Spy: A New Group of Eukaryotic DNA Transposons without Target Site Duplications

Min-Jin Han[1], Hong-En Xu[2], Hua-Hao Zhang[1,3], Cédric Feschotte[4], and Ze Zhang[1,*]

[1]School of Life Sciences, Chongqing University, China

[2]Department of Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, TU Muenchen, Freising, Germany

[3]College of Pharmacy and Life Science, Jiujiang University, China

[4]Department of Human Genetics, University of Utah School of Medicine

*Corresponding author: E-mail: zezhang@cqu.edu.cn, ze_zhang@126.com.

Accepted: June 18, 2014

Abstract

Class 2 or DNA transposons populate the genomes of most eukaryotes and like other mobile genetic elements have a profound impact on genome evolution. Most DNA transposons belong to the cut-and-paste types, which are relatively simple elements characterized by terminal-inverted repeats (TIRs) flanking a single gene encoding a transposase. All eukaryotic cut-and-paste transposons so far described are also characterized by target site duplications (TSDs) of host DNA generated upon chromosomal insertion. Here, we report a new group of evolutionarily related DNA transposons called *Spy*, which also include TIRs and DDE motif-containing transposase but surprisingly do not create TSDs upon insertion. Instead, *Spy* transposons appear to transpose precisely between 5′-AAA and TTT-3′ host nucleotides, without duplication or modification of the AAATTT target sites. *Spy* transposons were identified in the genomes of diverse invertebrate species based on transposase homology searches and structure-based approaches. Phylogenetic analyses indicate that *Spy* transposases are distantly related to *IS5*, *ISL2EU,* and *PIF/Harbinger* transposases. However, *Spy* transposons are distinct from these and other DNA transposon superfamilies by their lack of TSD and their target site preference. Our findings expand the known diversity of DNA transposons and reveal a new group of eukaryotic DDE transposases with unusual catalytic properties.

Key words: DNA transposon, transposition, *Spy*, target site duplication.

Introduction

Transposable elements (TEs) are the largest component of most multicellular genomes. They account for 15–47% of insect genomes (Holt et al. 2002; Kapitonov and Jurka 2003; Nene et al. 2007; Xu et al. 2013), 35–69% of mammalian genomes (Lander et al. 2001; Waterston et al. 2002; de Koning et al. 2011), and up to 90% of some plant genomes (Feschotte, Jiang, et al. 2002; Kidwell 2002). TEs are divided into two classes: Class 1 elements (retrotransposons) transpose through reverse transcription of an RNA intermediate, whereas class 2 elements (DNA transposons) transpose through a DNA intermediate (Finnegan 1989; Feschotte et al. 2002).

DNA transposons have deep evolutionary origins and are found in almost all eukaryotic genomes (Feschotte and Pritham 2007). They are classified into two major subclasses (cut-and-paste elements and rolling-circle or *Helitron* elements), which are distinguished by their transposition mechanism (Kapitonov and Jurka 2001; Feschotte and Pritham 2007; Wicker et al. 2007). Transposition of cut-and-paste DNA transposons involved excision and reinsertion catalyzed by an element-encoded transposase, whereas *Helitrons* are thought to transpose by a form of copy-and-paste mechanism involving DNA strand displacement akin to rolling-circle transposition. Cut-and-paste transposons usually have terminal-inverted repeats (TIRs) flanked by target site duplications (TSDs). TSDs are generated by autonomous element-encoded transposases, which makes staggered cuts in the target DNA and filled by the host repair machinery to complete the transposon's integration (Craigie and Mizuuchi 1985; Craig et al. 2002). Thus, the length and/or sequence of the TSDs reflect the enzymatic cleavage properties of transposases and can be used to classify cut-and-paste transposons into different superfamilies. For instance, the *Tc1/Mariner* superfamily is characterized by 5′-TA-3′ TSD (Shao and Tu 2001), the

piggyBac superfamily by 5′-TTAA-3′ (Sarkar et al. 2003; Mitra et al. 2008), the *hAT* (hobo-Ac-Tam3) superfamily by TSD of 8 bp (with any or little sequence specificity) (Kempken and Windhofer 2001), 3 bp in the *PIF/Harbinger* superfamily (Zhang et al. 2001), whereas the *Mutator/MuDR superfamily* is associated with TSD ranging in size from 9 to 12 bp (Lisch 2002; Marquez and Pritham 2010). Currently, 19 eukaryotic superfamilies are recognized in Repbase, the most comprehensive repository of eukaryotic TEs (Jurka et al. 2005). All eukaryotic cut-and-paste superfamilies so far described appear to belong to the "megafamily" of DDE/D recombinases, owing to the conserved amino acid triad of their catalytic domain, and they are all associated with the formation of TSD upon transposon insertion (Yuan and Wessler 2011).

Here, we report on the discovery of a new cut-and-paste transposon called *Spy*, first identified in the silkworm and subsequently in a variety of invertebrate genomes through transposase and structural similarity searches. These elements possess TIRs and appear to encode a DDE motif-containing transposase. Surprisingly, however, they do not generate TSDs, but integration occurs precisely between host 5′-AAA-3′ and 5′-TTT-3′ nucleotides, without deletion and duplication of the target sequence.

Materials and Methods

Identification and Characterization of *Spy*

The original *Spy* transposons were discovered based on a comprehensive and semiautomated annotation of TEs in the silkworm genome (July 2013). Multiple alignments were performed using MUSCLE (http://www.ebi.ac.uk/Tools/muscle/index.html, last accessed June 30, 2014) with default parameters. Aligned sequences were manually refined using the BioEdit program (Hall 1999). To estimate the abundance of each silkworm *Spy* family, the consensus sequence of each *Spy* family was used as query in BLASTN ($e < 10^{-5}$) search against new assembly silkworm genome that was downloaded from SilkDB (http://silkworm.swu.edu.cn/silkdb, last accessed June 30, 2014). A copy for *Spy* family was defined by an *e* value less than e^{-5}, length larger than 50 bp, and a minimum nucleotide identity of 80%. None of the *Bombyx mori Spy* families were previously in Repbase (v.18.08) (Jurka et al. 2005) or in the NCBI nonredundant (nr) databases.

Transposase-coding sequences were predicted with GeneMark.hmm (http://exon.biology.gatech.edu/eukhmm.cgi, last accessed June 30, 2014), GENESCAN (http://genes.mit.edu/GENSCAN.html, last accessed June 30, 2014), or GetORF software (Rice et al. 2000). Transposase domains were predicted with CD-search at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, last accessed June 30, 2014). Secondary structures of representative transposases were predicted using PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/, last accessed June 30, 2014) (Bryson et al. 2005). Putative helix-turn-helix (HTH) motifs were predicted by NPS@ software (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hth.html, last accessed June 30, 2014) (Dodd and Egan 1990).

