Rotifer rDNA-Specific R9 Retrotransposable Elements Generate an Exceptionally Long Target Site Duplication upon Insertion

Eugene A. Gladyshev[1,2] and Irina R. Arkhipova[1*]

[1]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA; [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: iarkhipova@mbl.edu.

Dr. Irina R. Arkhipova

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution

Marine Biological Laboratory

Woods Hole, MA 02543, USA

Tel. (508) 289-7120

Fax: (508) 457-4727

Abbreviations: non-LTR, non-long terminal repeat; rDNA loci, loci encoding the tandem array of rRNA genes; RT, reverse transcriptase; TE, transposable element; TPRT, target primed reverse transcription; TSD, target site duplication.

Keywords: bdelloid rotifers; ribosomal insertions; 28S rDNA; non-LTR retrotransposons.

Abstract

Ribosomal DNA genes in many eukaryotes contain insertions of non-LTR retrotransposable elements belonging to the R2 clade. These elements persist in the host genomes by inserting site-specifically into multicopy target sites, thereby avoiding random disruption of single-copy host genes. Here we describe R9 retrotransposons from the R2 clade in the 28S RNA genes of bdelloid rotifers, small freshwater invertebrate animals best known for their long-term asexuality and for their ability to survive repeated cycles of desiccation and rehydration. While the structural organization of R9 elements is highly similar to that of other members of the R2 clade, they are characterized by two distinct features: site-specific insertion into a previously unreported target sequence within the 28S gene, and an unusually long target site duplication of 126 bp. We discuss the implications of these findings in the context of bdelloid genome organization and the mechanisms of target-primed reverse transcription.

1. Introduction

Retrotransposons are ubiquitous genetic parasites inhabiting genomic DNA of most eukaryotes. They insert either at random genomic locations or at specific target sites, depending on the degree of sequence specificity of an element-encoded endonuclease. According to the presence or absence of long terminal

repeats (LTRs), they are subdivided into LTR and non-LTR retrotransposons. Sequence-specific retrotransposons belonging to the R2 clade are one of the most evolutionarily successful groups of non-LTR retrotransposons with regard to long-term persistence in the host genomes (reviewed in Eickbush, 2002). This is in part due to their high degree of specialization for insertion into highly conserved multicopy targets, such as DNA coding for the large subunit 28S ribosomal RNA (rDNA). Initially discovered in *Drosophila*, R2-like elements were later found in the majority of other arthropods, as well as in certain chordates, echinoderms, flatworms, and cnidarians. Although they are not found in mammalian rDNA, such lack is thought to result from vertical losses, as horizontal transfer of these elements has never been documented, and their origin arguably dates back to the early stages of metazoan evolution predating the split between protostomes and deuterostomes (Eickbush, 2002; Kojima et al., 2006). R2 elements represent a model system for studying target-primed reverse transcription (TPRT), whereby the cDNA copy of a retrotransposon is synthesized directly at the target site, primed by the 3'OH at the nick in chromosomal DNA which is introduced by the endonuclease moiety of the retrotransposon-encoded protein (see Christensen et al., 2006 for the latest TPRT model).

Rotifers of the Class Bdelloidea are microscopic freshwater invertebrates thriving in ephemerally aquatic habitats. They are best known for their ability to reproduce entirely asexually and to survive repeated cycles of desiccation and

rehydration at any stage during their life cycle. As part of the ongoing investigation of telomere structure in bdelloid rotifers (Gladyshev and Arkhipova, 2007; Gladyshev et al., 2007, 2008), in the course of screening of the genomic library from the bdelloid *Adineta vaga*, we obtained a number of fosmid clones containing, in addition to rDNA, several insertions of an R2-like element, which we named R9Av. While its overall structure corresponds to that of the other known R2 elements, its insertions are characterized by unusually long target site duplications (TSDs), and insertion site specificity differs from that of other rDNA insertion elements, which precludes it from being named R2 and triggers sequential numbering as a successor to the R8 element (Kojima et al., 2006). Here we report on the structure and phylogenetic placement of R9Av and discuss the implications of long TSDs for the process of target-primed reverse transcription (TPRT).

2. Materials and Methods

Screening of the *A. vaga* genomic fosmid library (Hur, 2006) and isolation of fosmid clones was done as described in Gladyshev and Arkhipova (2007). The R9 screening probe was obtained by PCR using the following primers: R2_F1, ATGTTCGACAATACGGCTCCT; and R2_R1, ATCTCCAGTGGTGTCAAGCAA. Each of the 32 hybridizing fosmids was sequenced with a set of seven custom primers spanning the entire length of R9 and the adjacent flanking DNA: seq1, TGTGATGACGAGAGTATCG; seq2, CGAGTTGTCTTATCAATATAGC; seq3,

ATTCCTTGAACACCTAGCC; seq4, ACAATGGAATCTACAGATTCACGG; seq5, CCACATTCTCAAAAGCTATCACA; seq6, GCTCTCGAATCGGCAAATTCA; seq7, GGCGAAATTAAGACAATGACAA. Sequences obtained in this study were deposited in GenBank under accession numbers GQ398057-GQ398061. Southern blot of *A. vaga* genomic DNA digested with *Xho*I was hybridized with the above R9 probe and with the ^{32}P-labeled 259-bp 28S rDNA PCR fragment obtained with primers AGCGAATGTGAGTGCCAAGT and ACTAGTCGATTCGGCAGGTG; this rDNA probe was also used to screen the genomic library. Phylogenetic analysis was done with MEGA4 (neighbor-joining or minimum evolution; p-distance; pairwise deletion; 1000 bootstrap replications) (Tamura et al. 2007). Amino acid sequence alignments are provided as Supplementary data.

