Discovery of numerous novel *Helitron*-like elements in eukaryote genomes using HELIANO

Zhen Li[1], Clément Gilbert[1], Haoran Peng[2,3], Nicolas Pollet[1*]

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France

[2]Crop Genome Dynamics Group, Agroscope, 1260 Nyon, Switzerland

[3]Present address: Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

*Corresponding author: nicolas.pollet@universite-paris-saclay.fr

Abstract

Helitron-like elements (HLEs) are widespread eukaryotic DNA transposons employing a rolling-circle transposition mechanism. Despite their prevalence in fungi, animals, and plant genomes, identifying *Helitrons* remains a formidable challenge. We introduce HELIANO, a software for annotating and classifying autonomous and non-autonomous HLE sequences from whole genomes. HELIANO overcomes several limitations of existing tools in speed and accuracy, demonstrated through benchmarking and its application to the complex genomes of frogs (*Xenopus tropicalis* and *Xenopus laevis*) and rice (*Oryza sativa*), where it uncovered numerous previously unidentified HLEs. In an extensive analysis of 404 eukaryote genomes, we found HLEs widely distributed across phyla, with exceptions in specific taxa. HELIANO's application led to the discovery of numerous new HLEs in land plants and identified 20 protein domains integrated within certain autonomous HLE families. A comprehensive phylogenetic analysis further classified HLEs into two primary clades, HLE1 and HLE2, and revealed nine subgroups, some of which are enriched within specific taxa. The future use of HELIANO promises to improve the global analysis of HLEs across genomes, significantly advancing our understanding of this fascinating transposon superfamily.

Introduction

Transposable elements (TEs) are ubiquitous selfish genetic elements characterized by their capacity to move and duplicate within genomes (1, 2). The nature and

abundance of TEs exhibit substantial variation across species (3, 4). This diversity of the TE landscape across genomes is associated with the absence of an ideal TE prediction tool, and manual curation will likely remain the best way to obtain the most accurate map of TEs in a given genome (5, 6). Yet, as complete genome sequences from all over the Tree of Life are produced, our global understanding of TEs diversity, evolution and impact on host genomes will continue to benefit strongly from improving automated algorithms that enable fast TE identification and classification (1, 4, 6, 7). Here, we present a new algorithm dedicated to fast and accurate *de novo* annotation of *Helitrons* while overcoming several limitations of existing tools (8–12).

Among eukaryote DNA transposons, *Helitrons* form a particular superfamily believed to use a rolling-circle transposition mechanism to spread in genomes (13). *Helitrons* have been reported in the genomes of numerous eukaryotic taxa, including fungi, animals, plants, and algae (8, 13–19). Among some well-studied species, they can contribute a considerable proportion of their genome sequence (18, 20). For example, *Helitrons* have been estimated to span around 6% of the little brown bat genome and about 4% of the silkworm genome (18, 20). Numerous reports revealed that *Helitrons* can capture genes and lead to horizontal transfer and genome shuffling, making them significant sources for genome dynamics and evolution (15, 21–23). While their evolutionary significance is undebated, *Helitrons* are still tricky to identify efficiently because they do not create target site duplication (TSD) upon transposition and lack classical structural features (1, 8).

All autonomous *Helitrons* encode Rep/Helicase (RepHel) transposase predicted to have both HUH endonuclease activity and 5’ to 3’ helicase activity (13, 19, 24). However, based on their terminal structure and coding potential differences, *Helitrons* were recently divided into two distinct groups: *Helitron*-like element 1 (HLE1) and *Helitron*-like element 2 (HLE2) (Figure 1) (14, 25–27). The HLE1 group corresponds to the canonical *Helitron* or *Helitron1*. In their terminal regions, HLE1 elements start at the 5’ end with the TC dinucleotide and terminate at the 3’ end with a short hairpin and a CTRR motif suffix (13, 19). In contrast, HLE2 elements have short terminal inverted repeats, making them structurally distinguishable from HLE1 (Figure 1A) (14, 17, 19, 26, 27). Elements from the two groups also vary in their insertion site preferences: the HLE1s usually insert between A and T nucleotides, while HLE2s generally insert between T and T nucleotides (14, 19, 27) (Figure 1A). Furthermore, HLE1s and HLE2s are phylogenetically distinguishable in their transposase sequence similarity (Figure

1B) (19, 24, 25). Previous studies indicated that the HLE2 group could be further classified into two variants: *Helitron2* and *Helentron* (14, 17, 19, 25–27) based on the presence or absence of the apurinic-apyrimidinic endonuclease (EN) domain (14, 17, 19, 26, 27). The *Helentron* variant was named to underscore the presence of the EN domain in its autonomous elements (14, 17, 19, 26, 27). However, whether *Helentron* and *Helitron2* are two distinct variants is still debated, as some studies indicated that *Helitron2* should include *Helentron* (14). Thomas and colleagues reported a proto-Helentron variant found only in the *Phytophthora* oomycete genomes (27). They proposed that this *proto-Helentron* variant might be an intermediate group between HLE1 and HLE2 because its terminal structure is more similar to the HLE1 group. At the same time, its transposase is phylogenetically closer to HLE2 (19, 27).

Besides the RepHel transposase domain, many additional gene sequences are recurrently found in HLEs (19). For example, the gene encoding a single-stranded DNA-binding protein homologous to the replication protein A (RPA) can be detected in HLE1s and HLE2s. Still, the Ovarian Tumor protein (OTU, homologous to predicted cysteine proteases) and apurinic-apyrimidinic endonuclease (EN) gene fragments could only be found in HLE2s (13, 17, 19, 27). These gene sequences tend to be fragmented and are not always detected in autonomous HLEs. This suggests they are not essential to HLE's transposition activity and might come from ancient gene capture events (19). Finally, autonomous HLEs often give birth to thousands of non-autonomous insertions, which share high similarity with their autonomous counterparts at both terminal regions (19) (Figure 1A).

Currently, tools for detecting HLEs are mainly structure-based, e.g., HelitronFinder, HelSearch, Helraizer, HelitronScanner, and EAHelitron (8–12). The primary strategy used in these tools is to search terminal signals of canonical *Helitrons* (HLE1): the TC signals for the left terminal region and the CTRR motif for the right terminal region (18). Because such terminal signals are widespread in genome sequences, these software tools suffer from a high rate of false positives, even if scores can help evaluate the prediction's quality, such as in HelitronScanner (8,14). Still, the terminal signals of HLE2 are pretty different from HLE1; therefore, these tools cannot detect HLE2s (19).

Our new software, HELIANO, a *Helitron*-like element annotator, was designed to comprehensively annotate all autonomous HLEs and their associated non-autonomous elements in a given genome. Unlike previously developed tools used for

HLE identification, HELIANO first relies on homology-based searches for detecting autonomous HLEs and then characterizes candidate element boundaries through a statistical approach, allowing the identification of significantly co-occurring left and right terminal signal pairs. We benchmarked HELIANO against the manually curated HLEs database of the *Fusarium oxysporum* genome, and we then used it to perform an in-depth prediction of HLE in three large genomes. Finally, we applied our new tool to scan the genomes of 404 eukaryotic species spanning the whole Tree of Life. We further annotated all predicted HLEs for their additional domains, built a new, largely extended phylogenetic tree of HLEs and proposed new perspectives on HLE classification. HELIANO is more accurate than previous HLE-annotation tools, is well-suited for large-scale, systematic analysis of HLEs in eukaryotes, and will thus be a valuable tool to further our understanding of HLE evolution and impact.

Material and Methods

Curation of *Helitron*-like elements from Repbase

Before detecting HLEs in genomes, we reasoned that having a global view of their structural features based on previously characterized elements would be helpful. We thus used four parameters to obtain a detailed description of the structural features of HLEs available in Repbase (28): the whole length of HLEs, the distance between Rep and Hel domains (d-RepHel), the distance between LTS and Rep domain (d-LTSRep), and the distance between Hel and RTS domain (d-RTSHel). We found that HLEs varied greatly in size from 53 nt to 39,893 nt, but about 75% of autonomous HLEs were shorter than 12,338 nt with an average length of 9,666 nt, while 75% of non-autonomous HLEs were shorter than 2,619 nt with an average length of 2,049 nt (Supplementary Figure S1A). The d-RepHel was shorter than 973 nt for about 75% of autonomous HLEs, with a maximum value of 2,439 nt (Supplementary Figure S1B). Moreover, we observed that the d-LTSRep value was shorter than 5,275 nt for 75% of autonomous HLEs with a maximum value of 27,501 nt, and the d-RTSHel value was shorter than 3,301 nt for 75% of autonomous HLEs with a maximum value of 16,811 nt (Supplementary Figure S1C, D). Based on the distribution of these HLE structural features, we set the default value of ‘dm’ in the HELIANO program as 2,500, which corresponds to the parameter d-RepHel; the default value of ‘w’ in the HELIANO program as 10,000, which is corresponding to parameters d-LTSRep and d-RTSHel.

HLE groups differ in their terminal structure and coding potential (19). On one hand, these differences could be used to classify different groups. On the other hand, various strategies are required to identify such a group. Although Repbase is a well-curated TE reference database where hundreds of autonomous HLEs have been collected, most of these collected HLEs have not been further classified (28). To improve the annotation of HLEs in Repbase, we initially collected Rep and helicase protein sequences from previous studies where HLE groups had been classified (21, 25, 29). Because these sequences represented a small subset of HLEs, we expanded this dataset by searching homologous sequences in Repbase (28) using NCBI blastp with default parameters (v2.13.0+) (30). Together with the query sequences from previous studies, we finally collected 239 helicase sequences: 167 for the HLE1 group, 72 for the HLE2 group, and 228 Rep endonuclease sequences: 155 for the HLE1 group, 73 for the HLE2 group. This expanded dataset represented a large diversity of species, including 13 fungi species, 20 land plants, 39 animals, two algae, and three Oomycota. Each homologous sequence found in Repbase was then classified into a specific group based on the highest blastp score obtained. The classification of collected HLE sequences was further checked and curated through a phylogenetic analysis. We computed multiple alignments using mafft (v7.475) with the parameter '--auto' (31), inferred phylogenetic tree using FastTree (v2.1.11 with default parameters) (32), and removed ambiguous leaves for which the phylogenetic position was inconsistent with the classification determined by blastp results. We provided the classification information in Supplementary Table S1. We observed in the phylogenetic trees that the transposase of HLE1s and HLE2s were distinctly separated, suggesting a reasonable classification (Figure 1B, Supplementary Figure S2-S3). Finally, we used this dataset to build the HMM model of HLE transposase used in HELIANO.

Training HMM models for Rep and Helicase domains

We computed multiple alignments for Rep and Helicase sequences for each HLE group using mafft with the '--auto' parameter. We then ran hmmbuild (v3.3) with the default parameter on each aligned file to obtain the four HMM models used in HELIANO (Supplementary material) (33).

HELIANO workflow

The HELIANO program follows a simple strategy: the first step is to search autonomous HLEs based on the transposase amino-acid sequence motifs, and the

second step is to identify their non-autonomous derivatives. We divided the pipeline into three main parts: transposase detection, LTS-RTS pair identification, and filtration (Figure 2). HELIANO relies on the prediction of ORFs in the genome sequence query and applies our pre-built HMM models to search for HLE joint Rep and Hel domains to find the transposase sequences (Figure 2A). HELIANO then scans the flanking region of transposases to identify significantly co-occurring LTS and RTS pairs (Figure 2B). Finally, HELIANO refines the candidates by checking the alignments of each subfamily's 50 nucleotides (nt) flanking sequence containing identical LTS and RTS pairs. (Figure 2C). As previous work suggested, we define families based on their RTS sequences and subfamilies based on their LTS sequences (19). The source code of HELIANO is available from Zenodo (https://zenodo.org/records/10625240) and GitHub (https://github.com/Zhenlisme/heliano/).

1) Search for potential HLE transposases in genomes

For a given genome, HELIANO uses the program getorf (EMBOSS:6.6.0.0) to predict open reading frames (ORF) with the parameter '-minsize 100' (34). Then HELIANO uses each trained HMM model to predict possible Rep endonuclease and helicase ORFs in the genome using hmmsearch (v3.3) with the parameter '--domtblout --noali -E 1e-3'. To avoid false positives, HELIANO filters out hits with 'c_Evalue' or 'i_Evalue' lower than 1e-5. The classification of every hit as HLE1 or HLE2 is further determined by selecting the group with the highest 'full-sequence score'. Because all HLE transposases contain both the Rep endonuclease domain and the Helicase domain, and because the Rep endonuclease domain is always upstream of the Helicase domain, we can deduce the genomic regions of HLE transposase. HELIANO uses bedtools window (v2.30.0) to find the genomic coordinates for the Rep-Helicase region (35). The potential HLE transposase is then classified based on the classification of the Rep and Helicase domains (Figure 2A).

