Article

Giant reverse transcriptase-encoding transposable elements at telomeres

Irina R. Arkhipova*, Irina A. Yushenova, Fernando Rodriguez

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

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

Short title**:** Giant retrotransposons at telomeres

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 co-oriented 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 inter-family 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; non-LTR 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 co-encoded 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 over-abundance 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, non-LTR, 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 non-LTR 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 (Table S1). 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 (Table S1).

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 pre-existing TE or gene was irreversibly lost by truncation, with subsequent addition of telomeric repeats and *Terminon* attachment (Table S1; Figs. S1A, S6). 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* 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* (Fig. S1B).

The 3' ends of *Terminons* are often present in higher copy numbers than the 5' ends, a pattern observed in non-LTR retrotransposons, which display well-defined 3' ends and frequent 5'-truncations (Wei, et al. 2001; Hayashi, et al. 2014). As seen in Fig. S2, 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 Fig. 2 (and an expanded codon-based version in Fig. S3A) shows that *Athena*-like RTs are subdivided into three major clades with intact catalytic RT motifs: ILOM (I, L, O, M families); NT (N, K, Q, P, R, S/T families); and W (W, W1, W2 families). The additional JVX and Y clades (Fig. S3A) are formed by inactive RT derivatives (J, X, V, Y, see below), which lack essential catalytic residues, and are co-localized 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* (Fig. S3A), 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 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) (Fig. S4). The frameshift site exhibits conservation in an otherwise rapidly evolving sequence context (Fig. S4A-B). It is also present in all members of the W clade (Fig. S4A-C), and in the catalytically inactive X/V ORFs from the JVX clade (Fig. S4B,D). 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 (Figs. S3A, S5). 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 (Figs. S3, S7). Intron acquisition can be followed by occasional intron losses, as follows from the intron presence-absence mapping on the phylogenetic tree (Fig. S3A). All members of the frameshifted ILOM clade are intronless.

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 (Fig. S5). 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 (Fig. S3A) 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; Fig. S5). 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; Figs. S3, S6, S7).

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; Fig. S6D). 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, TM) can be intronless or may contain introns in conserved positions (0-7 introns per ORF, as in GIY-YIG or Rep ORFs) (Fig. S3B-C, S5, S6A-C). With a few exceptions (see below), most ORFs are co-oriented, as in Fig. 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 non-enzymatic 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 non-LTR retrotransposons (CR1, RTEX, 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, e.g.

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 (Fig. S5, S6D).

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 array 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, 8 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 (Fig. S3C). 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 (Fig. S3C). 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 (Figs. S3, S7), there is no concordance between Rep-based and RT-based phylogenies between families (Fig. S3). 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, e.g. 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 (Fig. S7A,B). In the ILOM clade, it contains an excess of polar (S,T,Q,N,Y) with some basic (K,R) residues. In the catalytically dead JVX clade, it displays a high content of acidic residues (D,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 (Fig. S7). One of the conserved intron positions coincides in both types of CC-ORFs in the core region common to all clades (Fig. S7B). 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.

<u>*Transmembrane domain-containing ORFs*</u> (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

<u>*pLTRs*</u>. 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 (Table 1). Furthermore, no TSDs surrounding pLTRs, as in canonical *Penelope* elements, can be discerned. In addition, pLTR-like sequences can frame non-RT ORFs (e.g. DEDDy-like ORFs in the L family; Fig. S6D), 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.

<u>*Hammerhead ribozyme motifs and their relation to pLTRs*</u>. 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 co-transcripts, as do HDV-like ribozymes in R2 and L1Tc non-LTR 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. 3A). 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. 3A). The P family in the congeneric *Adineta sp. 11* (Fig. 3A: Fig. S6B) 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* (Fig. S2), implying that RNA structural properties are more important for proliferation than catalytic properties. Indeed, our tests for HHR activity *in vitro* (Fig. 3B) 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 non-functional 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. 3C). Tandem duplication of the HHR-bearing segments is thought to be important for functionality of minimalist HHRs in a dimeric 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 Fig. 1, Fig. S1. 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 co-localized 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 (Figs. 1, S2, S6) 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 10 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 pre-determined order, which often makes *Terminon* identification a non-trivial task, especially when the 3’-proximal ORFs are non-enzymatic. Nevertheless, blocks of synteny can be traced between some families, likely reflecting the degree of their evolutionary relatedness (Fig. S6). Some ORFs exhibit evidence of circular permutation, e.g. in WJ and JW families (Fig. S6A), which might occur during processing of a circular intermediate. Occasionally, ORFs can undergo partial deletion, so that only a fragment remains identifiable (Fig. S6A).

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, non-LTR, or DNA TEs) (Fig. S6). 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 Fig. S1A, S6).

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; Fig. S1C). 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 anti-correlated 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 non-canonical 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). 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 (Table S1) 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*; Fig. S6C), 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 (Fig. S3). Although non-RT *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 co-evolved within each *Terminon* family (Figs. S3B, S7). 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. 3C). 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, Fig. S1), we sought to investigate statistically the correlations between TE families, foreign genes, and telomeric repeats in window sizes ≤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, non-LTR, TIR and *Helitron* elements, for window sizes ≥ 2kb (Fig. 4B; Table S2). 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. 4C). 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. 4C). 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. 4D). Comparing Fig. 4B and 4D, 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, re-inspection 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 less than 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 overall 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. 3D). The characteristic 5’-truncation (Fig. S2), which in non-LTR 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 (Fig. S2, S6C) 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. Non-catalytic 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 non-LTR retrotransposons, which can utilize host RNase H for target-primed reverse transcription in the nucleus (Malik and Eickbush 2001). The non-enzymatic 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, e.g. *Agaricus bisporus* and *Tuber melanosporum* (Martin, et al. 2010; Foulongne-Oriol, et al. 2013). These EN(-) PLEs code for a single non-frameshifted *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. 3A,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 co-orientation; 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 (Fig. S6D). 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. 3A,D), which was likely selected to favor the optimal 3'-terminal configuration. Interestingly, a recent study associates HHR motifs with non-autonomous 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; Arkhipova, et al. 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 pre-existing 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 co-translational *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) (Fig. S6C), 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 Figures S1-S7, Supplementary Tables S1-S2, and Supplementary Methods are available online (http://www.mbe.oxfordjournals.org).

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 in Genetics 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:00-00.
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 Biology 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. The EMBO Journal 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.

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	Unknown ORFs	HHR
I	1009	I*#	-	-	-	-	-	-	-	GDSL(i)	-	-
M	1009	M#	-	-	-	-	-	-	-	-	1 (a)	I
O	574	O#	-	-	-	-	-	sCC	-	-	-	I
L1	1351-560	L1#	-	1-nc	-	-	-	sCC	1	GDSL(a)/	-	I
L2	791	L2#	V#	2-nc	-	-	-	sCC	1	GDSL(a)	-	I
L3	643	L3#	-	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#	J	1	2	RepNc	1	1	3	sCC	JN2,JD1	I.t
W	560	W#	-	1-nc	1	1	1	1	2	2 sCC	-	I.t
JW	643	W#	J; Y	-	3	-	2	2	6	2 sCC	1	I.t
W2	721-1061	W#	-	-	2	-	1	1	-	2 sCC	4	I.t
W1	14; 660	W#	-	-	-	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#;V#	1	3	-	2	1	1	Zn-ribbon	2	I.t
T	587	T	X#;V#	1	3	-	2	sCC	1	Zn-ribbon	1 (a)	I.t

#, ORF with a programmed -1 ribosomal frameshift; *, ORF with a defect in all family members; 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.

Figure Legends

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 non-metazoan 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.

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, 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 Fig. S6.

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 3 HHRs (boxes). Arrows, self-cleavage sites (s-c); scissors, linearization of the plasmid template; *, uniform T7 labeling. Time course, 0-30h; 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.

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-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.

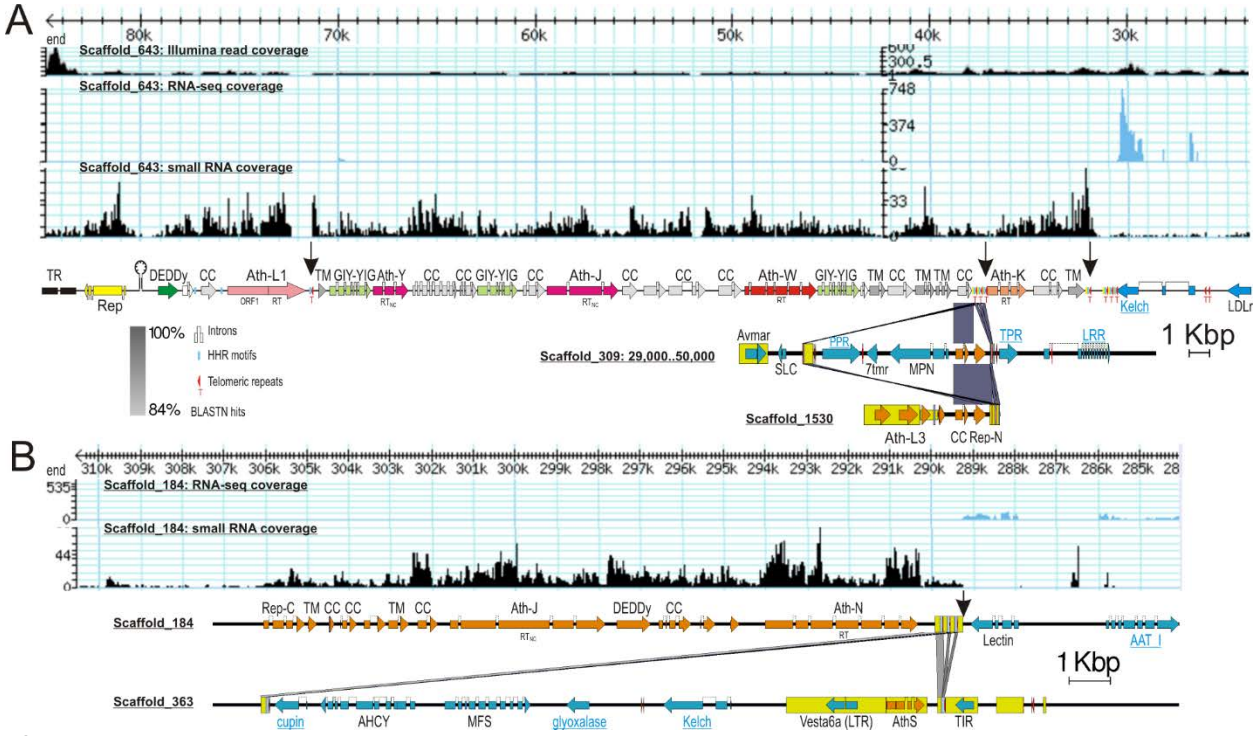
A
Scaffold_643: Illumina read coverage
Scaffold_643: RNA-seq coverage
Scaffold_643: small RNA coverage
Introns
HHR motifs
Telomeric repeats
BLASTN hits
Scaffold_309: 29,000..50,000
Scaffold_1530
1 Kbp
B
Scaffold_184: RNA-seq coverage
Scaffold_184: small RNA coverage
Scaffold_184
Scaffold_363
1 Kbp

Fig. 1.

