Drosophila Euchromatic LTR Retrotransposons are Much Younger Than the Host Species in Which They Reside

Nathan J. Bowen and John F. McDonald

Genome Res. 2001 11: 1527-1540
Access the most recent version at doi:10.1101/gr.164201

Letter

Drosophila Euchromatic LTR Retrotransposons are Much Younger Than the Host Species in Which They Reside

Nathan J. Bowen[1] and John F. McDonald[2]

Department of Genetics, University of Georgia, Athens, Georgia 30602, USA

The recent release of the complete euchromatic genome sequence of *Drosophila melanogaster* offers a unique opportunity to explore the evolutionary history of transposable elements (TEs) within the genome of a higher eukaryote. In this report, we describe the annotation and phylogenetic comparison of 178 full-length long terminal repeat (LTR) retrotransposons from the sequenced component of the *D. melanogaster* genome. We report the characterization of 17 LTR retrotransposon families described previously and five newly discovered element families. Phylogenetically, these families can be divided into three distinct lineages that consist of members from the canonical Copia and Gypsy groups as well as a newly discovered third group containing *BEL*, *mazi*, and *roo* elements. Each family consists of members with average pairwise identities ≥99% at the nucleotide level, indicating they may be the products of recent transposition events. Consistent with the recent transposition hypothesis, we found that 70% (125/178) of the elements (across all families) have identical intra-element LTRs. Using the synonymous substitution rate that has been calculated previously for *Drosophila* (.016 substitutions per site per million years) and the intra-element LTR divergence calculated here, the average age of the remaining 30% (53/178) of the elements was found to be 137,000 ±89,000 yr. Collectively, these results indicate that many full-length LTR retrotransposons present in the *D. melanogaster* genome have transposed well after this species diverged from its closest relative *Drosophila simulans*, 2.3 ± .3 million years ago.

Retrotransposons are the most abundant and widespread class of eukaryotic transposable elements. For example, >50% of the maize genome (SanMiguel et al. 1996) and >40% of the human genome (Smit 1999) are comprised of retrotransposons. The biological importance of retrotransposons ranges from their contribution to mutation (Green 1988) and disease (Deininger and Batzer 1999) to their postulated role in evolution (McDonald 1990, 1993; Kidwell and Lisch 1997). The genome sequencing of humans and selected experimental and agriculturally important species is providing an unprecedented opportunity to view the patterns of variation existing among the entire complement of retrotransposons in complete genomes.

Retrotransposons are made up of short interspersed nuclear elements (SINES), long interspersed nuclear elements [LINES, also known as non-long terminal repeat (LTR) retrotransposons], LTR retrotransposons, and retroviruses. LTR retrotransposons are named for their long terminal repeats, which contain transcriptional regulatory sites and flank the internal coding regions of the elements (Boeke and Stoye 1997). LTR retrotransposons are classically divided into two groups, the Copia/Ty1 group and the Gypsy/Ty3 group. The distinguishing characteristic between these groups is the order of the three protein domains — protease (PR), reverse transcriptase (RT), and integrase (IN) — encoded within the *polymerase* (*pol*) gene of the elements. The *pol* region of Copia/Ty1 elements has the order (PR, IN, RT) whereas the Gypsy/Ty3 group has the more familiar arrangement (PR, RT, IN), which is also the order found in retroviruses. Recently, a third major group of LTR retrotransposons has been described containing the *BEL* element from *Drosophila melanogaster* as well as the *Cer7-12* elements of *Caenorhabditis elegans* (Bowen and McDonald 1999; Malik et al. 2000). The IN domain is also found downstream of the RT domain in this third group of LTR retrotransposons.

LTR retrotransposons and retroviruses are nearly identical in structure and are clearly related phylogenetically (Xiong and Eickbush 1988). The main distinguishing characteristic is that some LTR retrotransposons, such as Ty1 in yeast, do not contain an *envelope* gene, which renders retroviruses infectious. Many LTR retrotransposons, such as *gypsy* from *D. melanogaster*, however, do encode Envelope proteins and are infectious (Song et al. 1994). Therefore, LTR retrotransposons also serve as excellent models for the study of the evolution of infectious retroviruses. Previous large-scale analyses of the LTR retrotransposons of *Saccharomyces cerevisiae* (Jordan and McDonald 1998, 1999a,b; Kim et al. 1998), *C. elegans* (Bowen and McDonald 1999), *Zea mays*, and *Hordeum vulgare* (Shirasu et al. 2000) have provided novel insights into the molecular evolution and phylogenetic distribution of these retrotransposons.

Because the long terminal repeats of LTR retrotransposons are synthesized from a single template during reverse transcription, they are identical at the DNA sequence level on integration. Therefore, if the nucleotide substitution rate for the host DNA polymerase is known, the relative integration

[1]Present Address: Section on Eukaryotic Transposable Elements, Laboratory of Gene Regulation and Development, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA.
[2]Corresponding author.
E-MAIL mcgene@arches.uga.edu; FAX (706) 542-3910.
Article published on-line before print: *Genome Res.*, 10.1101/gr.164201.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.164201.

time or age of the element can be estimated from the level of sequence divergence existing between an element's LTRs. Previously, LTR nucleotide identity has been used to estimate the time of insertion of LTR retrotransposons from *S. cerevisiae*, *Zea mays*, and humans. For example, the age of the Ty1 and Ty2 elements from *S. cerevisiae* has been estimated to be <100,000 years old (Jordan and McDonald 1999b; Promislow et al. 1999). In contrast, it has been reported that the LTR retrotransposons within the ADH-region of the maize genome are much older, having transposed in the past 2 to 6 million years (SanMiguel et al. 1998). In a similar study, it has been reported that most human endogenous retroviruses (HERVs) inserted into the human genome long before humans diverged from the Old World monkeys, more than 25 million years ago (Tristem 2000).

In an initial effort to characterize all of the LTR retrotransposons within the genome of *D. melanogaster*, we report the annotation, phylogenetic analysis, and estimated ages of 178 full-length elements (i.e., those containing two intact LTRs and intervening coding regions) from the nonredundant sequence found in GenBank (Benson et al. 2000). Our results indicate that there are three major groups of LTR retrotransposons found within the *D. melanogaster* genome. We find that these three groups consist of over 20 individual families of elements and that each family of elements is composed of a group of highly homologous individual elements (~99% identity at the nucleotide level). We conclude that many LTR retrotransposons from each family have resulted from evolutionarily recent episodes of transpositional activity.

Table 1. LTR Retrotransposons Used to Categorize Elements from the *Drosophila melanogaster* Genome

Element name	Accession no.	Inserted element length (bp)
From *D. melanogaster*		
17.6	X01472	7439
1731	X07656	4648
297	X03431	6995
412	X04132	6897
BEL	U23420	6126
blastopia	Z27119	5416
burdock	U89994	6411
copia	X04456	5146
gypsy	AF033821	7469
HMS Beagle	AH001020	~7300
Idefix	AJ009736	7411
mdg1	X59545	7480
mdg3	X95908	5519
micropia	X14037	5465
nomad	AF039416	7592
tirant	X93507	5184
transpac	AF222049	5249
From other organisms		
Anopheles gambiae Moose	AF060859	
Tribolium castaneum woot	U09586	
Drosophila buzzatii osvaldo	AJ133521	
Drosophila virilis Ulysses	X56645	
Ceratitis capitata yoyo	U60529	
Drosophila virilis gypsy	M38438	
Drosphila subobscura gypsy	X72390	
Trichoplusia ni TED	M32662	
Drosophila ananassae Tom	Z24451	
Drosophila virilis Tv1	AF056940	

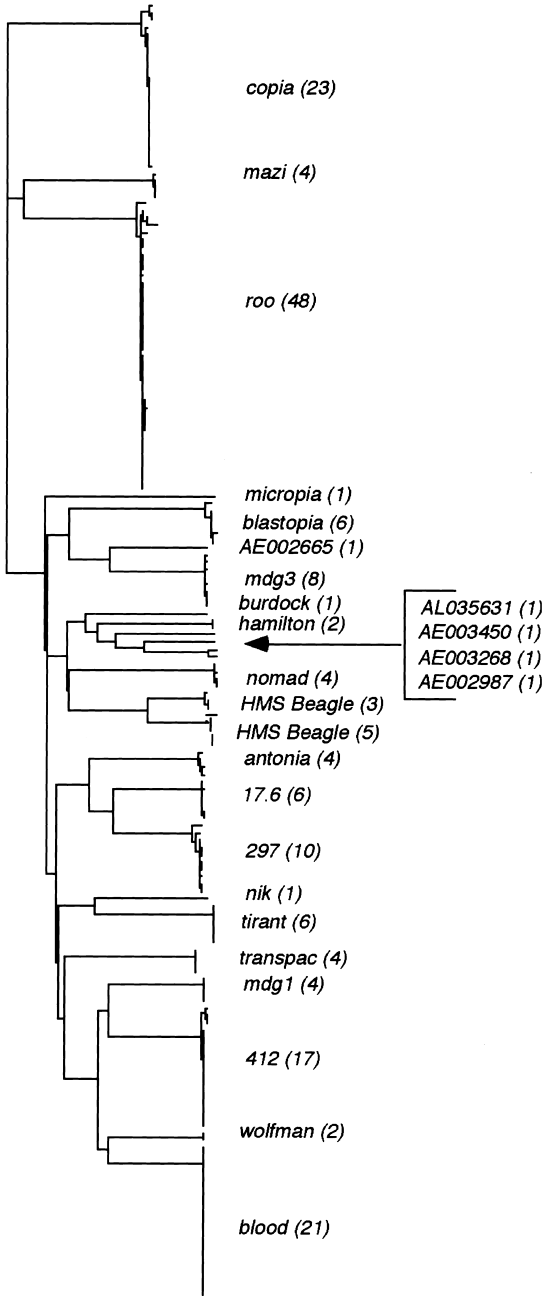

Figure 1 Neighbor-joining phylogenetic tree of the `ClustalW` alignment of LTR retrotransposon nucleotide sequences. This tree indicates the high level of sequence homology within a family of elements. For clarity, the element families are listed once with the number of elements in each family in parentheses. Unclassified elements are identified by the accession number in which they are located.

