Classification: Evolution

Telomere-associated, endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes

Eugene A. Gladyshev[*] & Irina R. Arkhipova[*†‡]

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

[‡]To whom correspondence should be addressed. E-mail: arkhipov@fas.harvard.edu.

Abbreviations: TE, transposable element; ORF, open reading frame; aa, amino acid; LTR, long terminal repeat; RT, reverse transcriptase; EN, endonuclease; TERT, telomerase reverse transcriptase.

Data deposition: Sequences reported in this paper were deposited in GenBank (accession nos. EF484951-EF485020).

Keywords: *reverse transcriptase* | *telomerase* | *transposable elements*

The authors declare no conflict of interest.

The evolutionary origin of telomerases, enzymes that maintain the ends of linear chromosomes in most eukaryotes, is a subject of debate. *Penelope*-like elements (PLEs) are a recently described class of eukaryotic retroelements characterized by a GIY-YIG endonuclease domain and by a reverse transcriptase domain with similarity to telomerases and group II introns. Here we report that a subset of PLEs found in bdelloid rotifers, basidiomycete fungi, stramenopiles, and plants, representing four different eukaryotic kingdoms, lack the endonuclease domain and are located at telomeres. The 5' truncated ends of these elements are telomere-oriented and typically capped by species-specific telomeric repeats. Most of them also carry several shorter stretches of telomeric repeats at or near their 3' ends, which could facilitate utilization of the telomeric G-rich 3' overhangs to prime reverse transcription. Many of these telomere-associated PLEs occupy a basal phylogenetic position close to the point of divergence from the telomerase-PLE common ancestor, and may descend from the missing link between early eukaryotic retroelements and present-day telomerases.

Genomic DNA in many eukaryotes is composed, to a large extent, of transposable elements (TEs), especially retrotransposons, which multiply *via* an RNA intermediate copied into DNA by reverse transcriptase (RT) and inserted into new sites by an endonuclease/integrase. While RT creates new copies, DNA cleavage is essential for TE proliferation, *i.e*. insertion into previously unoccupied sites. Integrases of retrovirus-like

(LTR) retrotransposons insert dsDNA into chromosomes, while endonucleases (EN) of non-LTR retrotransposons generate the 3'OH-end that primes cDNA synthesis directly onto the chromosome (target-primed reverse transcription). The only known eukaryotic RT-containing genes lacking EN domains are telomerase reverse transcriptases (TERTs), which are not TEs but specialized ribonucleoprotein enzymes maintaining telomeres by repeated copying of a short segment of an unlinked template RNA, primed by the 3' OH end of a linear chromosome (see 1-5 for review).

PLEs are a widespread but not very extensively studied class of eukaryotic TEs characterized by a single ORF coding for RT and an unusual GIY-YIG EN domain also found in bacterial group I introns, and by the presence of spliceosomal introns in several members (4,6). They occupy a special place in retroelement phylogeny by sharing a common ancestor with TERTs (4). PLEs insert relatively randomly throughout the genome, preferring AT-rich targets (6). Indeed, the element-encoded EN, in which the conserved residues are essential for transposition, exhibits some sequence preferences but no pronounced sequence-specificity (7).

Rotifers of the class Bdelloidea, a large taxon of multicellular freshwater invertebrates considered to be anciently asexual (8,9), contain a distinct group of PLEs, called *Athena* (4), carrying spliceosomal introns within highly conserved RT motifs. Two *Athena* copies initially obtained from a genomic library of the bdelloid *Philodina roseola* were missing the entire EN domain and contained short stretches of reverse-complement telomeric repeats, $(TCACCC)_{3\text{-}5}$, near their 3' termini. This finding prompted us to investigate the universality of the EN domain absence and possible telomeric associations in this special group of PLEs in two bdelloid species, *Adineta vaga* and *P. roseola*, representing two families that separated tens of millions of years ago (9).

Results

To find out whether *Athena* elements are indeed located at telomeres, we developed a method for constructing telomere-enriched plasmid mini-libraries containing inserts of chromosomal DNA, originally located either at telomeres or at sites of chromosome breakage, which does not rely on any prior knowledge of sequences at the chromosome ends (Fig. 1; Methods). Three independent mini-libraries were prepared for *A. vaga*, which has 12 chromosomes and *ca*. 500-Mbp genome (10,11). Random sequencing of mini-library clones identified $(TGTGGG)_n$ as *A. vaga* telomeric repeats. We obtained 44 different telomere sequences ending with $(TGTGGG)_n$ (Supporting Table 1), indicating chromosome end polymorphism. Notably, two telomeres (designated M and N) contained 5' truncated but otherwise intact ORFs of two *Athena* variants, designated *Athena-AvM* and *Athena-AvN* (Fig. 2a). Telomeric clones were also obtained by probing the *A. vaga* genomic fosmid library with $(TGTGGG)_4$. Fosmid sequencing revealed several *Athena* variants, forming head-to-tail interspersed tandem arrays at the chromosome termini (Fig. 3a). The variant *Athena-AvO* was first identified on fosmids, and its 3' UTR was then shown to match telomeres O1-O3 (Fig. 2a, 3a; Supp. Table 1). Subtelomeric *Athena* copies are 5' truncated by addition of reverse-complement telomeric repeats, and their coding sequences are typically followed by 1-3 shorter stretches of reverse-complement telomeric repeats (3-5 repeat units). Similarly oriented but decayed *Athena* copies were found in the adjacent proximal region of a fosmid, forming arrays up to 7 deep (not shown). All complete *Athena* elements code for both an RT and an upstream ORF1 with several nuclear localization signals (NLS) and a coiled-coil motif (Fig. 2a). In no case, however, could we detect an associated EN domain: the C-terminal region is *ca*. 100-150 aa shorter than in EN-containing PLEs (Supp. Fig. 6). While it is formally possible that the RT domain *per se* may exhibit a cryptic EN activity, this possibility appears unlikely.

We next sought to confirm by an alternative technique that the *Athena* elements are located at chromosome ends. We chose a PCR-based technique called STELA, which was developed to measure single telomere length variation (12). Primers were designed to amplify telomeres containing *Athena-AvM* and *Athena-AvO* (Fig. 2a,b; Fig. 3b). Sequencing of cloned amplicons confirmed their exact correspondence to the telomere M1 for *AvM* primers, and to telomeres O1-O3 for *AvO* primers. The length of the amplified telomeric repeat tracts (up to 65 repeat units) can be as short as 3-4 units, and occasional incorporations of a variant repeat were observed in the proximal region, indicating that telomeric tracts are subject to cycles of expansion and contraction, during which considerable telomere shortening may occur.

It was also of interest to find out whether the *Athena* variants that code for a full-length ORF can be transcribed, and whether the transcription start site is located at or near the 5' end of the element to give rise to a full-length protein. RT-PCR experiments yielded bands of the expected size and sequence for the three *Athena* variants depicted in Fig. 2a, including spliced forms of the intron-containing *AvN,* for which an unspliced product was also detected (Fig. 3d). Transcription start sites were determined for *AvO* and *AvN* by 5' RACE (Methods), and sequencing of individual amplicons confirmed that the RNA start sites in each case are positioned upstream of the first ATG codon of ORF1, with a single predominant start site for *AvN* and several start sites for *AvO* (Fig. 3c; Supp. Fig. 7).

