**PIF-like Transposons are Common in Drosophila and Have Been Repeatedly Domesticated to Generate New Host Genes**

Claudio Casola, A. Michelle Lawing, Esther Betrán, and Cédric Feschotte

Biology Department, University of Texas, Arlington

The P instability factor or PIF superfamily of DNA transposons constitutes an important group of transposable elements (TEs) in plants, but it is still poorly characterized in metazoans. Taking advantage of the availability of draft genome sequences for twelve Drosophila species, we discovered 4 different lineages of Drosophila PIF-like transposons, named DPLT1-4. These lineages have experienced a complex evolutionary history during the Drosophila radiation, involving differential amplification and retention among species and probable events of horizontal transmission. Like previously described plant and animal PIF transposons, full-length DPLTs encode a putative transposase as well as a second predicted protein containing a Myb/SANT domain. In DPLTs, this domain is most closely related to the MADF DNA-binding domain found in several Drosophila transcription factors. In addition, we identified 7 distinct genes distributed across the Drosophila genus that encode proteins related to PIF transposases, but lack the hallmarks of transposons. Instead, these sequences show features of functional genes, such as an intact coding region evolving under purifying selection, the presence of orthologs in at least 2 Drosophila species, and the conservation of intron/exon structure across orthologs. We also provide evidence that most of these genes are transcribed and that some are developmentally regulated. Together the data indicate that these genes derived from PIF-transposons that have been “domesticated” to serve cellular functions. In one instance the recruitment of the transposase gene was accompanied by the co-recruitment of the adjacent second PIF gene, which raises the hypothesis that both proteins now function in the same pathway. The second PIF gene has retained the capacity to encode a protein with an intact MADF domain, suggesting that it may function as a transcription factor. We conclude that PIF transposons are common in the Drosophila lineage and have been a recurrent source of new genes during Drosophila evolution.

**Introduction**

Transposable elements (TEs) are genetic units found in nearly all eukaryotes that are able to move and amplify within a host genome. In some group of organisms, like mammals and grasses, TEs represent the single largest component of the genome, accounting for 40% to 80% of the nuclear DNA (Lander et al. 2001; Vitte and Bennetzen 2006). Eukaryotic TEs are usually divided into 2 main classes according to their mechanism of transposition. Class I or retroelements move via an RNA intermediate that is reverse-transcribed, while class II elements or DNA transposons move directly as DNA. In eukaryotes, DNA transposons transpose through a cut-and-paste mechanism whereby the element is excised and reinserted elsewhere in the genome (Craig et al. 2002). For most DNA transposon systems, this reaction only requires a single enzyme, the transposase (TPase), which is encoded by autonomous copies. Copies that do not encode the transposase, and therefore are non-autonomous, can still transpose if they carry the binding sites for the transposase. The binding sites are generally located within the terminal inverted repeats (TIRs) of the transposon. Other typical features of DNA transposons include flanking target site duplications (TSD) of conserved length that result from the nicking activities of the TPase at the site of chromosomal integration (Craig et al. 2002).

Approximately 10 superfamilies of eukaryotic DNA transposons are currently recognized based on sequence similarity, motifs in their TPases, TIR sequence and TSD length. The PIF/IS family, also known as Harbinger (Kapitonov and Jurka 1999; Zhang et al. 2001), is a recently discovered superfamily of DNA transposons first identified in maize (Walker et al. 1997; Zhang et al. 2001). It has been successively detected in the genomes of many flowering plants, some fungi and diverse animals, such as nematode, mosquito, sea urchin, tunicate and fish (Le et al. 2001; Zhang et al. 2001, 2004; Kapitonov and Jurka 2004). Most PIF-like transposons (PLTs) and the related Tourist-like miniature inverted-repeat transposable elements (MITEs) possess relatively short TIRs (12–40 bp long). PLTs cause 3-bp TSD, whose consensus is often TWA (where W stands for A or T). All potentially autonomous PIF-like transposons characterized so far appear to contain 2 transcriptional units encoding 2 distinct proteins: (i) the putative transposase (TPase), and (ii) an accessory protein containing a Myb/SANT domain (hereafter referred to as PIFFp2) (Kapitonov and Jurka 2004; Zhang et al. 2004). The TPase displays a motif similar to the catalytic acidic triad “DDE” shared by other transposases and integrases and is distantly related to transposases of the IS5 group of bacterial insertion sequences. The Myb/SANT domain is found in proteins involved in transcriptional regulation and chromatin remodeling (Aasland et al. 1996; Boyer et al. 2004). Typically, this domain provides sequence-specific DNA binding activity, but it may also mediate protein-protein interaction (Sterner et al. 2002; Ding et al. 2004; Mo et al. 2005). The activities of either PIF-encoded proteins have not been functionally investigated, but their presence and conservation in putative autonomous PIF-like transposons from a broad range of species suggest that both proteins participate in the life cycle of these elements.

The evolution of animal PIF-like transposons has not been analyzed in detail, but previous works suggest that they have a patchy taxonomic distribution. For example, PIF-like transposons have been identified in several invertebrates, including mosquitoes (Kapitonov and Jurka 2004), but none have been detected in the fruit fly Drosophila melanogaster, despite the availability of a high-quality genome.
sequence and 2 decades of intense TE mining in this species (Kapitonov and Jurka 2003; Quesnevse et al. 2005). Similarly, PIF-like transposons were readily identified in the genome of the pufferfish Takifugu rubripes and the zebrafish Danio rerio, but they have not yet been found in mammals or any other amniote (Aparicio et al. 2002; Kapitonov and Jurka 2004; Zhang et al. 2004). However, a PIF-like TPase seems to have been present in the common ancestor of vertebrates to create a new gene, HARBI1, which is highly expressed in the chicken and mammals (Kapitonov and Jurka 2004). The HARBI1 gene belongs to a growing list of TPase genes that have been “domesticated” to perform cellular functions (Volff 2006). However, no domesticated PIF-like genes have been reported in other animals, plant or fungi genomes. Thus, it is unclear whether this group of transposons significantly contributes to the emergence of new coding sequences, as previously described for other superfamilies of DNA transposons such as\textit{P}-element, hAT and Tc1/mariner (Volff 2006).

Here we took advantage of the genome sequencing of\textit{D. melanogaster} (Adams et al. 2000), \textit{D. pseudoobscura} (Richards et al. 2005) and 10 additional Drosophila species to investigate the presence and evolutionary history of the PIF superfamily in these insects. We show that PIF-like transposons (\textit{PLT}s) have colonized the genome of most Drosophila species, albeit with various success. We also present evidence that PIF-like transposase genes gave rise to at least 7 different domesticated genes during the Drosophila radiation. Finally, we report the first case of domestication of a PIFp2 protein, which was recruited into a MADF-like protein. Together these results indicate that PIF-like transposases have been a recurrent source of coding sequences for the emergence of new genes in Drosophila.

Materials and Methods

Database Searches

PIF-like sequences in Drosophila and other insects have been identified by similarity searches (blastn, tblastn) using the FlyBase BLAST server (http://flybase.bio.indiana.edu/blast/) and the NCBI BLAST servers (http://130.14.29.110/blast/, nr, est, gss and wgs databases). We used as initial queries the TPases from \textit{PIF}\textit{-like transposons} (\textit{PLT}s) that have been identified by similarity searches (blastn, tblastn) into a MADF-like protein. Together these results indicate that PIF-like transposases have been a recurrent source of coding sequences for the emergence of new genes in Drosophila.

