Pif1 Helicases and the Evidence for a Prokaryotic Origin of *Helitrons*

Pedro Heringer and Gustavo C.S. Kuhn*

Departamento de Genética, Ecologia e Evolução, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

***Corresponding author:** E-mail: gcskuhn@ufmg.br.

Associate editor: Irina Arkhipova

Abstract

***Helitrons* are the only group of rolling-circle transposons that encode a transposase with a helicase domain (Hel), which belongs to the Pif1 family. Because Pif1 helicases are important components of eukaryotic genomes, it has been suggested that Hel domains probably originated after a host eukaryotic Pif1 gene was captured by a *Helitron* ancestor. However, the few analyses exploring the evolution of *Helitron* transposases (RepHel) have focused on its Rep domain, which is also present in other mobile genetic elements. Here, we used phylogenetic and nonmetric multidimensional scaling analyses to investigate the relationship between Hel domains and Pif1-like helicases from a variety of organisms. Our results reveal that Hel domains are only distantly related to genomic helicases from eukaryotes and prokaryotes, and thus are unlikely to have originated from a captured Pif1 gene. Based on this evidence, and on recent studies indicating that Rep domains are more closely related to rolling-circle plasmids and phages, we suggest that *Helitrons* are descendants of a RepHel-encoding prokaryotic plasmid element that invaded eukaryotic genomes before the radiation of its major groups. We discuss how a Pif1-like helicase domain might have favored the transposition of *Helitrons* in eukaryotes beyond simply unwinding DNA intermediates. Finally, we demonstrate that some examples in the literature describing genomic helicases from eukaryotes actually consist of Hel domains from *Helitrons*, a finding that underscores how transposons can hamper the analysis of eukaryotic genes. This investigation also revealed that two groups of land plants appear to have lost genomic Pif1 helicases independently.**

***Key words:* *Helitrons*, transposon, Pif1, helicase.**

Introduction

Helitrons are DNA transposable elements (TEs) found in a wide variety of species from all eukaryotic kingdoms but make up variable genomic proportions across different taxa. For instance, they constitute between 0.1% and 6.6% of the genomic DNA in plants and between 0% and 10% in animals (reviewed in Kapitonov and Jurka [2007] and Thomas and Pritham [2015]). These TEs have been shown to mobilize within a genome by a process known as rolling-circle (RC) transposition (RCT) (Grabundzija et al. 2016, 2018) which could be viewed as a variation of the RC replication (RCR) process employed by several groups of plasmids and viruses from prokaryotes and eukaryotes (reviewed in Chandler et al. [2013] and Wawrzyniak et al. [2017]). In *Helitrons*, the RCT is executed by the Rep/Helicase (RepHel) transposase, which is composed by two major domains: an endonuclease (Rep) domain and a superfamily 1 helicase (Hel) domain (Thomas and Pritham 2015) (fig. 1).

Helitrons can be classified into four structural and coding variants, namely *Helitron*, *Helentron*, *Helitron2*, and *Proto-Helentron* (Thomas and Pritham 2015). In contrast to the first three variants, which have been shown to represent distinct phylogenetic groups (Poulter et al. 2003; Thomas et al. 2014; Heringer and Kuhn 2018), *Proto-Helentron* elements seem to constitute a subtype of *Helentrons* with derived *Helitron*-like structural features (Thomas et al. 2014). Although all *Helitrons* have RepHel proteins with two major domains, distinct variants, or specific variant lineages, can encode additional domains in their transposase or/and additional genes. Likewise, specific sets of structural features, like inverted repeats, can be used to identify major lineages or variants (fig. 1).

The Hel domain present in *Helitron* transposases is a superfamily 1 helicase, more specifically from the Pif1 family (Kapitonov and Jurka 2001; Thomas and Pritham 2015). Pif1 helicases have been found in essentially all eukaryotes studied to date (Bochman et al. 2010) and are involved in several processes, like DNA replication and repair, telomere maintenance, Okazaki fragment maturation, disruption of protein–DNA complexes, resolution of nucleic acid secondary structures, mitochondrial DNA maintenance, among others (reviewed in Boule and Zakian [2006]; Bochman et al. [2010]; and Muellner and Schmidt [2020]). Although typically known as eukaryotic proteins, Pif1-like helicases can

Article

also be found in some prokaryotic species, bacteriophages, and eukaryotic viruses (Bochman et al. 2011). We henceforth refer to eukaryotic and prokaryotic proteins that perform genomic-related tasks as genomic Pif1 helicases, in order to distinguish them from Pif1-like viral helicases or Hel domains found in *Helitron* transposases.

The structural and mechanistic similarities between eukaryotic and prokaryotic RC transposons initially prompted the hypothesis that *Helitrons* could be descendants of bacterial elements (e.g., IS91 family). Furthermore, it was suggested that *Helitron* ancestors could have given rise to eukaryotic RCR viruses, as these viruses were only found in plant species at that time (Kapitonov and Jurka 2001). Conversely, because geminiviruses had been found integrated into plant chromosomes, it was also proposed that *Helitrons* could likewise be derived from an ancient genomic integration of a eukaryotic RCR virus (Feschotte and Wessler 2001). However, as revealed by recent findings, Rep domains from *Helitrons* are distantly related to proteins from prokaryotic TEs and eukaryotic viruses, and share more similarities with RCR plasmids and viruses from bacteria (Heringer and Kuhn 2018; Kazlauskas et al. 2019). In spite of these similarities, the prokaryotic plasmid and viral elements which are more closely related to *Helitrons* do not encode a helicase domain (Heringer and Kuhn 2018), what makes the origin of Hel domains a still unsolved issue. The absence of helicases on the coding sequences of prokaryotic RC TEs, together with the presence of introns in some Hel domains from plants and *Caenorhabditis elegans Helitrons*, have been considered as tentative evidences that a *Helitron* ancestor acquired its Hel domain by capturing a helicase gene from its eukaryotic host (Kapitonov and Jurka 2001, 2007; Thomas and Pritham 2015). However, we still lack information about the evolutionary origins of *Helitron* Hel domains and their relationship with other helicases, as these issues have never been investigated in detail.

The fact that Pif1 family helicases are present in virtually all eukaryotes but absent in RC mobile genetic elements (MGEs), except *Helitrons*, renders the investigation about the origin of Hel domains more difficult. Moreover, to our knowledge there are no automated methods to clearly distinguish genomic Pif1 helicases from *Helitron* Pif1-like helicases. Regarding the later issue, both genomic and *Helitron* Pif1-like sequences can be found in eukaryotic genomes and sometimes is not possible to discriminate them without a more detailed analysis. For instance, Blastp searches on eukaryotic genomes using Pif1 proteins as queries often result in multiple significant hits, even though most eukaryotic species apparently have only one or two genomic Pif1 helicases (Bochman et al. 2010). Therefore, although up to few hits are expected to represent genomic Pif1 helicases in eukaryotic species, most of them often constitute *Helitron* Pif1-like protein sequences. In addition, some eukaryotes apparently have multiple genomic Pif1 paralogs (Bochman et al. 2010, 2011; Harman and Manna 2016), which makes their distinction from *Helitron* Pif1-like helicases even more complex.

In the present study, we retrieved prokaryotic, eukaryotic and viral Pif1-like proteins in silico using a stepwise searching method to avoid classifying *Helitron* coding sequences as genomic helicases. After doing so, we were able to investigate the relationship between Hel domains and Pif1-like genes from a wide variety of organisms and MGEs. Our results reveal further valuable information about the evolution of RepHel transposases, indicating that Hel domains are only distantly related to genomic Pif1 helicases and were likely present in *Helitrons* before they invaded eukaryotic hosts. We discuss the general implications of our findings considering the known mechanistic features of RepHel transposases and Pif1 helicases, also demonstrating how the similarities between these proteins can interfere with their classification and analysis.

Results

Finding Genomic Helicases

Before conducting searches to retrieve genomic Pif1-like helicases, we first expanded our sample of *Helitrons* from different variants (*Helitrons*, *Helentrons*, and *Helitron2*) selected previously (Heringer and Kuhn 2018). Consensus sequences from the helicase domains (Hel) found in those Helitrons were used as queries to obtain Pif1-like helicases from a wide diversity of organisms (see Materials and Methods). Because *Helitrons* are found throughout a large portion of eukaryotic genomes, the distinction between genomic Pif1 and *Helitron* Pif1-like helicases (Hel domains) across individual species is highly prone to identification errors (supplementary fig. S1, Supplementary Material online). For that reason, we initially selected only organisms lacking *Helitron* Rep sequences in their genomes, so that genomic Pif1 helicases could be correctly identified before our analyses. *Helitron* Rep sequences can be used as unique identifiers for the presence of *Helitrons* as they are exclusive of these RC elements and do not have genomic counterparts in eukaryotes.

The larger or smaller representation of specific taxonomic groups in the Pif1 helicases selected initially, depended on the number of available genomes and on the presence or absence of *Helitrons* in each taxon. For instance, although our searches on Embryophyta (land plants) revealed the presence of Pif1-like proteins in most species, only the common liverwort *Marchantia polymorpha* was devoid of Rep sequences from *Helitrons*, thus being the single representative of land plants selected in the first round of searches.

Although almost all retrieved sequences from prokaryotes and eukaryotes were annotated as genomic Pif1 helicases, one of the hits from the searches on archaea was a TraA relaxase annotated as belonging to a species from the *Methanothrix* genus (*Methanothrix* sp., accession number: TFH49976.1). This hit displays a relatively low sequence coverage (62%) and identity (24%) to the query (*Helentron* Hel consensus) (supplementary data S1, Supplementary Material online). Nevertheless, as TraA relaxases constitute a group of proteins involved in conjugation of bacterial plasmids and are also known to have a helicase domain (Alt-Mörbe et al. 1996; Pérez-Mendoza et al. 2006), we decided to include additional TraA relaxase representatives in our analysis. To do that, the *Methanothrix* TraA relaxase (TFH49976.1) was used as query

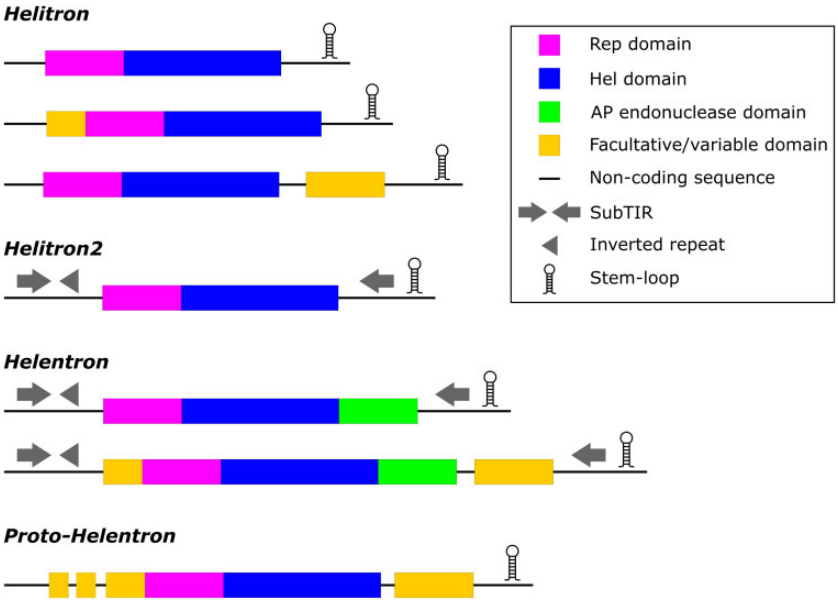

Fig. 1. *Helitron* structural and coding variants. Each variant can be identified by a set of structural (symbols) and coding sequences (colored boxes). *Helitrons*, *Helitron2*, and *Helentrons* are major phylogenetic variants, with *Proto-Helentrons* representing an internal group of *Helentrons* that have intermediate features found in *Helitrons* and *Helentrons*. Adapted from Thomas and Pritham (2015).

in Blastp searches on the nonredundant protein sequences (nr) database from GenBank (Sayers et al. 2019). Interestingly, the best hits from this search consisted of TraA sequences from the phylum Proteobacteria (supplementary table S1, Supplementary Material online), with no hits from archaeal species, indicating that TFH49976.1 could either represent a horizontally transferred gene (from a bacterium to an archaeon) or a misannotated sequence from a bacterium species (discussed in the next topic).