The telomere cloning procedure was also applied to *P. roseola*, a species with 13 chromosomes, two of which are dot chromosomes (10). Its estimated genome size of ~2,000 Mbp (11) exceeds that of *A. vaga*, and exhaustive cloning of chromosome ends is more challenging because of the lower ratio of chromosome ends to random breaks. From three independent *P. roseola* mini-libraries we obtained 20 $(TGAGGG)_n$-containing telomeres (Supp. Table 1), one of which matched the *Athena-PrT* variant

found on two of three sequenced cosmids (Fig. 2b). Two other telomeric clones had weaker matches to the transcribed and spliced (4) *Athena-PrR* variant, also present on two cosmids (*PrR**, Fig.1b; Supp. Fig. 7). In addition, we recovered five *Athena* clones not capped with telomeric repeats (Fig. 2b; Supp. Table 1), which may have originated either from sites of chromosome breakage, or from exposed chromosome termini not yet capped by telomeric repeats.

The sequenced *P. roseola* cosmids, similar to *A. vaga* telomeric fosmids, exhibited a high density of *Athena* elements, all characteristically lacking an EN domain. Two cosmids carried 4 variants each, together with various DNA transposons (13), and one consisted almost entirely of 6 *Athena* variants, intact followed by decayed. As in *A. vaga*, many *Athena* copies were truncated at the 5' end with reverse-complement telomeric repeats, and carried short stretches of such repeats downstream of the RT ORF. The *Athena*-containing cosmid inserts, which in this case do not carry terminal telomeric repeats because of the library construction method, were employed as probes for fluorescent *in situ* hybridization to *P. roseola* embryo nuclei (Supp. Fig. 8). Each cosmid yielded, on average, four strong and two weak telomeric hybridization signals, the latter at the two ends of a dot chromosome. No hybridization to internal sites was detected, although the sensitivity of the technique allows one to visualize only fragments as large as 30-40 kb (14). Labeling of several ends agrees with the telomere cloning data, while other, more diverged *Athena* variants that may be present at other ends may have insufficient homology to the probe to generate observable signal. Six additional cloned *A. vaga* and four *P. roseola* telomeres (Supp. Table 1) were also suspected to be formed by terminal addition of as yet unknown diverged variants: they contain identical subterminal segments 0.3-2 kb in length.

To find out how many copies of each *Athena* variant are present in the *A. vaga* genome, we performed an exhaustive screen of the genomic library with *Athena* probes

and compared the number of positive fosmids with the number of fosmids containing the *A. vaga hsp82* gene, of which there are four copies (15). This method, in contrast to *in situ* and telomeric mini-library screening, is biased against chromosome termini, which are strongly under-represented in genomic libraries, but would detect all internal copies, even short ones. We find that, for each tested *Athena* variant, the number of hybridizing fosmids per genome is even less than that for *hsp82* (Supp. Table 3A). Most of these fosmids, however, also hybridize to the telomeric repeat probe, indicating that they likely originate from subtelomeric locations and contain remnants of former telomeres.

To find out whether telomere-associated, EN-deficient retroelements are a unique feature of bdelloid genomes or represent a more general phenomenon, we searched publicly available databases for PLEs with similar properties. Among numerous PLE ORFs assembled from diverse eukaryotes, we identified EN-deficient ORFs in genomes of representatives of three other kingdoms: fungi (inky cap mushroom *Coprinus cinereus* and the white rot fungus *Phanerochaete chrysosporium*); plants (spike moss *Selaginella moellendorffii*); and stramenopiles, or heterokonts (pennate diatom *Phaeodactylum tricornutum*) (Fig. 2c-f). Strikingly, all of them exhibit the same connections with species-specific telomeric repeats: most of the copies contain short stretches of such repeats at or near the 3' termini, and are 5' truncated by a longer stretch of telomeric repeats comprising the chromosome end (Fig. 2c-f; Supp. Table 2). The fungal *Coprina* elements are somewhat distinct in having a single long ORF and a slightly extended C-terminus (Fig. 2 c,d; Supp. Fig. 6), while the protist and plant elements, like *Athena*, possess an upstream ORF1 which exhibits poor conservation (as opposed to RT), low amino acid sequence complexity, and no discernable sequence motifs other than NLS and coiled-coil domains (Fig. 2c-f). In all of these elements, the 5’end is apparently present at a single genomic location, so that the full-length elements may essentially be regarded as single-copy genes (Supp. Table 3B).

Remarkably, comparison of available PLE sequences shows that sequence similarities between PLEs and TERTs can be extended beyond the seven core RT1-RT7 motifs into the N-terminal and C-terminal domains, with the N-termini alignable for at least 200 aa, and the C-termini of TERTs and EN(-) PLEs ending at approximately the same position, which serves as the EN addition point in EN(+) PLEs (Supp. Fig. 6). The extended alignment provides an opportunity to refine PLE-TERT phylogenetic relationships, previously investigated at the level of core RT only (4,16). An initial snapshot of the phylogenetic data structure within the combined PLE-TERT dataset was obtained by NeighborNet analysis (Supp. Fig. 9A), and the suggested topology was then evaluated by other phylogenetic methods such as likelihood distance-based analysis with bootstrap networks (Fig. 4) and maximum likelihood analysis under the best fitting model (Supp. Fig. 9B). Of the two major PLE groups with the GIY-YIG domain found in animals, *Penelope/Poseidon* and *Neptune* (6,17), the *Penelope* group forms a well-supported late-branching clade, while the position of the *Neptune* group is less certain. All telomere-associated EN(-) PLEs can be roughly assigned to two major groups, *Coprina* and *Athena*, with *Coprina* elements appearing as the earliest-branching clades since the divergence of PLEs and TERTs from the common ancestor, possibly predating EN acquisition. In our previous analysis of the core RT domain (4), *Athena* elements formed a sister clade to *Neptune*, but this placement by Bayesian analysis may have been overconfident, since it is not observed in neighbor-joining or maximum likelihood analyses, and statistical tests demonstrate that the branching order of *Athena* and *Neptune* elements cannot be determined with confidence (Supp. Table 4). These tests also reject late-branching position for *Coprina* elements, thereby placing their origin early in eukaryotic evolution. Two alternatives for *Athena* origin may be considered: initial lack of EN, or its secondary loss. The latter appears somewhat less likely, since several independent EN losses by precise truncation would have had to occur in each of the *Athena* variants.

Discussion

Several telomere-associated non-LTR retrotransposons have been described previously: *HeT-A*, *TAHRE*, and *TART* in Drosophila (18,19), *SART* and *TRAS* in *Bombyx mori* (20), *GilM* and *GilT* in *Giardia lamblia* (21). Most of them have an intact EN domain, raising the possibility of EN-mediated specific insertion into a subterminal target, shown directly for *SART* and *TRAS* (20). In our case, however, the lack of an associated EN domain, characteristic patterns of telomeric repeat distribution at the 5' and 3' termini, orientation preference, and similarity to TERTs strongly argue in favor of terminal addition to exposed chromosome ends. The lack of EN activity leaves these elements with little choice other than using the available 3'-OH at the chromosome ends to prime reverse transcription. The shortness of the telomeric repeat stretch between PLEs and the adjacent genomic DNA (Supp. Fig. 7) indicates that, prior to PLE addition, telomere length is considerably reduced, which is likely associated with loss of the normal capping structure. Utilization of free chromosome ends would not completely rule out occasional insertion at internal sites, *e.g.* in the course of double-strand DNA break repair, as observed for mammalian L1 non-LTR retrotransposons with a disabled EN domain (22), at replication forks (23), or upon action of endonucleases coded elsewhere. All of these processes, however, would be insufficient for effective spread of EN-deficient PLEs, and the overwhelming majority of insertions do occur at telomeres.