3. Results

3.1. Identification of retrotransposon insertions in the A. vaga 28S rRNA genes

A nearly complete copy of the R9Av element was identified in a fosmid clone from the *A. vaga* genomic library, which was initially selected by hybridization to the telomeric repeat probe. One of the fosmid-end sequencing reads contained an ORF with homology to the coding sequence of known R2 elements, while the other end was 100% identical to the known 28S rDNA sequence from *A. vaga* (Garcia-Varela and Nadler, 2006). The original R9Av copy, however, was C-terminally truncated by cloning. In order to obtain the

sequence of full-length copies and to determine their insertion site specificity, we synthesized a 792-bp internal R9Av probe by PCR (Fig. 1; see Methods) and used it to screen the *A. vaga* genomic library. After screening *ca*. 4 *A. vaga* genome equivalents, we obtained 32 fosmid clones, each being 35-40 kb in length. Since rDNA-containing clones are poorly assembled by the shotgun approach because of their intrinsically repetitive nature, we chose to determine the complete sequence of R9 elements on each fosmid by primer walking. Assembly of all sequenced copies revealed that they fall into four groups of sequences indistinguishable by nucleotide polymorphisms. Each of these groups represented an insertion into the same site of 28S rDNA which, however, differed from the canonical R2 insertion site (see below). We therefore named the element R9Av, in agreement with the previous practice for naming rDNA insertion elements, whereby consecutive numbering is used for elements in previously undescribed specific insertion sites in 28S or 18S rDNA (Burke et al., 1995, 2003; Kojima et al., 2006).

3.2. Copy number and site occupancy

A rough estimate of the R9 copy number and the percentage of rDNA units occupied by R9 insertions can be made from comparing the number of R2-containing fosmids with the total number of rDNA-positive fosmids obtained from screening of the same genomic library membranes. Screening of three different membranes (*ca*. 12 genome equivalents) with the R9/28S probes yielded 32/221, 27/224, and 21/197 hybridizing fosmids, respectively. Since the variability of the

intergenic spacer region between rDNA units appears to be minor, the length of the rDNA unit could be estimated as 10.4 kb from the assembled rDNA contig containing 40 out of 64 end-reads from the 32 sequenced R9-containing fosmids, so that up to 4 unoccupied rDNA units could be located on a typical 40-kb fosmid. Thus, while the upper estimate of rDNA units per genome is about 200 copies, less than 3% of these units are occupied by R9 insertions (assuming that the latter are present on each fosmid clone as single-copy insertions).

We sought to validate these estimates by restriction analysis, Southern blotting, and hybridization with the rDNA/R9 probes. Digestion of *A. vaga* genomic DNA with *Xho*I, which does not have recognition sites within R9, but which cuts about 0.4 kb downstream and 2.9 kb upstream from the R9 insertion site, and subsequent hybridization with the 28S rDNA probe amplified from the region downstream of the R9 insertion site (see Fig. 1; Methods) yielded, in addition to the high-intensity 3.3-kb 28S rDNA-derived band, a longer 7.3-kb band resulting from insertion of the 4-kb R9 element (Fig. 2). The intensity of the R9-derived band comprises only a small percentage of the intensity of the major rDNA band, in agreement with the above estimates based on genomic library screening.

Finally, while an R9 copy number estimate based on sequence analysis would not discriminate between 100% identical copies, it does provide the absolute minimum, which equals 4. Furthermore, defective variants 2 and 3, which are likely to be unique, are present on a smaller number of fosmids (4-5 each), in agreement with genome coverage, while the intact copy 4 is present on

10 fosmids and may exist as two or at most three identical copies. Screening of the same membrane with the single-copy *hsp82* probes from two divergent alleles previously yielded either 4 or 5 fosmids per single copy of this gene, while two identical *hsp82* alleles were present on 13 fosmids (bdelloids are degenerate tetraploids) (Hur, 2006). Taken together, all of the above estimates make it unlikely that the number of R9 copies in the *A. vaga* genome is below 4 or exceeds 6.

3.3. Structural organization of R9Av

The structure of R9Av is very similar to the canonical structure of R2 elements. It harbors a single open reading frame coding for a 1102-aa polypeptide, which contains all of the domains expected for a typical R2 element (Figs. 1, 3). The central core is occupied by the reverse transcriptase (RT) domain (cd01650; pfam00078). At the C-terminus there is a restriction enzyme-like (REL) endonuclease (EN) domain, which confers sequence specificity to R2 insertion (Yang et al., 1999). At the N-terminus there are three Zn finger domains - two CCHH and one CCHC (Fig. 3A) - at least some of which likely are responsible for binding of the enzyme to the target DNA (Christensen et al., 2005). There is no poly(A) tail at the 3' end, and the poly(A) signal is located only 6 bp upstream from the end of the 3' UTR, which is 303 nt in length.

Not all copies appear to contain a fully functional R9 element. Despite the apparent lack of 5' truncated copies, two out of four sequence variants exhibit the same frameshift between the N-terminal domain and the core RT, which occurs

within a frameshift-prone T8 stretch and is expected to split the single ORF into the nucleic acid-binding and enzymatic moieties. We do not know, however, if the N-terminally truncated RT-EN polypeptide would be able to exhibit enzymatic activity. Most likely, these two copies would be inactive due to the lack of the N-terminal domain in the holoenzyme, which is thought to be important for nucleic acid binding and the TPRT reaction (Chistensen et al., 2005, 2006).

While two out of four variants are apparently inactive, the other two variants may be capable of reverse transcription and transposition. The intact copy 1 is 99% identical to frameshift-containing copies 2 and 3, differing by only 13 nucleotides. Copy 4, while intact, differs from all the others by 6%, thus representing a distinct subfamily. An additional apparent deficiency common to all copies is that the first ATG codon within the element is positioned in the middle of the first CCHH Zn finger motif (Figs. 1, 3A). This is not necessarily a defect, since many R2 elements from other subclades have either two or one Zn finger motifs at the N terminus, and several known R2 elements without proper ATG codons have been reported and proposed to utilize an internal ribosome entry site (Burke et al., 1999; George and Eickbush, 1999; Kierzek et al., 2009). However, it does raise some questions about the capacity of R9Av for efficient transposition, as copy number estimates (see 3.2) demonstrate that only a minor fraction of rDNA units is occupied by R9Av insertions.