RESULTS

Isolation and Characterization of *D. melanogaster* LTR Retrotransposon Families from the Genome Sequence

The majority of the *D. melanogaster* genome sequence now available in GenBank is from the euchromatic regions of the

Table 2. Element Family Average Pairwise Nucleotide Identities

Element (sample size)	Average pairwise identities	Calculated divergence time (millions of years)
17.6 (6)	99.676525 ± 0.000862996	0.1078252 ± 0.028766517
297 (8)	99.846149 ± 0.000986654	0.051283679 ± 0.032888452
412 (19)	99.6790517 ± 0.001212138	0.106982767 ± 0.040404609
antonia (4)	98.8624153 ± 0.00499742	0.379194889 ± 0.166580662
blastopia (6)	99.7571929 ± 0.000753808	0.080935711 ± 0.025126928
blood (21)	99.8221958 ± 0.001247047	0.059268056 ± 0.041568245
copia (23)	99.7509078 ± 0.001175924	0.083030736 ± 0.039197471
wolfman (2)	99.889276	0.036908
hamilton (2)	99.973008	0.008997333
HMS Beagle (3)		
#'s 4, 5 and 6	99.448482 ± 0.00441751	0.172349167 ± 0.138047175
HMS Beagle (4)		
#'s 1, 2, 3, 7 and 8	99.852204 ± 0.0008726	0.04618625 ± 0.02726868
mazi (4)	99.8201147 ± 0.000824226	0.059961778 ± 0.027474189
mdg1 (5)	99.6931778 ± 0.000748225	0.102274067 ± 0.024940837
mdg3 (8)	99.6225424 ± 0.001529709	0.12581919 ± 0.050990305
nomad (4)	99.9075055 ± 0.000276762	0.0308315 ± 0.009225413
roo (40)	99.6887925 ± 0.001617852	0.103735823 ± 0.053928391
tirant (6)	99.9374788 ± 0.000253444	0.0208404 ± 0.008448124
transpac (4)	99.882341 ± 0.000408498	0.039219778 ± 0.013616585

genome (Adams et al. 2000). In contrast, only 2.5% of the genome sequence is derived from heterochromatic clones (Myers et al. 2000). Constitutive heterochromatin, which comprises roughly one-third of the *D. melanogaster* genome, is poorly represented in the genome sequence because these regions are not easily cloned into large inserts (Myers et al. 2000). Likewise, the assembly of DNA sequence from genomic regions that contain many tandemly arranged repetitive elements can result in the omission of internal sequences (E. Myers, pers. comm.). These issues are important to our study because *D. melanogaster* heterochromatin is thought to contain a substantial number of transposable elements (TEs) (Pimpinelli et al. 1995). Also LTR retrotransposons have been shown to exist in nested arrays in other species (SanMiguel et al. 1996a). Consequently, any LTR retrotransposons located in these regions of the genome are precluded from our analysis. Further sequencing and gap-filling efforts being conducted by Celera and the Berkeley *Drosophila* Genome Project (Myers et al. 2000) will likely identify additional elements within both the euchromatic and heterochromatic portions of the genome. Therefore, our results represent a large sampling of LTR retrotransposons from the euchromatin of *D. melanogaster*.

Following the initial characterization of each LTR retrotransposon (see Methods section), `ClustalW` (Thompson et al. 1997) was used to align the nucleotides of each element to known full-length LTR retrotransposons of *D. melanogaster* and other related organisms listed in Table 1. Information concerning all elements identified previously can be obtained through Flybase (http://flybase.harvard.edu). This initial alignment was done to group elements into known and unknown families. The phylogram generated from this preliminary alignment is shown in Figure 1. For clarity, each family is labeled once followed by the number of elements in each family. Because of the low level of interfamily nucleotide sequence identity, this initial phylogram may not accurately represent all interfamily relationships, but it does allow us to classify elements into distinguishable groups. The long interfamily branches and the large cluster of nearly identical elements at the termini of the family lineages apparent in this initial phylogram indicate that most families of *D. melanogaster* LTR retrotransposons consist of a group of highly homologous elements. Subsequently, computed pairwise nucleotide identities confirmed this finding in that each family was found to consist of elements with average pairwise nucleotide identities of ≥99% (Table 2).

The nucleotide sequences of the LTR retrotransposons that did not group with known elements were translated and their RT motif was aligned to the RT of the known elements (Fig. 2A). Only those RT motifs that were uninterrupted by frame shifts or stop codons were used in the characterization of novel families. This alignment was then used to generate pairwise amino acid identities. Consistent with criteria established previously (Bowen and McDonald 1999), if an element had a pairwise identity of <90% to a known RT, it was classified as a new family. These novel elements are shown in boldface type in the first column of Table 3. Additional RT sequences from other *Drosophila* and invertebrate species were included in this analysis to ensure that novel elements from *D. melanogaster* did not represent previously characterized elements from other related species. Elements with RT pairwise identities >90% to a previously characterized element were given the name of that element followed by a number.

All elements characterized in this study are listed individually in Table 3. Included in this table are other distinguishing characteristics of the LTR retrotransposons, including their accession numbers, chromosomal locations, inverted terminal repeats (ITRs), direct terminal repeats (DTRs), LTR length, complete element length, and estimated age of each element (see below). The DTRs result from a duplication of the unoccupied insertion site following proviral or element insertion (Coffin et al. 1997). In our study, the DTRs served as internal controls for the assembly process following the whole genome shotgun sequencing of *D. melanogaster* (Myers et al. 2000). If proviral elements located at different loci were incorrectly assembled, they would contain a mixed set of DTRs. For the elements that contained unique DTR sequences, 93% were identical. The other 7% are either incor-

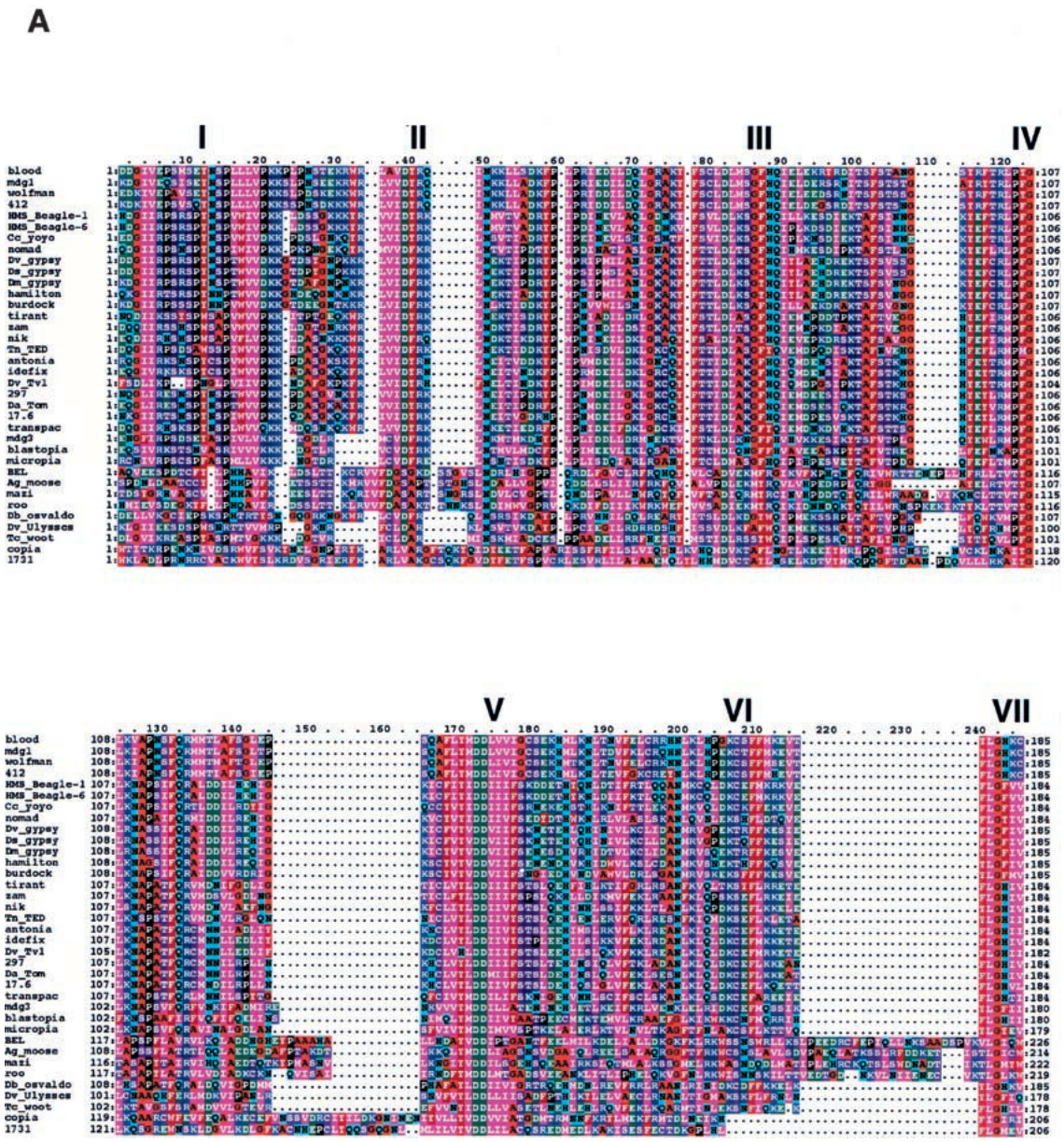

Figure 2 (*A*) Amino-acid alignment of the RT motif of *Drosophila melanogaster* LTR retrotransposons found in this study. The sequences were aligned using `ClustalX` as described in Methods. The seven conserved domains of RT (Xiong and Eickbush 1988) are indicated *above* the alignment. Residue coloring was performed by `MacBoxshade` v2.01 and is based on the similarity scheme (F, W, Y), (I, L, M, V), (P), (D, E), (G, A), (S, T, C), (N, H), (R, K), (Q). Members of a similar residue group are shaded identically. (*B*) Unrooted neighbor joining tree of sequences shown in *A*. Numbers found adjacent to branches indicate bootstrap support from 100 replicates. (*C*) Neighbor joining phylogenetic tree of the alignment shown in *A*rooted with the branch leading to the *1731* and *copia* lineages. Branches with bootstrap values <50 were collapsed to indicate only well-supported groups. Novel element families first characterized in this study are in boldface type and indicated by asterisks.

rectly assembled or are possibly the result of ectopic recombination between proviral elements at different loci. This hypothesis is currently under further investigation. In summary we identified 23 *copia*, six *17.6*, 10 *297*, 18 *412*, four *antonia*, six *blastopia*, 21 *blood*, one *burdock*, two *hamilton*, eight *HMS Beagle*, four *mazi*, five *mdg1*, eight *mdg3*, one *micropia*, one *nik*, four *nomad*, 40 *roo*, six *tirant*, four *transpac*, two *wolfman*, and five unclassified elements (Table 4). The elements' combined lengths totaled 1, 279, 046 nucleotides or nearly 1% of the sequenced component of the genome.