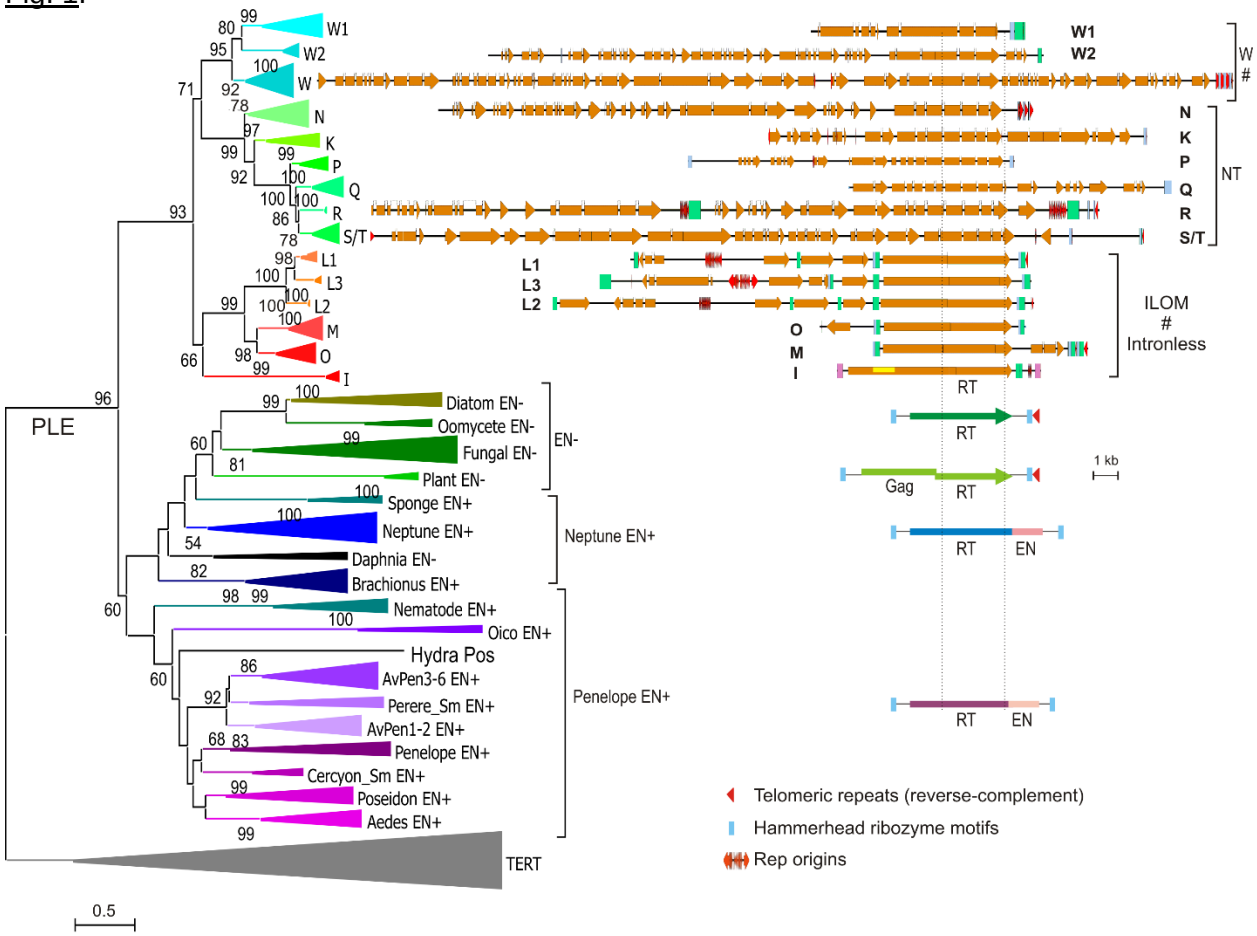
PLE
W1
W2
W
N
K
P
Q
R
S/T
L1
L3
L2
M
O
I
Diatom EN-
Oomycete EN-
Fungal EN-
Plant EN-
Sponge EN+
Neptune EN+
Daphnia EN-
Brachionus EN+
Nematode EN+
Oico EN+
Hydra Pos
AvPen3-6 EN+
Perere_Sm EN+
AvPen1-2 EN+
Penelope EN+
Cercyon_Sm EN+
Poseidon EN+
Aedes EN+
TERT
EN-
Neptune EN+
Penelope EN+
0.5
W #
NT
ILOM # Intronless
RT
Gag
EN
1 kb
Telomeric repeats (reverse-complement)
Hammerhead ribozyme motifs
Rep origins

Fig. 2.

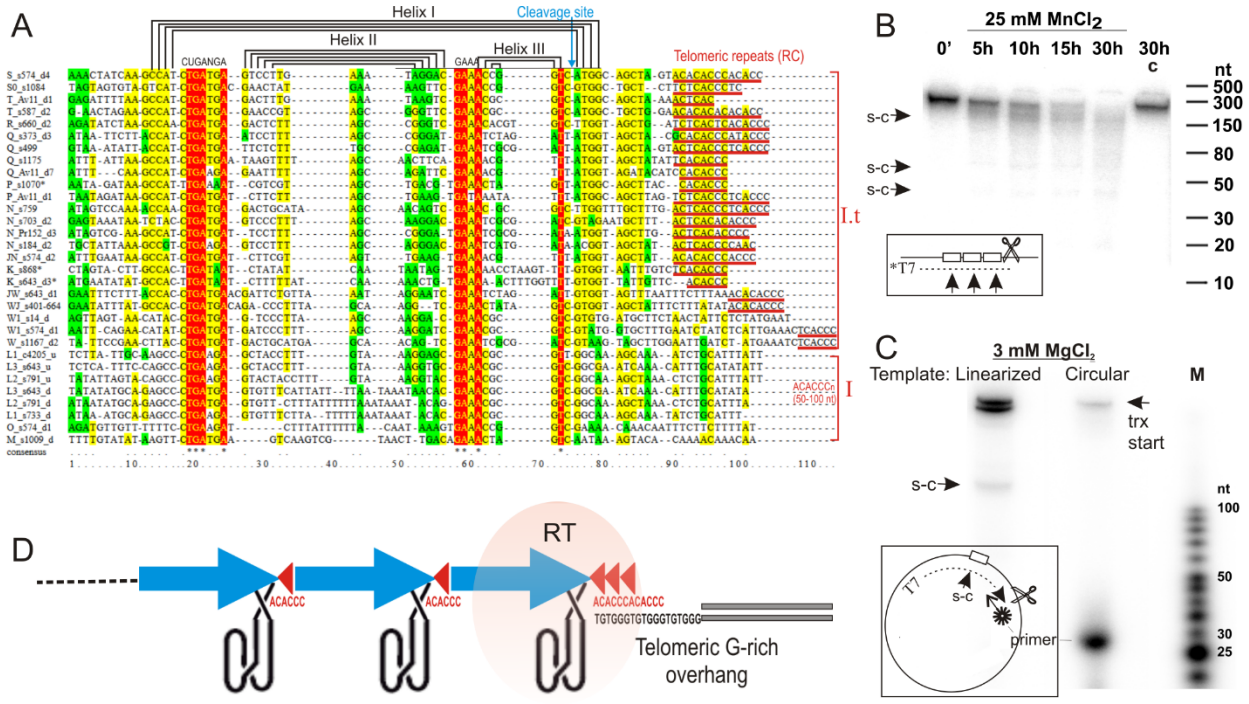
A
Helix I
Cleavage site
Helix II
Helix III
CUGANGA
GAAA
Telomeric repeats (RC)
B
25 mM MnCl2
0'
5h
10h
15h
30h
30h c
nt
500
300
150
80
50
30
20
10
s-c
*T7
C
3 mM MgCl2
Template: Linearized
Circular
M
trx start
nt
100
50
30
25
T7
primer
D
RT
ACACCC
ACACCCACACCC
TGTGGGTGTGGGTGTGGG
Telomeric G-rich overhang

Fig. 3.

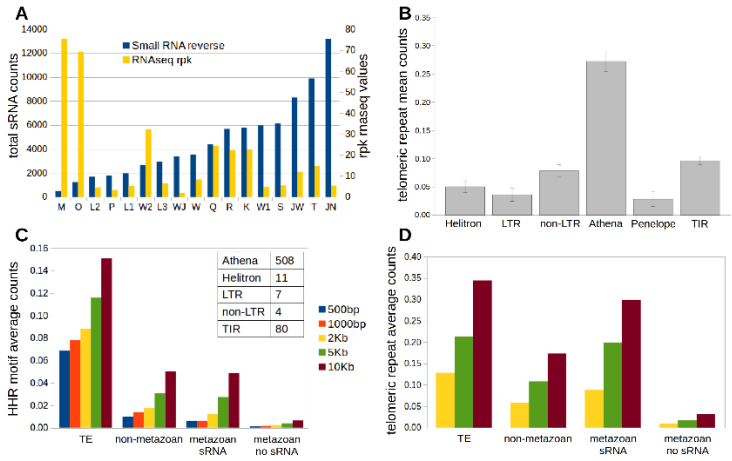
A
Small RNA reverse
RNAseq rpk
total sRNA counts
rpk maseq values
M O L2 P L1 W2 L3 WJ W Q R K W1 S JW T JN
B
telomeric repeat mean counts
Helitron
LTR
non-LTR
Athena
Penelope
TIR
C
HHR motif average counts
Athena 508
Helitron 11
LTR 7
non-LTR 4
TIR 80
500bp
1000bp
2Kb
5Kb
10Kb
TE
non-metazoan
metazoan sRNA
metazoan no sRNA
D
telomeric repeat average counts
TE
non-metazoan
metazoan sRNA
metazoan no sRNA

Fig. 4.

Supplementary Material

Giant reverse transcriptase-encoding transposable elements at telomeres

Irina R. Arkhipova, Irina A. Yushenova, Fernando Rodriguez

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

Supplementary Figure Legends

Fig. S1. Examples of terminal DNA loss and recent *Terminon* transposition in *A. vaga*.
(**A**) Two allelic scaffolds, one of which underwent loss of *CzcO* cyclohexanone monooxygenase (now present only in single copy in the *A. vaga* genome), and was healed by telomeric repeat addition and by *Ath-W* retrotransposition. Truncation occurred within a MITE TE, leaving the thioester reductase intact. The allelic regions, including the common MITE fragment, are 97% identical. (**B**) Telomere M1 (fosmid 184A11 from (Gladyshev and Arkhipova 2007)) compared to scaffold 868 from the same clonal culture four years later (Flot, et al. 2013). Blue arrows, PCR primers initially used to amplify sphingomyelin phosphodiesterase (GDPD), which failed to amplify it in 2010. GNAT, N-acetyltransferase GNAT family. Each separate transposed unit in the head-to-tail array is marked with a bracket. Other notations are as in Fig. 1. (**C**) Regions of HHR-Q homology surrounding the ORFs coding for a TPR protein and a seven-transmembrane receptor (7tmr). An HHR-W2-containing fragment is present in direct orientation.

Fig. S2. Intragenomic coverage of *Terminon* families in the *A. vaga* reference assembly.
Reference scaffolds from Table 1, each diagrammed on the top, were used in BLASTN searches of the *A. vaga* genome assembly scaffolds at NCBI. The graphical overview of the search results is displayed in red. The coding potential for each family is shown in detail in Fig. S6.

