Giant Reverse Transcriptase-Encoding Transposable Elements at Telomeres

Irina R. Arkhipova,*[,1] Irina A. Yushenova,[1] and Fernando Rodriguez[1]

[1]Marine Biological Laboratory, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Woods Hole, MA

*Corresponding author: E-mail: iarkhipova@mbl.edu.

Associate editor: Jeffrey Townsend

Abstract

Transposable elements are omnipresent in eukaryotic genomes and have a profound impact on chromosome structure, function and evolution. Their structural and functional diversity is thought to be reasonably well-understood, especially in retroelements, which transpose *via* an RNA intermediate copied into cDNA by the element-encoded reverse transcriptase, and are characterized by a compact structure. Here, we report a novel type of expandable eukaryotic retroelements, which we call *Terminons*. These elements can attach to G-rich telomeric repeat overhangs at the chromosome ends, in a process apparently facilitated by complementary C-rich repeats at the 3′-end of the RNA template immediately adjacent to a hammerhead ribozyme motif. *Terminon* units, which can exceed 40 kb in length, display an unusually complex and diverse structure, and can form very long chains, with host genes often captured between units. As the principal polymerizing component, *Terminons* contain *Athena* reverse transcriptases previously described in bdelloid rotifers and belonging to the enigmatic group of *Penelope*-like elements, but can additionally accumulate multiple cooriented ORFs, including DEDDy 3′-exonucleases, GDSL esterases/lipases, GIY-YIG-like endonucleases, rolling-circle replication initiator (Rep) proteins, and putatively structural ORFs with coiled-coil motifs and transmembrane domains. The extraordinary length and complexity of *Terminons* and the high degree of interfamily variability in their ORF content challenge the current views on the structural organization of eukaryotic retroelements, and highlight their possible connections with the viral world and the implications for the elevated frequency of gene transfer.

Key words: **retrotransposons, bdelloid rotifers, hammerhead ribozymes, horizontal gene transfer.**

Introduction

Transposable elements (TEs) are segments of DNA with the ability to relocate within or between genomes, which is conferred by the element-encoded enzymatic functions. Traditionally, TEs are divided into two major classes: class I (retrotransposons) code for a reverse transcriptase (RT) capable of making a cDNA copy of the template RNA, which serves as a transposition intermediate; and class II (DNA TEs) code for a transposase, which can mobilize DNA in the absence of RNA intermediates (Finnegan 1989; Wicker et al. 2007; Kapitonov and Jurka 2008). Retrotransposons are in turn subdivided into four subclasses: LTR (long terminal repeat) retrotransposons, which are closely related to retroviruses; nonLTR retrotransposons, also called LINEs; DIRS, or YR-retrotransposons; and *Penelope*-like elements (PLEs). These subclasses are phylogenetically distinct, and their RTs usually operate together with the respective types of coencoded phosphotransferase/endonuclease (EN) DNA-cleaving enzymes: IN (DDE-integrase); APE (apurinic-apyrimidinic EN) or REL (restriction enzyme-like EN); YR (tyrosine recombinase); and GIY-YIG (nickase initially identified in prokaryotic group I introns). It is the concerted action of the RT and the phosphotransferase that determines the retroelement's ability to insert into internal genomic locations. Enzymatically active domains are typically fused into a single polyprotein called *pol*, which may undergo proteolytic processing or function as a multi-domain protein. Formation of a ribonucleoprotein (RNP) particle is ensured by the structural ORF1 (*gag*), which is usually separated from the downstream ORF2 (*pol*) by a programmed ribosomal frameshift or by in-frame stop codons. Downstream of *pol*, many retrovirus-like TEs have incorporated *env* genes of various origins coding for envelope glycoproteins responsible for membrane fusion and interaction with cell surface receptors during viral entry and egress. Two families of RT-containing viruses, hepadnaviruses and caulimoviruses (collectively called pararetroviruses), differ from retroviruses in encapsidating their DNA instead of RNA, and do not regularly integrate into chromosomes (Glebe and Bremer 2013; Hohn and Rothnie 2013).

Penelope-like elements (PLEs) are an enigmatic group of retroelements whose RTs share a common ancestor with telomerase reverse transcriptases (TERTs) (Arkhipova et al. 2003). Canonical PLEs are 3–4 kilobases (kb) in length; are framed by terminal repeats called pLTRs, which may be either direct or inverted; encode an RT with a C-terminal GIY-YIG EN domain; and yield target-site duplications (TSD) of variable length upon insertion (Evgen'ev and Arkhipova 2005). Of special interest is the unique group of PLE RTs named *Athena* (Gladyshev and Arkhipova 2007). Previously described *Athena* retroelements are 4–6 kb in length; are

phylogenetically distinct from canonical *Penelope* RTs; do not carry EN domains; and contain stretches of telomeric repeats at the junctions with host DNA. Such EN-deficient RTs are found at or near telomeres in many basidiomycete fungi and in a few plants and protists, but are particularly abundant at telomeres of bdelloid rotifers (Gladyshev and Arkhipova 2007).

Bdelloid rotifers are microscopic freshwater invertebrates that reproduce clonally, are highly resistant to desiccation and ionizing radiation, and contain numerous horizontally transferred genes in their genomes (Gladyshev and Meselson 2008; Gladyshev et al. 2008; Mark Welch et al. 2008). Genome sequencing of the first bdelloid representative, *Adineta vaga*, revealed that over 8% of its genes originate from bacteria, fungi, plants, or protists (Flot et al. 2013). Known TE families make up over 3% of the *A. vaga* genome, and are characterized by low copy numbers and high family diversity. Recently, we described canonical *Penelope* retrotransposons from *A. vaga*, which integrate into internal chromosomal locations with the aid of the C-terminal GIY-YIG EN domain (Arkhipova et al. 2013). Here we investigate *Athena*-containing retroelements in *A. vaga*, compare their organization in related species separated by tens of millions of years, and discover that they possess an extraordinarily complex structure not yet described in retroelements. We also uncover the basis for their affinity to telomeres and identify putative *cis*-acting elements that may play a role in mobilizing genes of foreign origin and members of multigene families.

Results

Athena RTs Belong to Giant Transposable Units Spanning Tens of kb and Encoding Multiple ORFs

We first sought to verify the boundaries of *Athena* retroelements in the *A. vaga* genome assembly. The commonly used TE detection pipelines perform poorly on *A. vaga* due to overabundance of low-copy-number families with one or two members (Flot et al. 2013). Most of the computer-generated *Athena* consensi were represented by RT and some adjacent sequences, but their boundary verification was far from straightforward. Specifically, while the 3′ boundary, at least in some families, was relatively easy to define from comparison between inserts, the 5′ boundaries were mostly formed by variably positioned 5′-truncations of apparently longer units, which included a variety of ORFs shared by some families but different in others. All other bdelloid TEs (LTR, nonLTR, DNA TEs) form well-defined host-TE boundaries (Flot et al. 2013).

To facilitate boundary definition, we employed small RNA coverage as a proxy for delimiting host-TE junctions (El Baidouri et al. 2015). *Piwi*-interacting RNAs (piRNAs) are a class of small RNAs typically expressed from specialized loci termed piRNA clusters, which in many genomes are composed of multiple adjacent TEs or their fragments, and ensure silencing of homologous TEs in the germ line (Weick and Miska 2014). Our piRNA libraries, as expected, were highly enriched in known *A. vaga* TEs (Rodriguez and Arkhipova 2016). Notably, for most *Athena* RTs, we observed that piRNA coverage extends well beyond the RT ORF (fig. 1). However, inspection of RT flanks did not reveal any sequences that might correspond to adjacent TEs in a piRNA cluster. Instead, the zone of piRNA coverage includes a multitude of densely spaced ORFs, which display the same polarity as *Athena* RTs, and apparently constitute parts of very large transposable units.

The most surprising observation is the length of these units, which can exceed 40 kb, far more than any of the known retroelement types. A highly variable and diversified gene content is also not typical of retroelements, which display relatively simple and well-defined ORF composition (e.g., *gag-pol-env* in LTR retrotransposons, or *gag-pol* in nonLTR retrotransposons). Collectively, these observations indicate that *Athena*-containing units represent a previously undescribed type of TEs, and justify further inquiry into their characteristics.

5′- and 3′-Boundaries Define the Giant Terminon Units

In three previously described *Athena* families, the 3′- and 5′-boundaries were formed by short stretches of species-specific reverse-complement telomeric repeats (Gladyshev and Arkhipova 2007). A genome-wide inventory of host-TE boundaries near *Athena*-like RTs in *A. vaga* reveals that, although the immediate RT environment does not always include such repeats, they are invariably found at the actual TE-host junction (supplementary table S1, Supplementary Material online). In other words, the RT is not always positioned near the 3′-end of the entire unit, so that a series of intervening ORFs may appear between the RT and telomeric repeats at the host-TE boundary. The 5′-boundaries in most cases are also formed by stretches of telomeric repeats capping 5′-terminally truncated copies. The *Athena-M* family is somewhat of an exception: out of six contigs, only one had $(ACACCC)_2$ at the junction between the 3′-pLTR and the downstream *Athena-M* copy (supplementary table S1, Supplementary Material online).