Our model for EN-independent terminal retrotransposition, which accommodates most of the observed structural features, is presented in Fig. 5. Notably, terminal retrotransposition exhibits the same polarity as in telomeric repeat addition by TERTs. cDNA synthesis is accompanied by telomerase-mediated addition of telomeric repeats to the variably-truncated 5' end at sites with 3-4 nucleotide microhomologies to the telomeric repeat unit. At the target-priming stage, reverse-complement telomeric repeats

in the 3’ UTR could facilitate annealing between the template and the telomeric G-rich 3' overhang. Primer-template annealing is required for integration of non-LTR retrotransposons in *B. mori* (20), and may also facilitate L1 integration in mammals (24). The occurrence of several short telomeric repeat stretches within each 3’ UTR may have resulted from occasional acquisitions of additional downstream sequences after terminal transposition and readthrough transcription, similar to 3’ transduction in L1 elements (25). Elements that apparently do not require 3' telomeric repeats for attachment, such as *AvM*, might be capable of extending severely eroded telomeres, which have already lost their telomeric repeats. The ORF1 product may be hypothesized to play a role in targeting, as shown for Drosophila *HeT-A* and *B. mori SART* elements (20,26), and/or in primer-template annealing, as shown for mammalian L1 ORF1, which also contains a coiled-coil domain and a basic region (27).

Although EN(-) retroelements may simply be transposing to telomeres in order to minimize damage to host genes, their low replicative capacity, resulting from inability to generate insertion sites on their own, is not very likely to ensure their survival as “selfish DNA” (28), which should replicate more efficiently than host DNA. Rather, it may be hypothesized that these low copy number elements, essentially confined to the chromosome termini, were occasionally preserved in evolution as a supplement to the telomerase-based system, providing extra protection against terminal DNA loss. In the early days of eukaryotic evolution, when primordial RNA-dependent DNA polymerases have not yet become associated with endonucleases to give rise to “selfish” retrotransposons that later conquered most eukaryotic genomes, movement of reverse-transcripts could have been limited to the free DNA ends. Over time, an ancestral retroelement could have evolved into a telomerase catalytic subunit upon disruption of linkage between RT and its template RNA, which would then become a subunit of the telomerase holoenzyme. In the evolutionary history of eukaryotes, telomere-associated PLEs may therefore be regarded as descendants of the missing link between ancient

EN(-) retroelements and the present-day telomerases, shedding light on the fundamental problem of evolution of telomerase-based maintenance of linear chromosome ends.

Materials and Methods

Construction of telomere-enriched plasmid mini-libraries. High-molecular weight (HMW) chromosomal DNA was prepared by embedding rotifers into 0.7% LMP agarose blocks, digesting with Proteinase K (Invitrogen) at 55^{o}C for 30 h in 1x digestion buffer (50 mM NaCl, 50 mM TrisHCl, pH 8.0, 100 mM EDTA, 1% Sarcosyl, 2 mM spermine, 2 mM spermidine), and removing broken DNA by pulsed-field gel electrophoresis with the following parameters: 0.7% LMP agarose (SeaPlaque), 5 V/cm, switch time 50-250 sec, switch angle 120 degrees, run time 18 h, 0.5xTAE buffer at 12^{o}C (BioRad CHEF-DR III System). HMW DNA (>1.9 Mbp) was excised from the gel compression zone and stored in agarose blocks in 50 mM TrisHCl, pH 8.0, 50 mM EDTA. For cloning, blocks were dialyzed against 50 mM TrisHCl pH 7.5, 10mM $MgCl_2$ for 5 h at 4^{o}C on shaker, transferred to 0.75-ml tubes, and supplemented with a soaking solution of DTT, dNTPs, BSA, $MgCl_2$, TrisHCl, pH 7.5 and T4 DNA polymerase (NEB) to bring their concentrations in agarose blocks to 5 mM, 0.25 mM each, 100 μg/ml, 10 mM, 50 mM, and 3U/100 μl, respectively. After soaking for 4 hrs on ice, tubes were transferred to 14^{o}C for 1 h to activate T4 DNA polymerase, and then back on ice. Blocks were carefully removed and dialyzed against 25 mM TrisHCl, pH 8.0, 50 mM EDTA for 10 h to remove salt, dNTPs, and T4 DNA polymerase. Blocks were transferred to fresh 0.75-ml tubes, agarose was melted for 5 min at 65^{o}C, supplemented with 2μg/100 μl of pBluescript II SK- (Stratagene) linearized with *Hin*cII, and dephosphorylated with shrimp alkaline phosphatase (Promega). Extreme care was taken to add the vector as slowly and gently as possible to minimize HMW DNA breakage. Vector was allowed to diffuse in melted agarose for 3 h at 37^{o}C, and agarose was supplemented with a mixture of DTT, ATP, BSA, $MgCl_2$, TrisHCl, pH 7.5 and T4

DNA ligase (High-concentration, Invitrogen) to their final concentrations of 10mM, 1mM, 50 μg/ml, and 15 Weiss Units/100 μl, respectively. Again, extreme care was taken not to cause breakage of HMW DNA. The ingredients were allowed to diffuse in melted agarose for 30 min at 37^{o}C, and the tubes were transferred to 14^{o}C for 24 h to allow ligation. After ligation, extreme care is no longer necessary. Blocks were melted for 5 min at 65^{o}C, agarose was mixed by pipetting, transferred on ice, let to solidify, and equilibrated with 0.5xTAE buffer for 3 h. Unligated vector was removed from genomic DNA by four rounds of electrophoresis (two forward and two reverse) in 0.5% LMP agarose, 0.5xTAE at 4^{o}C. Genomic DNA in agarose was digested with β-agarase I (NEB) in 0.5xTAE supplemented with 1x NEBuffer III and 100 μg/ml BSA. DNA was digested to completion with *Hin*cII (10 U/100 μl), extracted with phenol-chloroform, chloroform, EtOH-precipitated and dissolved in 72 μl H_2O. The solution was supplemented with 10 mM DTT, 1 mM ATP, 10 mM $MgCl_2$, 50 mM TrisHCl, pH 7.5, and 15 Weiss units of T4 DNA ligase in the final volume of 100 μl. After ligation for 16 h at 14^{o}C, DNA was extracted with phenol-chloroform, chloroform, EtOH-precipitated, and dissolved in 4 μl of water to transform 20 μl of DH10B electrocompetent *E. coli* (Invitrogen) in BioRad Gene Pulser (2 kV, 25 μF, 200 Ohm, 2-mm-wide cuvette). Inserts were sequenced with M13 forward and reverse primers to determine the telomeric end, and, if no internal tandem repeats were present, sequenced to completion by primer walking from the non-telomeric end. The procedure was initially tested on *D. melanogaster* genomic DNA and resulted in cloning of a telomere-associated retrotransposon *HeT-A* (not shown).

Cloning, sequencing, and hybridization. Telomere sequences were also obtained by screening the *A. vaga* genomic fosmid library prepared from sheared embryo DNA (15) with $(TGTGGG)_4$ telomeric repeat probe. End-sequencing of hybridizing fosmids was employed to determine whether the insert contains telomeric repeats at one end. Genomic *P. roseola* cosmid library, prepared by partial *Sau*3AI digestion (14), was used

to select *Athena*-containing clones by hybridization to a PCR-generated mixed *Athena* probe described in (4). *Athena*-containing fosmids/cosmids were sheared by sonication, subcloned into pBluescript II SK-, and sequenced on ABI3730XL. Cosmids used as FISH probes were purified using NucleoBond® Maxi Kit (Clontech), labeled by nick-translation to incorporate the red fluorophore Alexa 568-dUTP (Molecular Probes), under conditions adjusted to yield 100-300-nt fragments, and FISH was performed as in (14). Cultures of *A. vaga* and *P. roseola* maintained in the laboratory descend from a single egg isolated 10 and 15 years ago, respectively.