3.4. Flanking 28S rDNA target sequences undergo an exceptionally long target site duplication

While there is nothing seemingly unusual about the R9Av structure, it does exhibit a striking feature when its rDNA flanks are examined. Canonical R2 elements always insert into the same, highly conserved, site in the 28S rDNA target, which is cleaved by the element-encoded EN activity (reviewed in Eickbush 2002). R9 represents a notable exception in this respect, as it inserts into a previously unknown site in 28S rDNA, 1436 bp upstream from the usual R2 insertion site (Fig. 1).

When the second strand of the 28S target undergoes cleavage by the element-encoded endonuclease during TPRT, it typically occurs in the vicinity of the first-strand cleavage site, and, depending on the location of the second cleavage site, insertion usually generates either a small deletion (*e.g*. 2 bp for R2Bm) if the second-strand cleavage occurs upstream from the first-strand nick, or a TSD up to 20 bp in length if the cleavage occurs downstream (R4, R6, R7, R8) (see Fig. 1 in Kojima et al., 2006). If the first and second strand cleavages occur at the same site, there is no TSD. In the case of R9Av, however, the observed size for the target site duplication in each of the four sequenced variants is as large as 126 bp (Fig. 4). Since it is thought that the subunit responsible for the second strand cleavage acts upon completion of the first strand synthesis by the first subunit in the vicinity of the insertion site (Christensen et al., 2006), such an unusual distance requirement poses interesting questions regarding the mechanistic aspects of the second strand synthesis (see Discussion).

Although none of the insertions were located outside of the rDNA cluster, the 5'-flanking sequence in copy 3 differs from all the others: while the sites for the top and bottom strand cleavage in the 28S rDNA remain the same, the upstream 28S rDNA segment is 126 bp long and is interrupted by another sequence distantly resembling the 3' UTR of R9 in the region immediately adjacent to the target site (Fig. 4A, copy marked U). Such tendency of R2-like elements to form clusters in rDNA, which are separated by TSD, was previously noted by Stage and Eickbush (2009) and can be explained by the fact that generation of a TSD upon insertion effectively restores the intact target site, which can be used for subsequent insertions not subject to further negative selection, since this particular 28S gene is already inactivated. Finally, one of the R9 copies occupying the R9 target site co-exists with a 789-bp apparently non-autonomous insertion element occupying the canonical R2 target site 1436 bp downstream from R9 (Fig. 1). The insertion is relatively short, causes a 1-bp TSD, has no terminal repeats of any kind, no coding potential, and does not exhibit any apparent homology to R2 or R9.

3.5. R9Av belongs to the R2-A "clade"

Based on the amino acid sequence phylogeny and on the number of N-terminal Zn fingers, which can vary between one and three, R2 retrotransposons were subdivided into four clades, R2-A, -B, -C and -D, by Kojima et al. (2005, 2006). Phylogenetic analysis (Fig. 5) confirms this division and places R2Av into the R2-A clade, together with other known members of this clade, such as R2Tc

from *Triops longicaudatus* (with which it groups with 99% support), R2Ci and R2Cs from *Ciona intestinalis* and *C. savignyi*, respectively, R2Lp from *Limulus polyphemus*, R2Dr from *Danio rerio*, and R8Hm from *Hydra magnipapillata*. This placement is confirmed by the presence of three N-terminal Zn finger motifs, as is typical for other member of this clade (Fig. 1,3). The clade also contains the newly identified R2Tg element from the zebrafinch, *Taenopygia guttata*, which, however, appears to harbor only two Zn fingers. Thus, the entire R2 clade is now shown to contain representatives from at least five different animal phyla.

R9Av exhibits much stronger similarity to R2Tc from the tadpole shrimp *Triops longicaudatus*, an arthropod crustacean, than to other R2's (42% amino acid identity, as opposed to 25% or less for an equivalent RT-EN R2 fragment from any other species). This, however, still does not constitute a violation of the "no-horizontal-transfer" rule for R2 elements (Burke et al., 1998; Malik et al., 1999), as certain other nuclear genes from *T. longicaudatus* exhibit comparable levels of divergence to the corresponding rotifer genes (data not shown). It does, however, split the "R2-A1 subclade" of Kojima & Fujiwara (2005), claimed to contain only members consistent with phylogenetic distribution of host species, into two more "sub-subclades" (Ascidian+Cnidarian and Rotifer+Crustacean), presumably originating in the Precambrian.

4. Discussion

In this study, we identified a new rDNA-specific retrotransposable element with previously undescribed insertion site specificity. This element was named

R9, succeeding the non-LTR retrotransposon R8Hm from *Hydra magnipapillata* (Kojima et al. 2006). R8Hm was noted for the change in insertion specificity from the canonical site in the 28S rDNA gene to a new site in the 18S rDNA gene. A similar change in sequence specificity apparently occurred in R9Av, since its phylogenetic neighbors, the R2-A elements from *Ciona* and *Triops* (Fig. 5), do insert into the standard R2 target. There is little similarity between the R2 and R9 targets, except for the five most proximal nucleotides, 5' G|TAGC 3'. It should be noted that the 28S rDNA sequence between the sites of the first- and second-strand cleavage (Fig. 4B) does not exhibit high levels of evolutionary conservation characteristic of the standard R2 insertion site (Kojima and Fujiwara, 2005), perhaps accounting for the fact that this site has not previously been reported to be occupied by any rDNA insertions. The sequence specificity, thought to be conferred by the element-encoded endonuclease and the associated DNA-binding domains, does not correlate with phylogenetic placement of R9, which is a *bona fide* member of the R2 clade (Malik et al., 1999) and belongs to its R2-A subclade ("clade", Kojima and Fujiwara, 2005), as defined by phylogenetic analysis and the number of Zn finger motifs in the N-terminal domain of the element-encoded protein. We should note that the now widespread use of the term "clade" with regard to retrotransposon nomenclature causes an increasing degree of confusion as deeper and deeper branching sub-clades are uncovered.

