTAF(II)47 and yTAF(II)65 and with SAGA component ySPT7. Mol Cell Biol. 21:1841–1853.

Gangloff YG, Werten S, Romier C, Carré L, Poch O, Moras D, Davidson I. 2000. The human TFIID components TAF(II)135 and TAF(II)20 and the yeast SAGA components ADA1 and TAF(II)68 heterodimerize to form histone-like pairs. Mol Cell Biol. 20:340–351.

Hernandez N. 1993. TBP, a universal eukaryotic transcription factor? Genes Dev. 7:1291–1308.

Hiller M, Chen X, Pringle MJ, Suchorolski M, Sancak Y, Viswanathan S, Bolival B, Lin TY, Marino S, Fuller MT. 2004. Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. Development. 131:5297–5308.

Hiller M, Lin TY, Wood C, Fuller MT. 2001. Developmental regulation of transcription by a tissue-specific TAF homolog. Genes Dev. 15:1021–1030.

Hochheimer A, Tjian R. 2003. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev. 17:1309–1320.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol. 7:900–904.

Isogai Y, Takada S, Tjian R, Keles S. 2007. Novel TRF1/BRF target genes revealed by genome-wide analysis of *Drosophila* Pol III transcription. EMBO J. 26:79–89.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Leurent C, Sanders SL, Demény MA, Garbett KA, Ruhlmann C, Weil PA, Tora L, Schultz P. 2004. Mapping key functional sites within yeast TFIID. EMBO J. 23:719–727.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature. 424:147–151.

Lin TY, Viswanathan S, Wood C, Wilson PG, Wolf N, Fuller MT. 1996. Coordinate developmental control of the meiotic cell cycle and spermatid differentiation in *Drosophila* males. Development. 122:1331–1341.

Liu WL, Coleman RA, Grob P, et al. (10 co-authors). 2008. Structural changes in TAF4b-TFIID correlate with promoter selectivity. Mol Cell. 29:81–91.

Mcdonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature. 351:652–654.

Metcalf CE, Wassarman DA. 2007. Nucleolar colocalization of TAF1 and testis-specific TAFs during *Drosophila* spermatogenesis. Dev Dynam. 236:2836–2843.

Rabenstein MD, Zhou S, Lis JT, Tjian R. 1999. TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family. Proc Natl Acad Sci USA. 96:4791–4796.

Richards S, et al. (52 co-authors). 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.

Saurin AJ, Shao Z, Erdjument-Bromage H, Tempst P, Kingston RE. 2001. A *Drosophila* polycomb group complex includes Zeste and dTAFII proteins. Nature. 412:655–660.

Shao H, Revach M, Moshonov S, Tzuman Y, Gazit K, Albeck S, Unger T, Dikstein R. 2005. Core promoter binding by histone-like TAF complexes. Mol Cell Biol. 25:206–219.

Struhl K, Moqtaderi Z. 1998. The TAFs in the HAT. Cell. 94: 1–4.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci USA. 98:7375–7379.

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol. 12:823–833.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. Mol Biol Evol. 21:36–44.

Thomas MC, Chiang CM. 2006. The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol. 41:105–178.

White-Cooper H, Schafer MA, Alphey LS, Fuller MT. 1998. Transcriptional and post-transcriptional control mechanisms coordinate the onset of spermatid differentiation with meiosis I in *Drosophila*. Development. 125:125–134.

Verrijzer CP. 2001. TFIID—not so basal after all. Science. 293:2010–2011.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA. 103:3220–3225.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555–556.

Marta Wayne, Associate Editor

Accepted February 3, 2009