STELA. Rotifer genomic DNA (0.5 µg) was used for STELA (12) with the following modifications: the total volume of ligation mix was 15 µl; 25 pmols of each telorette oligo (GTGACGCTATCATAACGCTCCCCACACCC, GTGACGCTATCATAACGCTCCCACACCCA) were used together; following ligation, genomic DNA was separated from unligated oligonucleotides on Sephacryl S500, extracted with phenol/chloroform and chloroform, precipitated with EtOH, and resuspended in 30 µl H_2O. 1 µl of resuspended DNA was used for PCR with Expand Long Template PCR system (Roche) with primers teltail (GTGACGCTATCATAACGCTC) and AvM (TGGTAGGCTTTCAAGGCTG) or AvO (ACGTTTCGTCCGTTCTACC). PCR products were separated in agarose gels and either analysed by Southern blotting, or cloned and sequenced.

RNA manipulations. Total RNA was extracted from ~10^4 rotifers with 1ml of TRIzol reagent (Invitrogen). Poly(A) fraction was prepared with Oligotex RNA Midi kit (Qiagen). Poly(A)$^+$ RNA (1 µg) was treated with DNaseI (Invitrogen), extracted with phenol/chloroform, precipitated with EtOH and reverse-transcribed with SuperScript III (Invitrogen) in the total volume of 10 µl, with or without RT added. After heat inactivation, reactions were diluted 5-fold, and 1 µl was used for PCR with Platinum Taq High Fidelity Polymerase (Invitrogen) using the same cycling conditions: 2'@94C; (20"@94C, 1'@53C, 30"@68C)x38; 5'@68C. The following pairs of primers were

used: *AvM*, CGAAGCAACGAAAACAATCA and GATAATTTCTTTCTTAATGCCG; *AvO*, ACGATATCTTCATCGCAGCA and CACAGTTCCGAAATCCAACA; *AvN* intron 1, TCGACAAAATGATGCCAAAG and CTGATTGTTTATTTGCTAACTC; *AvN* intron 3, TACGAGTCGTCCGCTTGTGT and GTGGTTGACCGGAGTTTGAC. PCR products were resolved on 1.2% LMP agarose gels, excised, digested with β-agarase 1 (NEB), extracted with phenol/chloroform, precipitated with EtOH and sequenced. For 5' RACE, poly(A)$^+$ RNA (100 ng) was used for first-strand synthesis with *Athena*-specific primers R1-AvO (CAGGAGGAGCACCAGGAAT) or R1-AvN (GATCATAATAACTTTGGTAGAGA). Upon extension, reactions were treated with RNase H and RNase T1 for 30' at 37C. Extension products were extracted with phenol/chloroform, EtOH-precipitated, and resuspended in H_2O. cDNAs were tailed with TdT (NEB) supplemented with 0.2 mM dCTP. Following heat inactivation, reactions were diluted 5-fold, and 1 µl was used for nested PCR with Platinum Taq as above, using primers RACE_AUAP (AGTGACCGTATCATTTGGCTG) and R2-AvO (GTCCTTGGCTTCAAGGTCTG) or R2-AvN (CTTTTTTCTTCTTGATTGGATGAT). PCR products were separated on agarose gels and sequenced.

Bioinformatics. The whole-genome shotgun (WGS) sequence (AACS00000000) of *Coprinus cinereus* (aka *Coprinopsis cinerea*) strain Okayama-7 #130 was produced by the Broad Institute of MIT and Harvard (http://fungal.genome.duke.edu/ and http://www.broad.mit.edu/annotation/fungi/coprinus_cinereus/). The WGS assembly (AADS00000000) of a homokaryotic *P. chrysosporium* strain RP-78 (29) and the WGS reads of *Phaeodactylum tricornutum* and *Selaginella moellendorffii* were produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/). PLEs were identified by TBLASTN searches of WGS assemblies and subsequent BLASTN searches of trace archives. Reads containing five or more telomeric repeat units were retrieved and sorted into telomeric clusters. Mate-pairs from every cluster were used as queries in BLASTN searches of WGS assemblies to verify that each cluster forms a scaffold in only one direction. A similar approach was used in Li *et al*. (30). The longest

Coprina fragments are contained in GenBank entries AACS01000397.1 (*Cc1*), AADS01000564, AY916132 (*Pc1*), and AADS01000820 (*Pc2*). Consensus sequences for *Cc1*, *Pc1*, *Pc2, Pt1, Sm1* and *Sm2* were deposited in Repbase Update (31).

Phylogenetic analysis. For phylogenetic inference, we used the region 540-1280 of the alignment shown in Supp. Fig. 6 and provided as Supporting Dataset. The best fitting model of protein sequence evolution was selected using PROTTEST 1.3 (32) among a set of 80 candidate models constituted by all combinations of the empirical amino acid substitution matrices (JTT, mtREV, mtMam, mtArt, Dayhoff, WAG, rtREV, cpREV, Blosum62, VT) with a gamma distribution with eight rate categories (+G_8), a proportion of invariable sites (+I), and observed amino acid frequencies (+F). All statistical criteria selected rtREV+I+G_8(+F) (33) as the best fitting model, with WAG+I+G_8(+F) (34) coming a close second; other models performed significantly worse. PROTTEST also calculated the observed amino acid frequencies and the rate heterogeneity parameter α. Evaluation of the phylogenetic data structure using phylogenetic networks was done with NeighborNet (35), implemented in SPLITSTREE 4.6 (36). Likelihood distance-based phylogenetic trees were inferred by applying the BioNJ algorithm (37) in SPLITSTREE 4.6 on ProteinML distances computed using the WAG model and the α and θ parameter values previously estimated by PROTTEST. Neighbor-Net networks (35) were constructed from the same distance estimates. Bootstrap proportions were also obtained from 1000 replicates using the same distance correction. Bootstrap networks were then constructed from all splits that occurred in any of the 1000 bootstrap replicates. Phylogenetic network construction allows one to visualize conflicting signals and areas of uncertainty in the dataset. The topology obtained by NEIGHBORNET was also obtained in neighbor-joining analyses by MEGA 3.1 (38) (JTT substitution model; pairwise deletion; gamma distributed rates; 100 bootstrap replications). For maximum likelihood analysis under the best fitting model, we used TREEFINDER (39) under rtREV+G_8+F, substituting the amino acid frequencies of rtREV

with observed frequencies calculated by PROTTEST. Likelihood-based statistical tests of alternative topologies were conducted with TREEPUZZLE 5.2 (40) under WAG+I+G_8+F model.

We thank M. Meselson for encouragement and support throughout the course of this work, critical reading of the manuscript, and financial support from NIH; A. Pokrovski for help with Supporting Table 2; J. Hur and J. Mark Welch for providing genomic libraries; and the U.S. National Science Foundation (MCB-0614142) for continued financial support.

1. Blackburn EH (2000) *Nat Struct Biol*. 7:847-850.

2. Eickbush TH (1997) *Science* 277:911-912.

3. Nakamura TM, Cech TR (1998) *Cell* 92:587-590.

4. Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB (2003) *Nat Genet* 33:123-124.