Orthology Assignment and Gene Structure Prediction

Orthology of DPLGs was determined by assessing the synteny of flanking genes using the University of California at Santa Cruz (UCSC) Genome Browser Database (http://genome.ucsc.edu/); DPLGs were considered orthologs when the microsynteny was conserved on at least one side of the gene. The structure of each DPLG coding sequence was initially predicted using FGENESH (http://www.softberry.com/berry.phtml) and refined by multialignment with orthologous genes.

Sequence Analysis and Phylogenetic Inferences

Protein and nucleotide multialignments were performed using MAFFT package (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/), T-Coffee (http://gs-server.cnrs-mrs.fr/Tcoffee/tcoffee.cgi/index.cgi) and CLUSTALX 1.83 (Chenna et al. 2003), and edited with Bioedit v7.0.5.3 (Hall 1999). Phylogenetic inferences were obtained using the neighbor-joining and parsimony methods implemented in MEGA 3.1 (Kumar, Tamura, and Nei 2004), and the Bayesian approach implemented in MrBayes (Ronquist and Huelsenbeck 2003). For the Bayesian analyses, we used the mixed amino acid model, with 4 chains running for 500,000 generations and sampling every 100 generations. Convergence was attained with standard deviation of split frequencies <0.01, and all branch potential scale reduction factors approached unity. A consensus tree was estimated by using a “burnin” parameter of 1250 trees (25% of 5,000 samples). Nucleotide divergence between DPLT2 elements from \textit{P. persimilis}, \textit{D. pseudoobscura}, \textit{D. willistoni} and \textit{D. mojavensis}, and \textit{Adh}, \textit{yellow} and \textit{RPL18} in the same 4 species, were calculated over the entire length of transposons (Tamura-Nei method) and the coding sequence of genes (synonymous sites, \textit{Kumar} method) using MEGA 3.1 (Kumar et al. 2004). Domain searches were carried out on protein sequences of PIF-like TPases and PI fp2, and PIF-derived genes using the SMART (http://smart.embl-heidelberg.de/) and InterPro (http://www.ebi.ac.uk/interpro/) databases. Putative helix-turn-helix motifs were predicted by the NPS@ software (Dodd and Egan 1990). Secondary structures were predicted using JPRED (http://www.compbio.dundee.ac.uk/~www-jpred/).

GC-content Analysis

GC-content for the whole coding sequence and first, second and third codon position was calculated by the FREQSQ software (http://bioinfo.hku.hk/services/analyseq/cgi-bin/freqsq_in.pl). Plots of GC percentages for DPLGs, DPLTs and average genome coding sequences for each
species, as well as the equiprobability ellipse for D. pseudoobscura genes, were drawn using STATISTICA (StatSoft 2001). To compare the GC-content of DPLGs and DPLTs to the rest of the genome coding regions, we performed a randomization test. The coding region sequences of the D. pseudoobscura FlyBase genes annotated in the November 2004 dp3 assembly were downloaded from the University of California at Santa Cruz Genome Browser Database (http://genome.ucsc.edu/). From the total 9,946 retrieved genes, we eliminated 98 sequences containing stretches of N (gaps). We calculated the difference (dDPLG) between the average GC-content for 5 DPLGs and the average GC-content of the rest of the genes in the genome for the whole gene and first, second and third codon positions. We wrote a C program to randomly sample 5 D. pseudoobscura coding sequences from the 9,848 retrieved genes and to calculate the statistic (d) that is the difference between the average GC-content of each random sample and each set of the remaining genes. The program performed 10,000 permutations and provided a distribution for the d statistic. We then calculated the p-value by counting how many times in the distribution we obtained a value of d smaller of equal to dDPLG and divided that by the number of permutations. The same randomization test was carried out for the 4 DPLTs changing the size of the sample to 4.

**DPLG Codon Substitution Pattern Analysis**

The evolutionary dynamics of codon substitutions were estimated using the CODEML program of PAML v3.15 package (Yang 1997). For each DPLG group, we obtained a multialignment of the coding region with the MAFFT package (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/), and eliminated ambiguity sites. We used an input unrooted tree and the equilibrium codon frequencies as calculated from the average nucleotide frequencies at the 3 codon sites (F3X4 option).

**Amplification, Cloning and Sequencing of D. persimilis and D. willistoni DPLG1**

D. persimilis and D. willistoni genome sequence strain were obtained from Tucson Stock Center. Genomic DNA was extracted from 15 females using the Puregene™ kit (Gentra Systems, Minneapolis, MN). PCRs were performed using the primers Dper_PLG1-F1 (5’-CAAGAGAACGCCAAGAGGTTT-3’) and Dper_PLG1-R1 (5’-CTTTGC-TGAACCGAACGATCC-3’) designed at position 1246–1268 and 1595–1616 of the D. persimilis DPLG1 ortholog, and the primers Dwil_PLG1-F1 (5’-GCCAAT-CAGAGAATCAAGTGCC-3’) and Dwil_PLG1-R1 (5’-GCCAAT-CAGAGAATCAAGTGCC-3’) designed at position 246–269 and 1227–1246 of the D. willistoni DPLG1 ortholog. Twenty ng of genomic DNA were used for the following amplification reactions: initial denaturation of 3’ 94°C, 35 cycles of amplification of 30’ 94°C, 30’ 52°C and 1’ 72°C, and final extension of 7’ 72°C. The single-band PCR product was purified using the QIAquick™ kit (QIAGEN Group, Valencia, CA), and sequenced by an ABI automated DNA sequencer (Applied Biosystems, Carlsbad, CA) with fluorescent DyeDeox terminator reagents.
Table 1 Characteristics of the Four Drosophila PIF-like Transposon Clades

<table>
<thead>
<tr>
<th>Lineage</th>
<th>TE Length</th>
<th>Copy Number</th>
<th>TIRs Length</th>
<th>TSD</th>
<th>TPase Length</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPLT1</td>
<td>2,300–2,900</td>
<td>&lt;10, several hundreds</td>
<td>28–64</td>
<td>TWA</td>
<td>~400</td>
<td>Dyak, Dpse, Dper, Dwil, Dmel, Dsim, Dsec, Dere, Dmoj</td>
</tr>
<tr>
<td>DPLT2</td>
<td>915–2740</td>
<td>&lt;10, 50</td>
<td>35–40</td>
<td>TWA</td>
<td>~410</td>
<td>Dpse, Dper, Dwil, Dmoj</td>
</tr>
<tr>
<td>DPLT3</td>
<td>~2,660</td>
<td>~100</td>
<td>15–23</td>
<td>TWA</td>
<td>~360</td>
<td>Dpse, Dper</td>
</tr>
<tr>
<td>DPLT4</td>
<td>~2,500</td>
<td>~100–150</td>
<td>18</td>
<td>AWW</td>
<td>~350</td>
<td>Dpse, Dper</td>
</tr>
</tbody>
</table>

* Referred to the consensus sequence.
* Species abbreviations: Dyak: Drosophila yakuba; Dpse: D. pseudoobscura; Dper: D. persimilis; Dwil: D. willistoni; Dmel: D. melanogaster; Dsim: D. simulans; Dsec: D. sechellia; Dere: D. erecta; Dmoj: D. mojavensis.
* Underlined species show only PIF miniature inverted-repeat transposable elements (MITEs).