The site occupancy by R9 insertions in *A. vaga* appears relatively low, which does not indicate massive disruption of 28S transcription units by

retrotransposon insertion, as occurs in Drosophila (reviewed in Eickbush, 2002). In addition, the already-disrupted rDNA units do tend to accumulate additional insertions, as was noted before (Stage and Eickbush, 2009), apparently because such insertions would not cause any additional damage to the host. Indeed, 10 out of 32 sequenced fosmids yielded double reads with a subset of sequencing primers, indicating that some of the primers were landing at more than one sequence. Two such adjacent insertions were directly revealed by sequence analysis, one located 1 TSD upstream of R9Av-3 with a similar 3' UTR, and another downstream from R9Av-2 at the canonical R2 insertion site, with no apparent similarity to R2.

The most unusual characteristic of R9Av uncovered in the present study is an exceptional length of the target site duplication it generates upon insertion. Each of the four sequenced variants is surrounded by the same-length, 126-bp TSD, indicating a very high degree of precision during second-strand cleavage at distances comparable to nucleosomal. An alternative explanation is that all copies originated by gene conversion, which is highly unlikely because copy 4, which is 6% divergent form the others, exhibits the same TSD length. Our finding calls for careful evaluation of distance requirements during the design of PCR screens for rDNA insertions at specific sites, as such insertions are likely to be missed if both primers fall onto a long TSD.

Naturally occurring insertions of R2 and similar elements are associated with TSDs not exceeding 30 bp in length (Eickbush, 2002; Kojima and Fujiwara, 2005; Kojima et al., 2006; Stage and Eickbush, 2009). R2Dm elements can also

cause variable deletions of the upstream rDNA ranging between 2 and 82 bp (Stage and Eickbush, 2009). Mammalian LINE-1 elements, which transpose by a similar mechanism, generate variable-sized TSDs, with lengths centering around 9 and 15 bp and never exceeding 60 bp in more than 16,000 analyzed L1 insertions in the human genome (Szak et al., 2002). In cultured cells, TSDs ranging from 50 to 323 bp have been reported, but large TSDs vary in length considerably and only are observed in a minor fraction of insertions, and were thought to be disfavored by selection in nature or to represent a peculiarity of retrotransposition assays conducted in cultured cells. (Moran et al., 1996; Symer et al., 2002; Gilbert et al., 2002, 2005). Occasional long TSDs of variable size have even been reported in human endogenous retroviruses (Mamedov et al., 2004), and it was suggested that such rare events may be associated with resection and DSB repair around the insertion site, involving other host systems in addition to retrotransposon-encoded enzymes. For R9, however, long TSDs of fixed length appear to be the rule rather than exception, indicating that it is the intrinsic property of the element which causes the permanent increase in TSD length. Since the R2 elements usually serve as a model system for studying TPRT mechanisms in general, and it is widely believed that these mechanisms are shared by all non-LTR retrotransposons, it is of interest to discuss our finding of exceptionally long TSDs in the context of the current TPRT model (Christensen et al., 2006).

The variability in TSD length for many non-LTR retrotransposons is thought to arise from imprecise positioning of the second-strand cleavage site

with respect to the first-strand cleavage site. Conversely, the lack of such variability, as observed in R9 insertions, clearly indicates a high degree of precision in second-strand cleavage. The current model of R2 retrotransposition postulates coupling between completion of the first-strand synthesis by the upstream subunit of the R2 enzyme, resulting in the release of the 5' end of the RNA from the downstream subunit. Release of the 5' RNA initiates top strand cleavage by the downstream subunit in the vicinity and initiation of the second-strand synthesis (Christensen et al., 2006). If this model is applicable to all non-LTR retrotransposons, the second-strand cleavage site of the R9 target needs to be brought in proximity to the first-strand cleavage site from a relatively large distance, which would likely involve sequence-specific recognition for both top and bottom strands, rather than the bottom strand only. R9 has an unusually large linker between the distal and proximal Zn fingers (Fig. 3A), which might contribute to recognition at a distance, perhaps in the context of target DNA wrapped onto a nucleosome. However, since a typical nucleosome is wrapped by 146 bp of DNA, which is 20 bp longer than the distance between the top and the bottom strand cuts, it is unlikely that proximity of the two cleavage sites is mediated by a canonical nucleosome. It is also unlikely that microhomologies between the 5’ end of R9 and the top strand of rDNA target are utilized to prime the R9 second strand synthesis, as hypothesized for R2Dm by Stage and Eickbush (2009), since no apparent microhomologies surrounding the insertion site could be found. It might be of interest to study biochemical properties of the

R9-encoded enzyme in order to gain insight into molecular peculiarities of the TPRT reaction in these elements.

Overall, the ability of bdelloids to survive repeated desiccation and their long-term asexuality had apparently created a unique genomic environment in which some TEs thrive and others do not. Localization of the *A. vaga* rDNA at the chromosome termini agrees well with our previous findings that bdelloid TEs are mostly concentrated at telomeres (Arkhipova and Meselson, 2005; Gladyshev and Arkhipova, 2008; Gladyshev et al., 2007, 2008), and the convenience of having a specific multicopy target permits R9 elements to survive without inserting into random genomic locations. The overall low copy number is also typical of other bdelloid retrotransposons and never permitted their detection by PCR screens designed for high copy number elements (Arkhipova and Meselson, 2000). While there is a distinct possibility that bdelloids engage in a desiccation-induced parasexual process by means of horizontal exchange, TE insertions may be kept at low numbers in the absence of meiotic recombination by synergistic selection against dispersed repeats in central gene-rich regions of bdelloid genomes imposed by deleterious translocations and deletions occurring during DNA breakage and repair associated with repeated cycles of desiccation and rehydration (Gladyshev and Meselson, 2008; Gladyshev et al., 2008). Bdelloid telomeres, which accumulate TE insertions (including those specifically adapted to inserting into telomeric regions) and genes of foreign origin in large numbers, will undoubtedly continue to serve as a rich source of novel TEs with unusual properties.

Acknowledgments

We wish to thank J. Hur for providing the *A. vaga* genomic fosmid library. This work was supported by the U.S. National Science Foundation grant MCB-0821956 to I.A.

