Merlin, a New Superfamily of DNA Transposons Identified in Diverse Animal Genomes and Related to Bacterial IS*1016* Insertion Sequences

Cédric Feschotte[1]

Departments of Plant Biology and Genetics, The University of Georgia, Athens

Several new families of DNA transposons were identified by computer-assisted searches in a wide range of animal species that includes nematodes, flat worms, mosquitoes, sea squirt, zebrafish, and humans. Many of these elements have coding capacity for transposases, which are related to each other and to those encoded by the IS*1016* group of bacterial insertion sequences. Although these transposases display a motif similar to the DDE motif found in many transposases and integrases, they cannot be directly allied to any of the previously described eukaryotic transposases. Other common features of the new eukaryotic and bacterial transposons include similarities in their terminal inverted repeats and 8-bp or 9-bp target-site duplications. Together, these data indicate that these elements belong to a new superfamily of DNA transposons, called *Merlin*/IS*1016,* which is common in many eubacterial and animal genomes. We also present evidence that these transposons have been recently active in several animal species. This evidence is particularly strong in the parasitic blood fluke *Schistosoma mansoni*, in which *Merlin* is also the first described DNA transposon family.

Introduction

Transposable elements (TEs) represent the largest genomic component of most eukaryotic organisms. They account for 15% to 25% of the genetic material in fruit fly and mosquito (Holt et al. 2002; Kapitonov and Jurka 2003), 35% to 45% in mouse and humans (Lander et al. 2001; Waterston et al. 2002), and 50% to 80% in maize and barley (SanMiguel et al. 1996; Vicient et al. 1999). Even the relatively compact genomes of *Arabidopsis thaliana*, *Neurospora crassa*, *Caenorhabiditis elegans*, and *Fugu rubripes* harbor a wide diversity of TEs (*Caenorhabditis elegans* Genome Consortium 1998; The *Arabidopsis* Genome Initiative 2000; Aparicio et al. 2002; Galagan et al. 2003). Compelling evidence indicates that TEs are not just "junk DNA," but have played a central role in the structural organization and plasticity of genomes and participated in the establishment of new cellular functions during evolution (for reviews see Kidwell and Lisch [2001], Bowen and Jordan [2002], and Feschotte, Jiang, and Wessler [2002]). With the advent of large-scale sequencing, the identification and characterization of TE populations has become an important facet of genome biology.

Traditionally, TEs are identified and classified upon sequence similarities with previously established TE families (Capy et al. 1998; Craig et al. 2002; Feschotte, Jiang, and Wessler 2002). As for most genetic entities, the coding regions of TEs evolve with more functional constraints than their noncoding regions. Therefore, it is possible to recognize and classify the elements only when they have preserved a significant fraction of their encoded products (e.g., reverse transcriptase, integrase, and transposase). Nonautonomous elements that do not contain significant coding capacity are usually classified on the basis of sequence similarities with autonomous elements present in the same genome (Feschotte, Zhang, and Wessler 2002).

TEs are divided into class 1 (retrotransposons) and class 2 (DNA transposons and rolling-circle transposons) elements. DNA transposons are further divided into superfamilies upon sequence similarities and/or specific signatures in the encoded transposases (Capy et al. 1998). Based on these criteria, eight superfamilies of DNA transposons have been recognized so far in eukaryotes: Tc1/*mariner*, hAT, CACTA, P, *Mutator*, *piggyBac*, PIF/*Harbinger*, and, recently, Transib (Doak et al. 1994; Capy et al. 1998; Robertson 2002; Kapitonov and Jurka 2003). Several of the superfamilies include members from two or three different eukaryotic kingdoms, which suggests that they have diverged early during or before eukaryotic evolution. Data gathered from the functional study of a limited number of elements indicate that eukaryotic DNA transposons adopt a "cut-and-paste" mechanism of transposition (for review see Craig et al. [2002]). During this process, transposase molecules bind to the ends (usually to the terminal inverted repeats, or TIRs) of their respective transposon(s) and catalyze both the DNA cleavage and strand transfer steps of the transposition reaction. Transposon integration results in the duplication of a short host sequence at the insertion site or target-site duplication (TSD). The length of the TSD is determined by the enzymatic properties of each transposase (Haren, Ton-Hoang, and Chandler 1999; Craig et al. 2002). Hence, elements responding to the same superfamily of transposases generally create TSD of the same length and share similarities in their TIRs sequences (Chandler and Mahillon 2002; Feschotte, Zhang, and Wessler 2002; Feschotte, Swamy, and Wessler 2003).

Here, I describe several new families of DNA transposons from several animal species identified by computer-assisted searches. Many of these elements have coding capacity for transposases, which are phylogenetically related to each other and to those encoded by the IS*1016* group of bacterial insertion sequences (IS). Elements that belong to different families share also similarities in TIRs and create 8-bp (or 9-bp) TSDs upon insertion, a characteristic shared by the IS*1016* bacterial elements. Together, these data indicate that the different TE families belong to a new superfamily of DNA

[1] Present address: Department of Biology, The University of Texas, Arlington, TX 76019, USA.

Key words: Transposable elements, DNA transposons, transposase, insertion sequences, IS*1016*.

E-mail: cedric@plantbio.uga.edu.

Mol. Biol. Evol. 21(9):1769–1780. 2004
doi:10.1093/molbev/msh188
Advance Access publication June 9, 2004

transposons, called *Merlin*/IS*1016,* which is commonly found in eubacterial and animal genomes.

Methods

Data Mining and Sequence Availability

The original *Merlin_Cb1* element was fortuitously discovered in October 2001. Most database searches were performed online from this date through October 2003 and updated while this article was being prepared, in March 2004. Blast searches (predominantly BlastN and TBlastN) were conducted through various Web servers essentially by use of default parameters and without filtering for simple or complex repeats. To optimize the probability of detection of all *Merlin* elements within the same species, TBlastN searches were performed iteratively by using first the sequence of *Merlin* transposase(s) available from the closest species (e.g., Merlin_Dr1p from the zebrafish against the human genome), and then by using the newly identified *Merlin* sequences against the same database. Generally, a hit was considered significant when the e-value was lower than 10^{-4}. For extremely distant species, such as eukaryotes and bacteria, hits with e-values up to 0.01 were also considered but validated only after closer inspection for their coding potential and the presence of conserved protein motifs. Reiterative PSI-Blast searches were also carried out, although their value was limited by the fact that they can only be performed against protein databases available through NCBI.

Most Blast searches were conducted against the various GenBank databases (nr, GSS, WGS, and EST) through the NCBI server (http://www.ncbi.nlm.nih.gov/blast/). Other searches were conducting by use of the following material and servers. For *Caenorhabditis briggsae*, the Jim Mullikin's assembly (6/24/02, 98% coverage) produced by the Sanger Institute and the Genome Sequencing Center (GSC) of Washington University, St Louis was searched at http://www.genome.wustl.edu/projects/cbriggsae. For *Schistosoma mansoni*, the shotgun genome assembly (12/10/2003) produced by the Sanger Institute was searched at http://www.sanger.ac.uk/Projects/S_mansoni. For *Ciona intestinalis*, the WGS assembly (unmasked, version 1.0, >80% coverage) produced by the DOE Joint Genome Institute (JGI) was searched at http://www.jgi.doe.gov/. The JGI server was also used for searching numerous other species for which extensive sequence data were recently produced by the JGI as WGS assembly (e.g., *Takifugu rubripes*) or raw sequencing reads (e.g., *Phytophthora sojae*). For *Danio rerio*, the Zv2 preliminary assembly generated by Ensembl (July 2003, supercontig coverage: 95%) was searched at http://www.ensembl.org/Danio_rerio.

Complete or consensus *Merlin* sequences reported in this article were deposited in Repbase Update, www.girinst.org/Repbase_Update.html (Jurka 2000). Accessions numbers, positions and sequences of individual elements mined from the various databases are available from the author upon request.

