Two families of non-LTR retrotransposons, *Syrinx* and *Daphne,* from the Darwinulid ostracod, *Darwinula stevensoni*

Running title: Non-LTR elements in ostracods

Keywords: mobile genetic elements; asexual reproduction.

Isabelle Schön[a,1] and Irina R. Arkhipova[b,c,1,*]

[a]*Freshwater Biology Section, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, B-1000 Brussels, Belgium;*

[b]*Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA;*

[c]*Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA*

[1]*Both authors contributed equally to this work.*

Abbreviations: aa - amino acid(s); APE - apurinic-apyrimidinic endonuclease; bp - base pair(s); dsRNA – double-stranded RNA; EN – endonuclease; kb - kilobase(s); LTR – long terminal repeat; Myr - million years; ORF - open reading frame; rDNA - ribosomal DNA; REL - restriction enzyme-like endonuclease; RT - reverse transcriptase; SMC - structural maintenance of chromosomes; TE - transposable element; UFW - Universal Fast Walking; UTR – untranslated leader region.

Address for correspondence: *Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. Tel. (617) 495-7899. Fax: (617) 496-2444. E-mail: arkhipov@fas.harvard.edu.

Abstract

Two novel families of non-LTR retrotransposons, named *Syrinx* and *Daphne,* were cloned and characterized in a putative ancient asexual ostracod *Darwinula stevensoni*. Phylogenetic analysis reveals that *Daphne* is the founding member of a novel clade of non-LTR retroelements, which also contains newly described families from the sea urchin and the silkworm and forms a sister clade to L2-like elements. The *Syrinx* family of non-LTR retrotransposons exhibits evidence of relatively recent activity, manifested in high levels of sequence similarity between individual copies and a three- to ten-fold excess of synonymous substitutions, which is indicative of purifying selection. The *Daphne* family may have very few copies with intact open reading frames, and exhibits neutral within-family ratio of non-synonymous to synonymous substitutions. It can additionally be characterized by formation of inverted truncated head-to-head structures. All of these features make recent activity less likely than in the *Syrinx* family. Our results are discussed in light of the evolutionary consequences of long-term asexuality in general and in *Darwinula stevensoni* in particular.

1. Introduction

The paradox of sex remains the queen of evolutionary problems (Bell, 1982) until today, and the existence of the so-called ancient asexual scandals (Judson and Normark, 1996) provides one of the most exciting possibilities to investigate the long-term consequences of the absence of sex and meiosis. To date, only three examples of putatively ancient asexual animal taxa are being studied in detail at the molecular level – bdelloid rotifers (Mark Welch and Meselson, 2000), oribatid mites (Maraun *et al*., 2003) and Darwinulidae, a family of non-marine ostracods (Crustacea) (Martens *et al*., 2003). Although darwinulid molecular data on the whole are still scarce, *Darwinula stevensoni*, the type species of this ostracod family is supposed to have reproduced asexually for at least 25 Myr (Straub, 1952). It can further be characterized by an exceptionally low mutation rate in both the *hsp82* nuclear gene and the ribosomal ITS1 region (Schön and

Martens, 2003; Schön *et al*., 1998), an exceptionally long generation time of 1 to 4 years depending on the latitude (McGregor, 1969; Ranta, 1979; Van Doninck *et al*., 2003) and low fecundity (Ranta, 1979; Van Doninck *et al*., 2003). Furthermore, ecological investigations have shown that *D. stevensoni* has an extremely wide tolerance for temperature, pH (Van Doninck *et al*., 2002) and oxygen (Rossi *et al*., 2002). Likelihood permutation tests (McVean *et al*., 2002) on sequence data from *hsp82*, a calmodulin intron and ITS1 did not provide any evidence for recombination, whereas homogenizing mechanisms such as gene conversion appear to be restricted to the multi-copy ITS1 region only (Schön and Martens, 2003).

Several studies hypothesized that long-term asexuals should be encumbered by lesser amounts of parasitic DNA (Hickey, 1982; Zeyl and Bell, 1995) or that part of their genome could be genetically silenced (Schön and Martens, 2000). Moreover, an important contributor to the maintenance of sexual lineages and to the early extinction of asexuals could be the operation of meiosis-dependent mechanisms that may prevent the unchecked increase of deleterious transposable elements (TEs), in addition to general TE control mechanisms (Arkhipova and Meselson, 2000; 2005). We therefore initiated studies of TE content in *Darwinula stevensoni*, with the expectation that the patterns of distribution and evolution of TEs in these ostracods, in light of their putative ancient asexuality, may reveal interesting features with respect to TE transmission properties, proliferative and mutagenic potential, and transposon-host interactions that could result in increases or decreases in host fitness.

TEs are ubiquitous components of eukaryotic genomes, and in animal species they may comprise from only a few to more than fifty percent of total genomic DNA (Kidwell and Lisch, 2001; Kazazian, 2004). Those TEs that insert into or near genes could cause mutations, while those found in specialized chromosome compartments such as centromeric or telomeric heterochromatin could be actively participating in the establishment and function of such compartments (Pardue and DeBaryshe, 2003; Lippman *et al*., 2004; Sun *et al*., 2004). Two major categories of TEs may be distinguished by their mode of transposition and coding capacity: retrotransposons move *via* an RNA intermediate and code for reverse transcriptase (RT), while DNA transposons do not have an RNA intermediate and code for transposase.

The genome structure and TE content in species belonging to the Phylum Arthropoda, the most species-rich phylum on Earth, is primarily inferred from sequence information accumulated during genome sequencing studies of model insects, such as fruit flies, mosquitoes, and silkworm (Kaminker *et al*., 2002; Boulesteix and Biemont, 2005; Goldsmith *et al*., 2005). Little information, however, is available regarding genome structure and TE content in crustaceans. Scarce reports of crustacean TEs include such host species as an isopod and horseshoe crab (rDNA-specific non-LTR retrotransposon *R2*; Burke *et al*., 1999), waterflea (rDNA-specific DNA transposon *Pokey* from the *piggyBac* superfamily; Penton *et al*., 2002), and hydrothermal crab (*mariner*-like DNA transposon *Bytmar1*, Halaimia-Toumi *et al*., 2004; they also mention jockey-like *gag* and CR1-like RT fragments).Thus, the vast majority of crustacean TEs remains unexplored.

In view of the lack of prior information on any existing ostracod TEs, we chose the search method that is most broadly targeted towards detection of RT-related sequences and is applicable to most eukaryotic organisms (Arkhipova and Meselson, 2000). This technique targets the most conserved catalytic domains of RT with nested degenerate primers, and was shown to amplify short RT-specific products in 23 animal phyla from protists to vertebrates (*loc. cit*.). Using these fragments as a starting point, sequences can be extended in the 5' and 3' direction to yield complete copies and their adjacent flanking DNA. In this study, we focus on RT-encoding non-LTR retrotransposons, also called LINEs, which were detected by the above technique in all animals tested, with the sole exception of rotifers of the Class Bdelloidea, for which there is evidence of ancient asexuality (Mark Welch and Meselson, 2000; Mark Welch *et al*., 2004). TEs of this type rarely, if ever, undergo horizontal transmission between species and are inherited vertically from parents to offspring in most animal, plant, and fungal species (Malik, Burke and Eickbush, 1999; Eickbush and Malik, 2002). Here, we report on diversity, genomic environment, phylogenetic placement, and patterns of evolution of two different families of non-LTR retrotransposons identified in *D. stevensoni*.

2. Materials and Methods

2.1. DNA manipulations.

Living specimens of *Darwinula stevensoni* were collected from the saline lake Hollandersgatkreek (Belgium) in early summer of 2003. This population has also been subjected to ecological (Van Doninck *et al*., 2002) and genetic (Schön *et al*., 1998, Schön and Martens, 2003) studies and appears to be monoclonal (Van Doninck *et al*., 2004). DNA was extracted from 50 pooled adult females with the DNA Minikit (Qiagen) following the manufacturer's protocol, yielding about 0.5 µg of DNA. Prior to DNA extraction, animals were thoroughly washed in distilled water to avoid contamination from gut content and surface. To test for the presence of non-LTR (LINE-like) retrotransposons, we utilized nested degenerate primers corresponding to the highly conserved RT domains (Arkhipova and Meselson, 2000; Table 1). Genomic DNA of *D. stevensoni* was used in a two-step PCR reaction, and the resulting amplicons were cloned with the aid of the Topo-TA kit (Invitrogen) and sequenced. The resulting 120-130 bp sequences displaying homology to LINE-like elements were used to design four primers in each direction (Table 1), as required for Universal Fast Walking (UFW) (Myrick and Gelbart, 2002). This PCR-based technique allows amplification of unknown sequences flanking a known segment of genomic DNA without relying on the presence of restriction sites in the vicinity, and does not involve restriction or ligation steps. LA Taq DNA polymerase (error rate 8.7×10^{-6}, www.takaramirusbio.com) and custom oligonucleotides (Invitrogen) were used for UFW, with the following modifications: the annealing time in Step 1 was increased from 30 to 45 sec and the extension time from 15 sec to 1 min 30 sec in order to obtain longer first-step extension products. UFW products were subjected to Topo-TA cloning and sequencing, using custom internal primers when needed (Table 1), and the resulting sequences were assembled into contigs. Sequences were examined for divergence, intactness of open reading frames (ORFs), and used to design additional sets of primers for subsequent rounds of UFW, until the flanking host sequences were reached and an uninterrupted ORF was assembled.

