Evolution of the *copia* Retrotransposon in the *Drosophila melanogaster* Species Subgroup

I. King Jordan and John F. McDonald

Department of Genetics, University of Georgia

We report the results of a phylogenetic survey of the retrotransposon *copia* in the *melanogaster* subgroup of the *Drosophila* genus. The polymerase chain reaction was used to amplify the *copia* 5′ long terminal repeat and the adjacent untranslated leader region from representative *melanogaster* subgroup species. Restriction and sequence analyses of this region reveal discrete classes of *copia* size variants within the *melanogaster* subgroup. Phylogenetic comparisons of *copia* sequence data indicate that the size variants represent different *copia* subfamilies which diverged prior to their distribution in the *melanogaster* subgroup. Our results also suggest that *copia* elements have been subject to horizontal and vertical transmission during their evolution.

Introduction

Long terminal repeat (LTR)-containing retrotransposons are a class of repetitive, mobile elements which transpose via the reverse transcription of an RNA intermediate. Retrotransposons are ubiquitous components of eukaryotic genomes, and insertions of these elements are known to be a major source of spontaneous mutations (Berg and Howe 1989). In addition, there is a growing body of evidence that retrotransposons have played a significant role in host genome evolution (McDonald 1993, 1995; White, Habera, and Wessler 1994; Wessler, Bureau, and White 1995; Miller, Kruckenhauser, and Pinsker 1996; Pardue et al. 1996; San-Miguel et al. 1996). Despite the biological importance of retrotransposons, relatively little is known about the factors which influence their evolution. Information on the distribution and variation existing among retrotransposons in natural populations can provide valuable insight into their evolutionary history and the manner in which they coevolve with host genomes.

The *Drosophila copia* LTR retrotransposon is an ideal system for the study of retrotransposon evolution. *Drosophila* species have long been used as model systems for genetic studies, and the systematics of the genus is thoroughly documented and supported by morphological, chromosomal, and molecular evidence (Thockmorton 1975; Beverley and Wilson 1984; Ashburner 1989; DeSalle and Grimaldi 1991; Russo, Takezaki, and Nei 1995). *Copia*, which has been fully sequenced and characterized (Mount and Rubin 1985), has been extensively studied on the molecular level, and the distribution of *copia* in the *Drosophila* genus is well characterized (Martin, Wiernasz, and Schedl 1983; Stacey et al. 1986).

*Copia* is ~5 kb in length and consists of two LTRs flanking a single open reading frame (fig. 1A). The *copia* transcript, which begins in the 5′ LTR and terminates in the 3′ LTR, contains homology to the *gag* and *pol* loci of retroviruses (Mount and Rubin 1985). The region of *copia* focused on in this study consists of the 5′ LTR and the untranslated leader region (ULR). The 5′ LTR-ULR contains sequences responsible for controlling *copia* transcription, which is a rate-limiting step in the retrotransposition process (Boeke et al. 1985). The 5′ LTR contains promoter sequences and the start site of transcription (Flavell et al. 1981; Emori et al. 1985). The ULR contains a series of repeated sequence motifs which function as enhancers (Sneddon and Flavell 1989; Matyunina, Jordan, and McDonald 1996; Wilson, Matyunina, and McDonald 1998). Despite the functional importance of these regulatory sequences, the noncoding LTR-ULR is the most rapidly evolving region of LTR retrotransposons (Arkhipova, Lyubomirskaya, and Ilyin 1995). A survey of *copia* LTR-ULR sequences can therefore provide valuable phylogenetic information as well as a better understanding of retrotransposon regulatory variation. In this study, we employed a PCR-based assay using primers which flank the 5′ LTR-ULR to screen *melanogaster* subgroup populations for the presence of *copia*. Restriction and sequence analyses were carried out on *copia* 5′ LTR-ULR amplification products from the different populations. A number of *copia* sequences available on the database were also incorporated into our analysis. A diverged class of *copia* sequences isolated from *repleta* group species was used as an outgroup in this study. Our results demonstrate the existence of subfamilies of *copia* elements which diverged prior to their distribution throughout the *melanogaster* subgroup and suggest that *copia* sequences have been subject to both horizontal transfer and vertical transmission during their evolution.