In general, the number of individual elements that we have characterized for each LTR retrotransposon family is consistent with average copy numbers estimated previously by in situ hybridization (Table 4). In situ hybridization also detects only those elements located in the polytenized, euchromatic component of the genome. For example, the most

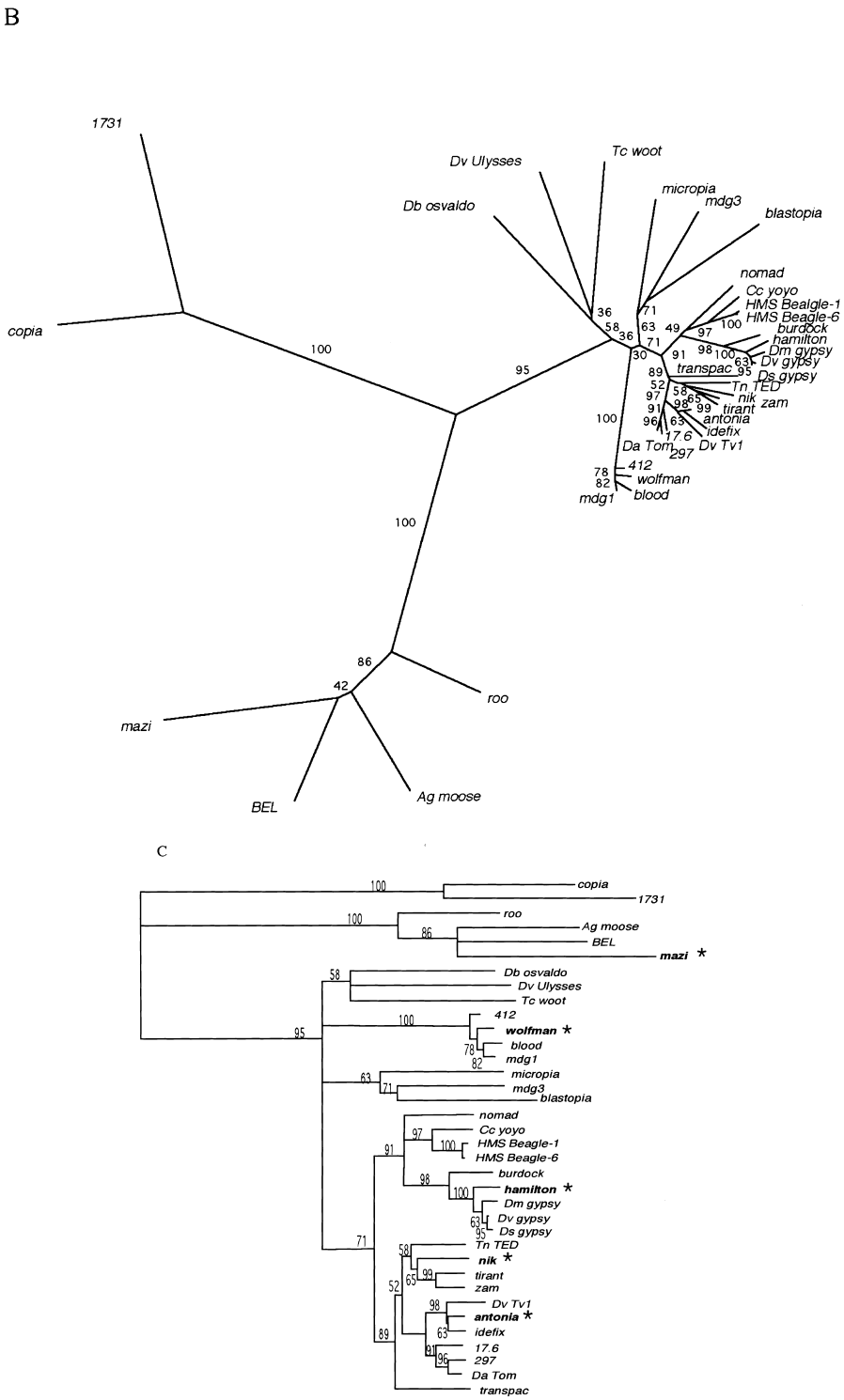

Figure 2 (continued)

abundant element found is the *roo* element, which occupies, on average, 68 ± 14 sites within the polytene chromosomes of natural *D. melanogaster* populations (Vieira et al. 1999). We also found that *roo* was the most abundant element with at least 40 full-length copies in the genome of the sequenced *D. melanogaster* lab strain. In contrast, the least abundant elements in natural populations are *1731*, *gypsy*, and *zam*. Each of these elements has, on average, less than two copies in natural populations (Vieira et al. 1999). We did not find any copies of these elements in our analysis of the *D. melanogaster* genome sequence. This is not surprising in that both *gypsy* and *zam* are known to be most abundant in constitutive heterochromatin and are in low abundance or absent from the euchromatic regions of some *D. melanogaster* strains (Pimpinelli et al. 1995; Baldrich et al. 1997).

Phylogenetic Characterization of *D. melanogaster* LTR Retrotransposons

The aligned RTs of all *D. melanogaster* LTR retrotransposons shown in Figure 2A were used to generate the phylogenetic trees presented in Figure 2, B and C. Additional RT sequences from other *Drosophila* and invertebrate species as well as *D. melanogaster* elements that were not identified in our analysis are also included in this phylogeny. The RT phylogeny indicates that there are three major groups of LTR retrotransposons within the *D. melanogaster* genome.

Copia Group

To date, members of the Copia group found in *D. melanogaster* include only *copia* and *1731*. In our study, we did not find any representatives of the *1731* family. In one instance we found a *copia* element, *copia-8*, inserted into another element (*mdg3-2*). Similar composite insertions have been observed previously in maize (San-Miguel et al. 1996) and barley (Shirasu et al. 2000).

Gypsy Group

Previously characterized Gypsy group members that we identified in this study include *17.6*, *297*, *412*, *blastopia*, *blood*, *burdock*, *HMS Beagle*, *mdg1*, *mdg3*, *micropia*, *nomad*, *tirant*, and *transpac*. Novel Gypsy group members first identified and named here are *antonia*, *hamilton*, *nik*, and *wolfman*. Five additional elements were identified that are closely related to Gypsy group elements and not characterized previously at the level of RT amino-acid identity. These elements are listed by accession number only in Figure 1. The RTs of these elements contain frame shifts or

Table 3. LTR Retrotransposons Characterized in This Study

Element	Accession #	Genomic location	ITR	DTR	LTR length (bp)	Inserted element length	LTR % identity	Element age in millions of years
17.6-1	AC005860	2R/47D1	AG/TT	ATAT/ATAT	512	7202	100	0
17.6-2	AC005894	2R/46A1-46B2	AG/TT	ATAT/ATAT	512	7200	100	0
17.6-3	AC004333	3L/62A1-62A4	AG/TT	ATAT/ATAT	512	7245	100	0
17.6-4	AE003471	3L	AG/TT	ATAT/ATAT	514	6505	100	0
17.6-5	AE003082	2R	AG/TT	ATAT/ATAT	512	6329	99.419	0.181563
17.6-6	AE003545	3L	AG/TT	ATAT/ATAT	517	6509	99.189	0.253438
297-1	AC006303	2L/35E-36A	AG/CT	ATAT/ATAT	414	6808	100	0
297-2	AC007176	2L-2R/37C1-46B2	AG/CT	ATAT/ATAT	414	6791	100	0
297-3	AC005558	2L/23A1-23B2	AG/CT	ATAT/ATAT	414	6336	100	0
297-4	AC005720	3L/79F1-80A2	AG/CT	ATAT/ATAT	414	6730	100	0
297-5	AC004306	2R/51E1-51E2	AG/CT	ATAT/ATAT	414	6805	100	0
297-6	AC002441	2L/35F3-35F6	AG/CT	ATAT/	414/3′ LTR DELETION	6809	100	
297-7	AE003413	2L/34C4-36A7	AG/CT	ATAT/ATAT	414	6808	100	0
297-8	AE003574	X	AG/CT	ATAT/ATAT	414	5708	100	
297-9	AE003415	SL/34C4-36A7	AG/CT	ATAT/ATAT	415	6809	100	0
297-10	AE003599	3L	AG/CT	ATAT/ATAT	414	5549	99.193	0.252188
412-1	AC005847	3L/61F3-62A2	TG/CT	ATAT/ACAG	571	7592	99.789	0.065938
412-2	AC005639	2R/59E3-59F4	TG/CA	AGGG/AGGG	481	7519	100	0
412-3	AC005267	3L/60A2-60B2	TG/CA	ATAT/ATAT	514	7566	100	0
412-4	AL050231	X/1A-1B	TG/TA	CAAT/TATC	512	7572	100	0
412-6	AE003417	X	TG/TA	CCAAT/TCGAC	514	7578	99.415	0.182813
412-7	AE003429	X	TG/CA	TGAA/TGAA	515		99.610	0.121875
412-8	AE003553	3L	TG/CA	GGGAGA/TGGCGA	514	7582	99.805	0.060938
412-9	AE003709	3R	TG/CA	CTAC/CTAC	517	7558	99.368	0.197500
412-10	AE003461	2R	TG/CA	AGGG/AGGG	481	7452	100	0
412-11	AE003462	2R	TG/CA	ATAT/ATAT	518	7492	99.364	0.198750
412-12	AE003463	2R	TG/CA	ATAG/ATAG	482	7454	100	0
412-13	AE003463	2R	TG/CA	CGAT/CGAT	484	7461	99.583	0.130313
412-14	AE003464	2R	TG/CA	AGTATG/CTAGAG	472	7446	100	0
412-15	AE003470	3L	TG/CA	GTCCTT/CAAGTC	518	7529	100	0
412-16	AE003470	3L	TG/CA	GATATA/ACAGCA	572	7576	99.786	0.066875
412-17	AE003511	X	TG/CA	AAAC/AAAC	514	7520	100	0
412-18	AE003523	3L	TG/CA	GTGG/GTGG	514	7520	99.790	0.065625
412-19	AE003718	3R	TG/CA	ACAG/ACAG	483	7550	100	0
antonia-1	AC005891	2L/33F3-34A2	AG/TT	ATAT/ATAT	659	6953	100	0
antonia-2	AC004362	2L/35E3-35E6	AG/TT	ATAT/ATAT	659	6943	100	0
antonia-3	AE003640	2L	AG/TT	ATAT/ATAT	661	6170	99.697	0.094688
antonia-4	AE003786	2R	AG/TT	ATAT/ATAT	661	6116	98.573	0.445938
blastopia-1	AC007177	2R/59C1-59C5	TG/CA	TATA/TATA	276	5029	100	0
blastopia-2	AC006402	2L/38C-38D	TG/CA	TATA/TATA	276	5029	100	0
blastopia-3	AC005638	2R/47E1-47E3	TG/CA	TTTA/TTTA	276	4264	100	0
blastopia-4	AL033125	DISTAL X	TG/CA	TATA/TATA	276	4645	100	0
blastopia-5	AE003417	X	TG/CA	TACA/TACA	276	4272	100	0
blastopia-6	AE003489	X	TG/CA	TGTA/TGTA	275	3851	100	0
blood-1	AC002474	2L/38D1-38D2	TG/CA	GTAC/GTAC	399	7411	100	0
blood-2	AC005734	2L/31A1-31B1	TG/CA	AAGG/AAGG	401	7413	100	0
blood-3	AC004442	2L/22A3-22A7	TG/CA	GTGG/GTGG	399	7408	100	0
blood-4	AC005130	2L/39C1-39D1	TG/CA	AGCA/AGCA	399	7415	100	0
blood-5	AC007082	2L/37D	TG/CA	CCTG/CCTG	401	7417	100	0
blood-6	L49394	2L/35E1-35E2	TG/CA	AGGG/AGGG	399	7411	100	0
blood-7	AC004115	2L/27B7-21C2	TG/CA	GAAT/GAAT	415	7443	99.758	0.07563
blood-8	AC003727	3R	TG/CA	ATAT/AGTC	395	7425	99.500	0.15625
blood-9	AE003413	2L/34C4-36A7	TG/CA	AGGG/AGGG	399	7411	100	0
blood-10	AE003647	2L	TG/CA	AGGG/AGGG	399	7421	100	0
blood-11	AE003718	3R	TG/CA	ATAC/ATAC	399	7421	100	0
blood-12	AE003775	3R	TG/CA	GTAC/GTAC	399	7421	100	0
blood-13	AE003555	3L	TG/CA	ACAT/ACAT	400	7422	99.750	0.078125
blood-14	AE003667	2L	TG/CA	GTAC/GTAC	400	7420	99.748	0.078750
blood-15	AE003670	2L	TG/CA	AGCA/AGCA	400	7420	99.751	0.077813
blood-16	AE003633	2L	TG/CA	GCAG/GCAG	400	7422	100	0
blood-17	AE003709	3R	TG/CA	ACAG/ACAG	401	7425	100	0
blood-18	AE003598	3L	TG/CA	CTGC/CTGC	401	7423	100	0
blood-19	AE003780	3R	TG/CA	CTAC/CTAC	402	7427	100	0
blood-20	AE003589	2L	TG/CA	GAAT/GAAT	415	7452	99.758	0.075625
blood-21	AE003638	2L	TG/CA	GTAC/GTAC	377	7403	99.24	0.237500

