





# An Ancient Clade of *Penelope*-Like Retroelements with Permuted Domains Is Present in the Green Lineage and Protists, and Dominates Many Invertebrate Genomes

Rory J. Craig <sup>\*,†,1</sup> Irina A. Yushenova <sup>†,2</sup> Fernando Rodriguez <sup>2</sup> and Irina R. Arkhipova <sup>\*,2</sup>

<sup>1</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom

<sup>2</sup>Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding authors: E-mails: iarkhipova@mbl.edu; rory.craig@ed.ac.uk.

Associate editor: Jeffrey Townsend

## Abstract

***Penelope*-like elements (PLEs) are an enigmatic clade of retrotransposons whose reverse transcriptases (RTs) share a most recent common ancestor with telomerase RTs. The single ORF of canonical endonuclease (EN)+ PLEs encodes RT and a C-terminal GIY–YIG EN that enables intrachromosomal integration, whereas EN– PLEs lack EN and are generally restricted to chromosome termini. EN+ PLEs have only been found in animals, except for one case of horizontal transfer to conifers, whereas EN– PLEs occur in several kingdoms. Here, we report a new, deep-branching PLE clade with a permuted domain order, whereby an N-terminal GIY–YIG EN is linked to a C-terminal RT by a short domain with a characteristic CxC motif. These N-terminal EN+ PLEs share a structural organization, including pseudo-LTRs and complex tandem/inverted insertions, with canonical EN+ PLEs from *Penelope/Poseidon*, *Neptune*, and *Nematis* clades, and show insertion bias for microsatellites, but lack canonical hammerhead ribozyme motifs. However, their phylogenetic distribution is much broader. The *Naiads*, found in numerous invertebrate phyla, can reach tens of thousands of copies per genome. In spiders and clams, *Naiads* independently evolved to encode selenoproteins containing multiple selenocysteines. *Chlamys*, which lack the CCHH motif universal to PLE ENs, occur in green algae, spike mosses (targeting ribosomal DNA), and slime molds. Unlike canonical PLEs, RTs of N-terminal EN+ PLEs contain the insertion-in-fingers domain (IFD), strengthening the link between PLEs and telomerases. Additionally, we describe *Hydra*, a novel metazoan C-terminal EN+ clade. Overall, we conclude that PLE diversity, taxonomic distribution, and abundance are comparable with non-LTR and LTR-retrotransposons.**

**Key words:** transposable elements, retrotransposons, reverse transcriptase, GIY–YIG endonuclease, selenoproteins, microsatellites.

## Introduction

Transposable elements (TEs) are characterized by their intrinsic ability to move within and between genomes. In eukaryotes, TEs contribute not only to the structural organization of chromosomes and variation in genome size, but also to the genetic and epigenetic regulation of numerous cellular processes (Wells and Feschotte 2020). TEs are traditionally divided into two classes, based on the presence (retrotransposons, class I) or absence (DNA transposons, class II) of an RNA intermediate in the transposition cycle. Retrotransposons, in turn, are divided into subclasses based on the presence or absence of long terminal repeats (LTRs): LTR-retrotransposons are framed by direct repeats; phylogenetically close DIRS elements by inverted or split direct repeats; non-LTR retrotransposons (aka LINES) lack terminal repeats; and *Penelope*-like elements (PLEs) have a special kind of repeats called pseudo-LTRs (pLTRs), which may be direct or inverted. Intrachromosomal integration in each subclass is associated with the combined action of phylogenetically

distinct clades of the reverse transcriptase (RT) domain fused to different types of endonuclease/phosphotransferase (EN) domains: DDE-type integrases (IN) and tyrosine recombinases (YR) in LTR-retrotransposons and DIRS, respectively; restriction enzyme-like (REL) or apurinic/apyrimidinic (AP) EN in non-LTR retrotransposons; and GIY–YIG EN in PLEs (Arkhipova 2017). The EN–RT fusion is either C-terminal or N-terminal, with the latter arrangement found in *copia*-like LTR-retrotransposons, in AP-containing non-LTR retrotransposons, and, as an exception, in the *gypsy*-like retrotransposon *Gmr1* (Eickbush and Malik 2002; Goodwin and Poulter 2002). The concerted action of RT and EN, which combines cleavage and joining of DNA strands with cDNA synthesis during retrotransposition, results in characteristic terminal structures that define the boundaries of new insertions.

The GIY–YIG EN domain typically associated with PLEs may have its evolutionary origins in bacterial group I introns, which are not retroelements (Stoddard 2014). The group I intron-encoded homing ENs are characterized by long

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

recognition sequences, and act essentially as monomeric nicksases, cleaving DNA on one strand at a time. The relatively short GIY–YIG cleavage module (~70 aa) is often tethered to additional DNA-binding domains for target recognition (Derbyshire et al. 1997; Van Roey et al. 2002). In eukaryotic PLEs, the activity of the recombinant GIY–YIG EN has been studied in vitro for *Penelope* elements of *Drosophila virilis*, where it displayed several properties expected from homology to prokaryotic enzymes, such as functional catalytic residues, nicking activity producing a free 3'-OH for RT priming, and moderate target preferences (Pyatkov et al. 2004). Variable distance between first-strand cleavage of DNA during target-primed reverse transcription (TPRT) and second-strand cleavage upon TPRT completion dictates the variable length of the target-site duplication, which is observed at the integration site. Phylogenetically, PLE ENs form a distinct cluster within a large GIY–YIG nuclease superfamily, where diverse homing ENs occupy a central position (Dunin-Horkawicz et al. 2006). PLE ENs are distinguished from those of homing ENs by the presence of a highly conserved CCHH Zn-finger motif, where the two cysteines are located directly between the GIY and YIG motifs (Arkhipova 2006).

Phylogenetic history of the longer RT domain is much more informative and reveals a sister relationship between PLEs and eukaryotic telomerase RTs (TERTs), which use a specialized RNA template to add G-rich repeats capping telomeres (Arkhipova et al. 2003). All described PLEs form two major groups: endonuclease-deficient (EN–) PLEs, retroelements found in several eukaryotic kingdoms at or near telomeres, and endonuclease-containing (EN+) PLEs, which harbor a C-terminal GIY–YIG EN enabling retrotransposition throughout the genome (Gladyshev and Arkhipova 2007). Three large EN+ PLE clades have been named *Penelope/Poseidon*, *Neptune*, and *Nematis*, the latter two being characterized by the presence of an additional conserved Zn-finger-like motif in the linker between RT and EN (Arkhipova 2006) (fig. 1). Two EN– PLE clades, *Athena* and *Coprina*, lack the EN domain entirely, but display a unique ability to attach to the exposed G-rich telomeric repeat overhangs, assisted by stretches of reverse-complement telomeric repeats combined with adjacent hammerhead ribozyme motifs (HHR) (Gladyshev and Arkhipova 2007; Arkhipova et al. 2017). Despite the ancient origin of PLEs predating their divergence from TERTs, which are pan-eukaryotic, the phylogenetic distribution of EN+ PLEs has so far been restricted to animals, with one exception of documented horizontal transfer to conifers (Lin et al. 2016). Here, we report the discovery of a novel deep-branching EN+ PLE clade, where the GIY–YIG EN is unexpectedly positioned N-terminally to the RT. A clade of these elements present in animals, termed *Naiad*, contains the GIY–YIG domain bearing the characteristic Zn-fingers found in canonical EN+ PLEs, whereas members of a second group, termed *Chlamys*, are present in green algae, spike mosses, and the slime mold *Physarum polycephalum*, and lack both Zn-finger motifs in EN. These results uncover the hitherto unknown PLE diversity, which spans all eukaryotic kingdoms, testifying to their ancient origins. We also report that *Naiads* from species as diverse as spiders and clams can

code for selenoproteins, which have not previously been described in any TEs.

## Results

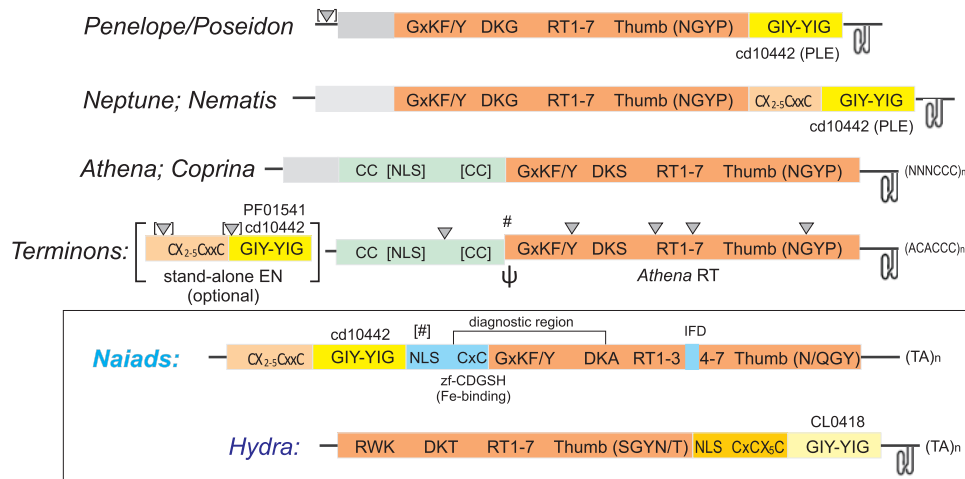
### Novel PLEs with N-Terminal Location of the GIY–YIG Endonuclease Domain

While cataloguing PLEs in several recently sequenced genomes, such as an acanthocephalan (*Pomphorhynchus laevis*) and a bdelloid rotifer (*Didymodactylos carnosus*), as well as a darwinulid ostracod (*Darwinula stevensoni*) (Mauer et al. 2020; Nowell et al. 2021; Schön et al. 2021), we noticed the absence of the GIY–YIG domain at the C-terminus of several PLEs, which is typically indicative of EN– PLEs. In these cases, however, extending the 5'-end of the frequently truncated PLE copies revealed a conserved N-terminal GIY–YIG EN domain, typically 220–275 aa in length. A high degree of 5'-truncation apparently precluded earlier identification of this novel type of PLEs. For instance, Repbase, a comprehensive database of eukaryotic TEs (Bao et al. 2015), contains two PLEs consistently appearing as top RT matches to the novel PLEs, yet having no N-terminal EN domain (*Penelope-2\_CGi* from the Pacific oyster *Crassostrea gigas* and *Penelope-1\_EuTe* from the Texas clam shrimp *Eulimnadia texana*). We extended the 767-aa *Penelope-2\_CGi\_1p* consensus in the 5'-direction and compared it with two sibling species, *Crassostrea virginica* and especially *Saccostrea glomerata*, where this element is mostly intact, revealing an N-terminal GIY–YIG domain which brings the total ORF length up to 1,024 aa in *S. glomerata* and to 1,000 aa in *C. gigas*.

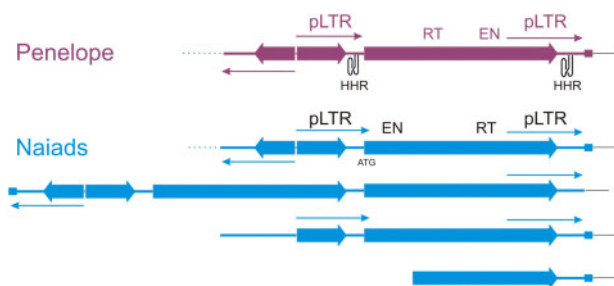
We then conducted an extensive database search for representatives of this previously undescribed type of PLEs in sequenced genomes, relying primarily on the N-terminal position of the GIY–YIG domain and several diagnostic motifs (see below) to discriminate between novel and canonical PLEs (fig. 1). Our search revealed a surprising diversity of hosts from eight animal phyla, including ctenophores, cnidarians, rotifers, nematodes, arthropods, mollusks, hemichordates, and vertebrates (fish) (supplementary file S1, Supplementary Material online). Additionally, about a dozen hits on short contigs were annotated as bacterial, however, upon closer inspection these were discarded as eukaryotic contaminants from metagenomic assemblies with an incorrect taxonomic assignment (Arkhipova 2020). Out of 36 animal host species, most were aquatic (26), six were parasitic, and only four were free-living terrestrial species (two spiders and two nematodes). We therefore chose the name *Naiad* for this newly discovered type of PLEs.