2) Detection of LTS and RTS of HLEs

HELIANO then scans two windows at both ends of HLE transposases up to a given distance (default is 10,000 nt). The left terminal sequences (LTS) are searched in the left or upstream window, and the downstream right terminal sequences (RTS) are searched in the right or downstream extended window. For the HLE group, HELIANO applies the LCV model developed by HelitronScanner to detect the LTS, which starts with the dinucleotide TC (8). The RTS is expected to form a stem-loop structure

containing a 'CTRR' motif, and this structure is searched using rnabob (v2.2.1). For HLE2, HELIANO uses tirvish of the GenomeTools package (v1.6.1) (36) to detect TIRs whose left and right pairs will be taken as LTS and RTS, respectively. We set the size of TIRs to be longer than 11 nt and shorter than 18 nt according to previous studies (Figure 2B) (14, 27).

3) Identification of autonomous and non-autonomous HLE*s*

For all transposase regions, HELIANO collects sequences from all detected LTS and RTS sequences with an extension of 30 nt. These sequences are clustered to obtain unique sets of LTS and RTS using cd-hit (v4.8.1) (37). Next, HELIANO searches all homologous sequences using unique LTS and RTS sequences as queries against the genome using NCBI blastn (v 2.13.0+) (bitscore >= 32 by default). Finally, HELIANO retrieves all LTS-RTS pairs whose LTS should be upstream of its RTS using bedtools (v2.30.0) (35). By default, HELIANO searches LTS-RTS pairs whose LTS and RTS originate from the same transposase. To identify the best LTS-RTS pair, we test whether each pair's sets of LTS and RTS sequences colocalize in the whole genome, i.e., are located less than *dn* bp apart from each other. HELIANO takes advantage of the Fisher's exact test wrapped in the program bedtools to find such pairs (35). LTS and RTS sequences that significantly co-occur in a given genome would be taken as potential terminal sequences of HLEs, including autonomous and non-autonomous copies. We further classify these pairs into families based on their RTS sequences and subfamilies based on their LTS sequences. For example, two candidates could be classified into the same family if their RTS sequences share at least 90% identity. The candidates from the same family could be further classified into the same subfamily if their LTS sequences share at least 90% identity (Figure 2B).

4) Selection of the representative candidates from all possibilities

Inner LTS-RTS pairs existing within the intervals defined by the selected LTS-RTS pairs can also pass Fisher's exact test introduced in the last step. To examine such cases, for each subfamily, HELIANO samples up to 20 sequences, including 50 nt of flanking nucleotides and performs a multiple alignment using mafft. We reasoned that flanking nucleotides are conserved if they belong to the transposon, while flanking nucleotides of the real LTS-RTS are not expected to be conserved. HELIANO evaluates the average identity of aligned sequences using the R package seqinr (v4.2.30) (38). Ultimately, subfamilies with less than 70% identical flanking regions are

selected as representative HLE candidates, while the remaining constitute alternative candidates (Figure 2C).

5) Predict HLE candidates based on pre-identified LTS-RTS pairs

In some species whose autonomous HLEs do not exist, HELIANO supports the search for their HLE insertions based on pre-identified LTS-RTS pair sequences. These sequences will be added to downstream procedures (Figure 2).

Benchmarking

We needed a reliable database as 'ground truth' for benchmarking HELIANO and the other tools for HLE annotation. We identified the study of Chellapan and collaborators as suitable for this benchmarking because they manually curated HLEs in ten *F. oxysporum* genomes (14). As a result, they characterized five families of the HLE2 group and 26 consensus sequences that can be found in Repbase. We selected the genome of the Fo5176 strain (GCA_030345115.1) for benchmarking because it represents the most contiguous *F. oxysporum* genome with 4.5 Mbp for N50, 7 Mbp for L50, and 70.1 Mbp for genome size. To ensure that all complete insertions were fully recovered, we collected all LTSs and RTSs of HLE2 described in that study: 25 unique LTSs and 24 unique RTSs (14). Next, we used blastn using an e-value cutoff of 1e-2 to find all their homologous sequences in the Fo5176 genome. Then, we recovered full insertions by pairing all LTSs and RTSs using the window function of bedtools. After manual curation, we finally recovered 253 full insertions (Supplementary Table S2), used as a 'ground truth' in the following benchmarking process. HELIANO was run with the parameter '-w 15000 -is2 0 -p 1e-5 -n 20'; HelitronScanner was run with the default parameters; EAHelitron was run with the parameter '-r 4 -p 20 -u 20000'; RepeatModeler2 (v2.0.5) was run with the default parameter (8, 9, 39). As RepeatModeler2 only outputs consensus sequences, we then recovered their corresponding full copy insertions (>= 80% length of consensus) with the blastn program. We executed each program using a computer operated under Ubuntu GNU/Linux 22.04 LTS system with 20 threads and reported the real time of execution.

We then designed four benchmarking matrices for each program to evaluate their performance, including precision, sensitivity, FDR, and F1 score, computed using standard formulae (40). True positive (TP) was defined as the number of predicted insertions with more than the cutoff overlap in length with real insertion. The remaining

predicted insertions were defined as false positives (FP). False negative (FN) was defined as the number of real insertions with less than the cutoff overlap in length with any predicted insertions (Figure 3A). Eight cutoffs were further tested to calculate FP, FN, and TP: 65%, 70%, 75%, 80%, 85%, 90%, 95%, and 100%.

Prediction of HLEs from *F. oxysporum* 4287 strain genome

We downloaded the genome assembly of *F. oxysporum* 4287 (Fo4287) from NCBI (GCA_000149955.2). Its previously identified HLEs were recovered similarly to what we did for the Fo5176 strain. We then run HELIANO to search the HLE insertions in Fo4287 genomes by using the verified LTS-RTS pair sequences of Fo5176 genome with the following parameter ‘‘-w 15000 -is2 0 -p 1e-5 -n 20 --dis_denovo -ts PairFile’.

Comparison between HELIANO prediction and Repbase dataset for *X. tropicalis, X. laevis* and *O. sativa*

We downloaded the *Xenopus tropicalis* (GCF_000004195.4) and *Xenopus laevis* (GCF_017654675.1) genomes from NCBI, the *Oryza sativa* genome (version 7.0) from the RGAP website (http://rice.uga.edu/). For *X. tropicalis* and *X. laevis* genome, we ran HELIANO with the parameter ‘-s 30 -is1 0 -is2 0 -sim_tir 90 -n 20 -p 1e-5’. For *O. sativa* genome, we ran HELIANO with the parameter ‘-s 30 -is1 0 -is2 0 -n 20 -p 1e-5’. We manually examined each HLE subfamily's insertions by aligning the predicted insertion sequence with its genomic loci using *dialign2* (41). For *O. sativa* and *X. tropicalis*, we searched homologous sequences using the corresponding Repbase HLE consensus as queries against their genomes using NCBI blastn. We identified complete copies from the hits that shared at least 80% identity and 80% coverage to the query. These full-copy datasets were named Rbfull-XT for *X. tropicalis* and Rbfull-OS for *O. sativa* (Supplementary Table S3 and S4). We then used bedtools intersect to compare HELIANO and Rbfull insertions. An insertion was considered present in Rbfull and HELIANO if the Rbfull insertion was covered by at least 80% of its length.

HELIANO annotation on selected genomes

We established a selection of eukaryotic genomes as follows: 1) we downloaded the taxonomic information from 2,302 species whose genome assembly level reached the chromosomal level from the NCBI assembly database. 2) we then randomly selected the species by ensuring every order has at most two species, which will not represent the same family or genus. We removed the species *Trichomonas stableri* from the list because we could not find its genome assembly in NCBI. As a result, we

finally identified 404 well-assembled genomes (Supplementary Table S5). We then ran HELIANO with default parameters for each of these sampled genomes. Four thousand four hundred ninety-one bacterial genomes downloaded from NCBI were also tested as negative controls (Supplementary Table S6). We run HELIANO for the genome of *Phytophthora infestans* (GCA_026225685.1) with the parameter '-sim90 -p 1e-5 -is1 0 -is2 0'.

Detection of additional protein domains in HLEs from sampled genomes

For each detected autonomous HLE, we used the program getorf (EMBOSS:6.6.0.0) to predict their ORFs with the parameter '-minsize 100'. Then we used hmmsearch to identify additional domains in HLEs with the parameter '-E 1e-3' from Pfam downloaded from the InterPro website https://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.dat.gz (42). We reasoned that a domain captured by a family of HLE should be found in most copies belonging to the same HLE family. Moreover, we expected that random TE insertions could contribute to domains in HLEs, and we needed to exclude such cases from our analysis. To do so, the first filter we used was to remove domains found in less than five or 50% of HLE copies. Since many domains remained scattered along the whole length of HLEs, we empirically determined that filtering out the domains more than 4,000 bp away from the RepHel domain and removing the domains that occurred upstream and downstream of RepHel proved effective. Because of the lack of specific annotation for HLE transposases in Pfam, their Rep and Helicase domains have been annotated as different Pfam families. The Rep domain was annotated as families of Helitron_like_N (N-terminal of HLE transposase) and RepSA (replication initiator protein). The Helicase domain was annotated as families of PIF1/Pif1_dom_2B (Pif1-like Helicase). We thus ignored RepSA domains from the result and merged the name of Pif1_dom_2B with PIF1.

Construction of phylogenetic trees of HLEs from genomes of sampled species

For each species whose HLEs could be detected by HELIANO, we used the program getorf (EMBOSS:6.6.0.0) to predict ORFs of all its HLEs with the parameter '-minsize 400 -find 1'. Then, all Rep and Helicase domains were predicted via hmmsearch (v3.3) with the parameter '-E 1e-3' based on the same hmm model used in HELIANO. Predicted Rep and Helicase amino acid sequences of each HLE were

extracted and concatenated into single sequences (RepHel). For each species, we ran the program cd-hit with the parameter '-c 0.7' to get the representatives of RepHel sequences. An outgroup sequence was made by concatenating the geminivirus rep catalytic protein (NCBI accession number: WP_015060107.1) and helicase protein of *Myroides phaeus* (NCBI accession number: WP_090404604.1). All representatives of RepHel sequences and the outgroup sequence were aligned using mafft (v7.475). Finally, FastTree (v2.1.11) was applied to reconstruct a phylogenetic tree with default parameter (32).

Definition of autonomous and non-autonomous HLEs

This work defined HLE insertions as autonomous elements based on detecting the RepHel transposase domain. This means that HLE with a RepHel domain without identified terminal sequences were defined as autonomous, but they will be tagged as "orfonly" in the HELIANO output. Similarly, we defined non-autonomous HLEs as HLEs that do not contain a detectable RepHel domain but that contain HLE terminal sequences.

Results

HELIANO benchmarking and comparison with other tools

We used the published HLE2 dataset of *F. oxysporum* as a 'ground truth' for benchmarking (14). We selected it because it is the only accessible manually curated dataset for HLEs, as far as we know. We recovered 253 full HLE2 insertions, which were taken as genuine in the following benchmarking process (14). To estimate the performance of HELIANO, we calculated its precision, sensitivity, FDR, and F1 under different overlap cutoffs. We evaluated the performance of HelitronScanner, EAHelitron and RepeatModeler2 using the same method (Figure 3).

HELIANO had the highest sensitivity and F1 score among all test software (Figure 3B). HELIANO could detect 224 of the 253 genuine insertions (88.53%) with a coverage cutoff of 95%. EAHelitron found 40 insertions (15.81%) at the same cutoff level. Similarly, RepeatModeler2 identified 14 insertions (5.53%), and HelitronScanner only had eight (3.16%, Supplementary Table S2). The low sensitivity of EAHelitron and HelitronScanner on the *F. oxysporum* genome could be attributed to their design targeting the HLE1 group specifically. Overall, HELIANO had the lowest FDR and the highest precision. HELIANO predicted 68 complete insertions absent in the genuine insertion dataset, with an FDR of 23.29% at a 95% cutoff level (Supplementary Table

S2). We found that most of these predictions were characterized by clear terminal signals, indicating they might be new families of HLEs that have not been discovered yet (Supplementary Table S2). The EAHelitron software annotation had the highest FDR value, with 98.30% insertions absent in the genuine dataset. The software RepeatModeler2 had the second highest precision (69.70%), close to HELIANO at a 65% cutoff level. But when the cutoff was increased, RepeatModeler2 precision reduced while HELIANO kept a better precision level (76.71% ~ 78.77%).