Fig. S3. Phylogenetic analysis of *Terminon*-associated enzymatic ORFs.
Unrooted maximum-likelihood (ML) phylogenetic trees were generated using codon-based nucleotide sequence alignments. Clade support values >50% are shown at the nodes; scale bars, nucleotide substitutions per site. Taxon names are as follows: *Av, Adineta vaga*, black; *As, Adineta sp. 11* natural isolate, green; *Pr, Philodina roseola*, red. Asterisks indicate a defect in ORF; 0i-9i, number of introns. Intron gains and losses, when definable, are shown by black and gray triangles, respectively. (**A**) ML tree of catalytically active (ILOM, W, NT) and catalytically inactive (JVXY nc) RT ORFs. Frameshift-containing clades are marked by #. The alignment spans the extended core RT including the C-terminal thumb domain and the conserved N1-N3 motifs at the N-termini (Arkhipova 2006); ORF1 is not included in its entirety, and is presented separately in Fig. S7. (**B**) ML phylogram of GIY-YIG-like ORFs. Clades marked with purple and magenta brackets are both present in their respective families and have been co-evolving with the corresponding RTs. (**C**) ML trees of the full-length Rep-like ORFs (Rep-FL) were split into the N-terminal (HUH-Y2) and C-terminal (S3H) domains, and analyzed together with the respective Rep-N and Rep-C ORFs containing only one of the domains (underlined). The non-catalytic Rep-Nc clade is marked in cyan; /, 5'-truncation. Rep-FL ORFs were assigned to arbitrary clades a-d; number of introns varied from 0 to 7, and is not shown for simplicity.

Fig. S4. Programmed ribosomal frameshifting in *Athena* RTs.
(**A**) Limited sequence conservation in the region spanning the frameshift site in the ILOM and W clades (using MView visualization tool at EMBL-EBI). Note that in the O clade, a shift of the slippery sequence to the left (underlined) has occurred in *A. vaga* (see also panel D). (**B**) Secondary structure-based alignment showing the stem-loop conservation in the W and VX clades. (**C-E**) KnotInFrame *in silico* prediction of -1 frameshift sites, including slippery sequences and simple pseudoknots, in representative *A. vaga* families.

Fig. S5. Graphical representation of pairwise similarities between RT coding sequences from all families.
Each family was plotted as nucleotide (**A**) and amino acid (**B**) sequence to highlight the contrasting degrees of divergence between ORF1 (N-termini) and RT. Representatives from each family (left column) from each of the four clades (right column, square brackets) on reference scaffolds (center column) were compared to their neighbors on the phylogram (Fig. 2) with Easyfig2.2.2. Identifiable cases of intron gains and losses are shown in (A) by yellow and blue triangles, respectively. The yellow box marks the location of the GDSL domain in the I clade; nc, lack of catalytic residues in RT-derived ORFs. The position of the frameshifting pseudoknot is marked by # on the top, and by vertical lines connected by a dotted line in other sequences in (A). The region corresponding to the core RT motifs 1-7 (Xiong and Eickbush 1990) is marked by a square bracket on the top (panel A, dashed line); motif RT5 is shown in white font.

Fig. S6. Structural organization of *Terminon* families.
Pairwise nucleotide sequence similarities are shown for individual scaffolds containing representatives of the W clade (**A**), NT clade (**B-C**), and ILOM clade (**D**). Scaffolds are grouped by RT phylogenetic relatedness (mini-phylograms on the left, with scale bars in nucleotide substitutions per site). Reference scaffold/contig IDs from *A. vaga* and *P. roseola* are shown in red type; *Adineta sp.* contigs (As), in green. Contiguous units from the same family on each scaffold are shown against the yellow background; from other families, against the green background. Blue arrows against the green background denote ORFs from other TE classes; blue arrows with no background, host genes. Other notations are as in Fig. 1.

Fig. S7. Properties of coiled-coil (CC) motif-containing ORFs.
(**A**) Sequence logos of different (gray bar) and similar (blue bar) regions from 67 NWT-like (top) and 46 JVX-like (bottom) ORFs. Light gray arrows mark family-specific intron positions also shown in (B); pink bar, weak homology to HTH motif. (**B**) Prediction of coiled-coil motifs in NWT-like (top) and JVX-like (bottom) ORFs. The gray and blue bars mark the regions shown in detail in (A). Black arrow, intron position conserved in both CC-ORF types; dark gray arrow, intron conserved in all NWT-like ORFs; light gray arrows, family-specific introns in P and J families. (**C**) Maximum likelihood (ML) analysis of NWT-like CC-ORF amino acid sequences, including: stand-alone CC_{NWT} ORFs located upstream (u) or downstream (d) from the catalytically intact RT; ORF1's from frameshifted W clade RTs; and N-terminal moieties from RTs of the NT clade. ORFs from *P. roseola* cosmids (Pr) are in red; from *A. vaga* scaffolds, in black; and from *Adineta sp. 11* (As), in green. Asterisks mark a defect in ORF. (**D**) ML analysis of JVX-like CC-ORFs, including stand-alone CC_{JVX} from NT and W clades; ORF1's from the frameshifted VX clade; and N-terminal moieties from the non-frameshifted J clade RT-like amino acid sequences. Scale bars, amino acid substitutions per site.

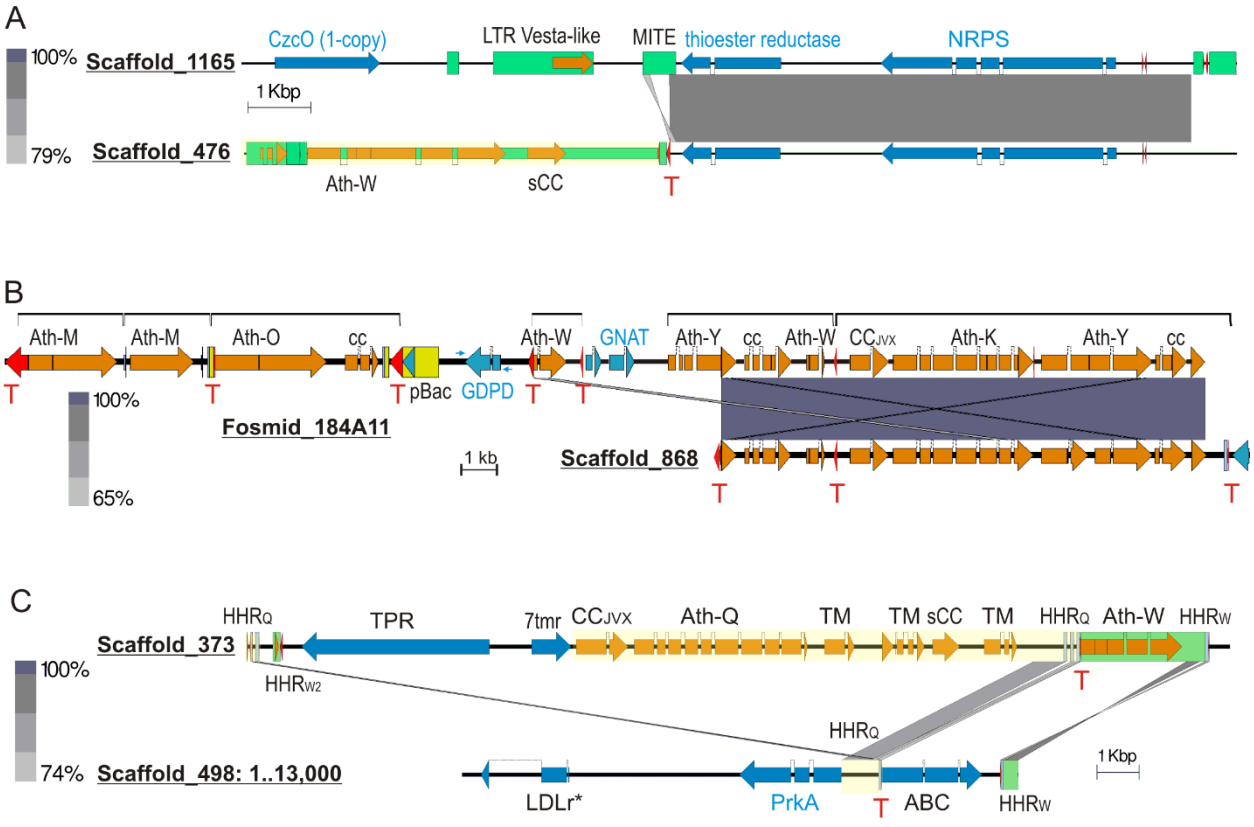

Fig. S1

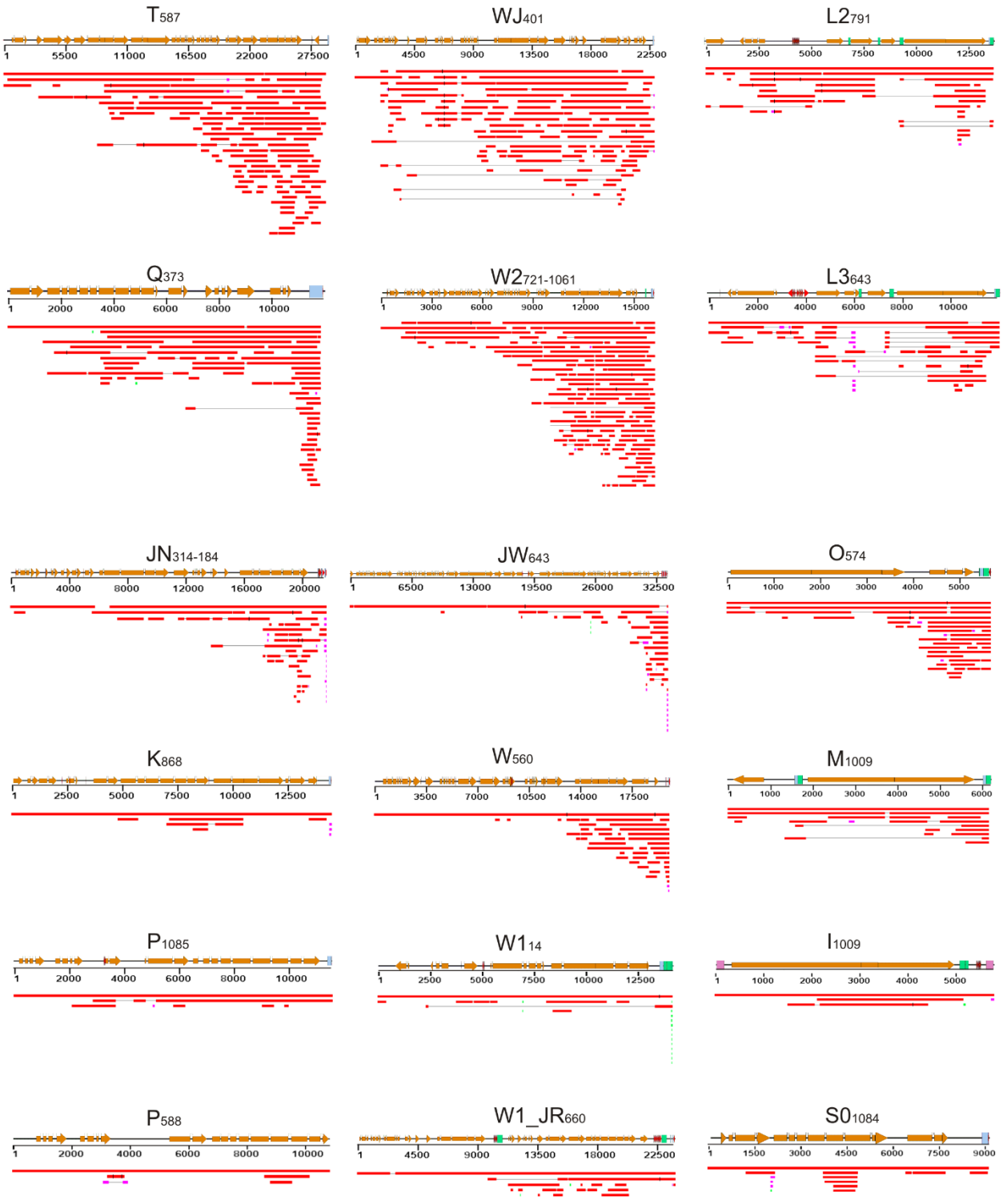
T587
1 5500 11000 16500 22000 27500
WJ401
1 4500 9000 13500 18000 22500
L2791
1 2500 5000 7500 10000 12500
Q373
1 2000 4000 6000 8000 10000
W2721-1061
1 3000 6000 9000 12000 15000
L3643
1 2000 4000 6000 8000 10000
JN314-184
1 4000 8000 12000 16000 20000
JW643
1 6500 13000 19500 26000 32500
O574
1 1000 2000 3000 4000 5000
K868
1 2500 5000 7500 10000 12500
W560
1 3500 7000 10500 14000 17500
M1009
1 1000 2000 3000 4000 5000 6000
P1085
1 2000 4000 6000 8000 10000
W114
1 2500 5000 7500 10000 12500
I1009
1 1000 2000 3000 4000 5000
P588
1 2000 4000 6000 8000 10000
W1_JR660
1 4500 9000 13500 18000 22500
S01084
1 1500 3000 4500 6000 7500 9000