5. Kazazian HH Jr. (2004) *Science* 303:1626-1632.

6. Evgen'ev MB, Arkhipova IR (2005) In: *Retrotransposable elements and genome evolution*, J.-N.Volff, ed.; Karger AG: Basel. *Cytogenet. Genome Res*. 110:510-521.

7. Pyatkov KI, Arkhipova IR, Malkova NV, Finnegan DJ, Evgen'ev MB (2004) *Proc Natl Acad Sci USA* 101: 14719-14724.

8. Normark BB, Judson O, Moran N. (2003) *Biol. J. Linn. Soc*. 79: 69-84.

9. Mark Welch D, Meselson M (2000) *Science* 288:1211-1215.

10. Mark Welch JL, Meselson M (1998) *Hydrobiologia* 387/388:403-407.

11. Mark Welch D, Meselson M (2003) *Biol. J. Linn. Soc*. 79:85-91.

12. Baird DM, Rowson J, Wynford-Thomas D, Kipling D (2003) *Nat Genet* 33:203-207

13. Arkhipova IR, Meselson M (2005) *Proc. Natl. Acad. Sci. USA* 102:11781-11786.

14. Mark Welch J, Mark Welch D, Meselson M. (2004) *Proc Natl Acad Sci USA* 101:1618-1621.

15. Hur JH (2006) Ph.D. Dissertation, Harvard University, Cambridge, MA.

16. Doulatov S, Hodes A, Dai L, Mandhana N, Liu M, Deora R, Simons RW, Zimmerly S, Miller JF (2004) *Nature* 431:476-481.

17. Arkhipova IR (2006) *Syst. Biol*. 55:875-885.

18. Danilevskaya ON, Arkhipova IR, Traverse KL Pardue ML (1997) *Cell* 88:647-655.

19. Pardue ML, DeBaryshe PG (2003) *Annu. Rev. Genet*. 37:485-511.

20. Fujiwara H, Osanai M, Matsumoto T, Kojima KK (2005) *Chromosome Res*. 13:455-467.

21. Arkhipova IR, Morrison HG (2001) *Proc. Natl Acad. Sci. USA* 95:11495-11502.

22. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV (2002) *Nat Genet*. 31:159-165.

23. Zhong J, Lambowitz AM (2003) *EMBO J*. 22:4555-4565.

24. Kulpa DA, Moran JV (2006) *Nat Struct Mol Biol*. 13:655-660.

25. Moran JV, DeBerardinis RJ, Kazazian HH (1999) *Science* 283:1530-1534.

26. Rashkova S, Athanasiadis A, Pardue ML (2003) *J Virol*. 77:6376-6384.

27. Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, Hodges RS, Williams MC (2005) *J Mol Biol*. 348:549-561.

28. Doolittle WF, Sapienza C (1980) *Nature* 284:601-603.

29. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein K, Ramaiya P, Detter J, Larimer F Coutinho P, Henrissat B, Berka R, Cullen D, Rokhsar D (2004) *Nat Biotechnol*. 22:695-700.

30. Li W, Rehmeyer CJ, Staben C, Farman ML (2005) *Bioinformatics* 15:1695-1698.

31. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) *Cytogenet. Genome Res*. 110:462-467.

32. Abascal F, Zardoya R, Posada D (2005) *Bioinformatics* 21:2104-2105.

33. Dimmic MW, Rest JS, Mindell DP, Goldstein RA (2002) *J. Mol. Evol*. 55:65-73.

34. Whelan S, Goldman N (2001) *Mol. Biol. Evol*. 18:691-699.

35. Bryant D, Moulton V (2004) *Mol. Biol. Evol*. 21:255-265.

36. Huson DH, Bryant D 2006. *Mol. Biol. Evol*. 23:254-257.

37. Gascuel O (1997) *Mol. Biol. Evol*. 14:685-95.

38. Kumar S, Tamura K, Nei M (2004). Briefings in Bioinformatics 5:150-163.

39. Jobb G, von Haeseler A, Strimmer K (2004) *BMC Evol. Biol*. 4:18.

40. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) *Bioinformatics*, 18:502-504.

41. Gladyshev EG, Meselson M, Arkhipova IR (2006) *Gene* 390:136-145.

42. Lue NF, Lin YC, Mian IS (2003) *Mol. Cell. Biol*. 23:8440-8449.

FIGURE LEGENDS

Fig. 1. Flowchart of the chromosome end enrichment procedure (see Materials and Methods for details).

Fig. 2. Structural organization of telomere-associated retroelements. Each red letter **T** indicates point of PLE 5' truncation and addition of reverse-complement telomeric repeats at a chromosome end; 5' truncation points within individual copies are shown by thin diagonal lines; reverse-complement telomeric repeat units are specified for each species. Non-coding sequences are shown by a thin line; PLE ORFs by an open rectangle with the N-terminal and C-terminal domains (N,C) and the central region which includes the seven core RT motifs (RT1-RT7) and the thumb domain (TH). **J**, 5' truncation point in an upstream copy when joined to a full-length downstream copy, forming a "pseudo-LTR" (see also Supp. Fig. 7); **O**, point of addition of *Athena-O* to *Athena-N* at the O1, O3 and N1 telomeres containing both elements in the same orientation. Small red boxes mark the position of short internal telomeric repeat stretches; larger boxes mark longer tandem repeats (shown in Supporting Fig. 7); introns are denoted by triangles. Telomeric mini-library clones from telomeres M1-M2, O1-O3, N1-N2 in *A. vaga* and C, K in *P. roseola* (also listed in Supp. Table 1) are aligned with the corresponding *Athena* sequences. Also shown is the position of *Athena*-specific primers used for RT-PCR (black, paired), STELA (orange), and 5' RACE (purple) (see Fig. 3,b-d for experiments). Only *Athena* variants found at telomeres are shown; additional diverged variants were identified on

sequenced cosmids/fosmids, but have not yet been found at telomeres, and are not presented here. ♦, nuclear localization signals; cccc, coiled-coil domains; LZ, leucine zipper motif. Scale bar, 1 kb.

Fig. 3. Characterization of bdelloid *Athena* elements. **a**. Structure of telomeres M1, O3 and O4 in *Athena*-containing fosmids obtained from the *A. vaga* genomic library. Color codes and ORFs are as in Fig. 2; telomeres are in red; truncated *Athena* copies are delimited with ~ (vertical or horizontal). There are 10 and 8 48-bp repeats between *AvO* and *AvN* in the O3 and O4 telomeres, respectively. *Juno1.4* is a slightly 3' truncated copy of an LTR retrotransposon in an inverse orientation (41). Scale bar, 1 kb. **b**. Single telomere length analysis (STELA). The rationale (12) is shown on the top: a telorette oligo is annealed to the G-rich overhang and, following ligation, a specific telomere is amplified with the teltail primer and the primer in the subtelomeric region. The EtBr-stained gel shows amplification of telomeres M and O with the corresponding *Athena* primers (Fig. 2a; Methods); below is the same gel probed with $(TGAGGG)_4$ for visualization of telomeric repeat-containing amplicons. As a control, lanes marked (Telorette -) contained no telorette oligos in the ligation mix. Amplification of telomeres M1, O1 and O3 was confirmed by cloning and sequencing of total PCR products. **c**. Rapid amplification of cDNA ends (5' RACE) for *AvN* and *AvO*. Arrows indicate the position of RNA start sites relative to ORF1, obtained by sequencing of the corresponding amplicons. The level of *AvM* transcription (d) was insufficient to generate a RACE product. **d**. RT-PCR of *A. vaga* poly$(A)^+$ RNA with *AvM*, *AvO* and *AvN* primers (see Fig. 2a). All