2.2. Sequence analysis.

For alignments and and calculations of Ka/Ks ratios, we used the Wisconsin Package (Accelrys). Protein secondary structure prediction was done on the JPRED server (http://www.compbio.dundee.ac.uk/~www-jpred/). Phylogenetic studies and estimates of genetic divergences were performed with the Beta10 version of PAUP 4.0 (Swofford, 1998), MEGA 3.0 (Kumar *et al*., 2004), MrBayes 3.1.1 (Ronquist and Huelsenbeck, 2003), Tree-Puzzle (Schmidt *et al*., 2002) and Phylip 3.6 (Felsenstein, 2004). Neighbor-joining trees constructed with PAUP were based on relative distances (without any special model for molecular evolution) and random data input, parsimonious trees were constructed with branch-and-bound search and furthest taxa input. Maximum likelihood trees in Phylip were run with proml, the Jones model for protein evolution and variable rates among sites. Trees in Tree-Puzzle were constructed with the Adachi-Hasegawa substitution model for amino acids (Adachi and Hasegawa, 1996), gamma distributed rates, and both clock and no-clock assumptions. Bayesian analysis was run with the following settings: 4 Markov chains, 10^6 generations, Hasegawa-Kishino-Yano (DNA) or Jones (protein) model, invariable plus gamma across-site rate variation, each 100th tree sampled, and the first 200 trees discarded as burn-in. Statistical support was verfied by bootstrapping of 100 (Phylip) or 1000 (PAUP) replicates, by checking the total number of resolved quartets, or by likelihood mapping (Tree-Puzzle, Strimmer and Von Haeseler, 1997). Sequences obtained in this study were deposited in GenBank (accession Nos. XXX-YYY). Consensus sequences were deposited in Repbase.

3. Results

3.1. Initial screening for the presence of LINEs.

Two-step PCR amplification of *D. stevensoni* genomic DNA designed to amplify regions between the most conserved RT motifs (see 2.1) yielded 120-140 bp products corresponding to the region between motifs 4 and 5, which were cloned and sequenced (Fig. 1A-B). Sequence analysis of 32 individual clones revealed two groups, containing 11 and 3 sequences, which had an uninterrupted reading frame and differed by a few bp within each group. Because of the expected low sequence similarity at the DNA level, we conducted BLASTP searches of protein databases with translated amino acid sequences. These yielded low-scoring matches (as expected for such short fragments) to RT sequences of known non-LTR retrotransposons, in one case belonging to the jockey clade found in insects, and in the other case - to the CR1 clade which includes representatives from insects, vertebrates, flatworms, and nematodes (Malik *et al.*, 1999; Eickbush & Malik, 2002). We therefore focused on these two sets of sequences as possibly originating from recently active multicopy non-LTR retrotransposons from two different families. The remaining 18 sequences were shorter than the expected size range and encountered only in single clones. They carried frameshifts and/or in-frame stop codons, did not yield any matches to known RTs, and were not analyzed further.

3.2. UFW strategy.

Nucleotide sequences of the short RT fragments were used to design primers for UFW in the 5' and 3' direction (see 2.1) for each of the two families, which we named *Syrinx* and *Daphne*. UFW products from each round were cloned and sequenced from both ends, and additional primers corresponding to internal TE sequences were synthesized to fill in the gaps in longer clones (Table 1). Since the initial, most conserved RT4-RT5 fragment is not centrally located within the large RT-containing ORF, we had to employ an additional round of 5' UFW to obtain the sequence of the N-terminally located endonuclease (EN) domain, while the C-terminus of the RT, the 3' UTR, and the adjacent flanking sequences were easily reached with a single round of 3' UFW. The results of the sequence assembly for *Syrinx* and *Daphne* are presented in Fig. 1A-B and Fig. 2A. Since most of the cloned copies were more than 95% identical, it is reasonable to assume that the assembled consensus sequence differs only slightly from the sequence of the active progenitor copy.

3.3. Coding sequences.

The assembled consensus ORF sequences of *Syrinx* and *Daphne* are 871 and 902 aa in length, respectively, and consist of the ~450 aa RT domain (rvt, pfam00078) and ~350 aa N-terminally located

APE (apurinic-apyrimidinic endonuclease) domain (Exo_endo_phos, pfam03372) (Fig. 2A). All of the expected conserved motifs in each domain can be readily identified in both elements; only the EN motifs I-II could not be reached by UFW and are missing from the consensus (see 3.4). The two families exhibit only 29% and 19% amino acid identity in the RT and EN domains, respectively. The extreme C-terminal RT regions, hypothesized to be involved in recognition of different 3' terminal sequences during retrotransposition (Kajikawa *et al*., 2005), do not exhibit any similarity between the two elements. In the APE domain, the distances between motifs V-VI and VII-VIII are shorter than the corresponding AP-pinches in AP endonucleases, and are similar to most other non-LTR elements, indicating that, like other LINEs, they should not be able to recognize apurinic/apyrimidinic DNA (Kajikawa *et al*., 2005). Secondary structure predictions for the EN domain of *Syrinx* and *Daphne* also correspond to that for other non-LTR retrotransposons (Kajikawa *et al*., 2005) and do not reveal an extra β-sheet implicated in sequence-specific interaction of TRAS1 EN with *Bombyx mori* telomeric DNA (Maita *et al*., 2004) (Fig. 2A).

Out of 12 *Syrinx* 3' UFW products, five were 3' incomplete, either as truncated genomic copies or as a result of a short UFW (Figs. 1, 3B). Of the remaining seven, six had long uninterrupted ~250 aa coding sequences, somewhat diverging in the most C-terminal 30-40 aa, and one had a frameshift resulting from a $(TA)_4$->$(TA)_6$ slippage polymorphism. Out of 10 *Syrinx* 5' UFW products, three contained long uninterrupted coding sequences, three had frameshifts resulting from small (1-5 bp) indels, three were 5' truncated and had flanking sequences (Fig. 3A), and the last one, with intact RT ORF, belongs to a closely related but different *Syrinx'* subfamily, with 55% nucleotide (51% amino acid) identity to *Syrinx* (Fig. 2B).

Similarly, out of 9 *Daphne* 3' UFW products, three had an intact 270-aa coding sequence, one had a single in-frame stop codon, three had frameshifts resulting from 4-17 bp insertions, and two were incomplete at the 3' end as a result of a short 3' UFW (Fig. 3D). Out of 6 *Daphne* 5' UFW products, only one had an intact ORF, two shared a 2-bp deletion, and three were highly similar and shared the same truncation points (Figs. 1C, 3C). Two more 5' UFW products were amplified from a *Daphne*-related subfamily, which had 43% nucleotide (35% amino acid) sequence identity to *Daphne* (Table 2), and two additional decayed copies had only 24/43% amino acid identity/similarity (Fig. 2B).

3.4. 3' and 5' termini.

The 3' UTR of *Daphne* is about 440 bp in length and includes the AATAAA signal immediately followed by $(TTA)_{4\text{-}6}$. Such microsatellite-like endings are thought to originate from RT slippage during its engagement, and are often observed at the 3' termini of non-LTR retrotransposons instead of the more common oligo(A). Interestingly, *Syrinx* has a much longer 3' UTR totaling about 1 kb in length, which includes two internal polymorphic oligo(A) stretches ($A_{6\text{-}12}$ and $A_{12\text{-}20}$) and ends in a typical variable-length oligoadenylate stretch $(A)_{10\text{-}25}$, which is, however, separated by 425 bp from the nearest preceding AATAAA signal. Long 3' UTRs are generally not typical of non-LTR retrotransposons, the only exception being telomere-associated retrotransposons for which 3' UTRs may reach several kb in length (Pardue and DeBaryshe, 2003).