To investigate the distribution of *Spy* transposons in other species, the transposase of silkworm *Spy* transposons was used as query in TBLASTN searches against various GenBank nucleotide databases (nr, WGS, GSS, and EST). A hit was considered as a candidate element when the *e* value was lower than 10^{-4}. For extremely distant species, such as bacteria, hits with *e* value up to 0.01 were also considered as a candidate element. Reiterative PSI-BLAST searches were also performed. Candidate elements were inspected to verify *Spy* transposon features, including DDE domain, TIRs, and target sequences.

Phylogenetic Analysis

To investigate the evolutionary relationships among the *PIF/Harbinger*, *ISL2EU*, *IS5*, and *Spy* transposons, the transposase sequences of 17 *ISL2EU*, 24 *PIF/Harbinger*, and 13 *IS5* transposons were downloaded from Repbase and the insertion sequences (ISs) database (Kichenaradja et al. 2010) (https://www-is.biotoul.fr/, last accessed June 30, 2014). Multiple alignments of the predicted transposases were performed using MUSCLE with default parameters. Aligned sequences were manually refined using BioEdit (Hall 1999). Phylogenetic trees were constructed using the Bayesian approach using the DDE domains of the transposase multiple alignment. Bayesian inferences were performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) with the WAG model estimated using prottest-3.2 software (Darriba et al. 2011). We performed 3,000,000 generations and other parameters were set as default.

Target Site Verification through Identification of Paralogous and Orthologous Empty Sites

To search for orthologous empty sites, four loci from four silkworm *Spy* families in three domesticated silkworm strains (DaZao, BiBo, and HeiGao) were assayed by polymerase chain reaction (PCR) with primer pairs flanking each element followed by DNA sequencing. DNA was extracted from individual pupae and moths using a standard phenol–chloroform extraction approach (Nagaraja and Nagaraju 1995). The four pairs of primers used for these assays are listed in supplementary table S1, Supplementary Material online. To identify paralogous empty sites (paralogous genomic sites devoid of the element), the *Spy* flanking sequences (100 bp) were used as queries in BLASTN search against the corresponding genome.

Results

Discovery and Characterization of *Spy* in the Silkworm

Recently, we have initiated a comprehensive and semiautomated annotation of TEs in the silkworm genome (Xu et al.

2013). The results revealed 269 putative cut-and-paste transposon families that could not be readily affiliated to known superfamilies. After investigating "manually" the characteristics of these families, we were intrigued by seven related families (BmTEdb ids are Bmori_102.674, Bmori_62.1257, Bmori_428.1384, Bm_503, Bm_374, Bm_682, and Bm893) with clearly identifiable TIRs but no apparent TSDs in their flanking regions (fig. 1*A* and supplementary fig. S1, Supplementary Material online). These seven families were designated as *Spy-1_BMo* to *Spy-7_BMo* (table 1). A multiple alignment of individual copies of each family was performed using MUSCLE. We found that all these seven *Spy* families are characterized by flanking 5′-AAA-3′ and 5′-TTT-3′ terminal trinucleotide. At this point, we cannot distinguish whether the AAA and TTT motifs are part of the TIRs or the host flanking sequences. To distinguish between these two possibilities, we searched for paralogous empty sites, which occur when the transposon inserted within another repetitive element in the genome. This analysis unambiguously revealed that the AAA and TTT motifs were actually from the host DNA; that is, an uninterrupted AAATTT motif remained at the paralogous empty site (fig. 1*B*). Thus, the AAATTT motifs represent host target sequences rather than the terminal nucleotides of the TIRs and *Spy* elements precisely inserted between the central A and T nucleotides without any alteration of the target motif.

To estimate the abundance of these seven *Spy* families in *B. mori*, the consensus of each family was used as query in BLASTN searches (*e* value $<e^{-5}$, the size >50 bp and identity >80%) against the silkworm genome assembly (International Silkworm Genome Consortium 2008) deposited in GenBank. In total, we identified 2,073 *Spy* elements, which constitute about 1.66 Mb (~0.36%) of the silkworm genome assembly (466 Mb). The copy number of each silkworm *Spy* family ranges from 51 to 688, the size of TIRs ranges from 10 to 12 bp, and the size of individual elements ranges from 160 to 5,443 bp (table 1 and supplementary table S2, Supplementary Material online). The size of typical elements ranges from 1,000 to 1,500 bp (supplementary fig. S2*A*, Supplementary Material online). Moreover, copies of the same family are