Results

Several Lineages of PIF-like Transposons Are Present in Drosophila

We initiated this study by carrying out reiterative similarity searches with queries representing PIF-like TPases from the mosquito A. gambiae, the sea squirt Ciona intestinalis and the zebrafish Danio rerio, deposited in Repbase as “Harbinger” elements (Jurka et al. 2005), against the twelve Drosophila genomes using the FlyBase BLAST server (http://flybase.bio.indiana.edu/blast/). These searches led to the identification of numerous PIF-like transposons (PLTs) in the genome of several Drosophila species, which we named Drosophila PIF-like transposons or DPLTs. Four species (see table 1) were found to contain PLTs with TPase coding sequences, while 5 other species had related MITEs and other incomplete elements with no detectable coding capacity. Only the longest sequences were further characterized. Within each species, PLT sequences were grouped into families on the basis of (i) their sequence similarity (members of the same family share >85% similarity over their entire length) and (ii) phylogenetic clustering, where members of the same family form a monophyletic group supported by at least 75% of bootstrap values (data not shown). This step resulted in the definition of 11 distinct families. For each PLT family, the retrieved copies were aligned to derive a consensus sequence (available upon request). To infer the phylogenetic relationships among the different DPLT families and with other members of the PIF superfamily, we aligned the 11 putative PLT TPases from Drosophila and those from other insects with previously described animal and plant PIF-like TPases (fig. 1A). Phylogenetic trees obtained using different methods (neighbor-joining, parsimony and Bayesian) reveal very similar topology, wherein PLTs from animals (deuterostomes and protostomes) fall into either 1 or 2 well-supported clades that are distantly related to plant PLTs (fig. 1A). According to the phylogeny, the 11 DPLT families can be grouped into 4 distinct lineages. Three lineages (DPLT1-3) fall within 1 of the 2 animal clades together with several PLTs from the 2 mosquito species, the zebrafish Danio rerio and the tunicate Ciona intestinalis. A 4th lineage of Drosophila elements (DPLT4) falls within the second animal clade together with PLTs from various insects such as A. gambiae (but not A. aegypti), B. mori and T. castaneum, and 2 additional PLT lineages from zebrafish. Thus, distantly related PLT lineages are found to co-exist within the same genome in both deuterostome and protostome species, suggesting that the PIF superfamily underwent ancient episodes of diversification in an early animal ancestor.

Diversification of PLTs in Drosophila

Within species, DPLTs are relatively young, with pairwise nucleotide sequence divergence ranging from 2% to 15% between copies of the same family. D. yakuba and D. willistoni seem to harbor the most recently active elements (all from the DPLT1 lineage) because some copies located at different chromosomal locations are almost identical. When DPLTs from different species are compared, a wide range of sequence diversity is observed, either between but also within the same DPLT lineage. For example, TPases from the same DPLT lineages but from different species share from 40% to 99% amino acid identity and there is only 13% to 29% identity between TPases from different DPLT lineages. Likewise, the TIRs of DPLT are relatively well conserved within the same lineage, but greatly diverge when different lineages are compared (fig. 2). These data are consistent with an ancient diversification of PLTs in animals and a complex history of these elements during the Drosophila radiation, involving vertical propagation and subsequent diversification. DPLTs have also experienced differential amplification and retention during Drosophila evolution (table 1). For instance, DPLT3 and DPLT4 are present only in the sibling species D. pseudoobscura and D. persimilis, while members of the DPLT1 lineage occur in 9 Drosophila species and show a higher level of diversity. The abundance of DPLTs and the success of individual families within a species are also highly variable, with copy number ranging from less than 10 copies in the DPLT2 lineage to several hundred for the DPLT1a subfamily in D. willistoni (table 1).

Horizontal Transfers of PLTs Between Drosophila Species

Horizontal transfer events also appear to have contributed to the propagation of DPLTs. To illustrate this, we turn our attention to the DPLT2 lineage. Members of this lineage are found in distantly related species like D. pseudoobscura, D. willistoni and D. mojavensis, but the level of identity between copies from different species that diverged about 60 to 63 Mya (Tamura et al. 2004) is unexpectedly high. For instance, the DPLT2 consensus sequences of D. pseudoobscura and D. willistoni are 93% identical over
their entire nucleotide sequence. Sequence similarity is elevated throughout the entire sequence of the elements, including non-coding subterminal regions (suppl. fig. 1), which are known to evolve relatively rapidly in DNA transposons (Zhang et al. 2004; Feschotte et al. 2005; Diao et al. 2006). A similar level of conservation (~90%) is observed when the DPLT2 elements from D. mojavensis are compared to those from either D. pseudoobscura or D. willistoni (note, however, that in this case the D. mojavensis consensus is only 914 bp long). Indeed, the nucleotide divergence of DPLT2 elements among the 3 species is 1.6 to 4.7 times lower than the nucleotide divergence of 3 orthologous nuclear genes evolving under strong purifying selection (Adh, yellow and RPL18) from the same species (see Materials and Methods, data not shown). Two of these genes, Adh and yellow, were chosen because their substitution rate has been extensively studied in Drosophila (see, for example, Tamura et al. 2004). The third gene, RPL18, encodes a ribosomal protein that is highly conserved among the 3 species. Thus, the most parsimonious hypothesis to explain the high level of sequence conservation between DPLT2 elements invokes recent horizontal transfer(s) of these elements among D. pseudoobscura, D. willistoni and D. mojavensis or their close relatives or their proximate ancestors. In support of this hypothesis, we note that the geographical range of D. persimilis, D. pseudoobscura and D. mojavensis is overlapping in the southwestern of United States, and D. pseudoobscura occurs in sympathy with D. willistoni in central America (Ashburner et al. 1982; Ruiz et al. 1990).

**Coding Capacity of DPLTs**

In previously described PIF-like transposons, the predicted TPase gene is interrupted by 1 to 3 introns (Kapitonov and Jurka 2004; Zhang et al. 2004), a feature shared by putative autonomous Drosophila PIF-like transposons (fig. 1B). The predicted TPases encoded by animal PIF transposons, comprising DPLTs, vary in length from 340 to 420 amino acids, and share a 35–45% of inter-clade similarity (table 1).

In addition to the TPase, putative autonomous PIF transposons encode a second protein, PIFp2, which contains a N-terminal region with similarity to the Myb/SANT domain (Kapitonov and Jurka 2004; Zhang et al. 2004). Gene prediction tools revealed that each DPLT group also contains a second putative gene on the opposite strand relative to the TPase gene. In DPLT1 and DPLT2 lineages, this gene seems to be formed by 2 exons, with the most downstream exon nested in the TPase gene intron (fig. 1). The same overlapping organization of TPase and PIFp2 genes has been found in the A. gambiae Harbinger element (Kapitonov and Jurka 2004), but is not observed in other animal or in plant PIF transposons (data not shown) (Zhang et al. 2004; Jurka et al. 2005). Searches of the protein domain databases (SMART) indicate that the second ORF is predicted to encode a peptide with significant similarity to the MADF domain (Myb/SANT-like domain in Adf-1). The MADF domain is a distant relative of the Myb/SANT domain and it is found in a family of proteins that has
mostly expanded in arthropods (England et al. 1992; Bhaskar and Courey 2002; Zimmermann et al. 2006). In sum, DPLTs seem to contain 2 separate genes, 1 of which would encode for the putative TPase, while the other could encode a MADF-containing protein, which we refer to as PIFp2, following the annotation of other PIF-like transposons in Repbase (Jurka et al. 2005).