Fig. S2

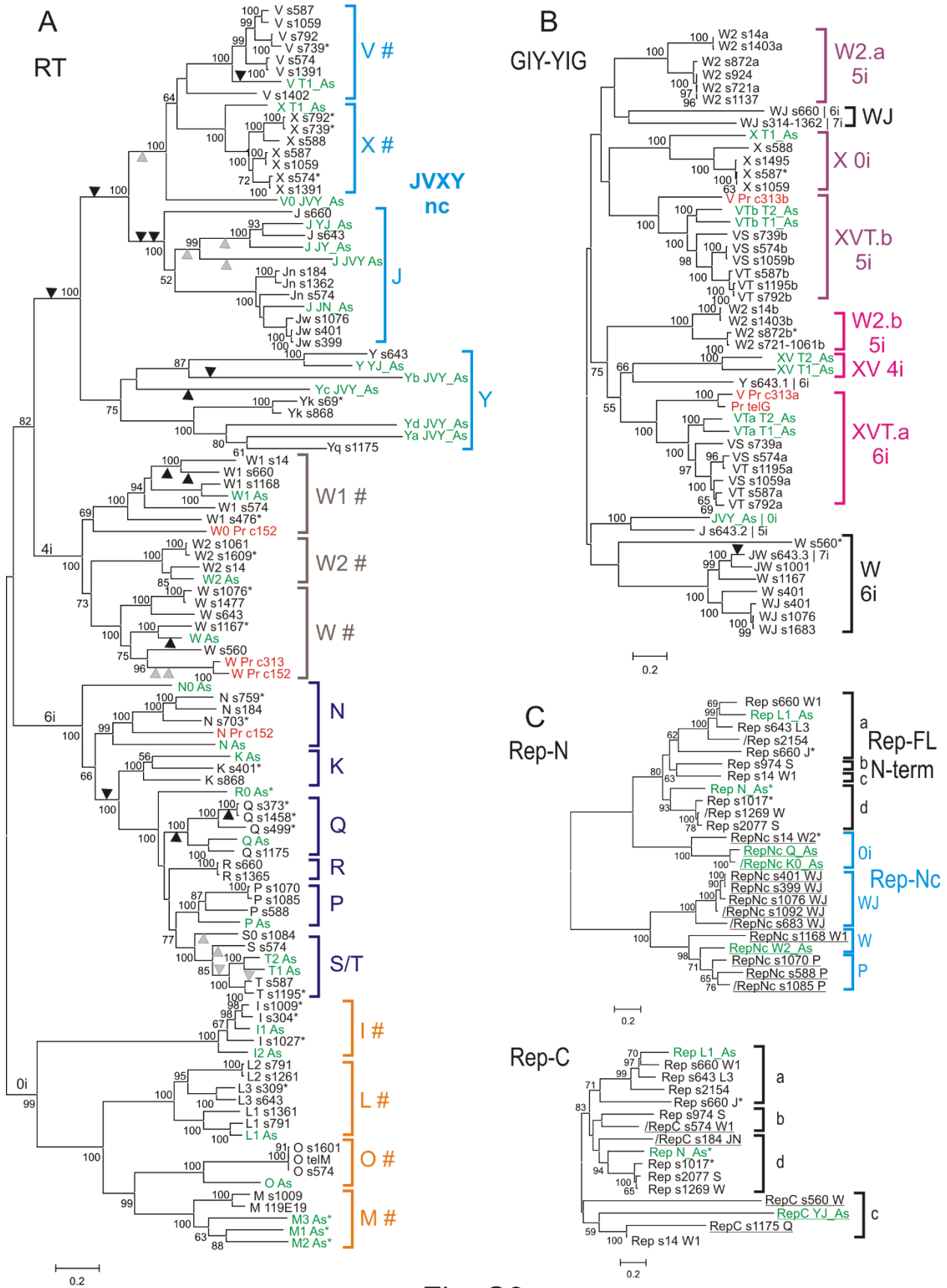
A
RT
V #
X #
JVXY
nc
J
Y
W1 #
W2 #
W #
N
K
Q
R
P
S/T
I #
L #
O #
M #
4i
6i
0i
0.2
B
GIY-YIG
W2.a
5i
WJ
X 0i
XVT.b
5i
W2.b
5i
XV 4i
XVT.a
6i
W
6i
0.2
C
Rep-N
a
b
c
d
Rep-FL
N-term
0i
Rep-Nc
WJ
W
P
0.2
Rep-C
a
b
d
c
0.2

Fig. S3

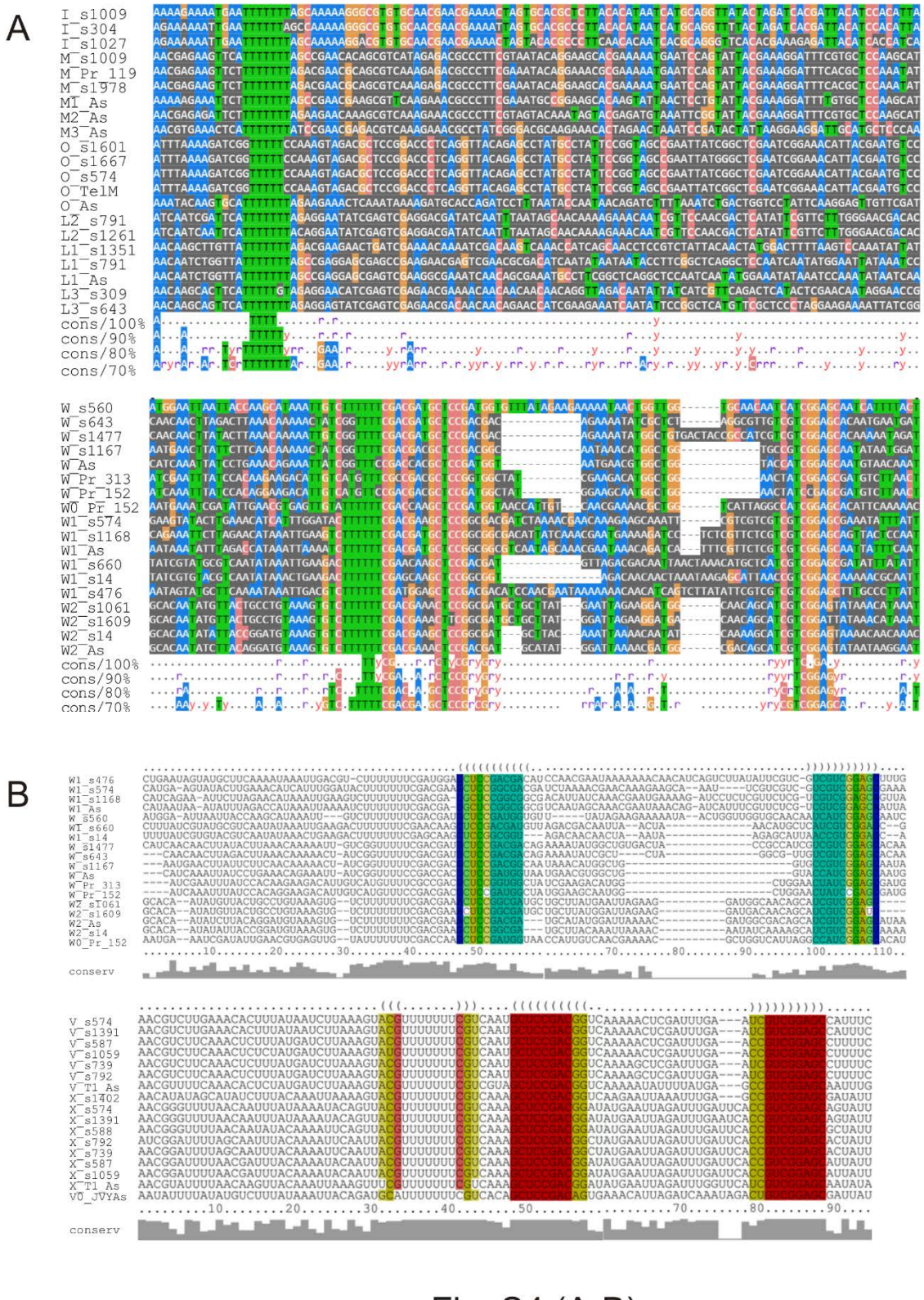
A
I_s1009
I_s304
I_s1027
M_s1009
M_Pr_119
M_s1978
MI_As
M2_As
M3_As
O_s1601
O_s1667
O_s574
O_TelM
O_As
L2_s791
L2_s1261
L1_s1351
L1_s791
L1_As
L3_s309
L3_s643
cons/100%
cons/90%
cons/80%
cons/70%
W_s560
W_s643
W_s1477
W_s1167
W_As
W_Pr_313
W_Pr_152
W0_Pr_152
W1_s574
W1_s1168
W1_As
W1_s660
W1_s14
W1_s476
W2_s1061
W2_s1609
W2_s14
W2_As
cons/100%
cons/90%
cons/80%
cons/70%
B
W1_s476
W1_s574
W1_s1168
W1_As
W_s560
W1_s660
W1_s14
W_s1477
W_s643
W_s1167
W_As
W_Pr_313
W_Pr_152
W2_s1061
W2_s1609
W2_As
W2_s14
W0_Pr_152
conserv
V_s574
V_s1391
V_s587
V_s1059
V_s739
V_s792
V_T1_As
X_s1402
X_s574
X_s1391
X_s588
X_s792
X_s739
X_s587
X_s1059
X_T1_As
V0_JVYAs
conserv

Fig. S4 (A-B)