Gene Predictions, Alignments, and other Sequence Analyses

Conceptual translations were performed with the Translate program (www.expasy.org/tools/dna.html) or with MacVector version 7.0 (http://www.accelrys.com/products/macvector/). Transposase coding sequences were assembled by removing introns predicted with more than 85% confidence by NetGene2 (http://www.cbs.dtu.dk) and/or FGENESH (http://genomic.sanger.ac.uk/gf/gf.html). When necessary, frame shifts were judiciously introduced according to nucleotide alignments of closely related elements. Putative initiation codons were predicted by NetStart (http://www.cbs.dtu.dk). The resulting transposase sequences were aligned by ClustalW with default parameter setting and edited manually in MacVector.

In this study, two TE copies are defined as being members of the same family when they display at least 85% pairwise similarity over their entire nucleotide sequence. Pairwise sequence comparisons were carried out by ClustalW or LFASTA, and percentage similarities were calculated by excluding gaps that spanned more than 3 nucleotides. Alignment of multiple family members and computation of consensus sequences was performed by ClustalW in MacVector with the majority rule application. Other sequence manipulations were performed by MacVector and through the NCBI and Infobiogen (www.infobiogen.fr) servers.

Results

Discovery of *Merlin* in Nematodes

Merlin_Cb1-1 was discovered as a 1,914-bp insertion nested into a copy of a previously uncharacterized repeat family (*hAT_Cb1m*) from the nematode *Caenorhabiditis briggsae* (Feschotte, unpublished data). Alignment of multiple members of the *hAT_Cb1m* family shows that the insertion of *Merlin_Cb1-1* created an 8-bp TSD (fig. 1*A*). The TSD and the presence of long TIRs in *Merlin_Cb1-1* (see below) was indicative of the insertion of a DNA transposon. BlastN searches using *Merlin_Cb1-1* as a query against the whole-genome shotgun (WGS) assembly of *C. briggsae* produced six highly significant hits (e-values < 2e-107). The six hits started precisely from one or both ends of *Merlin_Cb1-1* and display 95% to 100% sequence similarity to the query over their whole length. In contrast, the flanking sequences were totally unrelated to each other. Together, these observations suggest that the six hits correspond to different copies of the same TE family inserted at various positions in the genome. A consensus sequence was constructed from an alignment of the six copies; it is 1,915 bp long and has 141-bp TIRs with only three mismatches (figs. 2 and 3). The *Merlin_Cb1* consensus contains a predicted gene interrupted by one intron, which can encode a 338-aa protein (Merlin_Cb1p). Blast searches revealed that the C-terminal half of this protein has strong similarity with the putative transposases of IS*1016* from *Haemophilus influenzae* (and related bacterial IS) and with predicted or hypothetical proteins from a wide range of eukaryotes, such as other nematodes, flat worms, mosquito, ascidians,

A

```
CAAC01000127  94877 ACATTGTTATGCTACTTGTTAAGAAATAAA/MERLIN_CB1-1/GAAATAAATTCAAATAATTCTTCAATAATT  96856
CAAC01000124 403263 ACATTGTTATGCTACTTGTTAAGAAATAAA                    TTCAAATAATTCTTCAATAATT 403212
```

B

```
C002525074       38 GGTTTAGAGAGCTTTTTGTAAAAGTCATCT/MERLIN_SM1-9/AGTCATCTTTCACTCCATCATAGCTGCAGT   1197
C002144262     5628 AGTTCAGAGAGCTTTATGTAAGAGTCATCT                    TTCACTTCATCATGGCTGCAGT   5679
```

C

```
AABS01001053  14298 CAAAAAAGCCGGTTGAGAAAGCAAGAGATT/MERLIN_CI1-1/CAAGAGATTTTCTATAGTAAAAATAGTCAT  15935
AABS01001403   2948 CAAAAAGGCCGTGTGAGAAAGCATAAGATT                    TTCCATAGTAAAAATAGTCAT   2897
```

D

```
CTG01089     173744 ACATTAAACACTAACATTATTATAACTTTT/MERLIN_DR2M8/TAACTTTTAAACGTATTGTTTTATTTCACT 175084
CTG22761     316787 ACATTAAACACTAACATTATAATAACTTTT                    AAACATATTGTTTTTTTTCACT 316838
```

FIG. 1.—Target-site duplications (TSD) created upon the insertion of *Merlin* elements. Shown are four examples (*A–D*) of alignments of the flanking sequences of *Merlin* insertions with a paralogous sequence found within the same genome but devoid of the transposon. The paralogous "empty" sequence presumably corresponds to the target sequence before the insertion. They are generally derived from a repeat family found in multiple copies in the genome, for which only one copy has suffered the insertion (see text). Paralogous sequences were identified by BlastN with *Merlin* flanking sequences as queries. The transposon sequence is represented between backslashes. The TSD created upon insertion of the element is underlined. The database accession numbers and coordinates of the aligned segments are given. (*A*) Insertion of *Merlin_Cb1-1* from *C. briggsae*; (*B*) insertion of *Merlin_Sm1-9* from *S. mansoni*; (*C*) insertion of *Merlin_Ci1-1* from *C. intestinalis*; (*D*) insertion of *Merlin_Dr2m8* from *D. rerio*.

zebrafish, *Xenopus*, and humans (see table 1 and description below). This finding strongly suggests that Merlin_Cb1p is the transposase responsible for the spread of *Merlin_Cb1* elements.

A BlastN search that uses the ends of the *Merlin_Cb1* consensus shows approximately 10 related nonautonomous elements in the *C. briggsae* genome. These elements are short (<500 bp) and have no significant coding capacity. They are flanked by 8-bp TSD, have TIRs with strong similarity to those of *Merlin_Cb1* (see figure 2, *Merlin_Cb1m1, 2, 3*), but their internal sequences display little, if any, sequence similarities to each other or to *Merlin_Cb1*. These elements likely reflect the past activity of other *Merlin*-like transposons in the *C. briggsae* genome.

TBlastN searches did not reveal the presence of full-length *Merlin* transposase homologs in the *C. elegans* N2 genome sequence. However, a 388-bp element was identified that contains a short ORF (41 aa) with 46% identity (60% similarity) to the *Merlin_Cb1* transposase (see figure 3). This element, *Merlin_Ce1m1*, is flanked by an 8-bp direct repeat and has 23-bp TIRs with some similarities to those of *C. briggsae Merlin* elements (fig. 2). The TIRs of *Merlin_Ce1m1* also display strong similarity to those of five nonautonomous DNA transposons families previously identified in *C. elegans* as PAL8C_1-5 by Kapitonov and Jurka (2003) (Repbase Update, www.girinst.org, see figure 2). The PAL8C and related families form a relatively recent population of several hundreds of nonautonomous transposons in the *C. elegans* genome (Feschotte, unpublished data). Like *Merlin* elements, PAL8C are flanked by 8-bp TSD. Together, these data point to the recent presence and activity of *Merlin*-like transposase(s) in *C. elegans*.