Detection of 7 Different PIF TPase-derived Genes in Drosophila (DPLG)

In addition to the DPLT lineages described above, we identified 7 distinct (i.e. non-orthologous) single-copy sequences that can potentially encode a protein similar to the PIF TPase, but appear to represent stationary host genes (table 2). We designate these putative genes DPLG1-7 (Drosophila PIF-like gene 1-7). DPLG1-4 have been annotated in the D. melanogaster genome as genes CG12253, CG32187, CG32095 and CG7492, respectively, and the homologs predicted in the D. pseudoobscura genome as genes GA11511, GA16677, GA16674 and GA20390.

We systematically inspected the flanking sequences of highly similar sequences in conserved microsyntenic regions of the other 10 Drosophila species (see Materials and Methods), therefore likely representing orthologs of DPLG1-4. DPLG5-7 have not been annotated in any Drosophila genome, although some of them were predicted according to certain gene models depicted in the UCSC Genome Browser. We could identify orthologs for each of these 3 genes in at least 2 Drosophila species. They occur predominantly in D. pseudoobscura and D. persimilis, a distribution that mirrors those of the DPLT lineages (see below). The following sections each provide an independent line of evidence that DPLGs represent bona fide protein-coding genes derived from PIF transposons at different times during Drosophila evolution and that have now acquired a cellular function.

Absence of Structural Hallmark of Transposons Associated with DPLGs

We systematically inspected the flanking sequences of all DPLGs for typical structural hallmark of PIF-like transposons, such as TIRs and TSD and in all cases we were unable to detect any of these features or their remnants. In contrast, these features could be readily identified for all DPLTs (table 1). Furthermore, blastn and tblastn searches of each species’ genome with individual DPLG sequences failed to retrieve any other closely related paralogous sequence, indicating that each DPLG, when present, likely occur in single copy per haploid genome. The only exception was a partial paralogous copy of DPLG2 in D. erecta (corresponding to the first 447 bp) present in another genomic region, which can be attributed to a segmental duplication that also encompasses an unrelated gene (the putative ortholog of D. melanogaster CG32191) located upstream of DPLG2. In contrast, all TPase-encoding DPLT families are represented by at least 3 and often many more copies interspersed in the genome, consistent with their recent mobility.

Structure and Sequence Conservation of DPLGs

A second line of evidence supporting the domestication of DPLGs resides in their high level of conservation both in sequence and structure across Drosophila species. Sequence conservation is evident from a neighbor-joining phylogenetic analysis of each DPLG protein across all the representative species (fig. 3). First, the topologies of the

---

**Table 2**

**Drosophila PIF-like Genes Features**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Location</th>
<th>Protein Length</th>
<th>Catalytic Triad</th>
<th>Distribution</th>
<th>Expression (No. of ESTs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG12253</td>
<td>GA11511</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPLG2</td>
<td>Chr3L: 17770199–17771624</td>
<td>409–423</td>
<td>F<a href="35">F/N/S</a> E/D/N</td>
<td>Drosophila</td>
<td>Dmel (3)</td>
</tr>
<tr>
<td>CG32187</td>
<td>GA16674</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7492</td>
<td>GA20390</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG4922</td>
<td>GA26920</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a FlyBase annotation in D. melanogaster (CG) and D. pseudoobscura (GA).

b DPLG1–4: D. melanogaster Muller elements. DPLG5 and DPLG7: D. pseudoobscura Muller elements and groups. DPLG6: scaffold of D. virilis for (Release droVir2).

c Conserved catalytic amino acids are in bold, alternative first residues are in square brackets.

d Species abbreviation as in table 1, except: Dana: D. ananassae; Dvir: D. virilis; Dgri: D. grimshawi; Gmor: Glossina morsitans.

---
resulting trees are in good agreement with the widely accepted species tree (Tamura, Subramanian, and Kumar 2004). This is in contrast to transposon gene phylogenies, which are often at odds with species trees due to horizontal transfers and frequent lineage sorting (Robertson and Lampe 1995; Capy et al. 1998; Sanchez-Gracia et al. 2005). Second, the branch lengths in each distance tree are comparable to those generated in phylogenies of well-conserved Drosophila genes of known cellular function (see example of Adh in fig. 3). Such a level of sequence conservation likely reflects strong functional constraints acting on DPLG-encoded proteins (see below).

Furthermore, different DPLGs have distinct exon/intron structure, but the structure is well conserved in DPLG orthologs. Gene structure predictions are supported by several spliced EST sequences and sequence alignments of intron/exon boundaries (fig. 3 and data not shown). The only substantial structural diversity was found among DPLG7 orthologs, which can be separated into 2 groups with distinct exon/intron organization (fig. 3). DPLG7A, which is found in D. pseudoobscura, D. persimilis and D. willistoni, has a single intron, while DPLG7B, present in the 3 species of the Drosophila subgenus D. virilis, D. mojavensis, and D. grimshawi, displays a second intron splitting the downstream exon. Presumably, this variation can be explained by a single intron gain/loss in one of the ancestor of these species. Note that DPLG7A and B are also found at different chromosomal positions, but this is most likely due to the relocation of DPLG7B in the common ancestor of the Drosophila subgenus (see below). A second minor structural change occurred in the D. willistoni/DPLG1 ortholog, where the second exon is split by a 58 bp intron. After re-sequencing this genomic region of the D. willistoni sequenced strain (see Materials and Methods), we found no difference from the deposited assembly and therefore we concluded that this specific gene organization is a derived trait of DPLG1 in D. willistoni.

Another significant observation that serves to distinguish DPLGs from the transposons is the fact that all 60 DPLG orthologs examined in this study display intact coding regions that seem to encompass the entire ancestral TPase sequence (from 374 to 588 amino acids), while almost all of the TPase genes examined in DPLTs had obvious disabling mutations introducing 1 or several premature stop codons. It should be noted that we initially detected 2 instances of single nucleotide insertion/deletion that had apparently disabled the coding region of 2 different DPLGs. First, the D. persimilis DPLG1 ortholog had an insertion of an adenosine at position 804 based on its comparison to the 98% identical D. pseudoobscura DPLG1 coding region. However, PCR amplification and re-sequencing on both strands of the 2 regions using DNA extracted from D. persimilis individuals of the same strain revealed no interruption in the DPLG1 ORF (see Materials and Methods). Second, the sequence assembly of the D. simulans DPLG3 ortholog shows a single base-pair deletion at position 1216 in the coding region in reference to D. melanogaster DPLG3. However, this deletion is absent from 3 out of 4 D. simulans raw sequence reads overlapping with DPLG3 that we retrieved from the NCBI traces database. We conclude that in both cases, the disabling mutations were sequencing or assembly artifacts and all DPLGs are therefore devoid of obvious disabling mutations. Considering the broad taxonomic distribution of some DPLGs and therefore their ancient origin, their coding integrity as transposon genes would be extremely unlikely in the absence of selective constraints. Thus, the most likely explanation is that they are not transposon genes anymore, but functional host genes.