### Structural Characteristics of *Naiad* Elements

Structurally, *Naiad* insertions exhibit most of the previously known characteristic features of PLEs (Evgen'ev and Arkhipova 2005; Arkhipova 2006). Insertions show a high degree of 5'-truncation and are often organized into partial tandems, so that a full-length copy would be preceded by a partially truncated copy, forming a pLTR. Often, there is also an inverted 5'-truncated copy found immediately adjacent at the 5'-end, leading to formation of inverted repeats



**Fig. 1.** Domain architecture of the major PLE types found in animals. The newly described *Naiad* and *Hydra* clades are boxed. Shown are the most conserved amino acid motifs in the RT domain (in addition to the core RT motifs 1–7) and the characteristic Zn-finger-like motifs upstream of the GIY–YIG EN. Conserved introns are denoted by triangles; ORF1 can be separated by a frameshift (#) with a pseudoknot ( $\psi$ ); CC, coiled-coil; IFD, insertion in fingers domain; NLS, nuclear localization signal;  $(TA)_n$ , preferred targets;  $(ACACCC)_n$  or  $(NNNCCC)_n$ , short stretches of reverse-complement telomeric repeats in EN–PLEs. Optional elements are in square brackets. Also shown are the highest scoring PFAM/CD matches for the GIY–YIG EN, and the secondary structure of HHR motifs near the 3′-ends. Not to scale.



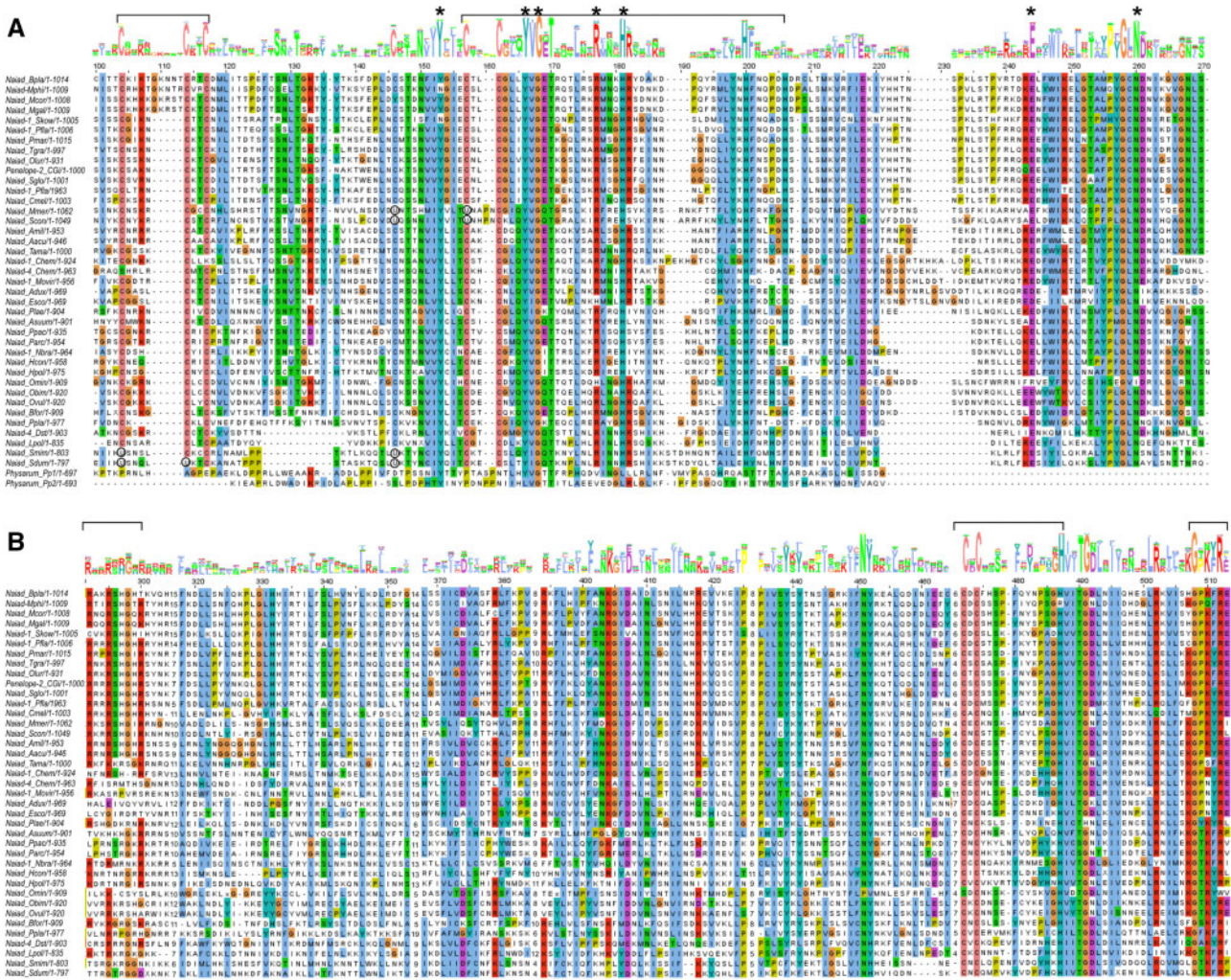
**Fig. 2.** Structural arrangements of PLE copies. Shown is the typical arrangement of two or more ORFs in partial tandems, forming pseudo-LTRs (pLTRs) denoted by arrows. An inverted 5′-truncated copy is often found adjacent to the upstream pLTR, forming inverted-repeat structures. Small squares denote a 30- to 40-bp extension (tail) that is often present only at one end of the insertion. HHR, hammer-head ribozyme motif. For *Penelope*, only the most typical structure is shown, but all other variants are also observed.

flanking the entire unit (fig. 2). Such complex structures of insertions often lead to problems in whole-genome shotgun (WGS) assemblies, especially short read-based. To further complicate boundary recognition, a 30- to 40-bp extension (tail) is often found at either end of the insertion unit, most likely resulting from EN-mediated resolution of the transposition intermediate. However, a notable difference between *Naiads* and other PLEs is the lack of canonical HHR, which are typically located within pLTRs (Cervera and de la Peña 2014; Arkhipova et al. 2017). Ignoring any tandemly inserted sequences, the main body of full-length *Naiad* copies are generally 3.4–4.4 kb.

Sequence conservation of the RT domain is strong enough to retrieve RTs of canonical PLEs in a BLAST search, thus for *Naiad* identification, it is practical to rely on several diagnostic regions, such as the CxC motif (showing weak homology to zf-CDGSH Fe-binding Zn-fingers) and the DKG motif

(Arkhipova 2006), which in *Naiads* is modified to DKA (fig. 1). In the core RT, the region between RT3(A) and RT4(B) is  $\sim 20$  aa longer than in other PLEs and corresponds in position to the IFD (insertion in the fingers domain) of TERTs (Lingner et al. 1997; Lue et al. 2003) (supplementary fig. S1, Supplementary Material online). Interestingly, the IFD is missing from *Naiads* in chelicerates (spiders and the horseshoe crab) and *D. stenosoni*, which resemble canonical PLEs in this region. Finally, between RT and the upstream EN domain, there is usually a large KR-rich block harboring a nuclear localization signal. This block, which is rich in adenines, is particularly prone to frameshift mutations resulting in detachment of the EN domain from RT and its eventual loss. Such mutations apparently prevented earlier recognition of the N-terminal EN domain in *C. gigas* and *E. texana* PLEs from Repbase (Bao et al. 2015).

The EN domain in *Naiads* displays most similarity to the GIY–YIG ENs of other PLEs (cd10442), especially to those in the *Neptune* and *Nematis* clades which harbor an additional conserved CX<sub>2-5</sub>CxC Zn-finger-like motif (lengthened by 5-aa insert in mussels) upstream from the GIY–YIG motif (figs. 1 and 3A; supplementary fig. S2, Supplementary Material online). Perhaps it may facilitate recognition of  $(TA)_n$  microsatellite sequences, which often serve as preferred targets for *Naiad* insertion. Its designation as a Zn-finger is tentative, as it shows variably nonsignificant matches to ZnF\_NFX, ZnF\_A20, ZnF\_TAZ, ZnF\_U1, or RING fingers in SMART database searches (Letunic et al. 2021). The CCHH Zn-finger-like motif with two cysteines inside the GIY–YIG core, characteristic of all canonical EN+ PLEs (Arkhipova 2006), is also present, and the catalytic domain beyond the GIY–YIG core is well conserved and includes the R, H, E, and N residues implicated in catalysis (Van Roey et al. 2002). Overall, despite the permuted arrangement of the RT and EN domains, *Naiads* share the peculiarities of structural



**FIG. 3.** Multiple sequence alignments of conserved domains characteristic for Naiads. (A) The GIY–YIG EN domain. The Zn-finger-like motifs are demarcated by square brackets; catalytic residues are denoted by asterisks; residues corresponding to selenocysteines (U) are circled. The position of the second H in the CCHH motif is variable. (B) The conserved diagnostic region between the EN domain and the GxKF/Y motif present in other PLEs. The KR-rich, CxC, and GxKF/Y motifs are marked by square brackets. The sequence order corresponds to that in figure 5.

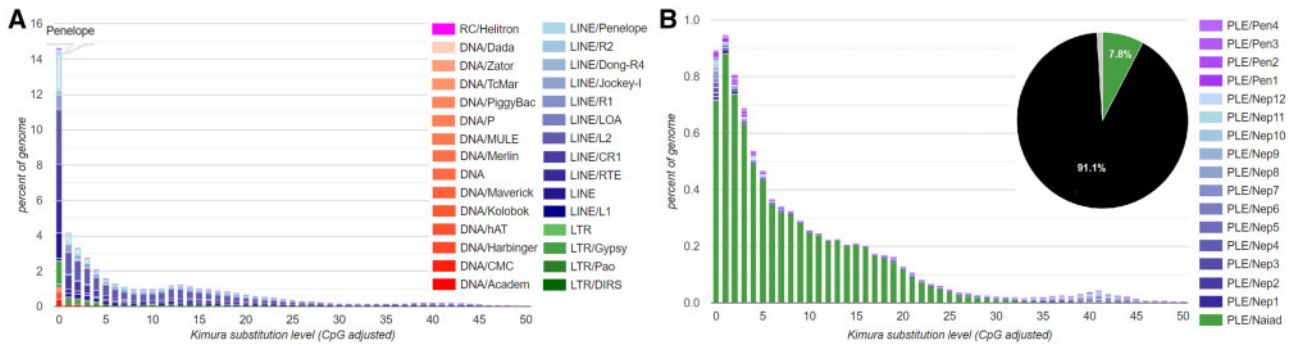
organization with other PLEs, indicating that their retrotransposition likely proceeds through a similar mechanism.

**Naiads Can Reach Exceptionally High Copy Numbers**  
 Although PLEs in animal genomes are typically outnumbered by non-LTR and LTR retrotransposons, we noticed that *Naiads* can be particularly successful in certain genomes in comparison to known PLE types. For example, inspection of TE landscape divergence profiles in the acanthocephalan *P. laevis* (fig. 4A) shows that PLE families are responsible for 8.9% of the genome, of which *Naiad\_Plae* occupies 7.8% (fig. 4B). The remaining 12 *Neptune* (PLE/Nep) and four *Penelope/Poseidon* (PLE/Pen) families combined occupy only 1.1% of the genome (fig. 4B).