Merlin Elements in *Schistosoma* Flatworms

The putative transposase sequence Merlin_Cb1p from nematode was used as query in Blast searches against the GenBank nucleic acid and protein databases as well as

```
Merlin_Cb1    GGSGCTAAGGTAGTTGGGGGGGGCCACTTTTTTTTTC…
Merlin_Cbm1   GGCGCTAACGTAGTTCGGAGGGCGGCCCAAA
Merlin_Cbm2   GGCGCTACCTTAGTTGGGGGGG
Merlin_Cbm3   GGCGCTAGCGTAGTTGGGAGGG

Merlin_Ce1m   GGTACTAGTCCTAAATCACCCCGCCACTTTTTTTT
PAL8C_1       GGTACTTATGGGTTTCGTTCCCCCCA
PAL8C_2       GGTACTTATGGGTTTCGTTCCCCCCAAAATGATTTTT
PAL8C_3       GGTACTTTTCCTTTTCTACCCCGCATTTTT
PAL8C_4       GGTACTAGTCCTAAATCCCCCCGCCACTTTTCC
PAL8C_5       GGTACTTTTCCGATTTCTGCCCCCCAAAATGTTTTTT

Merlin_Sm1    GGCGAGACTAAAGAACATGGACGA

Merlin_Ci1    GGTAATCCTGCCGCTGGTGACCAGTGAAAATTT…

Merlin_Dr1-1  GGTAACACTTTATTTTGATGGTCCCCCTTTAGATAGT…
Merlin_Dr3m   GGTMACACTTTATTTTGATGGTCCRTTTG
Merlin_Dr2-3′ GGTAACACTTTAGAATAACTATCCGTTATAGGTAGTT…
Merlin_Dr2m   GGTAACACTTTAGAATAACTATCCGTTATAACTAGTT…

Merlin_Hs1    GGATCAATTGAACAATTTGTCYCT

IS1016_Hin    GGGGCTGACGTAGATTAGC
IS1016_Pm     GGGGCTGACGTAGATTATCCCTAAATATC
IS1016_Av     GGCTATGTCTGAGTTAGCT
IS1016_Nm     GGGGCTGTCCTAGATAACTAGG
IS1016_Hp     GGGGCTGTCCTAGATAAC
```

FIG. 2.—Terminal inverted repeats (TIRs) of *Merlin* and IS*1016* elements. The transposon's name is followed by the sequence of the TIRs. Each sequence represents both the 5′ TIR and the reverse complement of the 3′ TIR. Mismatches between the two TIRs are shown as degenerate bases (S = C/G, M = A/C, and R = A/G). Sequences are majority-rule consensus derived from the alignment of multiple copies of each family, except for *Merlin_Ce1m1, Merlin_Dr1-1*, and IS*1016* elements, which are from individual copies extracted from the database, and for *Merlin_Dr2*-3′, which is from the putative 3′ TIR.

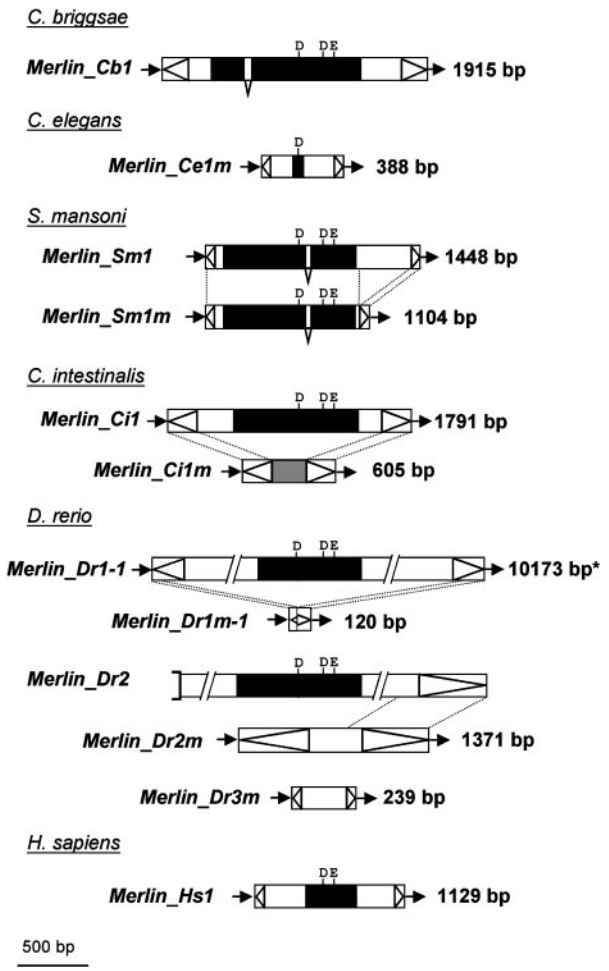

FIG. 3.—Structure of *Merlin* elements. Open triangles represent the TIRs. Transposase coding sequences are depicted as solid black boxes, and the position of the DDE triad is shown. Predicted introns are indicated by an open triangle below the element. TSDs are shown as arrows flanking the element. The relationship of full-length elements to deletion derivatives is shown as dashed lines connecting the homologous regions between the two elements. The bracket interrupting *Merlin_Dr2* marks the end of the contig.

partial sequences and whole-genome sequences (WGS) publicly available through the Web servers of various genomic institutes (see *Methods*). In this way, an abundant family of *Merlin*-like repeats was identified in the WGS assembly of the flatworm *Schistosoma mansoni*. This family, *Merlin_Sm1*, is represented by a 1,448-bp canonical element with perfect 24-bp TIRs (figs. 2 and 3). This element contains a predicted gene made up of two exons separated by a 32-bp intron, which can encode a 294-aa putative transposase (Merlin_Sm1p). This protein has 34% identity and 56% similarity with Merlin_Cb1p.

At least 500 elements have greater than 80% similarity to the *Merlin_Sm1* consensus in the current WGS assembly of *S. mansoni*. This abundant repeat family represents the first family of DNA transposons described from a flatworm species (Brindley et al. 2003). Most of the copies resemble internal deletion derivatives of the *Merlin_Sm1* canonical copy. An abundant and homogeneous population of elements is characterized by a 344-bp deletion (positions 1068 to 1411), whose breakpoints are precisely located between the direct repeat 5′-CATTTAAG-3′ in the *Merlin_Sm1* consensus sequence (*Merlin_Sm1m* in figure 3). This structure suggests that the internal deletion was likely caused by slippage during replication or repair of the element, a process that has been observed previously for other DNA transposons (e.g., Engels et al. 1990; Hsia and Schnable 1996; Rubin and Levy 1997; Yan et al. 1999; Brunet et al. 2002). *Merlin_Sm1* elements examined were generally flanked by an 8-bp DR, which likely represents TSD. This finding is evident when sequences flanking one of the *Merlin_Sm1* copies are compared with an empty paralogous site found elsewhere in the *S. mansoni* genome (fig. 1*B*).

TBlastN searches using Merlin_Cb1p and Merlin_Sm1p against the WGS of *S. mansoni* revealed the presence of other more distant lineages of *Merlin*-like transposases in this species (data not shown, and see phylogenetic analyses below). The alignments produced by TBlastN of these coding sequences with Merlin_Sm1p span at least 100 amino acids but show only 35% to 45% amino acid identity.

A BlastN search that uses *Merlin_Sm1* against the GenBank EST database yields 60 significant hits (e-values $<$ 2e-04) out of approximately 124,000 reads generated by the *S. mansoni* transcriptome project (Verjovski-Almeida et al. 2003). In addition, TBlastN searches that use Merlin_Sm1p against the same database yield 24 hits (e-values $<$ 9e-06) with EST from the related species *S. japonicum*. This result indicates that a closely related family of transposase genes is present and transcribed in *S. japonicum*. EST hits from *S. mansoni* can be divided into two categories, represented in roughly equal proportion: (1) EST reads that include the termini and subterminal region of a *Merlin_Sm1* copy and some of the adjacent flanking DNA and (2) EST reads that derive only from *Merlin_Sm1* sequences and predominantly from the coding region. The first category of ESTs is likely to result from read-through transcription from external promoters into adjacent *Merlin_Sm1* copies. At least 35 hits fall into this category, which suggests that *Merlin_Sm1* elements are frequently inserted in the vicinity of RNA polymerase II promoters and in actively transcribed regions of the *S. mansoni* genome. Whether the second category of ESTs results from the activity of adjacent "host" promoters or from an element's internal promoter is difficult to determine. Interestingly, three *S. mansoni* EST span the 33-bp intron predicted in the transposase gene, but the intron is retained in all of them. Yet, the same intron is removed from most, but not all, EST hits from *S. japonicum* (see figure 1 in Supplemental Material online). Retention of the intron in these transcripts results in the introduction of a premature stop codon in the translated protein and, therefore, would lead to the production of a severely truncated and likely inactive transposase. Speculation that such a truncated transposase could act as a repressor of transposition is tempting because it has been observed in the P element system of *Drosophila* (Laski, Rio, and Rubin 1986).