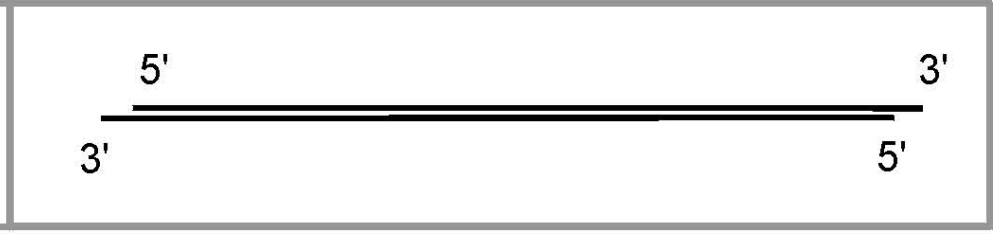
upper bands correspond in sequence to the unspliced message; lower bands are spliced at the predicted intron boundaries (*AvN*) or result from cryptic splicing (*AvO*).

Fig. 4. Bootstrap network of 46 PLE and TERT sequences based on maximum-likelihood (ML) distances estimated with a WAG substitution matrix plus an eight-category gamma rate heterogeneity correction. The dataset included 700 characters from the core RT and its N-terminal and C-terminal extensions (Supp. Fig. 6). A 370-aa RT fragment of an early-branching PLE was found in the slime mold, *Physarum polycephalum* (Amoebozoa), but no evidence is yet available for its association with telomeres because of insufficient genome coverage. This PLE contains an insertion between motifs RT3 and RT4 called IFD (insertion into the fingers domain), which is found only in TERTs and is important for TERT function, apparently stabilizing very short DNA-RNA hybrids (42). EN(-) retroelements shown in Fig. 2 (*AvM, AvO, AvN, PrR, Cc1, Pc1, Pc2, Pt1, Sm1, Sm2*) are underlined; EN(+) indicates the presence of EN domain in *Neptune* and *Poseidon/Penelope* groups (full element and species names are given in Supp. Fig. 6). The *Coprina* group may or may not be monophyletic. Triangle indicates the midpoint. For clade support values, see Supp. Fig. 9B.

Fig. 5. Model for endonuclease-independent terminal retrotransposition. Red, retroelement sequences; blue, chromosomal DNA; pale ovals, proteins that normally form caps at the telomeres; RT, reverse transcriptase; TERT, telomerase. Priming at the G-rich 3' overhang is facilitated by annealing with

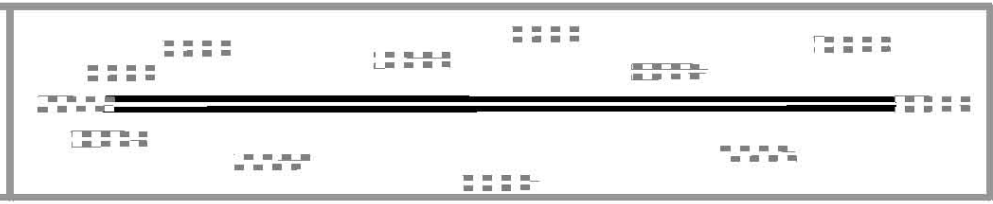
reverse-complement telomeric repeats in the 3' UTR of the RNA template. In the absence of telomeric repeats, annealing at microhomologies could be assisted by ORF1. Telomeric repeats are added by telomerase, after which the normal capping structure is restored. Note that the second-strand synthesis would not require special mechanisms other than routine DNA replication as occurs during C-rich strand synthesis. Not to scale.

Intact animals are digested in agarose, DNA is purified on a pulse-field gel and stored in agarose blocks



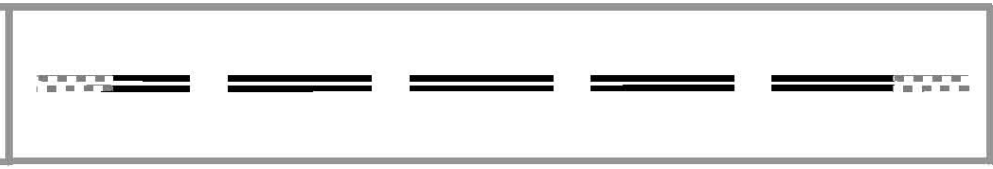
DNA (200-300 ng) is blunt-ended in agarose blocks with T4 DNA Pol

Vector (dashed gray) is ligated in excess to chromosomal ends

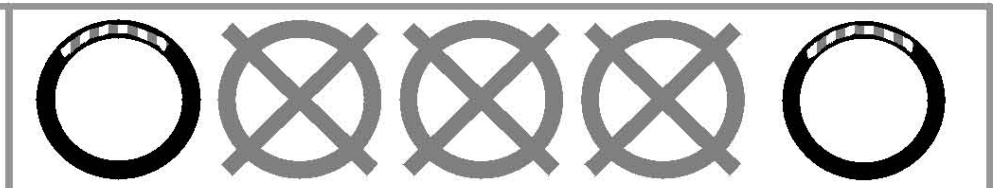


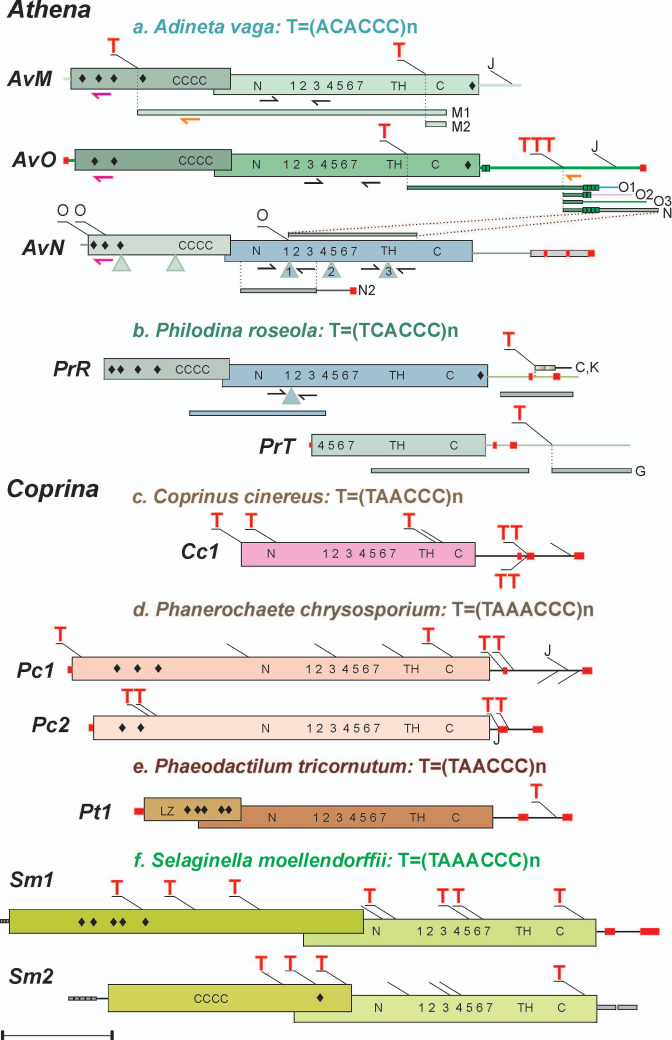
Unligated vector is removed *via* gel electrophoresis

