Mol Gen Genomics (2003) 270: 173–180
DOI 10.1007/s00438-003-0909-0

ORIGINAL PAPER

A. Sarkar · C. Sim · Y. S. Hong · J. R. Hogan
M. J. Fraser · H. M. Robertson · F. H. Collins

Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences

Received: 6 April 2003 / Accepted: 28 July 2003 / Published online: 29 August 2003

Abstract *piggyBac* is a short inverted-repeat-type DNA transposable element originally isolated from the genome of the moth *Trichoplusia ni*. It is currently the gene vector of choice for the transformation of various insect species. A few sequences with similarity to *piggyBac* have previously been identified from organisms such as humans (*Looper*), the pufferfish *Takifugu rubripes* (*Pigibaku*), *Xenopus* (*Tx*), *Daphnia* (*Pokey*), and the Oriental fruit fly *Bactrocera dorsalis*. We have now identified 50 *piggyBac*-like sequences from publicly available genome sequences and expressed sequence tags (ESTs). This survey allows the first comparative examination of the distinctive *piggyBac* transposase, suggesting that it might contain a highly divergent DDD domain, comparable to the widespread DDE domain found in many DNA transposases and retroviral integrases which consists of two absolutely conserved aspartic acids separated by about 70 amino acids with a highly conserved glutamic acid about 35 amino acids further away. Many *piggyBac*-like sequences were found in the genomes of a phylogenetically diverse range of organisms including fungi, plants, insects, crustaceans, urochordates, amphibians, fishes and mammals. Also, several instances of "domestication" of the *piggyBac* transposase sequence by the host genome for cellular functions were identified. Novel members of the *piggyBac* family may be useful in genetic engineering of many organisms.

Keywords *piggyBac* · Transposable element · Transposase · TTAA-specific · DDE domain

Communicated by G. P. Georgiev

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00438-003-0909-0.

A. Sarkar · C. Sim · Y. S. Hong · J. R. Hogan · M. J. Fraser
F. H. Collins (✉)
Center for Tropical Disease Research and Training,
Department of Biological Sciences, University of Notre Dame,
Notre Dame, IN 46556-0369, USA
E-mail: frank.h.collins.75@nd.edu
Tel.: +1-574-6318045
Fax: +1-574-6313996

H. M. Robertson
Department of Entomology,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA

Introduction

Transposable elements constitute an important part of most eukaryotic genomes (Craig et al. 2002). One apparent exception is the genome of the malaria parasite *Plasmodium falciparum* (Gardner et al. 2002). Class II transposable elements move via a DNA intermediate and have been extensively adapted for use as gene vectors (Horn and Wimmer 2000).

The *piggyBac* transposon from the cabbage looper moth *Trichoplusia ni* is a 2472-bp Class II transposon with 13-bp inverted terminal repeats (ITRs) (Cary et al. 1989; Fraser et al. 1996). It contains a single ORF in a 2.1-kb transcript , which encodes a 594-amino acid transposase that mediates cut-and-paste excision and reinsertion of the element into a TTAA target site in the genome, causing a target-site duplication that flanks the element (Fraser et al. 1995, 1996). The sequence originally published contained several errors that led to a frameshift in this long ORF, and the corrected version is available from GenBank as Accession No. J04364.2. A sequence related to the *T. ni piggyBac* element has since been identified in the genome of the tephritid fruit fly *Bactrocera dorsalis* by PCR (Handler and McCombs 2000). Neither the transposase nor the TTAA target site of these elements show any obvious similarity to those of other, better known Class II transposon families; thus *piggyBac* has been interpreted as the type member of a new Class II transposon family, the *piggyBac* family (Robertson 2002).

The *T. ni piggyBac* element is particularly interesting because it is mobile in the genomes of a wide range of

insect species tested. It has been used as a transformation vector in insects belonging to the orders Diptera, such as *Drosophila melanogaster* (Handler and Harrell 1999), *Bactrocera dorsalis* (Handler and McCombs 2000), *Aedes aegypti* (Lobo et al. 2002), *Anopheles stephensi* (Nolan et al. 2002) and *An. gambiae* (Grossman et al. 2001); Lepidoptera, including *Pectinophora gossypiella* (Peloquin et al. 2000) and *Bombyx mori* (Tamura et al. 2000), and Coleoptera (*Tribolium castaneum*; Horn and Wimmer 2000). Indeed *piggyBac* is currently the most widely used gene vector for insects of medical and economic importance (Fraser 2000; Handler 2002). Recently, *piggyBac* has been further developed as a gene tagging and enhancer trapping system for *Drosophila* and other arthropods (Horn et al. 2003).