Since terminal addition is the only plausible mechanism that can account for the presence of telomeric repeats at both 3′- and 5′-termini (see below and Discussion), we further refer to the giant *Athena*-containing transposable units as *Terminons*, reflecting their capacity to attach to chromosome ends. Indeed, we were unable to find a TE or a host gene interrupted by a *Terminon* insertion. On the contrary, we observed multiple cases in which part of a preexisting TE or gene was irreversibly lost by truncation, with subsequent addition of telomeric repeats and *Terminon* attachment (supplementary table S1 and figs. S1*A* and S6, Supplementary Material online). Remarkably, the added *Terminon* units can extend telomeres by tens of kb at a time. Such additions can effectively counteract the ongoing terminal erosion, the dynamic nature of which is seen from comparison of the same *A. vaga* telomere at different points in time: telomere M1 (Gladyshev and Arkhipova 2007) from a 2006 fosmid library, compared to the corresponding region in the genome of the same clonal culture in 2010 (Flot et al. 2013), underwent loss of the distal 11-kb chain of *Ath-M* and *Ath-O*

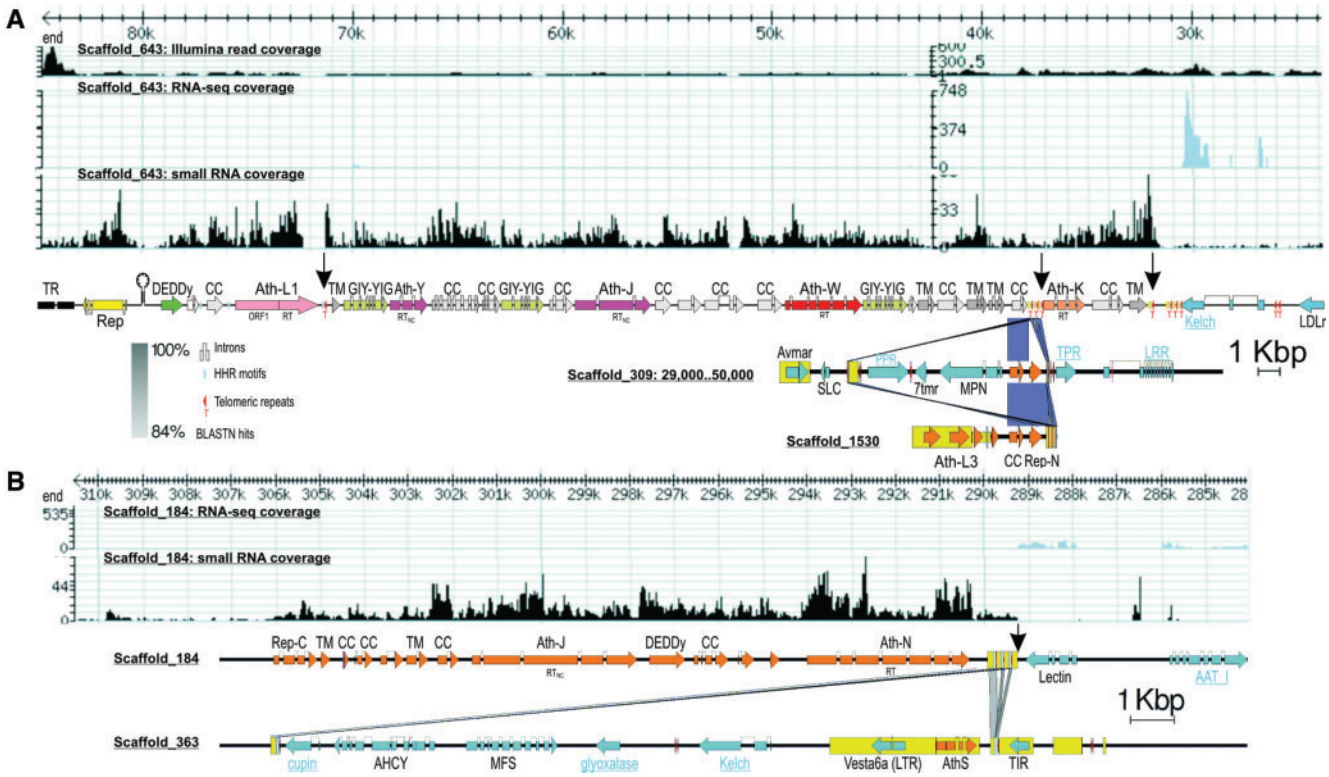

FIG. 1. Genomic organization of selected *A. vaga* scaffolds. Screenshots from the genome browser show custom tracks, as indicated, along scaffold 643 (*A*) and scaffold 184 (*B*). RT_{NC}, catalytically inactive RT domains. Unit boundaries are marked by arrows. Other TEs on scaffolds are shown by boxed arrows. Names of nonmetazoan genes are underlined. The 3′-end of scaffold 643 is aligned to scaffolds 309 (three host genes flanked by JW 3′-ends in inverted orientation) and 1530 (JW 3′-end) to illustrate the capture of host genes; similar pairwise BLASTN hits are shown for scaffolds 184 and 363. Gene abbreviations: LDLr, low density lipoprotein receptor; LRR, leucine-rich repeat protein; PPR, pentapeptide repeat protein; TPR, tetratricopeptide repeat protein; 7tmr, seven-transmembrane domain receptor; MFS, major facilitator superfamily transporter; AHCY, adenosylhomocysteinase; MPN, MPN superfamily deubiquitinase; AAT_I, putrescine aminotransferase.

elements and unique host genes (as validated by PCR), and ends with a stretch of telomeric repeats added *de novo* to the proximal *Ath-Y* (supplementary fig. S1*B*, Supplementary Material online).

The 3′ ends of *Terminons* are often present in higher copy numbers than the 5′ ends, a pattern observed in nonLTR retrotransposons, which display well-defined 3′ ends and frequent 5′-truncations (Wei et al. 2001; Hayashi et al. 2014). As seen in supplementary figure S2, Supplementary Material online, intragenomic *Terminon* replication follows the master-copy model, whereby shorter copies are derived from the longest resident copy. Indeed, WJ and W2 families propagated a subset of internally deleted copies, which is indicative of *trans*-complementation.

Nomenclature and Phylogenetic Relatedness of Athena RTs

To classify families with highly variable gene content, we relied on phylogenetic relationships of the element-encoded RTs, which are the longest and the most conserved ORFs in each unit. The phylogram in figure 2 (and an expanded codon-based version in supplementary fig. S3*A*, Supplementary Material online) shows that *Athena*-like RTs are subdivided into three major clades with intact catalytic RT motifs: ILOM (I, L, O, and M families); NT (N, K, Q, P, R, and S/T families); and W (W, W1, and W2 families). The additional JVX and Y clades (supplementary fig. S3*A*, Supplementary Material online) are formed by inactive RT derivatives (J, X, V, Y, see below), which lack essential catalytic residues, and are colocalized on the same unit with a catalytically intact RTs (W, N, S/T, or K). Interestingly, *Athena* RTs from *Philodina roseola*, a species from the bdelloid family Philodinidae (Gladyshev and Arkhipova 2007), can be confidently placed into N and W clades defined by RTs from *A. vaga* (supplementary fig. S3*A*, Supplementary Material online), indicating that the split between clades predates the divergence of bdelloid families (>60 Mya) (Mark Welch et al. 2008).

Table 1 presents the inventory of *A. vaga* families, with the scaffold harboring the longest and most intact representative designated as the reference scaffold for each family. Members of each family share common features and display between-clade variability in the RT ORF structure, manifested in the presence/absence of programmed ribosomal frameshifts, commonly found in retrovirus-like elements; gain and loss of introns; and acquisition of RT derivatives lacking catalytic residues.

Frameshifts

All members of the ILOM clade are characterized by a −1 programmed ribosomal frameshift, allowing the RT to be