Two rounds of UFW in the 5' direction allowed us to approach the N-terminus of the EN domain, but additional UFW rounds did not yield any further extensions in the 5' direction. In the case of *Daphne*, extensions became complicated because of its apparent propensity to form head-to-head structures, consisting of two 5'-truncated copies of different length in inverted orientation (Fig. 1C). Four identified types of these structures consist of a *Daphne* copy truncated within the RT domain and another inverted *Daphne* copy truncated within the EN domain; all truncations occurred at different points. In three cases, it was possible to identify microhomology overlaps at the inversion junction (Fig. 1C). Obviously, the existence of such structures represents a major obstacle to UFW, as the primers would invariably match both copies in inverted orientation and yield preferential amplification of inverted-repeat junctions. Although such structures were not encountered for *Syrinx*, UFW beyond the EN domain did not operate either, which could be explained by the presence of very few full-length master copies. Eventually, the primers began to amplify unrelated high-copy-number sequences, such as other DNA transposons, which had fortuitous matches with the terminal nucleotides of the primer.

3.5. Divergence patterns.

For *Syrinx*, we compared the sequences of 11 clones from 5' UFW and 12 clones from 3' UFW (0.5-4.5% difference) that contained exclusively coding sequences. The average distances based on the branch lengths from the Bayesian trees were 2.08% (± 0.88%) for the 5' and 2.57% (± 1.05%) for the 3' end of *Syrinx*. We observed three- to ten-fold excess of synonymous substitutions for each pairwise comparison, which is indicative of purifying selection (Li, 1997) acting on *Syrinx*-encoded RT (Table 3A,B). Such selection is likely to be element-based rather than host-based, as the commonly observed *cis*-preference in retrotransposition of non-LTR elements does lead to preferential proliferation of active copies (Wei *et al*., 2001). We did not observe evidence of purifying selection acting on *Daphne* sequences (Table 3C), which is not too surprising, as most of the cloned copies were apparently 5' truncated. Average distances for cloned *Daphne* sequences were 1.80% (± 0.74%) and 2.70% (± 1.83%) for the 5' and 3' end, respectively.

3.6. Adjacent flanking regions.

We examined up to a kilobase of flanking host DNA adjacent to *Syrinx* and *Daphne*, obtained in the course of UFW. For *Syrinx*, six UFW clones extended into the 3' flanking region. Two of these clones had very similar 3' flanks with only 13 differences over a 1.75-kb 3' terminal region of *Syrinx* (4 in the coding sequence and 9 in the 3' UTR) and 3 differences in a 400-bp flanking region, suggesting their origin from recent duplication, gene conversion, or allelic difference, since the number of differences is too high to be attributed to PCR errors. Another clone had a 250-bp segment of very limited nucleotide sequence similarity (59%) to these two, interrupted by $(TGTC_4)_3$. No coding sequences were found in the 3' flanks, except for a 5' truncated fragment of a decayed *R2*-like retrotransposable element in the same transcriptional orientation. In three 5’ truncated *Syrinx* copies, we found several minisatellite repeats (9-15 bp, repeated 2-7 times) in the 5' flanking regions, and two of these regions also contained long polyguanine tracts (G_{15-16}). No poly(G) tracts were found in the 3' flanks. Finally, one of the UFW products contained a copy of *Syrinx* that underwent an insertion of a *mariner*-like element in the opposite transcriptional orientation upstream of RT motif 1.

For *Daphne*, 3' flanking sequences were inspected in four different copies. Two of the flanks were highly similar, and differed by 10 single nucleotide substitutions and by a 87-bp indel. None of the 5' UFW products could have originated from an intact *Daphne* copy, since all of them exhibited 5' truncation. Two *Daphne'* copies from a related family contained 5' flanking sequences with different truncation points; the flanks did not exhibit any characteristic features such as repeats or known genes. Three 5' UFW *Daphne* products (L2a34,41,45) were arranged in a head-to-head truncated orientation, as described above, and had different 5' truncation points, indicating independent origin of each insertion event (Fig. 1C). Three more, highly similar, UFW products apparently originated from the same insertion event: the junctions between 5' truncated copies are identical and contain 0.5 kb of unrelated sequence (Fig. 1C; L2a28,29,43). There are 7 differences over 2 kb between the two most similar copies with this structure, and 15 differences between the more divergent pairs. Moreover, the longest intact *Daphne* ORF segment (L2a3B, Fig.3C) is also truncated at the same junction, with the same unrelated sequence, and differs from the other three by 1-2 bp throughout ~0.7 kb of overlapping sequence, indicating that all of them share the same origin. A combination of recent segmental duplication, gene conversion, and/or allelic divergence could account for such differences.

3.7. Phylogenetic placement.

While initial BLAST searches indicated affiliation of *Syrinx* and *Daphne* with the jockey and CR1 clades, respectively, phylogenetic analysis of the combined EN and RT domains reveals a more intriguing position for both elements (Fig. 4). Our analysis was conducted using non-LTR retrotransposons with the N-terminal APE domain, and did not include the early-branching lineages with the REL (restriction enzyme-like endonuclease) domain to avoid loss of resolution from limiting the analysis to the RT domain. In the combined EN+RT phylogeny spanning approximately 900 amino acids in a dataset of 45 non-LTR elements with representatives of the known clades, *Syrinx* appears as a sister element to the jockey clade, although it is not possible to say whether it is a basal branch of this clade or a new sister clade until other *Syrinx*-related elements are identified. Inclusion of both the RT and the accompanying AP-EN domain in the phylogenetic analysis, based on the assumption of monophyletic EN acquisition, allows better resolution of the overall phylogeny of AP-EN-containing non-LTR retrotransposons: the jockey clade assemblage (Eickbush and Malik, 2002), consisting of jockey and CR1 clades, which was poorly supported in the RT- or EN-based neighbor-joining analyses, is now supported with 100% clade credibility value.

Notably, *Daphne* is the founding member of a new clade, which appears as a sister clade to human *L2*, eel *UnaL2* and fugu *Maui* elements (Fig. 4) (Poulter *et al*., 1999; Kapitonov and Jurka, 2003; Kajikawa *et al*., 2005). We were able to identify two additional members of the Daphne clade, which we named *Sake* and *Urca*, by computer-assisted search of the genome sequencing project databases of the silkworm *Bombyx mori,* and the sea urchin *Strongylocentrotus purpuratus*, respectively. The complete consensus *Sake* element is 5.1 kb in length, has a 240- bp 3' UTR and a 280-bp 5' UTR, and ends with $(TTTGA)_n$. As judged by the presence of three different polymorphic variants of the *Sake* 5' UTR, there are at least three different full-length master copies present in the *B. mori* genome, while its 3' UTR is present in hundreds of copies, reflecting a high degree of 5' truncation. We also identified a 548-aa upstream ORF1 associated with the *Sake* element (see Discussion). The consensus *Urca* element codes for a 985-aa EN/RT ORF, has a 380-bp 3' UTR, and ends in $(CCAA)_n$, immediately preceded by the AATAAA signal. The presence of at least three copies of a tandem repeat upstream of the *Urca* ORF did not allow us to assemble a complete consensus sequence from the available trace reads, and its 5' end remains undefined until larger contigs become available in the whole-genome assembly.

Neighbour-joining trees showed similar tree topologies with good bootstrap support when compared to the Bayesian tree. Maximum-likelihood trees constructed with Phylip were alike although they lacked strong bootstrap support for deep branches. Maximum-likelihood trees constructed with Tree-Puzzle also confirm the topologies from Bayesian reconstructions. Quartets (groups of 4 sequences) are used as an alternative to bootstrapping in Tree-Puzzle to test for statistical support. Likelihood mapping provided additional evidence for a better-supported phylogeny if *Syrinx* and *Daphne* sequences were clustered within their groups from the Bayesian trees. With clustering, 87.7 and 95.2%, respectively, of all quartets favoured treelikeness. Without the two clusters, only 63.2% of all quartets supported treelike structures.

4. Discussion

4.1. The Daphne *clade.*

This study provides the first insights into diversity, structural organization, and phylogeny of retrotransposable elements in the poorly explored class Ostracoda of the arthropod subphylum Crustacea. Interestingly, we were able to identify a novel clade of non-LTR retrotransposons, called Daphne, which currently includes representatives from crustaceans, insects, and echinoderms. The fact that members of the Daphne clade can be found in both protostomes and in deuterostomes most likely indicates its ancient origin, since horizontal transfers of non-LTR elements are exceedingly rare. It has been previously argued that the origin of each non-LTR clade dates back to the pre-Cambrian era (Malik *et al*., 1999). The position of *Syrinx* with respect to the jockey clade also agrees with its vertical inheritance from the common ancestor of arthropods. It is also worth mentioning that the recently described *L2*-like non-LTR retrotransposons *Samurai* and *Abyss* from *B. mori* and *Anemonia sulcata*, respectively (Abe *et al*., 2005; Greenwood *et al*., 2005), clearly belong to the L2, and not to the *Daphne*, clade.