We estimated copy numbers in each host species by querying each genome assembly with the corresponding *Naiad* consensus sequence and counting the number of 3'-ends at least 80 bp in length. This approach avoids counting multiple internal fragments in lower quality assemblies. Among hosts, significant variation in *Naiad* copy number can be observed, even between

closely related species (fig. 5). Copy numbers mostly reflect activity levels: some *Naiads* are apparently intact and are still successfully amplifying, whereas in other species, they have been inactivated a long time ago and required numerous ORF corrections to yield an intact consensus. Surprisingly, several marine invertebrates, such as oysters, clams, and crabs, harbor tens of thousands of *Naiad* copies, with nearly 37,000 in the blue crab *Paralithodes platypus* (fig. 5). At the same time, copy numbers can differ wildly within phyla: within Mollusca, the genomes of bivalves are dominated by *Naiads*, whereas in cephalopods, they have been inactivated a long time ago and are present only as remnants. The lack of canonical HHR motifs apparently does not interfere with the proliferative capacity of *Naiads*, as they can outnumber canonical HHR-bearing PLE families in the same species by several orders of magnitude, as in *P. laevis* (fig. 4B).

Given the abundance of *Naiads* in certain species, it is pertinent to comment on how these elements have been classified by automated pipelines. Currently, the lowest level classification for PLEs in RepeatMasker and RepeatFinder is non-LTR/Penelope



**Fig. 4.** Landscape divergence plots showing TE activity over time and genome occupancy in the acanthocephalan *Pomphorhynchus laevis*. (A) All TEs, with PLEs on the top (note RepeatMasker classification as LINE/Penelope); (B) PLEs only, subdivided by families, with *Naiad\_Plae* family (PLE/*Naiad*) occupying most of the space on the divergence plot and on the inserted pie chart.

and LINE/Penelope (see fig. 4A), respectively, meaning that the most accurate automated classification for *Naiads* would be as generic PLEs. We found that RepeatModeler2 (Flynn et al. 2020) automated repeat models associated with our curated *Naiad* families were generally classified as a mixture of unknowns and generic PLEs. For example, for the four *Naiad* families curated from the hydrozoan *Clytia hemisphaerica*, five automated models were classed as LINE/Penelope and ten as unknown. Similarly, for the four families from *D. stvensoni*, two models were classed as LINE/Penelope and 13 as unknown. Although fewer models were classed as PLEs they were longer and overlapped with regions corresponding to the RT domain, enabling their classification via protein homology to *Penelope-2\_CGi* and *Penelope-1\_EuTe* from Repbase (see above). The models classified as unknown were short and captured the highly abundant 3'-ends, often with redundancy. Indeed, automated approaches were sufficient to call PLEs as the most abundant TEs in the bivalves *Mytilus galloprovincialis* and *Ostrea edulis* (Vera et al. 2015; Murgarella et al. 2016) and it is now clear that much of this sequence was contributed by *Naiads* (fig. 5B). Nonetheless, incorporation of our *Naiad* consensus sequences into TE databases, and ideally the extension of PLE classification to the clade level, will be expected to improve the detection of *Naiads* and other PLE types in future genome projects.

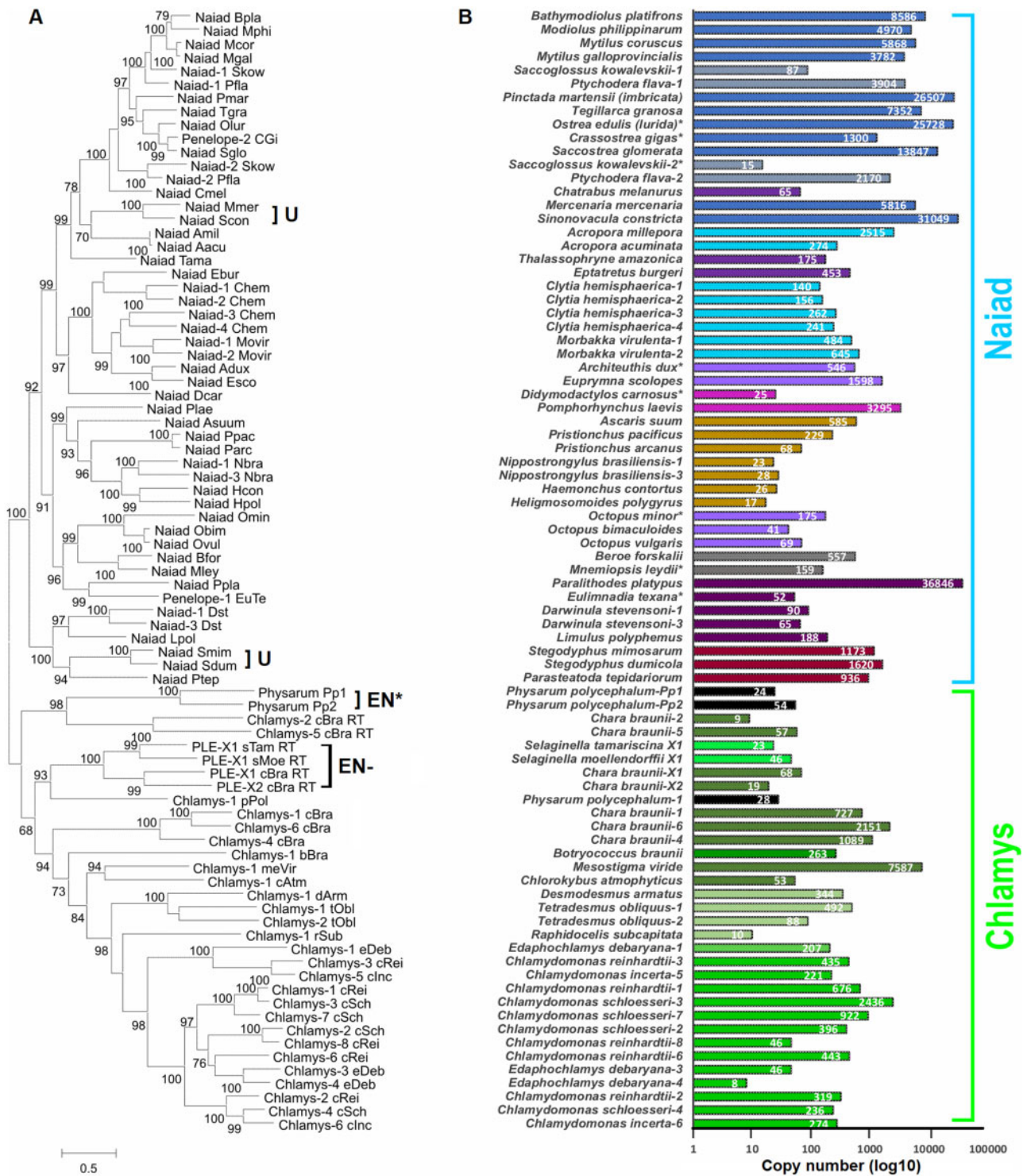
Finally, it is also evident that the *Naiad* phylogeny does not necessarily parallel that of the host species. Although some species, such as *C. hemisphaerica* or *D. stvensoni*, have experienced substantial within-species *Naiad* family diversification, others, such as hemichordates (*Saccoglossus kowalevskii*, *Ptychodera flava*), or cephalopods (*Architeuthis dux*, *Euprymna scolopes*, and *Octopus* spp.), harbor families belonging to different *Naiad* lineages (fig. 5 and supplementary file S1, Supplementary Material online). The fish *Naiads* (from *Chatrabus melanurus*, *Thalassophryne amazonica*, and *Eptatretus burgeri*) do not form a monophyletic clade, whereas the nematode or arthropod *Naiads* do (fig. 5A). The overall distribution pattern is suggestive of vertical inheritance punctuated by occasional horizontal transfer events and multiple losses.

### *Naiad* Selenoproteins in Clams and Spiders

The ORFs of four *Naiads*, from two clams (*Sinonovacula constricta* and *Mercenaria mercenaria*) and two spiders

(*Stegodyphus dumicola* and *Stegodyphus mimosarum*), each contained either three (*Naiad\_Smim* and *Naiad\_Mmer*) or four (*Naiad\_Sdum* and *Naiad\_Scon*) in-frame UGA codons. Except for one UGA codon in *Naiad\_Scon*, all UGA codons corresponded to highly conserved cysteines in the protein sequences of other *Naiads* (fig. 3A). In all families, UGA codons corresponded to the cysteine preceding the GIY-YIG motif, and to the cysteine eight aa downstream of the DKA motif (not shown in fig. 3). In spiders, UGA codons corresponded to either the first (*Naiad\_Smim*) or both the first and second (*Naiad\_Sdum*) cysteines in the CX<sub>2-5</sub>CxxC Zn-finger, whereas in clams, UGA codons corresponded to the first cysteine in the CCHH Zn-finger. The single remaining UGA codon in *Naiad\_Scon* corresponded to an aa in RT6(D) that was not strongly constrained.

Given the correspondence between the in-frame UGA codons and conserved cysteines, we hypothesized that the ORFs of these *Naiads* may encode selenoproteins, in which UGA is recoded from stop to selenocysteine (Sec). Recoding is achieved through a *cis*-acting selenoprotein insertion sequence (SECIS), a Sec-specific tRNA and additional *trans*-acting proteins (Berry et al. 1991; Tujebajeva et al. 2000). In eukaryotes, SECIS elements are located in the 3'-UTRs of selenoprotein mRNAs (Low and Berry 1996). Using SECISearch3 (Mariotti et al. 2013) to query each consensus sequence, we identified "grade A" (i.e., the highest confidence) type I SECIS elements in all four of the families (supplementary fig. S3A, Supplementary Material online). Except for *Naiad-Mmer*, the predicted SECIS elements were located immediately downstream of the inferred UAA or UAG stop codons (1–21 bp downstream, supplementary fig. S3B, Supplementary Material online), presumably placing the SECIS elements within the 3'-UTRs of each family. In *Naiad-Mmer*, the SECIS overlapped the first non-UGA stop codon, however, there was a UGA codon 7 bp upstream of the SECIS. The recoding of UGA is position dependent (Turanov et al. 2013) and a UGA codon in such close proximity to the SECIS is not expected to efficiently encode Sec (Wen et al. 1998), suggesting that this UGA codon may function as stop in *Naiad-Mmer*. Overall, the ORFs of each of the four families apparently encode selenoproteins that incorporate multiple Sec residues. Furthermore, following the phylogenetic



**Fig. 5.** Phylogenetic relationships between different *Naiad* and *Chlamys* families and copy number counts for each family. (A) Maximum likelihood phylogram based on the alignment of full-length ORFs with both EN and RT domains. Branch support values from 1,000 ultrafast bootstrap replications are shown. Families harboring selenocysteines are denoted by U, families lacking the EN domain by EN-, and families with EN remnants by EN\*. Scale bar, aa substitutions per site. (B) The copy number chart displays the counts for each family on a log scale. Similar shades denote similar taxonomic affiliations. Asterisks denote truncated or interrupted ORFs which are presumably nonfunctional. See [supplementary file S4, Supplementary Material](#) online, for protein sequences.

relationship of *Naiads* presented in [figure 5A](#), it is likely that the evolutionary transition to selenoproteins has occurred independently in spiders and clams.

### Structurally Diverse *Chlamys* Elements in the Green Lineage and Protists

As part of a recent annotation of TEs in the unicellular green alga *Chlamydomonas reinhardtii* and its close relatives ([Craig et al. 2021](#)), we identified novel PLE families with N-terminal GIY–YIG domains. These elements were termed *Chlamys*, although they were not further described. As with *Naiads*, the N-terminal EN+ PLEs in *Chlamydomonas* possess several of the defining features of canonical C-terminal EN+ PLEs, including genome-wide distributions, frequent 5′-truncation, and partial tandem insertions producing pLTRs. As introduced in the following text, *Naiad* and *Chlamys* share several features and collectively form a strongly supported N-terminal EN+ clade ([fig. 5A](#) and also [fig. 8](#) below), although *Chlamys* elements also possess characteristics that distinguish them from the metazoan *Naiad* clade.