We also asked if HELIANO can use only the pre-identified LTS-RTS pair sequences to predict the corresponding HLEs in another closely related genome. We selected the strain Fo4287 of *F. oxysporum* species as a test. The results obtained showed that among all 44 insertions recovered from a previously published study, HELIANO successfully recovered 27 of them (61.36%), which were further classified into two families (Supplementary Table S2) (14).

Regarding the execution time, RepeatModeler2 was the slowest software, with about two and a half hours, likely because it annotates all TEs. HelitronScanner took 35 minutes and EAHelitron 92 seconds. HELIANO ran the fastest with 70 seconds.

HELIANO uncovers overlooked HLEs in *Xenopus* frog genomes

As a first test case, we ran HELIANO on two frog genomes to further annotate their HLEs. The pipid frogs *X. tropicalis* and *X. laevis* are two important vertebrate model species with high-quality chromosomal scale genome assemblies in which TEs have been supposedly well annotated (43, 44). Moreover, these frog genome sequences are large and complex: 1.4 Gbp for *X. tropicalis* and 2.7 Gbp for the allotetraploid *X. laevis*, and their TE landscape is characterized by a majority of class II TE (43, 44). Only three non-autonomous *Helentron* consensus sequences have been reported for *X. tropicalis,* and none has been described for *X. laevis* in Repbase or previous studies (43). We annotated HLEs in these genomes using HELIANO in five minutes and 59 seconds for *X. tropicalis* and 16 minutes and 32 seconds for *X. laevis*.

Based on the 80-80 rule, we could map back the three Repbase non-autonomous HLE2 sequences in the *X. tropicalis* genome and identified 638 insertions (Rbfull-XT dataset, Supplementary Table S3). Using HELIANO, we annotated 82 HLE2 insertions and no HLE1 in the *X. tropicalis* genome. These *X. tropicalis* insertions included three autonomous and 79 non-autonomous HLE2s, further classified into three families based on the RTSs homology (Table 1, Supplementary Table S3). About 97% (72 out

of 74) of HELIANO-specific predictions belonged to the same family, HelenXT233. Only eight non-autonomous HLE insertions annotated by HELIANO were also in the Rbfull-XT dataset. These insertions were annotated as the HelenXT233 family in HELIANO prediction and the Helitron-N2_XT family in the Rbfull-XT dataset. However, we did not find the HELIANO-prediction counterparts for the other two Rbfull-XT families, Helitron-N1_XT (564 insertions) and Helitron-N1A_XT (59 insertions), which together made about 99% of the remaining Rbfull-XT-specific insertions. As expected, this difference stemmed from HELIANO's inability to detect families whose autonomous HLEs are absent from the genome.

Although there were no HLE sequences ever reported In the *X. laevis* genome (28, 43), HELIANO annotated 2,213 full insertions, including 15 autonomous and 2,198 non-autonomous insertions, which can be classified into three HLE2 and two HLE1 families (Table 1, Supplementary Table S7).

Table 1. HELIANO-predicted and Repbase full copies of HLEs in *Xenopus* genomes.

Species	Family	Subfamily	Auto	Nonauto	Variant	Source
X. tropicalis	HelenXT233	2	1	79	HLE2	HELIANO
X. tropicalis	HelenXT102	1	1	0	HLE2	HELIANO
X. tropicalis	HelenXT365	1	1	0	HLE2	HELIANO
X. tropicalis	Helitron-N1_XT	-	0	564	HLE2	Repbase
X. tropicalis	Helitron-N1A_XT	-	0	59	HLE2	Repbase
X. tropicalis	Helitron-N2_XT	-	0	15	HLE2	Repbase
X. laevis	HeliXL45	2	8	408	HLE1	HELIANO
X. laevis	HeliXL2	2	3	1238	HLE1	HELIANO
X. laevis	HelenXL108	1	1	545	HLE2	HELIANO
X. laevis	HelenXL83	1	1	0	HLE2	HELIANO
X. laevis	HelenXL460	1	2	7	HLE2	HELIANO

We manually examined each HLE subfamily's insertions by aligning the predicted insertion sequence with its genomic loci. We found that HELIANO correctly annotated their near full-length insertions of autonomous and non-autonomous HLE1s and HLE2s (Figure 4). Their boundaries could be confirmed by T-T insertion sites for HLE2 and A-T for HLE1 and by the precise alignment at both terminal regions and discordant alignments in flanking regions (Figure 4C, D, E). We could identify each family's terminal features, such as the TC motif in the LTS and stem-loop with CTRR suffix for HLE1s and the terminal inverted repeats and stem-loop structures in RTSs for HLE2s (Figure 4C, D, E). However, for some families, like HelenXT102 and HelenXT365, we failed to identify their insertion sites, and their boundaries were hard to find, which might represent degenerated HLE2 insertions. Moreover, we found that the terminal sequences of the autonomous insertion HelenXT233 were almost identical to those of Helitron-N2_XT, described in Repbase for decades, while its autonomous origins had never been discovered. This further evidenced the robustness of HELIANO for identifying autonomous HLEs and their non-autonomous derivatives in the large and complex *Xenopus* genomes (Figure 4C, D).

It is well known that HLE LTSs are more diverse than their RTSs (8, 19). We made similar observations for HLEs in *Xenopus* genomes. For example, the families HelenXT233, HeliXL45 and HeliXL2 could be further classified into subfamilies based on their LTSs homology (Table 1). The autonomous HelenXT233 is characterized by two LTSs, LTS1 and LTS2, forming a direct repeat (Figure 4A). However, LTS1 and LTS2 are not identical, each hallmark a different subfamily of non-autonomous HLE2. LTS1 characterizes HelenXT233-1, and HelenXT233-2 is characterized by LTS2 (Figure 4B, C, D).

HELIANO uncovers overlooked HLEs in *O. sativa*

As a second test case, we ran HELIANO on the *O. sativa* genome, one of the most important plant models (45, 46). HELIANO ran the task in 18 minutes and 26 seconds. There are 310 HLE1 consensus sequences collected from *O. sativa* in Repbase, including 22 autonomous and 288 non-autonomous entries. From these, we could map back only 14 autonomous HLEs consensus sequences and 236 non-autonomous ones in the genome sequence based on the 80-80 rule (2). These consensus sequences contributed to 25 autonomous and 2,088 non-autonomous complete insertions in the rice genome (RbFull-OS dataset, Supplementary Table S4). We ran

HELIANO on this rice genome and predicted 79 autonomous and 1,769 non-autonomous HLE1 insertions without evidence for HLE2 (Supplementary Table S8).

We then asked how many insertions annotated by HELIANO could also be found in the Rbfull-OS dataset. HELIANO annotated 21 autonomous insertions out of 25 (84%) in RbFull-OS. We examined the four autonomous insertions that differed and found that they were annotated in the HELIANO dataset but did not match over their entire length with their RbFull-OS equivalent, indicating that HELIANO predicted different LTS and RTS. We asked if the 58 HELIANO-unique autonomous insertions were new predictions or drawbacks of homology searching methods used to compile the RbFull-OS dataset. We ran a phylogenetic analysis to compare these 58 HELIANO-unique predictions and all 14 RbFull-OS autonomous HLE1s. The result showed that while all RbFull-OS autonomous HLE1s had identical counterparts in HELIANO prediction, the converse was not true, i.e. HELIANO unveiled new insertions (Supplementary Figure S4). We confirmed this finding by clustering all these HLE insertion sequences at the 90% identity threshold and observing that 44 clusters were HELIANO-specific (Supplementary Table S9). For example, HELIANO successfully annotated the total insertions of family HeliOs772 absent in RbFull-OS (Supplementary Figure S4, Supplementary Table S8). Its actual insertions were confirmed by their canonical HLE1 structures (started with TC dinucleotide and stopped with stem-loop structure with CTRR suffix) and insertion sites between A and T (Figure 4F).

For non-autonomous HLEs, HELIANO and RbFull-OS results differed significantly since 1,796 insertions were RbFull-OS-specific and 1,484 HELIANO-specific. The HELIANO-specific insertions could be explained by the fact that RbFull-OS included only a few non-autonomous complete copies for each family. At the same time, by design, HELIANO recovers non-autonomous insertions corresponding to autonomous ones based on shared LTS and RTS. For example, HELIANO predicted four autonomous and 60 full non-autonomous insertions for the family HeliOs1603 that had only one autonomous (Helitron-9_OS) and six non-autonomous counterparts in RbFull-OS (Figure 4G, Supplementary Figure S4). About 75% of HELIANO-predicted HeliOs1603 insertions were shorter than 7 kb, indicating that they were not likely to be false positives. The RbFull-OS-specific insertions could be attributed to the absence of their autonomous counterparts in the genome because HELIANO, by design, can not detect non-autonomous HLEs whose autonomous counterparts are missing.

We concluded that HELIANO could be especially useful for classifying and identifying HLEs in complex and HLE-rich plant genomes such as rice.

Reliability of HELIANO annotations

In the complex genomes of *O. sativa* and the two *Xenopus* frogs, most predictions of HELIANO were non-autonomous HLEs. Since these non-autonomous elements are devoid of RepHel domain, we considered the possibility that they may correspond to spurious predictions. We thus further checked the reliability of non-autonomous HLEs by assessing their copy number and their overlap with other repetitive sequences.

We first reasoned that the repetitive occurrence of predictions should be a good indicator of the reliability of non-autonomous HLEs. We counted the number of insertions (i.e. the copy number) for annotated HLEs by clustering sequences using a cutoff of 80% identity and 80% coverage and counting the number of sequences in each cluster. We found the following proportions of repetitive (more than one sequence) HELIANO predictions: 68.29% in the *X. tropicalis* genome, 90.33% in *X. laevis*, and 56.66% in *O. sativa* (Figure 5A). In addition, we checked whether the HELIANO predictions overlapped with simple sequence repeats and found only one minisatellite overlapping an *X. laevis* HLE sequence. Thus, these results indicated that most HELIANO predictions are *bona fide* repetitive sequence elements (Figure 5A).

We then evaluated the possible extent of HELIANO mis-annotations by quantifying the overlap between HELIANO predictions and other TE superfamilies. We first built a TE database comprised of all TE consensus sequences from Dfam (v3.8) and RepBase RepeatMasker Edition, which contain TEs from rice and *Xenopus* frogs annotated de novo in previous studies (28, 47). We then used RepeatMasker (v4.1.2) to reannotate HELIANO predictions. We classified HELIANO predictions as “OtherTE” if non-HLE TEs masked more than 60% of their length. Similarly, we classified HELIANO predictions as “HLE” if known HLE TEs masked more than 60% of their length. The HELIANO predictions that did not fit the two previous criteria were classified as “unannotated”. The “OtherTE” accounted for 2.44% of the HELIANO predictions for *X. tropicalis*, 0.68% for *X. laevis*, and 16.99% for *O. sativa* (Figure 5B). The “HLE” accounted for about 84.15% of the HELIANO predictions for *X. tropicalis*, 24.67% for *X. laevis*, and 60.61% for *O. sativa* (Figure 5B). A majority (74.65%) of predictions in *X. laevis* were not annotated by any known TEs, and they all belong to the family HeliXL45 and HeliXL2, indicating that they likely represent so far

undescribed HLEs. Thus, a vast majority of HELIANO predictions indeed uncover HLEs in genomic sequences.

We further asked the possible reasons for some predictions labeled as “OtherTE”. We took the *O. sativa* HLE predictions as an example whose “OtherTE” proportion is the highest. We found that 69.75% of the “OtherTE”predictions are repetitive and can be further clustered in two groups based on their sequence homology (Figure 5B). Sequences from group “a” had an average length of 181 nt and contributed to 192 HLE predictions. Sequences from group “b” had an average size of 559 nt and contributed to 10 predictions. However, we found the two groups could be annotated (more than 80% identity and 80% coverage) by both known HLEs and non-HLEs from the TE database: Helitron-N118_OS and Mariner-N17 for group a, Helitron-84_OS and TNR11 for group b. Thus, to the best of our knowledge, the annotation of these sequences remains unclear. In the remaining single-copy HLEs of the “OtherTE” label, we found that their sequence length is significantly longer than all other predictions (Wilcoxon rank sum test, p = 5.37e-38), and a majority of them (~ 78.95%) are longer than 6000 bp. Therefore, these sequences that are rare and too large to be HLEs may correspond to nested TEs. Indeed, we found other predictions with the “OtherTE” label as nested TEs, e.g., the nested Gypsy-25E_OS in HELIANO prediction homologous to Helitron-N33B_OS, the nested CR1_1b_Xt in predictions homologous to Helitron-N2_XT, and the nested Harbinger-N12_XL in HeliXL2 (Supplementary Figure S5). Altogether, we concluded that predictions with “OtherTE” were mostly caused by ambiguous annotations in the TE database or non-HLE TEs insertions inside HLE locus (i.e. nested TEs) which greatly inflated the HLE size.