Using our stepwise search and selection method (schematic workflow depicted in fig. 2), we retrieved an initial sample of 76 putative genomic Pif1 helicases from a wide variety of eukaryotes, prokaryotes, and plasmids, all lacking *Helitron* sequences in their genomes. After retrieving this sample of genomic (and plasmid) helicases, we further expanded the number of proteins in our data set by selecting Pif1-like helicases in all major groups of eukaryotes, prokaryotes and viruses, without filtering taxa by the presence of *Helitron* sequences. In addition to Hel domain consensus sequences, this time we also used the *Saccharomyces cerevisiae* Pif1 (NP_013650.1) as a query in Blastp searches. The proteins identified and selected previously with the Rep-filtering procedure were used to aid in the classification of this new set of Pif1-like proteins as genomic helicases or Hel domains from Helitrons by their relationship revealed in the phylogenetic analysis. We also included eukaryotic and prokaryotic viruses in this step of Blastp searches. All taxa selected for further analyses are shown in supplementary table S1, Supplementary Material online.

Phylogenetic Analysis

We used our final sample of 310 aligned protein sequences from *Helitrons*, eukaryotic and prokaryotic organisms, plasmids and viruses, to infer their phylogenetic relationship using the Maximum Likelihood method. Our resulting phylogeny revealed seven well supported major clades (or groups), named as follows: 1) TraA, 2) *Myoviridae*, 3) nucleocytoplasmic large DNA viruses (NCLDV)/*Baculoviridae*, 4) *Helentron*/*Helitron2*, 5) *Helitron*, 6) Prokaryotic, and 7) Eukaryotic clade (fig. 3). The TraA clade included exclusively TraA relaxases and constitute a sister group of the *Myoviridae* clade, which is composed by helicases from a subset myoviruses. The NCLDV/*Baculoviridae* group included helicases from a subset of NCLDV and all retrieved baculoviruses. Together with the *Helentron*/*Helitron2* and *Helitron* clades, they represent a basal group relative to the Prokaryotic and Eukaryotic major clades, as shown in the rooted tree (supplementary fig. S2, Supplementary Material online). The Prokaryotic clade includes most bacterial, archaeal and bacteriophage sequences. In contrast, the Eukaryotic major clade, which formed a sister group with the Prokaryotic clade, included all eukaryotic sequences, plus some bacterial, archaeal, eukaryotic viruses, and bacteriophage sequences, being the most diverse group in the phylogeny.

Regarding the distribution of *Helitron* variants, we observed two distinct and well supported clades, one with *Helitron* and the other containing *Helentron* plus *Helitron2* sequences (fig. 3). However, the connection between these two clades, and between each one of them and other groups of helicases, have low branch support values, and thus are presented collapsed in the phylogeny (fig. 3; supplementary fig. S2, Supplementary Material online). Considering previous analyses involving the Rep domain (Poulter et al. 2003; Heringer and Kuhn 2018) and the fact that a monophyletic origin of all *Helitrons* seems more parsimonious, the observed paraphyletic distribution of two major *Helitron* groups in our phylogeny could represent a methodological artifact (see Discussion). Nevertheless, the

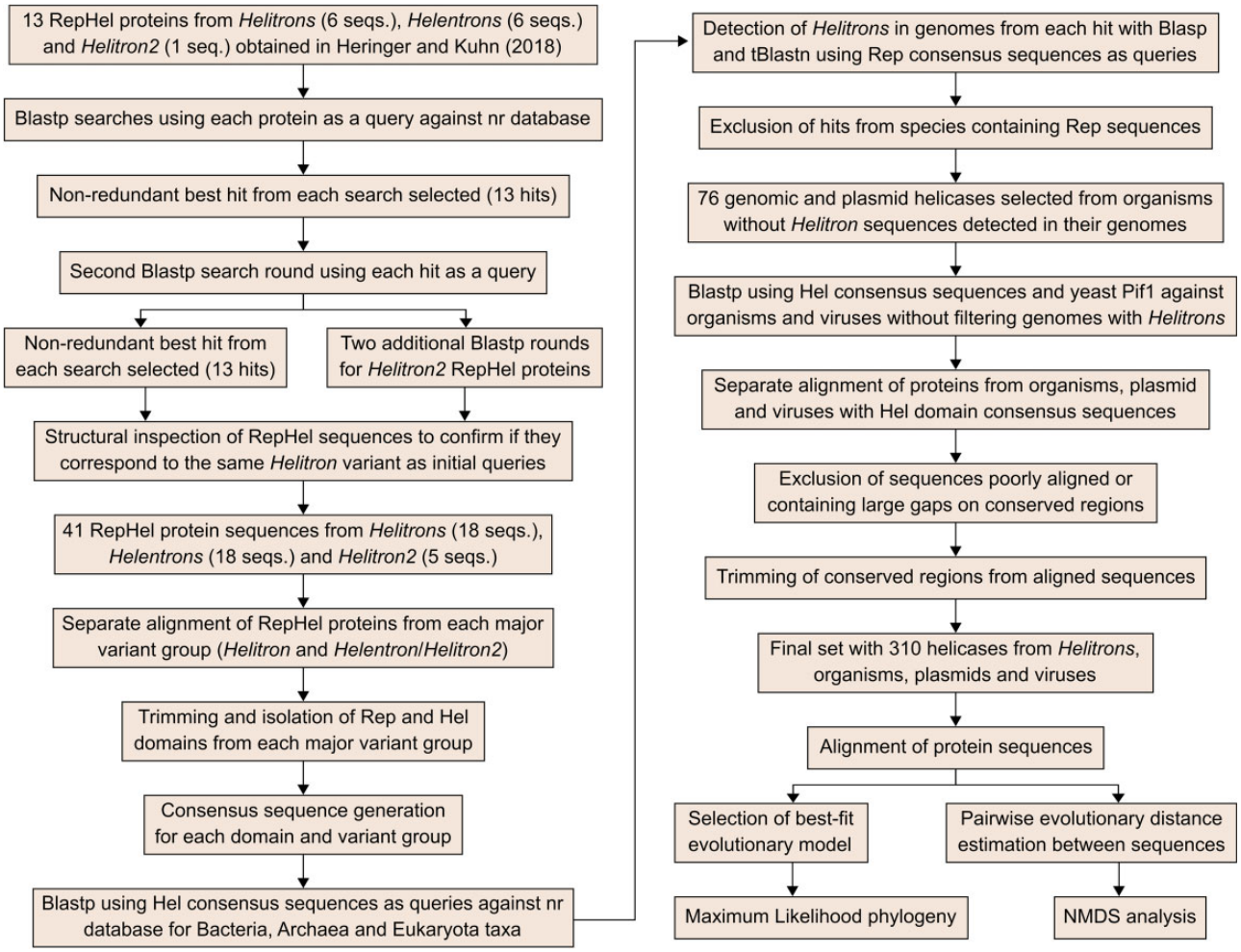

Fig. 2. Workflow with the methodology used in our study. See Materials and Methods for a more comprehensive description.

fact that *Helitrons* in general did not group closer to any other major clade, indicates that Hel domains are only distantly related to genomic Pif1 helicases and belong to completely independent lineages. An interesting aspect of the *Helentron*/*Helitron2* major clade is the presence of a Hel domain from the dinoflagellate *Symbiodinium microadriaticum* (CAE7237458.1) branching externally to the divergence of *Helitron2* and *Helentron* sequences (fig. 3; supplementary fig. S2, Supplementary Material online). This RepHel lacks the apurinic–apyrimidinic (AP) endonuclease domain typical of *Helentrons*, and the element corresponding to this transposase (CAJNJV010003184.1) is structurally more similar to a *Helitron2* variant (fig. 1). Hence, this *Helitron2*-like element appears to represent an intermediate variant that should be more closely related to the common ancestor of *Helentron* and *Helitron2* elements. To our knowledge, this is the first identification of a putative evolutionary intermediate between two *Helitron* variants. In this specific case, the putative intermediate variant was not identified before most likely because the *S. microadriaticum* sequence (CAE7237458.1) was submitted only recently (February 2021).

One of the prokaryotic sequences in the Eukaryotic major clade is a Pif1-like helicase from a Rickettsiales bacterium (MBO87943.1), positioned before the radiation including most eukaryotic Pif1 sequences (fig. 3). Most phylogenetic analyses conducted to date place the order Rickettsiales as the closest relative of mitochondria (reviewed in Roger et al. [2017]). Although this hypothesis has been challenged by some studies (Roger et al. 2017; Martijn et al. 2018), a recent analysis that used more robust methods confirmed the close relationship between Rickettsiales and the mitochondrion ancestor (Fan et al. 2020). Hence, the topology observed in our phylogeny seems to reflect the known evolutionary link between eukaryotic Pif1 proteins and their prokaryotic ancestor, which probably belonged to the symbiont that later gave rise to mitochondria (Bochman et al. 2011).