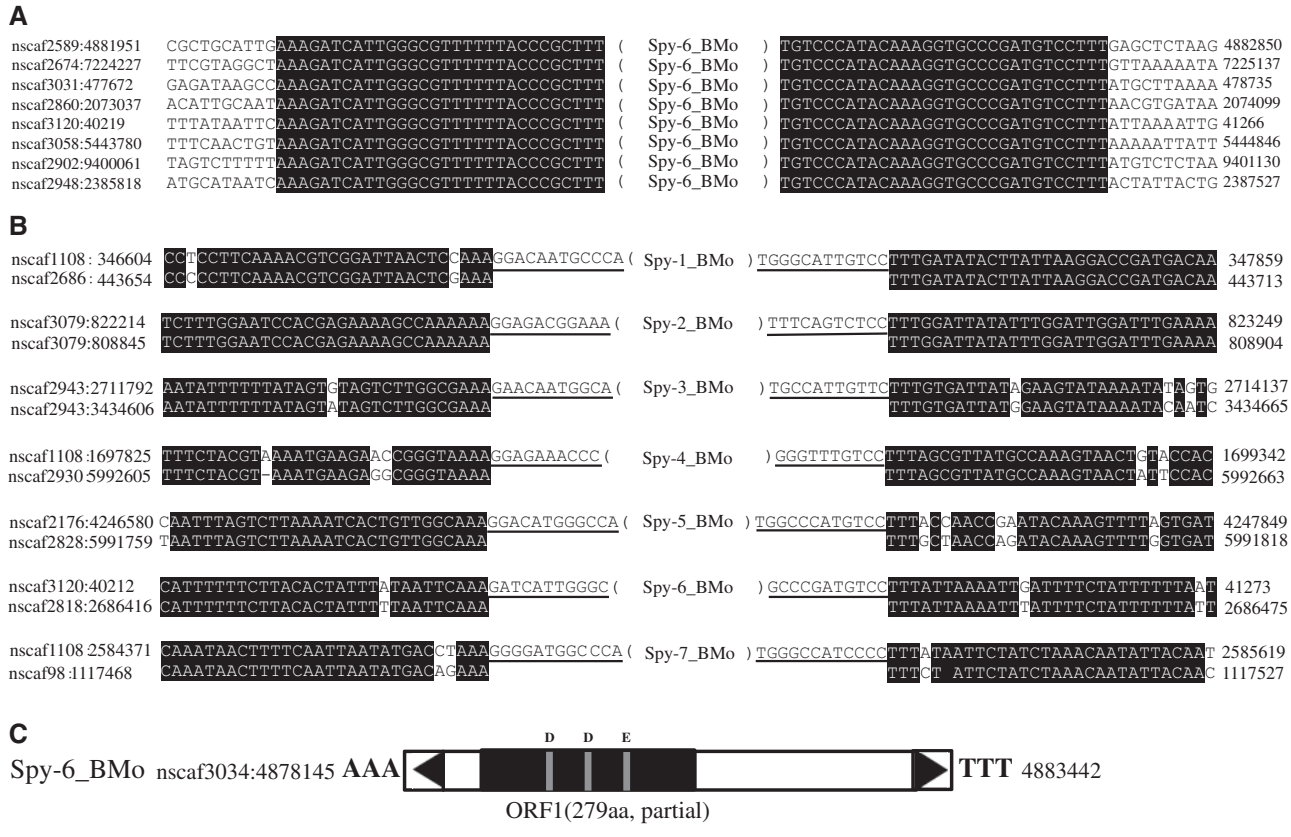


Fig. 1.—Characters of silkworm *Spy* transposons. (*A*) Sequence alignments for *Spy-6_BMo* family. The TIRs and flanking sequences are shown. (*B*) Seven examples of alignments of the flanking sequences of *Spy* insertions with a paralogous sequences found within the same genome but devoid of the transposon. The TIRs of the element is underlined. (*C*) Structure of *Spy-6_BMo*. Black triangles represent the TIRs. ORFs are depicted as solid black boxes and the position of the DDE triad is shown.

Table 1
Summary Information for the *Spy* Families in the Silkworm Genome

Species	TE Family	TIR (bp)	Copies	Length (bp)	Annotation
Bombyx mori	*Spy-1_BMo*	12	377	1,281	Novel
	Spy-2_BMo	11	184	978	Novel
	Spy-3_BMo	11	688	1,007	Novel
	Spy-4_BMo	10	204	1,608	Novel
	Spy-5_BMo	12	132	844	Novel
	Spy-6_BMo	11	51	5,376	Novel
	Spy-7_BMo	12	437	1,334	Novel

NOTE.—Length: The size of consensus sequence for each *Spy* family.

conserved (identity >80%), most expansion events appear to have happened within the past 2 Myr (supplementary fig. S2*B*, Supplementary Material online). However, there are no detectable similarities between *Spy* families besides similar TIRs and TSD (supplementary fig. S3, Supplementary Material online). The locations within the contigs of the *Spy* elements identified through these searches are shown in supplementary table S2, Supplementary Material online. For each silkworm *Spy* family, we derived consensus sequence, which were then used to a sensitive CENSOR search of Repbase (as of August 10, 2013). The results showed that none of these seven *Spy* families had a significant match to any of the transposons cataloged in Repbase.

To validate that these seven families belong to a related group of DNA transposons, ORFs of all 2,073 copies and of the seven consensus sequences were predicted using GetORF, GeneMark.hmm, or GENESCAN, then the predicted ORFs were annotated using homology search to the pfam and NCBI nr protein database. The results showed that a single copy of *Spy-6_BMo* encodes a DDE motif (pfam00665) containing transposase (length = 279 aa) (fig. 1*C*). However, further inspection suggested that the *Spy-6_BMo* transposase is truncated at its C-terminus. The size of consensus sequences for *Spy-1–7_BMo* is 1,281, 978, 1,007, 1,608, 844, 5,367, and 1,334 bp, respectively (table 1 and supplementary fig. S3, Supplementary Material online). Moreover, other *Spy* family (*Spy-1-5* and *7_BMo*) elements and consensus sequences display no significant similarity to other known transposases. Thus, almost all *Spy* elements identified in the silkworm genome likely represent deletion derivatives or nonautonomous elements.

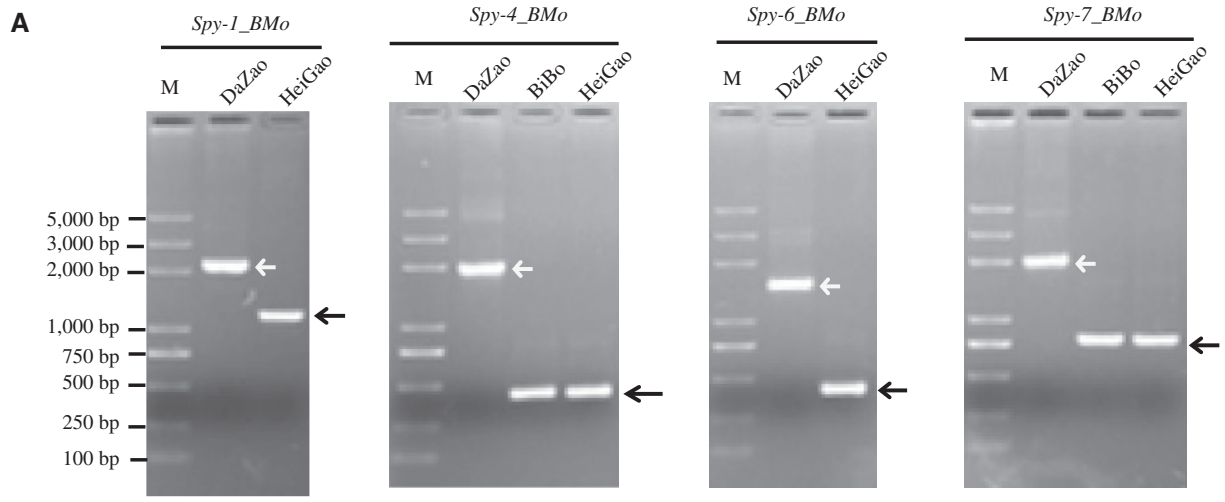
Verified TSDs of *Spy* Elements through Orthologous Empty Sites

Four *Spy* insertions from four distinct families (*Spy-2, -4, -6,* and *-7_BMo*) were selected for PCR assay for presence/absence across three silkworm strains (DaZao [the sequenced strain], BiBo, and HeiGao) using primer pairs flanking each of these insertions. The results indicated that each of these elements exhibited insertion dimorphism among the strains examined. For example, *Spy-7_BMo* is present in the DaZao strain but absent in the HeiGao and BiBo strains (fig. 2*A*). Sequences analysis of the PCR products corresponding to filled and empty *Spy* sites further confirmed that the integration of *Spy* occurs precisely between 5′-AAA-3′ and 5′-TTT-3′ host nucleotides, without deletion or duplication of target sequence (fig. 2*B*). These results also suggest that *Spy* elements have recently transposed in the silkworm genome.