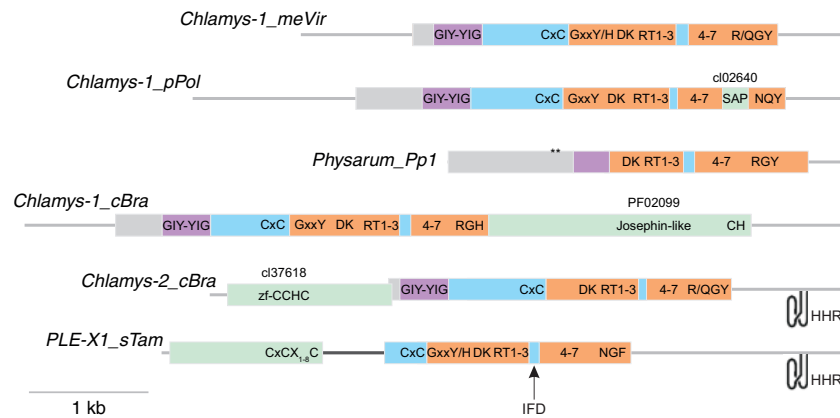
The predicted proteins of the *Chlamys* elements included the *Naiad*-specific Cx<sub>2</sub>zf-CDGSH-like Zn-finger motif and the IFD ([supplementary figs. S1 and S4, Supplementary Material](#) online). Additionally, *Chlamys* elements also lacked canonical HHRs (with two notable exceptions, [supplementary file S2, Supplementary Material](#) online and below), strengthening their evolutionary link to *Naiads*. As before, we used these conserved features to perform an extensive search for related PLEs in other taxa. We curated *Chlamys* elements from a wide diversity of green algae, including species from the Chlorophyceae order Sphaeropleales and the unicellular streptophyte algae *Mesostigma viride* and *Chlorokybus atmophyticus*. The Sphaeropleales and Chlamydomonadales are estimated to have diverged in the pre-Cambrian, whereas chlorophytes and streptophytes (which includes land plants) possibly diverged more than 1 Ga ([Del Cortona et al. 2020](#)). Additional curation identified more distantly related and structurally diverse families in the chlorophyte *Botryococcus braunii* (class Trebouxiophyceae), the multicellular streptophyte alga *Chara braunii*, two species of spike moss (genus *Selaginella*) and the myxomycete slime mold *P. polycephalum* (phylum Amoebozoa). Certain *Chlamys* families were also found in very high copy numbers, most notably in the genomes of the streptophytes *M. viride* and *C. braunii* ([fig. 5B](#)).

*Chlamys* elements were mostly longer (3.3–8.2 kb, not including pLTRs) and more structurally diverse than *Naiads* (see below). The length of several families was also increased by the presence of tandem repeats ([supplementary file S1, Supplementary Material](#) online). In the RT domain, the “DKG” motif was present as DK without a well-conserved third aa, and the IFD was generally longer (~20–40 aa) than that of *Naiads* ([supplementary figs. S1 and S4, Supplementary Material](#) online). Targeted insertion at (CA)<sub>n</sub> and (C)<sub>n</sub> repeats was observed for many *Chlamys* elements. Relative to *Naiads*, the most striking difference was in the EN domain. Although the GIY–YIG EN is N-terminal in both *Chlamys* and *Naiads*, in *Chlamys* both the linker domain harboring the CX<sub>2–5</sub>CxxC Zn-finger and the CCHH Zn-finger

motif are absent ([supplementary fig. S2, Supplementary Material](#) online). Thus, the EN of *Chlamys* differ from both *Naiads* and canonical C-terminal EN+ PLEs, all of which encode the CCHH motif (with the CX<sub>2–5</sub>CxxC Zn-finger absent in *Penelope/Poseidon*) ([fig. 1](#)). The conserved R, H, E, and N aa beyond the GIY–YIG core are all present in *Chlamys*. Finally, *Naiads* formed a well-supported clade to the exception of all *Chlamys* elements ([fig. 5A](#)). Collectively, *Naiad* and *Chlamys* are distinguished based on both taxonomic and structural features, and they can be considered as two major ancient groups that together comprise a wider N-terminal EN+ *Naiad/Chlamys* clade.

The minimal *Chlamys* domain organization, which is shared by most families in *Chlamydomonas*, the Sphaeropleales and the unicellular streptophytes, is represented by *Chlamys-1\_meVir* in [figure 6](#). Five families from *Chlamydomonas* encoded proteins with plant homeodomain (PHD) finger insertions, which were either located between RT2a and RT3(A) ([fig. 7](#)) or between the H and E conserved aa within EN. PHD fingers have been reported from TEs including CR1 non-LTR elements ([Kapitonov and Jurka 2003](#)) and *Rehavirus* DNA transposons ([Dupeyron et al. 2019](#)), where they may play a role in chromatin restructuring. PHD fingers are also present in several other retrotransposons and DNA transposons in *C. reinhardtii*, where it appears to be a common accessory domain ([Pérez-Alegre et al. 2005; Craig 2021](#)). Several additional domains were encoded by the more distantly related *Chlamys* elements. Two divergent organizations were observed in *P. polycephalum* families, the first of which included an SAP domain inserted between RT7(E) and the RT thumb (*Chlamys-1\_pPol*, [fig. 6](#)). SAP (SAF A/B, Acinus, and PIAS) is a putative DNA-binding domain that has previously been reported in *Zisupton* DNA transposons ([Böhne et al. 2012](#)). The second type included the element *Physarum\_Pp1*, which was first described from a 5′-truncated consensus as an unusual PLE with an IFD ([Gladyshev and Arkhipova 2007](#)). Extending the consensus sequence revealed a predicted protein with a reduced N-terminus that entirely lacked the CxC motif present in all other *Chlamys* and *Naiads*, and included a reduced EN domain in which the GIY–YIG motif was present but weakly conserved and the region containing the conserved R, H, E, and N aa was absent ([figs. 3A and 6](#)). Although the *P. polycephalum* genome is highly fragmented, *Physarum\_Pp1* does appear to be present genome wide.

EN+ PLEs were reported from the *C. braunii* genome project, although [Nishiyama et al. \(2018\)](#) did not further describe these elements. We observed three distinct types of *Chlamys* in *C. braunii*. The first possessed long ORFs (~1,800 aa) encoding peptides with a C-terminal extension including a motif with weak homology to Josephin and a second unknown motif with several well-conserved C and H aa (*Chlamys-1\_cBra*, [fig. 6](#)). Josephin-like cysteine protease domains are present in *Dualen* non-LTR elements, where they may play a role in disrupting protein degradation ([Kojima and Fujiwara 2005](#)). The second type featured a canonical 3′-HHR and an upstream ORF encoding a peptide with a gag-like zinc-knuckle domain (zf-CCHC, *Chlamys-*



**Fig. 6.** Structural diversity of *Chlamys* elements. Domain architecture of *Chlamys* elements is represented by to scale schematics. The thin lines represent sequences not present in ORFs. Domain designations are as in figure 1, except for the divergent GIY–YIG EN which in *Chlamys* elements lacks the CCHH motif present in *Naiad* EN. The most conserved amino acid motifs and the highest scoring PFAM/CD domain matches are also shown. The asterisks on the *Physarum* model represent in-frame stop codons, which may indicate the presence of an undetected intron. Note the *Physarum* EN-like domain is also reduced and weakly conserved (fig. 3A). The dark thick line in *PLE-X1\_sTam* represents an intron that was inferred from *Selaginella moellendorffii* annotated gene models.

*2\_cBra*, fig. 6). The third type was notable since related elements were also identified in spike mosses, and a small number of highly significant BlastP results were recovered from moss species, potentially indicating a wider distribution in “early diverging” plants. These families also feature canonical HHRs (supplementary file S2, Supplementary Material online), and their proteins contain the CxC motif but lack the GIY–YIG EN, with a unique N-terminal extension that is likely separated by an intron and includes a conserved CxCX<sub>1–8</sub>C motif (*PLE-X1\_sTam*, fig. 6). The families in *C. braunii* appeared to have genome-wide distributions, and remarkably, the two spike moss families exhibited targeted insertions at a precise location within 28S ribosomal RNA genes. The insertion target differed by only 4 bp between the families (supplementary fig. S5, Supplementary Material online), suggesting deep conservation of the target sequence at least since the divergence of *Selaginella moellendorffii* and *Selaginella tamariscina* ~300 Ma (Xu et al. 2018). Metazoan ribosomal DNA is a well-documented insertion niche for R-element non-LTRs and the piggyBac DNA transposon *Pokey* (Eickbush and Eickbush 2007), although to our knowledge this is the first example from both plants and PLEs. It remains to be seen how this group achieves either genome-wide or targeted ribosomal DNA insertion without an identified EN. Interestingly, these families form a well-supported clade with the EN+ family from *P. polycephalum* (*Chlamys-1\_pPol*, fig. 5A), potentially indicating secondary loss of the GIY–YIG EN.

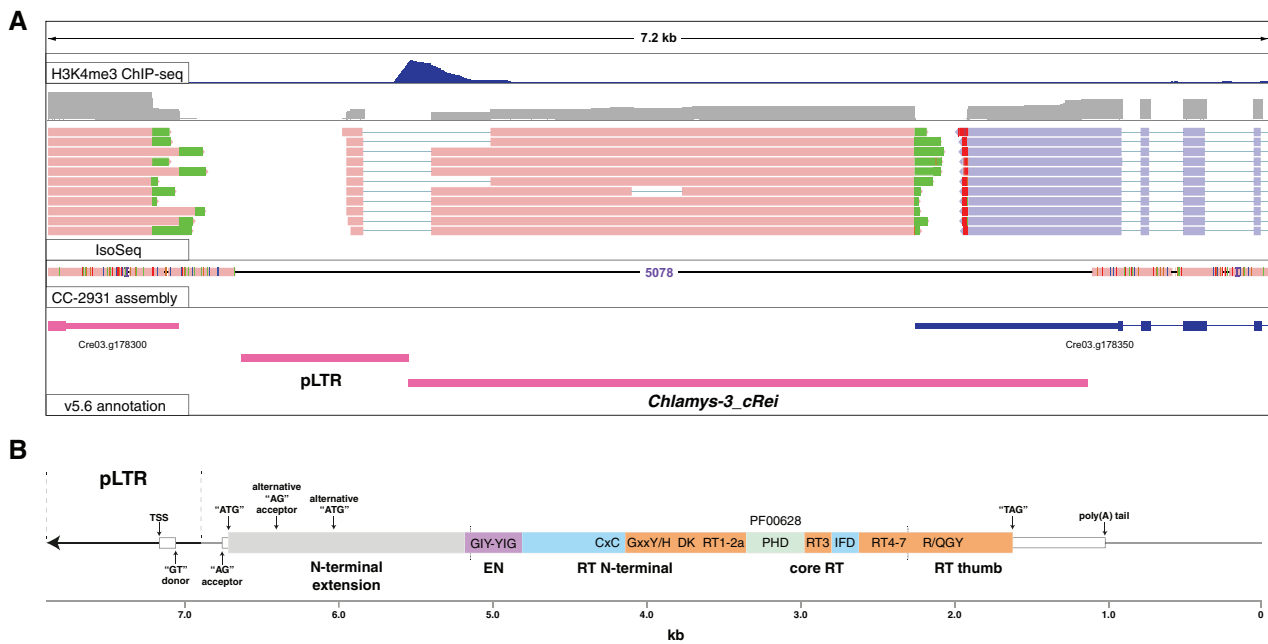
### Functional Characterization of an Active *Chlamys* Element

As high-quality functional data are available for *C. reinhardtii*, we further focused on the ten *Chlamys* families curated in this species. As with the case of two *Naiads* described above, four of these families are represented by truncated models in Repbase, two of which are annotated as unknown non-LTRs and two erroneously as non-LTR *L1* elements (supplementary file S1, Supplementary Material online). Notably, we

also identified putatively nonautonomous *Chlamys* elements, which produced pLTRs (and often multicopy head-to-tail insertions) and generally exhibited sequence similarity to autonomous families at their 3′-ends. These include *MRC1*, which was previously described as a nonautonomous LTR (Kim et al. 2006), most likely due to the observation of the pLTR, and is among the most active TEs in *C. reinhardtii* laboratory strains (Neupert et al. 2020). Further supporting recent activity, *Chlamys* copies exhibited minimal divergence from their respective consensus sequences (supplementary fig. S6, Supplementary Material online) and within-species polymorphic insertions were observed for copies of all ten autonomous families by comparison to a newly assembled PacBio-based genome of the divergent field isolate CC-2931 (supplementary file S3, Supplementary Material online). Cumulatively, *Chlamys* PLEs spanned ~1.6% of the 111 Mb *C. reinhardtii* genome and comprise ~15% of the total TE sequence.