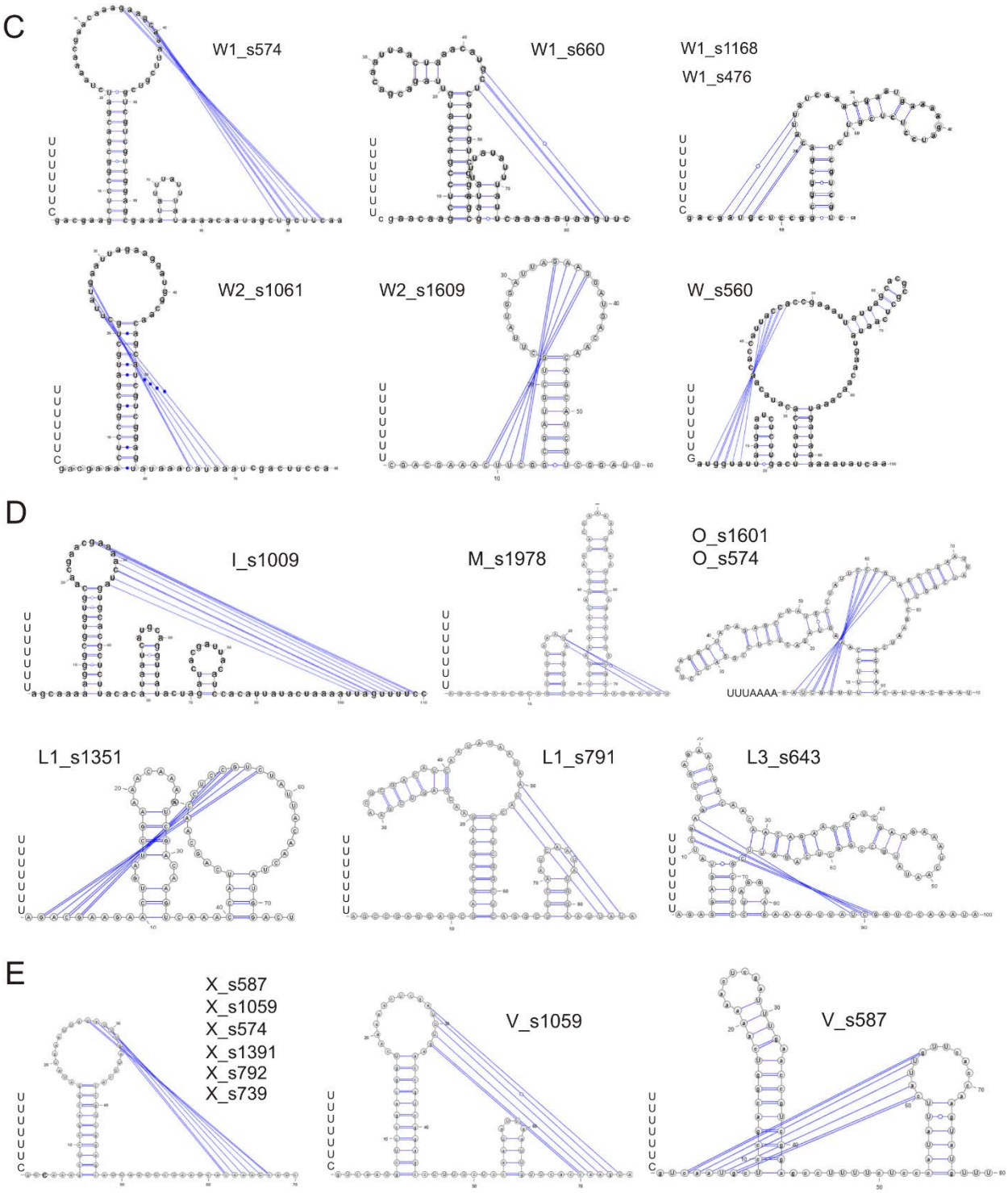
C
W1_s574
W1_s660
W1_s1168
W1_s476
W2_s1061
W2_s1609
W_s560
D
I_s1009
M_s1978
O_s1601
O_s574
L1_s1351
L1_s791
L3_s643
E
X_s587
X_s1059
X_s574
X_s1391
X_s792
X_s739
V_s1059
V_s587

Fig. S4 (C-E)

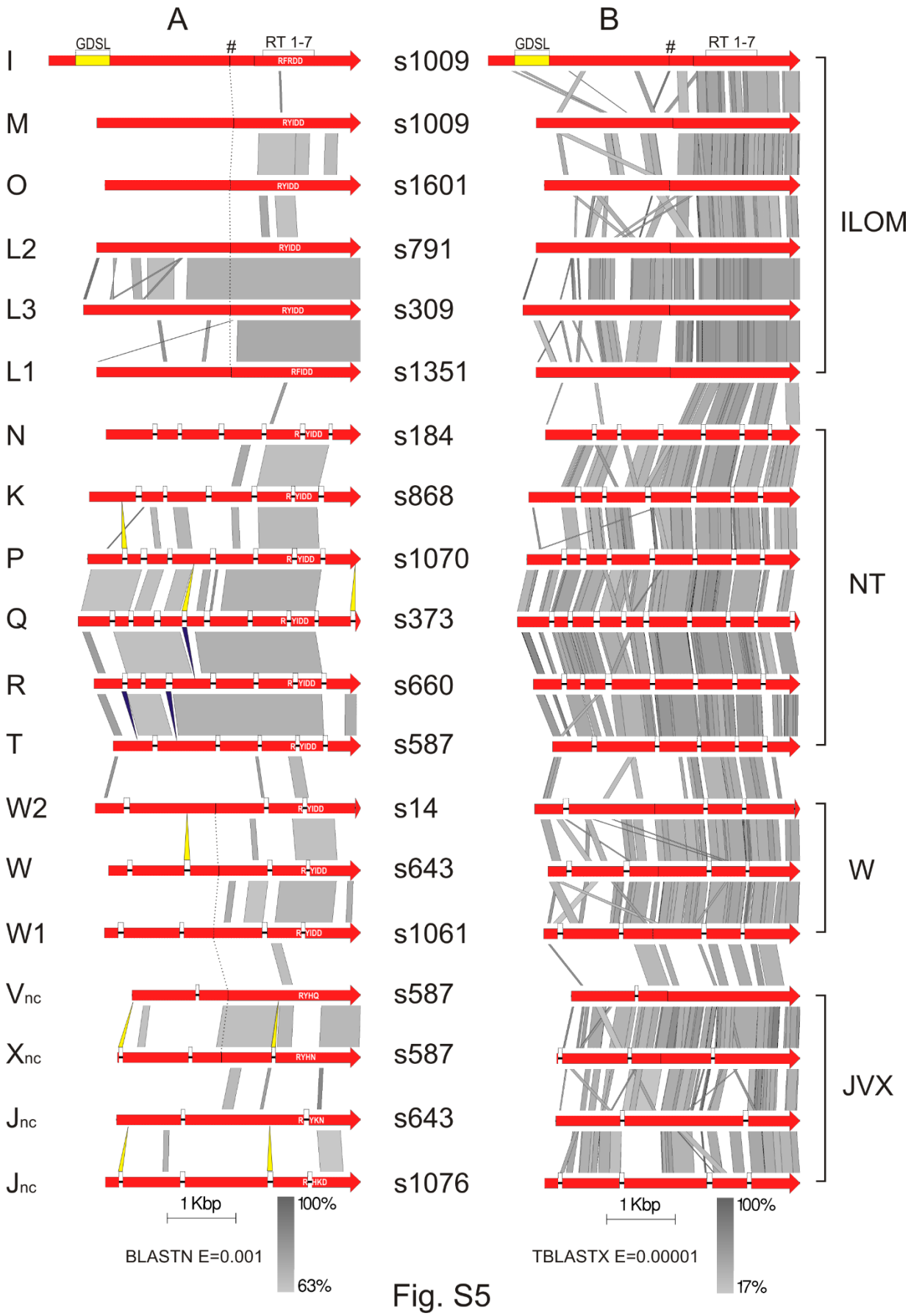
A
B
GDSL
#
RT 1-7
I
M
O
L2
L3
L1
N
K
P
Q
R
T
W2
W
W1
Vnc
Xnc
Jnc
Jnc
RFRDD
RYIDD
RFIDD
RYHQ
RYHN
RYKN
RHKD
s1009
s1009
s1601
s791
s309
s1351
s184
s868
s1070
s373
s660
s587
s14
s643
s1061
s587
s587
s643
s1076
ILOM
NT
W
JVX
1 Kbp
100%
63%
BLASTN E=0.001
1 Kbp
100%
17%
TBLASTX E=0.00001

Fig. S5

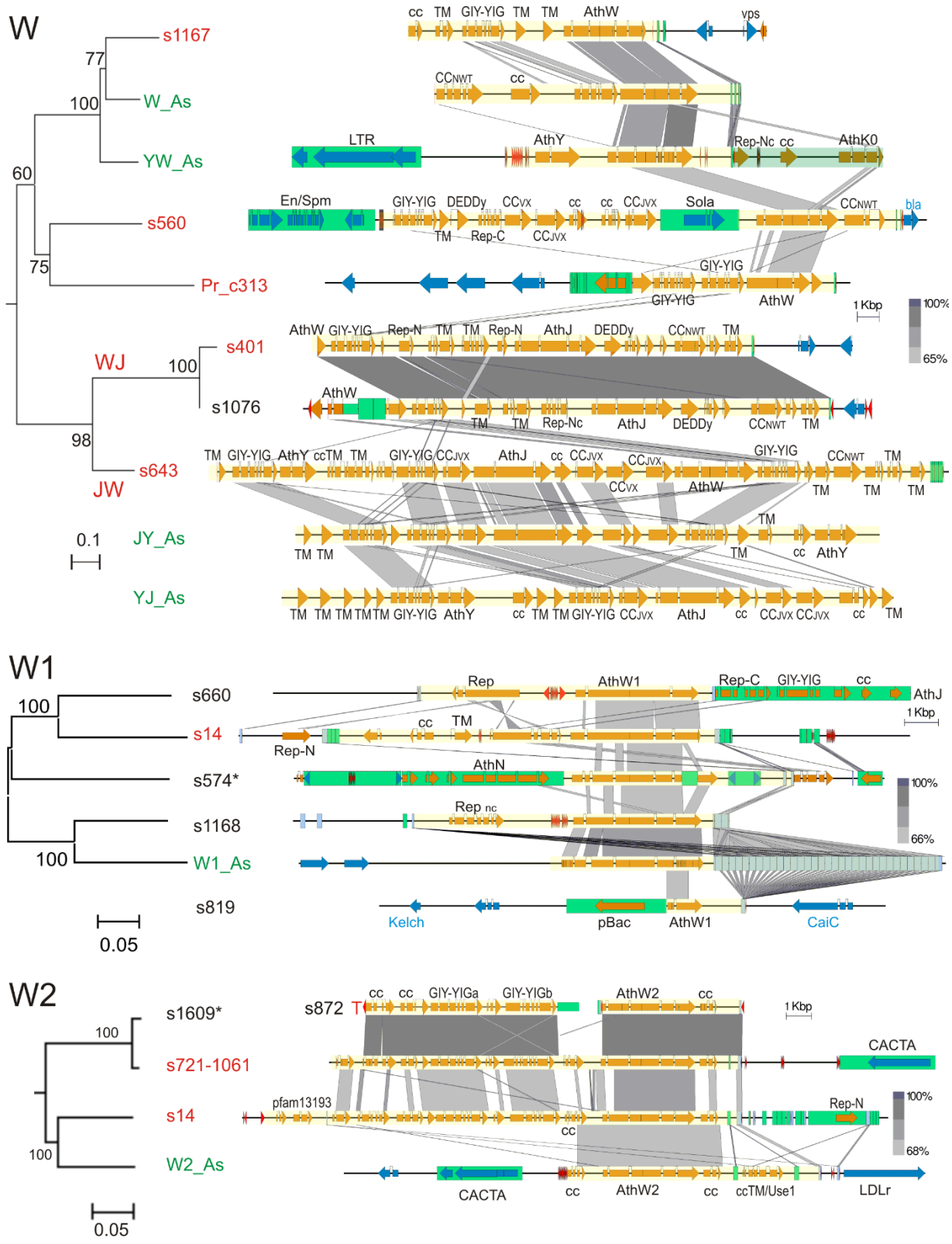
W
s1167
77
W_As
100
YW_As
60
s560
75
Pr_c313
WJ
100
s401
s1076
98
s643
JW
0.1
JY_As
YJ_As
W1
100
s660
s14
s574*
s1168
100
W1_As
s819
0.05
W2
100
s1609*
s721-1061
s14
100
W2_As
0.05
1 Kbp
100%
65%
66%
68%
LTR
AthY
Rep-Nc
AthK0
En/Spm
Sola
bla
Rep-C
AthW
AthJ
DEDDy
Rep-N
Rep-C
Rep
AthW1
AthN
Rep nc
Kelch
pBac
CaiC
s872
AthW2
CACTA
pfam13193
LDLr