References

Arkhipova I.R, Meselson M., 2005. Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci USA. 102, 1781-11786.

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

Burke W.D., Müller .F, Eickbush T.H., 1995. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. Nucleic Acids Res. 23, 4628-4634.

Burke W.D., Singh D., Eickbush T.H., 2003. R5 retrotransposons insert into a family of infrequently transcribed 28S rRNA genes of planaria. Mol Biol Evol. 20, 1260-1270.

Burke W.D., Malik H.S., Jones J.P., Eickbush T.H., 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. Mol Biol Evol. 16, 502-511.

Burke W.D., Malik H.S., Lathe W.C. 3rd, Eickbush T.H., 1998. Are retrotransposons long-term hitchhikers? Nature. 392, 141-142.

Christensen S.M., Bibillo A., Eickbush T.H., 2005. Role of the *Bombyx mori* R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. Nucleic Acids Res. 33, 6461-6468.

Christensen S.M., Ye J., Eickbush T.H., 2006. RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. Proc Natl Acad Sci USA. 103, 17602-17607.

Eickbush, T. H., 2002. R2 and related site-specific non-long terminal repeat retrotransposons, pp. 813–835. In: Mobile DNA II, N. L. Craig, R. Craigie, M. Gellart and A. M. Lambowitz, eds. American Society of Microbiology, Washington, DC.

Eickbush T.H., Eickbush D.G., 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics. 175, 477-485.

García-Varela M., Nadler S.A., 2006. Phylogenetic relationships among Syndermata inferred from nuclear and mitochondrial gene sequences. Mol Phylogenet Evol. 40, 61-72.

George J.A., Eickbush T.H., 1999. Conserved features at the 5’ end of Drosophila R2 retrotransposable elements: implications for transcription and translation. Insect Mol Biol. 8, 3-10.

Gilbert N., Lutz-Prigge S., Moran J.V., 2002. Genomic deletions created upon LINE-1 retrotransposition. Cell. 110, 315-325.

Gilbert N., Lutz S., Morrish T.A., Moran J.V., 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol. 25, 7780-7795.

Gladyshev E.A., Arkhipova I.R., 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci USA. 104, 9352-9357.

Gladyshev E.A., Meselson M., Arkhipova I.R., 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. 320, 1210-1213.

Gladyshev E.A., Meselson M., Arkhipova I.R., 2007. A deep-branching clade of retrovirus-like retrotransposons in bdelloid rotifers. Gene. 390, 136-145.

Gladyshev E., Meselson M., 2008. Extreme resistance of bdelloid rotifers to ionizing radiation. Proc Natl Acad Sci USA. 105, 5139-5144.

Hur, J., 2006. Duplication and divergence of regions containing *hsp82* before the separation of two bdelloid families, Adinetidae and Phylodinidae. Ph.D. Dissertation, Harvard University, Cambridge, MA.

Kierzek E., Christensen S.M., Eickbush T.H., Kierzek R., Turner D.H., Moss W.N., 2009. Secondary Structures for 5' Regions of R2 Retrotransposon RNAs Reveal a Novel Conserved Pseudoknot and Regions that Evolve under Different Constraints. J Mol Biol. 390, 428-442.

Kojima K.K., Fujiwara H., 2005. Long-term inheritance of the 28S rDNA-specific retrotransposon R2. Mol Biol Evol. 22, 2157-2165.

Kojima K.K., Kuma K., Toh H., Fujiwara H., 2006. Identification of rDNA-specific non-LTR retrotransposons in Cnidaria. Mol Biol Evol. 23, 1984-1993.

Malik H.S., Burke W.D., Eickbush T.H., 1999. The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol. 16, 793-805.

Mamedov I.Z., Lebedev Y.B., Sverdlov E.D., 2004. Unusually long target site duplications flanking some of the long terminal repeats of human endogenous retrovirus K in the human genome. J Gen Virol. 85, 1485-1488.

Moran J.V., Holmes S.E., Naas T.P., DeBerardinis R.J., Boeke J.D., Kazazian H.H. Jr. 1996. High frequency retrotransposition in cultured mammalian cells. Cell. 87, 917-927.

Stage D.E., Eickbush T.H., 2009. Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. Genome Biol. 10, R49.

Szak S.T., Pickeral O.K., Makalowski W., Boguski M.S., Landsman D., Boeke J.D., 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3(10):research0052.

Symer D.E., Connelly C., Szak S.T., Caputo E.M., Cost G.J., Parmigiani G., Boeke J.D. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. Cell. 110, 327-338.

Tamura K., Dudley J., Nei M., Kumar S.. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24, 1596-1599.

Yang J., Malik H.S., Eickbush T.H.. 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc Natl Acad Sci USA. 96, 7847-7852.

LEGENDS TO FIGURES

Fig. 1. Structural organization of R9Av (top) and the *A. vaga* rDNA unit (bottom). RT, reverse transcriptase; EN, restriction enzyme-like endonuclease; TSD, target site duplication; ZnF, three motifs containing CCHH, CCHC and CCHH fingers, respectively; myb, a putative region of homology to c-myb domain; IGS, intergenic spacer between rDNA units; ETS, external transcribed spacer; ITS1 and ITS2, internal transcribed spacers. ATG denotes the position of the fist start codon, and a dashed vertical line shows the position of the frameshift observed in two of the sequence variants. Hybridization probes used in this study are shown by thick horizontal lines below the drawings. R2 designates the position of the canonical insertion site used by R2 elements. X, *Xho*I recognition sites. Scale bar, 1 kb.

Fig. 2. Southern analysis of the *A. vaga* genomic DNA digested with *Xho*I restriction endonuclease. The blot was hybridized with a 259-bp 28S (rDNA) probe located 0.4 kb downstream from the R9Av insertion site, and with the R9 probe (see Fig. 1). Left, ethidium bromide stained agarose gel; right, hybridization to genomic DNA. M, 1 kb+ DNA ladder (Invitrogen); sizes in kb are indicated. Inserted and uninserted rDNA units are marked by arrows.