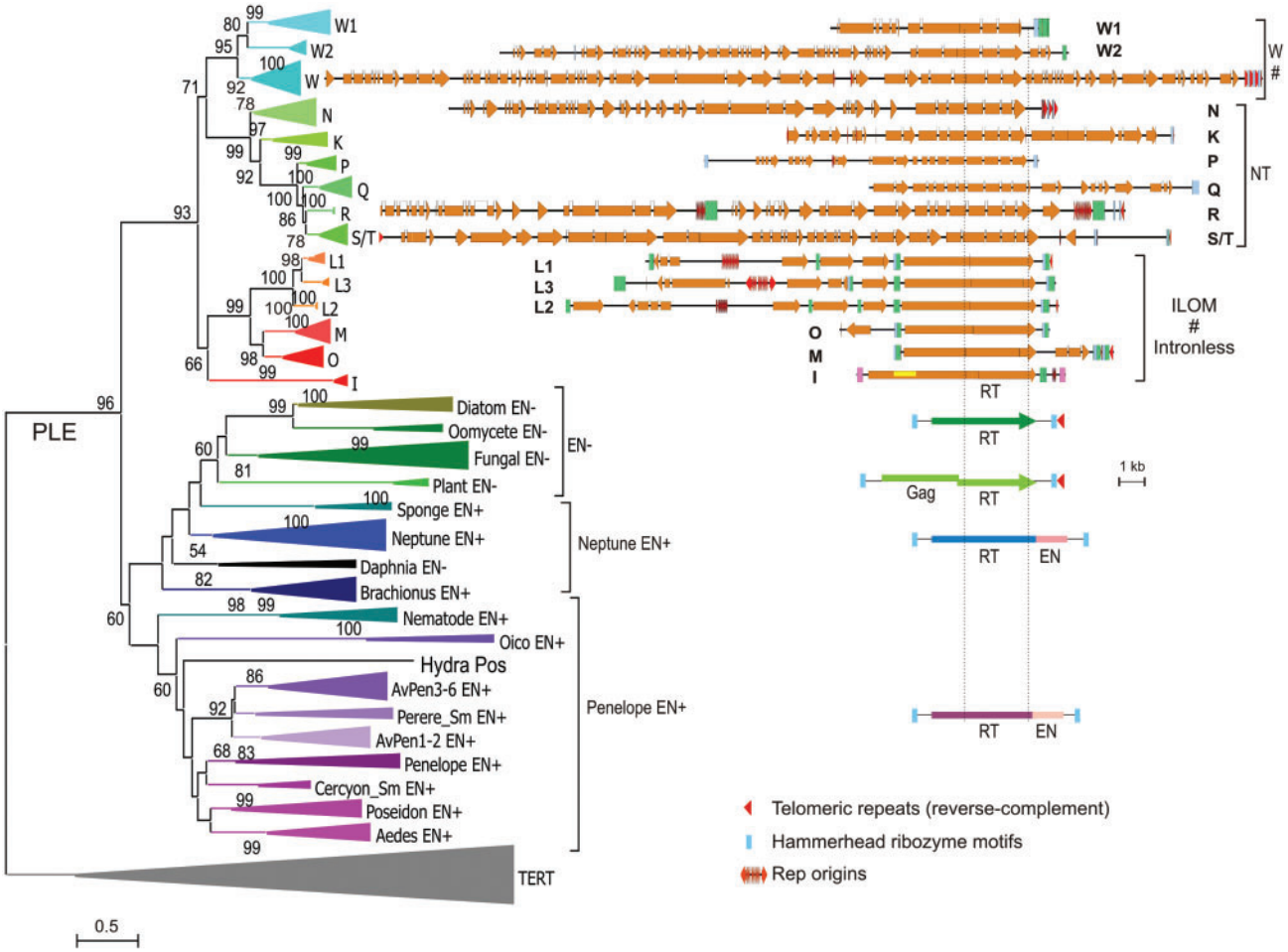

Fig. 2. Structural organization and phylogenetic relationships of *Penelope*-like elements (PLE). Shown is the maximum likelihood phylogram based on amino acid sequences of catalytically intact RTs, with TERTs as an outgroup. Structural diagrams are centered on the core RT domain framed by thin dotted lines. Major clades (W, NT, and ILOM) are marked with a bracket; clades with the programmed −1 frameshift are marked by #. PLE clades are as in (Arkhipova 2006; Arkhipova et al. 2013; Lin et al. 2016). Scale bar, amino acid substitutions per site. Detailed ORF composition for each family is shown in supplementary figure S6, Supplementary Material online.

expressed as a fusion with ORF1, which mostly consists of coiled-coil (CC) motifs. The −1 frameshift has a canonical structure, formed by a heptanucleotide "slippery sequence" (typically T_6G or T_6C, translating into consecutive Phe residues), and a downstream pseudoknot or hairpin (Caliskan et al. 2015) (supplementary fig. S4, Supplementary Material online). The frameshift site exhibits conservation in an otherwise rapidly evolving sequence context (supplementary fig. S4*A* and *B*, Supplementary Material online). It is also present in all members of the W clade (supplementary fig. S4*A*–*C*, Supplementary Material online), and in the catalytically inactive X/V ORFs from the JVX clade (supplementary fig. S4*B* and *D*, Supplementary Material online). The NT clade lacks a programmed frameshift and contains a conserved intron near this position.

Introns

Unlike other retroelements, PLEs, especially *Athenas*, possess an unusual ability to accumulate and retain spliceosomal introns (Arkhipova et al. 2003). Members of the NT clade have accumulated the largest number of introns, harboring 4–9 introns each (supplementary figs. S3*A* and S5, Supplementary Material online). Intron positions are highly conserved in the core motifs RT1-2 and RT5 (Xiong and Eickbush 1990), and an additional intron appears in the conserved NGY motif of the RT thumb domain in the NT clade. Members of the W clade have the frameshift and either 2 or 4 introns. In the JVX clade, V and X have the frameshift and either one (V) or two (X) introns, while J lacks frameshifts and contains 2 or 4 introns. Even in the poorly conserved N-terminal (ORF1) moiety, one of the intron positions is conserved between JVX, W, and NT clades, while two other positions are specific either for NT clade or for J/W clades (supplementary figs. S3 and S7, Supplementary Material online). Intron acquisition can be followed by occasional intron losses, as follows from the intron presence–absence mapping on the phylogenetic tree (supplementary fig. S3*A*, Supplementary Material online). All members of the frameshifted ILOM clade are intronless.

Table 1. ORF Content in *Terminon* Families.

Family	Reference Scaffold	RT_{CAT}	RT_{NC}	DEDDy	GIY-YIG	Rep	CC_{JVX}	CC_{NWT}	TM	Other known	UnknownORFs	HHR
I	1009	I[a,b]	—	—	—	—	—	—	—	GDSL(i)	—	—
M	1009	M[b]	—	—	—	—	—	—	—	—	1 (a)	I
O	574	O[b]	—	—	—	—	—	sCC	—	—	—	I
L1	1351-560	L1[b]	—	1-nc	—	—	—	sCC	1	GDSL(a)/	—	I
L2	791	L2[b]	V[b]	2-nc	—	—	—	sCC	1	GDSL(a)	—	I
L3	643	L3[b]	—	1-nc	—	1(a)	—	sCC	1	—	1	I
K	868	K	Y	—	—	—	1	1	—	—	1	I.t
JN	1362-184	N	J	1	1	Rep-C	1	1	2	—	JN1-2	I.t
WJ	1477-401	W[b]	J	1	2	RepNc	1	1	3	sCC	JN2,JD1	I.t
W	560	W[b]	—	1-nc	1	1	1	1	2	2 sCC	—	I.t
JW	643	W[b]	J; Y	—	3	—	2	2	6	2 sCC	1	I.t
W2	721-1061	W[b]	—	—	2	—	1	1	—	2 sCC	4	I.t
W1	14; 660	W[b]	—	—	—	1(a)	—	—	—	—	—	I.t
R	660	R	J	—	1	Rep-C	1	1	—	3 sCC	JR1-3	I.t
P	1085; 588	P	—	—	—	RepNc	1	—	1	—	—	I.t
Q	373	Q	—	—	—	Rep-C	1	sCC	3	—	QD1-3	I.t
S	1059-574	S	X[b];V[b]	1	3	—	2	1	1	Zn-ribbon	2	I.t
T	587	T	X[b];V[b]	1	3	—	2	sCC	1	Zn-ribbon	1 (a)	I.t

NOTE.—CAT, ORFs with catalytic residues; NC, ORFs with no catalytic residues;/, 5′- or 3′-terminal truncations; (a), ORF in antisense orientation to other ORFs; (i), internal domain within ORF; CC, coiled-coil motif-containing ORFs (>200 aa); sCC, small coiled-coil motif-containing ORFs (<200 aa); TM, transmembrane domain-containing ORFs; RepNc, N-terminal pseudo-catalytic domain of Rep; Rep-C, C-terminal SF3 helicase domain of Rep; I, Type I HHR motifs with telomeric repeats 50–100 nt downstream; I.t, Type I HHR with immediately adjacent telomeric repeats.
[a]ORF with a defect in all family members.
[b]ORF with a programmed −1 ribosomal frameshift.

Catalytic Residues

In the JVX clade, *Athena*-derived ORFs are characterized by complete loss of the RT catalytic residues: the core palm motifs RT3-5 (or A–C), encompassing the DDD catalytic triad, along with finger motifs RT1-2, are wiped out in the context of an otherwise intact, intron-containing ORF, rendering the RT domain catalytically inactive (supplementary fig. S5, Supplementary Material online). While their roles evidently cannot involve catalysis, these RT derivatives should have the potential to interact with a catalytically active RT (from W or N/T clades), which is usually present on the same unit (see below). The highly diverse Y clade (supplementary fig. S3*A*, Supplementary Material online) entirely lacks the N-terminus corresponding to ORF1, and contains barely recognizable RT derivatives.

ORF Content and Directionality

In each *Terminon* family, RTs and the associated CC-ORFs represent the obligatory components of these units (table 1; supplementary fig. S5, Supplementary Material online). However, *Terminons* can also harbor other ORFs, such as: Rep proteins most closely related to geminiviruses (circular ssDNA viruses of plants) (Hanley-Bowdoin et al. 2013); RNase D-like DEDDy-type endonucleases (Zuo and Deutcher 2001); GIY-YIG endonucleases resembling those in giant dsDNA viruses or virophages (Dunigan et al. 2012; Belfort and Bonocora 2014); GDSL esterases/lipases (Akoh et al. 2004) (in L1–L3; also found as part of ORF1 in the I family); stand-alone ORFs with one or more coiled-coil motifs (CC-ORFs); smaller ORFs with one or two transmembrane domains (TM-ORFs); and a few hypothetical ORFs of unknown origin (table 1; supplementary figs. S3, S6, and S7, Supplementary Material online).

Overall, each *Terminon* family is characterized by a core set of genes, all of which, however, are not necessarily present in each family (table 1). The intronless ILOM clade has the simplest ORF composition, with the most complex L3 family containing ORFs for a Rep (oppositely oriented), DEDDy endonuclease, two CC-ORFs, and a TM-ORF, while the M and O families encode only one additional ORF each (fig. 2; supplementary fig. S6*D*, Supplementary Material online). The most diverse ORF composition is observed in the intron-containing W and NT clades, where some of the ORFs can occur in more than one variant per unit. Each additional ORF (Rep, DEDDy, GIY-YIG, CC, and TM) can be intronless or may contain introns in conserved positions (0–7 introns per ORF, as in GIY-YIG or Rep ORFs) (supplementary figs. S3*B* and *C*, S5, and S6*A*–*C*, Supplementary Material online). With a few exceptions (see below), most ORFs are cooriented, as in figure 1. Such unidirectionality facilitates rapid assessment of *Terminon* boundaries at-a-glance, since the ORFs in the adjacent host DNA are distributed between Watson and Crick strands. The enzymatic potential of extra ORFs is described in the next section.