Table 3. (*Continued*)

Element	Accession #	Genomic location	ITR	DTR	LTR length (bp)	Inserted element length	LTR % identity	Element age in millions of years
burdock-1	AE002976	2R/41C-41D	AG/TT	TATA/TATA	276	6425	100	0
copia-1	AC007146	2L/33A	TG/CA	GATCC/GATCC	276	5146	100	0
copia-2	AC001657	2L/35B7-35C1	TG/CA	GTTTC/GTTTC	276	5145	100	0
copia-3	AC004445	2L/22A3-22A5	TG/CA	GTTGG/GTTGG	276	5145	100	0
copia-4	AC005554	2L/23C4-23D4	TG/CA	TTATC/TTATC	276	5145	100	0
copia-5	AC004276	2L/22E1-22E2	TG/CA	GTATC/GTATC	276	5147	100	0
copia-6	AC005647	2R/53C7-53C14	TG/CA	TATGT/TATGT	276	5151	100	0
copia-7	AC002501	2L/35F1-35F2	TG/CA	AAAAT/AAAAT	276	5149	100	0
copia-8	AC004735	2L/38A7-39C9	TG/CA	GAGCA/GAGCA	276	5144	100	0
copia-9	AC003053	2L/24A3-24C2	TG/CA	ATCAC/ATCAC	276	5145	100	0
copia-10	AC011662	2L/36D	TG/CA	ATAAC/ATAAC	275	5145	100	0
copia-11	AC006215	2L/38C1-38C2	TG/CA	ACTGC/ACTGC	275	5252	100	0
copia-12	AE003494	X	TG/CA	TTGAC/TTGAC	276	4835	99.638	0.113125
copia-14	AE003794	2R	TG/CA	GGTTC/GGTTC	276	4837	100	0
copia-15	AE003804	2R	TG/CA	ACGCC/ACGCC	276	4742	100	0
copia-16	AE003777	3R	TG/CA	GATGG/GATGG	277	4836	100	0
copia-17	AE003806	2R	TG/CA	TATGT/TATGT	277	4511	99.628	0.116250
copia-18	AE003579	2L	TG/CA	ATCAC/ATCAC	277	4747	100	0
copia-19	AE003729	3R	TG/CA	AAAAC/AAAAC	277	4506	100	0
copia-21	AE003466	2R	TG/CA	CTTTG/CTCAG	278	5176	100	0
copia-22	AE003783	2L	TG/CA	AACAC/AACAC	278	4509	100	0
copia-23	AE003526	3L	TG/CA	AATAT/AATAT	288	4564	100	0
copia-24	AE003778	3R	TG/CA	ATTAT/ATTAT	307	4741	98.536	0.457500
copia-25	AE003473	3L	TG/CA	AGAAC/AGAAC	276	4120	100	0
hamilton-1	AC006215	2L/38C1-38C2	AG/TT	TGTA/TGTA	495	6973	100	0
hamilton-2	AC006215	2L/38C1-38C2			495	6973	100	0
HMS Beagle-1	AC004563	2R/56D11-56E6	AG/CT	TATA/TATA	266	6883	100	0
HMS Beagle-2	AC007977	2L/27C	AG/CT	TATA/TATA	266	6885	100	0
HMS Beagle-3	AL121804	DISTAL X	AG/CT	TATA/TATT	266	6887	100	0
HMS Beagle-4	AE003116	3L	AG/CT	TATA/TATA	332	7180	99.034	0.301875
HMS Beagle-5	AE003574	X	AG/CT	TGTA/TGTA	332	7008	99.698	0.094375
HMS Beagle-6	AE003600	3R	AG/CT	TATA/TATA	332	7217	100	0
HMS Beagle-7	AE003679		AG/CT	TGCA/TGCA	266	4441	100	0
HMS Beagle-8	AE003651	2L	AG/CT	TGTA/TGTA	266	6906	100	0
mazi-1	AC004377	2R/58E8-58F4	TG/CA	ATGCTT/ATGGAA	224	6112	100	0
mazi-2	AE003482	3L	TG/CA	AAGGG/AAGGG	245	6150	100	0
mazi-3	AE003615	2L	TG/CA	GCTGG/GCTGG	245	6148	100	0
mazi-4	AE003777	3R	TG/CA	AGTGT/AGTGT	245	6149	100	0
mdg1-1	AL138971	DISTAL X	TG/CA	CTAG/CTAG	442	7355	100	0
mdg1-2	AE003736	3R	TG/CA	CTCC/CTCC	442	8170	99.582	0.130625
mdg1-3	AE003828	2R	TG/CA	TCAC/TCAC	444	7420	99.320	0.212500
mdg1-4	AE002943		TG/CA	CTAT/CTAT	425	8213	100	0
mdg3-1	AC004574	2L/22B1-22B2	TG/AA	GTAG/GTAG	267	5519	100	0
mdg3-2	AC004735	2L/38A7-38C9	TG/AA	GCAC/GCAC	267	5504	100	0
mdg3-3	AE003607	3R	TG/AA	ATTG/ATTG	265	5526	100	0
mdg3-4	AE003492	X	TG/AA	ATGT/ATGT	267	5399	100	0
mdg3-5	AE003514	3L	TG/AA	GTAT/GTAT	267	5331	100	0
mdg3-6	AE003604	3R	TG/AA	CCAT/CCAT	267	5330	100	0
mdg3-7	AE003686	3R	TG/AA	GTTG/GTTG	267	5526	100	0
mdg3-8	AE003502	X	TG/AA	TACT/TACT	268	5529	99.247	0.235313
micropia-1	AE003672	3R	TG/CA	TATA/TATA	514	5124	99.405	0.185938
nik	AE003485	X	AG/TT	CCAG/CCAG	430	6938	100	0
nomad-1	AE003411	2L/35B4-35C1	AG/CT	TTATA/TATA	518	7389	100	0
nomad-2	AC004716	2L/22C1-22C2	AG/CT	TTTA/TTTA	518	7473	100	0
nomad-3	AC005111	2L/31F5-32A1	AG/CT	TTCA/TTCA	518	7473	100	0
nomad-4	AL121805	DISTAL X	AG/CT	TATA/TATA	517	7475	100	0
roo-1	AC005333	2L/26A1-26A3	TG/CA	GGTAC/GGTAC	428	8067	100	0
roo-2	AC004728	2L/29D1	TG/CA	AGCGC/AGCGC	429	8218	100	0
roo-3	AC001647	2R/35A1	TG/CA	GTTAT/GTTAT	428	8211	100	0
roo-4	AC001659	2L/34D1-34D4	TG/CA	CTAAC/CTAAC	429	8209	99.528	0.147500
roo-5	AC007453	2R/49D1-49D3	TG/CA	TGGCC/TGGCC	428	7411	100	0
roo-6	AC004438	3L/68C2-68C3	TG/CA	CATTT/CATTT	428	8251	99.766	0.073125
roo-7	AC004767	3L/65A6-65A10	TG/CA	AACGG/AACGG	428	8211	99.766	0.073125
roo-8	AC004245	2L/21D1-21D2	TG/CA	GCCCC/GCCCC	428	8087	100	0
roo-9	AC005974	2R/46E1-46F6	TG/CA	TGTAT/TGTAT	428	8215	99.766	0.073125
roo-10	AC011662	2L/36D	TG/CA	CCCCC/CCCCC	428	8209	100	0