Materials and Methods

*Drosophila* Strains

The *D. melanogaster* Iquitos, Peru (SA), strain has been described previously (Csík and McDonald 1995). The *D. melanogaster* Bujumbura, Burundi (AF), strain was obtained from Dr. Jean David (C.N.R.S., France). Strains of the other *Drosophila* species used in the study were obtained from either the National *Drosophila* Stock Center, Bloomington, Indiana (*D. sechellia* #3590, *D. erecta* #1013, *D. teissier* #1015) or the National
Copia LTR-ULR sequences characterized in this survey correspond to GenBank accession numbers AF063868–AF063885. Previously characterized copia sequences analyzed here correspond to the following GenBank accession numbers: D. melanogaster X02599 (X), AC001281 (A), J01075 (J); D. simulans D10808 (D); D. keopfereae X96971; D. buzzatii X96972 (X).

Restriction and Sequence Analyses

Genomic DNA was prepared from adult flies as described (Bender, Spierer, and Hogness 1983). The primers “Cop Ltr” (5’-CTATTCAACCTACAAAAATAACG-3’) and “Cop Pcs” (5’-ATTACGT'TTA-GCCGTTGCCAT-3’) were used to PCR amplify copia LTR-ULRs, and the primers CS 249 (5’-TCGTAACAAGGTTTCCG-3’), CS 250 (5’-GTTT-3’) were used to PCR amplify rDNA ITS regions from the genomic DNA samples under the following conditions: 1 × PCR buffer (Fisher), 200 μM dNTPs (Pharmacia), 3 mM magnesium chloride (Fisher), 0.5 μM each primer, 2.5 U Taq polymerase (Fisher), and 100 ng genomic DNA. Thermal cycle conditions were as follows: 5 min at 94°C, followed by 30 cycles of 1 min at 94°C, 1 min at 50°C, and 1 min at 72°C, followed by a final 5 min at 72°C. PCR products were visualized after fractionation through 2% agarose gels by ethidium bromide staining and 1 min at 72°C.

Statistical comparisons of model versus optimal copia trees were performed using the POLEUP program of the Wisconsin GCG computer package followed by visual inspection and adjustment. The resulting sequence alignment was used to reconstruct the phylogeny of the copia elements using a number of different methods. PAUP 3.1.1 (Swofford 1993) and PHYLIP (Felsenstein 1991) computer programs were both used to reconstruct copia phylogenies. LTR and ULR gaps were treated as one character. Maximum-likelihood and parsimony as well as neighbor-joining and UPGMA distance-based methods were used in phylogenetic reconstruction. The results obtained using these various methods were all in agreement, with the exception of the placement of a few copia sequences within the major clades. The main implications of the reconstructed trees are supported by all methods. The tree reported here (figs. 4 and 5) was generated using the heuristic search option of PAUP 3.1.1. Ten separate heuristic searches using tree bisection reconnection branch-swapping were used. Each search terminated after the accumulation of 1,000 equally parsimonious trees. All 10 searches recovered most-parsimonious trees. PCR of copia genomic sequences resulted in a pool of LTR-ULR PCR products from multiple copia elements in a genome. These PCR products were then subcloned into pCRII or pCR2.1 using the TA cloning method (Invitrogen). Individual subclones were PCR-amplified as described above, and the copia LTR-ULR products were digested using ApaI (Promega), which cuts just 3’ to the end of the copia LTR (fig. 1B). The resulting two LTR- and ULR-containing fragments were fractionated through 2% agarose (Sigma) gels. Variability in fragment sizes indicated the presence or absence of gaps in the LTR and/or ULR (fig. 1B). Sequence analysis was subsequently performed to determine the positions of the gaps in these naturally occurring copia LTR-ULR size variants. Automated sequencing was carried out at the University of Georgia Molecular Genetics Instrumentation Facility. Copia LTR-ULR subclones were sequenced in both directions using either the T7 and SP6 or the T7 and M13 reverse primers. The resulting chromatograms were aligned and compared using the SeqEd program (Applied Biosystems) to resolve any sequence ambiguities.

PCR products were fractionated through 2% agarose gels and transferred to Hybond-N nylon membrane (Amersham). The membrane was prehybridized in 0.5 molar sodium phosphate buffer (pH 7.6), 7% SDS for 4 h at 65°C. A 5’ copia probe was generated with PCR using a genomic copia clone (M33-25) as a template with primers Cop Ltr and Cop Pcs. A D. melanogaster ITS probe was generated using D. melanogaster genomic DNA as a template with primers CS 633 and CS 262. Probes were labeled using 32P-labeled dATP (ICN Biochemicals) and the High Prime DNA labeling kit per the supplier’s instructions (Boehringer Manheim). The probe was added to the prehybridization buffer, and hybridization proceeded for 18 h at 65°C. Membranes were washed under the following conditions at 65°C: 20 min 2 × SSC, 0.1% SDS; 15 min 1 × SSC, 0.1% SDS; and 10 min 0.1 × SSC, 0.1% SDS.