ORFs with Enzymatic Functions and their Nonenzymatic Derivatives

DEDDy/DEDDh Single-Stranded 3′-Exonucleases

Eight *Terminon* families (table 1) contain ORFs with homology to the DEDD-type (or DnaQ-like) 3′–5′ exonuclease domain, which has three conserved sequence motifs (ExoI, ExoII, and ExoIII) with four acidic residues serving as ligands for the two metal ions required for catalysis (Zuo and Deutcher

2001). Two variants of the ExoIII motif are known: YX_3D (DEDYD) and HX_4D (DEDHD). These exonucleases perform 3′-end processing of structured RNAs (RNase D, RNase T, exosome subunit Rrp6), but may also act on single-stranded DNAs (WRN, DnaQ, and proofreading subunits of A- and B-type DNA polymerases). Six of the *Terminon*-encoded exonucleases are of the DEDYD-type, with the second D replaced with E (DEEYD); the remaining three derivatives in L1–L3 families lack the acidic residues, indicating catalytic inactivity. An additional EHCHC motif in all families, which is known to coordinate Zn^{2+} binding in Maelstrom proteins involved in piRNA biogenesis (Chen et al. 2015), suggests conservation of the RNA binding function. DEDDy exonucleases (ExoN) have also been found in nidoviruses, and were hypothesized to have a proofreading function in these large (+)ssRNA viruses (Ulferts and Ziebuhr 2014). Both DEDDy-like and GDSL-like *Terminon* ORFs exhibit similarity to ORF3's in bdelloid LTR retrotransposons (Rodriguez et al. 2017).

GDSL Esterases of the SGNH Hydrolase Family

GDSL esterases/lipases are hydrolytic enzymes with broad substrate specificity (Akoh et al. 2004). They contain five conserved blocks with a G-D-S-L or similar sequence with the catalytic Ser in the first block, and are also called SGNH hydrolases, after the letters specifying the invariant catalytic S, G, N, and H residues in the conserved blocks I, II, II, and V, respectively. Most members of the L1–L3 families encode these ORFs (table 1), which are similar to PC-esterases (pfam13839), enzymes predicted to have acyl esterase activity modifying cell-surface biopolymers such as glycans and glycoproteins (Anantharaman and Aravind 2010). However, the hydrolytic function of these ORFs must be impaired, as the Ser residues of the catalytic S-N-H triad are lacking. GDSL-esterase-like ORFs in a subset of nonLTR retrotransposons (CR1, RTEX, and BRIDGE1), which differ from PC-esterases, are thought to interact with membrane glycoproteins to facilitate entry or exit (Kapitonov and Jurka 2003; Schneider et al. 2013), and at least some of them possess hydrolytic activity, while others lack such activity but preserve binding properties (Montanier et al. 2009). In the viral world, analogous ORFs encode hemagglutinin-esterase fusion glycoproteins in ssRNA viruses, for example, orthomyxoviruses (influenza C) and coronaviruses (Zeng et al. 2008). Notably, the GDSL domain is found in the I family as an integral part of ORF1, although it may also lack catalytic activity (supplementary figs. S5 and S6*D*, Supplementary Material online).

GIY-YIG EN-Containing ORFs

GIY-YIG EN are nickases, with a single active site that hydrolyzes DNA by a one-metal ion mechanism, but can also generate double-strand breaks if DNA is nicked sequentially (Kleinstiver et al. 2013). The catalytic GIY-YIG module can be combined with various DNA-binding domains affecting DNA recognition and cleavage specificity (Derbyshire et al. 1997; Dunin-Horkawicz et al. 2006). In *Terminon*-encoded GIY-YIG ORFs, the central GIY-YIG domain includes the conserved catalytic R and N residues (Kowalski et al. 1999), and is framed by N- and C-terminal extensions averaging 300 and 130 aa, respectively. This arrangement does not match any of the known domain architectures, in which the GIY-YIG domain exhibits strong N-terminal preference. While the N- and C-terminal extensions lack known motifs, a characteristic arrangement of Cys residues ($CX_{2-4}CX_2CX_{33-35}CX_2CX_{10}CXCX_{59-63}HX_3C$), with a CXC motif embedded within the GIY-YIG motif, partially matches that in the PLE *Neptune* clade, where it is found between RT and EN (Arkhipova 2006). While this array of Cys residues does not match known Zn-finger-like profiles, it could still play a role in DNA binding, or form S-S bridges.

Rep (Replication Initiator Proteins)

In geminivirus Rep proteins, the N-terminal catalytic domain is critical for origin recognition and DNA cleavage/nucleotidyl transfer, while the C-terminal domain possesses helicase activity and belongs to superfamily 3 helicases (S3H or SF3), also classified as AAA+ ATPases (Campos-Olivas et al. 2002; Hickman and Dyda 2005; Clerot and Bernardi 2006). Out of 18 *Terminon* families, eight are associated with geminivirus-like Rep ORFs (table 1). Only three of them, however, are carrying the intact catalytic domain with two histidines (HxH or HUH) required for metal binding, the tyrosine performing DNA cleavage and ligation (YxxK motif), and the S3H domain (Chandler et al. 2013). Several families carry a shorter ORF derived from the N-terminal catalytic domain (Rep-Nc), which lost the metal-binding HUH and the catalytic YxxK motifs, but has persisted throughout evolution as a distinct clade (supplementary fig. S3C, Supplementary Material online). While some of the families retain only the helicase moieties (Rep-C), these are not phylogenetically distinct and likely correspond to random 5′-deletion products (supplementary fig. S3C, Supplementary Material online). Rep ORFs are more similar to geminiviruses than to *A. vaga Helitrons*, which also contain Rep domains (Kapitonov and Jurka 2007), and display the characteristic geminiviral domain structure that includes not only HUH-Y2 and S3H, but also the central domain Gemini_AL1_M (pfam08283). In contrast to other *Terminon* ORFs, such as GIY-YIG or CC-ORF (supplementary figs. S3 and S7, Supplementary Material online), there is no concordance between Rep-based and RT-based phylogenies between families (supplementary fig. S3, Supplementary Material online). Together with lack of unidirectionality, phylogenetic incongruence indicates that Reps are not integral components of *Terminons*, but instead can establish an association with them, possibly aided by intersecting steps of their replication mechanisms, for example, a putative single-stranded DNA intermediate, and once associated, tend to propagate as a unit.

Putatively Structural ORFs

Coiled-Coil Motif-Containing ORFs (CC-ORFs)

Table 1 shows the nearly universal occurrence of CC-ORFs in the families prone to expansion (W and NT clades in fig. 2). Most of the numerous CC-ORFs, which are 400–500 aa in

length and occur as stand-alone coding sequences on either side of *Athena* RTs, exhibit similarity to the N-terminal moieties of RTs (equivalent to ORF1 in the frameshifted W and VX clades). The extreme N-terminus of RTs is clade-specific and comes in three variants (supplementary fig. S7*A* and *B*, Supplementary Material online). In the ILOM clade, it contains an excess of polar (S, T, Q, N, and Y) with some basic (K and R) residues. In the catalytically dead JVX clade, it displays a high content of acidic residues (D and E) at the N-termini, and weak matches to BAR domains sensing membrane curvature (pfam03114). In the intron-rich, catalytically intact NT and W clades, it carries a conserved N-terminal KR-rich motif with an adjacent region of weak homology to helix-turn-helix dsDNA-binding motifs, while the central core occasionally shows similarity to DnaJ chaperones and surface antigens (pfam00226). The stand-alone CC-ORFs share a common ancestor either with NWT-like (with the KR-rich motif) or with JVX-like (DE-rich with a conserved SGTG motif) N-terminal RT moieties, however they have evolved and diversified as separate clades (supplementary fig. S7, Supplementary Material online). One of the conserved intron positions coincides in both types of CC-ORFs in the core region common to all clades (supplementary fig. S7*B*, Supplementary Material online). While ORF1's and CC-ORFs do not resemble classical orthoretroviral *gag* genes and lack Zn-knuckles, they are reminiscent of the *gag* genes forming nucleocapsids in foamy viruses (*Spumaretrovirinae*), which are similarly sized, do not undergo proteolytic processing, and contain up to four coiled-coil motifs (Goldstone et al. 2013; Mullers 2013). A combination of KR-rich and DE-rich N-termini in co-occurring CC-ORFs is highly likely to affect RNP properties, and might aid in raising the limits on RNA packaging.

Transmembrane Domain-Containing ORFs (TM-ORFs)

TM-ORFs are found in 12 *Terminon* families (table 1) and are typically small (200–300 aa in length), with one (or rarely two) predicted TM domain of type I membrane topology (single-pass N-exo/C-cyt). Some of these ORFs also contain a predicted coiled-coil motif and/or cysteine residues which may form disulfide bridges. The low conservation of TM-ORFs offers limited insight into their function other than possible interaction with membranes.

Cis-Acting Sequences

pLTRs

Initially, we attempted to assign *Athena* boundaries relying on terminal repeats known as pLTRs, which are characteristic of PLEs (Evgen'ev and Arkhipova 2005; Arkhipova et al. 2013). However, upon inspection of the larger *Terminon* units, it became evident that, in contrast to canonical *Penelopes*, pLTRs are not always found around the RT ORF. Even if present, they do not necessarily delimit the boundaries of the transposed unit, as the region of within-family homology may extend beyond pLTRs. Most *Terminon* pLTRs end in reverse-complement telomeric repeats $(ACACCC)_n$ forming the 3′-boundary of the unit (supplementary table S1, Supplementary Material online). Furthermore, no TSDs surrounding pLTRs, as in canonical *Penelope* elements, can be discerned. In addition, pLTR-like sequences can frame nonRT ORFs (e.g., DEDDy-like ORFs in the L family; supplementary fig. S6*D*, Supplementary Material online), and thus are not directly associated with RTs *per se*. Collectively, these observations confirm that the presence of pLTRs in *Terminons* is not a result of their regeneration during each retrotransposition cycle, as in retroviral LTRs. Instead, it may be that pLTR conservation results from the presence of hammerhead ribozyme (HHR) motifs described below, which were identified in pLTRs.

Hammerhead Ribozyme Motifs and Their Relation to pLTRs

The minimalist HHR motifs were previously found in pLTRs of diverse PLEs, including *Athena-M* in *A. vaga* (Cervera and De la Pena 2014). Their role remains unclear, since the predicted ribozyme cleavage site never coincides with TE-host boundary. It is conceivable that HHR motifs could aid in processing of longer cotranscripts, as do HDV-like ribozymes in R2 and L1Tc nonLTR retrotransposons, eliminating the need for internal promoters (Eickbush and Eickbush 2010; Sanchez-Luque et al. 2011) or terminators of transcription, and possibly enabling expression of multiple ORFs from a single RNA.