Table 3. (*Continued*)

Element	Accession #	Genomic location	ITR	DTR	LTR length (bp)	Inserted element length	LTR % identity	Element age in millions of years
roo-11	AC006415	2L/40B-40C	TG/CA	GCTTCC/GCTTCC	428	8212	100	0
roo-12	AL035311	DISTAL TIP OF X	TG/CA	GCATC/GCATC	427	7788	100	0
roo-13	AL121805	DISTAL TIP OF X	TG/CA	TTAAT/TTAAT	428	8197	100	0
roo-14	AC004349	3L/65B1-65B2	TG/CA	TAATA/	428/3′ END DELETION	5895	100	0
roo-16	AL031366	DISTAL TIP OF X	TG/CA	ATTA/ATTA	428	7758	100	0
roo-17	AL138971	DISTAL X	TG/CA	GGGAC/GGGAC	428	7593	100	0
roo-19	AE003606	3R	TG/CA	ATAAG/ATAAG	429	8211	100	0
roo-21	AE003428	X	TG/CA	CATAC/CCAGA	429	8086	99.532	0.146250
roo-22	AE003703	3R	TG/CA	AGAC/AGAC	429	8317	99.766	0.073125
roo-23	AE003497	X	TG/CA	GTCAT/GTCAT	429	8087	100	0
roo-24	AE003552	3L	TG/CA	TGGAG/TGGAG	429	8086	100	0
roo-25	AE003421	X	TG/CA	CAGCT/CAGCT	430	8251	100	0
roo-26	AE003773	3R	TG/CA	CAAAT/CAAAT	430	7332	99.767	0.072813
roo-27	AE003753	3R	TG/CA	AGGGC/AGGGC	430	8229	99.765	0.073438
roo-28	AE003756	3R	TG/CA	ATAGT/ATAGT	430	8229	99.766	0.073125
roo-29	AE003455	2R	TG/CA	ACTAC/ACTAC	431	8089	99.523	0.149063
roo-30	AE003544	3L	TG/CA	CATTT/CATTT	431	8086	99.769	0.072188
roo-31	AE003679	3R	TG/CA	GCCGG/GCCGG	431	8240	99.768	0.072500
roo-32	AE003738	3R	TG/CA	AGACC/AGACC	431	8240	99.766	0.073125
roo-33	AE003529	3L	TG/CA	ATAGG/ATAGG	431	8087	100	0
roo-34	AE003470	3L	TG/CA	ATAA/ATAA	432	8144	99.295	0.220313
roo-35	AE003455	2R	TG/CA	GCCAG/GCCAG	432	8086	99.766	0.073125
roo-36	AE003519	2R	TG/CA	GAAGG/GAAGG	432	8090	99.768	0.072500
roo-37	AE003520	3L	TG/CA	AATAG/AATAG	432	8089	99.766	0.073125
roo-38	AE003467	3L	TG/CA	TAAAC/TAAAC	433	8241	99.769	0.072188
roo-39	AE003515	3L	TG/CA	CTTGC/CTTGC	433	8091	99.761	0.074688
roo-40	AE003419	X	TG/CA	TTAC/TTAC	434	8090	99.529	0.147188
roo-41	AE003781	2L	TG/CA	ATAAC/ATAAC	431	8251	100	0
roo-42	AE003463	2R	TG/CA	CTTTA/CTTTA	430	8088	100	0
roo-43	AE003410	2L, REGION 34C4-36A7	TG/CA	ATAAC/ATAAC	428	8142	100	0
tirant-1	AC005444	2L/36D3	AG/CT	CGCG/CGCG	417	8109	100	0
tirant-2	AC004759	2L/38E1-38E9	AG/CT	CCCG/CCCG	417	7399	100	0
tirant-3	AE003829	2R	AG/CT	CGTG/CGTG	418	8016	100	0
tirant-4	AE003601	3R	AG/CT	CCCG/CCCG	418	8016	100	0
tirant-5	AE003593	3L	AG/CT	CGCG/CGCG	418	8016	100	0
tirant-6	AE003843	4	AG/CT	CGCT/CGAT	419	8017	100	0
transpac-1	AC006302	2L/34C4-34D2	AG/CT	ATAT/ATAT	330	5249	100	0
transpac-2	AE003426	X	AG/CT	ATAT/ATAT	330	5251	100	0
transpac-3	AE003762	3R	AG/CT	ATAT/ATAT	330	5247	100	0
transpac-4	AE003508	X	AG/CT	ATAT/ATAT	331	5253	100	0
wolfman-1	AC007146	2L/33A-33A	TG/CA	AAGC/AAGC	506	7345	100	0
wolfman-2	AE003439	X	TG/CA	CCTG/CCTG	506	7231	100	0
unclassified	AE002665	3L	TG/CA	AATG/AATG	266	7436	100	0
unclassified	AE002987	X	AG/TT	TATA/TATA	414	8437	99.510	0.153125
unclassified	AE003268		AG/CT	TATA/TATA	416	7185	99.760	0.075120
unclassified	AE003450	X	AG/TT	TATA/TATA	405	6538	100	0
unclassified	AL035631	DISTAL X	AG/TT	TCTA/TCTA	353	5023	99.433	0.177054

stop codons and are difficult to characterize (see above discussion). These five elements will require further analysis before they can be confidently placed phylogenetically with respect to their RT identity.

The Gypsy group found in *D. melanogaster* is composed of at least 20 different families that form three divergent clades seen in Figure 2, B and C. These clades all emerge from a central unsupported region deep within the Gypsy group. This is best illustrated in the unrooted phylogram shown in Figure 2B. One clade is composed of the elements *412*, *blood*, *mdg1* , and the novel element we named *wolfman*. This clade is well supported with a bootstrap value of 100. These four element families are closely related and form a very tight cluster at the end of a long branch that separates them from the rest of the Gypsy group.

A second clade that is less well supported (bootstrap value= 63) is composed of the elements *micropia*, *mdg3*, and *blastopia*. In contrast to the previously described group, these three elements are very distantly related to each other as indicated by very long branch lengths leading to each element.

The third clade that is found within the Gypsy group is better supported with a bootstrap value of 71. This is the most abundant clade within the *D. melanogaster* genome and contains, to date, 13 different families of elements. This clade can be divided further into two well-supported lineages containing five and eight families each. In addition to *gypsy*, *burdock*,

Table 4. The Content of LTR-Retrotransposons Analyzed from the *Drosophila melanogaster* Genome in This Study

LTR retrotransposon family	Total number of elements characterized in this study	Elements estimated by in situ hybridization in 10 natural populations[a]
copia	23	24 ± 4
17.6	6	3 ± 3
297	10	23 ± 6
412	18	28 ± 5
antonia	4	n.d.
blastopia	6	n.d.
blood	21	17 ± 3
burdock	1	10 ± 2
hamilton	2	n.d.
HMS Beagle	8	9 ± 3
mazi	4	n.d.
mdg1	4	21 ± 5
mdg3	8	14 ± 5
micropia	1	n.d.
nik	1	n.d.
nomad	4	n.d.
roo	40	68 ± 14
tirant	6	11 ± 3
transpac	4	n.d.
wolfman	2	n.d.
Unclassified	5	
Total	178	
Total nucleotide sequence length of LTR retrotransposons	1,269,435	

[a]Vieira et al. 1999

HMS Beagle, and *nomad*, the group of five contains one novel element we have named *hamilton*. Previously, only LTR nucleotide sequences were available for the *HMS Beagle* element. Here we describe the first full-length copies of this element family. *HMS Beagle* is most closely related to the *yoyo* element first characterized in the Mediterranean fruit fly, *Ceratitis capitata*.

As mentioned earlier, each LTR retrotransposon family that we have characterized consists of a group of nearly identical elements (≥99% identity at the nucleotide level). One exception to this is the *HMS Beagle* family, which contains elements that are highly related yet show some level of phylogenetic structure (Figure 1). *HMS Beagle* elements consist of two well supported phylogenetic groups that share 97% RT identity at the amino acid level (Figure 2C). An additional phylogenetic comparison based on the entire DNA sequence of the *HMS Beagle* elements supports the conclusion that *HMS Beagle* elements consist of two well-defined subgroups (Fig. 3).

The novel element found within the aforementioned group of five, *hamilton*, is most closely related to the *D. melanogaster* endogenous retrovirus *gypsy*. Interestingly, we found that the two members of the *hamilton* family of elements are tandemly duplicated and separated by a single LTR. LTR retrotransposons having this same duplicate structure have been identified previously in yeast (Roeder and Fink 1983) and flies (Csink and McDonald 1995) and are postulated to be the products of homologous recombination between two full-length elements within the LTR region.

The group of eight families within the third Gypsy group clade contains *tirant*, *zam*, *idefix*, *17.6*, *297*, *transpac*, and two novel elements we have named *antonia* and *nik*. *antonia* is most closely related to *idefix* and *Tv1* from *D. virilis*. *nik* is most closely related to *tirant*, *zam* and *TED*. *TED* was first found integrated into the genome of a baculovirus in the cells of the cabbage looper, *Trichoplusia ni* (Friesen et al. 1986).

BEL *Group*

In addition to the Gypsy and Copia clades, there is a third well-supported clade (bootstrap value = 100) that contains the *BEL*, *mazi*, and *roo* families.

The complete sequence of *BEL* has been published previously (Bell et al. 1985) and partial characterization of *roo* (first described as *B104*) has been reported (Scherer et al. 1982; Lerat and Capy 1999). *mazi*, however, is a novel element. This is the first phylogenetic characterization of both *roo* and *mazi*. Previously, BEL was the only described member from *D. melanogaster* belonging to this third group of LTR retrotransposons now referred to as the BEL group (Malik et al. 2000). Interestingly, the *roo* element seems to encode a single, long open reading frame (ORF) of 2360 amino acids that contains homology to all of the previously described motifs found in LTR retrotransposons and retroviruses (McClure 1991) (Fig. 4). In most instances a frame shift is present after the *gag* (*group specific antigen*) gene to regulate differential expression of the *gag* and *pol* regions of LTR retrotransposon and retrovirus genomes.