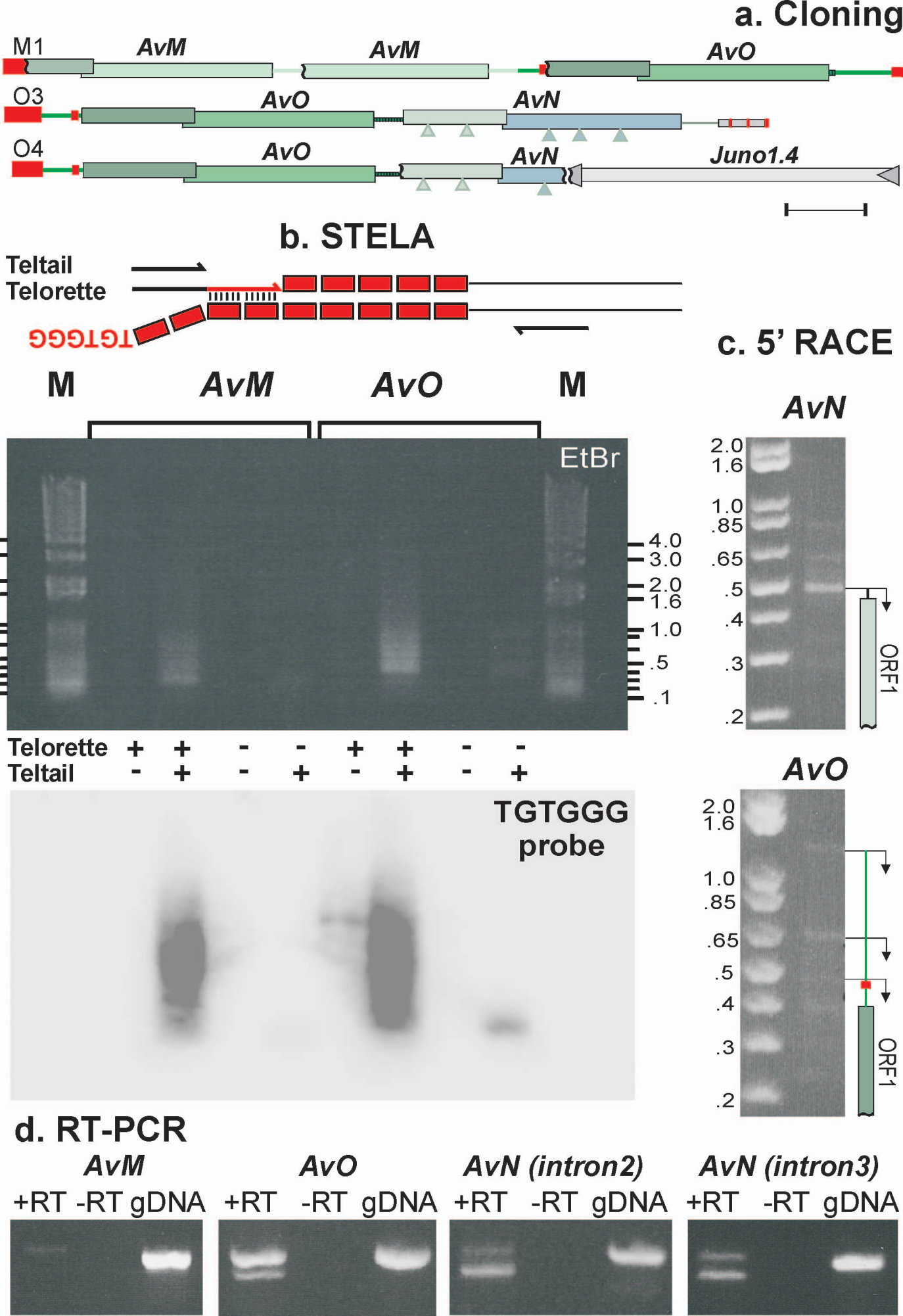
DNA is digested with a blunt six-cutter restriction enzyme