We searched for HHR motifs in *A. vaga* genomic DNA with the parameters used in (Cervera and De la Pena 2014), which were based on empirical criteria and tested on datasets of functionally active HHRs. A total of 497 HHR motifs fitting these descriptors were detected in *A. vaga* scaffolds, and assigned to M, O, W1, N, Q, R, S, and T families. Inspection of the remaining families revealed the essential core motifs (CUGANGA. . .GAAA) in the L and W families, albeit with a slightly different spacing (a much longer loop 2) (fig. 3*A*). When we modified the descriptor to accommodate these families, HHR-like motifs were detected in all PLE families, except for K, W2, and P. However, all members of the K, W2, and P families contain substitutions in the core HHR motif, although the sequence can easily be aligned with other HHRs and apparently preserves the structural helices (fig. 3*A*). The P family in the congeneric *Adineta sp. 11* (fig. 3*A*: supplementary fig. S6*B*, Supplementary Material online) carries a mutation in a different part of the catalytic core. The substitution in the core catalytic HHR motif, however, did not prevent successful expansion of W2 family in *A. vaga* (supplementary fig. S2, Supplementary Material online), implying that RNA structural properties are more important for proliferation than catalytic properties. Indeed, our tests for HHR activity *in vitro* (fig. 3*B*) demonstrate that self-cleavage in the JW HHR (three identical tandem units) is seen only in 25 mM $MnCl_2$ and does not occur under physiological conditions or in $MgCl_2$, reinforcing the idea that efficient catalysis in the HHR motif is not required for *Terminon* transposition. Although HHR motifs in a few other PLEs are also nonfunctional under physiological conditions (Cervera and De la Pena 2014), our experiments show that the HHR in a canonical *A. vaga* PLE, *AvPen3a* (Arkhipova et al. 2013), can efficiently self-cleave as a monomer in 3 mM $MgCl_2$ (fig. 3*C*). Tandem duplication of the HHR-bearing segments is thought to be important for functionality of minimalist HHRs in a dimeric

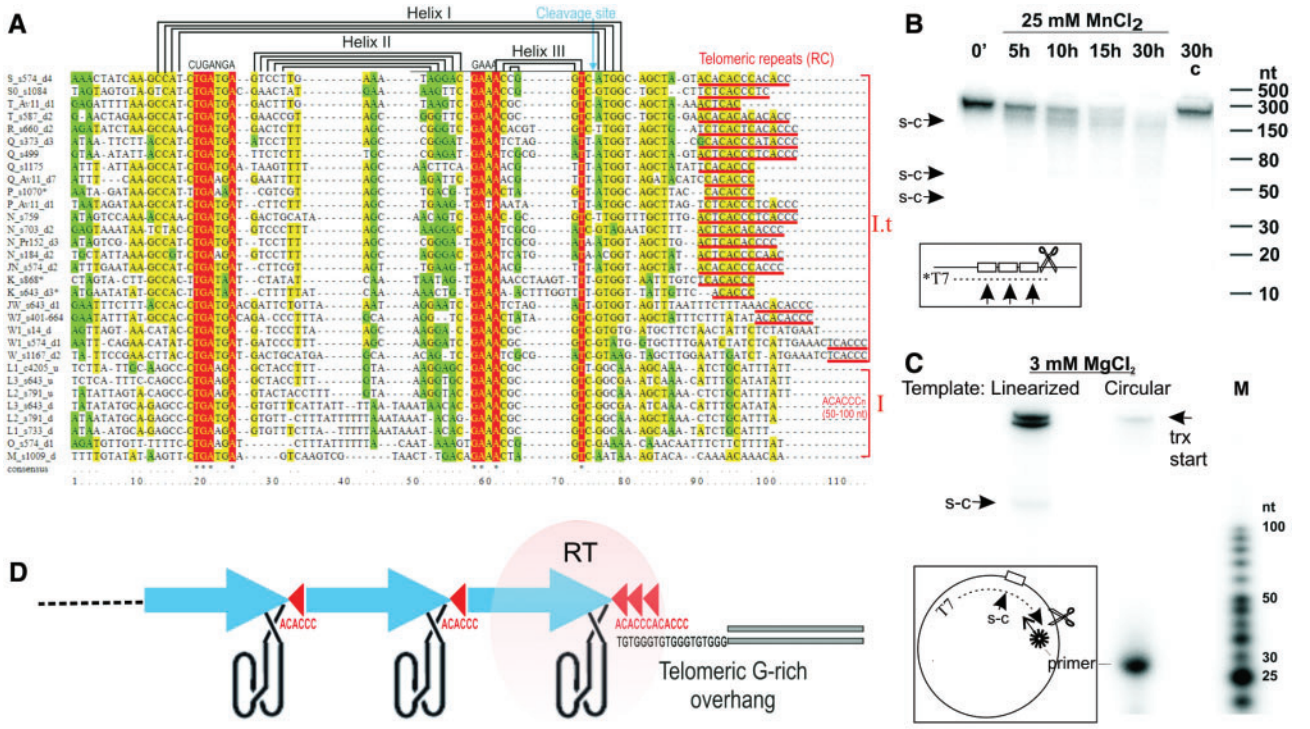

Fig. 3. Properties of PLE-associated HHRs. (*A*) Alignment of *Terminon* HHR motifs. Brackets separate W-NT and MOL clades differing by placement of reverse-complement (RC) telomeric repeats (50–100 nt downstream in type I; adjacent in type I.t). Sequence IDs include: family, scaffold #, HHR location downstream (d) or upstream (u) of RT, and its number in a series of tandem units. (*B*) Self-cleavage of the T7-driven 463-nt *JW_s643* transcript (inset, dotted line) with three HHRs (boxes). Arrows, self-cleavage sites (s-c); scissors, linearization of the plasmid template; *, uniform T7 labeling. Time course, 0–30 h; C, control 30-h incubation without $MnCl_2$. (*C*) Self-cleavage of the T7-driven 306-nt *AvPen3a* pLTR transcript with a single HHR. A 29-nt end-labeled primer was used for extension (inset); other notations are as in (*B*). (*D*) Model for end recognition and terminal attachment, with HHRs in a schematic 3-D configuration.

configuration (Cervera and De la Pena 2014), although, contrary to the dimer requirement reported for PLEs (Lünse et al. 2016), the *AvPen3a* HHR functions as a monomer.

The HHR-bearing repeats represent the most conserved region of the pLTRs, possibly reflecting their role as *cis*-acting elements for 3′-end recognition by RT, analogous to 3′-end stem-loop recognition in LINEs (Hayashi et al. 2014). Such *cis*-acting elements could participate in *trans*-mobilization of genic regions unrelated to *Terminons*, as seen in examples shown in figure 1 and supplementary figure S1, Supplementary Material online. While the variable nature of 5′-termini and the lack of TSDs precludes unambiguous identification of such transduction events, it is noteworthy that foreign genes and members of host multigene families are often colocalized with telomeric repeats and HHRs, as are *Terminon* ORFs (see below).

The HHR motif can be positioned within pLTRs in two ways: intron-containing clades (NT-W) harbor the HHR motif near the 3′-end, with telomeric repeats directly adjacent to helix I (fig. 3A, type I.t), while the intron-less families (L,O,M) carry the HHR motif in the 5′-terminal part of the pLTR, as do canonical *Penelope* elements. Note that helix I is the outermost helix of the type I HHR fold, and the expected cleavage site is always located in the center of the HHR, never coinciding with the TE-host boundary.

Rep Origins

The putative Rep-associated origins of replication represent yet another type of *cis*-acting elements often found near full-length Rep ORFs. They usually consist of a hairpin structure (fig. 1), often in combination with a series of tandem repeats, which are reminiscent of "iterons" in geminiviruses and contribute to the specificity of Rep binding to the hairpin (Londono et al. 2010). Such sequences often mark the point separating two divergent ORFs, as seen in geminiviruses (Hanley-Bowdoin et al. 2013).

ORF Polarity, Syntenic Blocks, and Gene Capture

The directionality of ORFs within each unit (fig. 1 and supplementary figs. S2 and S6, Supplementary Material online) implies that transcript continuity is important for function, and distinguishes *Terminons* from self-synthesizing *Polintons*/*Mavericks*, a class of virus-like DNA TEs of comparable size (15–20 kb, encoding up to ten ORFs in both orientations) (Kapitonov and Jurka 2006; Pritham et al. 2007). Such ORF unidirectionality is typical for retroelements, but not for DNA TEs. Spacing between *Terminon* ORFs can be very close, consistent with residing on a single long transcript rather than on individual transcriptional units.

Except for directionality, ORFs in different families are not arranged in any predetermined order, which often makes

Terminon identification a nontrivial task, especially when the 3′-proximal ORFs are nonenzymatic. Nevertheless, blocks of synteny can be traced between some families, likely reflecting the degree of their evolutionary relatedness (supplementary fig. S6, Supplementary Material online). Some ORFs exhibit evidence of circular permutation, for example, in WJ and JW families (supplementary fig. S6A, Supplementary Material online), which might occur during processing of a circular intermediate. Occasionally, ORFs can undergo partial deletion, so that only a fragment remains identifiable (supplementary fig. S6A, Supplementary Material online).

We did not detect internal capture of host genes within boundaries of any of the *A. vaga Terminon* units: if a *Terminon* sequence is interrupted, it is usually by insertion of a different TE type (LTR, nonLTR, or DNA TEs) (supplementary fig. S6, Supplementary Material online). Neither could we find a known TE or host gene to be interrupted by *Terminon* insertion: if broken, TEs or host genes are subject to end healing by telomeric repeats followed by *Terminon* attachment, so that the missing gene part is not found at the other end of *Terminons* or elsewhere in the genome (as in examples shown in supplementary figs. S1A and S6, Supplementary Material online).