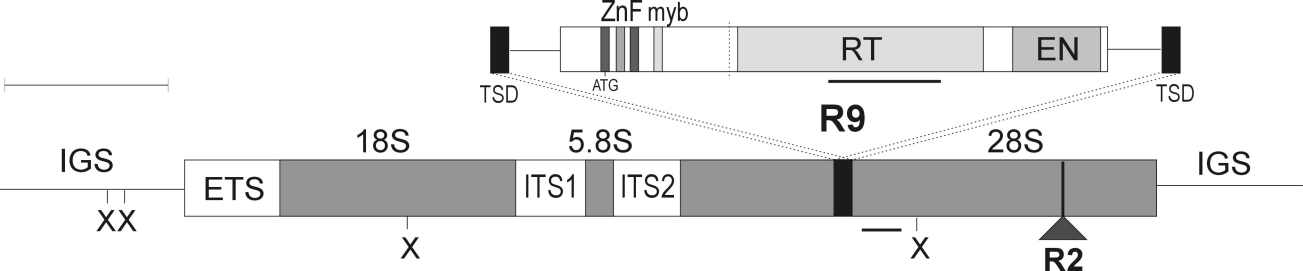
Fig. 3. Amino acid sequence alignments of selected domains from the R9Av-encoded ORF and the corresponding domains from other members of the R2 clade. Shown are the N-terminal regions containing three Zn finger motifs (**A**), a

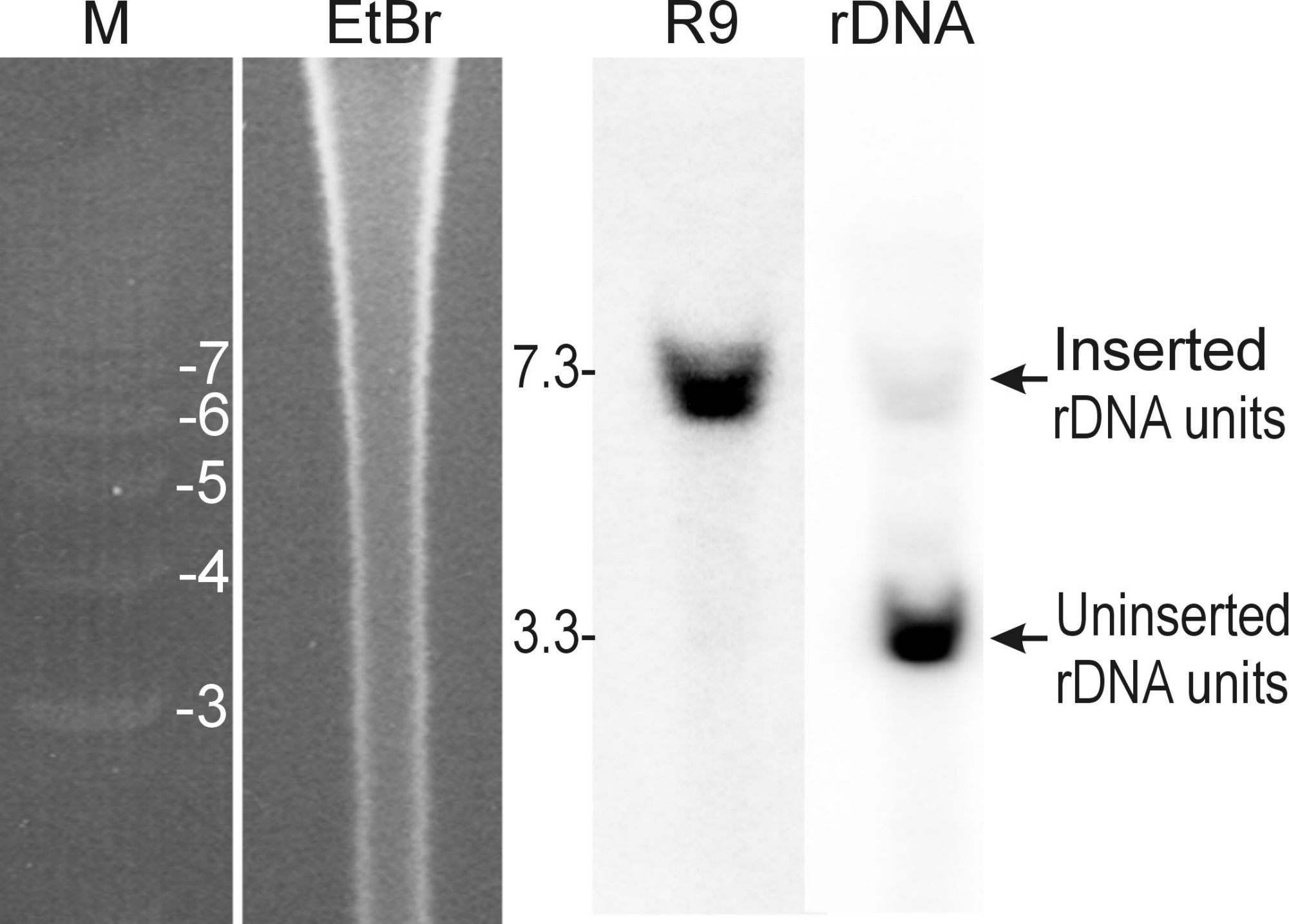
segment of the core RT domain containing motifs RT5 through RT7 (**B**), and a segment of the restriction enzyme-like endonuclease domain containing the CCHC finger motif and the first conserved motif of the endonuclease moiety. The species abbreviations are as follows: *Hm*, *Hydra magnipapillata*; *Tl*, *Triops longicaudatus*; *Ci, Ciona intestinalis*; *Lp, Limulus polyphemus; Dr, Danio rerio*; *Sp, Strongylocentrotus purpuratus;Pm, Petromyzon marinus;Tg, Taenopygia guttata; Nv, Nasonia vitripennis; Am, Apis mellifera*.