Another marked feature observed in our phylogeny is the presence of Pif1-like sequences from three eukaryotic species (*Perkinsela* sp., *Phytomonas* sp., and *Strigomonas culicis*) preceding the prokaryotic radiation within the Eukaryotic major clade (fig. 3; supplementary fig S2, Supplementary Material online). These sequences belong to kinetoplastids from the phylum Euglenozoa which, accordingly, is considered the group that diverged earliest during eukaryotic evolution (Cavalier-Smith et al. 2014). Although other kinetoplastid species are grouped separately from these three basal taxa (fig. 3), this distribution could be explained by the presence of

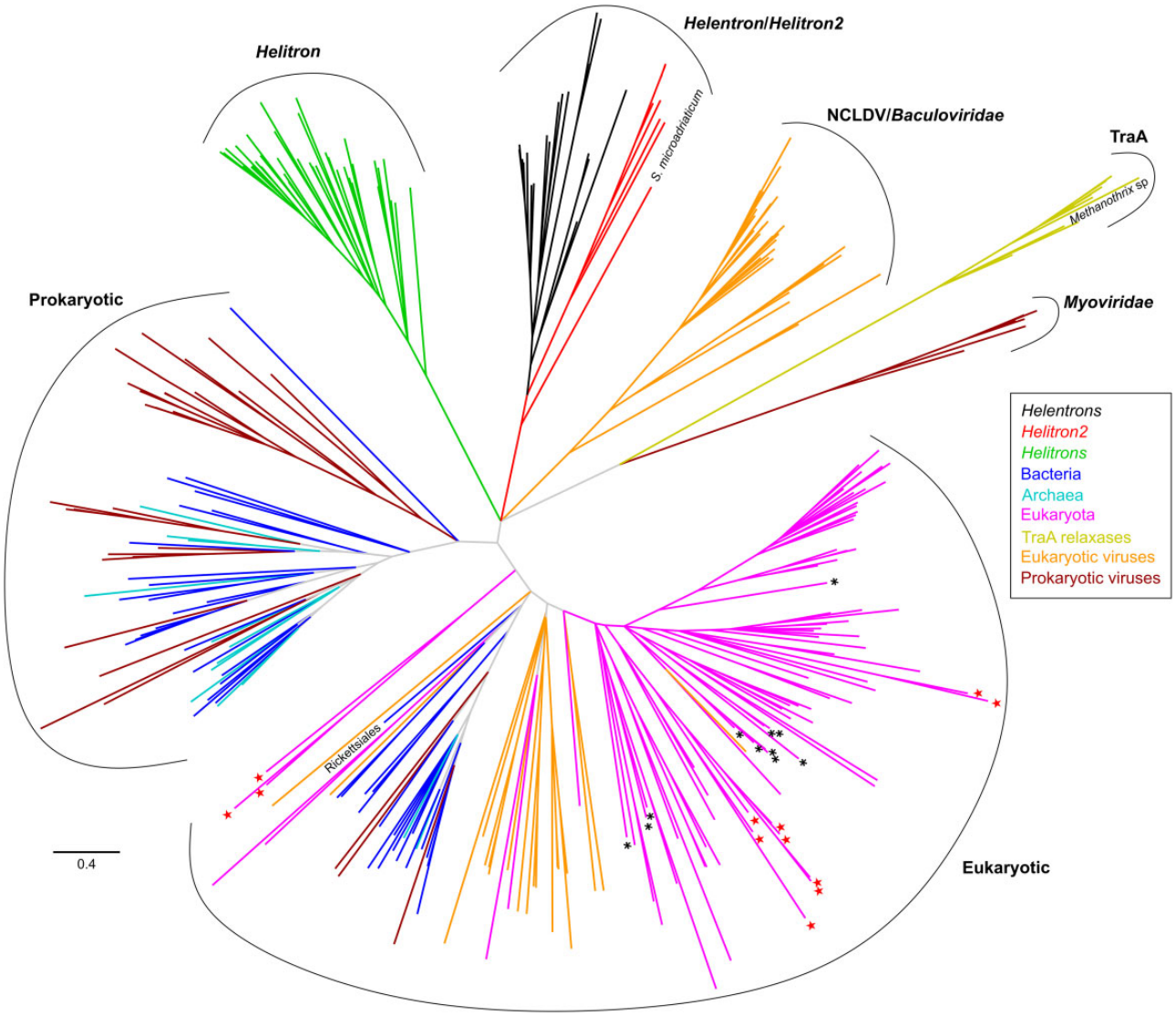

Fig. 3. Maximum-likelihood phylogeny of Pif1-like helicases. The resulting phylogeny includes Pif1-like helicases from *Helitron* variants, viruses, plasmids, and organisms, with seven major clades indicated around the tree. Specific taxa mentioned in the text are shown in branch tips. Kinetoplastids are marked with red stars and amoebae are marked with asterisks. Branches with <0.7 SH-aLRT statistical support were collapsed. The rooted tree with all taxa names and branch support values is shown in supplementary figure S2, Supplementary Material online.

multiple Pif1 paralogs in species from this class, which have been shown to encode up to eight Pif1-like genes (Liu et al. 2009; Bochman et al. 2010). If these three basal sequences represent some of the Pif1 paralogs adapted for kinetoplastid-specific functions (Bochman et al. 2010), a process of positive evolution following subfunctionalization, might have caused them to be artificially positioned externally in relation to other eukaryotic Pif1 helicases. In addition to kinetoplastids, other taxa also displayed a somewhat scattered distribution on the Eukaryotic major clade, instead of forming monophyletic clusters. For instance, amoebal Pif1 helicases were grouped in five separate clades (fig. 3). Interestingly, a scattered distribution of amoebal Pif1-like proteins was also observed in a previous study and it was explained as the result of horizontal gene transfer (HGT) and duplication events (Harman and Manna 2016). Also in the Eukaryotic major clade, eukaryotic viruses, mostly NCLDVs, were found dispersed in different clades, sometimes closer to eukaryotic and prokaryotic organisms than to other groups of viruses (fig. 3; supplementary fig. S2, Supplementary Material online). Although noteworthy, this result agrees with the growing evidence for multiple HGT events between these large viruses and a variety of organisms (reviewed in Barreat and Katzourakis [2021]).

Overall, the scattered topology observed for several taxa from the Eukaryotic major clade might have been the consequence of two main factors. First, as a result of our searching and selection method designed to retrieve Pif1-like helicases with the highest similarity to specific queries. Because we only selected the best results from each taxonomic group, and eukaryotes may have multiple Pif1 genes adapted for distinct functions, it is likely that our sampled sequences represent a mixture of paralogs and orthologs. Second, as a consequence of several HGT events between eukaryotes, prokaryotes, and

viruses. Eukaryotes have been involved in HGT exchanges not only with viruses, as mentioned above, but also with multiple prokaryotic groups and sometimes with distinct eukaryotic taxa (reviewed in Husnik and McCutcheon [2018] and Van Etten and Bhattacharya [2020]). Thus, it is possible that Pif1 genes have been horizontally transferred several times during the evolution of eukaryotes.

In the Prokaryotic major clade, cases of interspersed branches from bacteria, archaea, and phages were also abundant, and indicate that several HGT events involving Pif1-like genes have occurred between these taxa (fig. 3). Although horizontally transferred sequences represent a relatively small fraction of eukaryotic genomes, in prokaryotes, HGT has long been considered a primary source of new genes and a major driver of evolution. These gene exchanges are not limited to closely related organisms, as they have been shown to cross prokaryotic domains and sometimes occur between bacteria, archaea and viruses (reviewed in Koonin [2016]). Hence, based on our phylogenetic analysis, it is reasonable to conclude that Pif1-like helicases are also members of the large set of gene families that have been horizontally transferred among prokaryotic organisms. Regardless of the particular explanations for each case, the frequent grouping of relatively distant taxa observed in the Eukaryotic and Prokaryotic major clades indicates that, in addition to ordinary vertical inheritance of genes, other events (e.g., HGTs and gene duplications) have shaped the evolution of genomic Pif1 helicases extensively.

Other interesting results were also revealed by the phylogenetic analysis. For instance, the TraA and *Myoviridae* clades formed sister groups with good branch support (fig. 3; supplementary fig. S2, Supplementary Material online). This result suggests a closer than expected relationship between replicons with completely distinct modes of propagation, underscoring the highly dynamic modularity that is typical of MGEs. Finally, as previously indicated in our Blast results, a protein annotated as belonging to the archaeon genus *Methanothrix* (TFH49976.1) grouped with TraA relaxases from Proteobacteria species, more specifically in the Desulfobacteraceae family (*Desulfobacteraceae bacterium* and *Desulfosarcina cetonica*) (fig. 3; supplementary fig. S2, Supplementary Material online). To verify whether this TraA gene derives from an HGT event or misannotation, we first used its protein sequence (TFH49976.1) as a query in separate Blastp searches against bacteria and archaea in the nr database. In this case, the query was significantly more similar to bacterial than archaeal sequences. We also used the nucleotide sequence corresponding to the protein (accession number: SPBB01000211.1) as a query in Blastn searches against bacteria and archaea in the nucleotide collection (nr/nt) and Whole Genome Shotgun (WGS) contigs databases. In this case, no hits with significant similarity were found in archaea. The query displays a significant identity (up to 75%) to bacterial genes, although limited to short stretches that cover up to 15% of the query length. Furthermore, the contig corresponding to the query only contains the TraA gene without flanking sequences that could be used to determine if this gene was integrated into an archaeal genome. Therefore, this putatively archaeal TraA gene is significantly more similar to bacterial than archaeal sequences, both at the amino acid and nucleotide level. Because this sequence is part of a metagenome assembly (BioSample: SAMN11127048), the possibility of misannotation or contamination in this case is very likely. Together, our analyses indicate that this TraA gene is likely from a bacterial plasmid misannotated as belonging to an archaeon. Regardless of those considerations, knowing the host species of this protein sequence does not change the interpretation of our results.

NMDS Analysis

The estimated evolutionary divergence between sequences were used to represent their distances in two dimensions with nonmetric multidimensional scaling (NMDS) analysis. By doing so, we intended to visualize their spatial arrangement without assuming cladistic relationships, and also verify if their distribution replicates the overall topology observed in the phylogeny.

The arrangement of Pif1-like helicases in the resulting NMDS ordination showed an overall segregation of proteins into seven major clusters (fig. 4). It also displayed a large divergence between Hel domains from the two major clades previously observed in our phylogeny (fig. 3), with *Helentron* and *Helitron2* sequences forming a single group distinctly segregated from *Helitron* variant sequences. In addition, *Helitron* Pif1-like domains from all variants did not appear to be more closely associated with any other specific major group, being roughly equidistant from genomic and viral helicases found in prokaryotes and eukaryotes (fig. 4).

Pif1 helicases from the Eukaryotic and Prokaryotic major groups formed two separate, albeit closely related clusters. Although genomic Pif1 helicases in the Eukaryotic group showed a tendency for clustering with sequences from more closely related taxa, in the Prokaryotic group, sequences from bacteria and archaea displayed a highly interspersed distribution. In both major groups viral sequences were mostly scattered among genomic Pif1 helicases (fig. 4). These distinct arrangements in the Eukaryotic and Prokaryotic major groups confirm the taxonomic incongruences and complex evolutionary history of genomic Pif1 helicases indicated by the phylogenetic analysis.

In sum, the resulting NMDS ordination recapitulates the main features observed in the phylogeny, that is, the segregation of seven major clades, the distant relationship between Hel domains from *Helitrons* and genomic helicases, and the indication of multiple HGT events involving Pif1-like helicases from eukaryotes, prokaryotes, and viruses.