Distribution and Characteristic of *Spy* in Other Species

To investigate the distribution of *Spy* in other species, we used the predicted transposase of *Spy-6_BMo* as a query in BLASTP and TBLASTN searches against the NCBI nr protein database and various GenBank nucleotide databases (nr, WGS, GSS, and EST), respectively. Significant hits (e value $> 10^{-4}$) were manually inspected to look for the presence of features indicative of DNA transposons, including the presence of transposase-coding sequences with a DDE domain, TIRs, and the AAA|TTT target sequence (where | marks the insertion site). We were able to confirm the presence of *Spy* -like elements in 21 invertebrate species. These include two arachnids (*Metaseiulus occidentalis* and *Tetranychus urticae*), 16 insects (one hemiptera [*Rhodnius prolixus*], four lepidopterans [*Bombyx mori*, *Plutella xylostella*, *Manduca sexta*, and *Danaus plexippus*], six hymenopterans [*Acromyrmex echinatior*, *Camponotus floridanus*, *Harpegnathos saltator*, *Linepithema humile*, *Megachile rotundata*, and *Solenopsis invicta*], one strepsipteran [*Mengenilla moldrzyki*], one coleopteran [*Anoplophora glabripennis*], three dipterans [*Drosophila takahashii*, *Phlebotomus papatasi*, and *Mayetiola destructor*]), one bivalve (*Crassostrea gigas*), one hydrozoan (*Hydra magnipapillata*), and one rotifer (*Adineta vaga*) (fig. 3*A* and supplementary table S3, Supplementary Material online). For each species, closely related elements (e value $< e^{-5}$, identity >80%, and sequence length >50 bp) were clustered into families and consensus sequences were derived for each family. The genomic abundance and copy number of each family in each species were estimated (see Materials and Methods and supplementary table S3, Supplementary Material online).

Furthermore, a multiple alignment of representative copies, their coding capacity, and predicted protein secondary structure of the putative transposase, as well as paralogous empty sites are presented for each *Spy* family identified in these various species in supplementary figure S4, Supplementary Material online. Taken together, these analyses reveal the following shared characteristics: 1) Virtually all *Spy* elements are inserted into a 5′-AAA|TTT-3′ target site, 2) all candidate autonomous *Spy* elements contain a single long ORF predicted to encode a D(79–80)D(44–62)E motif-containing transposase, 3) the transposase of most *Spy* elements is predicted to contain a HTH motif at its N-terminus, 4) the TIRs of different *Spy* families are highly variable in length (9–1,474 bp) and sequence, except for a terminal 5′-GGANNNNG-3′ consensus



B

```
DaZao  : GCACTTTCATGTAATAAAATTGATGCTAAAGGAGACTGAAA (Spy-2_BMo) TTTCCGTCTCCTTTGTTTATTTTTTTATGATAGTCATATAT
HeiGao : GCACTTTCATGTCATAAAATTGATACTAAA                        TTTGTATATTTTTTTATGATAGTCATATAT

DaZao  : CTACATAATCCTGACATGATAAAGATGAAAGGACAAACCC  (Spy-4_BMo)  GGGTTTCTCCTTTAATATATACAATATCATCATGAGAAAT
BiBo   : CTACATAATGCTGACATGATAAAGATAAAA                        TTTAATAAATACAATATCATCATGAGAAAT
HeiGao : CTACATAATGCTGACATGATAAAGATGAAA                        TTTAATAAATACAATATCATCATGAGAAAT

DaZao  : TGATCGGCAAACCAAAGACATTGCAATAAAGATCATTGGGC (Spy-6_BMo) GCCCGATGTCCTTTAACGTGATAAATAAAATCCAAAAATAC
HeiGao : TGAGCTTCAAAGCAAAGACATTGCAATAAA                        TTTAACGTGATAAATAAAATCCAAAAATAC

DaZao  : TAGTTTCATCTAATTAACATCAACCGTAAAGGGGATTGCCCA(Spy-7_BMo)TGGGCCATCCCCTTTCATACTTCAATTTAGCAATTTTGAAAT
BiBo   : TAGTTTCATCTAATTAACATCAACCGTAAA                        TTTCATACTTCAATTTAGCAATTTTGAAAT
HeiGao : TAGTTTCATCTAATTAACATCAACTGTAAA                        TTTCATACTTCAATTTAGCAATTTTGAAAT
```

Fig. 2.—Target sites of the *Spy* family were tested using PCR and sequencing. (*A*) Results of PCR search for indel silkworm strains (DaZao, BiBo, and HeiGao) among four insertion sites of four silkworm *Spy* families (*Spy-2, -4, -6*, and *-7_BMo*). The black arrow points to the corresponding *Spy* lack at this genomic location. The white arrow represents the corresponding *Spy* occupied at this genomic location. (*B*) Results of sequencing for above locations. The target sequences are marked with black bold font, TIRs are marked with underline, and flank sequences are in blank.

motif that is relatively well conserved across families (fig. 3*B*). Extreme variability in TIR length is not unprecedented within a given superfamily of DNA transposons (e.g., Marquez and Pritham 2010). We found no obvious association between the occurrence of *Spy* elements with long TIRs and their phylogenetic distribution (fig. 3*A*). For example, we found that *Spy* with long TIRs occur in species of insects, arachnida, and rotifers, whereas those with short TIRs are found in species of insects, molluscans, and hydrozoans. The sequence of the long TIRs do not display any apparent subrepeat structure, dinucleotide compositional bias (data not shown), or significant sequence similarities across *Spy* families, except for the conserved terminal 5′-GGANNNNG-3′ consensus motif. Thus, *Spy* elements appear to have undergone repeated episodes of TIR expansion and/or contraction, but the underlying mechanism and biological significance, if any, are unclear. Conservation of the terminal nucleotides within a DNA transposon superfamily is thought to reflect similar cleavage specificity of related transposases (Feschotte et al. 2002).