Phylogenetic Analysis

Copia LTR-ULR sequences were aligned using the PILEUP program of the Wisconsin GCG computer package followed by visual inspection and adjustment. The resulting sequence alignment was used to reconstruct the phylogeny of the copia elements using a number of different methods. PAUP 3.1.1 (Swofford 1993) and PHYLIP (Felsenstein 1991) computer programs were both used to reconstruct copia phylogenies. LTR and ULR taxa from a particular host species were grouped into a single polytomy. Relationships between species polytomies were then made to reflect the phylogeny of the host species. Constraint trees were made progressively less conservative by relaxing topological constraints. This was done by extending the polytomies such that they contained more taxa with no predetermined relationships. For example, the least constrained strict vertical transmission tree consisted of three polytomies, each of which included copia LTR-ULR taxa.
A

**Copia**

```
5' LTR | gag | pol | 3' LTR
```

- **Cop Ltr**
- **Cop Pcs**

D. melanogaster, iquitos
D. simulans
D. mauritiana
D. tesselari

![Image showing 440 bp](image)

B

```
U3  R  U5
```

- **38 bp gap**
- **Apa I**
- **28 bp gap**

- **Full length**
- **LTR gap**
- **ULR gap**
- **Double gap**

- LTR band 274/236 bp
- ULR band 166/138 bp
from every species within each of the three melanogaster subgroup clades. Constraint trees were loaded into PAUP* 4.0d61 and allowed to constrain heuristic parsimony or maximum-likelihood searches to generate model trees. Model tree scores resulting from these constrained searches were compared with the scores of the optimal copia tree using the Kishino-Hasegawa test (Kishino and Hasegawa 1989).

Results

Discrete Classes of copia LTR-ULR Size Variants are Present in melanogaster Subgroup Species

We employed a PCR-based assay to survey representative species of the melanogaster subgroup. The melanogaster subgroup is a clade within the Sophophora subgenus consisting of D. melanogaster and seven sibling species. This subgroup originated in Africa (David and Capy 1988). Six of the eight species in the subgroup are endemic to Africa, while D. melanogaster and D. simulans are cosmopolitan species. The melanogaster subgroup is divided into three distinct clades, which we refer to as the melanogaster clade, the yakuba clade, and the erecta clade, respectively. Primers designed from D. melanogaster copia sequences which flank the long terminal repeat (LTR) and untranslated leader region (ULR) were used to amplify genomic DNA (fig. 1A).

Numerous populations representing seven of the eight species of this subgroup were assayed for the presence of copia. Consistent with previously reported Southern hybridization results (Martin, Wiernasz, and Schedd 1983; Stacey et al. 1986), all of the melanogaster group species assayed were positive for the presence of copia.

PCRs carried out using genomic DNA templates isolated from populations within the melanogaster subgroup indicated that there is significant size variation in the copia LTR-ULR region both within and between species. Strains representing some populations, for example, D. melanogaster Iquitos, are monomorphic for a particular size variant, while others, like D. simulans, are polymorphic with multiple size variants (fig. 1A). Southern hybridization of genomic PCR products confirmed the nature of this variation (data not shown).

PCRs performed using genomic DNA as a template yield amplification products from multiple copia elements within a genome. In order to isolate LTR-ULR regions from a single element, PCR products were subcloned. Individual subclones were reamplified and digested with ApaI in order to partition size variation to the LTR and/or the ULR (fig. 1B). The results of these experiments showed discrete classes of size variants present in different combinations in all of the populations surveyed (table 1). Full-length variants have no gaps. ULR gap variants have a gap in the ULR, and double gap variants have gaps in both the LTR and ULR.

In order to map the location of the size variation, representative size variants from each species surveyed were sequenced. Sequences were aligned using the PILEUP program of the Wisconsin GCG package followed by visual inspection and adjustment. All gap variants were characterized relative to the full-length variants (fig. 2). ULR gap variants have ~28-bp gaps in their ULRs. Double gap variants have a ULR gap as well as ~38-bp gaps within their LTRs. A unique size variant isolated from the D. sechellia population had a 20-bp insert at the 5’ end of the LTR in addition to the standard LTR and ULR gaps. Additionally, two D. melanogaster, Burundi, elements contained identical 34-bp regional duplications at the LTR-ULR boundary.