Reassessing the Classification and Number of Pif1 Genes in Eukaryotes

As we have mentioned, Blastp searches on eukaryotic genomes using Pif1 helicases as queries often result in multiple significant hits. Because *Helitrons* are pervasive in most eukaryotic groups and their transposase includes a Pif1-like Hel domain, it is always possible that some of those hits constitute *Helitron* coding sequences, instead of genomic helicases. For example, during our preliminary analyses we

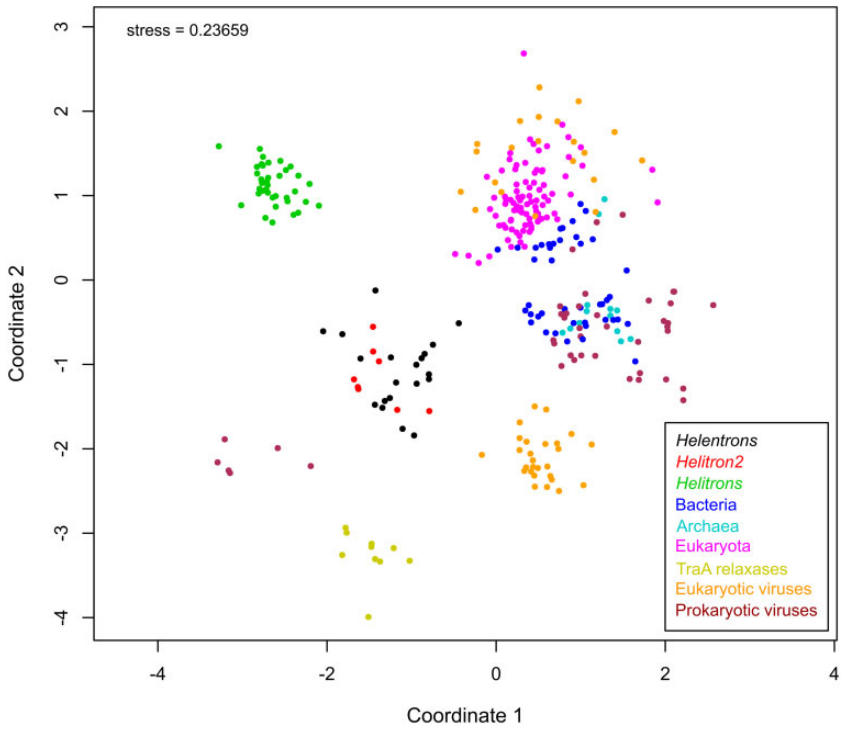

Fig. 4. NMDS plot of Pif1-like helicases. NMDS ordinations representing Euclidean distances between Pif1-like helicase sequences in two dimensions.

performed a Blastp search to identify putative genomic Pif1 helicases in the fungus *Rhizophagus clarus*, using the human Pif1 domain (6HPH_A) as a query, and found many candidate genes, together with RepHel sequences. However, a more detailed inspection revealed that some putative genomic Pif1 helicases are in fact Hel domains from *Helitron* coding sequences lacking the Rep domain in the same ORF (supplementary fig. S1, Supplementary Material online). Thus, without more careful analyses, the structural resemblance between genomic and *Helitron*-derived Pif1 domains can hinder the proper identification of sequences from this protein family. Indeed, to avoid classifying Hel domains as genomic Pif1 helicases, we excluded all species with *Helitrons* in their genomes from our initial Blast searches.

Although some eukaryotes are thought to have multiple genomic Pif1 helicases (Bochman et al. 2010, 2011; Harman and Manna 2016), most species from this domain of life apparently encode one or two Pif1 genes (Bochman et al. 2010). Considering that distantly related eukaryotes like *Schizosaccharomyces pombe* and humans only need one Pif1 helicase to carry out genomic functions, species with supposedly multiple Pif1 paralogs should be evaluated carefully. Thus, we reassessed three cases in the literature referring to genomic Pif1 genes from eukaryotes, which could have included *Helitron*-derived sequences inadvertently.

In the first example, *Arabidopsis thaliana* was described as having three genomic Pif1 helicases (CAB91581, NP_190738, and CAB63155) (Bochman et al. 2010). After examining the structure and sequence of these proteins we found that all of them are either RepHel proteins or Pif1-like sequences with significant identity to *Helitron* transposases (supplementary table S2, Supplementary Material online). Interestingly, a phylogeny of Pif1 sequences presented in the same work (Figure 1 in Bochman et al. 2010) displays a single Pif1 helicase from *Oryza sativa* (ABB47755) grouped together with the three *A. thaliana* proteins mentioned above. Because these three proteins were shown to be derived from RepHel transposases, and *Helitrons* are known to be abundant in the genomes of *A. thaliana* and *O. sativa* (Yang and Bennetzen 2009; Xiong et al. 2014), we examined this Pif1-like sequence from *O. sativa*. After inspecting its structure, we found that this *O. sativa* Pif1-like protein represents a RepHel transposase containing both of its major domains (supplementary table S2, Supplementary Material online). Hence, all these four proteins classified as genomic Pif1 helicases from *A. thaliana* and *O. sativa* constitute either RepHel transposases or Pif1-like Hel domains from *Helitrons*.

In the second example, the fungal pathogen of insects *Metarhizium robertsii* ARSEF 23 (formerly *M. anisopliae* ARSEF 23) was described as the eukaryote harboring the largest number of Pif1 genes, with 23 paralogs (Bochman et al. 2011). We conducted a Blastp search on the genome of this species using the human Pif1 domain (6HPH_A) and the *S. cerevisiae* Pif1 (NP_013650.1) as queries and found that, although *M. robertsii* appears to have up to 25 proteins with some similarity to Pif1 helicases, only 16 of them cannot be readily classified as RepHel transposases, that is, do not contain a Rep domain sequence. Of these 16 proteins, 11 either display significant similarity to RepHel transposases or belong to a cryptic RepHel ORF (truncated transposase with a Rep sequence upstream the Pif1 ORF), and one does not correspond to a Pif1 helicase (supplementary table S3,

Supplementary Material online). Hence, only four helicases from *M. robertsii* could represent genomic Pif1 candidates, with the other 20 Pif1-like sequence clearly being derived from *Helitron* transposases.

In the third example, it was suggested based on in silico analyses that *A. thaliana* could have up to 11 Pif1 genes (Knoll and Puchta 2011), with this large number of paralogs being attributed to *Helitrons* capturing and multiplying genomic Pif1 sequences. However, after inspecting all *A. thaliana* Pif1-like proteins on GenBank, retrieved after a Blastp searches using the human Pif1 domain (6HPH_A) and the *S. cerevisiae* Pif1 (NP_013650.1) as queries, we found that all of them either represent RepHel proteins directly or derive from *Helitron* transposases (supplementary table S4, Supplementary Material online). Although we anticipated that some sequences would derive from *Helitrons*, the fact that all retrieved *A. thaliana* Pif1-like proteins appear to represent RepHel transposases directly or indirectly was unexpected, considering the widespread distribution of genomic Pif1 helicases in eukaryotes. To investigate whether this apparent lack of genomic Pif1 homologs is exclusive from *A. thaliana*, we conducted a Blastp search using the same method on *O. sativa*, which is estimated to have diverged from *A. thaliana* ~163 Ma (Li et al. 2019). Like what was observed in *A. thaliana*, we found many Pif1-like sequences in *O. sativa*, with all results representing RepHel transposases directly or indirectly (supplementary table S5, Supplementary Material online).

Given the distant relationship between *A. thaliana* and *O. sativa*, we tried to estimate when genomic Pif1 helicases could have been lost during the evolution of these land plant lineages. To do that, we conducted a series of Blastp searches on taxonomic ranks above *A. thaliana* and *O. sativa* using the human Pif1 domain (6HPH_A) and the yeast Pif1 (NP_013650.1) as queries. Interestingly, genomic Pif1 homologs appear to have been lost in Brassicales and commelinids, the taxonomic groups from which *A. thaliana* and *O. sativa* belong, respectively (fig. 5). The best hits within these groups corresponded to RepHel proteins (supplementary table S6, Supplementary Material online). Conversely, the best hits from searches in taxa outside Brassicales (malvids) and commelinids (Liliopsida) were Pif1 proteins with low similarity to RepHel transposases, despite some of the species with putative genomic Pif1 helicases also having *Helitron* proteins (supplementary table S6, Supplementary Material online). To further confirm the absence of genomic Pif1 homologs in the mentioned groups, we first used the best hits from searches in malvids (EOX92974.1) and Liliopsida (MQL92731.1) as queries in Blastp searches against Brassicales and commelinids, respectively. The results still indicated a lack of genomic Pif1 homologs in Brassicales and commelinids, as the best hits also corresponded to *Helitron* sequences (supplementary table S7, Supplementary Material online). Additionally, we conducted Blastn searches using the nucleotide sequences corresponding to EOX92974.1 (CM001879.1) and MQL92731.1 (NMUH01001479.1) as queries against Brassicales and commelinids, respectively. Although the search against commelinids did not retrieve hits with significant similarity to the genomic Pif1 from Liliopsida, the result from Brassicales revealed a hit in *Bretschneidera sinensis* (JACXJD010000007.1) with 74% identity to the genomic Pif1 nucleotide sequence from malvids. This hit from *B. sinensis* translates to an ORF that appears to be intact, therefore representing a Pif1 gene that has not been annotated yet, which explains its absence in Blastp results. Interestingly, *B. sinensis* (family Akaniaceae) belongs to the most basal clade from Brassicales (Edger et al. 2018), indicating that genomic Pif1 homologs were probably lost shortly after the origin of this order and before the major radiation that gave rise to most extant families of Brassicales.

Although regions flanking genomic Pif1 helicases from malvids and Liliopsida up to tens of kilobase pairs on both sides display similarity to Brassicales and commelinids sequences, this similarity covers only limited portions of their length, as indicated by Blastn searches. Because this observed similarity is not contiguous over the whole span of flanking sequences, it is not possible to define whether they correspond to homolog regions, and therefore we could not determine what caused Pif1 genes to be lost in Brassicales and commelinids. However, it is noteworthy that most of the genomic Pif1-flanking regions with significant identity to sequences from both groups correspond to TEs, particularly LTR retrotransposons, as determined by searches using the Censor tool in Repbase (Kohany et al. 2006). Although with the current data presented it is not possible to ascertain what caused genomic Pif1 helicases to be lost in Brassicales and commelinids, the presence of long TE sequences in the vicinity of those genes in the closest taxonomic groups could be related to these events. For instance, TEs flanking these Pif1 genes could have promoted ectopic recombinations between insertions, leading to the deletion of large chromosome segments in Pif1 gene loci (Kent et al. 2017). However, more extensive analyses would be necessary to pinpoint the precise boundaries of these deleted chromosomal segments and to describe the mechanisms responsible for those events. Nonetheless, our results indicate that at least two major groups of land plants appear to have lost genomic Pif1 homologs independently (fig. 5) and that usual functions performed by this gene might be carried out by different proteins in species from these taxa.