The activity of *piggyBac* in such a diverse group of insects suggests that the *piggyBac* transposon family might have a phylogenetic distribution that extends beyond the insect families Diptera and Lepidoptera (Fraser 2000). However, in a review of the molecular evolution of DNA transposons, one of us (Robertson 2002) was unable to find relatives of *piggyBac* in the genome sequences of the nematode *Caenorhabditis elegans* or the fruit fly *D. melanogaster* that became available in 2000. Nevertheless, several indications of such a broader distribution have been published recently. First, the International Human Genome Sequencing Consortium (2001) noted that, in addition to *Looper*—a short consensus of multiple diverged sequences in the human genome thought to represent an ancient *piggyBac* relative (see RepBase v7.7; Jurka 2000), there are five putative genes in the human genome that appear to code for proteins with significant similarity to the *piggyBac* transposase, which are thought to have been derived from *piggyBac* relatives in ancient mammalian genomes (International Human Genome Sequencing Consortium 2001, Table 13). Second, the recently published draft sequence of the *Takifugu rubripes* (pufferfish) genome was reported to contain a *piggyBac*-like transposon called *Pigibaku* (Aparicio et al. 2002). Third, the recently completed draft sequence of the genome of the African malaria mosquito *An. gambiae* was reported to contain five copies of *piggyBac* relatives (Holt et al. 2002). Fourth, a distant relative of *piggyBac* called *Pokey* was recently described from two copies sequenced from the crustacean *Daphnia pulicaria* (Penton et al. 2002). These authors also noted three of the human sequences putatively derived from *piggyBac*-like transposons.

We first sought to confirm and expand these observations on the published human, *T. rubripes*; *Ciona intestinalis* and *An. gambiae* genomes, and then extended them to the other publicly available completed, draft, or incomplete genome sequences and expressed sequence tags (ESTs). This approach led to the identification of many *piggyBac* -related sequences in phylogenetically diverse organisms, and also uncovered instances of "domestication or co-option of the *piggyBac* transposase sequence by the host genome for cellular functions.

Materials and methods

Database searches and identification of *piggyBac*-like sequences

We searched all available sequence databases, and particularly genome databases, using the TBLASTN algorithm, initially with the canonical *T. ni piggyBac* transposase as the query, using the NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST). Novel *piggyBac*-like sequences identified in these searches were then used to query the databases again to identify more *piggyBac*-like sequences. Because only the original *T. ni piggyBac*, *Pokey* , two *D. melanogaster* genes, and some of the human sequences are annotated as encoding proteins, PSI-BLASTP searches for highly divergent sequences (Altschul et al. 1997) were not particularly informative. Thus it is possible that we have missed some highly divergent sequences that might eventually be recognized upon thorough annotation of the current and additional genome sequences. Conserved domain searches were carried out using the NCBI CDS server (http://www.ncbi.nlm.nih.gov/cdd/cdd.shtml).

Phylogenetic analyses

Transposase sequences were aligned using CLUSTALX 1.8 (Thompson et al. 1997), distances were corrected with the maximum-likelihood model in TREE-PUZZLE v5.0 (Schmidt et al. 2002) using the BLOSUM 62 amino acid substitution matrix, and phylogenetic analyses were carried out using PAUP*4.0b10 (Swofford 2002) and TREE-PUZZLE.

Results and discussion

piggyBac- like sequences in the human genome

The consensus sequence deposited in RepBase for the human *Looper* element is based on a few highly divergent sequences available from a few early human BAC clone sequences, and its potential protein product is only 278 amino acids long with 25% similarity to the N-terminal segment of the *T. ni piggyBac* transposase. Efforts to extend this consensus towards the C-terminus using the currently available, essentially complete, human genome sequence have not been successful. Because there is a high level of divergence between the approximately 115 fragmented copies that are roughly evenly distributed throughout the human genome (NCBI Build 31), we were unable to recover the original *Looper* consensus ORF. *Looper* therefore represents the remains of a *piggyBac*-like transposon that was active in mammalian genomes at least as long ago as the mouse-human divergence—variously dated from 70–100 Myr ago (Kumar and Hedges 1998; Mouse Genome Sequncing Consortium 2002)—and probably considerably longer; and perhaps most copies represent internally deleted versions. Phylogenetic analyses of only the N-terminal regions of the transposase do not reliably ally *Looper* with any other particular member of the family. Only four *Looper* copies are evident in the mouse draft genome sequence at NCBI. Jerzy Jurka has added consensus sequences for six such transposons from the zebrafish genome, called *LOOPERN (1–6)_DR* to RepBase, but these too are defective.

None of these sequences are included in our phylogenetic analyses.