Copia LTR-ULR Size Variants Show High Levels of Sequence Identity

The copia LTR-ULR size variants sequenced all show identities of 90% or greater despite the fact that some of the species within the group last shared a common ancestor >6 MYA (Russo, Takezaki, and Nei 1995). In many cases, copia sequences from within a population share less sequence identity with one another than with copia sequences isolated from the most distantly related species within the subgroup.

A series of PCR controls were instituted in order to eliminate the possibility that our results are due to contamination during the PCR process. For every PCR carried out, a blank reaction using ddH2O as a template was used as a negative control. This ensured against contamination of our results. Additional controls were performed to guard against contamination of genomic DNA templates with D. melanogaster DNA. In this series of controls, primers which amplify the internal spacer regions (ITSs) of the multicopy Drosophila rDNA genes were employed. ITS regions are variable and show species-specific size polymorphisms (Schlot-

![Fig. 1.—A, Copia element and the LTR-ULR region amplified by Cop Ltr and Cop Pcs primers. The gel shows four examples of PCR products amplified from genomic DNA templates isolated from different melanogaster subgroup species. Genomic PCR products consist of LTR-ULR fragments from multiple copia elements. B, LTR-ULR region of copia, showing the positions and sizes of the LTR and ULR gaps. The location where Apal cuts is also shown. Apal was used to digest individual copia LTR-ULR PCR products. PCR products amplified from different copia LTR-ULR size variants yield diagnostic restriction fragment patterns. The unique patterns of four LTR-ULR size variants can be seen in the gel.](image-url)
Fig. 2.—Alignment of representative sequences of three classes of copia LTR-ULR size variants. The sequence and orientation of the primers used to amplify copia transcription are shown at the ends of the alignment. Sequences important in the regulation of copia transcription are underlined in the full-length sequence. These include the start sites of transcription, initiator, and downstream elements (Flavell et al. 1981; Mount and Rubin 1985). Repeated sequence motifs located in the ULR which function as enhancers for the element are also underlined (Matyunina, Jordan, and McDonald 1996; Wilson, Matyunina, and McDonald 1998). The region of dyad symmetry tandemly duplicated in the full-length elements is shown. The positions and lengths of gaps which characterize copia LTR-ULR size variants are shown in the alignment.

In order to gain further insight into the evolution of copia, we utilized the LTR-ULR sequence alignment to reconstruct a copia phylogeny. Results reported here are based on a parsimony analysis performed with the PAUP 3.1.1 phylogenetic program (fig. 4). Neighbor-joining and maximum-likelihood-based methods were also used in phylogenetic reconstruction. The results obtained using all methods were in agreement, with the exception of the placement of a few copia sequences within the major clades. The main implications of the reconstructed trees are supported by all methods.

Phylogenetic Distribution of copia LTR-ULR Size Variants Is Incongruent with Host Species Phylogeny

In order to determine the relative amounts of sequence evolution which has occurred within each of the three gap variant clades, average p distances (fraction of sites that differ) between sequences were calculated for each clade. The double-gap clade shows the highest levels of sequence divergence; the average p distance for sequences in this clade (0.0231) is significantly greater (P < 0.001, t-test) than for sequences of either the ULR (0.0072) or full-length (0.0108) clades. The ULR gap clade sequences also have a higher average p distance between them than do the full-length sequences; however, this difference is not significant.

In order to determine the relative amounts of sequence evolution which has occurred within each of the three gap variant clades, average p distances (fraction of sites that differ) between sequences were calculated for each clade. The double-gap clade shows the highest levels of sequence divergence; the average p distance for sequences in this clade (0.0231) is significantly greater (P < 0.001, t-test) than for sequences of either the ULR (0.0072) or full-length (0.0108) clades. The ULR gap clade sequences also have a higher average p distance between them than do the full-length sequences; however, this difference is not significant.