The detection of a novel kind of an upstream ORF1 in one of the members of the Daphne clade, the *Sake* element from *B. mori*, is also of interest, as it increases the diversity of additional ORFs found in non-LTR retrotransposons. The N-terminal half of *Sake* ORF1, with a coiled-coil domain, reveals a statistically significant homology to SMC proteins (structural maintenance of chromosomes; domain COG1196 in the database of clusters of orthologous groups) found in many eukaryotes, with E-values of similarity to SMC3 proteins from PSI-BLAST iterations reaching 2x10^{-84}. SMC3 is a subunit of the heterodimeric SMC1/SMC3 cohesin complex, which is essential for sister chromatid cohesion and also acts in recombination and double-strand break repair (Losada and Hirano, 2005). However, it is also possible that it is the coiled-coil domain *per se* that yields high E-values, as no other motifs expected for SMC proteins can be identified in *Sake* ORF1. The only other case of a coiled-coil domain in non-LTR ORF1 is known in mouse L1 elements, where it is essential for L1 retrotransposition, ORF1 trimerization, and nucleic acid chaperone activity (Martin *et al*., 2003). There is a +1 frameshift between ORF1 and the EN/RT-containing ORF2, and ORF2 expression might occur *via* ribosomal frameshifting, similar to Ty1 (Kawakami *et al*., 1993), rather than by re-initiation, since the first available ATG codon is located between the EN motifs I and II.

Although the presence of a similar ORF1 in other members of the Daphne clade is now only a matter of speculation, it is not unreasonable to suggest that its presence may be a characteristic feature of the clade, similar to the esterase- or PHD-domain-containing ORF1 from members of the CR1 clade (Kapitonov and Jurka, 2003). Members of the jockey clade, to which *Syrinx* may belong, usually contain a different type of ORF1, with several Zn-knuckle motifs characteristic of retroviral gag proteins, which is believed to participate in the replication cycle by binding to the template RNA and participating in formation of an RNP complex consisting of template RNA, ORF1p, and ORF2p (Eickbush and Malik, 2002).

4.2. Non-LTR retrotransposons Syrinx *and* Daphne *from Darwinula stevensoni.*

We characterized in detail two families of non-LTR, or LINE-like, retrotransposable elements, *Syrinx* and *Daphne,* which occupy different positions in the overall phylogeny of non-LTR retrotransposons. One of these families, *Daphne,* is the founding member of a novel clade of non-LTR retrotransposons, which also includes elements from organisms as diverse as sea urchins and silkworms.

Most of the *Syrinx* sequences analyzed are 95-99% similar, suggesting their relatively recent activity. Moreover, these copies exhibit strong evidence of purifying selection, manifested in a three- to ten-fold excess of synonymous over non-synonymous substitutions. This pattern is typical of other non-LTR retrotransposons, which usually exhibit a strong *cis*-preference in transposition, leading to preferential proliferation of active copies (Pelisson *et al*., 1991; Wei *et al*., 2001). The difficulties in obtaining 5' terminal sequences may indicate that proliferation of these elements depends on only a few master copies, perhaps even a single copy, resulting in poor representation of 5' ends in genomic DNA and overabundance of 3' ends. Such organization is known in *Caenorhabditis elegans* and *C. briggsae*, in which one or a few master copies give rise to a large number of 5' truncated copies (Marin *et al*., 1998; Zagrobelny *et al*., 2004). Another possible explanation for UFW failure in the 5' direction would be terminal localization of *Syrinx* copies, similar to *HeT-A* and *TART* elements of *D. melanogaster* (Pardue and DeBaryshe, 2003), which would lead to difficulties in amplifying terminally exposed ends. If *Syrinx* is associated with telomeres, its atypically long and A-rich 3' UTRs may share functional similarities with long 3' UTRs of *TART* and *HeT-A*, which are thought to play a role in formation of telomeric heterochromatin. In this respect, it is worth noting that telomere-associated retrotransposons in *Giardia lamblia* also possess long 3’ UTRs and, moreover, give rise to both sense and antisense small RNAs (Arkhipova and Morrison, 2001; Ullu *et al*., 2005).

Daphne, on the other hand, is less likely to be active, although a few clones do contain intact reading frames and may originate from active elements. However, the neutral pattern of amino acid substitutions in comparisons between copies indicates that most of the copies could be defective, which agrees with the fact that most of the 5' UFW products were in a head-to-head 5' truncated arrangement and could no longer be active. Such inverted structures are not very typical of LINE-like elements in general, although they were described previously for two elements (Ostertag and Kazazian, 2001; Burke *et al*., 2002). In mammalian L1 elements, the junctions between inverted segments contain no extra sequences and have a microhomology overlap (Martin *et al*., 2005). The structures we observed in three head-to-head inverted *Daphne* copies are very similar and were likely formed by the same mechanism as the mammalian L1 inversion junctions, with the added complexity of utilization of two different templates (as in L2a34, Fig. 1C). A likely explanation for the presence of unrelated DNA at the junction of four highly similar *Daphne* copies is its origin from readthrough transcription initiated by a fortuitous upstream promoter. In this case, the *Daphne* RT should act mostly in *trans* (on other copies) and not in *cis*, since at least one of the two copies found in the inverted arrangement could not have originated from the RT-providing copy because of its incompleteness. This is also supported by finding of a shared 2-bp deletion in two *Daphne* ORF clones, which apparently originated from different insertion events, having formed different inversion junctions (Fig. 1C). In combination with a high degree of 5' truncation, the tendency to form inverted repeats and the high efficiency of *trans*-action may preclude the spread of full-length intact elements throughout the genome. It may also contribute to efficient silencing of all homologous elements and to heterochromatin formation by facilitating dsRNA production from such hairpin-forming structures.

Comparison of adjacent flanking sequences in both elements did not reveal any pronounced insertion site preferences, although certain trends may be pointed out. For *Syrinx*, each of the 5' flanks carried minisatellite-like repeats, and in two cases they were also associated with long oligo(G) tracts. The remnant of an *R2*-like retrotransposon detected in the 3' flanking sequence of *Syrinx* is not associated with rDNA, as is expected for this group of retrotransposons; most rDNA-specific elements in non-rDNA locations are inactive and undergo decay (e.g. Xiong *et al*., 1988). The presence of either simple repeats or other transposons (*R2, mariner*) in the vicinity and the absence of any recognizable genes argue in favor of *Syrinx* association with heterochromatic regions, as observed, for example, for a number of plant retrotransposons (Jiang *et al*., 2003; Lippman *et al*., 2004). Weak sequence similarity observed in some of the flanking regions may reflect affinity for certain targets, although their nature remains obscure.

Within each family, two copies were found to have nearly-identical or very similar 3' flanks, with the number of differences too high to be attributed to PCR errors. The simplest explanations for the appearance of such copies are either allelic divergence or recent segmental duplication. Specific insertion into the same target is more difficult to imagine, as utilization of other target sites indicates no strict target-site preference. Duplications, however, could have played a role in formation of extra copies, since for

Daphne we could identify three or perhaps even four highly similar copies of the same insertional event, as judged by the identity of the junction segment between two inverted copies.

As expected for non-LTR retrotransposons, no signs of horizontal transfer were detected so far in overall phylogenetic analysis of RT / EN domains, although rigorous tests would require additional analysis of other ostracod species. Members of both families could have existed in darwinulid ostracods throughout their evolutionary history, as average genetic distances amongst among different copies of *Syrinx* or *Daphne* would equal at least 2-10 Myr of molecular evolution. These estimates, however, are rather speculative, as the only molecular clock available for the nuclear genome of *Darwinula stevensoni* is based on the ribosomal ITS1 region (Schön *et al*., 2003). It remains to be determined whether these elements have acquired any host functions, possibly related to maintenance of such essential structural elements of chromosomes as centromeres and telomeres (Hall *et al*., 2003; Lippman *et al*., 2004; Sun *et al*., 2004).