Table 1
Distribution of *Merlin*-like Elements

Taxa	Abbreviation[a]	Accession Numbers[b]	DB[c]
Eukaryotes			
Nematodes (round worms)			
C. briggsae[d]	Cb	CAE74230	WGS
C. elegans[d] (partially coding)	Ce	AF003130	GEN
Trichuris muris	Tm	BG577820	EST
Meloidogyne incognita	Mi	AW571227	EST
Heterodera glycines	Hg	CA939780	EST
Trematodes (flat worms)			
Schistosoma mansoni[d]	Sm	AL621730	WGS
Schistosoma japonicum	Sj	BU779421	EST
Ascidians (tunicates)			
Ciona intestinalis[d]	Ci	AABS01001688	WGS
Ciona savignyi	Cs	AACT01060475	WGS
Halocynthia roretzi	Hr	AV383621	EST
Boltenia villosa	Bv	AF483024	mRNA
Vertebrates			
Danio rerio[d]	Dr	AL845359	GEN
Xenopus laevis	Xl	BQ737326	EST
Homo sapiens[d] (Partially coding)	Hs	AC091607	GEN
Insects (mosquitoes)			
Anopheles gambiae	Ag	EAA02656	WGS
Anopheles albimanus	Aa	AF293351	GEN
Microsporidians			
Nosema bombycis	Nb	NBU28045	GEN
Stramenopiles (oomycetes)			
Phytophthora sojae	Pj	AAWT24398.b1	GSS
IS*1016* (Bacteria)			
μ-Proteobacteria			
Rickettsia conorii	Rc	AAL03226	GEN
Rickettsia sibirica	Rs	EAA26457	GEN
β-Proteobacteria			
Neisseria meningitides (IS*1016-Nm*)	Nm	AAP44503	GEN
Neisseria gonorrohae	Ng	AAK08026	GEN
δ-Proteobacteria			
Desulfovibrio desulfuricans	Dd	ZP_00130090	GEN
γ-Proteobacteria			
Haemophilus paragallinarum	Hp	AAD01406	GEN
Haemophilus influenzae (IS*1016*)	Hin	CAA42428	GEN
Haemophilus somnus	Hso	ZP_001221	GEN
Mannheimia haemolytica	Mh	AAQ19214	GEN
Azotobacter vinelandii (IS*Azvi1*)	Av	AF322366	GEN
Spirochaetes			
Treponema denticola (ISTde3)	Td	AAS13222	GEN

NOTE.—The names of bacterial IS previously described in the literature are indicated into parentheses.

[a] Species abbreviation.

[b] Accession number for one representative hit per species is given. A more complete list of accession numbers is available upon request.

[c] Type of database where the hit is deposited: GEN = genomic sequences; EST = expressed sequence tag; WGS = whole-genome sequencing; GSS = genomic survey sequence (here, a shotgun sequencing read). Sequences can be accessed at GenBank (www.ncbi.nlm.nih.gov) through their accession number, except sequences from *P. sojae*, which are available through the DOE/JGI (www.jgi.doe.gov).

[d] Species for which complete *Merlin* elements are described in the present study (i.e., elements with both TIRs and apparent coding capacity for a full-length protein), except where otherwise noted (*C. elegans*, and *H. sapiens*).

Merlin Elements in the Sea Squirt *Ciona intestinalis*

TBlastN searches that used Merlin_Cb1p against the WGS assembly of the chordate species *Ciona intestinalis* (Dehal et al. 2002; see *Methods*) revealed at least three distinct ORFs with similarity to the putative transposase from the nematode (e-values $<$ 9e-08). Gene structure prediction, conceptual translation, and multiple alignments of the corresponding genomic regions suggest that each hypothetical protein represents a different lineage of *Merlin*

transposases and is associated with different transposon families (data not shown). We further characterized one of these families, *Merlin_Ci1*, and reconstructed a consensus based on the alignment of a presumably full-length complete copy, with three additional copies containing large ORFs but partially truncated because of gaps in the assembly. The *Merlin_Ci1* consensus sequence is 1,791 bp long with 155-bp TIRs (eight mismatches [figs. 2 and 3]). It is predicted to contain an intronless gene that encodes a 273-aa protein (Merlin_Ci1p), which is 34% identical and 50% similar to Merlin_Cb1p.

Merlin_Ci1 has been recently active in the sea squirt genome because the four copies are approximately 94% to 99% similar to the consensus. The only complete copy available in the WGS, *Merlin_Ci1-1*, has inserted into a genomic region that was presumably duplicated before the insertion. Comparison of the two paralogous regions shows that the integration of the transposon was accompanied by a 9-bp TSD (fig. 1*C*). This finding seems to be a characteristic of the putative *Merlin_Ci1* transposase because all related *Merlin* elements examined in *Ciona* were flanked by a 9-bp TSD (if any).

BlastN searches with the ends of *Merlin_Ci1* revealed the presence of approximately 250 related nonautonomous elements in the *C. intestinalis* WGS (*Merlin_Ci1m* elements). These elements can be grouped into discrete subfamilies that range in size from 400 to 700 bp and have 90% to 99% similarity to each other. Some subfamilies have recently expanded because they are highly homogeneous in length and sequence (>98% similarity, data not shown). A consensus representing the whole population of nonautonomous elements was derived from 20 copies randomly sampled in the WGS. The *Merlin_Ci1m* consensus sequence is 605 bp long with 155-bp TIRs (five mismatches) and has no significant coding capacity (fig. 3). The TIRs are 90% similar to those of the *Merlin_Ci1* consensus, but the internal 295-bp region is of unknown origin (hatched box in figure 3). All *Merlin_Ci1m* elements examined were flanked by a 9-bp TSD. These data suggest that *Merlin_Ci1m* elements have amplified by use of the *Merlin_Ci1* transposase, although the progenitor of the *Merlin_Ci1m* family was probably not a simple deletion derivative of a *Merlin_Ci1* copy. This situation is reminiscent of the relationship of miniature inverted-repeat transposable elements (MITEs) families with their putative autonomous partner transposons, as described in plants and nematodes (Oosumi, Garlick, and Belknap 1996; Feschotte, Zhang, and Wessler 2002; Feschotte, Swamy, and Wessler 2003).

Merlin Elements in Zebrafish

Several sequences with strong similarity to the putative transposase of *Merlin_Cb1* were also identified in several vertebrate species, such as zebrafish, *Xenopus*, and humans. TBlastN searches of the zebrafish genome with the putative transposase Merlin_Ci1p from *C. intestinalis* yields hits to 21 different contigs with e-values ranging from 2.2e-15 to 4.7e-05. All hits span residues 160 to 300 of Merlin_Ci1p, which indicates that they may represent complete or nearly complete *Merlin* transposase ORFs. The corresponding protein sequences display 27% to 89% identity to each other, which suggests that multiple divergent families of *Merlin* transposons coexist in the zebrafish genome. Three families, *Merlin_Dr1, Merlin_Dr2*, and *Merlin_Dr3m*, are described here in more details.