Fig. S6 (A)

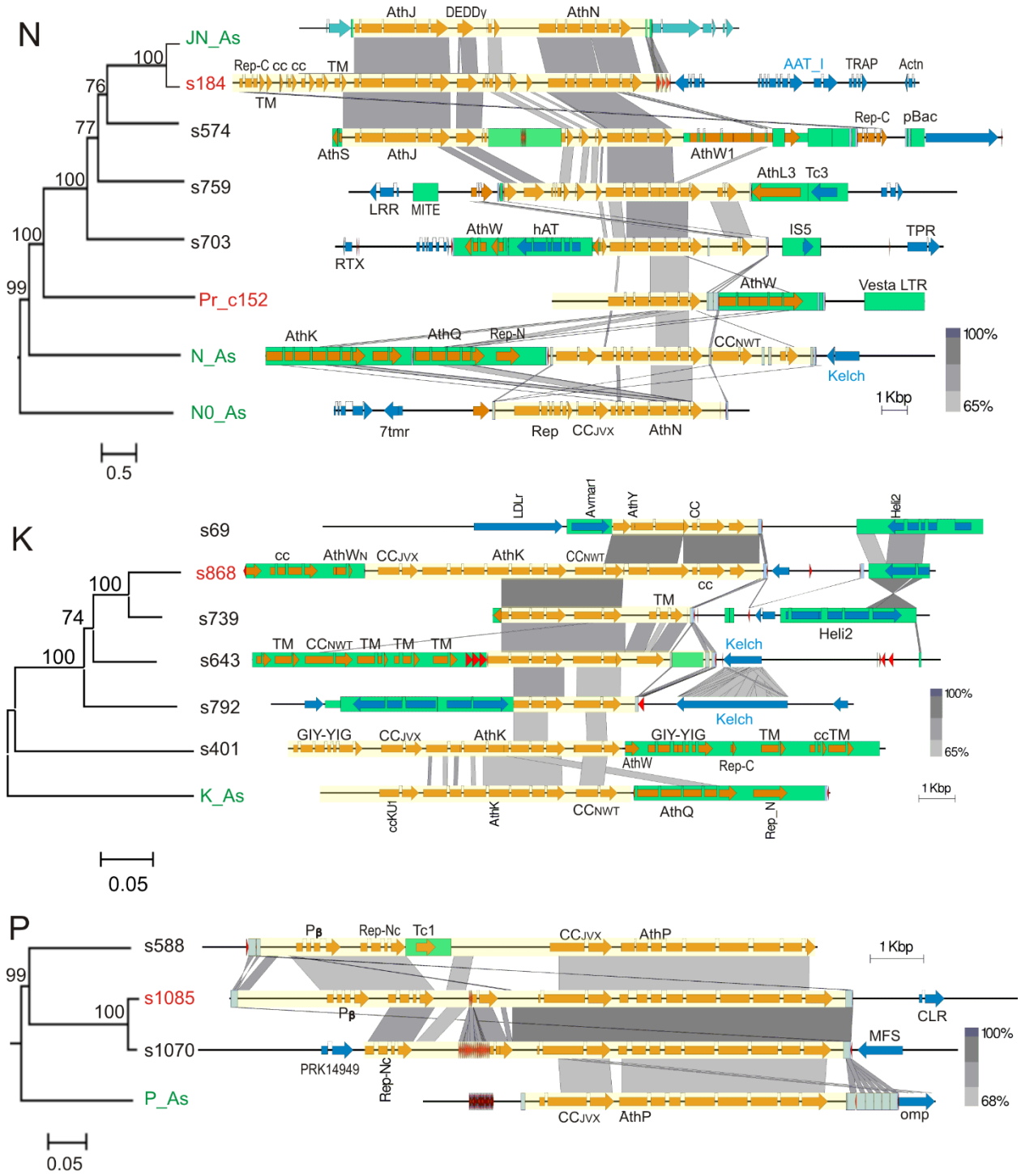
N
JN_As
s184
s574
s759
s703
Pr_c152
N_As
N0_As
100
76
77
100
100
99
0.5
AthJ
DEDDy
AthN
Rep-C cc cc
TM
TM
AAT_I
TRAP
Actn
Rep-C
pBac
AthS
AthJ
AthW1
AthL3
Tc3
LRR
MITE
AthW
hAT
IS5
TPR
RTX
AthW
Vesta LTR
AthK
AthQ
Rep-N
CCNWT
Kelch
100%
65%
1 Kbp
7tmr
Rep
CCJVX
AthN
K
s69
s868
s739
s643
s792
s401
K_As
100
74
100
0.05
LDLr
Avrmar1
AthY
CC
Heli2
cc
AthWN
CCJVX
AthK
CCNWT
cc
TM
Heli2
TM
CCNWT
TM
TM
TM
Kelch
Kelch
100%
65%
GIY-YIG
CCJVX
AthK
GIY-YIG
TM
ccTM
AthW
Rep-C
1 Kbp
ccKU
AthK
CCNWT
AthQ
Rep_N
P
s588
s1085
s1070
P_As
99
100
0.05
Pβ
Rep-Nc
Tc1
CCJVX
AthP
1 Kbp
Pβ
CLR
MFS
100%
68%
PRK14949
Rep-Nc
CCJVX
AthP
omp

Fig. S6 (B)

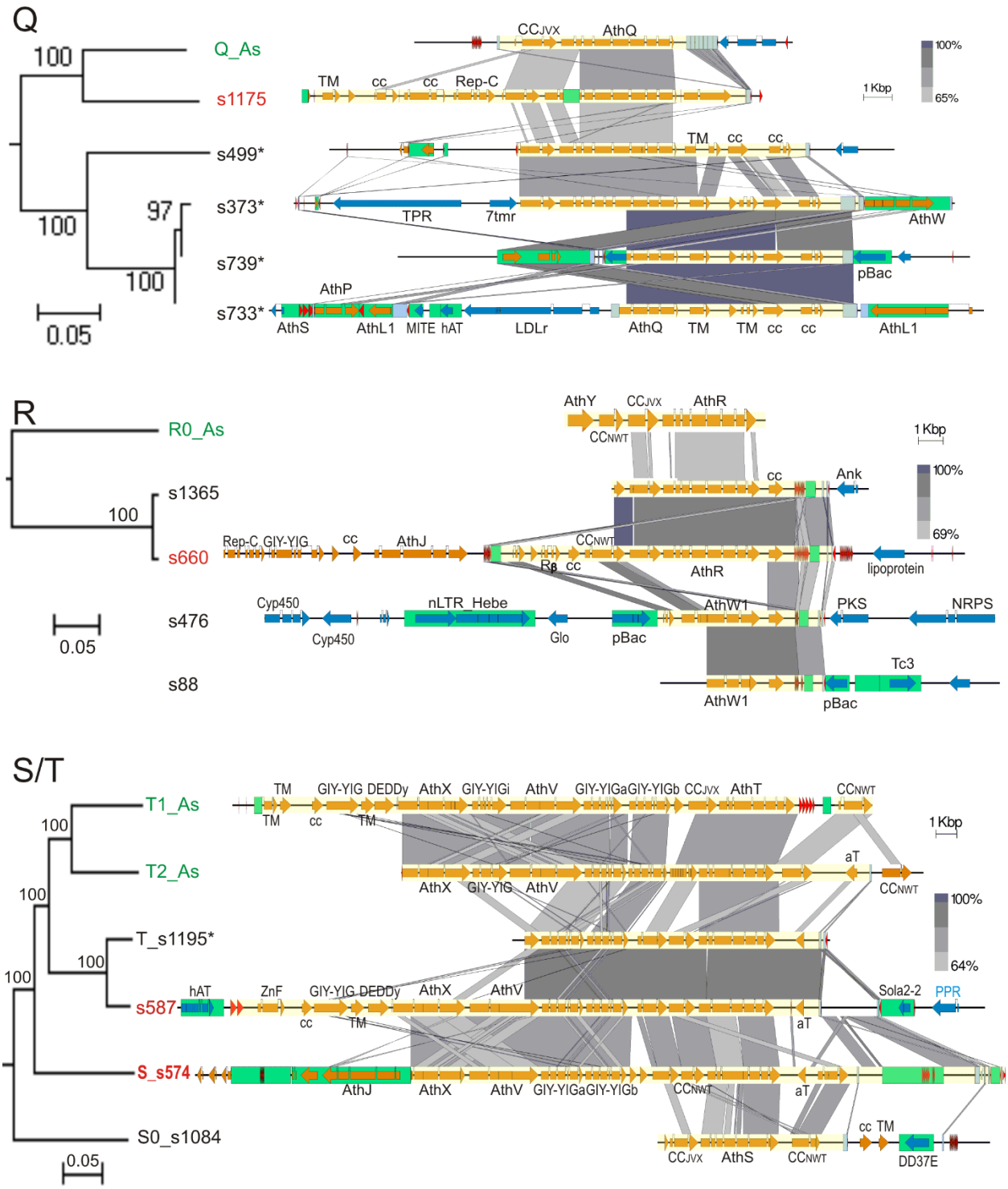
Q
100
Q_As
s1175
s499*
97
100
s373*
100
s739*
s733*
0.05
CCJVX
AthQ
TM
cc
cc
Rep-C
1 Kbp
100%
65%
TM
cc
cc
TPR
7tmr
AthW
pBac
AthP
AthS
AthL1
MITE
hAT
LDLr
AthQ
TM
TM
cc
cc
AthL1
R
R0_As
s1365
100
s660
s476
0.05
s88
AthY
CCJVX
AthR
CCNWT
1 Kbp
cc
Ank
100%
69%
Rep-C
GIY-YIG
cc
AthJ
CCNWT
Rβ
cc
AthR
lipoprotein
Cyp450
nLTR
Hebe
AthW1
PKS
NRPS
Cyp450
Glo
pBac
Tc3
AthW1
pBac
S/T
T1_As
100
T2_As
100
T_s1195*
100
100
s587
S_s574
S0_s1084
0.05
TM
GIY-YIG
DEDDy
AthX
GIY-YIGi
AthV
GIY-YIGa
GIY-YIGb
CCJVX
AthT
CCNWT
TM
cc
TM
1 Kbp
aT
AthX
GIY-YIG
AthV
CCNWT
100%
64%
hAT
ZnF
GIY-YIG
DEDDy
AthX
AthV
Sola2-2
PPR
cc
TM
aT
AthJ
AthX
AthV
GIY-YIGa
GIY-YIGb
CCNWT
aT
cc
TM
CCJVX
AthS
CCNWT
DD37E

Fig. S6 (C)