Aging the LTR-Retrotransposons of *D. melanogaster*

As described previously, LTR nucleotide identity can be used to estimate the time of integration (SanMiguel et al. 1998) of LTR retrotransposons and retroviruses. We have found that 125 of the LTR retrotransposons described here have identical LTRs, whereas the remaining 53 have low levels of nucleotide divergence. Identical LTRs indicate that the elements have inserted recently and have not had time to accumulate mutations between the LTRs. Using the synonymous substitution rate for *Drosophila* (Li 1997) of .016 substitutions per site per million years and the intra-element LTR divergence calculated here, we have calculated the integration time of the 53 elements with LTR nucleotide divergence. The average age of the remaining 30% (53/178) of the elements was found to be 137,000 ± 89,000 yr. These results are shown in Table 3 and Figure 5A. Our data indicate that all of the *D. melanogaster* LTR retrotransposons analyzed in this study have integrated within the last 500,000 years. Moreover, the level of divergence for most elements indicates integration times of <200,000 years.

A second method for dating TEs is to calculate the average pairwise nucleotide identity across the complete sequences of the elements that are very closely related at the phylogenetic level (Kapitonov and Jurka 1996; Costas and Naveira 2000). The assumption underlying this method is that phylogenetically related elements are identical at the time of integration and have subsequently accumulated differences attributable to host DNA polymerase substitutions. This method also assumes that no homogenization of the element sequences by molecular mechanisms related to gene conversion has occurred subsequent to their integration. Most elements we characterized were found to contain unique flanking sequences in the DTRs (see above). This indicates that gene conversion has not affected any of the se-

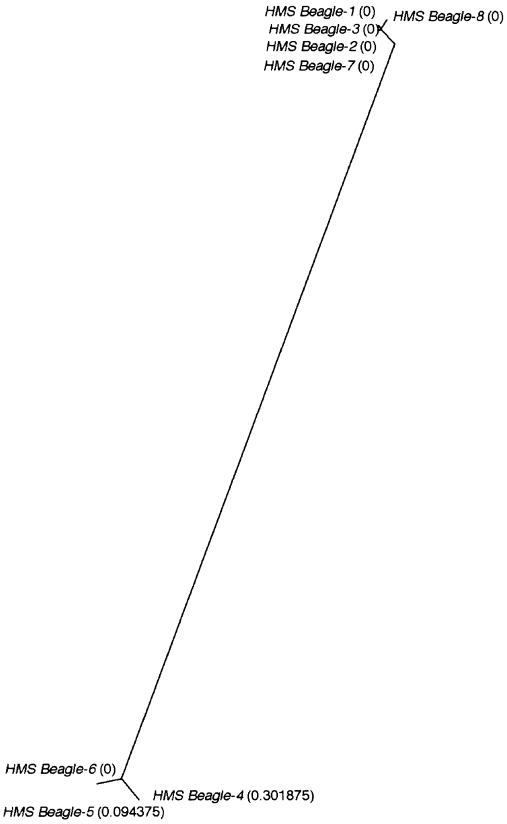

Figure 3 Unrooted neighbor joining phylogram of the nucleotide alignments of the *HMS Beagle* elements. Bootstrap values are shown on branches. The age of the individual elements as calculated by LTR sequence divergence is shown in parentheses following the name of the individual elements.

quence directly adjacent to the elements since their insertion. Although it is a formal possibility that gene conversion may have some role in homogenizing repetitive sequences, available data indicate that the magnitude of its influence is not sufficient to account for the degree of similarity we observe (Nevo-Caspi and Kupiec 1996). We analyzed each independent family of elements using this second method. The results of this independent method of aging elements also indicate that the full-length *D. melanogaster* LTR retrotransposons have integrated within the last 500,000 years (Table 2; Fig. 5B).

Therefore, both available methods of computing the age of LTR retrotransposon integration are consistent and indicate that many full-length LTR retrotransposons in *D. melanogaster* are much younger than the age of the genome in which they reside. The estimated divergence time of *D. melanogaster* from its closest relative *D. simulans* is 2.3 ± .3 million years ago (Li et al. 1999).

DISCUSSION

We have identified 178 full-length LTR retrotransposons from the sequenced, euchromatic component of the *D. melanogaster* genome. We have characterized the *D. melanogaster* LTR retrotransposons phylogenetically with respect to other known LTR retrotransposon families. In doing so, we have identified five novel families of LTR retrotransposons within the genome of *D. melanogaster* that we have named *antonia*, *hamilton*, *mazi*, *nik*, and *wolfman*. Four of these elements fall into the canonical Gypsy group of LTR retrotransposons. *mazi* groups with a third well-defined group of LTR retrotransposons present within the genome of *D. melanogaster*. Also found within this third group is the abundant element *roo*, which we found encodes a single polyprotein that contains all of the enzymes necessary for LTR retrotransposon replication. We have previously characterized six families of elements from *C. elegans* (*Cer7-12*) belonging to this newly defined third clade (Bowen and McDonald 1999), which also contains *Pao* from *Bombyx mori* and *Tas* from *Ascaris lumbricoides* (Xiong et al. 1993). This group is most closely related in structure to the Gypsy group of elements in that its *integrase* gene is found downstream or 3′ of *reverse transcriptase*. In the Copia group, *integrase* is found upstream or 5′ of *reverse transcriptase*. Judging from its almost equal phylogenetic distance from both Copia and Gypsy groups, however, this third clade likely diverged at or near the time of divergence of the Copia and Gypsy groups and represents an ancient group of LTR retrotransposons. Additional elements belonging to this third clade have since been characterized from the genomes of *Anopheles* mosquitoes (Cook et al. 2000). Even more recently, it has been claimed that elements belonging to this clade have been identified in the pufferfish *Fugu rubripes*, the ascidian urochordate *Ciona intestinalis*, and the blood fluke *Schistosoma mansoni* (Malik et al. 2000). Therefore, this third major group of LTR retrotransposons is likely to be widespread within the metazoan lineage.

Most LTR retrotransposons and retroviruses contain at least one translational frame shift following the *gag* gene to regulate the necessary overproduction of Gag relative to the other element proteins (Coffin et al. 1997). In addition to *roo*, other elements with single ORFs include *copia* from *D. melanogaster* as well as the Gypsy group members *Cer1* from *C. elegans* (Britten 1995) and *Tf1* from *Schizosaccharomyces pombe* (Levin et al. 1990). In the case of *Tf1*, a differential protein degradation process regulates the overproduction of Gag (Atwood et al. 1996). The presence of a long, single ORF in the *roo* element (Fig. 4) indicates that this characteristic is present within all three major groups of LTR retrotransposons.

Perhaps the most intriguing result to appear from our study is the fact that the *D. melanogaster* genome contains many families of full-length LTR retrotransposons, all of which have been transpositionally active in the very recent evolutionary past. Interestingly, this finding is similar to what has been observed previously for the LTR retrotransposons in *S. cerevisiae* (Jordan and McDonald 1998, 1999b) and *C. elegans* (Bowen and McDonald 1999). As shown in our results, the age of the full-length LTR retrotransposons in the *D. melanogaster* genome is substantially younger than the *melanogaster* species itself. Interestingly, the average ages of all full-length LTR retrotransposons in yeast (<100,000 yr) (Promislow et al. 1999) and nematode (<500,000 yr) (N. Bowen, unpubl.) are also much younger than the age of the species in which they are contained. In contrast to these findings, it has been reported using the same criteria we have used here that several full-length LTR retrotransposons within the ADH-region of the maize genome are much older, having transposed in the past two to six million years (SanMiguel et al. 1998). Likewise, the average age of full-length HERVs (>25

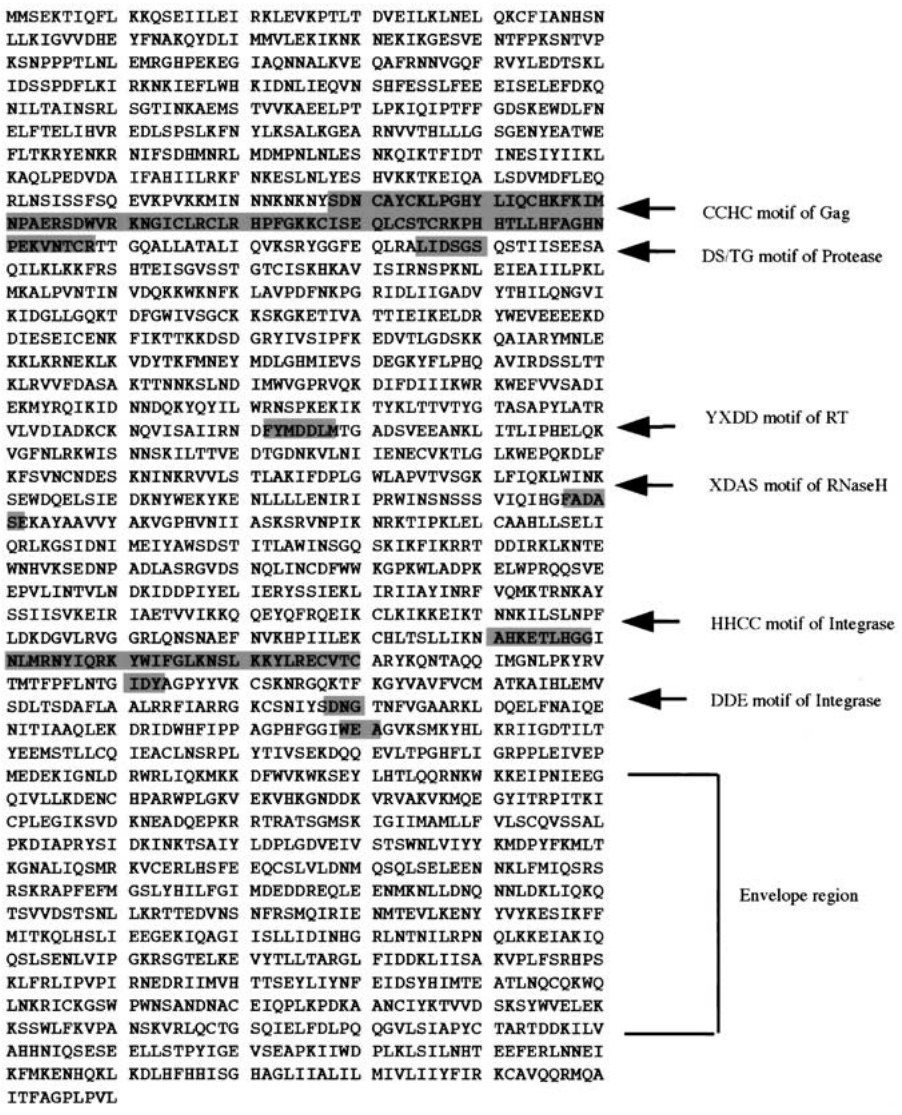

Figure 4 Translation of *roo* reveals one long ORF. The characteristic amino acid motifs of Gag, Protease, RT, Ribonuclease H (Rnase H), and Integrase of LTR retrotransposons and retroviruses (McClure 1991) are shaded and labeled in the margin. The region containing motifs similar to other Envelope proteins (Lerat and Capy 1999) is indicated near the end of the ORF.

million years) (Tristem 2000) is significantly older than the age of the human species (4–6 million years) (Yang 1996; Goodman et al. 1998).