Expression Pattern of DPLGs

Based on the presence of matching cDNA and ESTs in various Drosophila species, we could find evidence for the transcription of 6 out of 7 DPLGs (all but DPLG5) (table 2). Overall, transcription data is much more abundant for D. melanogaster and relatively scarce for the other species, and therefore it is not surprising that the 4 genes present in D. melanogaster received the most supporting evidence for transcription. We focused on the expression data of DPLG1-4 in D. melanogaster and could draw several interesting points. First, the 4 genes received different amounts of EST support, from 3 matching ESTs (DPLG2) to 21 (DPLG3). Based on the tissue and developmental stages from which the ESTs were cloned, DPLG1 and 2 appear to be mostly (if not only) transcribed during larval development, while DPLG3 and DPLG4 ESTs cover a broader developmental spectrum, ranging from embryos, larvae, metamorphic stages to adult head and gonads. Developmental profiling of D. melanogaster derived from microarray analysis retrieved from the UCSC Genome Browser is in good agreement with the EST data. It shows a marked down-regulation of DPLG1 activity in most stages, except during the mid-phase of larval development, while both DPLG3 and DPLG4 are intensively expressed during early embryogenesis and most subsequent developmental stages, as well as in the adult (fig. 4). Together, the data suggests that at least some of the DPLGs are transcribed and are likely subject to distinct developmental regulation.

GC-content of DPLGs and DPLTs

Previous analyses highlighted that the coding regions of transposable elements contain a lower percentage of GC than the genes of their host species. This discrepancy is particularly significant in the GC-rich genome of D. melanogaster (Lerat et al. 2002). In this species, this bias is also accompanied by a strikingly different codon usage between TE genes and host genes, regardless of their level of expression (Lerat et al. 2002). Therefore we reasoned that a comparison of the GC-content (%GC) of DPLTs, DPLGs and other Drosophila genes might bring further support to the notion that DPLGs are domesticated genes. We computed the %GC of the entire coding sequence of all DPLGs and DPLTs (TPase gene) separately for the 3 codon positions and compared them with the average %GC values calculated for all known genes (non-TEs) of the same Drosophila species (available from the “Codon Usage Database” at http://www.kazusa.or.jp/codon/). As DPLT clades are formed by closely related elements within each species, we used the reconstructed consensus for this analysis. However, we tested coding regions from several
Fig. 3.—Phylogenetic tree and the gene structure of each DPLG gene family with transposases from the four DPLT lineages and the human protein HARBI1. The multialignment of the proteins have been created using the L-INS-i algorithm of the MAFFT package, and edited to remove regions poorly conserved and gaps, giving a final alignment of about 200–250 residues. The phylogeny of Adh orthologs from the twelve Drosophila species is shown at the bottom right corner, together with four Adhr genes as outgroups. Each tree has been built using the neighbor-joining method implemented in the MEGA3.1 software package. Numbers on the nodes show bootstrap values after 1000 replicates. Exons are represented by bars, introns by the symbol “^”. Abbreviations as in table 1, figure 1 and 2, except: Hsap:HARBI1: Homo sapiens protein encoded by the HARBI1 gene. The trees are drawn to scale.
transposon copies, and observed no significant difference from the analyses carried out on the consensus sequences (data not shown).

A comparison of the GC-content revealed that DPLGs and the species-specific genes average group together, while DPLTs TPase genes form a separate cluster (fig. 5, suppl. fig. 2). The unusually low GC-content of coding regions in D. willistoni is probably responsible for the less striking difference in %GC between DPLTs, some DPLGs and its gene average observed in this species (suppl. fig. 2). Interestingly, DPLG1 behaves differently from the other DPLGs, showing GC values comparable to DPLTs. However, we noticed that several other genes located in the same genomic environment of DPLG1 were also characterized by a similarly low GC-content (data not show). Thus, the different nucleotide composition of DPLG1 may reflect peculiar selective forces acting locally to maintain a relatively low GC-content in this region of the genome. In order to determine the statistical significance of the observed difference between the GC-content of DPLGs, and DPLTs TPase genes, we performed 2 different analyses on the sequences obtained from D. pseudoobscura (see Materials and Methods). D. pseudoobscura is the only species with relatively accurate gene annotation where sufficient amount of DPLGs, and DPLTs TPase genes were available to perform these analyses. First, we drew a 95% equiprobability ellipse of the GC-content (in %) for the first and the third codon position of 9,848 D. pseudoobscura genes (see Materials and Methods). This is the ellipse that gives the 95% equiprobability contour for the bivariate distribution. We observed that all DPLTs as well as DPLG1 fall outside of the ellipse (data not shown). Second, we calculated the difference (dDPLG) between the average GC-content (in %) for 5 DPLGs and the average GC-content of the rest of the genes in the genome either for the entire gene (dDPLG (whole) = −3.18%) or separately for the first, second and third codon positions (dDPLG (first) = −0.12%; dDPLG (second) = −3.55%; dDPLG (third) = −5.84%). DPLG1 has not been included in this analysis as its GC-content deviates from the other DPLGs due to the local genomic environment as discussed above. The randomization test (see Materials and Methods) reveals that these genes do not behave significantly differently from the rest of the other genes in the genome (suppl. table 1). We also calculated the difference (dDPLT) between the average GC-content for the 4 predicted DPLT TPase genes and the average GC-content of the genes in the genome for the whole gene and for the first, second and third codon positions. The randomization test reveals that DPLT genes significantly differ from host genes (suppl. table 1). They have significantly lower GC-content for the whole gene, and for the first and third codon position (dDPLTs (whole) = −18.77%; dDPLT (first) = −16.54%; dDPLT (second) = −5.17%; dDPLT (third) = −34.57%) (see suppl. table 1).

Selection Regime Operating on DPLGs

Previous studies have shown that after their propagation within a genome, TPase genes evolve under no functional constraints following a neutral model, akin to pseudogenes, and therefore they rapidly accumulate mutations that lead to their inactivation (Witherspoon 1999; Lampe et al. 2003; Silva and Kidwell 2004). In contrast, if DPLGs are bona fide host genes with a cellular function, they are expected to be evolving under either purifying or positive selection. To test this hypothesis, we evaluated the ratio of non-synonymous substitutions (Ka) to synonymous substitutions (Ks) within each gene lineage using maximum-likelihood analyses (Yang 1997). A Ka/Ks value close to 1 is considered a valid indicator of neutral evolution, whereas Ka/Ks<1 or Ka/Ks>1 indicates that the analyzed sequences underwent purifying (negative) or diversifying (positive) selection, respectively. Using the CODEML algorithm implemented in the PAML package (Yang 1997), we applied a likelihood ratio test (LRT) to compare the likelihood of 2 different evolution models.
for each group of DPLG orthologs in the Drosophila lineage. The first model, which assumes that the DPLG orthologs are neutrally evolving coding sequences (Ka/Ks fixed to 1), was rejected for every gene group. The second model, which assumes a single Ka/Ks value for each gene tree (1-ratio model) was statistically more likely than the neutral model and Ka/Ks estimates take values between 0.05 and 0.177 for each orthologous gene group (suppl. tables 2 and 3). Together, these results indicate that all 7 DPLGs have evolved under strong purifying selection.