Only one active C-terminal EN+ PLE has been experimentally characterized, the archetypal *Penelope* of *D. virilis* (Pyatkov et al. 2004; Schostak et al. 2008). In an attempt to characterize a *Chlamys* element, we searched for an actively transcribed copy using recent PacBio RNA-seq (i.e., Iso-Seq) and H3K4me3 ChIP-seq data sets (Gallaher et al. 2021), with the H3K4me3 modification reliably marking active promoters in *C. reinhardtii* (Ngan et al. 2015). Due to frequent 5′-truncation, only two families were found with full-length copies, and transcription was observed for only a single copy of the *Chlamys-3\_cRei* family (fig. 7A). Unfortunately, this copy features a 2.8 kb deletion, although this is entirely within the ORF and the copy presumably retains a functional promoter, transcription start site (TSS), and terminator. Strikingly, the derived gene model of *Chlamys-3\_cRei* (fig. 7B) shared several features with *Penelope*, in which the pLTR harbors the TSS and a 75 bp intron within the 5′-UTR that overlaps the internal promoter (Arkhipova et al. 2003; Schostak et al. 2008). In *Chlamys-3\_cRei*, the TSS is also located in the pLTR and a 398 bp intron within the 5′-UTR spans the





**Fig. 7.** Functional characterization of an active *Chlamys* element in *Chlamydomonas reinhardtii*. (A) IGV browser view (Robinson et al. 2011) of a *Chlamys-3\_cRei* copy that is polymorphic between the reference genome and the divergent field isolate CC-2931. The mismatched bases on the extremities of Iso-Seq reads represent poly(A) tails of transcripts. The annotated gene on the right is on the reverse strand. (B) Schematic of the inferred gene model and structural organization of *Chlamys-3\_cRei*. Note that this represents the full-length element and the transcribed copy above contains a 2.84 kb internal deletion, the boundaries of which are shown by the dashed vertical lines. Domains are denoted as in figure 6 and conserved motifs are textually represented as shown for *Chlamys-1\_meVir* in that figure.

boundary between the pLTR and downstream main body. The H3K4me3 ChIP-seq supports an internal promoter coinciding with the intron. Additionally, three Iso-Seq reads supported an alternative isoform with a 751 bp intron. This isoform initiates at a downstream start codon and results in a peptide truncated by 293 aa, although as the predicted *Chlamys-3\_cRei* peptide includes an N-terminal extension both isoforms encode complete EN and RT domains. The similarities between *Penelope* and *Chlamys-3\_cRei* potentially indicate an ancient and deeply conserved organization and perhaps mechanism shared by canonical PLEs and the N-terminal EN+ PLEs described herein.

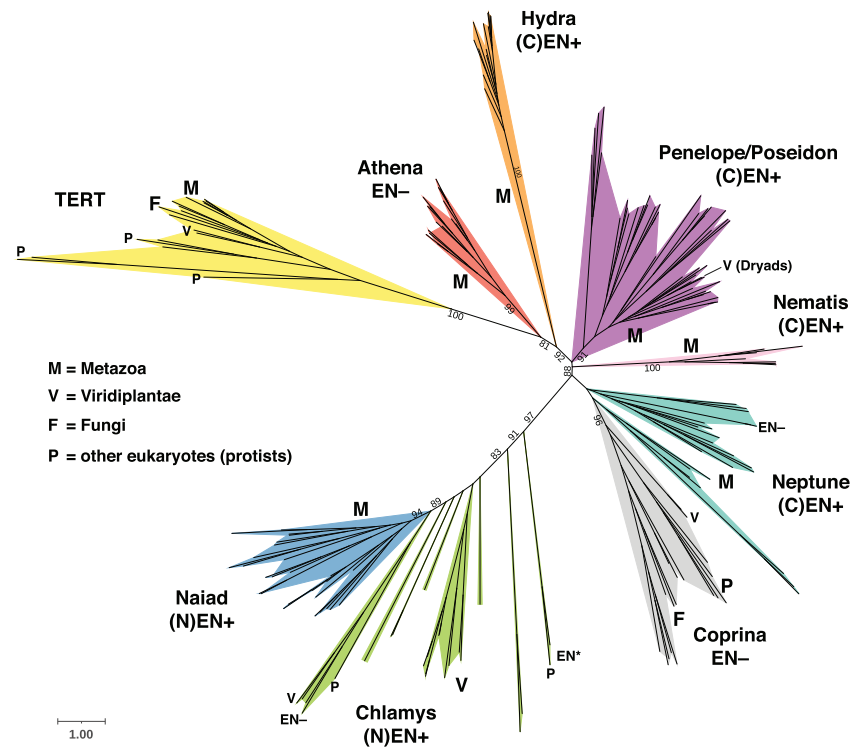
### *Hydra*: A Novel C-Terminal EN+ Clade

While performing an updated phylogenetic analysis of all PLEs (see below), we noticed that 21 C-terminal EN+ families in Repbase formed an isolated group highly divergent from *Neptune*, *Penelope/Poseidon*, and *Nematis*. Except for two families from the hemichordate *S. kowalevskii*, all were annotated from Cnidarian species, with 14 from the freshwater polyp *Hydra magnipapillata* and four from the starlet sea anemone *Nematostella vectensis* (supplementary file S1, Supplementary Material online). Using protein homology searches, we identified a small number of additional families in other aquatic invertebrates spanning four phyla (Cnidaria, Mollusca, Echinodermata, and Arthropoda), notably in species such as the stony coral *Acropora millepora* and the sea cucumber *Apostichopus japonicus*. These elements were generally short (<3 kb) and contained single ORFs encoding peptides with several similarities to canonical C-terminal EN+ PLEs, that is, no CxC motif, no IFD, and a C-terminal

GIY-YIG EN (supplementary fig. S7, Supplementary Material online). Canonical HHRs were readily detected (supplementary file S2, Supplementary Material online), strengthening the relationship with canonical PLEs. However, these families also exhibited unique features. The N-terminal GxKF/Y motif was not conserved and a RWK motif was present, the DKG motif was modified to DKT, and RT4(B) was particularly divergent and challenging to align (fig. 1 and supplementary fig. S7, Supplementary Material online). Most notably, in the EN domain the CCHH motif universal to C-terminal EN+ PLEs (and *Naiads*) was absent (supplementary figs. S2 and S7, Supplementary Material online). A linker domain was present that was most similar to that of *Nematis*, although the typical CX<sub>2-5</sub>CxC Zn-finger was modified to a CxCX<sub>5</sub>C motif. Furthermore, possible CxC and CxH motifs were found on either side of the conserved E and N aa (supplementary fig. S7, Supplementary Material online). Interestingly, almost all families exhibited insertions into (TA)<sub>n</sub>, strengthening the association between the linker domain and targeted insertion. In a phylogenetic analysis of these new elements, all but one family from *H. magnipapillata* formed a weakly supported clade, although generally Cnidarian families were distributed across the phylogeny (supplementary fig. S8, Supplementary Material online). We name this new clade of C-terminal EN+ PLEs *Hydra*, in line with both their aquatic hosts and their discovery in *H. magnipapillata*.

### Evolution of the RT and GIY-YIG EN Domains

As seen in figure 5, the newly discovered types of PLE span much of the well-sequenced taxonomic diversity in Eukarya,



**Fig. 8.** Core RT maximum likelihood phylogeny of PLE RTs and TERTs. Support values from 1,000 ultrafast bootstrap replications are shown, with all values from nodes within major clades and any values  $<70$  at deeper nodes excluded to aid visualization. The taxonomic range of clades and subclades is shown by letters. Note that the “V” marking the “*Dryad*” subclade within *Penelope/Poseidon* points at three conifer families from the presumed horizontal transfer event (Lin et al. 2016). The location of EN in EN+ groups is provided by the prefixes (N) and (C) for N-terminal and C-terminal, respectively. Subclades with EN remnants or no EN that are within EN+ clades are shown by EN\* and EN– tags, respectively. Scale bar, aa substitutions per site. The phylogeny was annotated using iTOL (Letunic and Bork 2019).

including protists, plants, and animals. We placed *Naiad*, *Chlamys*, and *Hydra* representatives into a reference PLE data set that included the previously known EN+ *Penelope/Poseidon*, *Neptune*, *Nematis*, and EN– *Athena* and *Coprina* clades, as well as representatives of the sister clade to PLEs, the TERTs (Arkhipova 2006; Gladyshev and Arkhipova 2007). The combined phylogeny of the extended core RT domain, which also includes the RT thumb and the previously identified N-terminal conserved motifs N1–N3 (Arkhipova 2006), is presented in figure 8. Except for *Neptune*, all of the above clades were recovered with ultrafast bootstrap support values  $>90\%$ . *Neptune* elements formed a paraphyletic group in a weakly supported clade with the taxonomically diverse EN– *Coprina* elements, as also occurred in a previous analysis (Arkhipova et al. 2017). The novel N-terminal EN+ elements (i.e., *Naiad* + *Chlamys*) formed a strongly supported clade, although *Chlamys* was paraphyletic with respect to the *Naiad* clade and the internal topologies of the more structurally diverse *Chlamys* elements were not well supported (although inclusion of the EN domain in fig. 5A supported *Chlamys* monophyly). Despite its potential paraphyly, we still consider *Chlamys* to be a useful grouping given its unique structural features. The rotifer-specific EN– *Athena* elements formed the most basal PLE clade when rooting the phylogeny on TERTs, although the deep branches linking the major PLE clades generally received weak support. As seen in both *Chlamys* and *Neptune*, EN– families occasionally emerge

within EN+ clades, apparently as a result of EN loss accompanied by acquisition of an alternative way of employing accessible 3'-OH ends for RT priming. These results were broadly supported by a complementary CLANS protein clustering analysis (Frickey and Lupas 2004) of the same sequences (supplementary fig. S9, Supplementary Material online). RTs from both EN+ and EN– canonical PLEs formed a single cluster, with each clade clearly defined as a subcluster. Similarly, *Naiad* and *Chlamys* formed subclusters within a larger cluster, although the *Chlamys* elements were less strongly linked. Most notably, *Hydra* formed a cluster highly distinct from both canonical PLEs and *Naiad/Chlamys*, highlighting the divergence of this group and the uncertainty of its phylogenetic placement in figure 8. Overall, it is evident that, with inclusion of the new superfamilies, PLE RTs display an astonishing level of clade diversity, which is comparable with that of non-LTR and LTR retrotransposon RTs, and will undoubtedly increase with the number of sequenced genomes from underrepresented eukaryotic branches of the tree of life.