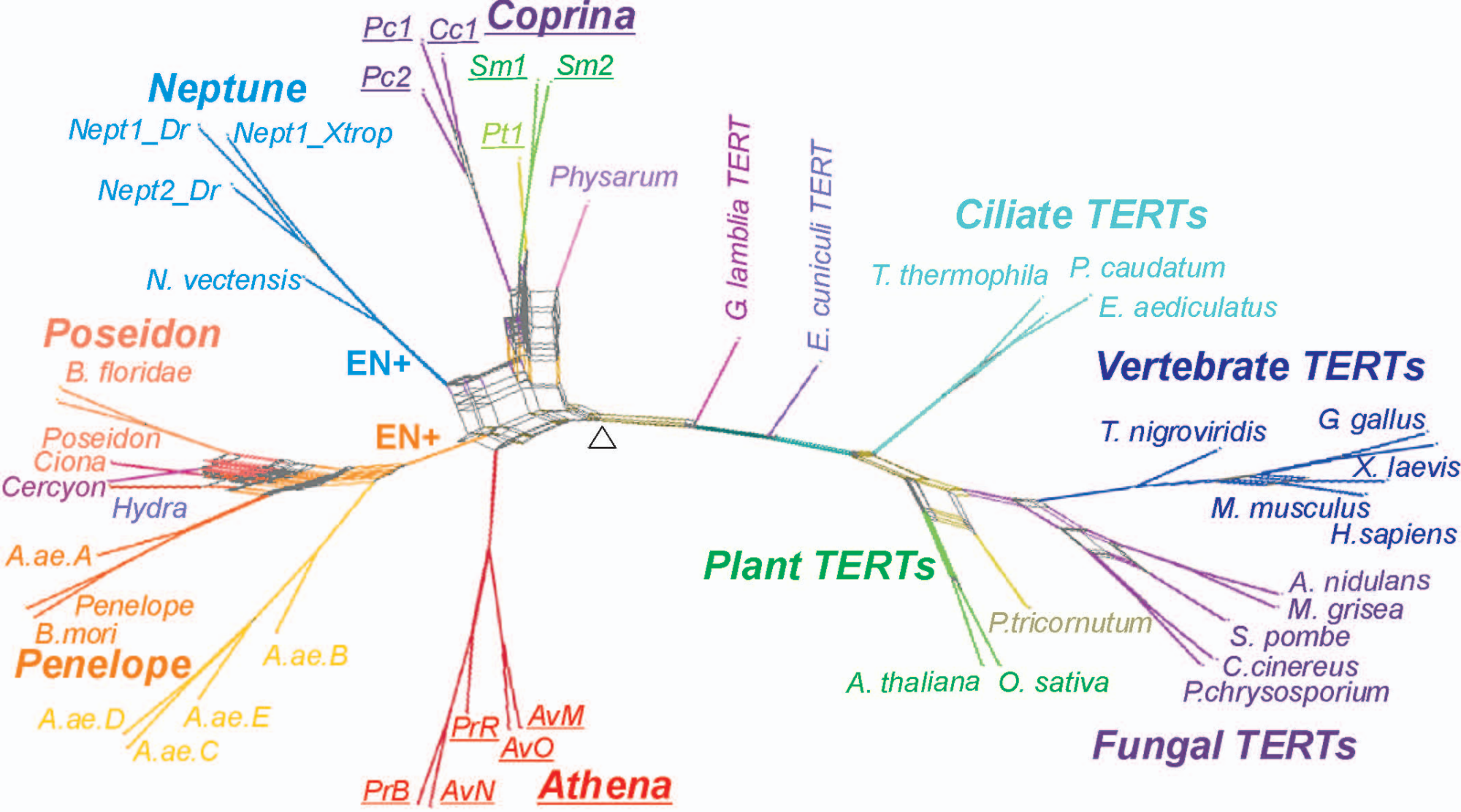


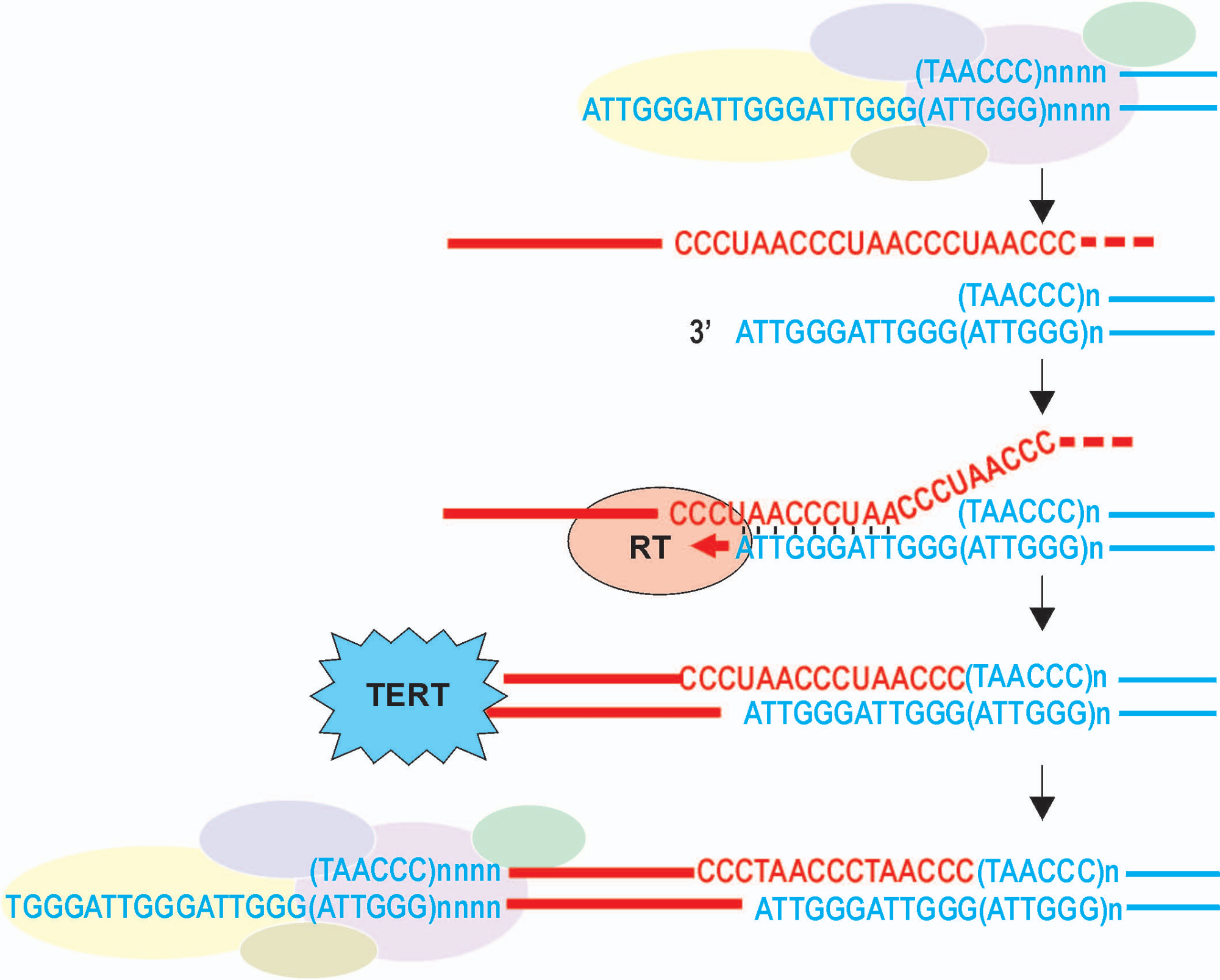
Restriction fragments are circularized to transform *E.coli* DH10B strain




Athena
a. Adineta vaga: T=(ACACCC)n
AvM
CCCC
N 1 2 3 4 5 6 7 TH C
J
M1
M2
AvO
TTT
O1 O2 O3
N
AvN
O O
O
1 2 3
N2
b. Philodina roseola: T=(TCACCC)n
PrR
C,K
PrT
4 5 6 7 TH C
G
Coprina
c. Coprinus cinereus: T=(TAACCC)n
Cc1
TT
d. Phanerochaete chrysosporium: T=(TAAACCC)n
Pc1
Pc2
e. Phaeodactilum tricornutum: T=(TAACCC)n
Pt1
LZ
f. Selaginella moellendorffii: T=(TAAACCC)n
Sm1
Sm2


a. Cloning
M1
AvM
AvM
AvO
O3
AvO
AvN
O4
AvO
AvN
Juno1.4
b. STELA
Teltail
Telorette
TGTGGG
c. 5' RACE
M
AvM
AvO
M
EtBr
4.0
3.0
2.0
1.6
1.0
.5
.1
Telorette + + - - + + - -
Teltail - + - + - + - +
TGTGGG probe
AvN
2.0
1.6
1.0
.85
.65
.5
.4
.3
.2
ORF1
AvO
2.0
1.6
1.0
.85
.65
.5
.4
.3
.2
ORF1
d. RT-PCR
AvM
+RT -RT gDNA
AvO
+RT -RT gDNA
AvN (intron2)
+RT -RT gDNA
AvN (intron3)
+RT -RT gDNA


Coprina
Pc1
Cc1
Pc2
Sm1
Sm2
Neptune
Nept1_Dr
Nept1_Xtrop
Pt1
Physarum
Nept2_Dr
N. vectensis
G lamblia TERT
E. cuniculi TERT
Ciliate TERTs
T. thermophila
P. caudatum
E. aediculatus
Poseidon
B. floridae
EN+
Vertebrate TERTs
EN+
T. nigroviridis
G gallus
X. laevis
Poseidon
Ciona
Cercyon
Hydra
M. musculus
H.sapiens
A.ae.A
Plant TERTs
A. nidulans
Penelope
B.mori
M. grisea
P.tricornutum
S. pombe
Penelope
A.ae.B
C.cinereus
A. thaliana
O. sativa
P.chrysosporium
A.ae.D
A.ae.E
A.ae.C
PrR
AvM
AvO
Fungal TERTs
PrB
AvN
Athena


(TAACCC)nnnn
ATTGGGATTGGGATTGGG(ATTGGG)nnnn
CCCUAACCCUAACCCUAACCC
(TAACCC)n
3’ ATTGGGATTGGG(ATTGGG)n
CCCUAACCCUAACCCUAACCC
(TAACCC)n
RT
ATTGGGATTGGG(ATTGGG)n
TERT
CCCUAACCCUAACCC(TAACCC)n
ATTGGGATTGGG(ATTGGG)n
(TAACCC)nnnn
TGGGATTGGGATTGGG(ATTGGG)nnnn
CCCTAACCCTAACCC(TAACCC)n
ATTGGGATTGGG(ATTGGG)n