Importantly, while foreign genes and host genes from multigene families are rarely internalized within units, they are often found between *Terminons* or their 3′-ends, either in direct or inverse orientation (fig. 1 and supplementary fig. S1C, Supplementary Material online). Thus, *Terminons* are likely to participate in gene amplification and transfer by providing *cis*-acting elements for transduction of genes *via trans*-action of the RT. Such complex structures could additionally propagate *via* rolling-circle replication, if combined with Rep origins of replication.

Transcription and piRNA Production

Most of the *Terminon* families in *A. vaga*, as judged by RNA-seq counts, exhibit measurable levels of transcriptional activity, which is largely anticorrelated with small RNA counts (fig. 4A). Members of the M, O, and W2 families could represent recent additions which have not yet established a robust piRNA response. It should be noted that *Terminons* are ideally suited for establishment of piRNA clusters (Weick and Miska 2014). These genomic loci, which give rise to noncanonical Pol II transcripts processed into piRNAs, are characterized, *inter alia*, by extended transcript length and intron retention (Sapetschnig and Miska 2014; Chen et al. 2016).

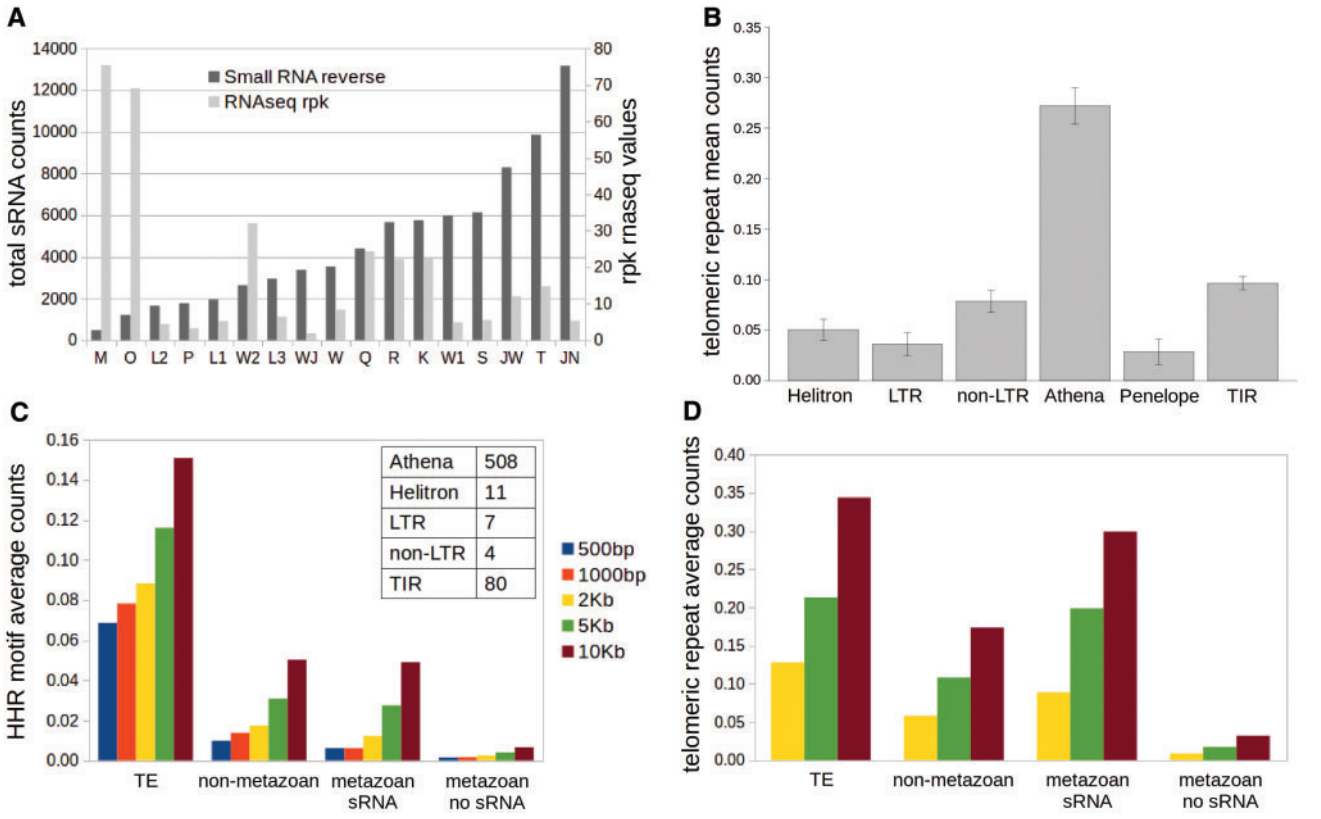

Fig. 4. RNA profiling and genomic environment of *Terminons*. (*A*) Distribution of RNA-seq reads with rpk values (reads per kb, right *Y*-axis), and small RNAs in reverse orientation (total counts, left *Y*-axis, sorted by increasing coverage) uniquely mapped to *A. vaga Terminons* from table 1. (*B*) Mean telomeric repeat counts around each type of annotated TE in a 10-kb window for genomic scaffolds >10 kb. A 5-kb window yields a similar profile (not shown). (*C* and *D*) Average HHR motif (*C*) and telomeric repeat (*D*) counts in indicated window sizes around each type of annotated features defined as in (Rodriguez and Arkhipova 2016). The inset in (*C*) counts the closest/overlapping TE annotations by TE type for each HHR motif.

Furthermore, not only can *Terminons* generate long sense-strand transcripts, but many copies are flanked at the 3′-end by an adjacent host gene in antisense orientation (supplementary table S1, Supplementary Material online) with the potential to provide a promoter for antisense transcription, which could stimulate formation of a dual-strand piRNA cluster (Sapetschnig and Miska 2014). If a flanking gene is 3′-truncated by terminal erosion and loses transcription termination signals, the resulting transcriptional readthrough would yield antisense *Terminon* transcripts, and hence RNA-mediated silencing (Kowalik et al. 2015).

Terminons in Other Members of the Class Bdelloidea

In earlier work, we amplified intron-containing *Athena* RT fragments by PCR of genomic DNA from representatives of three different bdelloid families, which diverged tens of millions of years ago (Mark Welch et al. 2008): *Philodina roseola* (Philodinidae), *Habrotrocha constricta* (Habrotrochidae), and *A. vaga* (Adinetidae). These fragments can now be reliably assigned to W, K, and N *Terminon* clades. To evaluate the degree of *Terminon* conservation in bdelloids, we inspected sequenced cosmid inserts from *P. roseola*, as well as *Athena*-containing contigs from a draft PacBio assembly of a natural isolate *Adineta sp.11*. In *P. roseola*, *Terminons* are joined to host DNA *via* a different telomeric repeat hexamer $(TCACCC)_n$ while in *Adineta* sp. *11* junctions are mostly formed by a variant octamer $(TCACACCC)_n$. Strikingly, ORF composition and even syntenic blocks have been preserved in several families, and thus can be traced back to their common ancestor. For example, the extended ORF block in the S/T family, which includes the catalytically dead *AthX* and *AthV* (*GIY-YIG*, *DEDDy*, *AthX*, *AthV*, 2x*GIY-YIG*, CC_{JVX}, *AthT*, *aORF*; supplementary fig. S6*C*, Supplementary Material online), appears in both *Adineta spp.*, which diverged over 10 Mya, and phylogenetic analysis of *Terminon* ORFs confirms the presence of virtually every described family throughout each species' evolutionary history (supplementary fig. S3, Supplementary Material online). Although nonRT *Terminon*-associated ORFs are shorter and less conserved than RTs, yielding less reliable branch support, their phylogeny is broadly congruent with the RT-based phylogeny, indicating that these ORFs have largely coevolved within each *Terminon* family (supplementary figs. S3*B* and S7, Supplementary Material online). The apparent exception is the Rep-related ORFs, for which a discordant phylogeny hints at the more transient character of their association with *Terminons*, albeit sufficiently prolonged to allow intron accumulation (fig. 3*C*). Overall, while the prevailing mode of inheritance for each *Terminon* family appears to be vertical, they may also persist within genera and species via horizontal mobility if a master copy is lost.

Terminons, Telomeric Repeats, and Foreign Genes

To reinforce the connections between host gene relocation and *Terminon* addition observed in isolated examples (fig. 1 and supplementary fig. S1, Supplementary Material online), we sought to investigate statistically the correlations between TE families, foreign genes, and telomeric repeats in window sizes $\leq$10 kb. At such distances, the overall TE density in *A. vaga* was shown to be significantly higher around foreign genes, and *vice versa* (Flot et al. 2013). We began by counting the number of telomeric repeats in windows of 2, 5, and 10 kb around each TE type. Using single-factor ANOVA (Tukey's test), we investigated distribution of telomeric repeats around different TE families for each window size. It is evident that the number of telomeric repeats is significantly higher near *Athena* RTs than near LTR, nonLTR, TIR, and *Helitron* elements, for window sizes $\geq$2kb (fig. 4*B* and supplementary table S2, Supplementary Material online). Significant differences between *Athena* and *Penelope* start with the window size of 5 kb.

Regarding HHR motifs, a clear association is again observed with PLEs but not any other TE type, with 508 out of 614 HHR counts showing the association (fig. 4*C*). HHR motifs also tend to occur close to foreign genes and host gene families with piRNA coverage, known to accumulate in the extended subtelomeric regions (Flot et al. 2013), but are almost never found near the bulk of host genes in the core genome (fig. 4*C*). Similar patterns are observed for telomeric repeats, which yield elevated counts near TEs and foreign genes, but low counts near the bulk of host genes (fig. 4*D*). Comparing Figure 4*B* and *D*, it is worth noting that accumulation of other TE types in subtelomeres is likely due to the reduced deleterious effects of their insertion in these regions.