In light of the increased diversity uncovered by *Naiad*, *Chlamys*, and *Hydra* PLEs, we also attempted to further elucidate the evolutionary relationships of the GIY–YIG EN domain. Dunin-Horkawicz et al. (2006) found that the most similar ENs to those from canonical C-terminal EN+ PLEs belonged to the HE\_Tlr8p\_PBC-V\_like group (cd10443), which includes homing ENs from bacteria, chloroviruses

(e.g., *Paramecium bursaria* chlorella virus 1, PBCV-1), and iridoviruses, as well as an EN from the *Tlr8* Maverick/Polinton element from *Tetrahymena thermophila*. Using CLANS (Frickey and Lupas 2004), we performed an updated clustering analysis with all available PLE ENs (supplementary fig. S10, Supplementary Material online). *Neptune*, *Nematis*, and *Penelope/Poseidon* ENs formed distinct although strongly connected clusters, with *Naiad* ENs essentially indistinguishable from *Neptune*. These results largely follow expectations from the shared presence of the CCHH motif and the presence/absence of the CX<sub>2-5</sub>CxxC Zn-finger linker (supplementary fig. S2, Supplementary Material online), and these domains are collectively representative of the canonical PLE EN described in NCBI (cd10442). *Hydra* ENs formed a distinct and well-resolved cluster that was nonetheless related to other PLE ENs, in line with the absence of the CCHH motif and the alternative configuration of the linker motif. Interestingly, the ENs sometimes associated with the giant *Terminons* (fig. 1), which contain RTs from the otherwise EN– *Athena* group (Arkhipova et al. 2017), were also recovered as a distinct cluster related to other PLE ENs. These ENs include both the CX<sub>2-5</sub>CxxC and CCHH motifs, suggesting shared ancestry with EN+ PLEs, although they also contain large unique insertions (supplementary fig. S2, Supplementary Material online). Finally, the *Chlamys* ENs were diffusely clustered between all other PLE ENs and several ENs from the HE\_Tlr8p\_PBC-V\_like group. The lack of strong clustering can likely be explained by the lack of both the CX<sub>2-5</sub>CxxC and CCHH motifs resulting in fewer conserved sites, and the possible link between *Chlamys* and HE\_Tlr8p\_PBC-V\_like ENs should be interpreted tentatively. Overall, the GIY–YIG ENs of all PLEs appear to be related, and in line with the results of Dunin-Horkawicz et al. (2006), PLE ENs are most similar to particular homing ENs from bacteria and viruses.

## Discussion

### A New Major PLE Clade with N-Terminal EN and Its Impact on Genome and Transposon Annotation

*Penelope*-like elements are arguably the most enigmatic type of retrotransposable elements inhabiting eukaryotic genomes. Due to their absence from the best-studied genomes such as mammals, birds, and angiosperms, and the complex tandem/inverted structures brought about by still undefined features of their peculiar transposition cycle, PLEs have largely been neglected and overlooked by most computational pipelines used in comparative genomics. Current approaches distinguish PLEs by the presence of a PLE-related RT, and classify them only to the “order” level as a clade of non-LTR elements (Bao et al. 2015) without subdivision into groups differing by domain architecture and phylogenetic placement, as is commonly done for non-LTR (LINE) and LTR retrotransposons (Storer et al. 2021). Here, we show that the degree of PLE structural and phylogenetic diversity matches that of non-LTR and LTR retrotransposons, emphasizing the need for updating current classification schemes and TE-processing computational pipelines.

Our data also underscore the need to adjust computational pipelines to incorporate searches for GIY–YIG EN either upstream or downstream from PLE RT, due to the high degree of polymorphisms (especially frameshifts) in the connector region, which complicates identification of full-length elements. This is especially relevant at a time when increasing numbers of invertebrate genomes are being sequenced, with *Naiad* elements often contributing tens of thousands of copies to metazoan genomic DNA. Underannotation of poorly recognizable TEs poses a serious problem to gene annotation. This is especially well-illustrated in host-associated and environmental metagenome analyses, where understudied eukaryotic TEs, including PLEs, become mis-assigned to bacterial genomes and are propagated in taxonomy-aware reference databases, jeopardizing future automated annotations (Arkhipova 2020).

Of special interest is the dominance of *Chlamys* PLEs in the plant kingdom, where their ancient nature is supported by their presence in the most basal members of the green lineage, by a high degree of divergence between *Chlamys* elements, and by distinctive features of the associated EN. In contrast to the documented case of horizontal transfer of a canonical C-terminal EN+ PLE into conifer genomes (Lin et al. 2016), their early branching position in the PLE phylogeny argues that they constitute ancestral genomic components of the green lineage, and does not support recent introduction. Nevertheless, their ongoing activity and diversification in *Chlamydomonas* indicates that *Chlamys* elements are actively participating in algal genome evolution.

### Common and Distinctive Features of *Naiad* and *Chlamys* Retrotransposition

Consistent association of all PLE RTs with a special type of endonuclease/nickase (GIY–YIG EN), which may have occurred several times in early eukaryotic evolution to form distinct lineages characterized by N- or C-terminal EN domains, underscores the importance of this EN for efficient intragenomic proliferation mediated by PLE RT, and emphasizes the need for further mechanistic investigations of the nontrivial PLE transposition cycle in representatives of each PLE lineage. It is very likely that the unique EN cleavage properties determine the formation of complex tandem/inverted pLTRs and the “tail” extension on either side of PLEs, not observed during TPRT of non-LTR elements, but seen in *Naiad/Chlamys*.

Further, PLEs are highly unusual among retroelements in their ability to retain introns after retrotransposition, sometimes even retrotransposing intron-containing host genes *in trans* (Arkhipova et al. 2003, 2013). Although most of the *Naiad/Chlamys* ORFs are not interrupted by introns, the functionally characterized active *Chlamys-3\_cRei* element shares an intron position within the 5′-UTR with the functionally studied *Penelope* from *D. virilis*, overlapping with the internal promoter (Schostak et al. 2008). This suggests that other PLEs may share this organization and harbor introns upstream of the ORF. The significance of intron retention is unknown, although it is likely a consequence of the unusual retrotransposition mechanism.

We failed to detect canonical HHR motifs in *Naiad* or *Chlamys* elements, except for two subclades of *Chlamys* elements from *C. braunii* and the spike mosses. These subclades are not closely related (fig. 5A), implying that HHRs may have been independently acquired or frequently lost from other *Chlamys*. Conversely, canonical HHRs are universally present in the *Hydra* clade and other PLEs. HHR function in other EN+ PLEs is still unclear, and while they have been hypothesized to help cleave the tandemly arranged long precursor RNAs (Cervera and de la Peña 2014), their absence from *Naiads* and most *Chlamys* elements obviously does not interfere with success in intragenomic proliferation, and may even facilitate it, if HHRs simply parasitize PLEs.

In many cases, it was not possible to discern target-site duplications in *Naiads* and *Chlamys* due to a strong insertion bias toward microsatellite repeats, with (CA)<sub>n</sub> and (C)<sub>n</sub> commonly observed in *Chlamys* and (TA)<sub>n</sub> in *Naiads*. The CX<sub>2-5</sub>CxC EN linker was hypothesized to mediate such bias in *Neptune* PLEs (Arkhipova 2006) and could do so in *Naiads*, but its absence from *Chlamys* suggests that the novel CxC domain may also play a role in targeting EN activity to specific DNA repeats. Also of interest are the EN– “PLE-X” families from two species of spike moss, which are the first known TEs to exhibit targeted insertion into the 28S ribosomal RNA gene in plants, as is observed in certain non-LTRs and DNA transposons of arthropods and other animals (Eickbush 2002; Penton and Crease 2004; Gladyshev and Arkhipova 2009).

Finally, it is unknown what role the IFD may play in *Naiads* and *Chlamys*. In TERTs, the IFD aids the stabilization of telomerase RNA (TER) and DNA during the extension of telomeric DNA (Jiang et al. 2018). The IFD domain in *Naiads* and *Chlamys* is shorter than that of TERTs, and its loss from a specific *Naiad* subclade demonstrates that it is not necessarily a functional requirement.

Establishment of an in vitro system to study PLE retrotransposition mechanisms would be the next important task required to achieve full understanding of PLE-specific TPRT features that distinguish them from LINEs, such as formation of complex tandem/inverted repeat structures and microsatellite insertion bias.

### *Naiad* Selenoproteins

The *Naiads* that encode selenoproteins are notable for two reasons. First, almost all described selenoproteins include a single Sec, whereas the *Naiads* contain either three or four. Bclaocos et al. (2019) performed analysis of selenoprotein P (SelP), one of the few selenoproteins including multiple Sec residues, finding that in bivalves SelP contains the most Sec residues of any metazoan group, and that spider SelP proteins contain a moderate number of Sec residues. Bivalves in particular are known for their high selenium content (Bryszewska and Måge 2015), and it may be that the *Naiads* represent cases of TEs adapting to their host cellular environments. However, even in bivalves selenoproteins are incredibly rare (e.g., the pacific oyster selenoproteome encompasses 32 genes; Bclaocos et al. 2019), suggesting a more specific role for the incorporation of Sec in these families. Sec residues are involved in numerous physiological processes and are

generally found at catalytic sites, where in many cases they have a catalytic advantage relative to cysteine (Labunskyy et al. 2014). All but one of the Sec residues in *Naiad* peptides correspond to highly conserved sites in the CCHH Zn-finger, CX<sub>2-5</sub>CxC Zn-finger and the DKA motif, and although the precise physiological role of these motifs in PLEs is unknown, it may be that the incorporation of Sec provides both a catalytic and evolutionary advantage.

Second, the *Naiad* families are the first described selenoprotein-encoding TEs. It is currently unclear whether these represent highly unusual cases, although the fact that they appear to have evolved independently in spiders and clams hints that other examples may be found in the future. This has potential implications for TE annotation in general, and selenoprotein-encoding TEs may have previously been overlooked in taxa such as bivalves because of apparent stop codons. Additionally, this result may provide insight into the evolution of new selenoproteins. The transition from encoding Cys to Sec is expected to be a complex evolutionary process, since a gene must acquire a SECIS element and near-simultaneously undergo a mutation from TGT/TGC (encoding Cys) to TGA (Castellano et al. 2004). The insertion of TEs carrying SECIS elements into the 3'-UTRs of genes could provide a pathway for SECIS acquisition, especially for TEs that undergo 5'-truncation and may insert with little additional sequence. It remains to be seen if the selenoprotein-encoding *Naiads*, or indeed any other TEs, have contributed to the evolution of new selenoproteins in their host genomes.

### Evolutionary Implications for PLE Origin and Diversification

As the branching order of major PLE clades diverging from TERTs is not exceptionally robust, it may be difficult to reconstitute evolutionary scenarios which were playing out during early eukaryogenesis. It is possible that an ancestral EN– PLE, similar to *Athena* or *Coprina* but lacking the extended N-terminus, was present at telomeres (Gladyshev and Arkhipova 2007) before undergoing either multiple domain fusions to give rise to TERTs, or fusions with GIY–YIG EN, either at the N- or the C-termini, to form the contemporary *Naiad/Chlamys*, *Neptune*, *Nematis*, and *Penelope/Poseidon* superfamilies capable of intrachromosomal proliferation.

There are several plausible evolutionary scenarios that could explain the observed EN and RT diversity, and ENs may have been acquired or exchanged several times by different PLE clades. It is possible that *Chlamys* elements acquired an EN without the CX<sub>2-5</sub>CxC and CCHH motifs from a homing EN from the HE\_Tlr8p\_PBC-V\_like family, and that the Zn-finger motifs were later gained by *Naiads*. ENs with both Zn-fingers could have been transferred from *Naiads* to the C-termini of EN– animal PLEs (once or multiple times), giving rise to other EN+ clades. This scenario would imply that the internal CCHH was then lost in *Hydra*, and the upstream linker domain was either reduced (*Nematis*), reduced and modified (*Hydra*), or lost (*Penelope/Poseidon*). Alternatively, an EN containing one or both Zn-fingers could have been independently acquired by C-terminal EN+ PLEs (again once or multiple times) and exchanged

with *Naiads* replacing the *Chlamys*-like EN (or gained independently by *Naiads* from a similar homing EN). This scenario would imply the existence of homing ENs with Zn-finger motifs, which have not been found, however, both the CX<sub>2-5</sub>CxxC and CCHH motifs are present in the stand-alone ENs occasionally associated with *Terminons*. EN acquisition, either at the N- or C-terminus, may have been facilitated if RT and EN were brought in proximity either on a carrier virus or on a chimeric circular replicon allowing permutation. Any combination of events in the above scenarios could of course explain the observed diversity. Notably, early metazoans such as cnidarians exhibit the highest PLE clade diversity, with *Poseidon*, *Naiad*, and *Hydra* present in *H. magnipapillata* and *Neptune*, *Naiad* and *Hydra* in *A. millepora*, implying that the appropriate conditions existed for either multiple exchanges or acquisitions of ENs. Finally, EN losses are not unusual, and EN– elements can emerge within EN+ clades, as in *Chlamys* PLE-X families in *Selaginella* and *Chara* (figs. 5 and 6) or the *Neptune*-like *MjPLE01* from the kuruma shrimp *Marsupenaeus japonicus* (Koyama et al. 2013), if they adopt alternative means of securing 3'-OH groups for TPRT.