Finally, we evaluated different parameters to reduce the proportion of mis-annotation. We designed eight parameter groups that set the distance between LTS and RTS and the insertion preference sites (Supplementary Figure S6A). Using identical genomic sequences, we found that these parameters can be used to identify and reduce the “OtherTE” proportion significantly (Supplementary Figure S6B). For example, the “OtherTE”proportion was reduced to 1.65% for *O. sativa* by limiting the distance between LTS and RTS to 6000 bp and by setting HLE insertion preference sites as A-T for HLE1 and T-T for HLE2 (Supplementary Figure S6B).

HELIANO revealed a broad distribution of HLE1 and HLE2 sequences in the eukaryotic world

To further evaluate the applicability of HELIANO, we sampled 404 chromosome-level genome assemblies of eukaryotic species from the NCBI genome database. The tested genome size ranged from 7.30 Mbp to 40.05 Gbp, and their GC content varied from 16.59% to 78.37% (Supplementary Figure S7). The corresponding species list covered 27 phyla, 83 classes, and 281 orders, including fungi, animals, land plants, and algae (Supplementary Table S5). Thus, this dataset represents a wide range of genome complexity and species diversity. We included 4,491 bacterial genomes expected to lack HLEs and used them as a true negative dataset according to the current model on HLE evolution (Supplementary Table S6) (19, 25).

We did not detect any HLEs in the sampled bacterial species, while HLEs were widespread among eukaryote genomes (Figure 6A). In addition, we found that 139 species lack HLEs in their genomes. For example, HELIANO did not detect any HLE sequences among all 29 sampled bird genomes. Among 26 sampled mammalian genomes, we only found HLE presence in the bat genome, as previously reported (16) (Figure 6A, B). Among the 404 tested eukaryote genomes, we identified 265 cases (66%) containing at least one HLE, encompassing 22 phyla and 61 classes (Supplementary Table S10, Figure 6A). Previous studies suggested a much narrower distribution of HLE2s than HLE1s, with a seeming absence in land plant genomes (19). However, in our large-scale scan, we found that both variants were prevalent all over the eukaryotic world. HLE1s were detected in 179 genomes from 20 phyla and 53 classes, HLE2s in 173 genomes from 19 phyla and 44 classes, and the unclassified HLEs in 19 genomes from eight phyla and 14 classes (Figure 6A, Supplementary Table S10). Furthermore, we identified a significant number of HLE2s in six (13 species) out of the eight (74 species) sampled land plant classes (Figure 6). We present the insertion of the HelenSM92 family in the *Sphagnum magellanicum* genome as an example of HLE2 presence in land plants, where we observed clear HLE2 features: short terminal inverted repeats and the ‘TT’ and ‘TT’ insertion sites (Supplementary Figure S8A). We also analyzed the protein domains of this HelenSM92 family using the Conserved Domain Database (CDD) search tool (48). Besides the RepHel domains, we found that a GIY-YIG domain was captured and transposed in this HLE2 family (Supplementary Figure S8B, C, D).

We conclude that HLEs are widely distributed in eukaryote genomes and that the prevalence of HLE2 was underestimated in previous studies. Moreover, our results

showed that HELIANO is a robust tool for annotating HLEs for complex and large genomes of diverse compositions.

Capture of additional gene domains in HLEs

HLEs are well known for their ability to capture gene sequences (12, 19, 49, 50). Additional protein domains recurrently found in HLEs, such as the RPA, OTU and EN domains, were thought to originate from different ancient gene-capture events (19). To explore the potential gene-capturing events across HLEs, we annotated the protein domains for each detected autonomous HLE in each sampled species. We expected that all copies of the same HLE family would be characterized by a captured domain found at a conserved position if this domain was stably captured and transposed.

Besides the most recurrently detected HLE1 helicase-like domain at N-terminus (Helitron_like_N) and PIF1 domains that we considered as being part of the HLE transposase, we found 20 additional domains that were stably included in HLEs, including the three previously described domains, RPA, OTU, and EN (annotated as Exo_endo_phos in Pfam) (19) (Supplementary Table S11). Overall, most of these 20 protein domains are known to enable the binding or modification of DNA or proteins. For example, the three amino acid peptide repeats (STPRs) and the B3 DNA binding domain (B3) function as transcription factors (51, 52). RPA and Ssb-like_OB are involved in DNA replication by binding to single-strand DNA (53, 54). The 2OG-Fe(II) oxygenase (2OG-FeII_Oxy_2) is reported to function as a DNA repair enzyme that removes methyl adducts and some larger alkylation lesions from endocyclic positions on purine and pyrimidine bases (55). The domain OTU, F-box associated domain (FBA_3), Ubiquitin carboxyl-terminal hydrolase (UCH), and C-terminal Ulp1 protease (Peptidase_C48) are involved in the regulation of protein degradation (56–60).

We further checked the position of these domains within HLE sequences. Globally, we found that all domains were enriched in certain regions, either downstream or upstream of RepHel, indicating their conserved position within HLEs (Figure 7). Some domains are likely to share the same open reading frame (ORF) with RepHel transposase, e.g., the domain DUF6570, Helicases from the Herpes viruses (Herpes_Helicase), N-terminal of large tegument protein of herpesviruses (Herpes_teg_N), and UvrD-like helicase C-terminal domain (UvrD_c_2) and provide evidence for molecular evolution events on the HLE transposase (Supplementary Figure S9A, B, S10, S11). However, the other 17 domains are encoded in different

ORFs. Moreover, most domains were not shared between HLE2 and HLE1 groups. For example, the B3, protein phosphatase 2A regulatory B subunit (B56), FBA_3, Fn3-like domain from Purple Acid Phosphatase (fn3_PAP), Herpes_Helicase and STPRs were almost exclusively found in HLE1. In contrast, 2OG-FeII_Oxy_2, DUF3106, DUF6570, EN, Herpes_teg_N, hemopoietic IFN-inducible nuclear protein (HIN), heat shock protein Hsp20 family (HSP20), Ssb-like_OB, and UCH were almost exclusively found in HLE2 (Figure 7). The remaining domains, e.g., RPA and OTU, were found in HLE2 and HLE1 variants (Figure 7). Interestingly, we found that the distribution features of RPA varied between HLE2 and HLE1. The RPA domain was enriched downstream of RepHel in HLE1 but upstream in HLE2, suggesting that HLE1 and HLE2 might have independently captured the RPA gene (Figure 7). Previous studies did not detect the presence of OTU in HLE1 (19). However, we noticed this domain in both HLE1 and HLE2 upstream of RepHel (Figure 7; Supplementary Figure S9C, D, S12-S13). Further phylogenetic analysis showed that the OTU domain of HLE1 was distinct from the OTU domain of HLE2 (Supplementary Figure S14). We conclude that gene capture events occurred repeatedly in HLEs and provided diverse molecular functions to these TEs.

Phylogenetic distribution of HLEs in eukaryotic genomes

Previous studies suggested that HLEs could be classified into HLE1 and HLE2 groups based on the difference in their coding potential and structural features (19, 25). HLE2 could be further classified into two different variants, *Helentron* and *Helitron2* (14, 19, 27). However, these analyses relied on a relatively small dataset (19, 25). Since we obtained a large number of HLEs from a wide diversity of genomes across the Tree of Life with HELIANO, we had the opportunity to study HLE diversity from a broader perspective. We reexamined this classification and further asked if we found additional variants of HLEs and what were their phylogenetic relationship. Our results showed that HELIANO accurately classified HLE1s and HLE2s following the phylogenetic classification. The accuracy of the HELIANO classification was estimated to be 99.17% (Figure 8A). Moreover, we discovered subgroups within the HLE1 and HLE2 clades. The HLE2 clade could be further classified into four subgroups (a, b, c, and d) and the HLE1 clade could be further classified into five subgroups (e, f, g, h, and i) (Figure 8A). In addition to the nine subgroups, we found a few HLEs in at least four additional clusters.

Known HLE1 sequences from Repbase were found across all five HLE1 subgroups. Most Repbase HLE sequences were found in subgroup b. Besides, Repbase HLEs were absent from subgroup d, indicating that HELIANO has identified a much broader diversity of HLEs than contained in the current Repbase collection.

In the HLE2 group, we found that the EN domain (shown as Exo_endo_phos in Figure 8B) characteristic of the previously identified variant *Helentron* was almost exclusively enriched in subgroup b (Figure 8A, B). Regarding the previously identified variant *Helitron2* sequences (14, 26), we found them within HLE2 subgroups a, b, and c, indicating that this *Helitron2* variant does not correspond to a monophyletic subgroup of HLE2 (Supplementary Figure S15, Figure 8A). The subgroups c and d together comprised a unique HLE2 clade consisting of a massive diversity of HLE2s, which mostly came from two species with giant genomes: the lungfish *Neoceratodus forsteri* (34.6 Gbp) and the newt *Ambystoma mexicanum* (28.2 Gbp).

Across all HLEs, we found that about five HLE subgroups (c, d, g, h, and i) are specific to their host types. For example, more than 90% of HLE1s in subgroups g, h, and i are hosted in diverse land plant species (Supplementary Figure S16). Conversely, we also observed a great variety of host types in some subgroups, highlighting the invasive nature of some HLEs. For example, at least four host types (Fishes, Mollusca, Cnidaria and land plants) could be found in subgroup b and five host types in subgroup f (Fishes, Arthropoda, Cnidaria and Amphibians and Mammals).

We then asked if any additional domains within HLEs could be used as signals to classify them. We selected the top ten most frequently detected domains to analyze their distribution across the HLE phylogenetic tree. The RepHel domains represented by Helitron_like_N and PIF1 were included as positive controls (Figure 8 B). Within our expectations, RepHel were prevalent in all HLE clades. Globally, we found the domain DUF6570 specifically in the HLE2 clade, suggesting it could be used as a marker to distinguish HLE1s from HLE2s. Moreover, many other domains were found to be limited within specific subgroups. For example, the UCH domain was enriched in subgroups c and d in the HLE2 clade, supporting their common origins (Figure 8B). The Exo_endo_phos and Herpes_teg_N domains were enriched in subgroup b. The RPA domain in HLE2 was limited to subgroup a. All these examples suggested that gene domains in HLEs could be potential signals to understand HLE evolution.

Revisiting proto-Helentron

Previous studies reported the variant proto-Helentron as an intermediate of HLE1 and HLE2 groups because of its LTS/RTS similarity to HLE1 (5' TT and 3' CTAG) and its coding potential similarity to HLE2 (19, 27). However, HLE1s usually have a 5' TC signal instead of a 5' TT signal. Besides, the proto-Helentron was only reported in *Phytophthora* genomes, and we did not detect any similar HLE sequences in the genomes of all 404 sampled species. We thus asked if HELIANO could detect the presence of proto-Helentron in *Phytophthora* genomes.

We annotated the *P. infestans* genome where the proto-Helentron was initially found. HELIANO predicted 1,428 full HLE insertions in this *Phytophthora* genome, including 613 autonomous and 815 non-autonomous elements. Among these predictions, 65 were annotated as HLE2s and 1,363 as HLE1s. We found one predicted HLE2 family named HelenPi572 almost identical to the initially published proto-Helentron (27). The HelenPi572 family contributed to three non-autonomous and 14 autonomous HLE2 insertions. The original proto-Helentron was ~ 3.6 kbp longer than autonomous HelenPi572 insertions (Supplementary Figure S17A). Further domain analysis showed that HelenPi572 additionally carried Toll-like receptors domain (TIR) and Su(var)3-9 and 'Enhancer of zeste' (SET) domain as described (Supplementary Figure S17B) (27). We then checked the structure information of the HelenPi572 family and observed short terminal inverted repeat sequences in most (16 out of 17) HelenPi572 insertions (Supplementary Figure S17A). The terminal inverted repeats were about 12 nt long with one mismatched base (Supplementary Figure S17A). Further multiple alignment analysis showed that the HelenPi572 family has a clear boundary at the 3' end (Supplementary Figure S17). However, the 5' boundary was about 3.6 kbp upstream of the predicted left boundary. In the 3.6 kbp extended region, we found a stem-loop structure at its 5' end (Supplementary Figure S17A). In conclusion, HELIANO detected the proto-Helentron insertions but missed the exact 5' terminal sequences, most likely because of its unusual structure and length.