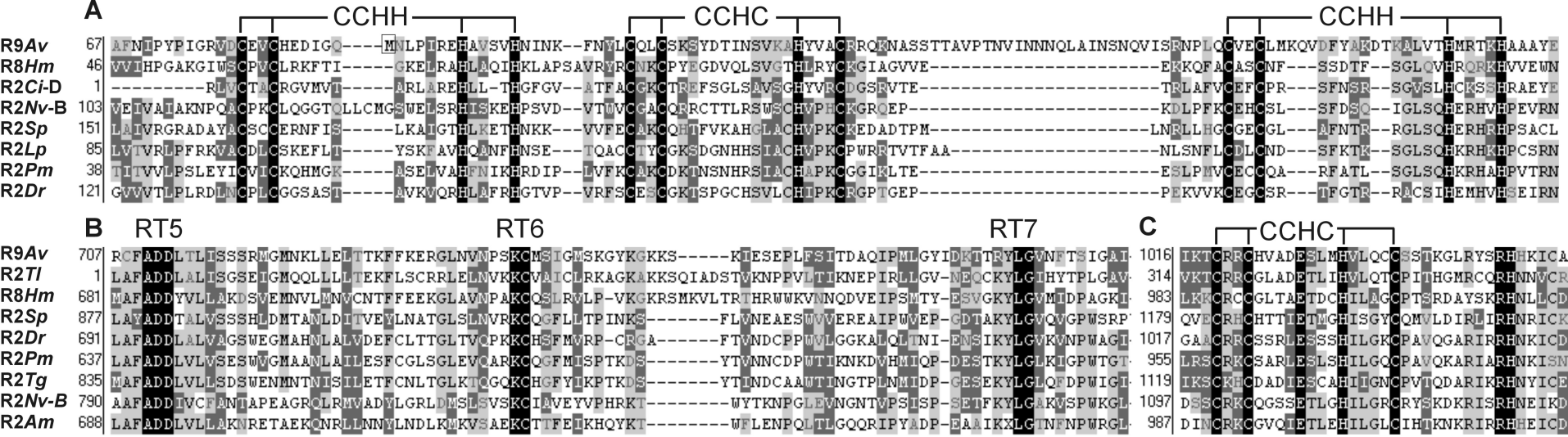
Fig. 4. Nucleotide sequence alignments between R9 extremities and 28S rDNA sequences. The rDNA sequences (R) are unshaded; the R9 sequences are shaded in gray. (**A**) Junction sequences of the R9 element with the 28S gene. The four R9Av sequence variants are designated 1, 2, 3, and 4; the insertion 126 bp upstream from copy 3, with similarity to the 3' UTR of R9Av, is designated U. The 126-bp target site duplication in rDNA (R) is underlined. Boundaries between R9Av and rDNA are denoted by **|**. (**B**) Nucleotide sequence conservation between 28S rDNA sequences from different species in the 126-bp region spanning the R9Av TSD. Sites of the bottom (^) and top (^{V}) strand cleavage are indicated. Sequences from four of the species listed in Fig. 3 are shown, with addition of *Drosophila melanogaster* (Drome), *Anopheles gambiae* (Anoga), *Schistosoma mansoni* (Schma), and *Xenopus laevis* (Xenla).

Fig. 5. Phylogenetic placement of R9Av within the R2 clade. Shown is the neighbor-joining unrooted phylogram including R9Av and 33 other members of

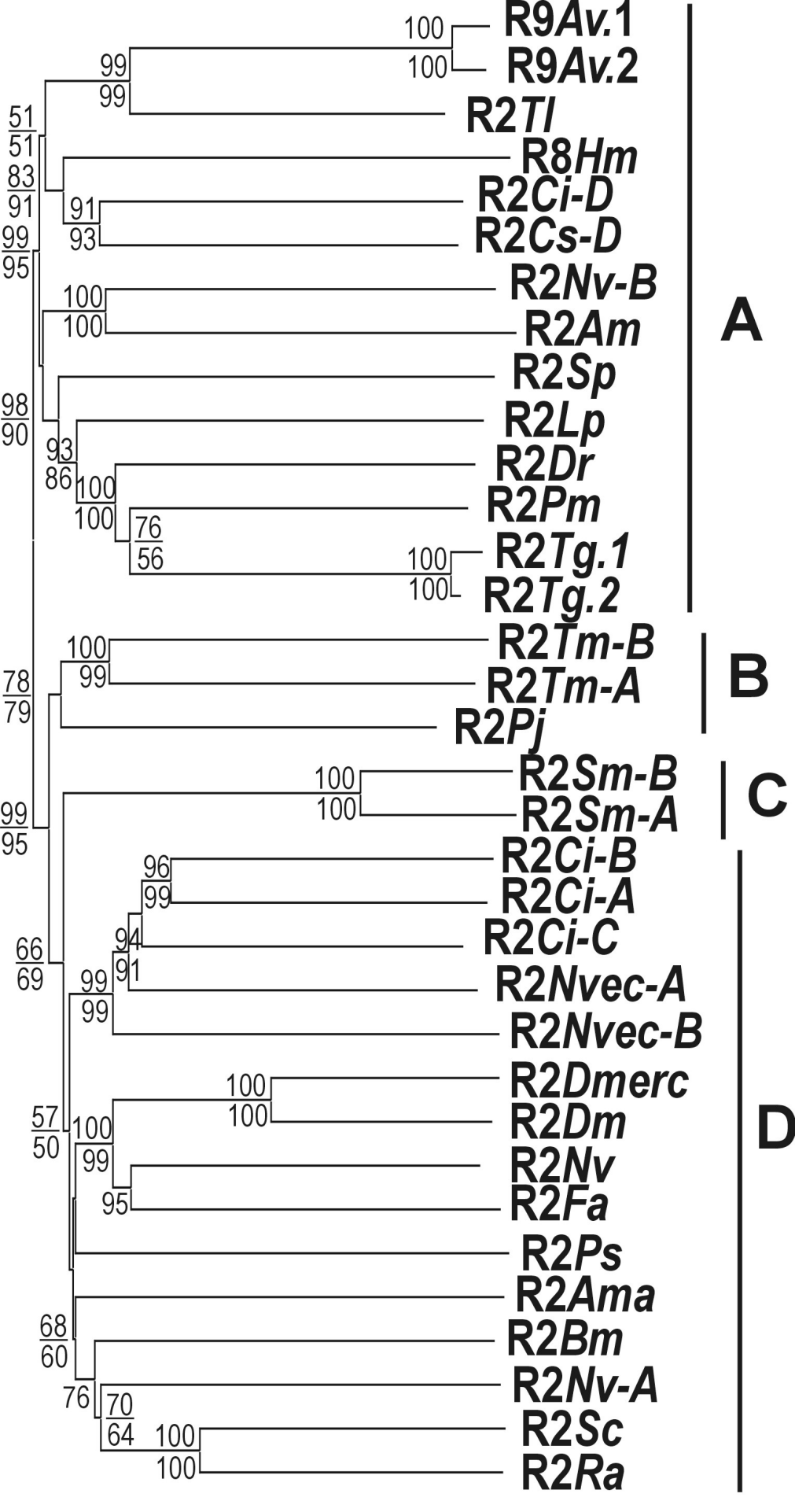
the R2 clade, with representatives from each subclade (A, B, C, and D) (Kojima and Fujiwara, 2005). Bootstrap support values exceeding 50% from neighbor joining/minimum evolution analyses are indicated. Species abbreviations are as in Fig. 3, with addition of *Tm (Tenebrio molitor), Pj (Popilia japonica), Sm (Schistosoma mansoni), Nvec (Nematostella vectensis), Dm (D. melanogaster), Dmerc (D. mercatorum), Fa (Forficula auricularia), Ps (Porcellio scaber), Ama (Anurida maritima), Bm (Bombyx mori), Sc (Sciara coprophila)*, and *Ra (Rhynchosciara americana)*. Amino acid sequence alignment with accession numbers is provided as Supplementary data. Scale bar, 0.1 amino acid substitutions per site.