The incongruence between host species and element-generated phylogenies can be taken as prima facie evidence for horizontal transfer. However, the fact that the topology of the copia and melanogaster species subgroup trees differs in the placement of numerous taxa, coupled with the low levels of sequence divergence between all copia elements examined within the subgroup, suggests that if horizontal transfer has occurred within the melanogaster subgroup, it has been frequent and relatively recent. Under such a scenario, incongruent element-generated trees may be expected to have low bootstrap values due to high levels of sequence similarity between elements from different species. This is the case for the copia phylogeny presented here (fig. 4), much of which is not well supported by bootstrapping. This lack of support of individual branches, while consistent with the scenario described above, prevents a statistically robust claim for the discordance between the host species phylogeny and the copia-generated phylogeny.
In order to more thoroughly evaluate the phylogenetic evidence for horizontal transfer in this case, statistical phylogenetic comparisons (Hibbett 1996; Huelsenbeck and Rannala 1997) were performed to compare the overall topology of the \textit{copia}-generated tree, which represents a horizontal-transfer hypothesis, with trees representing strict vertical-transmission hypotheses.

In order to generate vertical-hypothesis model trees for these statistical phylogenetic comparisons, \textit{copia} LTR-ULR sequences were forced to evolve under topological constraints representing vertical-transmission hypotheses. The taxa under consideration (\textit{copia} LTR-ULR sequences) were grouped phylogenetically into a series of constraint trees, with topologies representing hypotheses of vertical transmission, using MacClade (Maddison and Maddison 1992). These trees were then used to constrain heuristic searches of the \textit{copia} sequence alignment using PAUP* 4.061 (Swofford 1998) to generate a number of optimal vertical-hypothesis model trees. This series of model trees was progressively less conservative in that the topological constraints were relaxed, and subsequently, each successive model tree more closely resembled the optimal \textit{copia}-generated tree.

The first set of vertical-hypothesis trees (vertical hypotheses I-1 to I-3; table 2) represent a scenario of strict vertical transmission (fig. 6A). In these trees, \textit{copia} sequences isolated from a particular population or species are most closely related to each other, and the topology reflects the relationships of the host species phylogeny. In progressively less conservative versions of this hypothesis, topological constraints were relaxed until only \textit{copia} sequences from within each of the three \textit{melanogaster} subgroup clades were constrained to group together. The second set of vertical hypotheses (vertical hypotheses II-1 to II-3; table 2) are generally less conservative than the first in that they more closely approximate the \textit{copia}-generated tree, which represents a hypothesis of horizontal transfer (fig. 6B). These vertical-hypothesis trees take into consideration the presence of gap variant subfamilies and allow for ancestral polymorphism. In these trees, \textit{copia} sequences of a particular gap variant were grouped together. The host species phylogeny is then reflected within each gap variant
clade to varying degrees. Topological constraints were relaxed in successive trees in the same way discussed above.

Comparisons of trees representing vertical- and horizontal-transfer hypotheses were implemented using PAUP* version 4.0d61 (Swofford 1998). Statistical phylogenetic comparisons were performed using both parsimony and maximum likelihood. Copia LTR-ULR sequences were forced to evolve under the topological constraints representing alternative vertical or horizontal hypotheses as described above. The resulting trees were then assigned either a length in the case of parsimony or a log likelihood in the case of maximum likelihood. For each vertical-hypothesis tree, a pairwise comparison with the optimal copia-generated phylogeny was performed. For each of these pairwise comparisons, the Kishino-Hasegawa test was used to determine if the lengths or the log likelihoods of the trees were significantly different. Table 2 shows a summary of the results obtained. In the case of parsimony, all but the most conservative (vertical hypothesis II-3) of the vertical-hypothesis trees was significantly different in length from the copia-generated tree, which represents a horizontal-transfer hypothesis. The maximum-likelihood method yielded significantly different tree scores for all of the comparisons. This analysis lends statistical support to the claim that the incongruence between the host species phylogeny and the copia-generated phylogeny is significant and constitutes putative evidence for the horizontal-transfer hypothesis.

Discussion
Discrete LTR-ULR Size Variants Represent copia Subfamilies

Our survey of naturally occurring copia LTR-ULR variation reveals two levels of sequence organization. An alignment of a repleta group D. buzzatii copia LTR-ULR with a copia sequence from D. melanogaster reveals only 58.5% identity between the two sequences. This is a much lower level of sequence identity than is seen among any of the melanogaster group copia ele-
Evolution of the *copia* Retrotranspon

**Fig. 5.**

A. Phylogeny of the *melanogaster* subgroup host species surveyed for the presence of *copia*. The tree is adopted with modifications from Jeffs, Holmes, and Ashburner (1994) and Russo, Takezaki, and Nei (1995). The phylogenetic relationships depicted here are supported by a number of independent lines of evidence, including nucleotide sequence data, immunological distances, allozyme data, and morphological characters. B. The same tree shown in figure 4 presented as a cladogram to emphasize the topological relationships of the *copia* LTR-ULR sequences reported here. Taxa names represent the host species the *copia* sequences were isolated from, followed by the clone number. The *copia* sequence-generated tree is incongruous with the host species tree.

ments. This result suggests that the *repleta* group sequences may represent a different family of *copia* elements.