We noticed that the structure of proto-Helentron resembled that of FoHeli3/4/5 (renamed here as HLE2-reverse) discovered in the *F. oxysporum* genome. This structure differs from the canonical HLE2 structure with short terminal inverted repeats at both ends (12~16 nt) and a stem-loop structure at the 3' end after the right repeat. The stem-loop structure of HLE2-reverse is similar to proto-Helentron and located at the 5' end upstream of the left terminal inverted repeats, the difference being the distance between the 5' stem-loop structure and the left terminal inverted repeat (14)

(Supplementary Figure S18). Further phylogenetic analysis for proto-Helentron and HLE2-reverse confirmed that they belong to the HLE2 group (Supplementary Figure S19). This is supported by the presence of the DUF6570 domain in their autonomous elements, which our result indicated as a marker for distinguishing the HLE1 and HLE2 groups. Still, proto-Helentron and HLE2-reverse are separated in the phylogenetic tree, suggesting they belong to different subgroups (Supplementary Figure S19).

Discussion

Accurate TE annotation from genomic sequences is essential to genome annotation, especially in large eukaryote genomes (6, 39). Moreover, the accelerated pace of complete genome sequencing requires scalable methods to perform comparative analysis and to shed light on TE biology and evolution. Among DNA TE, HLEs stand out as relatively large mobile elements, ~ 10 kbp, characterized by their ability to incorporate host gene DNA, but their annotation remains challenging (12, 22, 50).

Our HELIANO software provides a comprehensive solution to address HLE annotation in complete genomes, enabling large-scale comparative analysis. Due to the lack of species with a completely and perfectly annotated genome for HLEs, assessing the validity of HLEs annotation remains a complex task. In this study, we presented various analyses to support the relevance of HELIANO output. Using a manually curated set of HLE on the *Fusarium* genome, we show that HELIANO outperforms HelitronScanner, EAHelitron and RepeatModeler2. On this benchmark, our HELIANO software obtained the best precision, sensitivity, FDR and F1 metrics for all coverage values and was the fastest. While these results validated our algorithmic choices to develop HELIANO, they showed that full-length HLE annotation with precise boundaries at the base level is still challenging to obtain in some genomic loci.

In selected cases drawn from the analysis of two high-quality complex frog genomes, we showed the potential of HELIANO to predict HLE1s and HLE2s that had been undetected so far. We found a strikingly different landscape of HLEs with autonomous and non-autonomous insertions in the diploid *X. tropicalis* and tetraploid *X. laevis* genomes that diverged 45-50 MYA (43, 44). Like HLEs annotated using *Fusarium* genomes, we could reproduce HELIANO performance on detailed annotation at the nucleotide-level resolution of TE boundaries, even though this

depended on the genomic environment. To continue benchmarking, we targeted the *O. sativa* genome, a complex and HLE-rich plant genome (13). Based on existing annotations, we identified 25 autonomous and 2,088 non-autonomous full insertions of HLEs in the rice genome. We ran HELIANO and predicted 82 autonomous and 1,766 non-autonomous insertions. These annotation results gave the same picture of an HLE-rich genome dominated by non-autonomous transposons. HELIANO not only spotted all the autonomous HLE1s listed in Repbase but also uncovered many others. As expected, HELIANO predictions on non-autonomous HLEs were limited to families for which both autonomous and non-autonomous transposons were identified. Thus, HLE1 annotation on the rice genome is a target for further methodological improvements, especially to detect non-autonomous HLEs for which no cognate autonomous elements exist. We further evaluated HELIANO annotations by quantifying the copy number and the overlap with known TEs. While we observed that a vast majority of HELIANO predictions indeed uncover HLEs in genomic sequences, there were some misannotations due to ambiguous cases and nested TEs events. We defined a set of parameters that enable the investigation of possible misannotations so that users can optimize the level of annotations according to any given genomic sequence.

Using a large set of eukaryote genomes, we found that HELIANO could quickly produce a range of annotations, from a lack of full-length HLE to predictions of thousands of full-length and non-autonomous copies. We did not predict any HLE using HELIANO on 4,491 bacterial genomes, in accordance with the current model on the evolution of HLEs (25). Similarly, we did not detect HLEs in 139 of the 404 screened eukaryote genomes, a finding underscoring that false positives are not a central issue. We found that both HLE1s and HLE2s were much more widespread across eukaryotes than expected. Among 27 sampled phyla, we identified HLE1s in 20 phylum genomes and HLE2s in 19. Moreover, previous studies suggested that land plants lacked HLE2s (19). Yet, we verified HLE2's presence in many land plant genomes (Figure 6, Supplementary Figure S8, S16).

We also explored additional gene domains within the predicted HLEs. Besides the previously described domains (OTU, Exo_endo_phos, and RPA), we detected 17 more gene fragments incorporated into the HLE coding regions. These domains have various biochemical functions, such as transcription factors, protein degradation, etc. However, further work is required to investigate whether they are used in HLE

transposition or involved in the host's gene regulatory network. By checking the relative location of these domains to RepHel, we found that many domains had different distribution patterns between HLE1s and HLE2s, suggesting this information could be used as phylogenetic signals for their classification.

Further phylogenetic analysis of RepHel domains from all detected HLEs allowed us to re-examine the current classification of HLEs. Our results supported the existence of the two clades, HLE1 and HLE2 (Figure 8A). Besides, we identified four subgroups in HLE2s (subgroup a-d) and five in HLE1s (subgroup e-i). One previously described variant, *Helentron,* was found within subgroup b, suggesting that *Helentron* are only one of the four HLE2 subgroups. Furthermore, we found that many subgroups were dominated by certain host types (Figure 8A), suggesting a vertical inheritance of these HLEs as described previously (18, 19, 61, 62). Conversely, we observed a great diversity of host types in some subgroups, which could be partly explained by the ability of HLEs to undergo horizontal transfer (15, 63). Many domains were limited within specific subgroups, further supporting the classification of subgroups and suggesting their potential as phylogenetic markers. Some groups seemed devoid of HLEs, e.g., no HLE was detected among 29 sampled distinct birds. The mechanisms explaining this observation appear to be worth exploring in future research.

Furthermore, in the *A. mexicanum* giant genome (28.2 Gbp), we observed a remarkable divergence of HLE2s that formed the subgroup d, indicating the success of this subgroup in this species. However, in the larger giant lungfish *N. forsteri* genome (34.6 Gbp), we did not observe a comparable divergence of HLE2s. Future investigations on HLEs in giant genomes could be done to analyze how HLE evolved in these genomic landscapes.

In conclusion, this work provides a comprehensive and robust solution for improving HLE annotations in genomes. In particular, HELIANO's ability to generate a novel annotation on full-length HLEs from a large set of samples makes it a unique and valuable tool for the scientific community.

Data and Resource availability

Data and tools to prepare all supplementary data are available at https://zenodo.org/records/10625090. The source code of HELIANO is available from https://zenodo.org/records/10625240 and on GitHub (https://github.com/Zhenlisme/heliano/).

Authors contributions

Conceptualization: ZL CG HP NP; Data Curation: ZL NP; Funding Acquisition: ZL NP; Investigation: ZL; Methodology: ZL HP; Project Administration: CG HP NP; Resources: HP NP; Software: ZL; Supervision: CG NP; Validation: ZL CG HP NP; Writing – Original Draft: ZL; Writing – Review & Editing: ZL CG HP NP.

Funding

This work was supported by a China Scholarship Council – Université Paris Saclay PhD fellowship to ZL (202106760020).

References

1. Wells,J.N. and Feschotte,C. (2020) A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet*, **54**, 539–561.

2. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., 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.

3. Sotero-Caio,C.G., Platt,R.N.,II, Suh,A. and Ray,D.A. (2017) Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biology and Evolution*, **9**, 161–177.

4. Kojima,K.K. (2019) Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet. Syst.*, **94**, 233–252.

5. Platt,R.N.,II, Blanco-Berdugo,L. and Ray,D.A. (2016) Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biology and Evolution*, **8**, 403–410.

6. Goubert,C., Craig,R.J., Bilat,A.F., Peona,V., Vogan,A.A. and Protasio,A.V. (2022) A beginner's guide to manual curation of transposable elements. *Mobile DNA*, **13**, 7.

7. Makałowski,W., Gotea,V., Pande,A. and Makałowska,I. (2019) Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In Anisimova,M. (ed), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology. Springer, New York, NY, pp. 177–207.

8. Xiong,W., He,L., Lai,J., Dooner,H.K. and Du,C. (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences*, **111**, 10263–10268.

9. Hu,K., Xu,K., Wen,J., Yi,B., Shen,J., Ma,C., Fu,T., Ouyang,Y. and Tu,J. (2019) Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. *BMC Bioinformatics*, **20**, 354.

10. Yang,L. and Bennetzen,J.L. (2009) Structure-based discovery and description of plant and animal Helitrons. *Proceedings of the National Academy of Sciences*, **106**, 12832–12837.

11. Du,C., Caronna,J., He,L. and Dooner,H.K. (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics*, **9**, 51.

12. Barbaglia,A.M., Klusman,K.M., Higgins,J., Shaw,J.R., Hannah,L.C. and Lal,S.K. (2012) Gene Capture by Helitron Transposons Reshuffles the Transcriptome of Maize. *Genetics*, **190**, 965–975.

13. Kapitonov,V.V. and Jurka,J. (2001) Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, **98**, 8714–8719.

14. Chellapan,B.V., van Dam,P., Rep,M., Cornelissen,B.J.C. and Fokkens,L. (2016) Non-canonical Helitrons in Fusarium oxysporum. *Mobile DNA*, **7**, 27.

15. Han,G., Zhang,N., Xu,J., Jiang,H., Ji,C., Zhang,Z., Song,Q., Stanley,D., Fang,J. and Wang,J. (2019) Characterization of a novel Helitron family in insect genomes: insights into classification, evolution and horizontal transfer. *Mobile DNA*, **10**, 25.

16. Kosek,D., Grabundzija,I., Lei,H., Bilic,I., Wang,H., Jin,Y., Peaslee,G.F., Hickman,A.B. and Dyda,F. (2021) The large bat Helitron DNA transposase forms a compact monomeric assembly that buries and protects its covalently bound 5′-transposon end. *Molecular Cell*, **81**, 4271-4286.e4.

17. Poulter,R.T.M., Goodwin,T.J.D. and Butler,M.I. (2003) Vertebrate helentrons and other novel Helitrons. *Gene*, **313**, 201–212.

18. Thomas,J., Phillips,C.D., Baker,R.J. and Pritham,E.J. (2014) Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. *Genome Biology and Evolution*, **6**, 2595–2610.

19. Thomas,J. and Pritham,E.J. (2015) *Helitrons* , the Eukaryotic Rolling-circle Transposable Elements. *Microbiol Spectr*, **3**, 3.4.03.

20. Han,M.-J., Shen,Y.-H., Xu,M.-S., Liang,H.-Y., Zhang,H.-H. and Zhang,Z. (2013) Identification and Evolution of the Silkworm Helitrons and their Contribution to Transcripts. *DNA Res*, **20**, 471–484.

21. Grabundzija,I., Messing,S.A., Thomas,J., Cosby,R.L., Bilic,I., Miskey,C., Gogol-Döring,A., Kapitonov,V., Diem,T., Dalda,A., *et al.* (2016) A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nat Commun*, **7**, 10716.

22. Morgante,M., Brunner,S., Pea,G., Fengler,K., Zuccolo,A. and Rafalski,A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet*, **37**, 997–1002.

23. Heringer,P., Dias,G.B. and Kuhn,G.C.S. (2017) A Horizontally Transferred Autonomous Helitron Became a Full Polydnavirus Segment in Cotesia vestalis. *G3 (Bethesda)*, **7**, 3925–3935.

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

25. Heringer,P. and Kuhn,G.C.S. (2022) Pif1 Helicases and the Evidence for a Prokaryotic Origin of Helitrons. *Molecular Biology and Evolution*, **39**, msab334.

26. Bao,W. and Jurka,J. (2013) Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mobile DNA*, **4**, 12.

27. Thomas,J., Vadnagara,K. and Pritham,E.J. (2014) DINE-1, the highest copy number repeats in Drosophila melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). *Mobile DNA*, **5**, 18.

28. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

29. Heringer,P. and Kuhn,G.C.S. (2018) Exploring the Remote Ties between Helitron Transposases and Other Rolling-Circle Replication Proteins. *International Journal of Molecular Sciences*, **19**, 3079.

30. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

31. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, **30**, 3059–3066.

32. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix | Molecular Biology and Evolution | Oxford Academic.

33. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLOS Computational Biology*, **7**, e1002195.

34. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

35. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

36. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations.

37. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

38. Charif,D. and Lobry,J.R. (2007) SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In Bastolla,U., Porto,M., Roman,H.E., Vendruscolo,M. (eds), *Structural Approaches to Sequence Evolution*, Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 207–232.

39. Flynn,J.M., Hubley,R., Goubert,C., Rosen,J., Clark,A.G., Feschotte,C. and Smit,A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, **117**, 9451–9457.

40. Hoen,D.R., Hickey,G., Bourque,G., Casacuberta,J., Cordaux,R., Feschotte,C., Fiston-Lavier,A.-S., Hua-Van,A., Hubley,R., Kapusta,A., *et al.* (2015) A call for benchmarking transposable element annotation methods. *Mobile DNA*, **6**, 13.

41. Al Ait,L., Yamak,Z. and Morgenstern,B. (2013) DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Research*, **41**, W3–W7.

42. Paysan-Lafosse,T., Blum,M., Chuguransky,S., Grego,T., Pinto,B.L., Salazar,G.A., Bileschi,M.L., Bork,P., Bridge,A., Colwell,L., *et al.* (2023) InterPro in 2022. *Nucleic Acids Research*, **51**, D418–D427.

43. Session,A.M., Uno,Y., Kwon,T., Chapman,J.A., Toyoda,A., Takahashi,S., Fukui,A., Hikosaka,A., Suzuki,A., Kondo,M., *et al.* (2016) Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, **538**, 336–343.

44. Hellsten,U., Harland,R.M., Gilchrist,M.J., Hendrix,D., Jurka,J., Kapitonov,V., Ovcharenko,I., Putnam,N.H., Shu,S., Taher,L., *et al.* (2010) The Genome of the Western Clawed Frog Xenopus tropicalis. *Science*, **328**, 633–636.

45. Song,S., Tian,D., Zhang,Z., Hu,S. and Yu,J. (2018) Rice Genomics: over the Past Two Decades and into the Future. *Genomics Proteomics Bioinformatics*, **16**, 397–404.

46. Yu,J., Hu,S., Wang,J., Wong,G.K.-S., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X., *et al.* (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.

47. Hubley,R., Finn,R.D., Clements,J., Eddy,S.R., Jones,T.A., Bao,W., Smit,A.F.A. and Wheeler,T.J. (2016) The Dfam database of repetitive DNA families. *Nucleic Acids Res*, **44**, D81–D89.

48. Wang,J., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R., Gwadz,M., Lu,S., Marchler,G.H., Song,J.S., Thanki,N., Yamashita,R.A., *et al.* (2023) The conserved domain database in 2023. *Nucleic Acids Res*, **51**, D384–D388.

49. Li,Y. and Dooner,H.K. (2012) Helitron Proliferation and Gene-Fragment Capture. In Grandbastien,M.-A., Casacuberta,J.M. (eds), *Plant Transposable Elements: Impact on Genome Structure and Function*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 193–217.

50. Garrigues,J.M., Tsu,B.V., Daugherty,M.D. and Pasquinelli,A.E. (2019) Diversification of the Caenorhabditis heat shock response by Helitron transposable elements. *eLife*, **8**, e51139.

51. Ulmasov,T., Hagen,G. and Guilfoyle,T.J. (1997) ARF1, a Transcription Factor That Binds to Auxin Response Elements. *Science*, **276**, 1865–1868.

52. Yu,L.-Y., Cheng,W., Zhou,K., Li,W.-F., Yu,H.-M., Gao,X., Shen,X., Wu,Q., Chen,Y. and Zhou,C.-Z. (2016) Structures of an all-α protein running along the DNA major groove. *Nucleic Acids Res*, **44**, 3936–3945.

53. Ren,W., Chen,H., Sun,Q., Tang,X., Lim,S.C., Huang,J. and Song,H. (2014) Structural basis of SOSS1 complex assembly and recognition of ssDNA. *Cell Rep*, **6**, 982–991.

54. Bochkarev,A., Pfuetzner,R.A., Edwards,A.M. and Frappier,L. (1997) Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA. *Nature*, **385**, 176–181.

55. Clifton,I.J., McDonough,M.A., Ehrismann,D., Kershaw,N.J., Granatino,N. and Schofield,C.J. (2006) Structural studies on 2-oxoglutarate oxygenases and related double-stranded β-helix fold proteins. *Journal of Inorganic Biochemistry*, **100**, 644–669.

56. Mossessova,E. and Lima,C.D. (2000) Ulp1-SUMO Crystal Structure and Genetic Analysis Reveal Conserved Interactions and a Regulatory Element Essential for Cell Growth in Yeast. *Molecular Cell*, **5**, 865–876.

57. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis | PNAS.

58. Mevissen,T.E.T., Hospenthal,M.K., Geurink,P.P., Elliott,P.R., Akutsu,M., Arnaudo,N., Ekkebus,R., Kulathu,Y., Wauer,T., El Oualid,F., *et al.* (2013) OTU Deubiquitinases Reveal Mechanisms of Linkage Specificity and Enable Ubiquitin Chain Restriction Analysis. *Cell*, **154**, 169–184.

59. Jentsch,S., Seufert,W. and Hauser,H.-P. (1991) Genetic analysis of the ubiquitin system. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, **1089**, 127–139.

60. Barrett,A.J. and Rawlings,N.D. (2001) Evolutionary Lines of Cysteine Peptidases. **382**, 727–734.

61. Yang,H.-P. and Barbash,D.A. (2008) Abundant and species-specific DINE-1 transposable elements in 12 Drosophila genomes. *Genome Biology*, **9**, R39.

62. Ellison,C.E. and Bachtrog,D. (2013) Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. *Science*, **342**, 846–850.

63. Thomas,J., Schaack,S. and Pritham,E.J. (2010) Pervasive Horizontal Transfer of Rolling-Circle Transposons among Animals. *Genome Biology and Evolution*, **2**, 656–664.

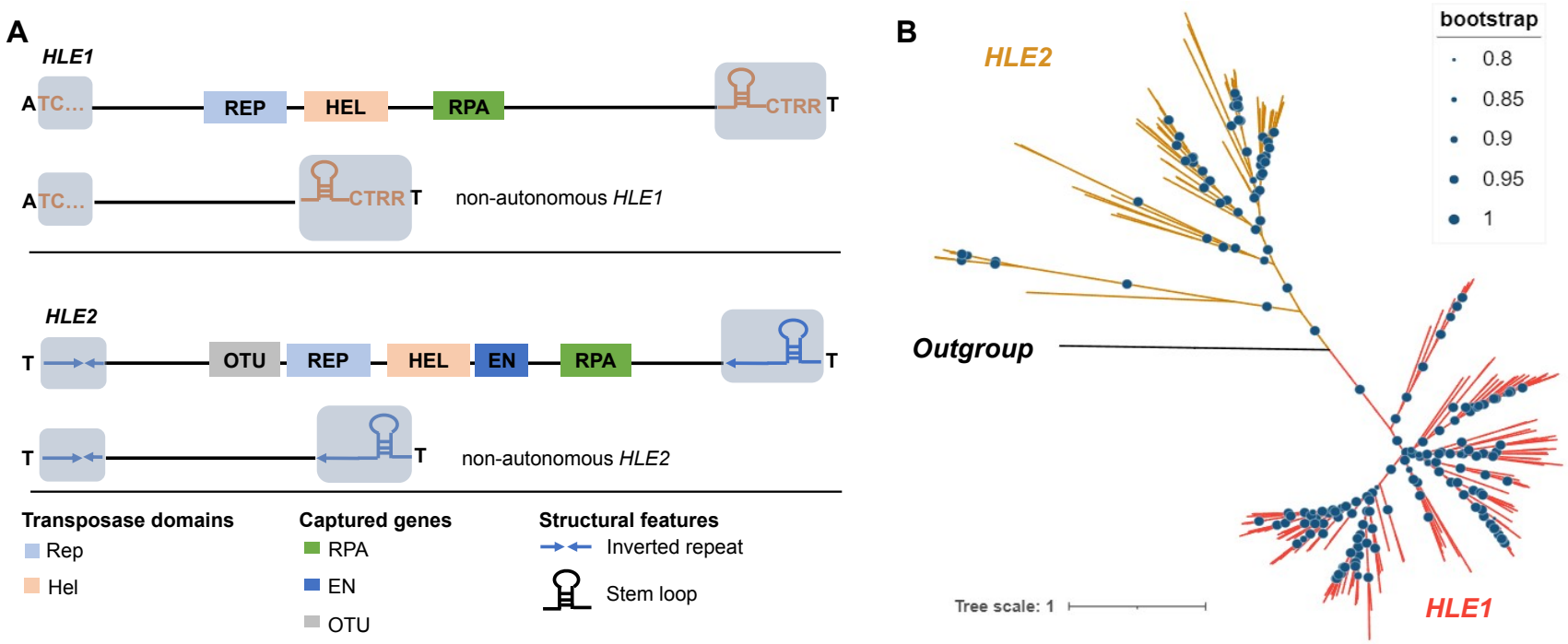

Figure 1. Introduction of *Helitron*-Like Elements (HLEs) groups. (A) Features of HLEs. Autonomous elements are pictured on top, with their non-autonomous derivatives just below. Non-autonomous HLEs share identical Left Terminal Sequences (LTS) and Right Terminal Sequences (RTS) with their autonomous counterparts. All autonomous HLEs encode the Rep (light blue) and Helicase (orange) domains. HLE1s might also carry the RPA domain (green), and HLE2s might have the EN domain (blue), RPA domain, and OTU domain (grey). HLE1s usually insert between A and T nucleotides, while HLE2s usually insert between T and T nucleotides. The scale of this scheme is relative. (B) Maximum likelihood estimation tree of HLE transposases from Repbase (LogLk = -153572.880). The clade highlighted in red corresponds to the HLE1 group, and the clade highlighted in orange corresponds to the HLE2 group (including *Helitron2* and *Helentron*). As an outgroup, we used a sequence made by concatenating geminivirus catalytic rep and helicase proteins of *Myroides phaeus*. Blue dots on the tree branches are bootstrap values greater than 0.8.