Discussion

The Evolutionary History of *Helitrons* Takes Shape

Because Pif1 helicases are known to be typically eukaryotic proteins (Bochman et al. 2010), and Hel domains found in some RepHel transposases have introns, it has been suggested that an *Helitron* ancestor likely captured a Pif1 gene from its eukaryotic host (Kapitonov and Jurka 2001, 2007; Thomas and Pritham 2015). However, our results indicate that *Helitrons* already encoded a Hel domain before invading eukaryotic genomes (fig. 6), as genomic Pif1 helicases from prokaryotes and eukaryotes formed sister groups in our analyses, with Pif1-like Hel domains being only distantly related to

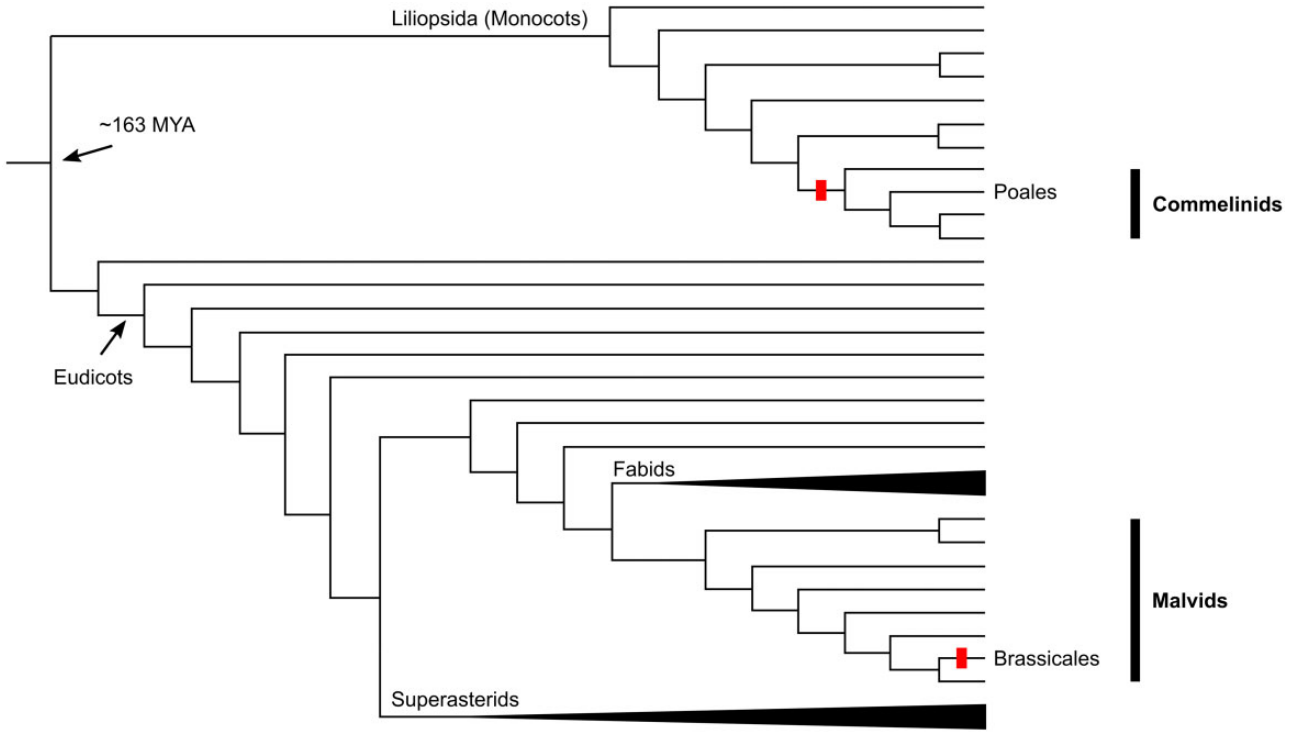

Fig. 5. Cladogram of plant groups that appear to have lost genomic Pif1 helicases. Only major clades are represented, with Poales and Brassicales indicating the orders of *O. sativa* and *A. thaliana*, respectively. Red bars mark the two branches that lack sequences with significant similarity to genomic Pif1 helicases. Phylogeny adapted from Li et al. (2019).

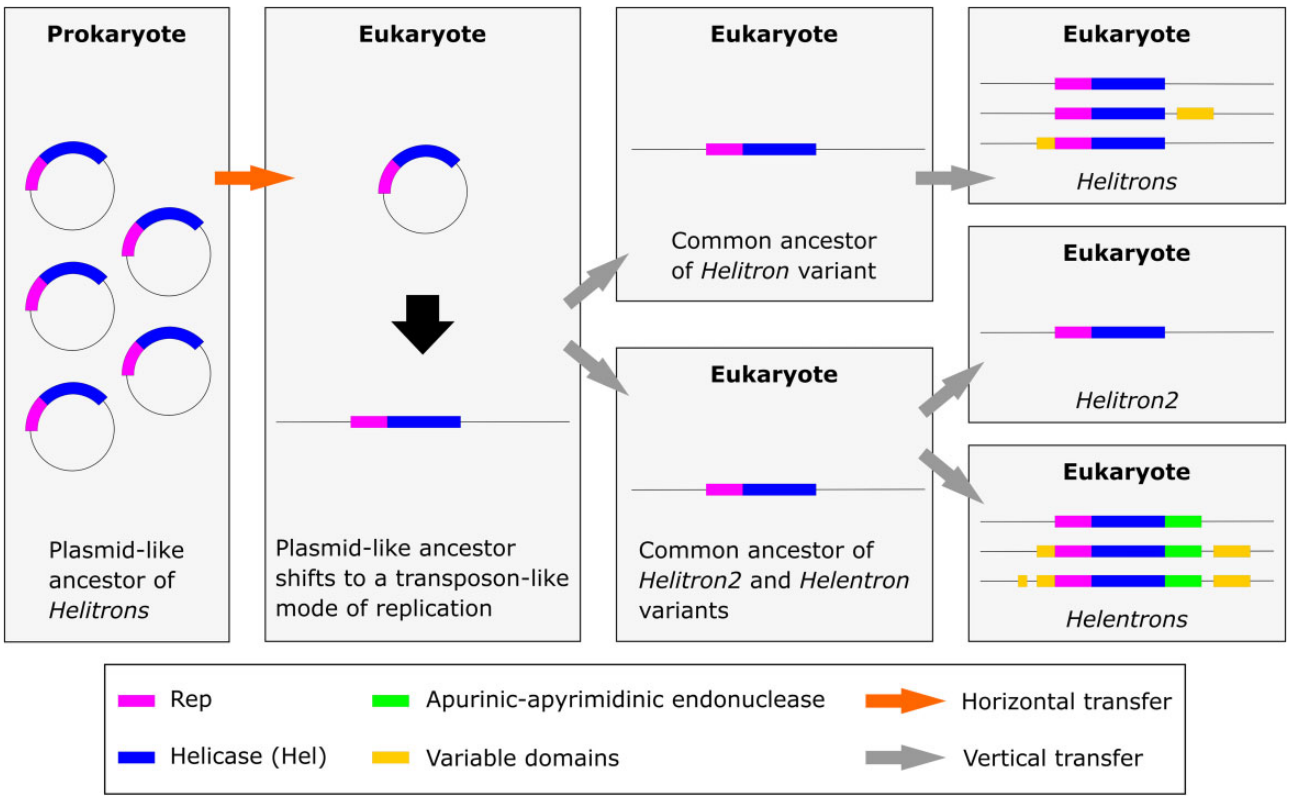

Fig. 6. A hypothesis for the evolution of *Helitrons*. We propose that *Helitrons* descend from prokaryotic plasmid-like elements (first box) that invaded eukaryotic cells during their early evolution. After invading eukaryotes, *Helitrons* shifted to a predominantly transposon-like mode of propagation. During their subsequent adaptation to specific hosts, *Helitrons* diverged into distinct variants (*Helitrons*, *Helentrons*, and *Helitron2*) and captured additional domains. Arrows represent major steps during the evolution of *Helitrons*.

them. Nonetheless, in addition to a RepHel with its archetypal double-domain structure, *Helentrons* also have an AP endonuclease domain in their transposase (fig. 1), which was probably captured from a non-LTR retrotransposon residing in the same eukaryotic host (Thomas and Pritham 2015). The capture of an AP endonuclease gene likely marked the evolutionary origin of *Helentrons* from *Helitron2*-like ancestors, which also gave rise to the *Helitron*2 variant. Our identification of an intermediate Hel domain from *S. microadriaticum* branching externally to *Helentron* and *Helitron2* sequences constitute the first direct evidence for a *Helitron2*-like element as the ancestor of both variants. Besides the AP endonuclease from *Helentrons*, several other domains have been incorporated to specific *Helitron* lineages during their evolution in eukaryotic genomes (Thomas and Pritham 2015) (fig. 6). However, the function of AP endonucleases and other coding sequences captured by *Helitrons* from eukaryotes have not been determined yet.

Although the evolutionary proximity of *Helentron* and *Helitron2* lineages was expected (Thomas and Pritham 2015; Heringer and Kuhn 2018), our results indicating that Hel domains from the *Helitron* variant form a distinct group from the *Helentron* and *Helitron2* variants (figs. 3 and 4) contrasts with the monophyletic distribution previously observed for *Helitron* Rep domains (Poulter et al. 2003; Heringer and Kuhn 2018). Assuming the more parsimonious scenario in which *Helitrons* constitute a monophyletic group, the resulting paraphyletic distribution of Hel domains might have been caused by faster evolutionary rates that occurred on this protein region. The same topology was not observed for Rep domains in previous studies, probably due to a higher tendency for amino acid sequence conservation in this portion of *Helitron* transposases. If Hel domains evolved under less constrained evolutionary pressures or went through a stronger positive selection imposed by their hosts, these processes could have potentially masked their monophyletic nature. Furthermore, the widespread distribution of *Helitrons* in eukaryotes (Thomas and Pritham 2015) and the overall similarity between RepHel and host phylogenies, indicate that *Helitrons* began to diverge before the emergence of most eukaryotic kingdoms (Poulter et al. 2003). As time estimates of major eukaryote radiations date back to approximately 1 billion years ago (Douzery et al. 2004; Berney and Pawlowski 2006), the first *Helitron* lineage divisions likely have a similar age. Thus, a rapid evolution of Hel domains that occurred through a very long period of time might have contributed to blur the monophyletic nature of *Helitrons* in our analyses.