To complement these analyses, we also tested a free-ratio model, which allowed for a separated estimation of Ka/Ks in each branch of the tree. This model is significantly better than the 1-ratio model for each DPLG group of orthologs except for DPLG5 (suppl. table 3). This data is indicative of heterogeneity in the rates at which different lineages are evolving. Nonetheless, in the trees obtained under the free-ratio model the Ka/Ks values were mostly lower than 0.1 (branches with an insufficient number of substitutions are not considered as they produce statistically not valuable Ka/Ks estimation), confirming that DPLGs evolved under strong purifying selection in most of the Drosophila lineages under consideration (suppl. fig. 3). However, we note that Ka/Ks can vary up to 10 fold between lineages under purifying selection, in the range of 0.02 to 0.2, which suggest that DPLGs have experienced alternate episodes of highly constrained evolution with episodes of more relaxed or positive selection.

Evolutionary History and Origin of DPLGs

The presence of DPLG1-4 at orthologous position in all 12 Drosophila species demonstrate that these genes originated at least prior to the Sophophora/Drosophila split, dated at −63 Mya (Tamura et al. 2004). Moreover, searches of all sequence databases currently available at GenBank revealed a likely homolog of DPLG1 and DPLG4 in the tse-tse fly Glossina morsitans. These are no genomic copies of these genes in the databases, but we identified 5 ESTs encoding for a protein closely related to the Drosophila DPLG1 and DPLG4 alignment shown later in figure 3 (Materials and Methods). These ESTs were aligned to reconstruct the complete coding region of a putative full-length DPLG1 homolog sharing 50% nucleotide identity and 63% amino acid similarity with the D. melanogaster DPLG1. This level of conservation together with phylogenetic analysis (fig. 3) suggests that the G. morsitans sequence is most likely an ortholog of the DPLG1 gene. Another EST from G. morsitans encodes a fragment of coding sequence that aligns with 70% similarity over 110 amino acids with the N-terminal region of Drosophila DPLG4 protein. Thus, DPLG1 and DPLG4 most likely originated from a PIF transposon domesticated prior to the divergence of the Drosophila and Glossina dipterans.

In contrast to DPLG1-4, DPLG5-7 have a more patchy phylogenetic distribution in Drosophila. However, if the phylogeny of the host species is correct (and it is currently well accepted), the current distribution suggests that these genes most parsimoniously arose at a relatively ancient time, but were subject to loss in certain lineages (see fig. 6). DPLG5 seems to have emerged in the Sophophora subgenus, prior to the divergence of the melanogaster and obscura species groups, but was subsequently lost from the melanogaster subgroup. DPLG6 is present as a seemingly intact gene only in D. grinshowi and D. virilis, but DPLG6 sequence relics are detectable at orthologous positions in D. mojavensis, D. pseudoobscura and D. persimilis, which indicates that DPLG6 may have originated prior to the Sophophora/Drosophila subgenus split, but was subsequently lost from most—if not all—lineages of the Sophophora subgenus. Finally, DPLG7 was likely recruited prior to the Sophophora/Drosophila subgenus, and seems to have been maintained in most lineages, except the melanogaster group. Hence, all DPLGs originated at least −55 Mya (Tamura et al. 2004).

In order to investigate the relationship of Drosophila PIF-like TPases and DPLG proteins, we used the multiple alignment shown later in figure 8 for phylogenetic reconstruction using different methods (see Materials and Methods). Neighbor-joining and parsimony methods provided trees where most of DPLG form a single or a few monophyletic clades with low statistical support and separated from PLTs (data not shown), providing poor phylogenetic resolution and little insight into the relationship of DPLG proteins with DPLT TPases. We interpret these results as a consequence of long-branch attraction artifacts that could not be resolved by these phylogenetic methods. In contrast, the Bayesian analysis (fig. 7) yielded a tree with a well-supported topology where DPLG form at least 3 distinct groups with different origins. DPLG1 groups with clade 2 of animal PLTs, while DPLG4, 5 and 6 are nested within the clade 1 of PLTs. The 3 remaining DPLG clusters together in a separate monophyletic group that cannot be directly allied with a particular group of PLTs. These results suggest that DPLGs arose from at least 3 independent domestication events. Diversification of DPLG2, 3 and 7 and of DPLG4, 5 and 6 may imply additional domestication events or may have occurred through gene duplication. Interestingly, none of the DPLGs appear to be directly descended from extant Drosophila PIF-like transposons, although DPLG1 and DPLG6 seem to share a common origin with PLTs from other insects (fig. 7). These observations indicate that DPLGs derived from PLTs that are now
extinct in the 12 Drosophila species examined in our study. This is not unexpected given the relatively ancient origin of DPLGs and the rapid turnover of TEsin Drosophila (Petrov 2002; Lerat et al. 2003).

Conserved Motifs and Domain Structure of PIF-like TPases and Possible Functions of the Derived DPLG Proteins

The predicted DPLG proteins have retained only 15–30% of sequence identity and 25–50% of sequence similarity to DPLT TPases. It was thus of interest to determine whether some of the original TPase regions or motifs have been preferentially preserved or eliminated in the DPLG proteins. To address this question, we first aligned 16 PIF TPases encoded by various DPLTs and 7 PIF-like transposons from vertebrates, A. gambiae and 2 plants (Oryza sativa and Arabidopsis thaliana) and use this alignment to identify 8 most conserved motifs scattered throughout the entire TPase sequences (a WebLogo consensus of each motif is reported in suppl. fig. 4). These 8 regions are largely overlapping with the 6 motifs previously identified in TPases from euakaryotic PIF transposons and bacteria IS5-like elements by Kapitonov and Jurka (2004), and the 4 motifs H, N2, N3 and C1 recognized in the plant PIF TPases by Zhang et al. (2004).

Next, we added the DPLG proteins to the alignment of PIF TPases and assess the presence and conservation of the 8 conserved motifs in the DPLG proteins (fig. 8). DPLG1-4 had retained only half of the 8 conserved motifs. DPLG2 and DPLG3 have lost part of conservation observed in the N-terminal region of PIF TPases, as pointed out by the absence of motif 2 and a highly divergent or incomplete motif 3. Four DPLG proteins also lack the first part of motif 4, which is one of the most highly conserved in PIF TPases. Thus, it appears that some conserved motifs that were presumably important for TPase function(s) have been repeatedly and independently lost during the evolution of DPLG proteins.