Each of the *Spy* consensus sequences defined in this study was subject to homology search against RepBase (as of October 15, 2013) using Censor. The results of these searches showed that none of the *Spy* families identified had a close match to any known TEs cataloged in Repbase except *Spy-1_CGi* and *Spy-2_CGi*, from the oyster *C. gigas*, which were deposited in RepBase under different names (*ISL2EU-4_CGi* and *ISL2EU-6_CGi*) and classified as members of the *ISL2EU* subgroup (Bao and Jurka 2013). However, as argued below, our data suggest that these and other *Spy* families actually define a separate clade of elements with unique properties and a distant relationship to *ISL2EU* elements.

Evolutionary Relationships among *PIF/Harbinger*, *ISL2EU*, *IS5*, and *Spy* Transposons

To investigate the relationships of *Spy* elements to known DNA transposons, the transposase sequences predicted for each consensus of *Spy* families using GetORF, GeneMark.

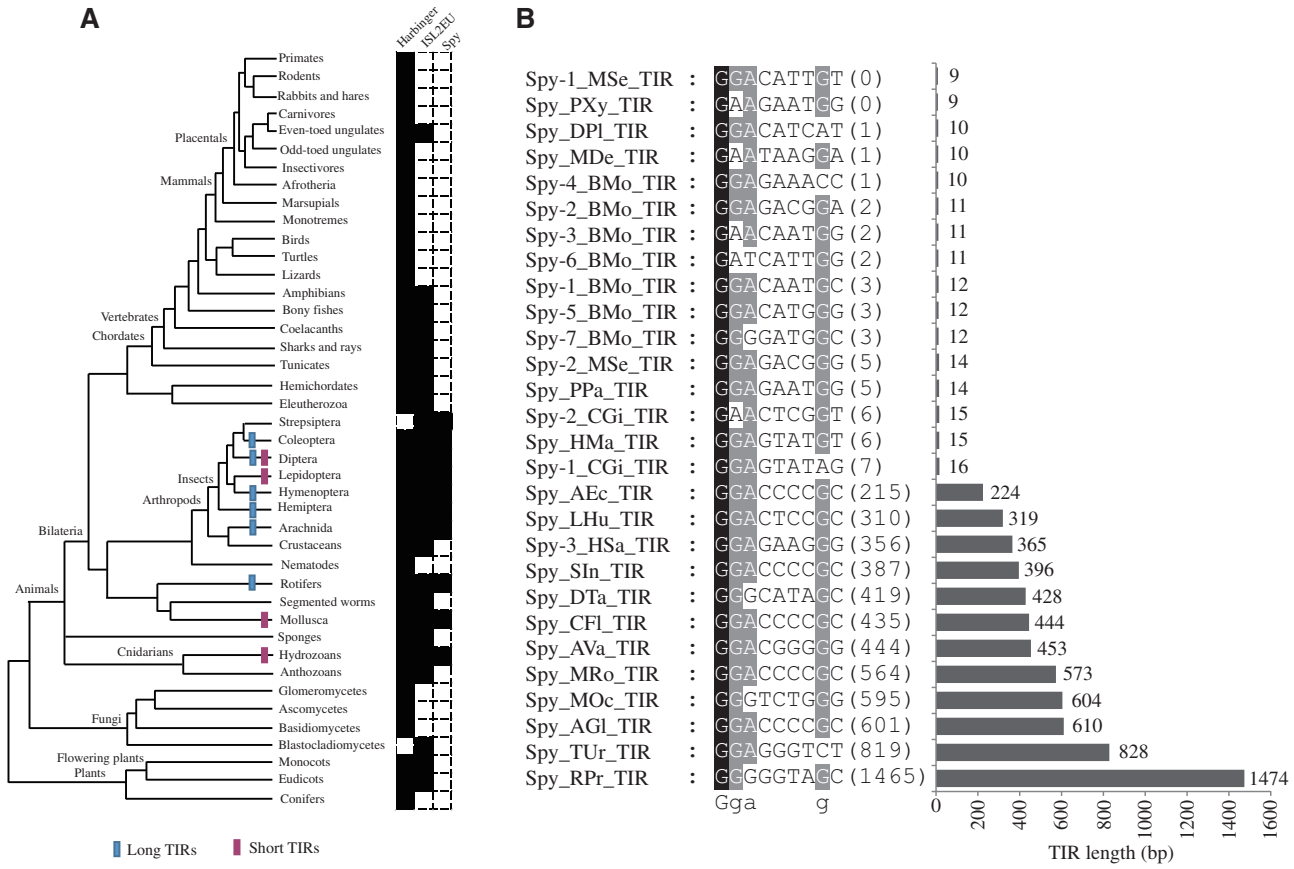


Fig. 3.—Taxonomic distribution of the *Spy*, *PIF/Harbinger*, and *ISL2EU* transposons as well as character of *Spy* TIRs. (*A*) Taxonomic distribution of three groups across the eukaryotic tree of life. Black and white boxes indicated presence and absence, respectively. (*B*) TIRs of all *Spy* identified in this study. Sequences are major-rule consensus derived from the alignment of multiple copies of each family.

hmm, and GENESCAN as well as 17 *ISL2EU* transposases (cataloged in Repbase), 24 transposases representatives of the *PIF/Harbinger* superfamily (cataloged in Repbase), which have been recently shown to be distantly related to *ISL2EU* (Yuan and Wessler 2011) and 13 bacterial *IS5* transposases downloaded from the IS database (Kichenaradja et al. 2010) and related to *PIF/Harbinger* (Zhang et al. 2001) were used to construct a multiple alignment of their core catalytic DDE domain (see Materials and Methods). The alignment (fig. 4*A*) reveals that all these transposases are characterized by a highly conserved set of amino acids: (D(19–29)K(29)D(18–40)D(11–19)P(18–28)R(3)E) where the numbers refer to the spacing between the conserved residues and the underlined residues represent the proposed catalytic DDE triad. In addition to these conserved residues shared by all four groups of transposases, each group is unified by a distinct set of additional conserved residues (marked with the black triangle below the alignment in fig. 4*A*). For example, we found 3 unique conserved residues (G, G, and N) in the *PIF/Harbinger* transposases, 6 unique conserved residues (Y, L, S, I, H, and R) in the *ISL2EU* transposases, 15 unique conserved residues (Y, S, N, L, P, G, P, A, R, D, Q, N, V, T, and W) in the *Spy* transposases, and 5 unique conserved residues (R, G, G, K, and L) in the bacterial *IS5* transposases.