Our data also suggest the existence of subfamilies of *copia* elements within the *melanogaster* group. As mentioned earlier, *melanogaster* group *copia* elements fall into discrete classes of size variants. These size variants may represent different subfamilies of *copia*. This possibility is supported by the fact that *copia* elements tend to cluster together in the element-generated phylogeny based on the class of size variant to which they belong. For example, double gap sequences group together regardless of the species from which they were isolated (fig. 4). Full-length and ULR gap elements generally group in a similar way. This type of clustering would be expected if the double gap variants actually represent a distinct *copia* subfamily separate from the ULR gap and full-length elements. Although it is also possible that the ULR and full-length variants represent additional separate *copia* subfamilies, the fact that these variants are less sequentially distinct from one another makes such a classification ambiguous at the present time.

The phylogenetic evidence of *copia* subfamilies is consistent with the hypothesis that full-length variants evolved from gap variants through regional duplication (Matyunina, Jordan, and McDonald 1996). The *copia* ULR contains a region of dyad symmetry which is pres-

### Table 2

**Horizontal Versus Vertical Tree Comparisons**

<table>
<thead>
<tr>
<th>TREE</th>
<th>Length</th>
<th>Difference</th>
<th>SD</th>
<th>P</th>
<th>−ln L</th>
<th>Difference</th>
<th>SD</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Copia</em> phylogeny</td>
<td>290</td>
<td>18</td>
<td>8.805</td>
<td>0.0414*</td>
<td>1893.6947</td>
<td>73.8906</td>
<td>27.2742</td>
<td>0.0070**</td>
</tr>
<tr>
<td>Vertical hypothesis I-1</td>
<td>308</td>
<td>17</td>
<td>8.637</td>
<td>0.0496*</td>
<td>1967.5853</td>
<td>68.3227</td>
<td>25.5727</td>
<td>0.0078**</td>
</tr>
<tr>
<td>Vertical hypothesis I-2</td>
<td>307</td>
<td>12</td>
<td>4.876</td>
<td>0.0142*</td>
<td>1962.0174</td>
<td>55.5758</td>
<td>17.6562</td>
<td>0.0017**</td>
</tr>
<tr>
<td>Vertical hypothesis I-3</td>
<td>302</td>
<td>14</td>
<td>6.908</td>
<td>0.0432*</td>
<td>1949.2525</td>
<td>69.5319</td>
<td>14.7495</td>
<td>0.0038**</td>
</tr>
<tr>
<td>Vertical hypothesis I-4</td>
<td>304</td>
<td>10</td>
<td>4.455</td>
<td>0.0252*</td>
<td>1936.5737</td>
<td>42.8790</td>
<td>14.7495</td>
<td>0.0038**</td>
</tr>
<tr>
<td>Vertical hypothesis II-1</td>
<td>300</td>
<td>10</td>
<td>5.278</td>
<td>0.0587</td>
<td>1949.1731</td>
<td>42.8790</td>
<td>14.7495</td>
<td>0.0038**</td>
</tr>
<tr>
<td>Vertical hypothesis II-2</td>
<td>300</td>
<td>10</td>
<td>4.455</td>
<td>0.0252*</td>
<td>1936.5737</td>
<td>42.8790</td>
<td>14.7495</td>
<td>0.0038**</td>
</tr>
<tr>
<td>Vertical hypothesis II-3</td>
<td>300</td>
<td>10</td>
<td>5.278</td>
<td>0.0587</td>
<td>1949.1731</td>
<td>42.8790</td>
<td>14.7495</td>
<td>0.0038**</td>
</tr>
</tbody>
</table>

*a* Difference between the optimal *copia* tree and the vertical-hypothesis model trees.

*b* Standard deviation of the differences between the trees.

*c* Probability that the difference between the trees is due to chance.

*P* < 0.05.