Notably, reinspection of our telomere-enriched mini-libraries from *A. vaga* and *P. roseola* (Gladyshev and Arkhipova 2007) shows that over 50% of sequenced plasmid clones represented various parts of *Terminons*, although <20% were previously recognized as *Athena*-containing. Together with fluorescent *in situ* hybridization data localizing *AthO*/*AthM*-containing cosmids to *P. roseola* telomeres (Gladyshev and Arkhipova 2007), our analysis underscores the capacity of *Terminons* to occupy terminal positions, forming multiple layers of "sacrificial DNA" at telomeres.

Discussion

For years, our knowledge of structural and functional TE diversity has remained relatively stable, with the understanding that we have largely grasped the major principles of their structural organization and the underlying basis for their mobility. It is therefore of special interest to identify taxonomic groups harboring hitherto unknown TE types. The principal subdivision between TEs rests upon the involvement of RNA into the replication cycle (class I TEs) or lack thereof (class II TEs). In class I autonomous TEs, the process of RNA copying into DNA requires that the TE codes for an RT, the enzyme capable of performing RNA-dependent DNA synthesis. On these grounds, the novel TEs described herein, named *Terminons*, can be unambiguously classified as retrotransposons. This, however, does not rule out the presence of enzymatic activities that may be involved in additional stages of the transposition cycle, which may even include rolling-circle replication. In total, the newly annotated *Terminons* occupy 1.1% of the *A. vaga* genome, increasing the known TE content from ~3% (Flot et al. 2013) to slightly over 4%.

Molecular signatures around *Terminons* clearly point at terminal addition as the primary integration mechanism (fig. 3*D*). The characteristic 5′-truncation (supplementary fig. S2, Supplementary Material online), which in nonLTR retrotransposons is often ascribed to premature RT fall-offs, in *Terminons* may also result from terminal DNA erosion, if the 5′-end is exposed to exonucleases before being capped by telomeric repeats. Long head-to-tail chains of sequentially added *Terminons* can exceed 60 kb in length, thereby greatly increasing the buffer zone that counteracts the ongoing terminal erosion. Site-specific integration into telomeric repeats, as observed in *Bombyx mori* for SART/TRAS retrotransposons (Fujiwara et al. 2005), is highly unlikely, because *Terminons* are often found attached to terminally truncated and healed host genes or TEs. The observed bias towards oppositely oriented transcriptional units, especially 3′-truncated ones, could indicate a shift from uni-strand piRNA-producing loci known to operate in somatic tissues to dual-strand loci known to operate in the germ line, *via* antisense transcriptional units often lacking a proper poly(A) signal (Mohn et al. 2014; Weick and Miska 2014; Kowalik et al. 2015). Members of the most prolific T family (supplementary figs. S2 and S6*C*, Supplementary Material online) have additionally incorporated a small transcriptionally active antisense ORF near the 3′ end.

Although *Terminons* can harbor a diverse set of enzymatic activities (table 1), none of these appear obligatory, except for RT itself, which is combined with CC-ORF of putatively structural nature. Some of the enzymatic ORFs (GIY-YIG, Rep) may have been recruited to facilitate transposition, while others (GDSL esterases; RNase D-like DEDDy exonucleases) may assist in RNP assembly and/or evading host defenses. It is of special interest that catalytically deficient ORFs derived from various enzymes are consistently found in *Terminon* units, often in combination with their catalytically intact counterparts, and have persisted throughout bdelloid evolution, indicating that their retention is not accidental. Moreover, those ORFs have evolved under purifying selection (data not shown), suggesting that their recruitment was not based on catalysis. Noncatalytic functions for such "pseudoenzymes" (Adrain and Freeman 2012) could be structural or regulatory, and may include utilization of their binding capabilities, or involvement in heteromeric complex formation. The observed difference in the extreme N-termini of KR-rich (NT/W-like) and DE-rich (JVX-like) ORFs, which carry strong positive and negative charges, respectively, could promote heteromeric complex formation, or the latter could act as nucleic acid decoys.

Interestingly, *Terminons* do not encode any protease-like ORFs, indicating that the CC-RT fusion polyproteins are either processed by host proteases, or can form large multimeric complexes, where RT moieties belong to polypeptide chains up to 1.3 kDa. Neither do they code for an RNase H-like activity, which removes RNA from DNA–RNA hybrid intermediates in cytoplasmically replicating retroviruses and LTR retrotransposons, but is optional in nonLTR retrotransposons, which can utilize host RNase H for target-primed reverse transcription in the nucleus (Malik and Eickbush 2001). The nonenzymatic CC-ORFs with coiled-coil motifs resemble in organization the gag proteins of certain reverse-transcribing viruses, which are dependent on eukaryotic cell membranes for their replication.

Despite the presence of exceptionally complex *Terminon* retroelements in all examined members of the class Bdelloidea, separated by tens of millions of years of evolution, we could not find *Athena*-like RTs in draft genomes of rotifers of the sister class Monogononta, which contain canonical EN(+) PLEs of the Neptune type (Arkhipova 2006; Arkhipova et al. 2013). Neither could we find any extra ORFs in EN-deficient RTs of telomeric *Coprina* PLEs in numerous sequenced filamentous fungi, where they occur in tandem arrays (Gladyshev and Arkhipova 2007; Arkhipova et al. 2013). Fungal EN(−) PLEs can be very efficient at terminal addition, occupying every telomere in some basidiomycetes, for example, *Agaricus bisporus* and *Tuber melanosporum* (Martin et al. 2010; Foulongne-Oriol et al. 2013). These EN(−) PLEs code for a single nonframeshifted *ca.* 1000-aa CC-RT ORF, but have no additional coding capacities. Thus, terminal addition *per se* does not require an extended ORF repertoire, although telomeric placement is clearly associated with the HHR motifs and reverse-complement telomeric repeats, which in bdelloids are uniquely exposed next to the HHR fold for optimal annealing to G-rich overhangs (fig. 3*A* and *D*).

Phylogenetic analysis of *Athena* RTs does not favor the scenario of shorter PLEs having evolved by reduction of longer ones. Rather, their phylogeny is more consistent with complex elements evolving from shorter ones *via* splitting of longer ORFs; acquisition of additional ORFs, possibly at the transcript level to account for coorientation; accumulation of introns; and loss of frameshifts. The L clade may serve as an example of recent expandability, as it still retains HHR remnants between neighboring ORFs (supplementary fig. S6*D*, Supplementary Material online). A split of ancestral elements into individual subdomains, perhaps by insertion of W-like ORFs, may have been accompanied by combination of elements with different subdomains, eventually giving rise to a highly complex structure preserving only one active RT compatible with a cognate HHR motif, which permits retrotransposition of the entire unit starting with HHR. The pLTR structure appears to have undergone a shift in HHR positioning relative to telomeric repeats (type I to type I.t; fig. 3*A* and *D*), which was likely selected to favor the optimal 3′-terminal configuration. Interestingly, a recent study associates HHR motifs with nonautonomous LTR retrotransposons, which may exist as short RNA circles (Cervera et al. 2016); however, all of those motifs belong to type III but not type I, differing in the topology of the open-ended helix, and possibly reflecting different structural requirements of PLE and LTR RTs.

It may be asked whether the unique structural characteristics of these retroelements could be associated with any biological features specific to the class Bdelloidea. In our view, the most relevant biological feature is the unusual susceptibility of bdelloid telomeric regions to acquisition of foreign genetic material and amplification of foreign genes and host multigene families (Flot et al. 2013; Rodriguez and Arkhipova 2016). PLEs are unique in their capacity to retain

introns after retrotransposition, which is also applicable to genes captured between pLTRs (Arkhipova et al. 2003, 2013). Retromobility of longer templates would be disfavored at internal chromosomal locations in the absence of a reliable integration mechanism, as it would largely depend on preexisting nicks or breaks. Bdelloid telomeres, however, apparently offer the opportunity to bypass intrachromosomal integration by supplying the exposed G-rich overhangs, and a TE which can take advantage of such overhangs for its proliferation can additionally provide the host with extra means of terminal DNA addition. The terminal attachment mode does not rule out occasional intrachromosomal integration events, which might be triggered by random DNA breakage or by the presumed nicking activity of GIY-YIG EN when present. However, such events would be rare in comparison with terminal addition, since the lack of RT-EN fusion eliminates the *cis*-preference effect based on cotranslational *cis*-recognition of structural elements near the 3′-end of the template by the RT.

In summary, we have identified and characterized a novel and ancient type of retroelements with unusually complex organization and variable gene content, which can be added to telomeric G-rich overhangs with the aid of the 3′-terminally positioned hammerhead ribozyme motif. Combination of putatively structural ORF types with differently charged N-termini within a unit suggests that they participate in formation of RNP particles with unusual properties. The associated *cis*-acting elements are strongly correlated with foreign genes and multigene families in the bdelloid rotifer *A. vaga*, suggesting their participation in intragenomic and/or intergenomic gene transfer.

It would be of interest to investigate *Terminon*-encoded ORFs for enzymatic activities *in vivo* and *in vitro*, as well as formation of putative RNP complexes, however such studies would be premature until transpositionally active copies are identified. While it is evident that *Terminons* have been recently transposing, as judged by the high degree of nucleotide sequence identity within some families (WJ, T) (supplementary fig. S2 and fig. S6C, Supplementary Material online), the sequenced *A. vaga* strain has been maintained in the laboratory for over 25 years and is no longer experiencing selective pressures to which natural populations are subjected, allowing ORFs to decay. Thus, we expect that further understanding of *Terminon* biology will come from comparative analysis of multiple bdelloid natural isolates. Although the mode of transmission for most families is predominantly vertical, as their interspecific divergence parallels that of host genes (data not shown), some families could exhibit horizontal mobility. The giant size and variable ORF composition of retroelements described herein pose new challenges to developers of automated TE annotation tools, and leave us wondering how many unknown TE types with the potential to give rise to novel intracellular or extracellular entities are lurking in the still poorly explored genomes of understudied taxonomic groups, and what unanticipated impacts they can have on their eukaryotic hosts.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank W. Reznikoff and M. Meselson for critical reading, A. Kondrashov, O. Vakhrusheva and E. Mnatsakanova for providing *Adineta sp. 11* isolate, and V. Srikanth for help with plasmid constructs. This work was supported by the National Institutes of Health (grant GM111917 to I.A.).