In addition, a comparison of secondary structure predictions of *Spy*, *ISL2EU,* and *PIF/Harbinger* transposases suggests that *Spy* transposases have a distinct architecture within their DDE catalytic core domain. For example, in *Spy* transposases, the first D of the proposed DDE triad is typically located in a predicted beta-sheet, the second D is located between two beta-sheets, and the last E within a beta-sheet, whereas for *PIF/Harbinger* and *ISL2EU* transposases, the first D occurs between two beta-sheets, the second D is typically between a beta-sheet and an alpha-helix, and the last E occurs within a predicted alpha-helix (fig. 5 and supplementary fig. S2, Supplementary Material online).

To further explore the evolutionary relationships between these four groups of transposases, we used the multiple alignment described above to perform a Bayesian phylogenetic analysis. The resulting tree (fig. 4*B*) shows that *Spy*, *ISL2EU*, *PIF/Harbinger*, and *IS5* transposases formed four distinct highly supported monophyletic clades, with *ISL2EU* and *Spy* forming

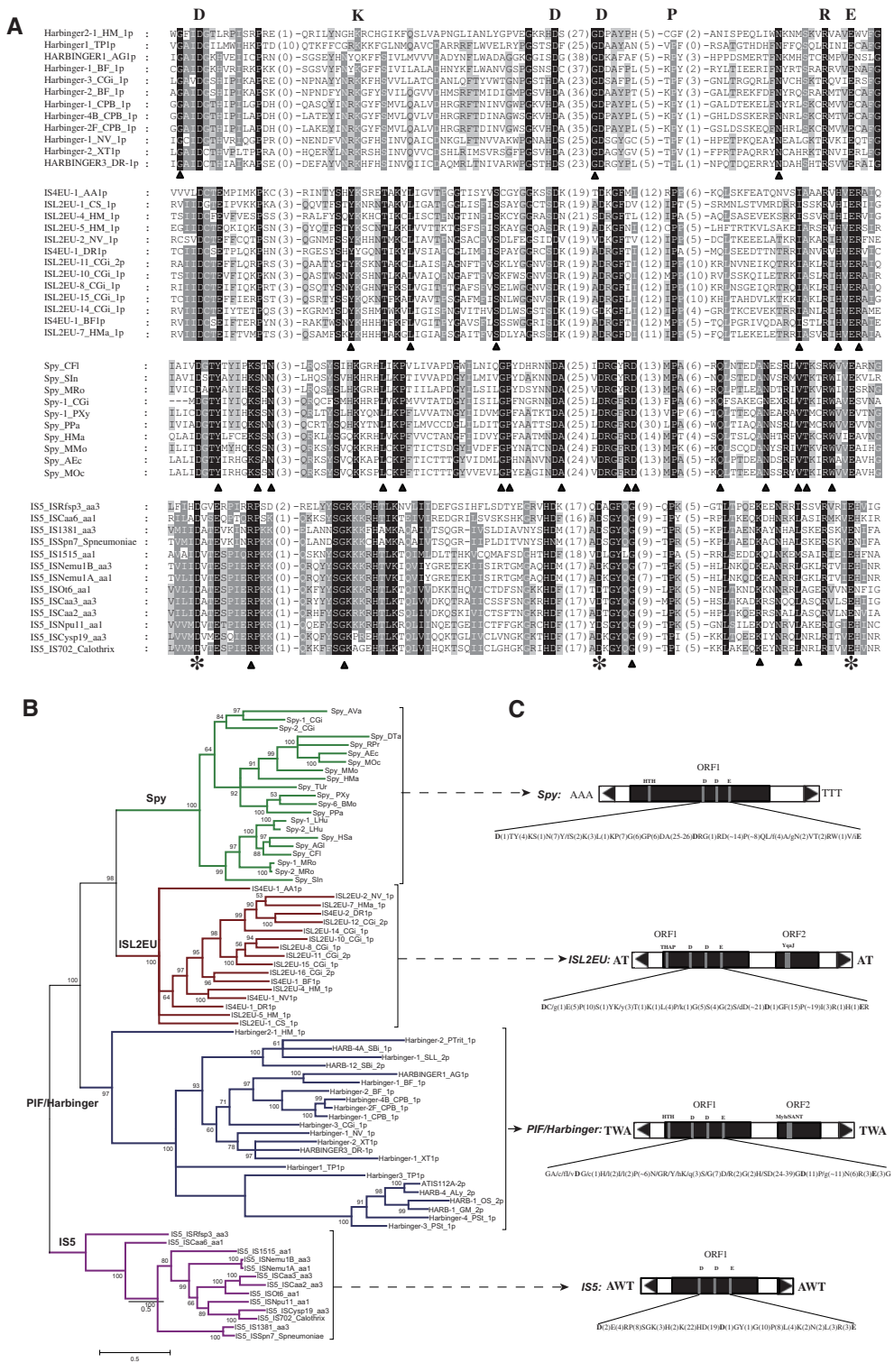


Fig. 4.—The results of phylogenetic analysis, coding capacity, and conserved transposase motifs of *IS5*, *Spy*, *ISL2EU*, and *PIF/Harbinger*. (*A*) The alignment of DDE domain of each superfamily after redundancy elimination. Distances between the conserved blocks are indicated in the number of amino acid residues. Conserved residues within each superfamily are highlight in black or gray. The DDE triad identified here is marked with asterisks below alignments. Common conserved residues among four superfamilies are marked with letter above the alignments. Unique conserved residues of each superfamily are marked with blank triangle below the alignment. (*B*) Phylogenetic tree based on DDE domain sequences of each superfamily. In front of the colon represents corresponding *IS5*, *Spy*, *ISL2EU*, or *PIF/Harbinger* elements name; behind the colon represents species. (*C*) Structure of each superfamily. Black triangles represent the TIRs. ORFs are depicted as solid black boxes, and the position of the DDE triad and additional domains is shown above. Target sequences are shown in flank.