**P** < 0.01.
ent once in the ULR gap and double gap variants and is tandemly duplicated in the full-length elements (fig. 2). Homology-dependent mechanisms such as replication slippage and unequal crossing over or template-switching errors during reverse transcription (Temin 1991) may have been responsible for the duplications within the ULR. Similar mechanisms may have generated full-length LTRs from the LTR gap variants (McDonald et al. 1997).

The finding that Copia LTR-ULR size variants isolated from the same species do not tend to group together phylogenetically indicates that these size variant subfamilies diverged prior to their distribution throughout the Drosophila genus. The subsequent distribution of these Copia subfamilies could have occurred via horizontal transfer or vertical transmission.

The average $p$ distances for the sequences of each clade suggest that the double gap clade is the oldest clade, followed by the ULR gap clade and the full-length clade. This finding is also consistent with the hypothesis that full-length variants evolved from gap variants. If Copia gap variants were transmitted strictly vertically, the uneven distribution of these variants (table 1) in species of the melanogaster subgroup could be due to stochastic loss of elements within species lineages. If gap variants were transmitted horizontally, then the data suggest that the distribution of gap variants could have occurred in waves, starting with the double gap variants, followed by the ULR gap and full-length variants. The present distribution of Copia sequences in the melanogaster group may be a result of both phenomena.

Copia Phylogeny Is Incongruent with Host Species Phylogeny

The topology of the Copia LTR-ULR sequence-based phylogeny does not match the topology of the host species phylogeny from which the Copia elements were isolated (fig. 5). Such phylogenetic discordance is considered one indication that elements may have evolved through horizontal transfer (Capy, Anxolabehere, and Langin 1994).

Horizontal transfer consists of nonsexual transmission of genes across species boundaries. Studies addressing the molecular evolution of transposable elements have revealed that these elements may be prone to horizontal transfer (Kidwell 1993). For example, there is particularly strong evidence which indicates that DNA type transposable elements have crossed species boundaries via horizontal transfer (Daniels et al. 1990; Maruyama and Hartl 1991; Robertson 1993; Hagemann, Haring and Pinsker 1996; Clark and Kidwell 1997). Claims of horizontal transfer of retrotransposons have been more controversial. Numerous reports have suggested that retrotransposons may horizontally transfer across species boundaries (Doolittle et al. 1989; Mizrokh and Mazo 1990; Xiong and Eickbush 1990; Flavell 1992; Alberola and de Frutos 1993). However, the evidence for horizontal transfer of retrotransposons in these cases is less direct. Further examination of this phenomenon requires a detailed examination of the molecular variation of retrotransposons within a phylogenetically well defined host species group such as that conducted in this study.
Evolutionary studies of transposable elements often yield element-generated phylogenies which are incongruent with host species phylogenies, such as we report, but these inconsistencies are not always indicative of horizontal transfer (Capy, Anxolabehere, and Langin 1994). Issues of sampling and experimental design are especially important when considering the results of such studies. Transposable elements are present in multiple copies in the genome. There are many factors which can obscure phylogenetic reconstruction of such multicopy genes. Comparisons of paralogous copies of elements and varying rates of sequence evolution of element copies within and between species are factors which can yield incongruous phylogenies even under conditions of strict vertical transmission (VanderWiel, Voytas, and Wendel 1993; Cummings 1994). In addition, it is also possible that ancestral polymorphism with differential rates of loss, and evolution of copia elements within and between species lineages, could obscure phylogenetic reconstruction of element sequences and lead to incongruous phylogenetic patterns.

In our comparison of vertical-transmission model trees versus the optimal copia phylogeny, we attempted to address these issues by generating sets of model trees which deal with the types of considerations discussed above. These types of statistical phylogenetic comparisons also allow a critical evaluation of the phylogenetic information contained in the entire tree topology, not just the support for individual branches (Hibbett 1996; Hillis, Mable, and Moritz 1996; Huelsenbeck and Rannala 1997). This method is well suited to a scenario in which multiple and recent horizontal transfers may yield trees with low bootstrap values.

In our comparisons of model trees versus the optimal copia tree, vertical-transmission model trees which take ancestral polymorphism of the gap variants into consideration are less different from the optimal copia-generated tree than are strict vertical-transmission model trees (table 2). This indicates that the presence of ancestral gap variant polymorphisms influences the copia tree topology and is relevant to the interpretation of our results. However, the finding that even these less conservative model trees were significantly different from the optimal copia-generated tree supports the horizontal-transfer hypothesis.