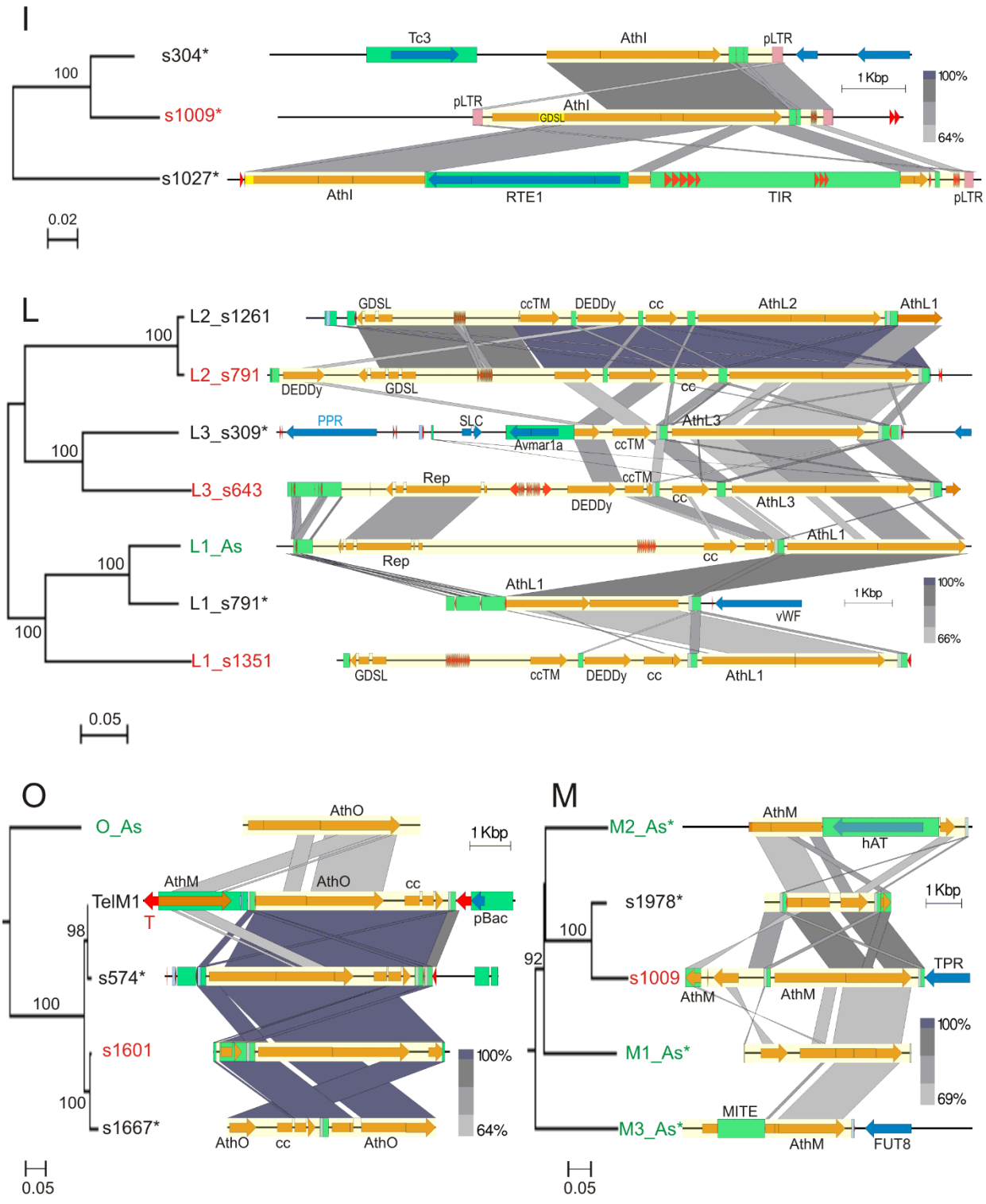
I
s304*
s1009*
s1027*
100
Tc3
AthI
pLTR
1 Kbp
100%
64%
GDSL
RTE1
TIR
0.02
L
L2_s1261
L2_s791
L3_s309*
L3_s643
L1_As
L1_s791*
L1_s1351
GDSL
ccTM
DEDDy
cc
AthL2
AthL1
PPR
SLC
Avmar1a
AthL3
Rep
vWF
66%
0.05
O
O_As
TelM1
s574*
s1601
s1667*
AthO
AthM
pBac
98
M
M2_As*
s1978*
s1009
M1_As*
M3_As*
hAT
TPR
MITE
FUT8
69%
92

Fig. S6 (D)

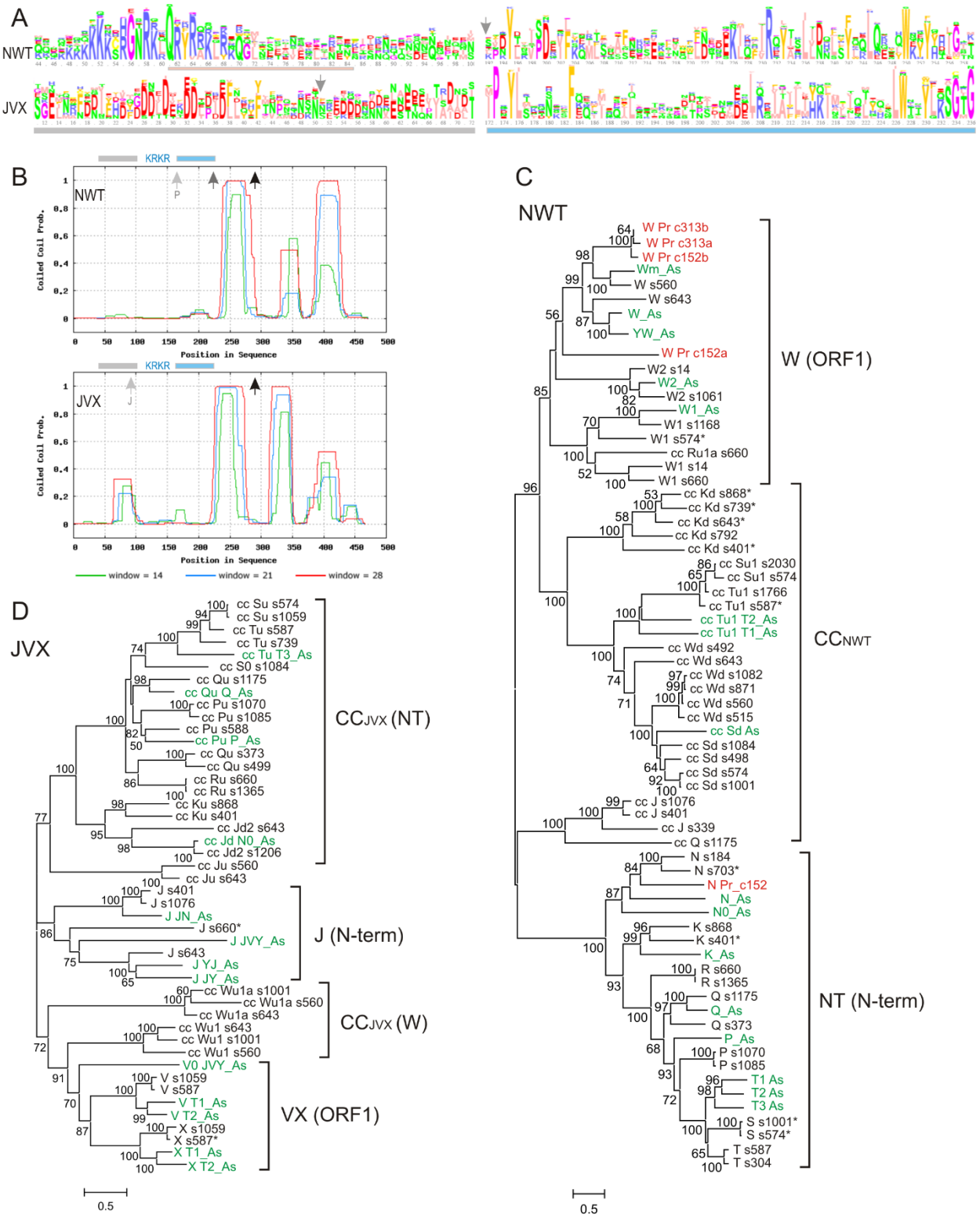
A
NWT
JVX
B
NWT
JVX
KRKR
Coiled Coil Prob.
Position in Sequence
window = 14
window = 21
window = 28
C
NWT
W (ORF1)
CCNWT
NT (N-term)
W Pr c313b
W Pr c313a
W Pr c152b
Wm_As
W s560
W s643
W_As
YW_As
W Pr c152a
W2 s14
W2_As
W2 s1061
W1_As
W1 s1168
W1 s574*
cc Ru1a s660
W1 s14
W1 s660
cc Kd s868*
cc Kd s739*
cc Kd s643*
cc Kd s792
cc Kd s401*
cc Su1 s2030
cc Su1 s574
cc Tu1 s1766
cc Tu1 s587*
cc Tu1 T2_As
cc Tu1 T1_As
cc Wd s492
cc Wd s643
cc Wd s1082
cc Wd s871
cc Wd s560
cc Wd s515
cc Sd As
cc Sd s1084
cc Sd s498
cc Sd s574
cc Sd s1001
cc J s1076
cc J s401
cc J s339
cc Q s1175
N s184
N s703*
N Pr_c152
N_As
N0_As
K s868
K s401*
K_As
R s660
R s1365
Q s1175
Q_As
Q s373
P_As
P s1070
P s1085
T1 As
T2 As
T3 As
S s1001*
S s574*
T s587
T s304
0.5
D
JVX
CCJVX (NT)
J (N-term)
CCJVX (W)
VX (ORF1)
cc Su s574
cc Su s1059
cc Tu s587
cc Tu s739
cc Tu T3_As
cc S0 s1084
cc Qu s1175
cc Qu Q_As
cc Pu s1070
cc Pu s1085
cc Pu s588
cc Pu P_As
cc Qu s373
cc Qu s499
cc Ru s660
cc Ru s1365
cc Ku s868
cc Ku s401
cc Jd2 s643
cc Jd N0_As
cc Jd2 s1206
cc Ju s560
cc Ju s643
J s401
J s1076
J JN_As
J s660*
J JVY_As
J s643
J YJ_As
J JY_As
cc Wu1a s1001
cc Wu1a s560
cc Wu1a s643
cc Wu1 s643
cc Wu1 s1001
cc Wu1 s560
V0 JVY_As
V s1059
V s587
V T1_As
V T2_As
X s1059
X s587*
X T1_As
X T2_As
0.5

Fig. S7

Table S1. Gene content at informative 3'-junctions of *Terminon* elements (within 10 kb of the 3'end). The remaining uninformative 3'-junctions contain little or no adjacent flanking DNA and are not shown. TR, telomeric repeats at 3'-ends. Color coding, genes of foreign origin (blue, bacterial; purple, fungal; green, plant; pink, protist) or TEs (yellow, PLE; orange, DNA TE). Highlighted copies are located on allelic scaffolds. +/-, sense or antisense CDS orientation; /, truncation.