Our findings on the presence of LINE-like elements in *Darwinula stevensoni* differ from those in anciently asexual bdelloid rotifers (Arkhipova and Meselson, 2000, 2005), where no such elements could be detected. This is not necessarily evidence against ancient asexuality of darwinulid ostracods, although rare sex still remains a possibility for both darwinulids and bdelloids. Both groups of putative ancient asexuals differ in several aspects, such as mutation rates (Schön and Martens, 2003), generation time, and number of offspring (Van Doninck *et al*., 2003). It is not unlikely that deleterious genetic elements were still present in the genomes of darwinulid ostracods when sexual reproduction was abandoned. The long generation time of *D. stevensoni* together with a low overall mutation rate might turn the purging of such elements into an exceptionally slow process. In addition, the processes balancing the overall load of transposable elements may reveal additional levels of complexity. If *Syrinx* or *Daphne* have become domesticated and acquired any host functions, they might have become part of darwinulid genomes despite ancient asexuality. Future research should reveal in more detail how reproductive mode and load of transposable elements are related in different taxonomic groups.

Acknowledgements

We thank Koen Martens (Brussels) and Matthew Meselson (Cambridge) for helpful discussions and critical reading of the manuscript. I.S. acknowledges funding from the EC (Marie-Curie grant BIO-4-CT-98-5086) and the Belgian OSTC (MO/36/005), and I.A. would like to thank the U.S. National Science Foundation.

References

Abe H, Seki M, Ohbayashi F, Tanaka N, Yamashita J, Fujii T, Yokoyama T, Takahashi M, Banno Y, Sahara K, Yoshido A, Ihara J, Yasukochi Y, Mita K, Ajimura M, Suzuki MG, Oshiki T, Shimada T, 2005. Partial deletions of the W chromosome due to reciprocal translocation in the silkworm *Bombyx mori*. Insect Mol Biol 14, 339-352.

Adachi J, Hasegawa M, 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42, 459-468.

Arkhipova I, Meselson M., 2000. Transposable elements in sexual and ancient asexual taxa. Proc Natl Acad Sci USA 97, 14473-14477.

Arkhipova I, Meselson M., 2005. Deleterious transposable elements and the extinction of asexuals. Bioessays 27, 76-85.

Arkhipova IR, Morrison HG, 2001. Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. Proc. Natl. Acad. Sci. USA 98,14497-14502.

Bell G., 1982. *The masterpiece of nature: the evolution and genetics of asexuality*. Berkeley: Univ. of California Press.

Boulesteix M, Biemont C., 2005. Transposable elements in mosquitoes. Cytogenet Genome Res. 110, 500-509.

Burke WD, Malik HS, Jones JP, Eickbush TH., 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Mol Biol Evol. 16, 502-511.

Burke WD, Malik HS, Rich SM, Eickbush TH., 2002. Ancient lineages of non-LTR retrotransposons in the primitive eukaryote, *Giardia lamblia*. Mol Biol Evol. 19, 619-630.

Eickbush TH, Malik HS., 2002. Origin and evolution of retrotransposons. In: Mobile DNA II, Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. Washington, DC: ASM Press.

Felsenstein J., 2004. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author.* Dept. Genome Sci., Univ. Washington, Seattle.

Feng Q, Moran JV, Kazazian HH Jr, Boeke JD., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87, 905-916.

Goldsmith MR, Shimada T, Abe H., 2005. The genetics and genomics of the silkworm, *Bombyx mori*. Annu Rev Entomol. 50, 71-100.

Greenwood AD, Leib-Mosch C, Seifarth W., 2005. Abyss1: a novel L2-like non-LTR retroelement of the snakelocks anemone (*Anemonia sulcata*). Cytogenet Genome Res 110, 553-558.

Halaimia-Toumi N, Casse N, Demattei M, Renault S, Pradier E, Bigot Y, Laulier M., 2004. The GC-rich transposon Bytmar1 from the deep-sea hydrothermal crab, *Bythograea thermydron*, may encode three transposase isoforms from a single ORF. J Mol Evol. 59, 747-760.

Hall IM, Noma K, Grewal SI., 2003. RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. Proc Natl Acad Sci USA 100, 193-198.

Hickey DA, 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101, 519-531.

Jiang J, Birchler JA, Parrott WA, Dawe RK., 2003. A molecular view of plant centromeres. Trends Plant Sci 8, 570-575.

Judson PO, Normark BB., 1996. Ancient asexual scandals. Trends Ecol Evol 11, 41-46.

Kajikawa M, Ichiyanagi K, Tanaka N, Okada N., 2005. Isolation and characterization of active LINE and SINEs from the eel. Mol Biol Evol 22, 673-682.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, Ashburner M, Celniker SE., 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3(12), RESEARCH0084.

Kapitonov VV, Jurka J., 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. Mol Biol Evol 20, 38-46.

Kawakami K., Pande S, Faiola B, Moore DP, Boeke JD, Farabaugh PJ, Strathern JN, Nakamura Y, Garfinkel DJ, 1993.. A rare tRNA-Arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in *Saccharomyces cerevisiae*. Genetics 135, 309–320.

Kazazian HH Jr., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626-1632.

Kidwell MG, Lisch DR., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution 55, 1-24.

Kumar, S, Tamura, K, Nei, M., 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics 5, 150-163.

Li WH, 1997. Molecular evolution. Sinauer Associates, Sunderland, MA.

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R., 2004. Role of transposable elements in heterochromatin and epigenetic control. Nature 430, 471-476.

Losada A, Hirano T., 2005. Dynamic molecular linkers of the genome: the first decade of SMC proteins. Genes Dev 19, 1269-1287.

Maita N, Anzai T, Aoyagi H, Mizuno H, Fujiwara H., 2004. Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. J Biol Chem 279, 41067-41076.

Malik HS, Burke WD, Eickbush TH., 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16, 793-805.

Maraun M, Heethoff M, Scheu S, Norton RA, Weigmann G, Thomas RH., 2003. Radiation in sexual and parthenogenetic oribatid mites (Oribatida, Acari) as indicated by genetic divergence of closely related species. Exp App Acarol 29, 175-187.

Marin I, Plata-Rengifo P, Labrador M, Fontdevila A., 1998. Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode *Caenorhabditis elegans*. Mol Biol Evol. 15, 1390-1402.

Martin F, Olivares M, Lopez MC, Alonso C., 1996. Do non-long terminal repeat retrotransposons have endonuclease activity? Trends Biochem Sci 21, 283-285.

Martin SL, Branciforte D, Keller D, Bain DL., 2003. Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci USA 100, 13815-13820.

Martin SL, Li WL, Furano AV, Boissinot S., 2005. The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenet Genome Res 110, 223-228.

Mark Welch D, Meselson M., 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science 288, 1211-1215.

Mark Welch JL, Mark Welch DB, Meselson M., 2004. Cytogenetic evidence for asexual evolution of bdelloid rotifers. Proc Natl Acad Sci USA 101, 1618-1621.

Martens K, Rossetti G, Horne DJ., 2003. How ancient are ancient asexuals? Proc R Soc Lond B 270, 723-729.

McGregor DL, 1969. The reproductive potential, life history and parasitism of the freshwater ostracod *Darwinula stevensoni*. In: Neale, J.W. (ed.) *The taxonomy, morphology and ecology of recent Ostracoda*. Oliver & Boyd, 194-221.

McVean G, Awadalla P, Fearnhead P., 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160, 1231-1241.

Myrick KV, Gelbart WM, 2002. Universal Fast Walking for direct and versatile determination of flanking sequence. Gene 284, 125-131.

Ostertag EM, Kazazian HH Jr., 2001. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11, 2059-2065.

Pardue ML, DeBaryshe PG., 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 37, 485-511.

Pelisson A, Finnegan DJ, Bucheton A., 1991. Evidence for retrotransposition of the I factor, a LINE element of *Drosophila melanogaster*. Proc Natl Acad Sci USA 88, 4907-4910.

Penton EH, Sullender BW, Crease TJ., 2002. Pokey, a new DNA transposon in *Daphnia* (Cladocera: Crustacea). J Mol Evol 55, 664-673.

Poulter R, Butler M, Ormandy J., 1999. A LINE element from the pufferfish (fugu) *Fugu rubripes* which shows similarity to the CR1 family of non-LTR retrotransposons. Gene 227, 169-179.

Ranta E., 1979. Population biology of *Darwinula stevensoni* (Crustacea, Ostracoda) in an oligotrophic lake. Ann. Zool. Fennici 16, 28-35.

Ronquist F, Huelsenbeck JP., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Rossi V, Todeschi EBA, Gandolfi A, Invidia M, Menozzi P., 2002. Hypoxia and starvation tolerance in individuals from a riverine and a lacustrine population of *Darwinula stevensoni* (Crustacea: Ostracoda). Arch Hydrobiol 154, 151-171.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A., 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502-504.

Schön I, Butlin RK, Griffiths HI, Martens K., 1998. Slow molecular evolution in an ancient asexual ostracod. Proc R Soc Lond B 265, 235-242.

Schön I, Martens K, 2000. Transposable elements and asexual reproduction. Trends Ecol Evol 15, 287-88.