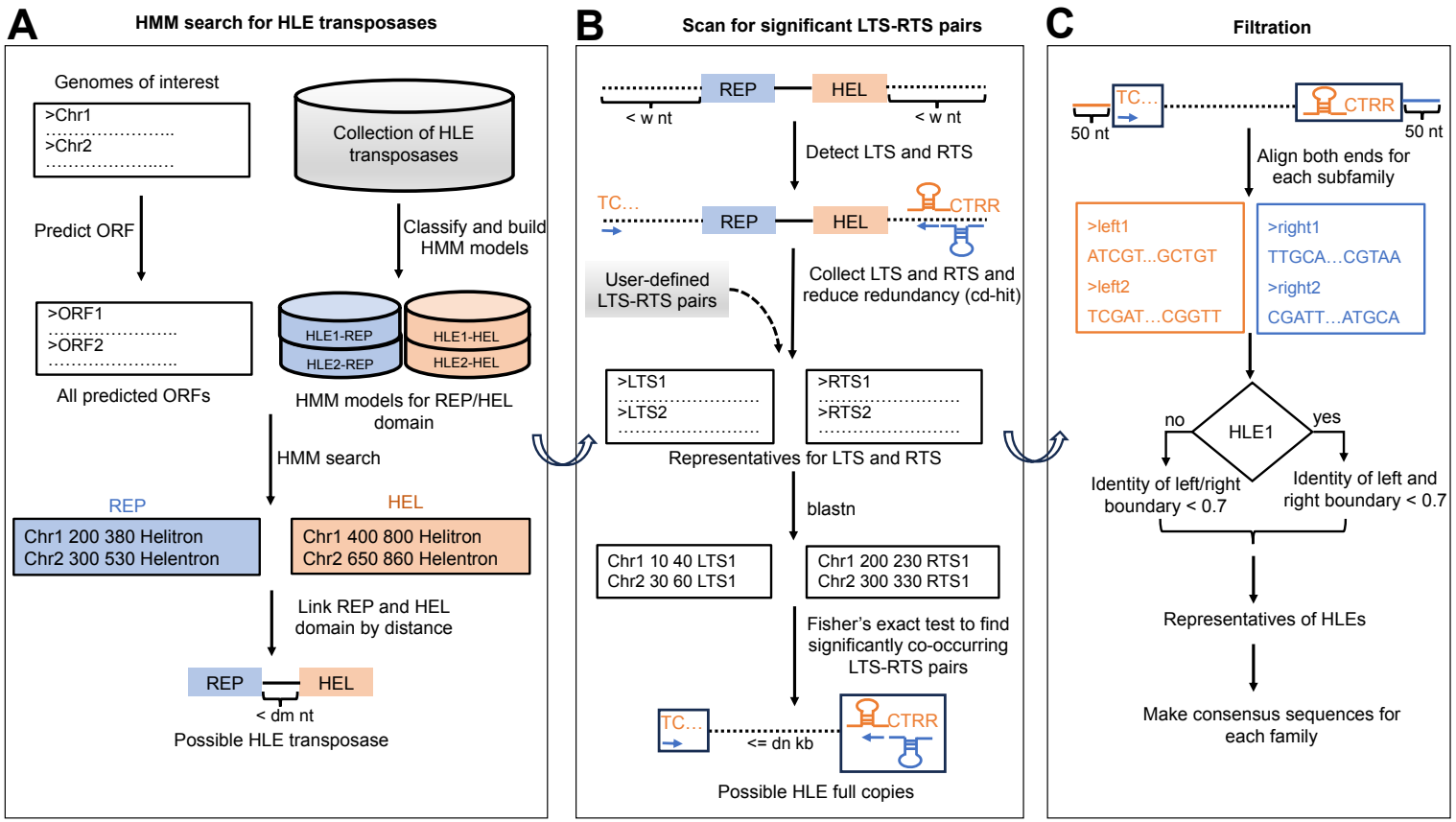

Figure 2. Workflow of HELIANO. (A) HMM searches for transposases of HLEs; dm denotes the distance between the Rep and Helicase domains. (B) Scan for significantly co-occurring LTS-RTS pairs; w indicates the length of the RepHel domain flanking sequences; dn denotes the distance between LTS and RTS. (C) Filtration to get representative insertions and make their consensus sequences.

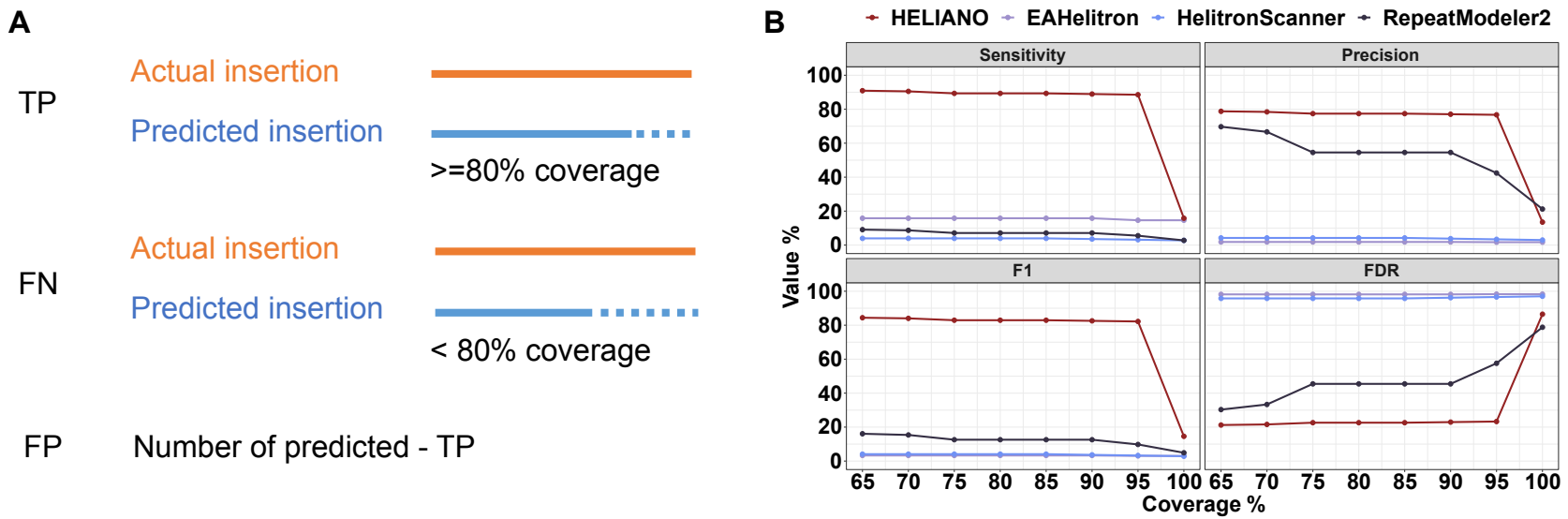

Figure 3. Benchmarking analysis of HELIANO. (A) Schematic representation of benchmarking metrics. TP: True positive; FN: False negative; FP: False positive. (B) Comparison of benchmarking metrics of all tested software. F1 is the score computed as the harmonic mean between sensitivity and recall. FDR: False Discovery Rate.

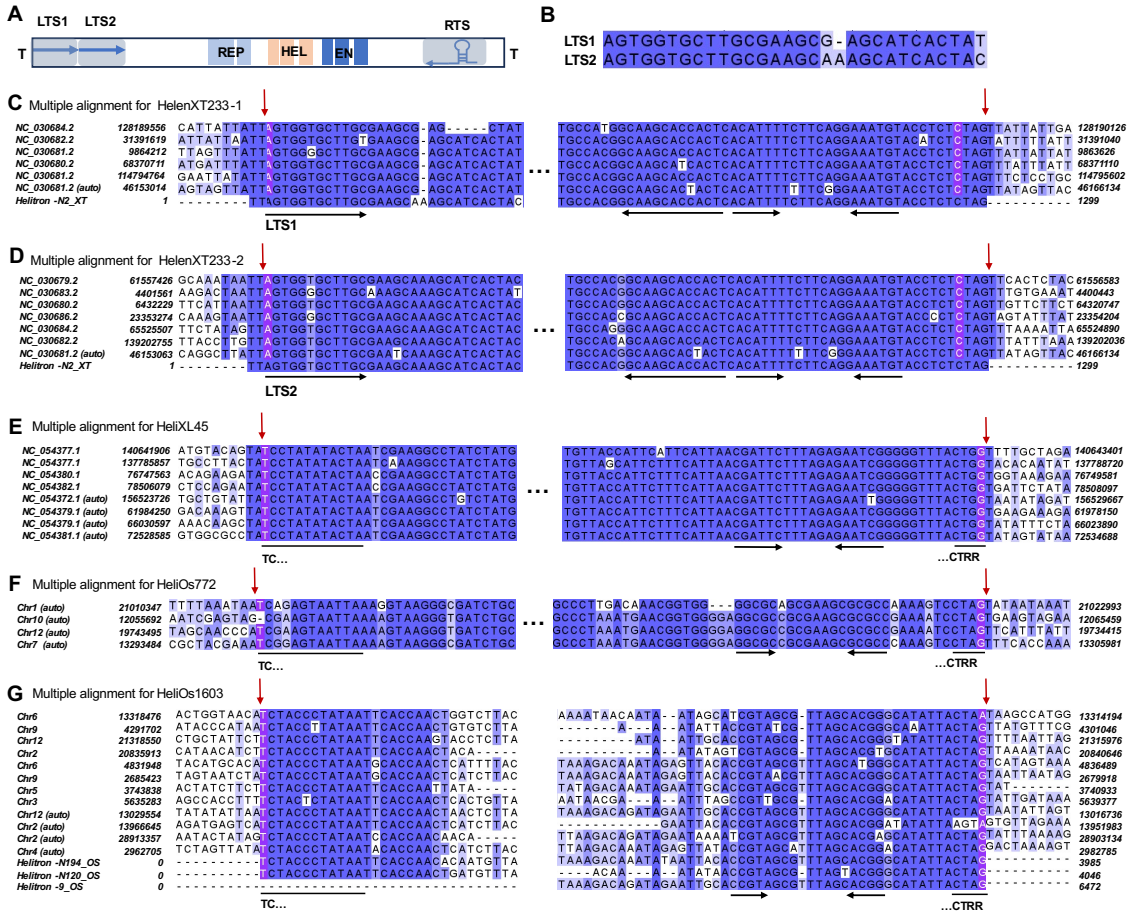

Figure 4. Multiple alignments of selected HLE1 and HLE2 insertions detected by HELIANO for *Xenopus* frog and *Oryza sativa* genomes. (A) Structure of the X. tropicalis autonomous HelenXT233. Two alternative LTSs were detected: LTS1 and LTS2. (B) Sequence alignment of LTS1 and LTS2 from the autonomous X. tropicalis HelenXT233. (C, D) Multiple alignments of insertions from HelenXT233 families (C) for HelenXT233-1 and (D) for HelenXT233-2. (E) A case for multiple alignments of X. laevis HLE1 insertions. (F, G) Cases of multiple alignments of HLE1 insertions in O. sativa genome. The autonomous insertions are labelled as 'auto' in each multiple alignment, and others are non-autonomous counterparts. The nucleotide highlighted in purple shows the predicted starts and stops by HELIANO. The down arrows in red indicate the precise insertion sites based on manual curation, using as a rule that HLE2 insert between T and T nucleotides and HLE1 between A and T nucleotides. Note the precise correspondences between the HELIANO annotation and the manual curation for HLE2 in E-F-G and the differences for HLE1 in C and D. The horizontal black arrows indicate terminal inverted repeats and stem-loop structures.

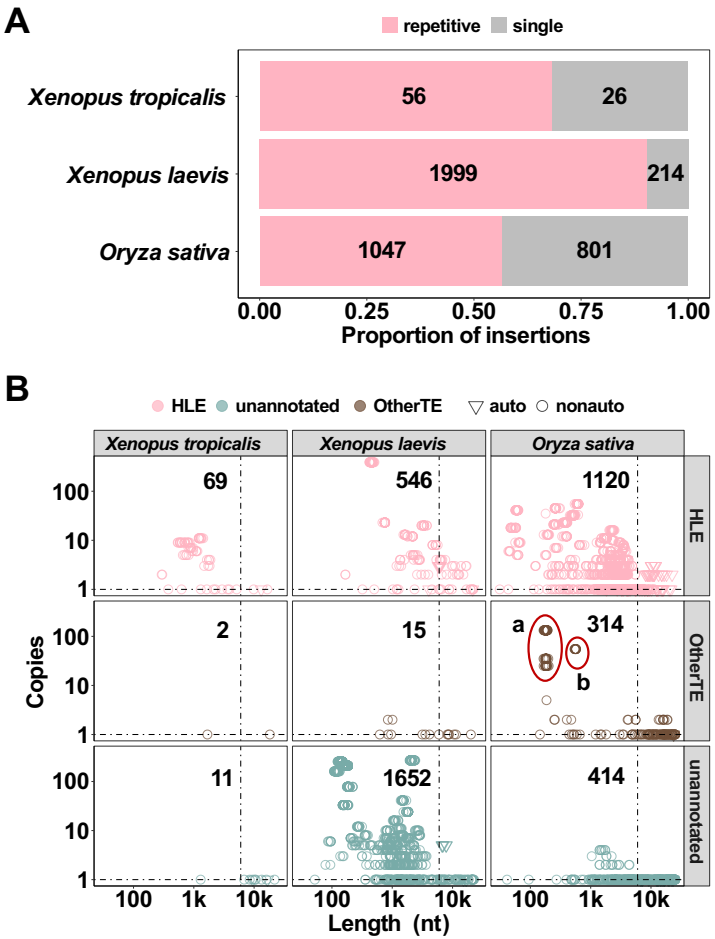

Figure 5. Reannotation of HELIANO predictions with known TEs. (A) The histogram shows the proportion of HELIANO predictions whose occurrence in the genome more than once (repetitive, pink) and once (single, grey). (B) Scatter plot shows the HELIANO predictions annotated as 'HLE' (pink), 'OtherTE' (brown), and 'unannotated' (cyan). Group a and b indicate ambiguous annotations in red circles.