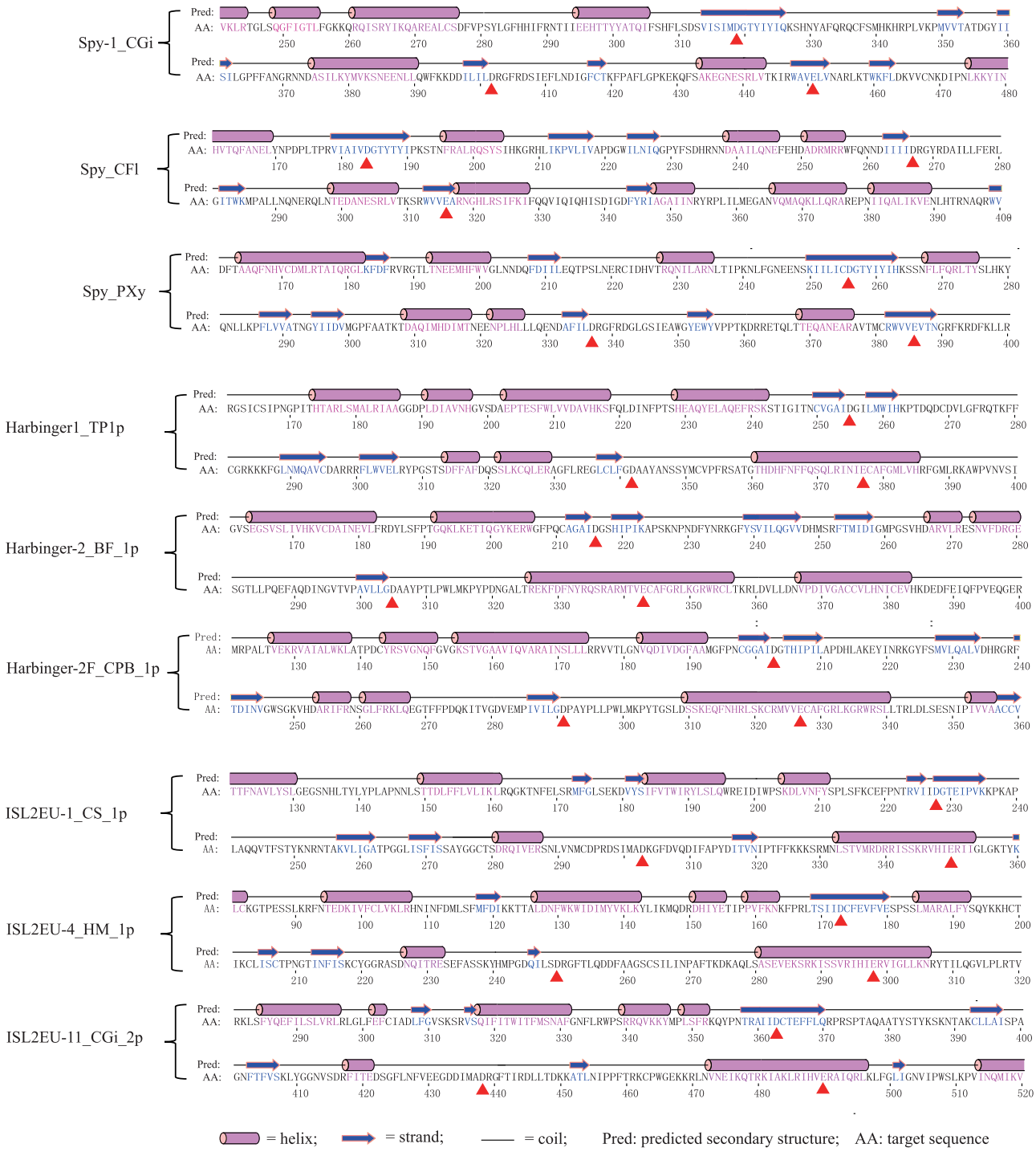


Fig. 5.—The secondary structure of DDE domain-containing transposase for *Spy*, *ISL2EU*, and *PIF/Harbinger* elements. The DDE triad is marked with red triangles below sequence.

sister clades. Thus, *Spy* can be considered a new group of transposons distinct from *IS5*, *PIF/Harbinger*, and *ISL2EU*.

Discussion

In this work, we report a new eukaryotic group of DNA transposons, called *Spy*. Most autonomous *Spy* transposons include only one DDE motif-containing transposase and TIRs. These are common features of cut-and-paste transposons (Yuan and Wessler 2011). In addition, all cut-and-paste transposons so far described are also characterized by TSDs (Wicker et al. 2007). However, *Spy* transposons have no TSDs in the flanking regions. The analyses of paralogous and orthologous

empty sites indicated that the integration sites of *Spy* transposons are precisely between the host AAA and TTT nucleotides, without deletion and duplication of host sequences upon *Spy* insertion. Typically, the TSDs of cut-and-paste transposons arise from the staggered joining of the 3′OH transposon ends, which are generated by double-strand breaks at the ends of the transposon. The 3′OH ends join to staggered positions on the top and bottom strands of the target DNA, followed by repair of the resulting gaps (Craigie and Mizuuchi 1985; Craig et al. 2002). In contrast, we speculate that the 3′OH ends of *Spy* that are exposed by blunt-ended double-strand breaks at each end of the transposon will join to non-staggered positions in the target DNA. Although other TEs do not create TSDs during transposition, these elements (Helitrons, Cryptons, and some class 1 transposons) do not belong to the subclass of cut-and-paste DNA transposons. Thus, *Spy* transposons are, to the best of our knowledge, unique among eukaryotic DNA transposons, in creating no TSD upon insertion. Biochemical studies would be needed to characterize the cleavage activities of *Spy* transposases.

To estimate the taxonomic distribution of *Spy* in other species, we used transposase homology search approach. The results indicated that *Spy* transposons are only detectable in invertebrate animals: Including arachnida (2 species), insecta (16), bivalvia (1), hydrozoa (1), and rotifer (1). The apparent predominance of *Spy* elements in insects could represent a bias in the databases for insect genomes. Meanwhile, it should be noted that *Spy* distribution was investigated using transposase homology search. The major limitation of this method is that it cannot identify nonautonomous *Spy* elements where transposase sequences are completely missing. Thus, we cannot exclude the possibility that related elements exist in many other taxa, including noninvertebrates.

Although most of the transposons identified here have not been previously reported in Repbase, *Spy* transposons of pacific oyster were previously identified and classified as members of the *ISL2EU* subgroup (Bao and Jurka 2013). In addition, previous studies showed that *ISL2EU* and *PIF/Harbinger* are evolutionarily related but clearly distinct from all other superfamilies, and *PIF/Harbinger* is distantly related to vast bacterial *IS5* group, which also include the ISL2 group of bacterial ISs (Zhang et al. 2001; Chandler and Mahillon 2002; Yuan and Wessler 2011). After aligning the DDE domains of *Spy*, *IS5*, *PIF/Harbinger*, and *ISL2EU* transposons using MUSCLE, we found that these transposases display a conserved set of residues, including the catalytic DDE triad (D(19–29)K(29)D(18–40)D(11–19)P(18–28)R(3)E) (fig. 4*A*). However, the resulting tree shows that *Spy*, *IS5*, *PIF/Harbinger*, and *ISL2EU* formed four separate clades. Because *Spy* and *ISL2EU* form sister clades in our phylogenetic analysis, one could propose that they form a single superfamily or subgroup. However, *Spy* transposons have several unique features that distinguish them from *ISL2EU* and from the other two groups of transposons.