An independent example supporting the hypothesis that each domain from RepHel proteins have evolved under distinct evolutionary pressures can be viewed in the phylogenies of *Helitron* Rep and Hel domains inferred by Poulter et al. (2003), which present distinct topologies. In their Rep domain phylogeny, *Helitron* sequences from the fungus *Phanerochaete chrysosporium* clustered with *Helentrons*, instead of *Helitrons*. Conversely, in the Hel domain phylogeny, all elements segregated into variant-specific clades, indicating that distinct *Helitron* variants display a more pronounced sequence divergence in this region. Furthermore, in the Hel phylogeny, *Helitron* clades were connected by relatively longer branches when compared with the Rep domain tree, similarly to the observed between our results presented here for Hel domains (supplementary fig. S2, Supplementary Material online) and on our previous study involving Rep domains (Heringer and Kuhn 2018). It is worth mentioning that, in contrast to our phylogeny, the one presented by Poulter et al. (2003) did not display a polyphyletic distribution for Hel domains. The reason for that might be related to the smaller sample size and diversity of *Helitrons* used in the latter analysis when compared with the one presented here.

Altogether, these observations suggest that each domain from RepHel transposases has evolved under distinct evolutionary rates. These differences could be derived from selective pressures that constrained the Rep amino acid sequence to a higher degree, and/or favored a more rapid evolution of the Hel domain to optimize its interaction with host components. Hence, a very early radiation of *Helitrons*, combined with relatively faster evolutionary rates that have occurred in Hel domains since they first invaded eukaryotes, probably explain the spurious paraphyletic distribution between major *Helitron* groups in our results. In this case, the observed topology could represent a result of long-branch attraction (Bergsten 2005).

In summary, our phylogenetic and NMDS analyses indicate that RepHel proteins evolved independently from genomic Pif1 helicases found in prokaryotes and eukaryotes. Thus, in spite of previous hypotheses about the origins of Hel domains, it is unlikely that a *Helitron* ancestor captured a Pif1 gene from its eukaryotic host. Instead, we suggest that, before entering eukaryotic cells, *Helitrons* already encoded RepHel proteins, branching into two major lineages after they invaded eukaryotic genomes (fig. 6). From there on, Hel domains probably evolved under relatively faster rates, which could explain their distribution into marked separate groups, in contrast to what was observed in analyses of Rep domains (Poulter et al. 2003; Heringer and Kuhn 2018).

Helitrons May Be Descendants of Plasmid-Like Elements

Although it seems clear that neither Rep nor Hel domains have originated from genomic proteins, the ancestor of *Helitrons* probably resided within a prokaryotic cell. If this ancestor already had a transposon-like mode of propagation, it is conceivable that their descendants (or their remnants) could still reside in genomes of some unknown prokaryote lineages. However, even assuming the hypothesis of a transposon ancestor as correct, it is unlikely that such elements would be found, as sequences that do not benefit cellular functioning directly (like TEs) are subject to extremely rapid turnover rates in prokaryotes (Sela et al. 2016; Wolf et al. 2016). A second possibility is that prokaryotic ancestors of *Helitrons* had a predominantly plasmid-like mode of replication before they became eukaryotic TEs. This scenario not only agrees with the current lack of *Helitron*-like sequences in prokaryotes, but with the close relationship found between Rep domains from *Helitrons* and RC bacterial plasmids (Heringer and Kuhn 2018; Kazlauskas et al. 2019) and the

fact that *Helitrons* generate plasmid-like intermediates during transposition (Grabundzija et al. 2018).

It is worth mentioning that a TraA relaxase was the only protein from a MGE retrieved in our Blast searches using Hel domains as queries. Similarly to RepHel transposases, TraA and other plasmid relaxases possess Rep-like and helicase domains within the same protein (Pérez-Mendoza et al. 2006; Chandler et al. 2013). Although Rep-like domains found in relaxases display an inverted orientation of their main catalytic motifs when compared with RepHel transposases, both enzymes have an overall similar architecture, consisting of a Rep followed by a helicase domain. In addition, despite their inverted orientation, the 3D topology of these motifs in relaxases and RCR proteins is essentially the same (Chandler et al. 2013). Interestingly, the cryo-EM structure of the RepHel in complex with the *Helitron* 5′-end ssDNA was solved only recently, revealing an even higher degree of organizational similarity with relaxases, particularly with TraI (Kosek et al. 2021). As mentioned by the authors, the structural similarity between these two classes of proteins does not imply a close evolutionary relationship, which is also supported by our results and previous studies involving the Rep domain (Heringer and Kuhn 2018; Kazlauskas et al. 2019). If these structural resemblances are most likely the result of convergent evolution, they would suggest the existence of functional parallels between relaxases and RepHel transposases. Nonetheless, the fact that a group of relaxases was retrieved in our searches by sequence similarity with Hel domains from *Helitrons* could still indicate a distant evolutionary relationship between these proteins.

Based on these considerations, we propose that *Helitrons* descend from prokaryotic plasmid-like elements that shifted to a transposon mode of propagation after invading eukaryotic cells (fig. 6). Importantly, a transition from an RCR plasmid to an RC TE would likely not require major adaptations, as the replicative processes employed in both types of MGEs work by the same basic enzymatic steps, only differing in the number of DNA substrates and type of final products involved (Chandler et al. 2013; Wawrzyniak et al. 2017).

What Is the Function of Pif1 Helicases in *Helitrons*?

Experimental assays revealed that *Helitrons* have to generate dsDNA circle intermediates in order to transpose, as ssDNA circular elements transfected into human cells were not viable substrates for host genome integration (Grabundzija et al. 2018). The formation of dsDNA intermediates could be achieved by the concomitant synthesis of leading and lagging strands while the element's leading strand is being "peeled-off," or by the addition of a short lagging strand primer on the unwound leading strand before an ssDNA circle is formed. In either case, these processes would require the recruitment of replication fork and DNA repair machinery components (Grabundzija et al. 2018), both of which Pif1 helicases are part of Bochman et al. (2010) and Muellner and Schmidt (2020). For instance, Pif1 stimulates the activity of DNA polymerase δ (Pol δ) during DNA repair and replication (Pike et al. 2009; Wilson et al. 2013; Koc et al. 2016) through its interaction with the proliferating cell nuclear antigen (PCNA) (Wilson et al. 2013; Buzovetsky et al. 2017; Dahan et al. 2018). In addition, Pif1 has a role in fork convergence, resolving the stalling of these structures, which are expected to occur in the final stages of linear and circular DNA replication (Deegan et al. 2019). Another relevant feature of Pif1 helicases is their preference for binding and unwinding forked structures (dsDNA with ssDNA overhangs) (Ramanagoudr-Bhojappa et al. 2013; Li et al. 2016), which are substrates expected to be formed in the first stages of RCT, when RepHel nicks the *Helitron*'s leading strand in its 5′-end (Dias et al. 2016; Grabundzija et al. 2016, 2018).

The combination of those Pif1 attributes suggests that the Hel domain could aid in the RepHel association to forked DNA structures during the initial steps of transposition and help to recruit replication machinery components from hosts (e.g., PCNA and Pol δ). Although prokaryotic RC TEs, which are thought to transpose similarly to *Helitrons*, do not encode helicases, it is possible that a Hel domain merged to a Rep protein confers mechanistic advantages for RCT in eukaryotic cells and maybe is essential in this environment. Indeed, it has been shown that a mutation in the Walker A motif from Hel domains causes *Helitrons* to lose their transposition activity in cells (Grabundzija et al. 2016). In addition, the RepHel cryo-EM structure reveals a considerable interface between the catalytic portion of Rep and the Hel domain, suggesting that they act in conjunction to unwind dsDNA and generate sufficient ssDNA to allow strand cleavage as transposition starts (Kosek et al. 2021). Thus, it is conceivable that a Hel domain also favored the invasion and colonization of eukaryotic genomes by *Helitrons*, which would explain their pervasiveness in this domain of life that lacks other groups of RC TEs.

Additionally, the Hel domain could facilitate the final stages of transposition, when the RepHel associated with a circular intermediate binds its target site before integration. In contrast to prokaryotic RC TE insertions, which are guided by site specificity (Garcillán-Barcia et al. 2002), *Helitrons* integrate between AT, TT, or TC dinucleotides, depending on the variant, with no preference for unique sequences (Thomas and Pritham 2015). Hence, the RepHel in complex with a *Helitron* intermediate could initially bind its target site by associating with specific DNA or chromatin structures, instead of using sequence guided recognition. In this case, an initial contact would be favored by the known affinity of Pif1 helicases to DNA secondary structures typically found in recombination sites and gene promoters (Bochman et al. 2012; Byrd and Raney 2015; Muellner and Schmidt 2020). Indeed, experimental assays revealed that active *Helitrons* appear to preferentially target highly expressed gene regions (Grabundzija et al. 2016). After a structure-based association mediated also by Hel, the Rep domain would be able nick the recipient strand at a nearby AT, TT or TC dinucleotide site, before transferring an ssDNA intermediate to the host's chromosome, forming a heteroduplex and completing transposition (Kapitonov and Jurka 2007; Thomas and Pritham 2015; Dias et al. 2016).

Taken together, these features of Pif1 helicases and *Helitrons* appear to agree with a scenario in which Hel domains play a more sophisticated role during RCT, beyond simply unwinding double-stranded DNA elements. The

presence of a Pif1-like Hel domain in *Helitron* transposases may have provided an advantage over the recruitment of host helicases, by concatenating the processes of DNA binding, leading strand nicking, and peeling-off, together with the formation of circular dsDNA intermediates, all conducted by the same enzyme. In addition, Hel domains could aid the association between RepHel–dsDNA intermediates and target sites on host chromosomes.

Helitrons Can Hamper the Identification of Eukaryotic Pif1 Helicases

The abundance of *Helitrons* in eukaryotic genomes, together with the general similarities between *Helitron* Pif1-like Hel domains and genomic Pif1 helicases from eukaryotes, make their distinction by in silico methods complicated. Our reevaluation of three examples in the literature describing Pif1 proteins from *A. thaliana*, *O. sativa*, and *M. robertsii* demonstrated how these problems have affected the classification and number estimation of genomic Pif1 helicases in eukaryotic species. In these cases, most, or all putative genomic Pif1 helicases described were shown to represent *Helitron*-derived sequences.

Interestingly, during our searches for genomic Pif1 candidates in *A. thaliana* and *O. sativa* we found that all Pif1-like proteins from these species either represent complete *Helitron* transposase sequences or Hel domains from broken RepHel ORFs. After investigating higher taxonomic ranks from which *A. thaliana* and *O. sativa* belong (Brassicales and commelinids, respectively), we found that both of them appear to have lost genomic Pif1 homologs independently (fig. 5). Even granting that Brassicales and commelinids may have genomic Pif1 homologs that went undetected in our searches, the fact that RepHel sequences represented the best hits to eukaryotic Pif1 helicases points to a similar evolutionary pattern in those distantly related groups. However, this issue should be further investigated to determine in more detail how the Pif1 family have evolved in land plants and if some of them have different proteins to perform the same functions of genomic Pif1 helicases.