Although the IFD domain may have been inherited by *Naiads/Chlamys* from a common ancestor with TERTs, IFD-like regions are also found sporadically in *Coprina* elements (e.g., in *Microbotryum violaceum*), arguing against its use as a synapomorphy. It is possible that the IFD has been lost multiple times in different PLE clades, as demonstrated by its loss in some *Naiads*, or gained independently in *Naiads/Chlamys* and certain *Coprina*. The presence/absence of canonical HHR motifs does not provide additional clues either: while found in only two basal *Chlamys* subclades and lacking in *Naiads* (perhaps even removing some constraints for *Naiad* amplification), they are present in all other PLEs, both EN+ and EN–. As with newly described retrozymes, they may exploit autonomous PLEs for their proliferation (Cervera and de la Peña 2020), or they could provide an unknown function.

Regardless of the exact sequence of events which led to PLE diversification in early eukaryotic evolution, it is now clear that the diversity of PLE structural organization, manifested in the existence of at least seven deep-branching clades (superfamilies) differing by domain architecture and found in genomes of protists, fungi, green and red algae, plants, and metazoans from nearly every major invertebrate and vertebrate phylum, can no longer be overlooked and should be reflected in modern genomic analysis tools. As more genomes from neglected and phylogenetically diverse lineages become available, it is likely that the diversity of PLEs will continue to expand, further supporting their increasingly important and unique position in TE biology and their contribution to shaping the amazing diversity of eukaryotic genomes.

## Materials and Methods

### Annotation and Curation of PLE Consensus Sequences

For general TE identification and annotation in metazoan genome assemblies (*D. stevensoni*, *P. laevis*), we used TEdenovo from the REPET package (Flutre et al. 2011) to

build de novo repeat libraries with default parameters. Although REPET-derived de novo TE consensus sequences are automatically classified under Wicker's scheme (Wicker et al. 2007), we additionally used RepeatMasker v4.1.0 (Smit et al. 2015) for TE classification, detection, and divergence plot building, using the initial TEdenovo repeat library. To specifically illustrate the composition on PLE families in the *P. laevis* genome, we used the corresponding consensus sequences of PLE families as a local library for divergence plot building.

Initial *Chlamys* consensus sequences from *C. reinhardtii* and its close relatives (*C. incerta*, *C. schloesseri*, and *Edaphochlamys debaryana*) were curated as part of a wider annotation of TEs in these species (Craig 2021; Craig et al. 2021). Inferred protein sequences from the initial metazoan and algal consensus models were then used as PSI-BLAST or TBLASTN (Camacho et al. 2009) queries to identify related *Naiad* and *Chlamys* PLEs in other species. PSI-BLAST was run using NCBI servers to identify putative PLE proteins that had been deposited in NCBI. TBLASTN was performed against all eukaryotic genome assemblies accessed at NCBI on 04/09/2020. Assemblies with multiple significant hits were selected for further curation, and where several related species had multiple hits the most contiguous assemblies were targeted. A Perl script was used to collect the nucleotide sequence of each TBLASTN hit from a given assembly, and the most abundant putative PLEs in each species were subjected to manual curation. This was performed by retrieving multiple copies by BlastN, extending the flanks of each copy and aligning the subsequent sequences with MAFFT v7.273 (Katoh and Standley 2013). The multiple sequence alignment of each family was then visualized and manually curated (removing poorly aligned copies, identifying 3' termini and pLTRs if present, etc.). Consensus sequences were produced for each family and protein sequences were inferred by identifying the longest ORF.

Copy number was estimated by performing BlastN against the assembly using the consensus sequence as a query, and counting the number of 3'-ends at least 80 nt in length to estimate actual insertion events, thus accounting for widespread 5'-truncation as well as for assembly fragmentation. NCBI BlastN optimized for highly similar sequences (megablast) was used with cutoff E-value 1e-5. WGS data sets were used for each species, with the best quality assembly used in case of multiple isolates. In most cases, the NCBI web interface was used to control for truncated and deleted copies and consensus quality via graphical summaries. If the maximum number of target sequences (5000) was exceeded, WGS data sets were created using blastn\_vdb from the SRA Toolkit and searched with BlastN 2.6.1+, or installed locally and searched with BlastN 2.10.1+. Note that the 3'-end-counting method gives a conservative but more precise estimate of insertion events, while counting all fragmented instances as copies could greatly inflate copy numbers: for example, 86,269 copies were reported in the highly fragmented *Mytilus galloprovincialis* assembly (Murgarella et al. 2016), as opposed to 3,782 3'-ends presented in figure 5.

Additional assessment of the correspondence between automated models and curated consensus sequences was

performed with RepeatModeler2, which performs automated classification using RepeatMasker. An automated model was deemed to be associated with a consensus sequence if it had >80% nucleotide sequence similarity over >80% of the length of the automated model.

Novel *Hydra* families were identified and curated using the same approach as above. Existing protein sequences from *H. magnipapillata* PLEs from Repbase (supplementary file S1, Supplementary Material online) were used as initial queries to search for related elements, alignments of which were then manually curated and used to produce consensus sequences.

### Motif Identification

SECS elements were identified in *Naiad* consensus sequences containing in-frame UGA codons using the SECSearch3 (Mariotti et al. 2013) online server (<http://gladyshevlab.org/SelenoproteinPredictionServer/>).

HHR motif searches were performed using secondary structure-based software RNAmotif (Macke et al. 2001). A general HHR descriptor (Cervera and de la Peña 2014) was used to detect HHR motifs in *Naiad*/*Chlamys* and *Hydra* elements. More relaxed descriptors were also employed as in Arkhipova et al. (2017) to accommodate different helices with longer loops and stem mispairing and more relaxed cores with mismatches, and with and without the presence of Helix III, however, it did not result in additional HHR motif detection. Although canonical HHRs are not represented in *Naiads*, some apparently possess the HHR catalytic core, but do not fit the current HHR stem/loop descriptors. To detect sequences with HHR core motifs, we used RNAmotif with single strings descriptors (seq="^cuganga" and seq="gaaa\$"), separated by a spacer (minlen = 7, maxlen = 17). Sequences with catalytic cores are denoted as HHR\_core in supplementary file S1, Supplementary Material online; additional eight cases with the first half of the catalytic core in the 3'-UTR and the second half found on either side at much longer distances (35–50 nt) are marked as HHR\_half\_core.

Matches to conserved protein domains were identified by searching the CDD (Marchler-Bauer et al. 2015) and PFAM (Mistry et al. 2021) databases. Some of the individual *Hydra* EN domains matched the cd10443 profile using CD-search at NCBI (Marchler-Bauer et al. 2015), whereas an HHpred search (Zimmermann et al. 2018) with the profile created from all *Hydra* EN sequences retrieved cd10446 as the top match.

### Functional Characterization of *Chlamys* Elements in *C. reinhardtii*

The divergence landscape (supplementary fig. S6, Supplementary Material online) and total abundance of *Chlamys* elements in *C. reinhardtii* were calculated using RepeatMasker v4.0.9 (Smit et al. 2015) and the highly contiguous assembly of strain CC-1690 (O'Donnell et al. 2020). The functional characterization represented in figure 7 was performed using the standard v5 reference genome. Iso-Seq (accession no. PRJNA670202) and H3K4me3 ChIP-seq (accession no. PRJNA681680) data were obtained from Gallaher et al. (2021). CCS (circular consensus sequence) Iso-Seq reads were mapped using minimap2 (-ax splice: hq -secondary no) (Li

2018). Within-species polymorphism was assessed by comparison to a de novo PacBio-based assembly of the divergent field isolate CC-2931, which exhibits ~3% genetic diversity relative to the standard reference strain (Craig et al. 2019). The sequencing and assembly of the CC-2931 assembly are described in supplementary file S3, Supplementary Material online. The CC-2931 assembly was mapped to the v5 reference using minimap2 (-ax asm10).

### RT Phylogeny and RT/EN Protein Clustering Analysis

Initial amino acid sequence alignments were performed with MUSCLE (Edgar 2004), with secondary structure assessed by inclusion of TERT PDB files (3kyl, 3du5) using PROMALS3D (Pei et al. 2008), and were manually adjusted to ensure the presence of each conserved RT motif at the proper position. Alignments in figure 3 were visualized in Jalview using the Clustal2 coloring scheme, and the sequence logos were created by AlignmentViewer (Waterhouse et al. 2009; Gabler et al. 2020). Phylogenetic analysis was performed with IQ-TREE v1.6.11 (Trifinopoulos et al. 2016), with the best-fit model chosen by ModelFinder according to Bayesian information criterion, and with 1,000 UFBoot replicates to evaluate branch support. The *Naiad* and *Chlamys* phylogram presented in figure 5 was based on an alignment of both EN and RT domains. Four *Naiads* and 16 *Chlamys* family proteins were excluded either due to high sequence similarity to other included proteins or to issues with determining the protein sequence (truncation, frameshifts, etc.). The *Hydra* phylogeny presented in supplementary figure S8, Supplementary Material online, was also based on alignment of both the RT and EN domains, using all 23 available proteins. The PLE and TERT phylogeny presented in figure 8 was based on the core RT and its N-terminal + thumb domains. A representative set of 27 *Naiad*, 24 *Chlamys*, and 13 *Hydra* proteins were selected for inclusion based on the subclade diversity revealed in figure 5 and supplementary figure S8, Supplementary Material online. All included/excluded families for each analysis are detailed in supplementary file S1, Supplementary Material online.

Protein clustering of the core RT and GIY–YIG EN domains was performed using CLANS (Frickey and Lupas 2004). The RT domain sequences used to produce figure 8 were used without any changes. For the EN analysis, GIY–YIG ENs from all superfamilies annotated at NCBI (cd00719) were combined with those from PLEs (canonical C-terminal EN+, N-terminal EN+, *Hydra*, and *Terminons*). All ENs were reduced to the core domain spanning from the "GIY" motif to the conserved N aa, unless a Zn-finger linker domain was present upstream of the "GIY," in which case this motif was also included (see supplementary fig. S2, Supplementary Material online). Several very distantly related EN superfamilies were excluded after a preliminary analysis. For both RT and EN domains, CLANS was run with a *P* value threshold of  $1 \times 10^{-8}$  until no further clustering changes were observed.

### Supplementary Material

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

## Acknowledgments

We thank Marcos de la Peña for advice on HHR motif searches. This work was supported by the U.S. National Institutes of Health (Grant No. R01GM111917) to I.R.A. R.J.C. was supported by the Biotechnology and Biological Sciences Research Council EASTBIO Doctoral Training Partnership and the project received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 694212).

## Author Contributions

All authors performed the research and analyzed the data, with R.J.C. focusing on algae/plants and I.A.Y., F.R., and I.R.A. on animals. R.J.C. and I.R.A. drafted the manuscript, with contribution to reviewing and editing from all authors.

## Data Availability

No new sequencing data were generated in this work; the employed data sets are listed throughout the text. *Naiad*, *Chlamys*, and *Hydra* consensus sequences are available as supplementary files S5–S7, [Supplementary Material](#) online, and the seed alignments were deposited in Dfam (Storer et al. 2021).