Schön I, Martens K, Van Doninck K, Butlin RK, 2003. Evolution in the slow lane: molecular rates of evolution in sexual and asexual ostracods (Crustacea: Ostracoda). Biol J Linn Soc 79, 93-100.

Schön I, Martens K., 2003. No slave to sex. Proc R Soc Lond B 270, 827-833.

Stimmer K, Von Haeseler A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of sequence alignment. Proc Natl Acad Sci USA 94, 6815-6819.

Sun FL, Haynes K, Simpson CL, Lee SD, Collins L, Wuller J, Eissenberg JC, Elgin SC., 2004. *cis*-Acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. Mol Cell Biol 24, 8210-8220.

Swofford DL., 1998. PAUP*. Phylogenetic Analysis Using Parsimony (and other methods)*, version 4.0. Sinauer Assoc., Sunderland, MA.

Ullu E, Lujan HD, Tschudi C., 2005. Small sense and antisense RNAs derived from a telomeric retroposon family in *Giardia intestinalis*. Eukaryot Cell 4, 1155-1157.

Van Doninck K, Schön I, De Bruyn L, Martens K., 2002. A general purpose genotype in an ancient asexual. Oecologia 132, 205-212.

Van Doninck K, Schön I, Martens K, Godderis B., 2003. The life cycle of the ancient asexual ostracod *Darwinula stevensoni* (Brady & Robertson, 1870) (Crustacea, Ostracoda) in a temperate pond. Hydrobiol 500, 331-340.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV., 2001. Human L1 retrotransposition: *cis* preference versus *trans* complementation. Mol Cell Biol. 21; 1429-1439.

Xiong Y, Burke W, Jakubczak J Eickbush T 1988 Ribosomal DNA insertion elements R1Bm and R2Bm transpose in a sequence specific manner to locations outside 28S genes. Nucl Acids Res 16, 10561-10573.

Zagrobelny M, Jeffares DC, Arctander P., 2004. Differences in non-LTR retrotransposons within *C. elegans* and *C. briggsae* genomes. Gene 330, 61-66.

Zeyl C, Bell G., 1995. Symbiotic DNA in the eukaryotic genome. Trends Ecol Evol 11, 10-15.

LEGENDS TO FIGURES

Fig. 1. Overview of structure and sequence coverage for the *Syrinx* **(A)** and *Daphne* **(B,C)** families. The endonuclease (EN) and reverse transcriptase (RT) domains in the coding sequence are shaded, and conserved motifs within each domain are indicated by corresponding numbers (Roman for EN, Arabic for RT). $(A)_n$, $(T)_n$ and $(TTA)_n$ denote polymorphic regions in the 3' UTR (n=4-25). The UFW strategy and the placement of individual clones along the consensus are shown below each element. Different flanking sequences are shown by dashed or dotted lines, with similar flanking sequences having the same dotting pattern. Thick lines with arrows in the flanking sequences denote fragments of *mariner*-like and *R2*-like elements in the opposite transcriptional orientation. **(C)** Detailed structure of four different 5' truncated head-to-head inversion junctions in six *Daphne* copies. Correspondence between direct/inverted segments and the consensus *Daphne* sequence is indicated by vertical lines. Arrows indicate the direction of the sense strand; squares denote 4-6 bp microhomology overlaps at the inversion junctions (also shown on the right), which may be assigned to either direct or inverted copy (arrows). The thick dashed line denotes a divergent *Daphne'* copy in L2a34. A thin dashed line represents unrelated sequence at the inversion junction in three nearly-identical copies. A space between two inverted copies is introduced for alignment purposes, and does not imply any additional sequence between two black squares.

Fig. 2. **(A)** Alignment of consensus amino acid sequences coded by *Syrinx* and *Daphne* in the BoxShade format, together with their top database matches: *Juan*-like non-LTR retrotransposon from the African malaria mosquito, *Anopheles gambiae* (gi:57920422;57911403), and *Urca* non-LTR retrotransposon from the sea urchin, *Strongylocentrotus purpuratus* (this study). Highly conserved residues are denoted by asterisks; identical residues are shaded in black, and chemically similar residues in gray. Shown are the most conserved motifs of the AP endonuclease domain (III-IX) (Martin *et al*., 1996; Feng *et al*., 1996) and the RT domain (core motifs RT0-RT7 and the thumb domain motifs RT8-RT9, designated in square brackets) (Malik *et al.*, 1999), as well as the C-terminal conserved region (CTCR) identified by Kajikawa *et al.* (2005) in the eel *UnaL2* and related elements. Also shown are the results of protein secondary structure prediction on the JPRED server for the first 400 amino acids of *Syrinx* (top) and *Daphne* (bottom) encompassing the EN domain. **(B)** Alignment of the N-terminal RT region (motifs RT0-RT4) of *Syrinx* and *Daphne* with the corresponding divergent lineages *Syrinx'* (clone L1a42) and *Daphne'* (clones L2a34 and L2a21) obtained in the course of 5' UFW.

Fig. 3. Divergence and phylogenetic relationships of individual *Syrinx* (L1 series, left) and *Daphne* (L2 series, right) nucleotide sequences, obtained in the course of 5' UFW (top) or 3' UFW (bottom). Four Bayesian phylograms are drawn on the same scale. Clade credibility values exceeding 50% are indicated. Copies that may originate from intact elements are underlined; asterisks denote in-frame stop codons; deletions and insertions within ORFs are shown by upward- and downward-pointed triangles, respectively, with the following number indicating their length; arrows pointing left and right indicate copies that are 5' or 3' truncated, respectively. L1asa45, 34, 41 are the inverted parts of L1a45, 34, 41, which could not be combined because there is no guarantee that they originate from the same copy. Square brackets indicate copies that share similar flanking sequences (not included in the alignments used for analysis).

Fig. 4. Phylogenetic placement of *Syrinx* and *Daphne* among non-LTR retrotransposons. AP-EN and RT domains of elements from the reference dataset (Eickbush and Malik, 2002) were used in Bayesian analysis. Additional elements included into the analysis were *TAHRE* from *D. melanogaster* (gi:48596584), *Juan*-like from *A. gambiae* (gi:55235907), *CR1*-like from *Branchiostoma floridae* (gi:17529698), *Sake* from *B. mori* (this study), and *Urca* from *S. purpuratus* (this study). Clade credibility values exceeding 50% are indicated; square brackets denote the names of the known clades.