The *Merlin_Dr1* family is represented by a copy of 10,173 bp with 177-bp TIRs (five mismatches) and a central intronless gene, which can potentially encode a 261–aa transposase (figs. 2 and 3). The TIRs are directly flanked by an 8-bp direct repeat (GATATTTA), which likely represents the TSD. The relatively large size of this copy compared with other *Merlin* elements can be attributed, in part, to the nested insertions of at least two uncharacterized transposons upstream and downstream of the predicted transposase gene (data not shown). *Merlin_Dr1m1* is one of the many elements closely related to *Merlin_Dr1-1* in the current zebrafish genome database (see figure 3). It is only 120 bp, yet flanked by an 8-bp putative TSD (CCAATGAT), indicative of a genuine transposition event. *Merlin_Dr1-2* resembles a perfect internal deletion derivative of *Merlin_Dr1-1* because it shares its first 36 and last 84 nucleotides with the 5′ and 3′ ends of *Merlin_Dr1-1*, respectively (fig. 3). In addition, the sequence homology breakpoint between the two elements is located between short direct repeat in *Merlin_Dr1-1*, as typically observed for transposon deletion derivatives (see references above and this study). Based on BlastN searches that used the TIRs of *Merlin_Dr1*, at least 50 different family members are estimated to be present in the zebrafish genome. Compared with the *Merlin_Dr1-1* copy, most of these elements are nonautonomous internal deletion derivatives of variable size but share more than 90% sequence similarity to each other and to *Merlin_Dr1-1*, which suggests the relatively recent activity of this transposon family.

The *Merlin_Dr2* family is primarily represented by a large population of nonautonomous elements, highly homogeneous in size, designated *Merlin_Dr2m*. A consensus for the subfamily was constructed from the alignment of multiple copies extracted from the database. The consensus is 1,371 bp long with 462-bp TIRs (with 96% identity between the TIRs [figs. 2 and 3]). Based on BlastN searches that used *Merlin_Dr2m*, the estimated copy number of the *Merlin_Dr2* family is approximately 500 per genome. Other homogeneous groups of repeats are closely related to the *MerlinDr2* family. For example, one of the repeat consensus sequences automatically generated by the program RECON (Z. Bao and S. Eddy, personal communication) and annotated in the zebrafish assembly as Drr000468 shares 85% similarity with *Merlin_Dr2m* over its entire sequence (691 bp). The Drr000468 repeat likely represents another subfamily of nonautonomous *Merlin* elements in zebrafish.

Although *Merlin_Dr2m* elements have no coding capacity, several lines of evidence suggest that they are mobilized by a *Merlin* transposase. First, the TIRs are very similar to the TIRs of *Merlin_Dr1* and begin with the same 4-bp motif as those of *Merlin_Ci1* (GGTAA [fig. 2]). Second, nine out of 12 copies examined are flanked by an 8-bp direct repeat and evidence that this repeat represents

the TSD created upon insertion is shown in figure 1*D*. Third, a sequence approximately 90% similar to the last 620 bp of the *Merlin_Dr2m* consensus (i.e., approximately its 3′ half) is located approximately 1 kb downstream of one of the putative *Merlin* transposase genes identified by TBlastN searches (fig. 3). This sequence may represent the 3′ terminus of an autonomous *Merlin_Dr2* element responsible for the origin and amplification of the *Merlin_Dr2m* subfamily. Unfortunately, the 5' terminus, which would then match the remaining part of *Merlin_Dr2m*, is not found on this contig, probably because of a gap in the current genome assembly (see figure 3).

BlastN searches with the TIRs of *Merlin_Dr1* identifed *Merlin_Dr3m*, another related family of non-autonomous elements. These elements are extremely homogeneous in size and sequence and a consensus was contructed based on 20 copies extracted from the database. The consensus is 239 bp long and displays 29-bp TIRs (with two mismatches) with strong similarity to those of *Merlin_Dr1* and *Merlin_Dr2m* (figs. 2 and 3). This consensus is 97% identical to the consensus for TDR11, an unclassified repeat family previously identified by Kapitonov and Jurka (2003) (Repbase Update, www.girinst.org). Most *Merlin_Dr3m* copies are flanked by an 8-bp TSD and they are 90% to 96% similar to the consensus. Based on the TIR sequences and the 8-bp TSD, we have few doubts that *Merlin_Dr3m* elements were propagated by a *Merlin*-like transposase. *Merlin_Dr3m* elements are extremely abundant in the zebrafish genome; the estimated number of copies is approximately 8,000. Assuming a random distribution of these elements and a genome size of 1,700 Mb, this copy number implies one copy for every approximately 200 kb. However, 538 copies were detected in 358 different BAC entries in the GenBank database (i.e., 47,972 Mb), which corresponds to an observed density of one copy per approximately 89 kb, which is more than twice the expected density. This finding suggests that the elements are neither randomly nor evenly distributed, but occur in dense clusters on the zebrafish chromosomes.

Relics of *Merlin* Elements in the Human Genome

Several fragments of human coding sequences share significant similarity to the putative *Merlin* transposases from other animal species. For example, a 507-bp fragment on human chromosome 3 (GenBank accession number AC091607, position 58546 to 59052) can be translated in a 169-aa product with 28% identity and 46% similarity to the Merlin_Cb1p transposase. This fragment is bracketed by 21-bp TIRs (two mismatches), which share some similarities with those of other *Merlin* elements and are immediately flanked by an 8-bp direct repeat (AGGAATTA). These features define a human *Merlin*-like transposon of 873-bp that was named *Merlin_Hs1-1*. Approximately 30 related elements can be identified in the draft of the human genome sequence by BlastN search that uses *Merlin_Hs1*. Only five copies seem to have retained their TIRs; the remaining copies seem to have suffered terminal deletions and are lacking one or both of their TIRs. Four of the five copies with recognizable TIRs are flanked by an 8-bp TSD. A consensus sequence for the *Merlin_Hs1* family was tentatively reconstructed from an alignment of 10 copies. The consensus is 1,129 bp with 21-bp perfect TIRs (fig. 2) but is unable to encode a complete and intact transposase because of a major deletion in the 5′ region of the element and because of multiple stop codons that interrupt the reading frame (fig. 3).

The *Merlin_Hs1* copies are 75% to 85% similar to each other at the nucleotide level, which suggests that this *Merlin* family amplified before the divergence of primates but after the divergence of eutherian mammals, an age comparable to most other DNA transposon families found in the human genome (Lander et al. 2001). Consistent with this idea, eight *Merlin* copies out of eight examined were detected at orthologous position in the draft genome sequence of chimpanzee (accession numbers and positions of chimpanzee *Merlin* elements are available upon request). Despite the relatively old age of the human elements, their relationship with *Merlin* family members is evident from similarities in the 21-bp TIR sequence (fig. 2) and the size of the TSD (8 bp). In addition, the short ORFs still detectable in several human elements show significant similarities with *Merlin* putative transposases from other species (see below). These ORFs are likely to represent remnants of a *Merlin*-like transposase once active in a mammalian genome.

Other *Merlin*-like Transposase Sequences in Eukaryotes

Blast searches with Merlin_Cb1p against all databases available at NCBI yield a number of significant hits (e-values $<$ 2e-09 over at least a stretch of 100 aa) with protein sequences from other animals that included two additional nematode species, two species of anopheline mosquitoes, two other ascidians (*Halocynthia roretzi* and *Boltenia villosa*), and the frog *Xenopus laevis* (table 1). All of these *Merlin*-like sequences identified were annotated as unknown or hypothetical proteins, with the exception of AF293351 from *Anopheles albimanus* (e-value = 1e-35) and AF483024 from *B. villosa* (e-value = 2e-09), which were both annotated as putative transposases without further explanation for this classification. In addition, more than 300 hits (e-values $<$ 8e-04) were obtained against a database of approximately 1.4 Gb of shotgun sequence reads from the oomycete *Phytophthora sojae*, available through the DOE/JGI Web site. Given that this database represents eightfold coverage of the genome, probably several dozens of distinct *Merlin*-like transposases are in this species. Unfortunately, the *P. sojae* sequences are short reads that are only available as trace files and not yet assembled into contigs. Thus, no attempt was made to reconstruct full-length *Merlin* transposons from this species. Finally, an additional significant hit (e-value = 5e-12, 42% identity over 100 aa) was with a short DNA fragment randomly isolated from the microsporidian *Nosema bombycis*. *P. sojae* and *N. bombycis* are presently the only nonanimal eukaryotic species represented in the databases for which *Merlin* sequences are detectable.