Copy	Scaffold	TR	3'-CDS	+/-	Copy	Scaffold	TR	3'-CDS	+/-
JW	643	✔	/Athena-K	-	W	373	✔	TPR	-
JW	309	✔	/TPR	+	W	373	✔	SET	-
JW	693	✔	NAD-ADPRT/	-	W	1354	✔	NRPS	-
JW	574	✔	UDP-GTase	+	W	922	✔	/LDL receptor	-
WJ	664	✔	lectin	+	W	560	✔	/β-lactamase	+
WJ	399	✔	Kelch	-	W	476	✔	NRPS	-
WJ	1070	✔	TPR	-	W	1082	✔	RfbB epimerase	-
WJ	1571	✔	NRPS	-	W	664	✔	ABC transporter	-
WJ	1292	✔	TPR	-	W2	1041	✔	LRR	-
WJ	1203	✔	ankyrin	-	W1	88	✔	Pokey4 TPase/	-
WJ	1386	✔	hypothetical Av	+	W2	536	✔	GTase	-
WJ	1058	✔	GTase	-	W2	339	✔	vcbs/SxtP	-
WJ	3004	✔	AIG1	-	W2	552	✔	zf-C2H2/	-
WJ	860	✔	MTase	+	W2	214	✔	Ig-like	-
WJ	648	✔	GTPase	-	W2	274	✔	Ig-like	-
JN	184	✔	lectin	-	S	852	✔	NRPS	+
N	695	✔	NRPS	-	S	1254	✔	NRPS	+
N	703	✔	IS5	-	S	1001	✔	/Athena-W	+
N	759	✔	Athena-M/	-	S	498	-	vcbs	-
N,W	736	✔	GTase	-	S	974	✔	LDL receptor	-
K	868	✔	hypothetical	-	T	1084	✔	hypothetical	-
K	643	✔	Kelch/	-	T	792	✔	PAT1	-
K	792	✔	Kelch	-	T	209	✔	ABC ATPase	-
P	1070	✔	folate transport	-	T	404	✔	/Athena-K	-
P	1085	✔	lectin receptor	+	T	979	✔	acyltransferase	-
Q	739	✔	Looper TPase/	-	T	587	✔	Sola2-2	-
Q	88	✔	Merlin TPase	-	T	809	✔	vcbs	-
Q	339	✔	hypothetical	-	T	41	-	/filamin	+
Q	156	✔	Merlin TPase	-	T	1041	✔	F-box/LRR	-
R	660	✔	lipoprotein	-	T	304	✔	Penelope2	-
R	1365	✔	ankyrin	-	T	974	✔	LDL receptor	-
L1	560	✔	/Helitron	+	T	1457	✔	Athena-L	+
L1	791	-	cell surface prot	-	T	612	✔	FkbM MTase	+
L1	733	+/-	Athena-Q	-	T	1027	✔	NHL	+
L1	404	-	Avmar TPase	+	T	798	✔	nitroreductase	-
L2	866	✔	peptidase	-	M	1009	-	TPR/	-
L2	24	✔	/NHL	+	M	1224	-	MuDR TPase	-
L2	1015	-	hypothetical Av	+	M	1036	-	TLR-2	-
L2	1193	-	NRPS	-	M	809	-	ISL2EU TPase	+
L2	476	✔	7tm receptor	+	M	13	-	NAD-ADPRT	+
L2	1261	-	/Athena-L	+	M	1978	✔	Athena-M	+
L3	309	✔	hypothetical Av	-	O	574	✔	hAT TPase	+
L3	1045	✔	Sola2d		O	399	✔	Athena-J	+
L3	643	✔	Athena-JW		O	404	✔	Tcb1 TPase	+
L3	664	✔	peptidase		O	703	✔	Athena-L	+

Table S2. Differences in telomeric repeat distribution around different TE types.

Window size	2Kb		5Kb		10Kb	
ANOVA Tukey's comparisons	t value	p value	t value	p value	t value	p value
LTR – Helitron	-1.357	0.731	-1.042	0.89209	-1.687	0.5125
non-LTR – Helitron	-1.268	0.783	-1.588	0.57897	-1.135	0.8523
Athena – Helitron	4.01	<0.001 ***	5.88	< 0.001 ***	5.678	<0.001 ***
Penelope – Helitron	1.244	0.796	-0.021	1	0.163	1
TIR – Helitron	-0.679	0.982	-0.298	0.99964	0.925	0.9326
non-LTR – LTR	0.12	1	-0.453	0.99723	0.541	0.9936
Athena – LTR	4.804	<0.001 ***	6.173	< 0.001 ***	6.587	<0.001 ***
Penelope – LTR	2	0.317	0.605	0.98929	1.167	0.8368
TIR – LTR	1.076	0.879	1.006	0.90585	2.753	0.0582 .
Athena – non-LTR	4.819	<0.001 ***	6.819	< 0.001 ***	6.219	<0.001 ***
Penelope – non-LTR	1.945	0.349	0.912	0.93636	0.823	0.9584
TIR – non-LTR	0.966	0.92	1.675	0.52085	2.152	0.2394
Penelope – Athena	-1.144	0.848	-3.462	0.00619 **	-3.166	0.0170 *
TIR – Athena	-5.442	<0.001 ***	-7.391	< 0.001 ***	-6.111	<0.001 ***
TIR – Penelope	-1.63	0.551	-0.111	1	0.242	0.9999

A one-way ANOVA followed by Tukey's post hoc test was used to test for differences in the mean counts of telomeric repeats between different TE types (Heliron, LTR, non-LTR, Athena, Penelope and TIR) in three window sizes. TE type was found to have an impact on the number of nearby telomeric repeats: 2 Kb window, $F(5, 15847) = 7.366$, ***$P < 0.001$; 5 Kb window, $F(5, 15847) = 12.788$, ***$P < 0.001$; and 10 Kb window, $F(5, 15847) = 11.461$, ***$P < 0.001$. Post-hoc Tukey's test showed that Athena RTs have significantly higher counts of nearby telomeric repeats than LTR, non-LTR, TIR and Helitron elements, while differences with Penelope elements start at the 5 Kb window. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Adjusted p values are reported.

Supplementary Methods

Genome and transcriptome datasets. We used the *Adineta vaga* reference genome (Flot, et al. 2013) available in GenBank as unannotated scaffolds (CAWI000000000.2), for homology searches, and in the browser format at www.genoscope.cns.fr/adineta/cgi-bin/gbrowse/adineta to download the annotated scaffolds containing matches to *Athena* RTs (Gladyshev and Arkhipova 2007). Matching scaffolds were individually reannotated using GeneWise (www.ebi.ac.uk/Tools/psa/genewise/) and manually adjusted to introduce corrections into *Athena* RTs and associated ORF sequences, which were mostly misannotated due to the poor approximation of the *C. elegans*-trained gene models to TE-encoded ORFs. For the same reason, this procedure was also applied to genes of non-metazoan origin. The Sanger-sequenced *A. vaga* and *P. roseola* large-insert library clones were from Av_184A11, Av_119E19, Pr_152C9, Pr_313N6, PrTEL_IV_4, Pr_TEL_G (EU643486, EF485018, DQ138288, MF143428, EF485015, EF485006). The PacBio draft partial assembly was obtained from 15 SMRT cells using the HGAP assembler in the SMRT® Portal (Pacific Biosciences). TE consensus sequences reported in this study have been submitted to Repbase (Bao, et al. 2015). The *A. vaga* transcripts were uniquely mapped to the reference genome using Bowtie2 (Langmead and Salzberg 2012) on 75-nt Illumina GAIIx reads for cDNA (ERX234948), and using Bowtie (Langmead, et al. 2009) on 50-nt Illumina HiSeq small RNA reads (SRP070765). Aligned sequence reads were counted by annotated genomic feature with htseq-count (Anders, et al. 2015).

Bioinformatics. Programmed ribosomal frameshifting sites with simple pseudoknots were identified using KnotInFrame (Janssen and Giegerich 2015) and graphically represented using VARNA (Darty, et al. 2009). Secondary-structure-based multiple RNA alignments were performed with LocARNA (Smith, et al. 2010). Domain architecture was assessed with CDART (Geer, et al. 2002). Graphical representation of pairwise BLASTN and TBLASTX similarities between genomic scaffolds was done with Easyfig 2.2.2 (Sullivan, et al. 2011), and graphical overviews of genome-wide BLASTN searches of *A. vaga* scaffolds were obtained from NCBI server (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with the megablast option. Multiple sequence alignments were done using MUSCLE (Edgar 2004); for protein-coding sequences, alignments were codon-based. Sequence logos for multiple alignments with AlignmentViewer, prediction of coiled-coil motifs with COILS/PCOILS, and profile-profile homology searches with HHPRED were done using the MPI toolkit (Alva, et al. 2016). HHR motif alignment was created with Boxshade 3.3.1. Maximum likelihood phylogenetic analysis was done using the GTR+F+G model for nucleotide sequences and the WAG+F+G model for amino acid sequences, with a discrete gamma distribution used to model evolutionary rate differences across sites (G=5), and the resulting trees were edited in MEGA 7.0.18 (Kumar, et al. 2016). To identify HHR motifs characteristic of the L and WJ families, we used RNAMotif 3.1.1 (Macke, et al. 2001) with the descriptor from (Cervera and De la Pena 2014) relaxed to accommodate core-I: ss (minlen=7, maxlen=9); longer loop 2: ss (minlen=3, maxlen=16); and split core-II: ss (minlen=3, maxlen=6, seq="gaa$"), ss (minlen=3, maxlen=10, seq="nuh$"). Telomeric repeats in *A. vaga* were annotated using the sequence $(TGWGGG)_n$ and counted in different window sizes around the annotated genomic features of interest. The *A. vaga* gene set was divided into non-metazoan (foreign) and metazoan subsets, with the latter additionally subdivided to genes with and without piRNA coverage, as described in (Rodriguez and Arkhipova 2016). Statistical significance of the frequency of association between telomeric repeats and different TE types was estimated using single-factor ANOVA (Tukey's test). Significant values were assumed at $p < 0.05$.

Nucleic acid manipulations. Genomic DNA from rotifer eggs purified as in (Flot, et al. 2013) and ground in liquid N_2 was extracted with the Qiagen Genomic-Tip 500/G tissue protocol according to manufacturer's instructions with minor modifications, and checked for integrity by pulsed-field gel electrophoresis. After BluePippin size selection, it was used to construct a 20-kb library, which was sequenced on PacBio RS II at the Johns Hopkins University Deep Sequencing and Microarray Core with P6-C4 chemistry. The 462-bp JW-643 fragment with three identical tandem HHR-containing units from scaffold_643:36798..37259 was chemically synthesized by GenScript and cloned into the KpnI-BamHI sites of pBluescript II SK+. The plasmid template was linearized with BamHI and used for *in vitro* transcription. The reaction mix containing 40 mM Tris–HCl, pH 8.0, 6 mM $MgCl_2$, 2 mM spermidine, 0.5 mg/ml RNase-free BSA, 0.1% Triton X-100, 10 mM dithiothreitol, 1 mM ATP, 1 mM CTP, 1 mM GTP, 0.1 mM UTP, 0.5 µCi/µl [α-^{32}P]-UTP (Perkin-Elmer), 20 U RNase inhibitor (Roche), 20 µg/ml DNA, and 4 U/µl T7 RNA polymerase (Stratagene) was incubated at 37°C for 8.5 hrs. Labeled full-length transcripts were extracted from polyacrylamide gels and purified using RNA Clean & Concentrator™-5 kit (Zymo Research). Purified *in*

vitro transcripts were used to set up self-cleavage reactions in 50 mM Tris-HCl, pH 7.5, 20 U RNase inhibitor (Roche), heated at 95°C for 1 min, and pre-incubated at 25°C for 15 min. Self-cleavage was initiated by addition of $MgCl_2$ or $MnCl_2$, aliquots were taken at different time points, immediately placed on ice, mixed with the stop solution (8 M urea, 50% formamide, 50 mM EDTA, 0.1% xylene cyanol, 0.1% bromophenol blue), and run on 15% urea PAGE. Self-cleavage assays for *AvPen3a* HHR were essentially similar, except that the plasmid template was obtained by cloning of a 310-bp PCR-generated *AvPen3a* (scaffold 1009) using primers F (CGGGGTACCTGCAGTAAAAACAAAACAGAATGAAT) and R2 (CGCGGATCCTGCAGACGGTCCTTGATGTT), and self-cleavage was monitored by extension of the R2 primer end-labeled with [γ-^{32}P]-ATP (Perkin-Elmer) and T4 polynucleotide kinase (NEB).

Supplementary References

Alva V, Nam S-Z, Söding J, Lupas AN. 2016. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res 44:W410-W415.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166-169.

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

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.

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

Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25:1974-1975.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

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.

Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: protein homology by domain architecture. Genome Res 12:1619-1623.

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.

Janssen S, Giegerich R. 2015. The RNA shapes studio. Bioinformatics 31:423-425.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Meth 9:357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Research 29:4724-4735.

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.

Smith C, Heyne S, Richter AS, Will S, Backofen R. 2010. Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. Nucleic Acids Res 38:W373-W377.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27:1009-1010.

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO Journal 9:3353-3362.