## References

- Arkipova IR. 2006. Distribution and phylogeny of *Penelope*-like elements in eukaryotes. *Syst Biol*. 55(6):875–885.
- Arkipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA*. 8:19.
- Arkipova IR. 2020. Metagenome proteins and database contamination. *mSphere* 5(6):e00854-20.
- Arkipova IR, Pyatkov KI, Meselson M, Evgen'ev MB. 2003. Retroelements containing introns in diverse invertebrate taxa. *Nat Genet*. 33(2):123–124.
- Arkipova 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(1):19.
- Arkipova IR, Yushenova IA, Rodriguez F. 2017. Giant reverse transcriptase-encoding transposable elements at telomeres. *Mol Biol Evol*. 34(9):2245–2257.
- Baclaococ J, Santesmasses D, Mariotti M, Biera K, Vetick MB, Lynch S, McAllen R, Mackrill JJ, Loughran G, Guigó R, et al. 2019. Processive recoding and metazoan evolution of selenoprotein P: up to 132 UGAs in molluscs. *J Mol Biol*. 431(22):4381–4407.
- Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 6:11.
- Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, Harney JW, Larsen PR. 1991. Recognition of UGA as a selenocysteine codon in Type I deiodinase requires sequences in the 3' untranslated region. *Nature* 353(6341):273–276.
- Böhne A, Zhou Q, Darras A, Schmidt C, Scharl M, Galiana-Arnoux D, Volff JN. 2012. Zisupton—a novel superfamily of DNA transposable elements recently active in fish. *Mol Biol Evol*. 29(2):631–645.
- Bryszewska MA, Måge A. 2015. Determination of selenium and its compounds in marine organisms. *J Trace Elem Med Biol*. 29:91–98.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Castellano S, Novoselov SV, Kryukov GV, Lescure A, Blanco E, Krol A, Gladyshev VN, Guigo R. 2004. Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep*. 5(1):71–77.
- Cervera A, de la Peña M. 2014. Eukaryotic *Penelope*-like retroelements encode hammerhead ribozyme motifs. *Mol Biol Evol*. 31(11):2941–2947.
- Cervera A, de la Peña M. 2020. Small circRNAs with self-cleaving ribozymes are highly expressed in diverse metazoan transcriptomes. *Nucleic Acids Res*. 48(9):5054–5064.
- Craig RJ. 2021. The evolutionary genomics of *Chlamydomonas*. Edinburgh: University of Edinburgh.
- Craig RJ, Bondel KB, Arakawa K, Nakada T, Ito T, Bell G, Colegrave N, Keightley PD, Ness RW. 2019. Patterns of population structure and complex haplotype sharing among field isolates of the green alga *Chlamydomonas reinhardtii*. *Mol Ecol*. 28(17):3977–3993.
- Craig RJ, Hasan AR, Ness RW, Keightley PD. 2021. Comparative genomics of *Chlamydomonas*. *Plant Cell* 33(4):1016–1041.
- Del Cortona A, Jackson CJ, Bucchini F, Van Bel M, D'hondt S, Škaloud P, Delwiche CF, Knoll AH, Raven JA, Verbruggen H, et al. 2020. Neoproterozoic origin and multiple transitions to macroscopic growth in green seaweeds. *Proc Natl Acad Sci USA*. 117(5):2551–2559.
- 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(5):494–506.
- Dunin-Horkawicz S, Feder M, Bujnicki JM. 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics* 7:98.
- Dupeyron M, Singh KS, Bass C, Hayward A. 2019. Evolution of Mutator transposable elements across eukaryotic diversity. *Mob DNA*. 10:12.
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
- Eickbush TH. 2002. R2 and related site-specific non-long terminal repeat retrotransposons. Washington (DC): ASM Press.
- Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175(2):477–485.
- Eickbush TH, Malik HS. 2002. Retrotransposon origin and evolution. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. *Mobile DNA II*. Washington (DC): ASM Press. p. 1111.
- Evgen'ev MB, Arkipova IR. 2005. *Penelope*-like elements – a new class of retroelements: distribution, function and possible evolutionary significance. *Cytogenet Genome Res*. 110(1–4):510–521.
- Fluttre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS One* 6(1):e16526.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 117(17):9451–9457.
- Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20(18):3702–3704.
- Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, Lupas AN, Alva V. 2020. Protein sequence analysis using the MPI bioinformatics toolkit. *Curr Protoc Bioinformatics*. 72(1):e108.
- Gallaher SD, Craig RJ, Ganesan I, Purvine SO, McCorkle S, Grimwood J, Strenkert D, Davidi L, Roth MS, Jeffers TL, et al. 2021. Widespread polycistronic gene expression in green algae. *Proc Natl Acad Sci U S A*. 118(7):e2017714118.
- Gladyshev EA, Arkipova IR. 2007. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A*. 104(22):9352–9357.
- Gladyshev EA, Arkipova IR. 2009. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* 448(2):145–150.
- Goodwin TJ, Poulter RT. 2002. A group of deuterostome Ty3/gypsy-like retrotransposons with Ty1/copia-like pol-domain orders. *Mol Genet Genomics*. 267(4):481–491.
- Jiang J, Wang Y, Sušac L, Chan H, Basu R, Zhou ZH, Feigon J. 2018. Structure of telomerase with telomeric DNA. *Cell* 173(5):1179–1190.e1113.
- Kapitonov VV, Jurka J. 2003. The esterase and PHD domains in CR1-like non-LTR retrotransposons. *Mol Biol Evol*. 20(1):38–46.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

- Kim KS, Kustu S, Inwood W. 2006. Natural history of transposition in the green alga *Chlamydomonas reinhardtii*: use of the AMT4 locus as an experimental system. *Genetics* 173(4):2005–2019.
- Kojima KK, Fujiwara H. 2005. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res.* 15(8):1106–1117.
- Koyama T, Kondo H, Aoki T, Hirono I. 2013. Identification of two Penelope-like elements with different structures and chromosome localization in kuruma shrimp genome. *Mar Biotechnol (NY)*. 15(1):115–123.
- Labunskyy VM, Hatfield DL, Gladyshev VN. 2014. Selenoproteins: molecular pathways and physiological roles. *Physiol Rev.* 94(3):739–777.
- Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47(W1):W256–W259.
- Letunic I, Khedkar S, Bork P. 2021. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49(D1):D458–D460.
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18):3094–3100.
- Lin X, Faridi N, Casola C. 2016. An ancient transkingdom horizontal transfer of Penelope-like retroelements from arthropods to conifers. *Genome Biol Evol.* 8(4):1252–1266.
- Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. 1997. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* 276(5312):561–567.
- Low SC, Berry MJ. 1996. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends Biochem Sci.* 21(6):203–208.
- Lue NF, Lin YC, Mian IS. 2003. A conserved telomerase motif within the catalytic domain of telomerase reverse transcriptase is specifically required for repeat addition processivity. *Mol Cell Biol.* 23(23):8440–8449.
- Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* 29(22):4724–4735.
- Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43(Database issue):D222–226.
- Mariotti M, Lobanov AV, Guigo R, Gladyshev VN. 2013. SECISearch3 and Sebastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res.* 41(15):e149.
- Mauer K, Hellmann SL, Groth M, Fröbuis AC, Zischler H, Hankeln T, Herlyn H. 2020. The genome, transcriptome, and proteome of the fish parasite *Pomphorhynchus laevis* (Acanthocephala). *PLoS One* 15(6):e0232973.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar Gustavo A, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49(D1):D412–D419.
- Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. 2016. A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PLoS One* 11(3):e0151561.
- Neupert J, Gallaher SD, Lu Y, Strenkert D, Segal N, Barahimipour R, Fitz-Gibbon ST, Schroda M, Merchant SS, Bock R. 2020. An epigenetic gene silencing pathway selectively acting on transgenic DNA in the green alga *Chlamydomonas*. *Nat Commun.* 11(1):6269.
- Ngan CY, Wong CH, Choi C, Yoshinaga Y, Louie K, Jia J, Chen C, Bowen B, Cheng H, Leonelli L, et al. 2015. Lineage-specific chromatin signatures reveal a regulator of lipid metabolism in microalgae. *Nat Plants.* 1:15107.
- Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D, et al. 2018. The *Chara* genome: secondary complexity and implications for plant terrestrialization. *Cell* 174(2):448–464.e424.
- Nowell RW, Wilson CG, Almeida P, Schiffer PH, Fontaneto D, Becks L, Rodriguez F, Arkhipova IR, Barraclough TG. 2021. Evolutionary dynamics of transposable elements in bdelloid rotifers. *Elife* 10:e63194.
- O'Donnell S, Chau F, Fischer G. 2020. Highly contiguous Nanopore genome assembly of *Chlamydomonas reinhardtii* CC-1690. *Microbiol Resour Announc.* 9(37):e00726–20.
- Pei J, Kim BH, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* 36(7):2295–2300.
- Penton EH, Crease TJ. 2004. Evolution of the transposable element Pokey in the ribosomal DNA of species in the subgenus *Daphnia* (Crustacea: Cladocera). *Mol Biol Evol.* 21(9):1727–1739.
- Pérez-Alegre M, Dubus A, Fernández E. 2005. REM1, a new type of long terminal repeat retrotransposon in *Chlamydomonas reinhardtii*. *Mol Cell Biol.* 25(23):10628–10638.
- Pyatkov KI, Arkhipova IR, Malkova NV, Finnegan DJ, Evgen'ev MB. 2004. Reverse transcriptase and endonuclease activities encoded by Penelope-like retroelements. *Proc Natl Acad Sci U S A.* 101(41):14719–14724.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol.* 29(1):24–26.
- Schön I, Rodriguez F, Dunn M, Martens K, Shribak M, Arkhipova IR. 2021. A survey of transposon landscapes in the putative ancient asexual ostracod *Darwinula stevensoni*. *Genes (Basel)* 12(3):401.
- Schostak N, Pyatkov K, Zelentsova E, Arkhipova I, Shagin D, Shagina I, Mudrik E, Blintsov A, Clark I, Finnegan DJ, et al. 2008. Molecular dissection of Penelope transposable element regulatory machinery. *Nucleic Acids Res.* 36(8):2522–2529.
- Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. Internet. Available from: <http://www.repeatmasker.org>.
- Stoddard BL. 2014. Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mob DNA.* 5(1):7.
- Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* 12(1):2.
- Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44(W1):W232–W235.
- Tujebajeva RM, Copeland PR, Xu XM, Carlson BA, Harney JW, Driscoll DM, Hatfield DL, Berry MJ. 2000. Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO Rep.* 1(2):158–163.
- Turanov AA, Lobanov AV, Hatfield DL, Gladyshev VN. 2013. UGA codon position-dependent incorporation of selenocysteine into mammalian selenoproteins. *Nucleic Acids Res.* 41(14):6952–6959.
- Van Roey P, Meehan L, Kowalski JC, Belfort M, Derbyshire V. 2002. Catalytic domain structure and hypothesis for function of GIY-YIG intron endonuclease I-TevI. *Nat Struct Biol.* 9(11):806–811.
- Vera M, Bello X, Alvarez-Dios JA, Pardo BC, Sanchez L, Carlsson J, Carlsson JE, Bartolome C, Maside X, Martinez P. 2015. Screening of repetitive motifs inside the genome of the flat oyster (*Ostrea edulis*): Transposable elements and short tandem repeats. *Mar Genomics.* 24(Pt 3):335–341.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.
- Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. *Annu Rev Genet.* 54:539–561.
- Wen W, Weiss SL, Sunde RA. 1998. UGA codon position affects the efficiency of selenocysteine incorporation into glutathione peroxidase-1. *J Biol Chem.* 273(43):28533–28541.
- 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(12):973–982.
- Xu Z, Xin T, Bartels D, Li Y, Gu W, Yao H, Liu S, Yu H, Pu X, Zhou J, et al. 2018. Genome analysis of the ancient tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Mol Plant.* 11(7):983–994.
- Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, Gabler F, Soding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol.* 430(15):2237–2243.