Hits are particularly abundant in the draft genome sequence of the malaria mosquito *Anopheles gambiae*, in which approximately 90 different *Merlin* transposase

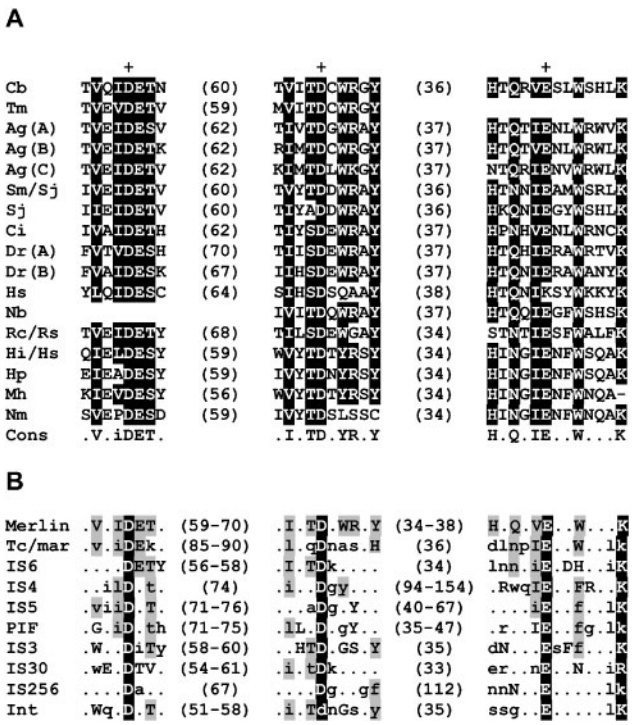

FIG. 4.—Alignment of the potential DDE motifs of *Merlin*/IS*1016* and other transposases. (*A*) Conservation of the D, D, and E residues, their spacing, and the surrounding residues in *Merlin* and IS*1016* transposases. The DDE triad is emphasized by plus signs above the alignment. The spacing (parentheses) refers to the number of residues between the DDE triad. For example, Merlin_Cb1p display a D(60)D(36)E motif. The sequences for *Merlin* are from a majority-rule consensus based on the alignment of a representative subset of the transposases from the same family of elements. When more than one family per genome was identified, a consensus is given for each family (e.g., three families, A, B, and C in *A. gambiae*). Note that the first block is missing for *N. bombycis* because the corresponding nucleotide sequence is not available. (*B*) Comparison of DDE block motifs and spacing of *Merlin*/IS*1016* with various DDE motifs in other transposase superfamilies and in the consensus retroviral integrase core (Int). For each superfamily, only the most conserved residues in each three block are shown, with the invariable or almost invariable residues in capitals and the predominant residues in lower letters. Data for the eukaryotic *PIF* superfamily are from Zhang et al. (2001) and Zhang et al. (2004), data for Tc1/*mariner* (Tc/mar) are from Shao and Tu (2001), and data for other transposases and the retroviral integrase are from Haren, Ton-Hoang, and Chandler (1999) and Chandler and Mahillon (2002). Residues conserved in all proteins are shaded in black (D, D, E, and K), whereas residues shaded in gray are conserved between the *Merlin*/IS*1016* motif and several other DDE motifs. This usage shows that the most conserved residues surrounding the DDE residues of *Merlin*/IS*1016* transposases are also the most conserved residues in other DDE transposase superfamilies. For species symbols, see table 1.

homologs can be detected (e-values $<$ 3e-11). The mosquito proteins can be divided in at least five divergent clades, based on sequence comparison and phylogenetic analysis (data not shown, but see the slightly different DDE motifs in figure 4 and multiple alignment in figure 2 of Supplementary Material online). Proteins from the same clade can share up to 100% identity, whereas interclade identity ranges from 30% to 65% over the most conserved region of the proteins ($\sim$150 aa at the C-terminal end). The majority of these mosquito sequences likely represent pseudogenes because conceptual translations frequently result in premature stop codons and frame shifts. Nonetheless, phylogenetic analysis indicates that these sequences have been rapidly and recently amplified, as judged by the high copy number and sequence homogeneity within certain clades (proteins with 96% to 100% similarity to each other [data not shown]). Together, these features strongly suggest that these sequences represent *Merlin*-like transposase (pseudo)genes and that this group of transposons has effectively colonized the mosquito genome.

Relationships of *Merlin* Elements with the Bacterial IS*1016* Group

As mentioned above, Blast searches that use Merlin_Cb1p yield significant hits with the protein encoded by the insertion sequence IS*1016* from *Haemophilus influenzae* (e-value = 2e-05; 24% identity and 49% similarity over 119 residues) and several related proteins from diverse eubacteria (table 1). Pairwise identities between these eubacterial proteins range from 19% (Rc_NP_360325 versus Ng_AAK08026) to 80% (Hso_ZP_00122161 versus Mh_AY32498). All these proteins appear to represent the transposases encoded by a relatively homogeneous group of IS elements, although only a subset of them have been annotated as such in the databases. None of these IS have been characterized in details, but direct evidence supports the mobility of IS*Azvi1* because this element was discovered as a spontaneous insertion that disrupts the *algU* gene in *A. vinelandii* (Page et al. 2001). The insertion of IS*Azvi1* was accompanied by an 8-bp TSD (see GenBank accession number AF322366). This finding appears to be a characteristic of the entire group because IS*1016* and all the other elements examined in the databases were flanked by an 8-bp DR (data not shown). I refer to this ensemble of bacterial IS as the IS*1016* group.

A multiple alignment of the IS*1016* and *Merlin* putative transposases was constructed by ClustalX. The most conserved region of these proteins is a region of approximately 150 aa closer to the C-terminal ends (figure 2 in Supplementary Material). This region is marked by three highly conserved blocks of residues, whose spacing and composition are reminiscent of the so-called DDE catalytic motif found in a variety of transposases, retroviral integrases, and other recombinases (fig. 4). Generally, each of the three acidic residues is embedded within a small block of other highly conserved residues, and the three blocks are spaced similarly in different groups of recombinases (Doak et al. 1994; Capy et al. 1998; Haren, Ton-Hoang, and Chandler 1999; Chandler and Mahillon 2002; Robertson 2002). The exact same pattern is seen in a multiple alignment of *Merlin* and IS*1016* proteins, with the D, D, and E residues and several surrounding residues invariable or strongly prevalent (figure 4*A*, and figure 2 in Supplementary Material online). In addition, the most conserved residues that surround the DDE motif of *Merlin*/IS*1016* transposases match some of the most conserved residues that surround the DDE motif of various transposase/integrase superfamilies (fig. 4*B*). The motif is D(59–70)D(34–38)E in the *Merlin*/IS*1016* transposases and is most similar in sequence and spacing to those of IS*6*, IS*30*,

IS*3* or the retroviral integrases (see figure 4*B* and alignments from Haren, Ton-Hoang, and Chandler [1999] and from Chandler and Mahillon [2002]).

Discussion

A Newly Recognized Group of DNA Transposons in Various Animal Genomes

Several novel families of DNA transposons, called *Merlin*, were detected by computational analysis in a wide range of animal genomes and shown to share common structural and sequence features. First, *Merlin* elements possess TIRs that range in length from 24 to 462 bp and display sequence similarities within species and across species (note the conservation of the terminal 5′-GG-3′ dinucleotide for all the families [fig. 2]). Second, most *Merlin* elements are flanked by an 8-bp direct repeat, or a 9-bp direct repeat in the case of the sea squirt elements. Identification of paralogous sites devoid of elements but retaining the 8-bp or 9-bp sequence at the insertion site (fig. 3) show that the flanking direct repeat result from duplication of the target sequence upon insertion, a characteristic of transposon insertions. This finding also provides evidence for the past mobility of the elements. Third, many elements identified in this study have coding capacity for approximately 300-aa proteins that have strong similarities with each other and with transposases encoded by a group of bacterial IS that I refer to as the IS*1016* group. Together, these data point to the existence of a previously unrecognized group of evolutionary-related DNA transposons that have colonized diverse animal genomes.