Figure 1

Click here to download high resolution image

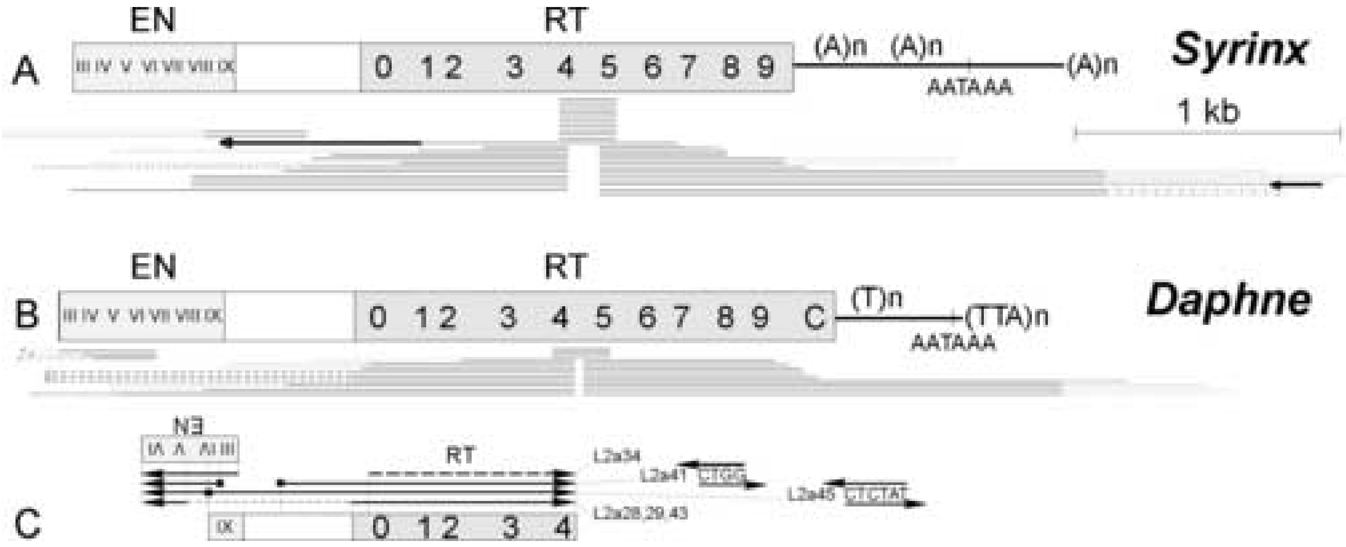

Figure 2

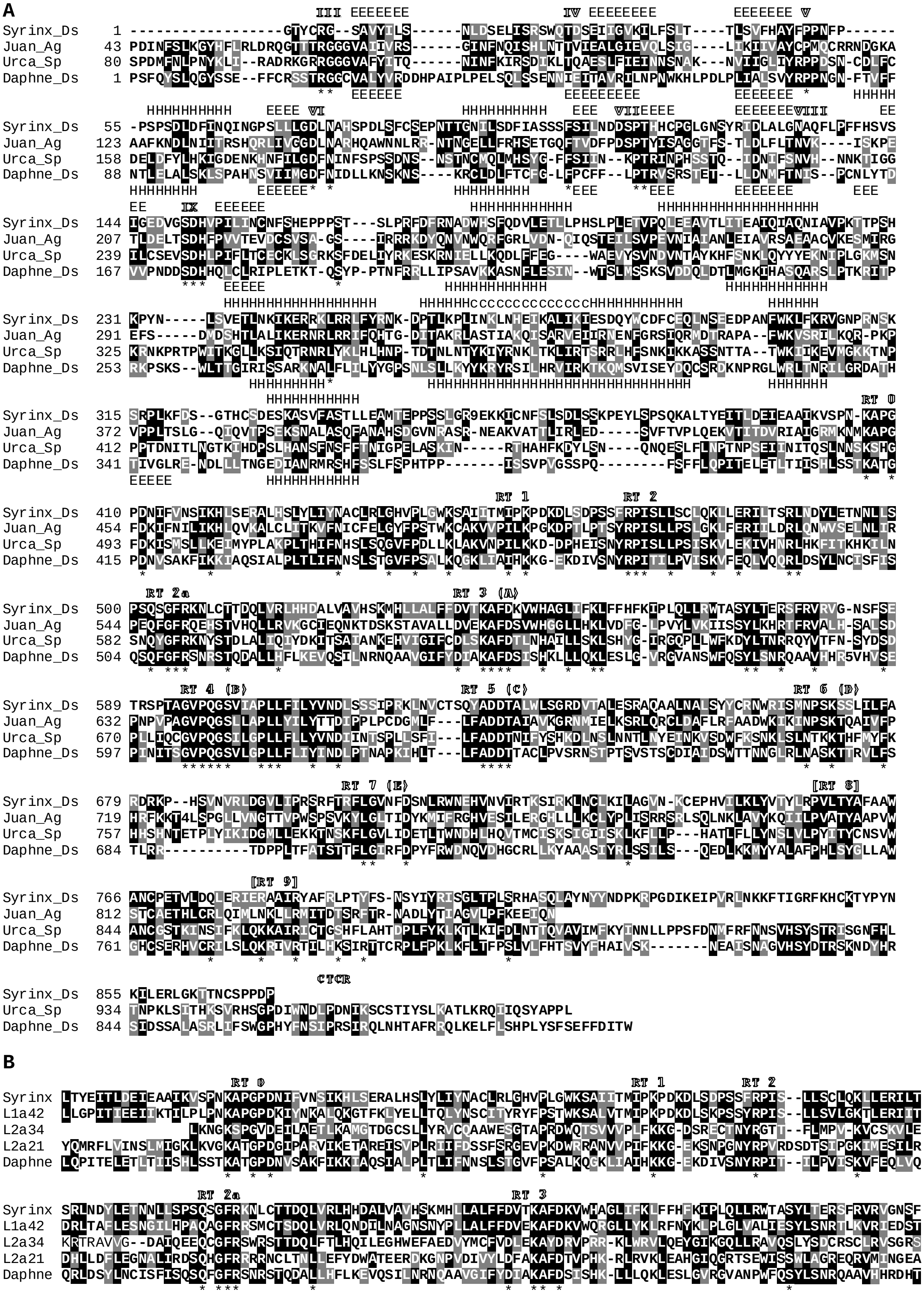
A
Syrinx_Ds
Juan_Ag
Urca_Sp
Daphne_Ds
III
IV
V
VI
VII
VIII
IX
RT 0
RT 1
RT 2
RT 2a
RT 3 (A)
RT 4 (B)
RT 5 (C)
RT 6 (D)
RT 7 (E)
[RT 8]
[RT 9]
CTCR
B
Syrinx
L1a42
L2a34
L2a21
Daphne
RT 0
RT 1
RT 2
RT 2a
RT 3