ZnF myb
RT
EN
ATG
TSD
TSD
R9
IGS
18S
5.8S
28S
IGS
ETS
ITS1
ITS2
XX
X
X
R2


M
EtBr
R9
rDNA
-7
-6
-5
-4
-3
7.3-
3.3-
Inserted
rDNA units
Uninserted
rDNA units


A
CCHH
CCHC
CCHH
R9Av
R8Hm
R2Ci-D
R2Nv-B
R2Sp
R2Lp
R2Pm
R2Dr
B
RT5
RT6
RT7
R9Av
R2Tl
R8Hm
R2Sp
R2Dr
R2Pm
R2Tg
R2Nv-B
R2Am
C
CCHC


A
R9 3'-end
rDNA
1 CTTTTCTACTGTGTTCTTTTTATCAGTTTTTTGTGGAAAAATTGAGAATAAATAAAGT|TAGCTGGTTCCGTCCGAAGTTTCCCTCAGGATAGCTGGCATTCAATTTTCACAGTTATAT
2 ...|...
3 ...|...
4 ...T...|...
U .....-GTG.....A.....-.....C.G.AC|...
TT T T A G 1
R CGTGCAAATCGATCGTCAAACATGCGTTAAGGGGCGAAAGACTAATCGAACCATCTAG|TAGCTGGTTCCGTCCGAAGTTTCCCTCAGGATAGCTGGCATTCAATTTTCACAGTTATAT
R CCGGTAAAGCGAATGTTTAGAGGCATTGGGCGTAAAATACGCTCAACCTATTGACAAACTTTAAAT|GGGTATGAAGTTCATTTTTCTACTTCATTGAGAATGAACAGCGAATGTGAGT
126
rDNA
R9 5'-end
1 CCGGTAAAGCGAATGTTTAGAGGCATTGGGCGTAAAATACGCTCAACCTATTGACAAACTTTAAAT|GAAATAGTTTGCAATGGTAGGTGTATGGCGCCTCTGTGTCTCTCTTTCGCTG
2 ...|...
3 ...|...
4 ...|----C.....-AT.....ATG.....C.....C.....
B
1
126
AVAGA TAGCTGGTTCCGTCCGAAGTTTCCCTCAGGATAGCTGGCATTC---AATTTT------CACAGTTATATCCGGTAAAGCGAATGTTTAGAGGCATTGGGCGTAAAATACGCTCAACCTATTGACAAACTTTAAAT
ANOGA .....C.....TGCA.GT-.GCG..TCGAAC.TT.T.CT....T.....A.....C..A..TTCG....GAT..T.....CT.....A.....
DROME .....T.....TGCATTTT...A..ATATAAA.T.A.CT....T.....A.....C..A..GTCG...CGAT..T.....CT.....
LIMPO .....C.....TGC..G--.GGG.-----AAAT....CTC.....A.....C.....GCCG...CGAC.....CT.....
SCHMA .....C.....TGAG.AA.A------A.....T.....A.....GAGG..TC.TC.....CT.....
SPURP .....C.....GC..A----.A--------G.....T....T.....A.....C.....GACG...CGAC.....CT.....
CIONA .....C.T.....GC..------.G.C-------G.....T....T.....A.....TC.....G.CG...CG.C.....CT.....
DANIO .....C.....GC..GA---.AC--------G.....T.....AC.....C.....GCCG...CGGC.....CT.....
XENLA .....C.....GC..GT--CCG.C-------G.....T.....A.....TC.....GCCG...CGAT.....CT.....


R9Av.1
R9Av.2
R2Tl
R8Hm
R2Ci-D
R2Cs-D
R2Nv-B
R2Am
R2Sp
R2Lp
R2Dr
R2Pm
R2Tg.1
R2Tg.2
R2Tm-B
R2Tm-A
R2Pj
R2Sm-B
R2Sm-A
R2Ci-B
R2Ci-A
R2Ci-C
R2Nvec-A
R2Nvec-B
R2Dmerc
R2Dm
R2Nv
R2Fa
R2Ps
R2Ama
R2Bm
R2Nv-A
R2Sc
R2Ra
A
B
C
D