A New Allied Superfamily of Eukaryotic and Prokaryotic DNA Transposases

As mentioned above and shown in this study, *Merlin* and IS*1016* transposons display broad sequence similarities in their encoded proteins (figure 2 in Supplementary Material online). Sequence similarities are particularly compelling in the C-terminal halves of the proteins (the last approximately 150 amino acids), over 25% identity and 40% similarity are commonly observed in pairwise comparisons between animal and eubacterial sequences. This region includes a motif strongly similar to the DDE motif found in the catalytic region of many transposases, integrases, and recombinases (figure 4, and see Doak et al. [1994], Capy et al. [1998], Haren, Ton-Hoang, and Chandler [1999], Chandler and Mahillon [2002], and Robertson [2002]). Beyond this motif and a few surrounding residues also well conserved in other transposases (fig. 4*B*), no obvious similarities with any previously established superfamily of transposases are evident. These data suggest that *Merlin* and IS*1016* proteins belong to a distinct monophyletic group of transposases that was differentiated from other transpoases before the divergence of eukaryotes and prokaryotes.

Besides similarities in coding sequences, *Merlin* and IS*1016* transposons also exhibit resemblances in their noncoding features, such as sequence similarities in their TIRs and a characteristic 8-bp TSD (except *Merlin_Ci* elements, which have a 9-bp TSD). The fact is well established for many DNA transposons that transposition is initiated by the recognition and binding of the transposase to the TIR sequences (for review see Craig et al. [2002]). Additional contacts between the transposon termini and the transposase also occur during the subsequent steps of the transposition process, the cleavage and strand-transfer reactions (e.g., Mizuuchi and Adzuma 1991; Beall and Rio 1998; Lee and Harshey 2003). Thus, conservation of the terminal nucleotides in different transposons is likely to reflect, in part, the common ancestry of their transposases as well as conserved biochemical processes during the transposition reaction. Similarly, the length of the TSD is determined by the catalytic activites of the transposases, acting as an endonuclease at the target DNA (for review see Craig et al. [2002]). Consequently, both TIR sequences and TSD length are expected to coevolve with the transposase's sequences (Lampe, Walden, and Robertson 2001; Naumann and Reznikoff 2002; Feschotte, Swamy, and Wessler 2003). Hence, the structural resemblances in TIRs and TSD of *Merlin* and IS*1016* elements provide further evidence for the common ancestry and evolutionary relationship of their transposases. On the basis of these data, I propose that *Merlin* and IS*1016* elements are the founding members of a newly recognized assemblage of eukaryotic and prokaryotic DNA transposons with a common ancestor, the *Merlin*/IS*1016* superfamily.

Distribution of *Merlin* Elements in Eukaryotes

This case is the fourth example of a close evolutionary link between eukaryotic DNA transposons and prokaryotic IS. Previous examples are the Tc1/*mariner*, PIF/*Harbinger*, and *Mutator* superfamilies of eukaryotic transposons, which show relationship to the IS*630*, IS*5*, and IS*256* prokaryotic groups, respectively (Doak et al. 1994; Eisen, Benito, and Walbot 1994; Kapitonov and Jurka 1999; Zhang et al. 2001). Members of each three superfamilies have been identified in a very wide range of eukaryotic lineages that include animals, fungi, plants, and some protozoans. This broad distribution is consistent with the ancient origin of these superfamilies (i.e., before the divergence of the eukaryotes and eubacteria) and their diversification in various branches of the eukaryotic tree. In comparison to these and other DNA transposon superfamilies (e.g., hAT), *Merlin* elements appear to have a much narrower and patchier distribution in eukaryotes (summarized in table 1). For example, no *Merlin*-like sequences could be identified in plant or fungal species, despite the increasingly large amount of genomic sequence available in the databases for several species (e.g., *Arabidopsis*, rice, *Brassica*, *Chlamydomonas reinhardti*, *Neurospora crassa*, *Aspergillus*, *Cryptococcus neoformans*, and yeasts). The only sequences with similarity to *Merlin*/IS*1016* transposases detected outside the animal kingdom are from the microsporidian *N. bombycis* and from the oomycete *P. sojae*. Microsporidians are most closely related to fungi, and oomycetes belong to the stramenopiles, a phylum that appears to have emerged relatively early in eukaryotic evolution (Baldauf et al.

2000). This finding indicates that *Merlin* elements are not exclusive to the animal kingdom and that members of the superfamily are likely to be discovered in a broader range of eukaryotes, albeit with a somewhat patchy distribution.

Several explanations can be advanced for the apparent patchy distribution of *Merlin* in eukaryotes. First, *Merlin*-like elements are possibly present in plants or fungi represented in the databases but have escaped detection by Blast searches because of sequence divergence. This explanation seems unlikely, however, because sequence similarities are readily detected by Blast between the animal *Merlin* proteins and diverse bacterial IS*1016* transposases. A second possibility is that *Merlin*-like elements once colonized the genome of plants and fungi but are now partially or completely extinct from these kingdoms or at least from the species represented in the databases. After all, *Merlin* elements exhibit a relatively patchy distribution even among the animals represented in the databases. For example, they are abundant and have been recently active in anopheline mosquitoes, but they are not detected in any other insect genomes although several are now completely sequenced (*D. melanogaster*, *D. pseudobscura*, and *Apis melifera*). Similarly, several *Merlin* families are readily identified in zebrafish, but none can be detected in the WGS of the pufferfish *Takifugu rubripes* and only a few remnants are apparent in the human and chimp genomes (but none in mouse, rat, or dog). Such a patchy distribution is not uncommon for other transposable elements (e.g., Robertson and Lampe 1995; Capy et al. 1998; Casavant et al. 2000; Goodwin and Poulter 2004), possibly in part because individual TE copies are generally present at low frequency in populations, so that whole families can become extinct from a species in a relatively short period of time (Charlesworth, Sniegowski, and Stephan 1994; Capy et al. 1998; Casavant et al. 2000). A third possibility is that horizontal transfers could also contribute to the patchy distribution of *Merlin* elements in eukaryotes. Several examples of horizontal transfers of TEs have been documented between animal species, in particular for the class 2 transposons, and these events are believed to be essential for the propagation and survival of these elements (for reviews see Capy et al. [1998] and Robertson [2002]). It may be that *Merlin* elements are also propagated through horizontal transfers and that such transfers occur more readily in some animals and in single-celled parasites (such as *Nosema* or *Phytophtora*) than in fungi and plants.

Applications and Concluding Remarks

With the advent of genome sequencing, TEs are increasingly seen to represent the predominant component of eukaryotic genomes. Identifying and classifying TEs in genomic sequences is a critical, albeit challenging, step of the annotation process. Currently, most TEs are identified on the basis of homology-based searches and classified according to similarities, predominantly in their coding sequences, with elements previously described in the same or another species. The identification of a novel superfamily of elements and their encoded transposases in diverse eukaryotic genomes is, thus, important for the annotation of the ongoing and future genome projects. As shown here for several species, *Merlin* families can amplify to substantial copy numbers (>500 per genome), which can represent a significant fraction of the genome content.