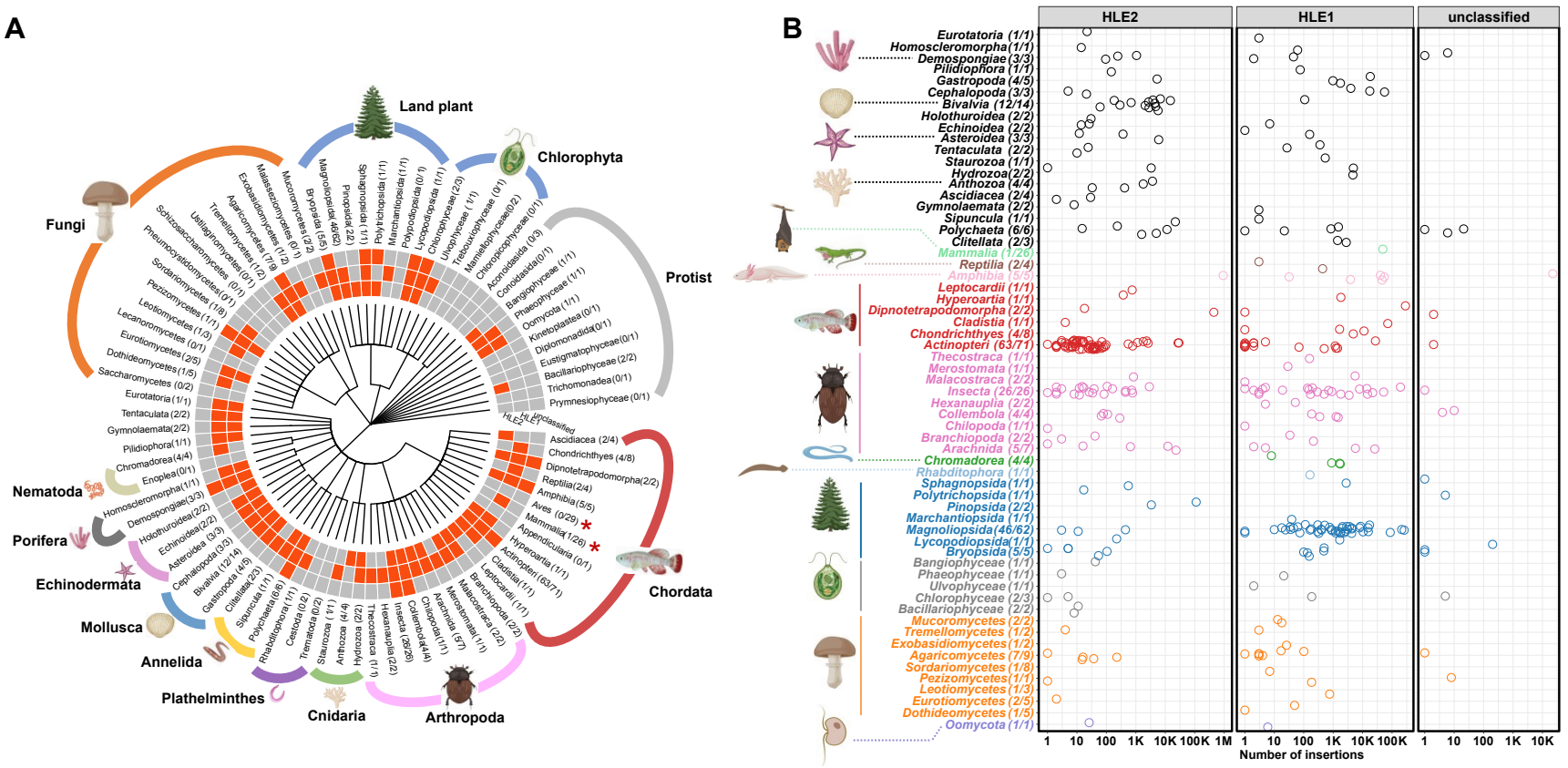

Figure 6. Distribution of HLEs among 404 eukaryote genomes. (A) A species tree obtained from NCBI indicated the phylogenetic relationship of sampled genomes. The fraction in each bracket represents the ratio of the number of species with HLEs to the number of all sampled species in a particular class. The heatmap indicates the presence (red) and absence (grey) of HLE groups in each sampled class. (B) The scatter plot shows the number of detected HLEs in each sampled class. Each point represents the number of corresponding HLE groups in a species. The fraction in each bracket represents the ratio of the number of species with HLEs to the number of all species in a particular class. The y-axis scale is log10 transformed.

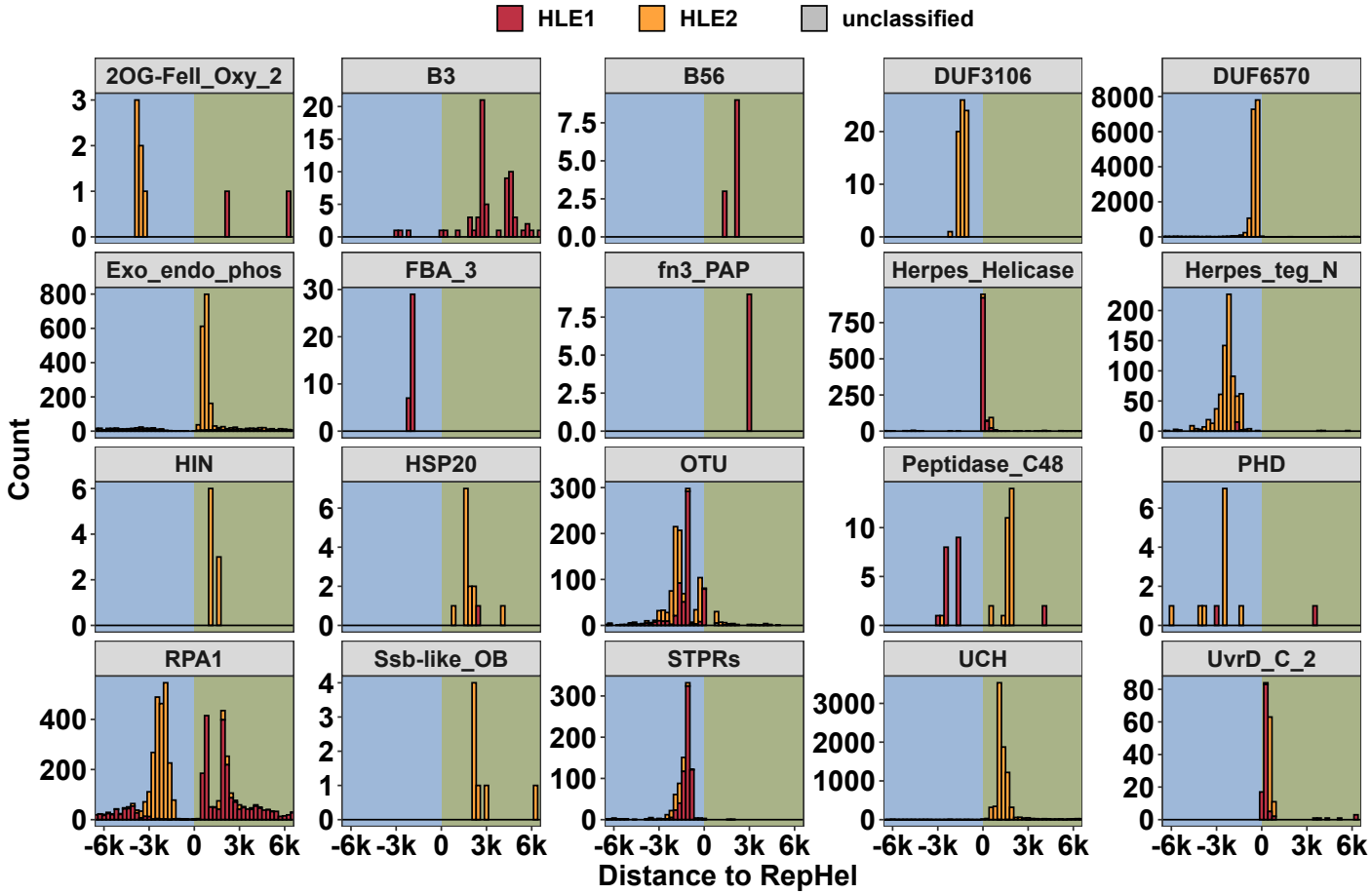

Figure 7. Distribution of the distance between RepHel and additional protein domains in HLEs. The zero value on the x-axis indicates the position of RepHel domains, negative values indicate the corresponding domains are upstream of RepHel, and positive value indicates their presence downstream of RepHel. The y-axis shows the count of HLEs.

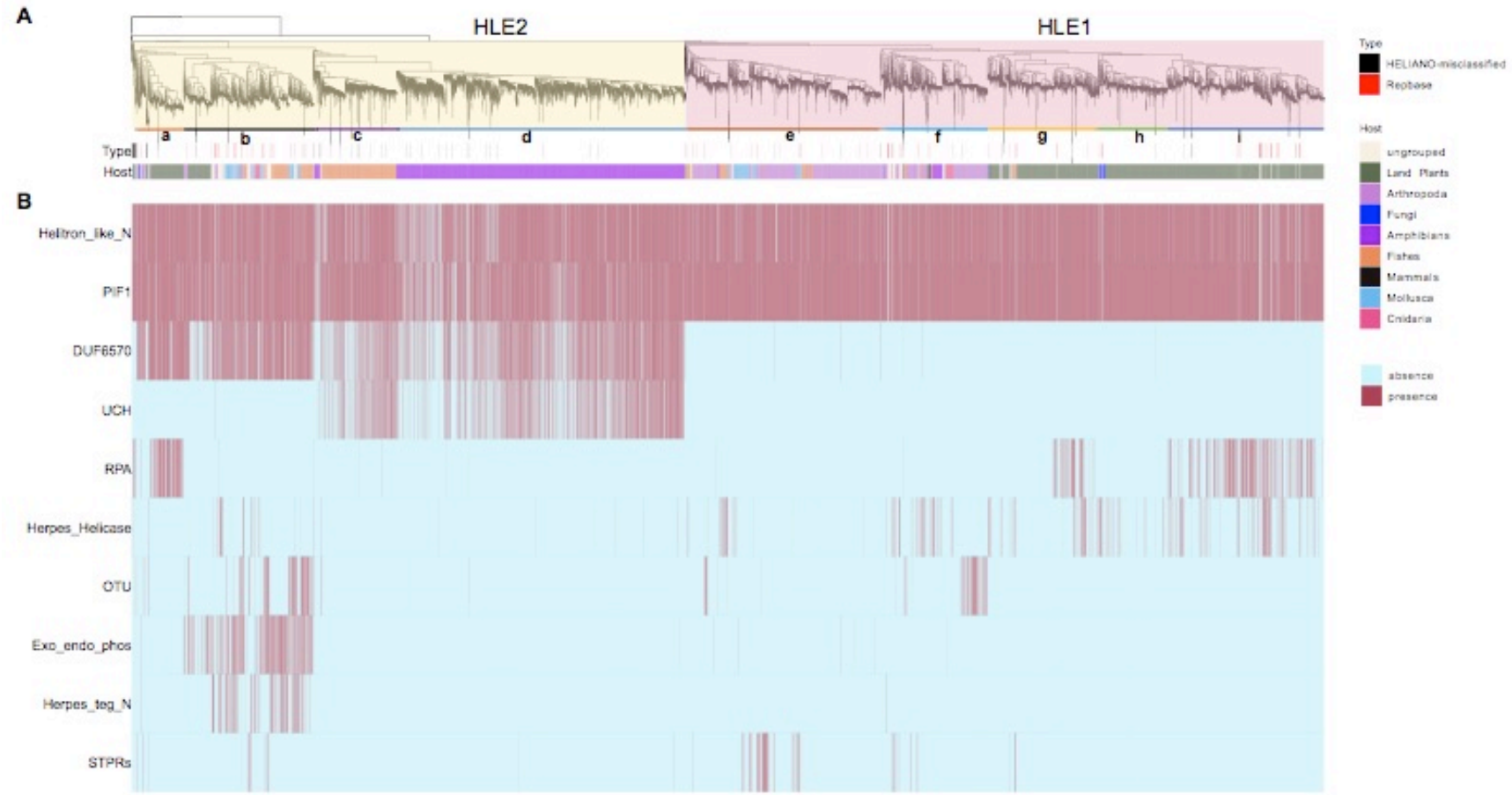

Figure 8. Distribution of HLEs and their captured domains in eukaryote genomes. (A) Maximum likelihood estimation tree of HLE transposases from sampled species (LogLk = -8062114.874). The HLE2 (light yellow block) and HLE1 (light red block) groups were further classified into subgroups: a-d for HLE2 and e-i for HLE1. Unclassified HLEs are in grey. The annotation below the tree entitled Type indicates the classification and source of HLEs. HLEs from Repbase are marked in red, and in black represent HELIANO misclassified HLEs. The annotation entitled host represents the species origin of HLEs. (B) The heatmap shows the presence or absence of additional domains in each corresponding HLE. Red indicates the presence of the domain, and light blue indicates its absence.