First, *Spy* transposases share a unique set of conserved residues that are not shared by those encoded by the other groups of transposons (fig. 4*A*). In fact, each of the four groups had a unique set of conserved residues in their catalytic domain. The number of conserved residues (15) is larger for *Spy* than for the other groups (fig. 4*A*), which could reflect their more recent divergence from a common ancestor. This hypothesis is supported by the apparently narrower taxonomic distribution of *Spy* elements among eukaryotes, being restricted to invertebrates (fig. 3*A*). In contrast, both *ISL2EU* and *PIF/Harbinger* groups include members in a wide range of animals and in plants (fig. 3*A*) and thus may have deeper evolutionary roots.

Second, most *Spy* transposons contain a single ORF encoding the putative transposase (fig. 4*C* and supplementary fig. S2, Supplementary Material online). In contrast, most *ISL2EU* and *PIF/Harbinger* transposons encode an additional ORF besides their transposase ORF, encoding a DNA-binding protein with a Myb/SANT domain in *PIF/Harbinger* elements (Zhang et al. 2004; Sinzelle et al. 2008), and a protein with an YqaJ exonuclease domain in *ISL2EU*. Furthermore, the transposase of *ISL2EU* contains a THAP DNA-binding domain at its N-terminus (a type of zinc-finger domain), whereas the transposase of most *Spy* elements does not appear to contain any zinc-finger domain but instead is predicted to contain a HTH motif at its N-terminus (supplementary table S3, Supplementary Material online). Moreover, the predicted secondary structure of the DDE catalytic core domain of *Spy* transposases appears distinct from that of *ISL2EU* and *PIF/Harbinger* transposases (fig. 5 and supplementary fig. S2, Supplementary Material online). Thus, *Spy* elements appear to have different coding capacity and transposase architecture than *ISL2EU* elements.

Finally, *Spy* transposons are distinct from all other groups of DNA transposons by their strong insertion preference within the AAATTT motif and the lack of TSDs upon insertion. *PIF/Harbinger* and *IS5* elements generate 3-bp TSD and *ISL2EU* generate 2-bp TSD (typically AT) (Zhang et al. 2001; Yuan and Wessler 2011). We note that the four groups of elements share a preference for insertion into AT-rich target sequences. In sum, on the basis of the above discussion we propose that *Spy* represents a distinct group within a larger assemblage of evolutionarily related transposons we propose to designate "PHIS" for *PIF/Harbinger*, *ISL2EU*, and *Spy*.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr Fang-Yin Dai for help in collecting the domesticated silkworm samples and all other members of

Zhang's laboratory for their laboratory assistance. We thank Dr Nancy Craig for her critical and helpful comments on our manuscript. Z.Z., C.F., and M.-J.H. designed the study; M.-J.H. performed the experiments; H.-E.X., H.-H.Z., and M.-J.H. analyzed the data; Z.Z. provided the reagents for experiments; Z.Z., C.F., H.-E.X., and M.-J.H. drafted and revised the manuscript. All authors read and approved the final manuscript. This work was supported by the Hi-Tech Research and Development (863) Program of China (2013AA102507); by a grant from Natural Science Foundation Project of CQ CSTC (grant cstc2012jjB80007); and by grant R01-GM077582 from the National Institutes of Health to C.F.

Literature Cited

Bao WD, Jurka J. 2013. DNA transposons from the Pacific oyster genome. Repbase Rep. 13:578–580.
Bryson K, et al. 2005. Protein structure prediction servers at University College London. Nucleic Acids Res. 33:W36–W38.
Chandler M, Mahillon J. 2002. Insertion sequences revisited, in mobile DNA II. Washington (DC): American Society for Microbiology.
Craig NL, Craigie R, Gellert M, Lambowitz AM. 2002. Mobile DNA II. Washington (DC): American Society for Microbiology Press.
Craigie R, Mizuuchi K. 1985. Mechanism of transposition of bacteriophage Mu: structure of a transposition intermediate. Cell 41:867–876.
Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.
de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 7:e1002384.
Dodd IB, Egan BJ. 1990. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucleic Acids Res. 18:5019–5026.
Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3:329–341.
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331–368.
Feschotte C, Zhang X, Wessler SR. 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons, in mobile DNA II. Washington (DC): American Society of Microbiology Press.
Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5:103–107.
Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 41:95–98.
Holt RA, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149.
International Silkworm Genome Consortium. 2008. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol. 38:1036–1045.
Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Kapitonov VV, Jurka J. 2001. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 98:8714–8719.
Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–6574.
Kempken F, Windhofer F. 2001. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110: 1–9.
Kichenaradja P, Siguier P, Pérochon J, Chandler M. 2010. ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 38:D62–D68.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Lisch D. 2002. Mutator transposons. Trends Plant Sci. 7:498–504.
Marquez CP, Pritham EJ. 2010. Phantom, a new subclass of mutator DNA transposons found in insect viruses and widely distributed in animals. Genetics 185:1507–1517.
Mitra R, Fain-Thornton J, Craig NL. 2008. piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J. 27:1097–1109.
Nagaraja GM, Nagaraju J. 1995. Genome fingerprinting of the silkworm, *Bombyx mori*, using random arbitrary primers. Electrophoresis 16: 1633–1638.
Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Sarkar A, et al. 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Mol Genet Genomics. 270:173–180.
Shao H, Tu Z. 2001. Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics 159: 1103–1115.
Sinzelle L, et al. 2008. Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. Proc Natl Acad Sci U S A. 105:4715–4720.
Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Xu HE, et al. 2013. BmTEdb: a collective database of transposable elements in the silkworm genome. Database (Oxford) 2013:bat055.
Yuan YW, Wessler SR. 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A. 108: 7884–7889.
Zhang X, Jiang N, Feschotte C, Wessler SR. 2004. PIF-and Pong-like transposable elements: distribution, evolution and relationships with Tourist-like MITEs. Genetics 166:971–986.
Zhang X, et al. 2001. *P* instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. Proc Natl Acad Sci U S A. 98: 12572–12577.

Associate editor: Esther Betran