Figure 3

Click here to download high resolution image

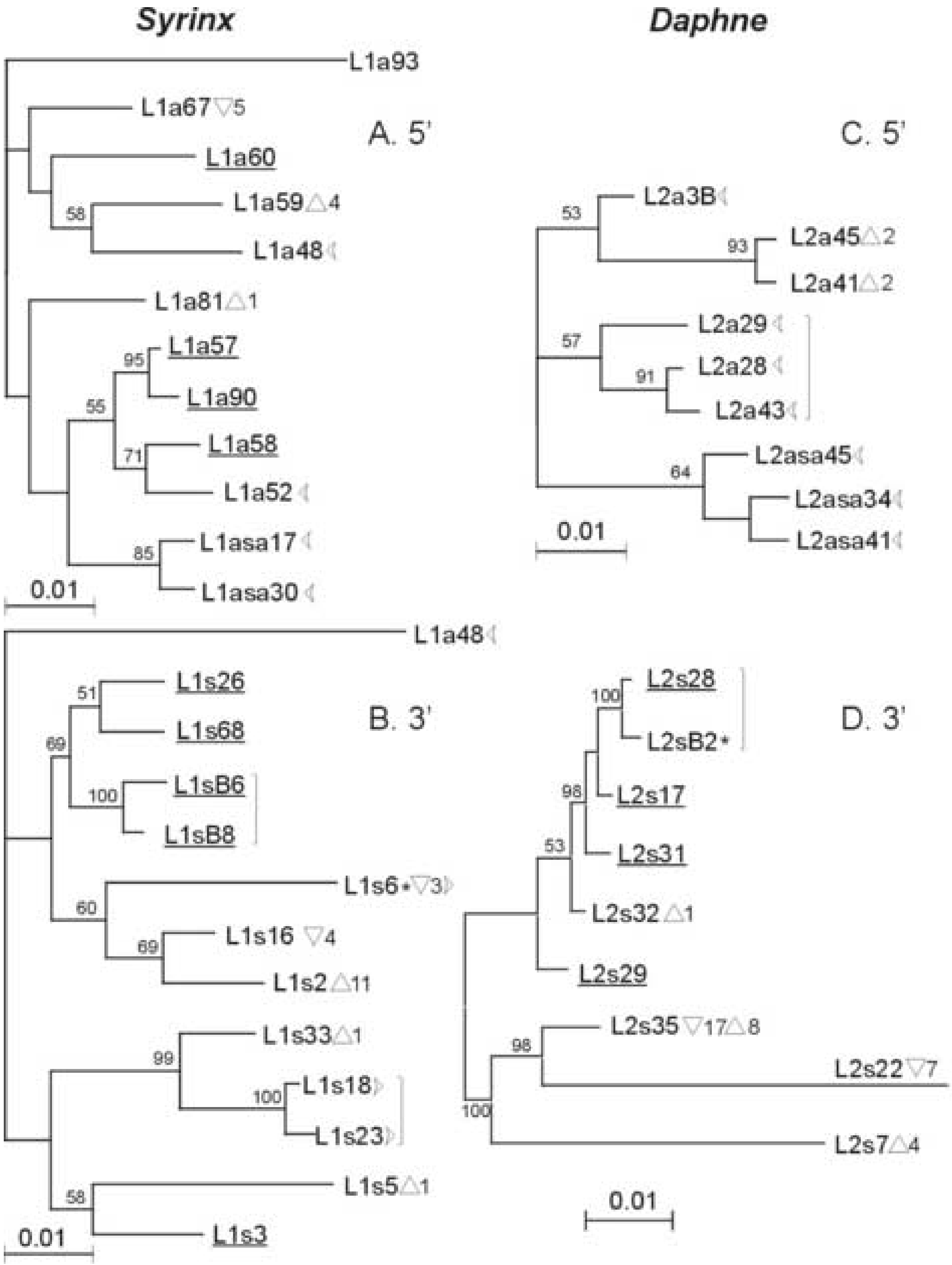

Figure 4

Click here to download high resolution image

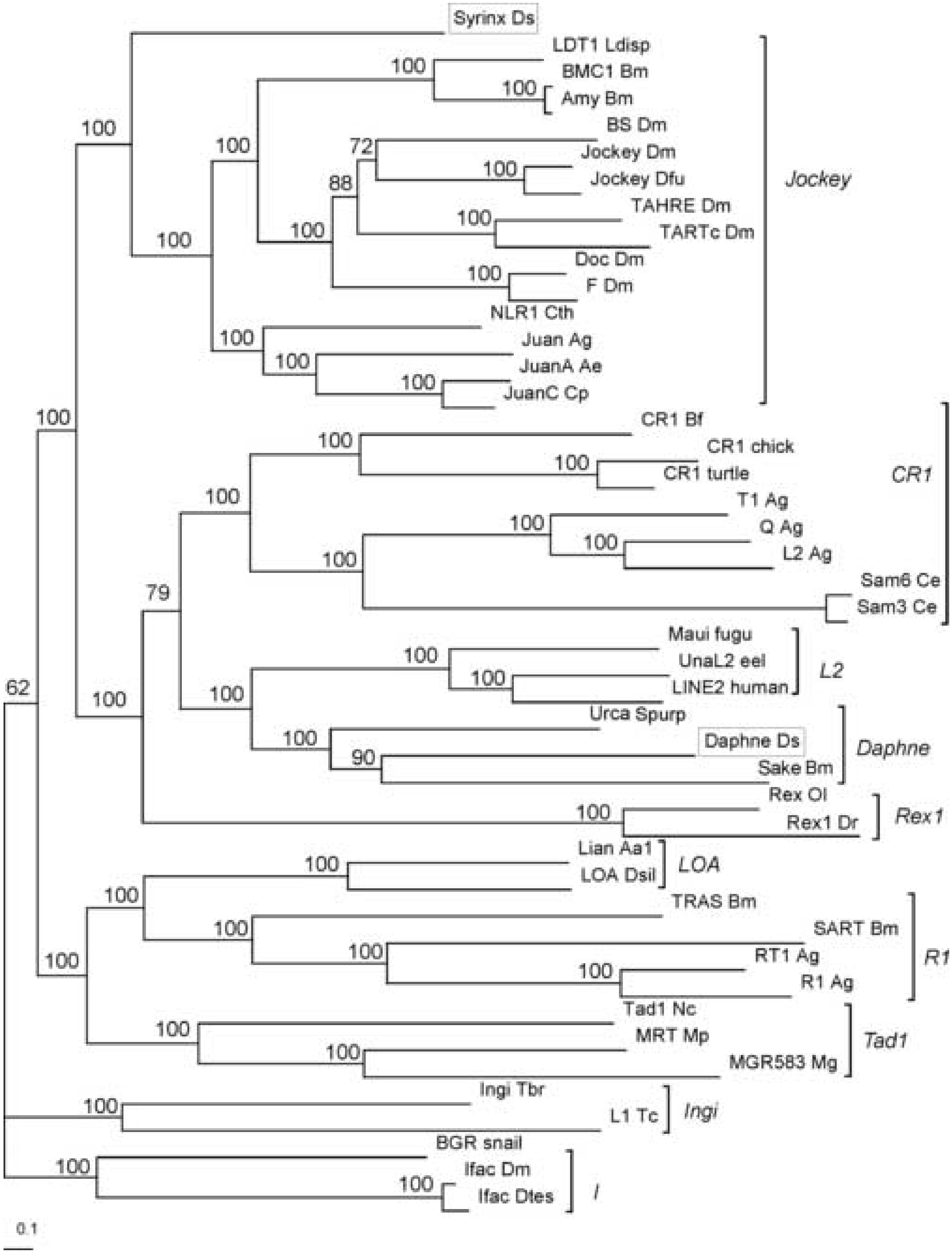

Table 1. Primers used for UFW and sequencing of UFW products. L1 = *Syrinx*, L2 = *Daphne*. Highly degenerate primers RT-3 – RT-7 were used for nested PCR of regions between the conserved core RT motifs (Arkhipova and Meselson, 2000); primers s1-s4 and as1-as4 were used for UFW in the sense and antisense direction, respectively, as described in Myrick and Gelbart (2002); primers labeled "internal" were used to close the internal sequence gaps.

Primer	**Orientation**	**Sequence**
RT-3	sense	GCTCTAGAYITIINNNVNGSNTWY
RT-7	antisense	GGAATTCAINIBIDNNCCNARRWM
RT-4	sense	GCTCTAGAINGGNIBNCSNCARGG
RT-5	antisense	GGAATTCRNNRNRTCRTCNGCRWA
DsL1-s1	3’ (sense)	GGCTCAGTAATAGCACCCCT
DsL1-s3	3’ (sense)	CCCCTTCTTTTCATTCTCTACG
DsL1-s2	3’ (sense)	CTCTACGTWAACGACCTCTCN_9A
DsL1-s4	3’ (sense)	AGCTCKATCCCTCGTAAACT
DsL1-as1	5’ (antisense)	AGTTTACGAGGGATMGAGCT
DsL1-as3	5’ (antisense)	GAGAGGTCGTTWACGTAGAG
DsL1-as2	5’ (antisense)	CGTAGAGAATGAAAAGAAGGGN_9A
DsL2-as4	5’ (antisense)	AGGGGTGCTATTACTGAGCC
DsL2-s1	3’ (sense)	GGCTCGGTCCTTGGTCCA
DsL2-s3	3’ (sense)	GGTCCACTCCTCTTTCTAAT
DsL2-s2	3’ (sense)	ATATTAACGACCTCCCCACCN_9A
DsL2-s4	3’ (sense)	CCAAGATTCACCTCACGCTC
DsL2-as1	5’ (antisense)	GCGAAGAGCGTGAGGTGAA
DsL2-as3	5’ (antisense)	GGTGAATCTTGSGAGCGTT
DsL2-as2	5’ (antisense)	GGTGGGGAGGTCGTTAAYATN_9A
DsL2-as4	5’ (antisense)	AGGAGTGGMCCAAGGACC
DsL1a	5’ (antisense) internal	GATGCAAAAACGGAGGCTTG
DsL2a	5’ (antisense) internal	GAGATCAATGTTGAAGTCACC
DsLI3	3’ (sense) internal	TAGATATTGGAGGTGTGTGA
DsL1SI1	3’ (sense) internal	CAGCCCCAAATCTACTAACC
DsL2-32	5’ (antisense) internal	GGTCCGATCGTTGGAACAATGC
DsL1as4a2	5’ (antisense) internal	GATGCAAAAACGGAGGCTTTG

Table 2. Features of amplified *Syrinx* and *Daphne* sequences.

Family	Consensus ORF	Similarity RT/EN (% aa)	No. of UFW products 3'/5'	No. of truncated products 3'/5'	No. of intact sequences 3'/5'	No. of frameshifts 3'/5'
Syrinx	871 aa	29 /19 (S/D) 51 (S/S')	12/10	5/3	6/3	1/3
Daphne	902 aa	29/19 (S/D) 35 (D/D')	9/6	2/3	3/1	3/2

aa = aminoacids. S = *Syrinx*. S' = *Syrinx*-related subfamily. D = *Daphne*. D' = *Daphne*-related subfamily. 3' = 3' UFW products. 5' = 5' UFW products.

Table 3. Ratios of non-synonymous (Ka) to synonymous (Ks) substitutions for *Syrinx* (L1) and *Daphne* (L2) sequences (3' UFW and 5' UFW products for *Syrinx*, 3' UFW products for *Daphne*).

A (L1s)	L1s18	L1s33	L1s23	L1s26	L1sB6	L1sB8	L1s16	L1s5
L1s33	0,24							
L1s23	*	0,24						
L1s26	0,19	0,11	0,19					
L1sB6	0,16	0,09	0,16	0,34				
L1sB8	0,23	0,15	0,23	0,77	0,57			
L1s16	0,26	0,20	0,26	0,36	0,28	0,37		
L1s5	0,27	0,14	0,27	0,21	0,31	0,42	0,17	
L1s3	0,24	0,13	0,24	0,22	0,09	0,17	0,20	0,23

B (L1a)	L1a58	L1a60	L1a90	**C (L2s)**	L2s29	L2sB2	L2s28	L2s17
L1a57	0,26	0,09	0	L2sB2	0,89			
L1a58		0,22	0,26	L2s28	1,07	*		
L1a60			0,09	L2s17	1,39	0,96	1,50	
L1a90				L2s31	0,69	0,31	0,49	0,53

*Ratio could not be calculated, because Ks=0.
Average is 0,25 ± 0,13 for *Syrinx* and 0,87 ± 0,41 for *Daphne*.