High Levels of copia Sequence Homology Seem Best Explained by Horizontal Transfer

Perhaps the most compelling evidence for horizontal transfer of copia elements reported in this paper is the extremely high level of sequence conservation between copia LTR-ULRs isolated from distantly related melanogaster subgroup species. For example, copia LTR-ULRs isolated from D. melanogaster and D. erecta show 99% sequence identity despite the fact that these species have been separated for >6 Myr (Russo, Takezaki, and Nei 1995). A priori, it seems unlikely that the noncoding copia LTR-ULR region could be evolving so slowly as to account for this low level of divergence. Previous studies which have looked at sequence evolution in the same region of various retrotransposon families have found much higher levels of divergence between elements of a particular family both within and between species (Kulguskin, Ilyin, and Georgiev 1981; Lankenau et al. 1990; Lyubomirskaya et al. 1990; Mizrokhi and Mazo 1991).

It is possible, however, that rates of evolution within the copia LTR-ULR are somehow much more conserved than are those for other LTR retrotransposons previously studied. The copia LTR-ULR contains many sequences responsible for regulating the expression and retrotransposition of the element. These include promoter and enhancer-like sequences (Sneddon and Flavell 1989; Cavarec and Heidmann 1993; Matyunina, Jordan, and McDonald 1996; Cavarec et al. 1997; Wilson, Matyunina, and McDonald 1998). Sequences critical for the reverse transcription and integration of the element, like the primer-binding site and inverted terminal repeats, are also present in this region (Mount and Rubin 1985; Kikuchi, Ando, and Shiba 1986). The presence of all of these critical sequences within a relatively short region of the element raises the possibility that much of the LTR-ULR is selectively constrained. This selective constraint could maintain high levels of sequence homology among copia elements from diverged species. However, similar control sequences are present in other LTR retrotransposons which do not show similar high levels of interspecific sequence homology (Kulguskin, Ilyin, and Georgiev 1981; Lankenau et al. 1990; Lyubomirskaya et al. 1990; Mizrokhi and Mazo 1991).

Another possibility is that the copia elements represented in this survey are maintained in a region of the genome which is undergoing very slow rates of sequence evolution. According to this hypothesis, certain genes or regions of the genome are replicated at very high rates of fidelity. This could be due to extremely high rates of DNA excision repair or other presently unexplained mechanisms (Martin and Meyerowitz 1986; Sullivan 1995). The extent to which such mechanisms may influence evolutionary rates of copia and other transposable elements is presently unknown.

Overall, the significant host versus element phylogenetic discordance, combined with the high level of sequence identity between copia LTR-ULR sequences isolated from phylogenetically diverged melanogaster subgroup species, seems to favor the hypothesis that copia elements have repeatedly crossed species boundaries via horizontal-transfer events within the relatively recent evolutionary past. The mechanism by which copia elements may be vectored between species is presently unknown. However, the fact that copia-like elements have been found within the genomes of insect viruses is highly suggestive (Miller and Miller 1982).

Conclusion

Our results indicate that at least three subfamilies of copia elements exist within the melanogaster subgroup. While this may not represent the full spectrum of copia variation within the subgroup, it does encompass all copia elements thus far known to be biologically active within melanogaster subgroup species. Interest-
ingly, the three phylogenetically defined copia subfamilies correspond with three previously characterized LTR-ULR size gaps. Our results indicate that the double gap variants are the most ancestral class, followed by the ULR gap and the most recently evolved full-length variants.

Models of transposable-element dynamics indicate that the ultimate fate of these elements within species lineages is inactivation and eventual extinction (Kaplan, Darden, and Langley 1985). It has been postulated that cross-species transfers may be an effective strategy by which transposable elements avoid inactivation over evolutionary time (Maruyama and Hartl 1991). Since copia is known to be subject to effective host-mediated repression, it seems reasonable that significant selective pressure might exist to favor horizontal transfer of copia over evolutionary time (Sneddon and Flavell 1989; Matyunina, Jordan, and McDonald 1996). However, vertical and horizontal transmission are clearly not mutually exclusive, and our data indicate that both phenomena have likely contributed to the present distribution of copia elements within the melanogaster subgroup.

LITERATURE CITED


Lankenau, D. H., P. Huisser, E. Jansen, K. Miedema, and W. Hennis. 1990. DNA sequence comparison of micropia


Daniel L. Hartl, reviewing editor

Accepted May 20, 1998