It has been proposed previously that PIF TPases contain a conserved DDE triad functionally similar to the catalytic acidic triad characteristic of the DDE TPase/integrate supergroup. This triad serves to coordinate metal ions that are involved in catalysis of the cleavage and strand transfer reactions. Almost all substitutions experimentally introduced at these conserved residues (especially in the first and second asparate) in a variety of TPases and integrases result in complete or partial loss of these activities (Haren et al. 1999; Craig et al. 2002). In metazoan PIF TPases, the last 2 residues are separated by 35, 36 or 37 amino acids in different transposon clades (Kapitonov and Jurka 2004; Zhang et al. 2004) (fig. 8), a spacing comparable to other TPases/integrases (Haren et al. 1999). On the other hand, the position of the first amino acid of the catalytic triad is ambiguous as all PIF TPases possess 2 different highly conserved aspartate residues in the correspondent position (fig. 8) (Kapitonov and Jurka 2004). Nevertheless, it is striking that all consensu PIF TPases possess an intact DDE triad, while none of the DPLG proteins display an intact DDE signature (table 2 and fig. 8). Hence, it is likely that DPLG proteins have lost at least some of their ancestral catalytic activities and thus may have been recruited for function unrelated to catalysis.

All TPases that have been functionally examined so far are known to use a N-terminal region to bind specifically to short DNA sites located near the termini of the cognate transposons (Craig et al. 2002). In several TPases, DNA-binding activity requires 1 or 2 helix-turn-helix (HTH) motifs located within the N-terminal region of the TPase (Feschotte et al. 2005). A putative HTH motif is computationally predicted in the N-terminal region of plant PIF TPases (Zhang et al. 2004), but no biochemical data are available concerning the actual DNA-binding activity of these proteins. We used the HTH prediction method of Dodd and Egan (Dodd and Egan 1990) to screen for the presence of potential HTH motif(s) in all the DPLG proteins and animal PIF TPases examined in this study. These analyses predict a single HTH motif with moderate to strong confidence score in 17 proteins out of 24. When predicted, the HTH motif is located at the same relative position in
a multiple alignment of the proteins (fig. 8), despite relatively weak conservation of the region at the primary sequence level. This observation strengthens the individual computational HTH predictions. To further validate the HTH predictions, we determined the putative secondary structure of PIF-like proteins using the JPRED program (Cuff et al. 1998). Two helices separated by a short linker are predicted at the same position than the predicted HTH motif in all the PIF-like proteins, except for the TPasses of the DPLT4 group. The first helix is 7–10 residues long and is located between conserved motifs 1 and 2, the second helix is usually 18 amino acid long and overlap almost perfectly with the second motif (fig. 8). These data indicate that most (if not all) DPLG proteins, despite strong sequence divergence, have preserved an HTH motif and therefore may have retained DNA binding activity.

Co-domestication of a PIFp2 gene in Drosophila

Because DPLT transposons encode both TPase and PIFp2 protein, the possibility exists that not only PIF TPases, but also Drosophila PIFp2 proteins could have been domesticated into cellular genes. However, the weak conservation of PIFp2 genes in DPLT and other PIF transposons (this study and Zhang et al. 2004) together with the presence of multiple host genes encoding Myb/SANT/MADF domain proteins makes it a more challenging task to uncover possible genes recruited from PIFp2 proteins using traditional similarity searches. Nevertheless, we reasoned that in the regions flanking the TPase-derived DPLGs it could still be possible to identify a domesticated PIFp2 gene derived from the same transposon. We identified an intact ORF potentially encoding a PIFp2 protein in a region immediately adjacent to the DPLG7A ortholog in D. pseudoobscura, D. persimilis and D. willistoni. We named this putative gene Drosophila PIF MADF-like protein-encoding gene 7, or DPM7. DPM7 is also present at orthologous position in D. viridis, D. mojavensis, and D. grimshawi, although in these species the DPLG7B gene is located 4.5 Mb downstream on the same chromosome arm (suppl. fig. 5). These data suggest a scenario whereby the DPM7 and DPLG7 from the same transposon copy were co-domesticated in the common ancestor of all these Drosophila species, but the DPLG7 was subsequently relocated in the common ancestor of D. viridis, D. mojavensis, and D. grimshawi. The orthology and conservation of a seemingly intact coding region in distantly related species strongly suggest that DPM7 is a functional gene in these species.

To confirm this hypothesis, we first compared the GC-content of DPM7 and DPLT PIFp2 coding regions. We found that the GC-content in DPM7 genes is very similar to other genes in all the Drosophila species, while in the transposon PIFp2 coding regions, the GC value is generally lower than host genes (suppl. fig. 6). To further assess the functionality of DPM7, we next carried out a selection analysis using CODEML (Yang 1997). The codon substitutions analysis revealed that DPM7 coding sequences have been affected by strong purifying selection in Drosophila. The LRT indicates that the 1-ratio model best fits the DPM7 genes evolutionary dynamics, with a \(K_{a}/K_{s}\) value of 0.0624 (suppl. tables 4 and 5, suppl. fig. 7). Thus, DPM7 together with its cognate DPLG7 gene, are functional genes likely derived from the same DPLT copy.

To shed light on the potential function of the predicted protein encoded by DPM7, we aligned the MADF domain from DPM7 and from PIFp2 proteins with related domains present in various Drosophila proteins, including other MADF-containing proteins and members of the Myb/SANT superfamly. The alignment reveals that the MADF domains of DPM7 and PIFp2 proteins contain the critical tryptophane residues characteristic of the Myb/SANT/MADF domains (Asland, Stewart, and Gibson 1996; Bhaskar and Courey 2002), but display several residues and features specific of the MADF domain family, such as extended flanking conserved regions (fig. 9). Moreover, secondary structure prediction of the DPM7 and PIFp2 MADF-like domains revealed that almost all have retained a HTH-like motif conserved in the Myb/SANT/MADF family (data not shown). Together these analyses lend support to the hypothesis that DPM7 has preserved the overall architecture of the MADF-like domain of the original PIFp2 protein from which it is derived, and therefore DPM7 might act as a DNA-binding protein.

Discussion

In summary, our results show that PIF-like transposons have been active in the genomes of several Drosophila species. These elements display all the characteristics of PIF/SIS superfamily DNA transposons: short TIRs, 3-bp TSD, and 2 separate genes encoding the putative TPase and PIFp2, an accessory protein with a Myb/SANT/MADF domain. While at least 4 distinct lineages of PIF-like transposons were initially present in the Drosophila common ancestor, these elements appear to have become extinct in some Drosophila species (e.g. D. melanogaster). DPLTs remain abundant, highly diversified in several species (table 1) and some families have recently expanded in D. pseudoobscura, D. persimilis and D. willistoni, as judged by the dispersion of almost identical copies within the same genome. It is possible that some DPLTs are still transpositionally active in these species or their close relatives.

The presence of very closely related elements in the distantly related species D. pseudoobscura, D. willistoni and D. mojavensis suggest that horizontal transmission has played a role in the evolutionary dynamics of PLTs among Drosophila species. Horizontal transfers of DNA transposons (primarily \(P\) and mariner-like elements) have been documented in Drosophila and other insects (Robertson and Lampe 1995; Brunet et al. 1999; Silva and Kidwell 2000; Lampe et al. 2003), but this is the first record of (probable) horizontal movement of PLTs in any species. This is somewhat surprising because vast number of PLTs have been previously isolated and characterized from many plant species belonging to a broad taxonomic range, but no obvious cases of horizontal transfer were apparent (Zhang et al. 2004). Likewise, hundreds of mariner-like sequences have been isolated from over 50 flowering plant species, and there is so far no clear indication for any horizontal movement of these elements among plants (Feschotte and Wessler 2002; C.F. unpublished data). In contrast, multiple cases of horizontal transfer of mariner-like elements have been
reported in various insects, including Drosophila (Roberson 2002). Together, these observations suggest that horizontal transfers of DNA transposons occur more readily among insects than among plants, for reasons that are presently unclear.