One possible explanation for these contrasting comparisons may be differential genome size constraints placed on these species. In this regard, Adrian Bird (Bird 1995) has postulated that large increases in genome size are necessarily associated with increases in informational noise. Bird believes that the evolution of global epigenetic control mechanisms, such as methylation, were prerequisite to the significant expansions in genome size observed over the evolutionary history of higher eukaryotes. Although methylation is known to have a key role in the silencing of LTR retrotransposons in plants and vertebrates (Yoder et al. 1997), it appears to be lacking this function in many invertebrate species, including yeast, nematodes, and *Drosophila* (Russo et al. 1996). We believe that it may be for reasons such as this that full-length LTR retrotransposons have not accumulated over evolutionary time within these invertebrate genomes. As a consequence of the lack of methylation-mediated silencing in invertebrates, there would be strong selective pressure to eliminate LTR retrotransposons from these genomes.

Evidence has been presented that supports the existence of an active mechanism for the deletion of TEs in *S. cerevisiae* (Jordan and McDonald 1999b) and *Drosophila* (Petrov et al. 1996). Numerous solo LTRs exist in the *S. cerevisiae* genome as the result of intra-element LTR recombination, which serves to eliminate *Ty* elements from the host's genome (Jordan and McDonald 1999b). In *D. melanogaster*, as well as in other *Drosophila* species, DNA deletions of <400 bp are thought to occur at an astonishingly high rate within the genome, leading to a very high incidence of DNA loss (Petrov and Hartl 1997). The level of DNA loss attributable to deletions in *Drosophila* is estimated to be 75 times higher than that produced by deletions in mammals (Petrov and Hartl 1997). Consistent with this hypothesis, many of the elements that we have characterized contain sequence deletions when compared to the length of the canonical elements found in the public database (Tables 1 and 3). For example, every *17.6* element that we characterized from the *D. melanogaster* genome is shorter than the 7439 bp reported for the canonical *17.6* element (Saigo et al. 1984). These active processes that eliminate elements from genomes supply selective pressure for these elements to continually replicate or risk elimination (Jordan and McDonald 1999b). In turn, this results in only young, full-length elements within these genomes. Our results indicate that the full-length elements from the *melanogaster* genome are very young. Further support that genome size constraints can limit the accumulation of older retrotransposons comes from the recent characterization of *BARE-1* insertion patterns in *Hordeum spontaneum* (barley) (Kalendar et al. 2000). These authors have shown that there is a positive correlation between full-length *BARE-1* elements and increased genome size in barley. They further suggest that, if needed, selection for increased genome size can be regulated by limiting the amount of intra-element LTR recombination as described above for *S. cerevisiae*.

A final question concerns the immediate source of the full-length LTR retrotransposons present within the *D. melanogaster* genome. One possibility is that the full-length LTR retrotransposons are descendants from older elements that

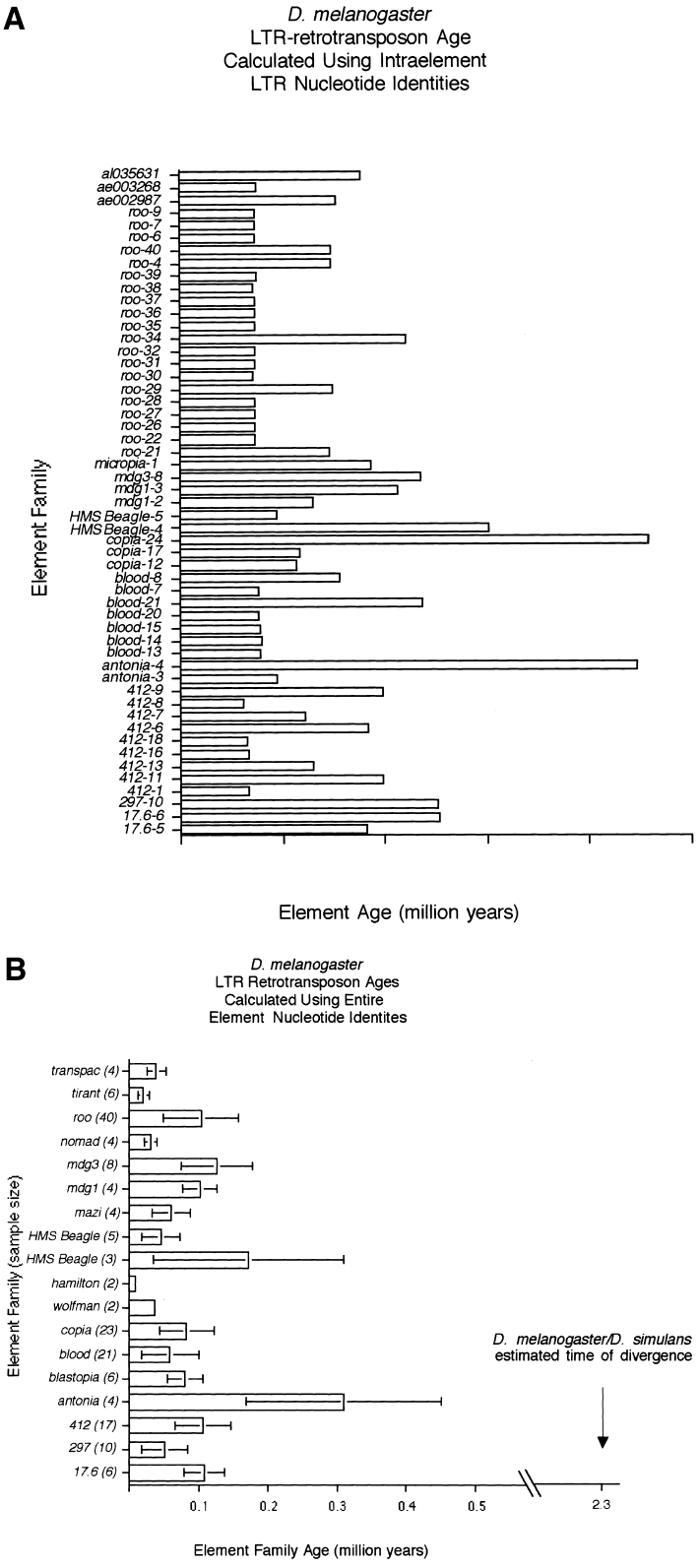

Figure 5 (*A*) Graph of LTR retrotransposon element ages of those elements that contain LTR nucleotide divergence values other than zero. (*B*) Graph of LTR retrotransposon family ages based on average pairwise identities of elements contained within a single family. The number of elements is shown in parentheses following the family name. Error bars indicate standard deviation of ages.

have been actively eliminated from the *D. melanogaster* genome or from older elements sequestered within the yet to be sequenced heterochromatin (see above discussions). An additional possibility is that the LTR retrotransposons currently present in the *melanogaster* genome have derived from elements recently introduced from other species via horizontal transfer. Recent analyses of specific families of *Drosophila* LTR retrotransposons indicate that horizontal transfer of LTR retrotransposons can occur (Jordan et al. 1999; Terzian et al. 2000). The extent of horizontal transfer and the degree to which it may have contributed to the overall composition of LTR retrotransposons that are present within the *D. melanogaster* genome remains to be determined.

Subsequent to the submission of this manuscript for publication, others (Frame et al. 2001) have reported a similar phylogenetic characterization for the members of the BEL clade. In their report, the element *Tinker* is identical to the element family that we call *mazi*. Similarly, a database of repetitive elements including a section for *Drosophila* has been made available by Genetic Information Research Institute (Jurka 2000) in which individual members of the families that we identify as *antonia*, *hamilton*, *mazi*, *nik*, and *wolfman* have been given the names *Quasimodo*, *Gtwin*, *Diver*, *Gypsy5*, and *Tabor*, respectively.

METHODS

Genome Query

Searches of the entire sequenced component of the *D. melanogaster* genome (using Advanced BLAST, http://www.ncbi.nlm.nih.gov/ blast/blast.cgi) were initiated by performing TBLASTN (Altschul et al. 1997) searches using the RT amino-acid sequence of the *Drosophila* LTR retrotransposons *BEL* (U23420), *copia* (M11240), and *mdg3* (X95908). Based on preliminary phylogenies we have constructed using the RT amino acid sequences of *D. melanogaster* LTR retrotransposons characterized previously, these three elements were chosen to represent the most divergent lineages. Nucleotide sequences with homology to the RTs were then subjected to a dot matrix (see below) analysis to reveal the presence of LTR sequences. Accession numbers that did not contain LTRs (as revealed by dot matrix analysis) were not included for further characterization. The characteristic ITRs as well as the DTRs that flank the LTRs were identified (Coffin et al. 1997). The region between LTRs was then translated to reveal coding sequences. Subsequently, the RT of each identified element was used to query the genome until all queries produced TBLASTN hits that overlapped into other element families.

Element Characterization

Each accession number containing a match to RT was retrieved from NCBI and ~10,000 bp on each side of the TBLASTN hit were subjected to further analysis. Sequences were characterized using SeqLab: The Graphical User Interface to the Wisconsin Package (GCG 1999), maintained, and made accessible by the Research Computing Resource (RCR) at the University of Georgia (UGA) (http://www.rcr.uga.edu/biosci/home.html). The dot matrix program COMPARE was used to identify regions of identity within each sequence. DOTPLOT was used to visualize the dot matrixes generated

with COMPARE. LTRs appeared as a line offset from and parallel to the identity diagonal. The terminal direct repeats were characterized from the flanking sequences of the LTRs. The terminal dinucleotides of each element LTR were also identified. RT motif amino-acid sequences of each element and the polyprotein of *roo* were predicted using TRANSLATE.