Several *Merlin* families display signs of recent transposition activity in some species. First, many families include multiple members with very high sequence and structural homogeneity. The genomes of *C. briggsae*, *S. mansoni*, and *C. intestinalis* each harbor identical or almost identical copies of different *Merlin* families with different flanking sequences, which indicate that they result from very recent transposition events in each of these species rather than segmental genomic duplications. Second, some *Merlin* TEs display long and apparently intact open reading frames that encode a potentially active source of transposase as well as the *cis*-sequences necessary for transposition, such as perfect TIRs (fig. 2). Evidence also supports transcription of *Merlin* transposases in some of these species (e.g., in flatworms [see table 1]). These data suggest that *Merlin* transposons may be currently able to transpose autonomously in some of these species. Identifying novel autonomous DNA transposons may allow the development of useful molecular tools, such as DNA delivery vectors and mutagenesis systems. This application may be particularly relevant for medically important species for which transposon-based tools are still lacking, such as the human parasite *S. mansoni*. *Merlin*_Sm1 family is the first DNA transposon family described in any flatworm species (Brindley et al. 2003), and its coding potential, transcriptional activity, and recent amplification makes it an outstanding candidate for further characterization.

Acknowledgments

The author expresses his gratitude to Sue Wessler for providing facilities in support for this study and continuous encouragement. I thank Ellen Pritham for suggestions on the manuscript and stimulating discussions. I also thank Eddie Holmes and two anonymous reviewers for their constructive comments. The cost of this publication was supported by funds from the Department of Biology, the University of Texas at Arlington. Some of the sequence data used in this work are from unpublished and/or unfinished projects produced by the Sanger Institute (http://www.sanger.ac.uk), The Institute for Genomic Research (www.tigr.org) and the DOE Joint Genome Institute (www.jgi.doe.gov) and are available through their web sites.

Literature Cited

Aparicio, S., J. Chapman, E. Stupka et al. (41 co-authors) 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297**:1301–1310.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972–977.

Beall, E. L., and D. C. Rio. 1998. Transposase makes critical contacts with, and is stimulated by, single-stranded DNA at the P element termini *in vitro*. EMBO J. **17**:2122–2136.

Bowen, N. J., and I. K. Jordan. 2002. Transposable elements and the evolution of eukaryotic complexity. Curr. Issues Mol. Biol. **4**:65–76.

Brindley, P. J., T. Laha, D. P. McManus, and A. Loukas. 2003. Mobile genetic elements colonizing the genomes of metazoan parasites. Trends Parasitol. **19**:79–87.

Brunet, F., T. Giraud, F. Godin, and P. Capy. 2002. Do deletions of mos1-like elements occur randomly in the drosophilidae family? J. Mol. Evol. **54**:227–234.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1998. Dynamics and evolution of transposable elements. Springer-Verlag, Austin, Texas.

Casavant, N. C., L. Scott, M. A. Cantrell, L. E. Wiggins, R. J. Baker, and H. A. Wichman. 2000. The end of the LINE?: lack of recent L1 activity in a group of South American rodents. Genetics **154**:1809–1817.

Chandler, M., and J. Mahillon. 2002. Insertion sequences revisited. Pp. 305–366 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, DC.

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371**:215–220.

Caenorhabditis elegans Genome Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science **282**:2012–2018.

Craig, N. L., R. Craigie, M. Gellert, and A. M. Lambowitz. 2002. Mobile DNA II. American Society for Microbiology Press, Washington, D.C.

Dehal, P., Y. Satou, R. K. Campbell et al. (87 co-authors) 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science **298**:2157–2167.

Doak, T. G., F. P. Doerder, C. L. Jahn, and G. Herrick. 1994. A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA **91**:942–946.

Eisen, J. A., M. I. Benito, and V. Walbot. 1994. Sequence similarity of putative transposases links the maize mutator autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res. **22**:2634–2636.

Engels, W. R., D. M. Johnson-Schlitz, W. B. Eggleston, and J. Sved. 1990. High-frequency P element loss in *Drosophila* is homolog dependent. Cell **62**:515–525.

Feschotte, C., N. Jiang, and S. R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. **3**:329–341.

Feschotte, C., L. Swamy, and S. R. Wessler. 2003. Genome-wide analysis of *mariner*-like transposable elements in rice reveals complex relationships with *Stowaway* MITEs. Genetics **163**:747–758.

Feschotte, C., X. Zhang, and S. Wessler. 2002. Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons. Pp. 1147–1158 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, DC.

Galagan, J. E., S. E. Calvo, K. A. Borkovich et al. (55 co-authors) 2003. The genome sequence of the filamentous fungus Neurospora crassa. Nature **422**:859–868.

Goodwin, T. J., and R. T. Poulter. 2004. A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. **21**:746–759.

Haren, L., B. Ton-Hoang, and M. Chandler. 1999. Integrating DNA: transposases and retroviral integrases. Annu. Rev. Microbiol. **53**:245–281.

Holt, R., A. G. M. Subramanian, A. Halpern et al. (41 co-authors) 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science **298**:129–149.

Hsia, A. P., and P. S. Schnable. 1996. DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, *MuDR*. Genetics **142**:603–618.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **16**:418–420.

Kapitonov, V. V., and J. Jurka. 1999. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. Genetica **107**:27–37.

———. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. USA **100**:6569–6574.

Kidwell, M. G., and D. R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evol. Int. J. Org. Evol. **55**:1–24.

Lampe, D. J., K. K. Walden, and H. M. Robertson. 2001. Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. Mol. Biol. Evol. **18**:954–961.

Lander, E. S. L. M. Linton B. Birren et al. (271 co-authors) 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Laski, F. A., D. C. Rio, and G. M. Rubin. 1986. Tissue specificity of *Drosophila* P element transposition is regulated at the level of mRNA splicing. Cell **44**:7–19.

Lee, I., and R. M. Harshey. 2003. Patterns of sequence conservation at termini of long terminal repeat (LTR) retrotransposons and DNA transposons in the human genome: lessons from phage Mu. Nucleic Acids Res. **31**:4531–4540.

Mizuuchi, K., and K. Adzuma. 1991. Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. Cell **66**: 129–140.

Naumann, T. A., and W. S. Reznikoff. 2002. Tn5 transposase with an altered specificity for transposon ends. J. Bacteriol. **184**:233–240.

Oosumi, T., B. Garlick, and W. R. Belknap. 1996. Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans*. J. Mol. Evol. **43**:11–18.

Page, W. J., A. Tindale, M. Chandra, and E. Kwon. 2001. Alginate formation in *Azotobacter vinelandii* UWD during stationary phase and the turnover of poly-beta-hydroxybutyrate. Microbiology **147**:483–490.

Robertson, H. M. 2002. Evolution of DNA transposons. Pp. 1093–1110 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. American Society for Microbiology Press, Washington, DC.

Robertson, H. M., and D. J. Lampe. 1995. Distribution of transposable elements in arthropods. Annu. Rev. Entomol. **40**:333–357.

Rubin, E., and A. A. Levy. 1997. Abortive gap repair: underlying mechanism for *Ds* element formation. Mol. Cell. Biol. **17**: 6294–6302.

SanMiguel, P., A. Tikhonov, Y.-K. Jin et al. (8 co-authors) 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science **274**:765–768.

Shao, H., and Z. Tu. 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.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

Verjovski-Almeida, S., R. DeMarco, E. A. Martins et al. (34 co-authors) 2003. Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*. Nat. Genet. **35**:148–157.

Vicient, C. M., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav, E. Nevo, and A. H. Schulman. 1999. Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell **11**:1769–1784.

Waterston, R., H. K. Lindblad-Toh, E. Birney et al. (222 co-authors) 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520–562.

Yan, X., I. M. Martinez-Ferez, S. Kavchok, and H. K. Dooner. 1999. Origination of Ds elements from Ac elements in maize: evidence for rare repair synthesis at the site of Ac excision. Genetics **152**:1733–1740.

Zhang, X., C. Feschotte, Q. Zhang, N. Jiang, W. B. Eggleston, and S. R. Wessler. 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. USA **98**:12572–12577.

Zhang, X., N. Jiang, C. Feschotte, and S. R. Wessler. 2004. Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. Genetics **166**:971–986.

Edward Holmes, Associate Editor

Accepted June 3, 2004