References

Adrain C, Freeman M. 2012. New lives for old: evolution of pseudoenzyme function illustrated by iRhoms. *Nat Rev Mol Cell Biol*. 13:489–498.

Akoh CC, Lee GC, Liaw YC, Huang TH, Shaw JF. 2004. GDSL family of serine esterases/lipases. *Prog Lipid Res*. 43:534–552.

Anantharaman V, Aravind L. 2010. Novel eukaryotic enzymes modifying cell-surface biopolymers. *Biol Direct*. 5:1.

Arkhipova I. 2006. Distribution and phylogeny of Penelope-like elements in eukaryotes. *Syst Biol*. 55:875–885.

Arkhipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. 2003. Retroelements containing introns in diverse invertebrate taxa. *Nat Genet*. 33:123–124.

Arkhipova IR, Yushenova IA, Rodriguez F. 2013. Endonuclease-containing Penelope retrotransposons in the bdelloid rotifer *Adineta vaga* exhibit unusual structural features and play a role in expansion of host gene families. *Mob DNA* 4:19.

Belfort M, Bonocora RP. 2014. Homing endonucleases: from genetic anomalies to programmable genomic clippers. *Methods Mol Biol*. 1123:1–26.

Caliskan N, Peske F, Rodnina MV. 2015. Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. *Trends Biochem Sci*. 40:265–274.

Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM. 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci U S A*. 99:10310–10315.

Cervera A, De la Pena M. 2014. Eukaryotic penelope-like retroelements encode hammerhead ribozyme motifs. *Mol Biol Evol*. 31:2941–2947.

Cervera A, Urbina D, de la Pena M. 2016. Retrozymes are a unique family of non-autonomous retrotransposons with hammerhead ribozymes that propagate in plants through circular RNAs. *Genome Biol*. 17:135.

Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 11:525–538.

Chen K-M, Campbell E, Pandey RR, Yang Z, McCarthy AA, Pillai RS. 2015. Metazoan Maelstrom is an RNA-binding protein that has evolved from an ancient nuclease active in protists. *RNA* 21:833–839.

Chen YC, Stuwe E, Luo Y, Ninova M, Le Thomas A, Rozhavskaya E, Li S, Vempati S, Laver JD, Patel DJ, et al. 2016. Cutoff suppresses RNA polymerase II termination to ensure expression of piRNA precursors. *Mol Cell*. 63:97–109.

Clerot D, Bernardi F. 2006. DNA helicase activity is associated with the replication initiator protein rep of tomato yellow leaf curl geminivirus. *J Virol*. 80:11322–11330.

Derbyshire V, Kowalski JC, Dansereau JT, Hauer CR, Belfort M. 1997. Two-domain structure of the td intron-encoded endonuclease I-TevI correlates with the two-domain configuration of the homing site. *J Mol Biol*. 265:494–506.

Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, Wulser K, Yanai-Balser GM, Gurnon JR, Vitek JC, et al. 2012. *Paramecium bursaria* chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. *J Virol*. 86:8821–8834.

Dunin-Horkawicz S, Feder M, Bujnicki JM. 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics* 7:98.

Eickbush DG, Eickbush TH. 2010. R2 retrotransposons encode a self-cleaving ribozyme for processing from an rRNA cotranscript. *Mol Cell Biol*. 30:3142–3150.

El Baidouri M, Kim KD, Abernathy B, Arikit S, Maumus F, Panaud O, Meyers BC, Jackson SA. 2015. A new approach for annotation of

transposable elements using small RNA mapping. *Nucleic Acids Res.* 43:e84.
Evgen'ev MB, Arkhipova IR. 2005. Penelope-like elements: a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res.* 110:510–521.
Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* 5:103–107.
Flot JF, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EG, Hejnol A, Henrissat B, Koszul R, Aury JM, et al. 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* 500:453–457.
Foulongne-Oriol M, Murat C, Castanera R, Ramirez L, Sonnenberg AS. 2013. Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus*. *Fungal Genet Biol*. 55:6–21.
Fujiwara H, Osanai M, Matsumoto T, Kojima KK. 2005. Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, *Bombyx mori*. *Chromosome Res*. 13:455–467.
Gladyshev E, Arkhipova IR. 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A.* 104:9352–9357.
Gladyshev E, Meselson M. 2008. Extreme resistance of bdelloid rotifers to ionizing radiation. *Proc Natl Acad Sci U S A.* 105:5139–5144.
Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science* 320:1210–1213.
Glebe D, Bremer CM. 2013. The molecular virology of hepatitis B virus. *Semin Liver Dis.* 33:103–112.
Goldstone DC, Flower TG, Ball NJ, Sanz-Ramos M, Yap MW, Ogrodowicz RW, Stanke N, Reh J, Lindemann D, Stoye JP, et al. 2013. A unique spumavirus Gag N-terminal domain with functional properties of orthoretroviral matrix and capsid. *PLoS Pathog.* 9:e1003376.
Hanley-Bowdoin L, Bejarano ER, Robertson D, Mansoor S. 2013. Geminiviruses: masters at redirecting and reprogramming plant processes. *Nat Rev Microbiol.* 11:777–788.
Hayashi Y, Kajikawa M, Matsumoto T, Okada N. 2014. Mechanism by which a LINE protein recognizes its 3′ tail RNA. *Nucleic Acids Res.* 42:10605–10617.
Hickman AB, Dyda F. 2005. Binding and unwinding: SF3 viral helicases. *Curr Opin Struct Biol.* 15:77–85.
Hohn T, Rothnie H. 2013. Plant pararetroviruses: replication and expression. *Curr Opin Virol.* 3:621–628.
Kapitonov VV, Jurka J. 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol Biol Evol.* 20:38–46.
Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.* 23:521–529.
Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A.* 103:4540–4545.
Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 9:411–412.
Kleinstiver BP, Wolfs JM, Edgell DR. 2013. The monomeric GIY-YIG homing endonuclease I-BmoI uses a molecular anchor and a flexible tether to sequentially nick DNA. *Nucleic Acids Res.* 41:5413–5427.
Kowalik KM, Shimada Y, Flury V, Stadler MB, Batki J, Bühler M. 2015. The Paf1 complex represses small RNA-mediated epigenetic gene silencing. *Nature* 520:248–252.
Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, Baxter SM, Derbyshire V. 1999. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: coincidence of computational and molecular findings. *Nucleic Acids Res.* 27:2115–2125.
Lin X, Faridi N, Casola C. 2016. An ancient transkingdom horizontal transfer of Penelope-like retroelements from arthropods to conifers. *Genome Biol Evol.* 8:1252–1266.
Londono A, Riego-Ruiz L, Arguello-Astorga GR. 2010. DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs. *Arch Virol.* 155:1033–1046.
Lünse CE, Weinberg Z, Breaker RR. 2016. Numerous small hammerhead ribozyme variants associated with Penelope-like retrotransposons cleave RNA as dimers. *RNA Biol*. doi: 10.1080/15476286.2016.1251002.
Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* 11:1187–1197.
Mark Welch DB, Mark Welch JL, Meselson M. 2008. Evidence for degenerate tetraploidy in bdelloid rotifers. *Proc Natl Acad Sci U S A.* 105:5145–5149.
Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, et al. 2010. Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* 464:1033–1038.
Mohn F, Sienski G, Handler D, Brennecke J. 2014. The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. *Cell* 157:1364–1379.
Montanier C, Money VA, Pires VM, Flint JE, Pinheiro BA, Goyal A, Prates JA, Izumi A, Stalbrand H, Morland C, et al. 2009. The active site of a carbohydrate esterase displays divergent catalytic and noncatalytic binding functions. *PLoS Biol.* 7:e71.
Mullers E. 2013. The foamy virus Gag proteins: what makes them different? *Viruses* 5:1023–1041.
Pritham EJ, Putliwala T, Feschotte C. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390:3–17.
Rodriguez F, Arkhipova IR. 2016. Multitasking of the piRNA silencing machinery: targeting transposable elements and foreign genes in the bdelloid rotifer *Adineta vaga*. *Genetics* 203:255–268.
Rodriguez F, Kenefick A, Arkhipova I. 2017. LTR-retrotransposons from bdelloid rotifers capture additional ORFs shared between highly diverse retroelement types. *Viruses* 9:78.
Sanchez-Luque FJ, Lopez MC, Macias F, Alonso C, Thomas MC. 2011. Identification of an hepatitis delta virus-like ribozyme at the mRNA 5′-end of the L1Tc retrotransposon from *Trypanosoma cruzi*. *Nucleic Acids Res*. 39:8065–8077.
Sapetschnig A, Miska EA. 2014. Getting a grip on piRNA cluster transcription. *Cell* 157:1253–1254.
Schneider AM, Schmidt S, Jonas S, Vollmer B, Khazina E, Weichenrieder O. 2013. Structure and properties of the esterase from non-LTR retrotransposons suggest a role for lipids in retrotransposition. *Nucleic Acids Res.* 41:10563–10572.
Ulferts R, Ziebuhr J. 2014. Nidovirus ribonucleases: structures and functions in viral replication. *RNA Biol.* 8:295–304.
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.
Weick EM, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* 141:3458–3471.
Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 8:973–982.
Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* 9:3353–3362.
Zeng Q, Langereis MA, van Vliet AL, Huizinga EG, de Groot RJ. 2008. Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution. *Proc Natl Acad Sci U S A.* 105:9065–9069.
Zuo Y, Deutcher MP. 2001. Exoribonuclease superfamilies: structural analysis and phylogenetic distribution. *Nucleic Acids Res.* 29:1017–1026.