Multiple Sequence Alignments and Phylogenetic Analyses

Se-Al (courtesy of Andrew Rambaut, andrew.rambaut@zoo.ox.ac.uk) was used for multiple sequence file format manipulation and labeling. The ClustalW (Thompson et al. 1997) extension to SeqLab (GCG 1999) and ClustalX (Thompson et al. 1997) were used to generate nucleotide and amino acid alignments as described previously (Bowen and McDonald 1999). The seven conserved domains of the RT motif (Xiong and Eickbush 1988), also known as the RT ordered series of motifs (OSM) (Hudak and McClure 1999), are shown boxed in Figure 1A. Amino-acid and nucleotide alignment files may be obtained from the authors by request. PHYLIP (Felsenstein 1993) was used for distance calculation, tree production and bootstrap analysis. Phylogenetic analyses were performed on the multiple sequence alignments using distance methods employed by PHYLIP (Felsenstein 1993). The PRODIST program of PHYLIP, employing the Categories model, was used to generate distance matrices that were analyzed with the NEIGHBOR program to generate neighbor-joining tree files. SEQBOOT was also used to generate 100 data replicates that were subsequently analyzed with PRODIST (Categories model), followed by NEIGHBOR, and finally with CONSENSE to generate an unrooted bootstrapped tree as presented in Figure 2B. The phylogram presented in Figure 2C was rooted with the *1731* and *copia* elements. All trees generated were visualized with TreeViewPPC version 1.5.3 (Page 1996).

LTR Retrotransposon Age Calculation

PAUP (Swofford 1999) was used to calculate intra-element LTR identities and entire element family pairwise identities using the Kimura-2 parameter method. Ages were calculated using the formula $T = K/2'r'$, where T = time of divergence, K =divergence, and r= substitution rate (Li 1997). The average synonymous or silent site substitution rate used was .016 substitutions per site per million years as calculated by E.N. Moriyama from 39 genes between the *melanogaster* and *obscura* groups where the time of divergence was set at 30 million years ago (Li 1997).

ACKNOWLEDGMENTS

We are grateful to Drs. John Avise, Susan Wessler, Kelly Dawe, Michael Bender, and members of our laboratory for comments on earlier drafts of this manuscript. We thank Maney Mazloom for assistance in searching Genbank for *Drosophila* elements. This work was supported by a National Institutes of Health grant to J.F.M.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Atwood, A., Lin, J.H., and Levin, H.L. 1996. The retrotransposon Tf1 assembles virus-like particles that contain excess Gag relative to integrase because of a regulated degradation process. *Mol. Cell. Biol.* **16:** 338–346.

Baldrich, E., Dimitri, P., Desset, S., Leblanc, P., Codipietro, D., and Vaury, C. 1997. Genomic distribution of the retrovirus-like element ZAM in *Drosophila*. *Genetica* **100:** 131–140.

Bell, J.R., Bogardus, A.M., Schmidt, T., and Pellegrini, M. 1985. A new copia-like transposable element found in a *Drosophila* rDNA gene unit. *Nucleic Acids Res.* **13:** 3861–3871.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. 2000. GenBank. *Nucleic Acids Res.* **28:** 15–18.

Bird, A.P. 1995. Gene number, noise reduction and biological complexity. *Trends Genet.* **11:** 94–100.

Boeke, J.D. and Stoye, J.P. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In *Retroviruses* (ed. Coffin, J.M., Hughes, S.H., and Varmus, H.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bowen, N.J. and McDonald, J.F. 1999. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.* **9:** 924–935.

Britten, R.J. 1995. Active gypsy/Ty3 retrotransposons or retroviruses in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **92:** 599–601.

Coffin, J.M., Hughes, S.H., and Varmus, H.E. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, NY.

Cook, J.M., Martin, J., Lewin, A., Sinden, R.E., and Tristem, M. 2000. Systematic screening of *Anopheles* mosquito genomes yields evidence for a major clade of Pao-like retrotransposons. *Insect Mol. Biol.* **9:** 109–117.

Costas, J. and Naveira, H. 2000. Evolutionary history of the human endogenous retrovirus family ERV9. *Mol. Biol. Evol.* **17:** 320–330.

Csink, A.K. and McDonald, J.F. 1995. Analysis of copia sequence variation within and between *Drosophila* species. *Mol. Biol. Evol.* **12:** 83–93.

Deininger, P.L. and Batzer, M.A. 1999. Alu repeats and human disease. *Mol. Genet. Metab.* **67:** 183–193.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Dept. of Genetics, University of Washington, Seattle, WA.

Frame, I.G., Cutfield, J.F., and Poulter, R.T. 2001. New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. *Gene* **263:** 219–230.

Friesen, P.D., Rice, W.C., Miller, D.W., and Miller, L.K. 1986. Bidirectional transcription from a solo long terminal repeat of the retrotransposon TED: Symmetrical RNA start sites. *Mol. Cell. Biol.* **6:** 1599–1607.

Genetics Computer Group (GCG). 1999. *Wisconsin Package Version 10.0*.

Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9:** 585–598.

Green, M.M. 1988. Mobile DNA elements and spontaneous gene mutation. In *Eukaryotic transposable elements as mutagenic agents* (ed. Lambert, M.E., McDonald, J.F., and Weinstein, I.B.), pp. 41–50. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Hudak, J. and McClure, M.A. 1999. A comparative analysis of computational motif-detection methods. *Pac. Symp. Biocomput.* 138–149.

Jordan, I.K., Matyunina, L.V., and McDonald, J.F. 1999. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. *Proc. Natl. Acad. Sci.* **96:** 12621–12625.

Jordan, I.K. and McDonald, J.F. 1998. Evidence for the role of recombination in the regulatory evolution of *Saccharomyces cerevisiae* Ty elements. *J. Mol. Evol.* **47:** 14–20.

———. 1999a. Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the S*accharomyces cerevisiae* genome. *Mol. Biol. Evol.* **16:** 419–422.

———. 1999b. Tempo and mode of Ty element evolution in *Saccharomyces cerevisiae*. *Genetics* **151:** 1341–1351.

Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A.H. 2000. From the cover: Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. *Proc. Natl. Acad. Sci.* **97:** 6603–6607.

Kapitonov, V. and Jurka, J. 1996. The age of Alu subfamilies. *J. Mol. Evol.* **42:** 59–65.

Kidwell, M.G. and Lisch, D. 1997. Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci.* **94:** 7704–7711.
Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A., and Voytas, D.F. 1998. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8:** 464–478.
Lerat, E. and Capy, P. 1999. Retrotransposons and retroviruses: Analysis of the envelope gene. *Mol. Biol. Evol.* **16:** 1198–1207.
Levin, H.L., Weaver, D.C., and Boeke, J.D. 1990. Two related families of retrotransposons from *Schizosaccharomyces pombe* [published Erratum appears in *Mol. Cell. Biol.* April, 1991. **11(4):** 2334]. *Mol. Cell. Biol.* **10:** 6791–6798.
Li, W. 1997. *Molecular Evolution*. Sinauer, Sunderland, MA.
Li, Y.J., Satta, Y., and Takahata, N. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: Application of the maximum likelihood method. *Genes Genet. Syst.* **74:** 117–127.
Malik, H.S., Henikoff, S., and Eickbush, T.H. 2000. Poised for contagion: Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10:** 1307–1318.
McClure, M.A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.* **8:** 835–856.
McDonald, J.F. 1990. Macroevolution and retroviral elements. *Bioscience* **40:** 183–191.
———. 1993. Evolution and consequences of transposable elements. *Curr. Opin. Genet. Dev.* **3:** 855–864.
Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.
Nevo-Caspi, Y. and Kupiec, M. 1996. Induction of Ty recombination in yeast by cDNA and transcription: Role of the RAD1 and RAD52 genes. *Genetics* **144:** 947–955.
Page, R.D. 1996. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12:** 357–358.
Petrov, D.A. and Hartl, D.L. 1997. Trash DNA is what gets thrown away: High rate of DNA loss in *Drosophila*. *Gene* **205:** 279–289.
Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.
Pimpinelli, S., Berloco, M., Fanti, L., Dimitri, P., Bonaccorsi, S., Marchetti, E., Caizzi, R., Caggese, C., and Gatti, M. 1995. Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. *Proc. Natl. Acad. Sci.* **92:** 3804–3808.
Promislow, D.E., Jordan, I.K., and McDonald, J.F. 1999. Genomic demography: A life-history analysis of transposable element evolution. *Proc. R. Soc. Lond. B. Biol. Sci.* **266:** 1555–1560.
Roeder, G.S. and Fink, G.R. 1983. Transposable elements in yeast. In *Mobile genetic elements* (ed. Shapiro, J.A.). pp. 299–328. Academic Press, New York.
Russo, V.E.A., Martienssen, R.A., and Riggs, A.D. 1996. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Saigo, K., Kugimiya, W., Matsuo, Y., Inouye, S., Yoshioka, K., and Yuki, S. 1984. Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. *Nature* **312:** 659–661.
SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nature Genet.* **20:** 43–45.
Scherer, G., Tschudi, C., Perera, J., Delius, H., and Pirrotta, V. 1982. B104, a new dispersed repeated gene family in *Drosophila melanogaster* and its analogies with retroviruses. *J. Mol. Biol.* **157:** 435–451.
Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P., 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **10:** 908–915.
Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.
Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D., and Corces, V.G. 1994. An env-like protein encoded by a *Drosophila* retroelement: Evidence that gypsy is an infectious retrovirus. *Genes & Dev.* **8:** 2046–2057.
Swofford, D.L. 1999. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinaver Associates, Sunderland, MA.
Terzian, C., Ferraz, C., Demaille, J., and Bucheton, A. 2000. Evolution of the Gypsy endogenous retrovirus in the *Drosophila melanogaster* subgroup. *Mol. Biol. Evol.* **17:** 908–914.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.
Tristem, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J. Virol.* **74:** 3715–3730.
Vieira, C., Lepetit, D., Dumont, S., and Biemont, C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol. Biol. Evol.* **16:** 1251–1255.
Xiong, Y. and Eickbush, T.H. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Mol. Biol. Evol.* **5:** 675–690.
Xiong, Y., Burke, W.D., and Eickbush, T.H. 1993. Pao, a highly divergent retrotransposable element from *Bombyx mori* containing long terminal repeats with tandem copies of the putative R region. *Nucleic Acids Res.* **21:** 2117–2123.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42:** 587–596.
Yoder, J.A., Walsh, C.P., and Bestor, T.H. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13:** 335–340.

Received December 28, 2000; accepted in revised form June 4, 2001.