Despite the examples described above, some eukaryotes have multiple bona fide genomic Pif1 helicases. As we have mentioned, kinetoplastids encode several Pif1 paralogs that likely participate in distinct functions related to their unique biology (Liu et al. 2009; Bochman et al. 2010). Furthermore, *Helitron* transposases are not found in kinetoplastid genomes, as indicated by our Blast searches and a previous analysis (Thomas and Pritham 2015). Hence, all Pif1 helicases found in this group might consist of genomic representatives derived from gene duplications. In addition to kinetoplastids, some amoebae also have multiple genomic Pif1 helicases, with *Acanthamoeba castellanii* encoding up to nine Pif1 genes (Harman and Manna 2016). Our Blast searches revealed that these amoebae species do not have RepHel sequences in their genomes, which confirms that these proteins indeed represent genomic Pif1 helicases. Thus, kinetoplastids and amoebae are the only eukaryotic groups so far in which there is solid evidence for species with more than two genomic Pif1 paralogs.

Altogether, it is clear that our knowledge about the distribution and number of genomic Pif1 helicases in eukaryotes is relatively limited to a small number of species. As we have shown, some of the attempts to identify genomic Pif1 proteins in eukaryotes have been hampered by the large amount of *Helitron* transposases found in this domain of life. It will be important to establish a reliable and efficient method to correctly discriminate between these two major groups of Pif1 helicases, before they are studied in large-scale analyses.

Conclusion

Although the similarity between Hel domains and genomic Pif1 helicases has been noted since the discovery of *Helitrons* 20 years ago, no study had explored their evolutionary connections. Despite previous suggestions that an *Helitron* ancestor likely acquired the Hel domain by capturing a Pif1 gene from its eukaryotic host, our results indicate that RepHel proteins already had their archetypal structure with two domains before invading eukaryotes. Furthermore, considering phylogenetic, structural, and mechanistic aspects of these elements, we propose that *Helitron* ancestors probably had a plasmid-like mode of replication in prokaryotic hosts, before invading eukaryotes and shifting into a transposon. Based on the known features of Pif1 helicases and RepHel proteins, we also hypothesize that Hel domains likely perform a more complex function during transposition, beyond simply unwinding *Helitron* double-stranded DNA.

In addition, our reassessment of the literature describing eukaryotic Pif1 helicases revealed that many of these examples actually represent complete or partial RepHel transposases from *Helitrons*, which are commonly abundant in eukaryotic genomes. This finding highlights the need for a careful inspection before classifying Pif1-like proteins as genomic helicases in eukaryotes, particularly in species that appear to harbor multiple Pif1-like genes. We also found that two distantly related groups of land plants appear to lack genomic Pif1 homologs, despite having multiple Pif1-like Hel domain sequences derived from *Helitrons*. This observation should be studied in more detail, as Pif1 helicases have been considered essential in many genomic processes that are conserved in all eukaryotes studied to date.

Materials and Methods

Selection of RepHel Sequences

We used RepHel protein sequences obtained in our previous study (Heringer and Kuhn 2018), belonging to the three main *Helitron* variants (*Helitron*, *Helentron*, or *Helitron2*) (Thomas and Pritham 2015), as initial queries in a series of Blastp searches on the nonredundant protein sequences (nr) database from GenBank (Sayers et al. 2019). With this strategy, we were able to retrieve a sample with a larger variety of RepHel representatives, thus enabling the generation of more accurate consensus sequences of each domain (Rep and Hel). Each one of the initial 13 *Helitron* protein sequences was used as a query to select an additional RepHel, which in turn, was used as a query to select another sequence in a second Blastp search round. In each of these searches the best hit, sorted

by Max Score, was selected, excluding sequences found in genomes of the same genus in a previous round. For the *Helitron2* variant we applied four rounds of consecutive searches to increase the number of sequences, as it had a single representative in our previous analysis (Heringer and Kuhn 2018). To determine whether the additional RepHel sequences belonged to the same variant as the initial queries, we visually inspected their structure with the Conserved Domain Database (CDD) search tool (Lu et al. 2020), following the classification provided by Thomas and Pritham (2015). This classification considers differences in amino acids within conserved regions from the Rep domain and the presence or absence of specific domains in the RepHel protein. A total of 41 RepHel protein sequences were selected for further analyses: 18 from *Helitrons*, 18 from *Helentrons*, and 5 from *Helitron2* elements. Sequences from *Helitron* and *Helentron*/*Helitron2* variants were aligned separately using the auto mode from the MAFFT online service (Katoh et al. 2019). *Helentron* and *Helitron2* sequences were aligned as a single group because these variants are known to be closely related (Thomas and Pritham 2015; Heringer and Kuhn 2018). Rep and Hel domains from each protein were isolated and trimmed, keeping only well-defined conserved regions among aligned sequences. These conserved regions were used to generate consensus sequences of each domain from *Helitron* and *Helentron*/*Helitron2* variants, considering the most common amino acid in each site (supplementary data S1, Supplementary Material online), using the Advanced Consensus Maker tool from the HIV Database (https://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html; last accessed November 16, 2021).

Stepwise Search and Selection of Helicase Protein Sequences

The Hel domain consensus sequences of *Helitron* and *Helentron*/*Helitron2* variants (supplementary data S3, Supplementary Material online) were used as queries in Blastp searches against the nr database from GenBank (Sayers et al. 2019), which includes all available annotated proteins for a given taxa. A sample of protein sequences representing a wide variety of organisms were retrieved from distinct taxonomic levels, depending on their number of resulting hits in preliminary Blastp searches. For example, in eukaryotes, searches were conducted from the kingdom down to the class level, as this domain displayed a large number of significant results distributed heterogeneously across thousands of genomes. Conversely, in bacteria we conducted searches at the phylum level, and in archaea the whole sample was retrieved at the domain level itself. The best hits (sorted by Max Score) from Blastp searches using consensus sequences of both *Helitron* and *Helentron*/*Helitron2* variants were selected. Each species containing best hits had one or two protein sequence representatives selected, depending on whether searches using different variant consensuses retrieved the same or different best hits, respectively. To verify if *Helitrons* were present in the genomes of species containing selected hits, we carried out a second round of searches in these taxa, this time using Rep consensus sequences as queries. Blastp searches were conducted against the nr database and tBlastn searches were conducted against the WGS contigs database. Because the aim of our study was to investigate the relationship between Hel domains from *Helitrons* and genomic Pif1 helicases, taxa containing hits corresponding to Rep sequences in any of the two searches (Blastp or tBlastn) were excluded at this stage. By doing so, we expected to have avoided the inclusion of helicases derived from *Helitrons* during the retrieval of putative genomic helicases, which could result in false phylogenetic inferences. Using these criteria, we were able to select 76 Pif1-like sequences from a wide variety of organisms lacking Rep sequences in their genomes. To expand our sample, we used Hel domain consensus sequences and the *S. cerevisiae* Pif1 (NP_013650.1) as queries in Blastp searches against the same groups of organisms from the previous analysis, this time without filtering taxa with Rep sequences in their genomes and including eukaryotic and prokaryotic viruses. Because Pif1-like proteins selected in the initial searches could be more readily identified as either genomic or *Helitron*-derived helicases, they were used to aid in the classification of sequences retrieved without the Rep-filtering procedure by their relationship revealed later in the phylogenetic analysis.

Alignment and Isolation of Helicase Domains

Helicase sequences from each major taxon group (Eukaryota, Bacteria, Archaea, plasmids, eukaryotic, and prokaryotic viruses) were aligned separately with the Hel domain consensus sequences from *Helitrons* and *Helentrons*/*Helitron2* using the auto mode from the MAFFT online service (Katoh et al. 2019) in order to identify a common region among them. Sequences that aligned poorly or displayed large gaps on conserved regions were excluded using the MAFFT data set refinement tool also available in the MAFFT online service (Katoh et al. 2019). Segments extending upstream and downstream the central conserved regions were visualized using MEGAX (Kumar et al. 2018) and trimmed to avoid spurious alignments between nonrelated portions of proteins. This procedure is important considering that a large majority of prokaryotic and eukaryotic proteins contain multiple domains that have evolved through modular rearrangements (Bornberg-Bauer et al. 2005; Wang and Caetano-Anollés 2009). Even among genomic Pif1-like domains from eukaryotes, there are low levels of sequence and size similarity in their N- and C-terminal regions extending beyond a conserved core (Boule and Zakian 2006). Thus, when conducting a phylogenetic analysis of highly divergent protein sequences, it is preferable to only consider limited domain regions as evolutionary units, because flanking segments can evolve through distinct selective constraints. A total of 310 helicases from *Helitrons* (65 sequences), eukaryotic (89 sequences) and prokaryotic organisms (56 sequences), plasmids (10 sequences), eukaryotic viruses (48 sequences), and prokaryotic viruses (42 sequences) were selected for the next step of our analyses (supplementary table S1, Supplementary Material online). Trimmed helicase domains from all taxa, including *Helitrons*, were aligned using the E-INS-i method combined with mafft-homologs in the MAFFT online service (Katoh

et al. 2019). The final alignment containing all sequences used in the following analyses are available in supplementary data S2, Supplementary Material online.

Phylogenetic and NMDS Analyses

The best-fit evolutionary model for the alignment (LG + G + I) was selected using the smart model selection in PhyML (Lefort et al. 2017). The maximum likelihood phylogeny of aligned amino acid sequences was inferred with the SPR method of tree topology search, six random plus one parsimony starting trees and six substitution rate categories across sites modeled with estimated gamma-shaped distribution parameter and proportion of invariant sites. Branch supports were estimated using the approximate likelihood ratio test (aLRT) with the nonparametric Shimodaira–Hasegawa correction (SH-aLRT). All these procedures were conducted on PhyML 3.1 (Guindon et al. 2010). Branches with <0.7 SH-aLRT statistical support were collapsed using TreeGraph 2 (Stöver and Müller 2010) and the final tree visualized using FigTree v.1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/; last accessed November 16, 2021). For the NMDS analysis, pairwise evolutionary distances between aligned sequences were estimated with the JTT matrix-based model and the rate variation among sites modeled with a gamma distribution on MEGAX (Kumar et al. 2018). NMDS ordinations with Euclidean distances of the sequences represented in two dimensions were generated using the R package vegan v2.5-6 (Dixon 2003). The NMDS analysis and plotting were executed in RStudio v1.3.959 (RStudio Team 2020) with R v4.0.0 (R Core Team 2020). All the methodology described heretofore is represented as a schematic workflow in figure 2.