We also reported on the identification of 7 distinct Drosophila genes (DPLG1-7), which appear to be derived from PIF-like TPase sequences. Each gene encodes a protein that shares moderate but significant similarity to a full-length PIF-like TPase, but seems to have originated independently from at least 3 distinct TPase sources (fig. 7). We showed that DPLGs share the characteristics of "host" genes encoding proteins with cellular function rather than TE-encoded genes. First, DPLG orthologs occur at the same relative chromosomal position in several Drosophila species, while TE insertions are typically not conserved in Drosophila (Biemont and Cizeron 1999; Caspi and Pachter 2006). This is in part because the turnover of TE sequences and non-functional DNA is extremely rapid in Drosophila (Petrov 2002; Lerat, Rizzon, and Biemont 2003) and also because a given TE insertion generally occurs at low frequency among individuals of different or same population (Charlesworth, Lapid, and Canada 1992; Petrov et al. 2003). Second, we found that each DPLG is essentially present in a single copy per haploid genome and is not flanked by TIRs and TSD, unlike all characterized PLTs. Third, we found that the nucleotide composition of DPLGs and DPLT TPase genes are dramatically different and that the GC-content of DPLGs, but not those of DPLTs, is comparable to other Drosophila (cellular) genes (suppl. fig. 7). This result is consistent with previous reports showing that TE-encoded genes in D. melanogaster and other plant and animal species are systematically more AT-rich than "host" genes and are not equally sensitive to codon bias (Lerat et al. 2002). Therefore, it appears that the domestication of DPLGs was accompanied by a shift in their nucleotide composition, leading to an enrichment of the GC-content at synonymous sites. The marked difference in the nucleotide composition of TE-encoded genes and domesticated TE genes may be applicable to other TE superfamilies and to other species to discriminate genes from TE and facilitate genome annotation (see also Zdobnov et al. 2005). Finally, we present evidence that all DPLG-encoded proteins are evolving under strong purifying selection in most—if not all—Drosophila lineages (suppl. tables 2 and 3, suppl. fig. 3). Again, this pattern is more reminiscent of host genes with cellular functions than TE genes, since the latter tend to evolve under no selective constraints, akin to pseudogenes (Witherspoon 1999; Lampe et al. 2003; Silva and Kidwell 2004).

At present, we can only speculate on the cellular function of the DPLG proteins. EST and microarray data suggest that some DPLGs have specific and distinct expression pattern and are likely to be developmentally regulated (fig. 4). These data remain preliminary and more detailed examination of the expression pattern of the different DPLG transcripts and proteins during development and in different tissues would certainly be enlightening. Nonetheless, these observations, combined with the fact that DPLGs have very little sequence similarity to each other and have not always preserved the same ancestrally conserved protein motifs, indicate that DPLG proteins probably function in distinct pathways and processes (fig. 8).

All TPases that have been biochemically characterized previously possess 2 distinct and separable functional domains: a N-terminal region that is responsible for specific DNA binding to the TIRs of the transposon and a C-terminal region involved in the catalytic activities of breakage, transfer and joining reactions (Craig et al. 2002). Sequence analyses showed that DPLG proteins have acquired mutations at positions known to be critical for catalytic activities of many TPases and other recombinases. In particular the DDE motif has been systematically altered in DPLG proteins and similar alterations are known to abolish or dramatically reduce catalytic activities of such
recombinases (Haren et al. 1999; Craig et al. 2002). In contrast, the PIF-derived gene HARBII1 present in vertebrates retains all the characteristic motifs of PIF TPases, comprising the catalytic signature DDE (Kapitonov and Jurka 2004). The predicted secondary structure of the N-terminal region of the ancestral DPLT TPases, including a putative HTH motif, has been apparently preserved (fig. 8). Thus, it is tempting to speculate that DPLG proteins have retained DNA binding capacities and could have been converted, for example, into transcription factors.

Several TPases are known to physically interact with other proteins. For example, the Sleeping Beauty TPase interacts with the Ku70 repair protein, with the DNA-binding, high-mobility group protein HMGB1 and with the transcription factor Miz-1 (Zayed et al. 2003; Izsak et al. 2004; Walisko et al. 2006). It is possible that some of the protein-protein interaction properties of the ancestral DPLT TPases might also have been co-opted. In this regard, the co-domestication of a PIFp2 gene, DPM7, along with its adjacent TPase-derived gene DPLG7 from the same transposon suggest the testable hypothesis that the respective proteins had an ancestral mutual interaction that has been maintained and both were co-opted for the same cellular function or pathway.

The recruitment of DPLG7 and DPM7 constitute, to our knowledge, the first reported case of multiple gene domestication from the same TE copy. The activities and possible role of PIFp2 proteins in the transposition cycle of PIF transposons have not been studied. Thus, it is difficult to predict the cellular function of the domesticated DPM7 protein. Nonetheless, we note that other MADF-containing proteins that have been biochemically and/or genetically characterized in D. melanogaster, such as Adf-1 and Mes2, act as transcriptional regulators in D. melanogaster and that the MADF domain in Dip3 is involved in sequence specific DNA-binding (Englander et al. 1992; Bhaskar and Courey 2002; Zimmermann et al. 2006). Since DPM7 and all other PIFp2 proteins contain a MADF domain (or a variant of the Myb/SANT domain), it is possible that DPM7 is a DNA-binding protein that functions in transcriptional regulation.

At first, it may seem surprising that the same superfamily of transposons would have repeatedly given birth to multiple ‘host’ genes in closely related species. In addition, a PIF transposon has also independently given rise to HARBII1, a gene of unknown function highly conserved in jawed vertebrates (Kapitonov and Jurka 2004). One interpretation is that PIF TPases possess peculiar features that make them prone to domestication. On the other hand, there are now multiple examples of domesticated TPase sequences from almost all recognized superfamilies and many more surely remain to be discovered (Cordaux et al. 2006; Volf 2006). Thus, the domestication of TPase sequences should not be viewed as a rare and odd phenomenon, but rather as a common path for the emergence of new genes.

Supplementary Material

Supplementary Tables 1 through 5 and Figures 1 through 7 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org).

Acknowledgments

We are grateful to Etsuko Moriyama for advice on the CG-content analyses, Alfredo Ruiz for data and discussion on the geographical distribution of Drosophila species and the Tucson Drosophila Stock center for providing D. per similis and D. willistoni stocks. We also thank 2 anonymous reviewers for their insightful comments. We thank Agencourt, Inc. (D. erecta, D. ananassae, D. mojavensis, D. virilis and D. grimshawi), Genome Sequencing Center, WUSTL School of Medicine (D. simulans and D. yakuba), TIGR (D. willistoni) and The Broad Institute (D. sechellia and D. persimilis) for prepublication access to their genome data. This work was supported by UTA start-up funds to E.B. and C.F., GM077582 grant from NIH to C.F., and GM 071813-01 grant from NIH to E.B.

Literature Cited


Jianzhi Zhang, Associate Editor

Accepted May 29, 2007