Search and Classification of Pif1-Like Proteins in Eukaryotic Species

To reexamine selected examples from the literature describing genomic Pif1 helicases, which could in fact constitute RepHel-derived sequences, we inspected the structure of those proteins using the CDD search tool (Lu et al. 2020). To reassess the description of species containing multiple genomic Pif1 helicases we conducted Blastp searches in the protein sequences from the corresponding taxa available in the nr database from GenBank (Sayers et al. 2019) using the human Pif1 domain (6HPH_A) and *S. cerevisiae* Pif1 protein (NP_013650.1) as queries. In order to verify if the resulting sequences corresponded to RepHel transposases, all hits had their structural features inspected with the CDD search tool (Lu et al. 2020). Hits that did not included a conserved Rep domain identified by the CDD search tool were used as queries in a second round of Blastp searches against the nr database from GenBank to check if they might constitute Hel domains from broken *Helitron* transposases (Hel domains highly similar to RepHel proteins) or cryptic RepHel proteins (truncated transposase with a Rep sequence upstream the Pif1 ORF). If the best hits (sorted by Max Score) from this second round of searches corresponded to RepHel proteins, queries were considered as derived from *Helitrons*. In contrast, if the resulting best hits did not correspond to RepHel sequences, queries were classified as putative genomic Pif1 helicases.

Supplementary Material

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

Acknowledgments

We would like to thank Dr Guilherme B. Dias for the helpful comments on an earlier version of the manuscript and three anonymous reviewers whose comments and suggestions helped to improve several aspects of the paper. We are also grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (fellowship 308386/2018-3 to G.C.S.K.) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) (doctoral fellowship to P.H.).

Data Availability

The data underlying this article are available in the article and in its supplementary material.

References

Alt-Mörbe J, Stryker JL, Fuqua C, Li PL, Farrand SK, Winans SC. 1996. The conjugal transfer system of *Agrobacterium tumefaciens* octopine-type Ti plasmids is closely related to the transfer system of an IncP plasmid and distantly related to Ti plasmid vir genes. *J Bacteriol*. 178(14):4248–4257.

Barreat JGN, Katzourakis A. 2021. Paleovirology of the DNA viruses of eukaryotes. *Trends Microbiol*. https://doi.org/10.1016/j.tim.2021.07.004.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21(2):163–193.

Berney C, Pawlowski J. 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc R Soc B*. 273(1596):1867–1872.

Bochman ML, Sabouri N, Zakian VA. 2010. Unwinding the functions of the Pif1 family helicases. *DNA Repair* 9(3):237–249.

Bochman ML, Judge CP, Zakian VA. 2011. The Pif1 family in prokaryotes: what are our helicases doing in your bacteria? *Mol Biol Cell*. 22(12):1955–1959.

Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet*. 13(11):770–780.

Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*. 62(4):435–445.

Boule JB, Zakian VA. 2006. Roles of Pif1-like helicases in the maintenance of genomic stability. *Nucleic Acids Res*. 34(15):4147–4153.

Buzovetsky O, Kwon Y, Pham NT, Kim C, Ira G, Sung P, Xiong Y. 2017. Role of the Pif1-PCNA complex in Pol δ-dependent strand displacement DNA synthesis and break-induced replication. *Cell Rep*. 21(7):1707–1714.

Byrd AK, Raney KD. 2015. A parallel quadruplex DNA is bound tightly but unfolded slowly by Pif1 helicase. *J Biol Chem*. 290(10):6482–6494.

Cavalier-Smith T, Chao EE, Snell EA, Berney C, Fiore-Donno AM, Lewis R. 2014. Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Mol Phylogenet Evol*. 81:71–85.

Chandler M, De La Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol*. 11(8):525–538.

Dahan D, Tsirkas I, Dovrat D, Sparks MA, Singh SP, Galletto R, Aharoni A. 2018. Pif1 is essential for efficient replisome progression through

lagging strand G-quadruplex DNA secondary structures. *Nucleic Acids Res*. 46(22):11847–11857.
Deegan TD, Baxter J, Bazán MÁO, Yeeles JT, Labib KP. 2019. Pif1-family helicases support fork convergence during DNA replication termination in eukaryotes. *Mol Cell*. 74(2):231–244.
Dias GB, Heringer P, Kuhn GCS. 2016. Helitrons in *Drosophila*: chromatin modulation and tandem insertions. *Mob Genet Elements* 6(2):e1154638
Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci*. 14(6):927–930.
Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A*. 101(43):15386–15391.
Edger PP, Hall JC, Harkess A, Tang M, Coombs J, Mohammadin S, Schranz ME, Xiong Z, Leebens-Mack J, Meyers BC, et al. 2018. Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *Am J Bot*. 105(3):463–469.
Fan L, Wu D, Goremykin V, Xiao J, Xu Y, Garg S, Zhang C, Martin WF, Zhu R. 2020. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat Ecol Evol*. 4(9):1213–1219.
Feschotte C, Wessler SR. 2001. Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proc Natl Acad Sci U S A*. 98(16):8923–8924.
Garcillán-Barcia MP, Bernales I, Mendiola MV, De La Cruz F. 2002. IS91 rolling-circle transposition. In: Craig N, Craigie R, Gellert M, Lambowitz A, editors. Mobile DNA II. Washington, DC: ASM Press. p. 891–904.
Grabundzija I, Messing SA, Thomas J, Cosby RL, 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.
Grabundzija I, Hickman AB, Dyda F. 2018. Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. *Nat Commun*. 9(1):1278.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.
Harman A, Manna S. 2016. Identification of Pif1 helicases with novel accessory domains in various amoebae. *Mol Phylogenet Evol*. 103:64–74.
Heringer P, Kuhn GCS. 2018. Exploring the remote ties between Helitron transposases and other rolling-circle replication proteins. *Int J Mol Sci*. 19(10):3079.
Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. 16(2):67–79.
Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*. 98(15):8714–8719.
Kapitonov VV, Jurka J. 2007. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet*. 23(10):521–529.
Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 20(4):1160–1166.
Kazlauskas D, Varsani A, Koonin EV, Krupovic M. 2019. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun*. 10(1):3425.
Kent TV, Uzunović J, Wright SI. 2017. Coevolution between transposable elements and recombination. *Phil Trans R Soc B*. 372(1736):20160458.
Knoll A, Puchta H. 2011. The role of DNA helicases and their interaction partners in genome stability and meiotic recombination in plants. *J Exp Bot*. 62(5):1565–1579.
Koc KN, Singh SP, Stodola JL, Burgers PM, Galletto R. 2016. Pif1 removes a Rap1-dependent barrier to the strand displacement activity of DNA polymerase δ. *Nucleic Acids Res*. 44(8):3811–3819.
Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: repbaseSubmitter and Censor. *BMC Bioinf*. 7:474.
Koonin EV. 2016. Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res*. 5:1805.
Kosek D, Grabundzija I, Lei H, Bilic I, Wang H, Jin Y, Peaslee GF, Hickman AB, Dyda F. 2021. The large bat Helitron DNA transposase forms a compact monomeric assembly that buries and protects its covalently bound 5′-transposon end. *Mol Cell*. 81(20):4271–4286.
Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 35(6):1547–1549.
Lefort V, Longueville JE, Gascuel O. 2017. SMS: smart model selection in PhyML. *Mol Biol Evol*. 34(9):2422–2424. Available from: http://www.atgc-montpellier.fr/phyml/. Accessed November 16, 2021.
Li J-H, Lin W-X, Zhang B, Nong D-G, Ju H-P, Ma J-B, Xu C-H, Ye F-F, Xi XG, Li M, et al. 2016. Pif1 is a force-regulated helicase. *Nucleic Acids Res*. 44(9):4330–4339.
Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y, et al. 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat Plants* 5(5):461–470.
Liu B, Wang J, Yaffe N, Lindsay ME, Zhao Z, Zick A, Shlomai J, Englund PT. 2009. Trypanosomes have six mitochondrial DNA helicases with one controlling kinetoplast maxicircle replication. *Mol Cell* 35(4):490–501.
Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 48(D1):D265–D268. Available from: https://www.ncbi.nlm.nih.gov/Structure/cdd/. Accessed November 16, 2021.
Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJ. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* 557(7703):101–105.
Muellner J, Schmidt KH. 2020. Yeast genome maintenance by the multifunctional PIF1 DNA helicase family. *Genes* 11(2):224.
Pérez-Mendoza D, Lucas M, Munoz S, Herrera-Cervera JA, Olivares J, de la Cruz F, Sanjuán J. 2006. The relaxase of the *Rhizobium etli* symbiotic plasmid shows nic site cis-acting preference. *J Bacteriol*. 188(21):7488–7499.
Pike JE, Burgers PM, Campbell JL, Bambara RA. 2009. Pif1 helicase lengthens some Okazaki fragment flaps necessitating Dna2 nuclease/helicase action in the two-nuclease processing pathway. *J Biol Chem*. 284(37):25170–25180.
Poulter RT, Goodwin TJ, Butler MI. 2003. Vertebrate helentrons and other novel Helitrons. *Gene*. 313:201–212.
R Core Team. 2020. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.R-project.org/. Accessed November 16, 2021.
Ramanagoudr-Bhojappa R, Chib S, Byrd AK, Aarattuthodiyil S, Pandey M, Patel SS, Raney KD. 2013. Yeast Pif1 helicase exhibits a one-base-pair stepping mechanism for unwinding duplex DNA. *J Biol Chem*. 288(22):16185–16195.
Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The origin and diversification of mitochondria. *Curr Biol*. 27(21):R1177–R1192.
RStudio Team. 2020. RStudio: integrated development for R. RStudio. Boston: PBC. Available from: http://www.rstudio.com/. Accessed November 16, 2021.
Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. *Nucleic Acids Res*. 47(D1):D94–D99. Available from: https://www.ncbi.nlm.nih.gov/genbank/. Accessed November 16, 2021.
Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci U S A*. 113(41):11399–11407.
Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinf*. 11:7.
Thomas J, Pritham EJ. 2015. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol Spectr*. 3(4):MDNA3-0049-2014.
Thomas J, Vadnagara K, Pritham EJ. 2014. DINE-1, the highest copy number repeats in *Drosophila melanogaster* are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). *Mob Dna*. 5:18.

Van Etten J, Bhattacharya D. 2020. Horizontal gene transfer in eukaryotes: not if, but how much? *Trends Genet*. 36(12):915–925.

Wang M, Caetano-Anollés G. 2009. The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure* 17(1):66–78.

Wawrzyniak P, Płucienniczak G, Bartosik D. 2017. The different faces of rolling-circle replication and its multifunctional initiator proteins. *Front Microbiol*. 8:2353.

Wilson MA, Kwon Y, Xu Y, Chung WH, Chi P, Niu H, Mayle R, Chen X, Malkova A, Sung P, et al. 2013. Pif1 helicase and Polδ promote recombination-coupled DNA synthesis via bubble migration. *Nature* 502(7471):393–396.

Wolf YI, Makarova KS, Lobkovsky AE, Koonin EV. 2016. Two fundamentally different classes of microbial genes. *Nat Microbiol*. 2:16208.

Xiong W, He L, Lai J, Dooner HK, Du C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci U S A*. 111(28):10263–10268.

Yang L, Bennetzen JL. 2009. Structure-based discovery and description of plant and animal Helitrons. *Proc Natl Acad Sci U S A*. 106(31):12832–12837.