Three of the five sequences identified as potentially "domesticated" *piggyBac* relatives in the human genome by the International Human Genome Sequencing Consortium (2001) are the same as the three identified by Penton et al. (2002) (all are available as NCBI RefSeqs; National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html), while the other two correspond to related pseudogenes (see Table 1). One of these three is referred to as cerebral protein 4 (HUCEP-4) in GenBank (Accession No. D88259), but we have formally named this gene *PGBD1* for *piggyBac - derived 1* (Table 1) (see the HUGO Human Gene Nomenclature Database at http://www.gene.ucl.ac.uk/nomenclature/). *PGBD1 / HUCEP-4* appears to be a chimeric or fusion gene, with five upstream exons encoding 290 amino acids that include a LER or SCAN domain commonly found in the N-terminal regions of zinc-finger transcription factors, followed by a single large exon encoding the *piggyBac* -derived portion of 519 amino acids, from which the extreme C-terminal cysteine-rich region has been lost (see below). We recognized two additional human genes that are apparently derived from *piggyBac*-like transposon sequences, giving a total of five. We reconstructed them all using a combination of genomic, cDNA and EST sequences, and have named these genes *PGBD1–5* (Table 1). *PGBD2* consists largely of a single long coding exon, with one intron near the extreme N-terminus and another in the 5′ UTR, while *PGBD3* has a single ORF and a 5′ UTR intron. *PGBD2* and *3* are related to the *piggyBac* domain of *PGBD1*, and in addition there are at least four apparent pseudogenes related to *PGBD3*. *PGBD4* and *5* are quite distinct (see Fig. 1). *PGBD4* appears to comprise a single ORF/exon without introns, and appears to be derived from the *MER75* transposon, which is now represented in the human genome by many internally deleted copies (Smit 1999). *PGBD5* is the most highly divergent of these human *piggyBac* -derived genes, and, remarkably, has acquired eight introns, roughly evenly spaced along the length of the gene, most of which are shared with both mouse and pufferfish orthologs. The presence of a clear pufferfish ortholog for *PGBD5*, but not *PGBD1–4*, indicates the antiquity of this domesticated gene, which must have gained these introns before the teleost/tetrapod split ~450 Myr ago (Kumar and Hedges 1998). Surprisingly, the mouse genome contains only single pseudogene versions of both *PGBD1* and *2*, and there are no orthologs of *PGBD3* and *4*, nor did we find any mouse-specific *PGBD* genes (Mouse Genome Sequencing Consortium 2002).

piggyBac- like sequences in pufferfish

A paracio et al. (2002) described a *piggyBac*-like transposon sequence in the *T. rubripes* genome (we will use *Takifugu* rather than *Fugu* as the genus name), called *Pigibaku*, which they reported to be present in approximately 200 copies in the genome (see their supplementary online text). We obtained the sequence for *Pigibaku* deposited in RepBase and found that it only corresponds to three long sequences in the genome, each of them highly defective in coding capacity. Searches with this and the *T. ni piggyBac* transposase indicated that in fact there are multiple, quite divergent, *piggyBac* transposon relatives in this genome, each of which is represented by just one to four relatively long sequences. We propose to name these sequences *Pigibaku1–9*, although the best "consensus" sequences that can be generated for *Pigibaku6–9* are too short to be useful in phylogenetic analyses. They form two clusters in our phylogenetic tree of *piggyBac*-like sequences (Fig. 1), and related sequences for each cluster are present in the partial genome sequences of the freshwater pufferfish *Tetraodon nigriviridis* and the zebrafish *Danio rerio*, indicating that these transposons were either inherited vertically amongst teleost fish or have been horizontally transferred among teleosts. Nevertheless, all of them in *T. rubripes* appear to be defective and are presumably now inactive. In addition, we identified five

Table 1 *piggyBac-derived* (PGBD) genes and pseudogenes in the human and mouse genomes

Gene name	RefSeq[a]	Number of amino acids	IHGSC No.[a]	Penton et al. No.[a]	Mouse
PGBD1	NP_115896	809 [b]	EST_1963278	NP_115896	Pseudogene
PGBD2	NP_733843	592	–	–	Pseudogene
PGBD3	NP_736609	593	pID_6453533	T42689/XP_046148	Absent
PGBD3P1	AC090017	–	–	–	
PGBD3P2	AC010374	–	EST_3594004	–	–
PGBD3P3	AC006065	–	BAC_4309921	–	–
PGBD3P4	AC114774	–	–	–	
PGBD4	NP_689808	585	EST_4073914	BAB71379/XP_091120	Absent
PGBD5	NP_078830	554	–	–	*pgbd5*

[a]NCBI RefSeqs are for the functional PGBD genes. The four *PGBD3P* pseudogenes are identified by the GenBank accession numbers of BAC clones that contain them. The numbers listed in Columns 4 and 5 are those assigned to sequences recognized by the International Human Genome Sequencing Consortium (2001) and by Penton et al. (2002), respectively (see text for details)

[b]*PGBD1 / HUCEP-4* is a chimeric gene with five upstream exons encoding 290 amino acids spliced to a single long exon encoding the*piggyBac* -derived portion of 519 amino acids, from which the cysteine-rich C-terminus is missing

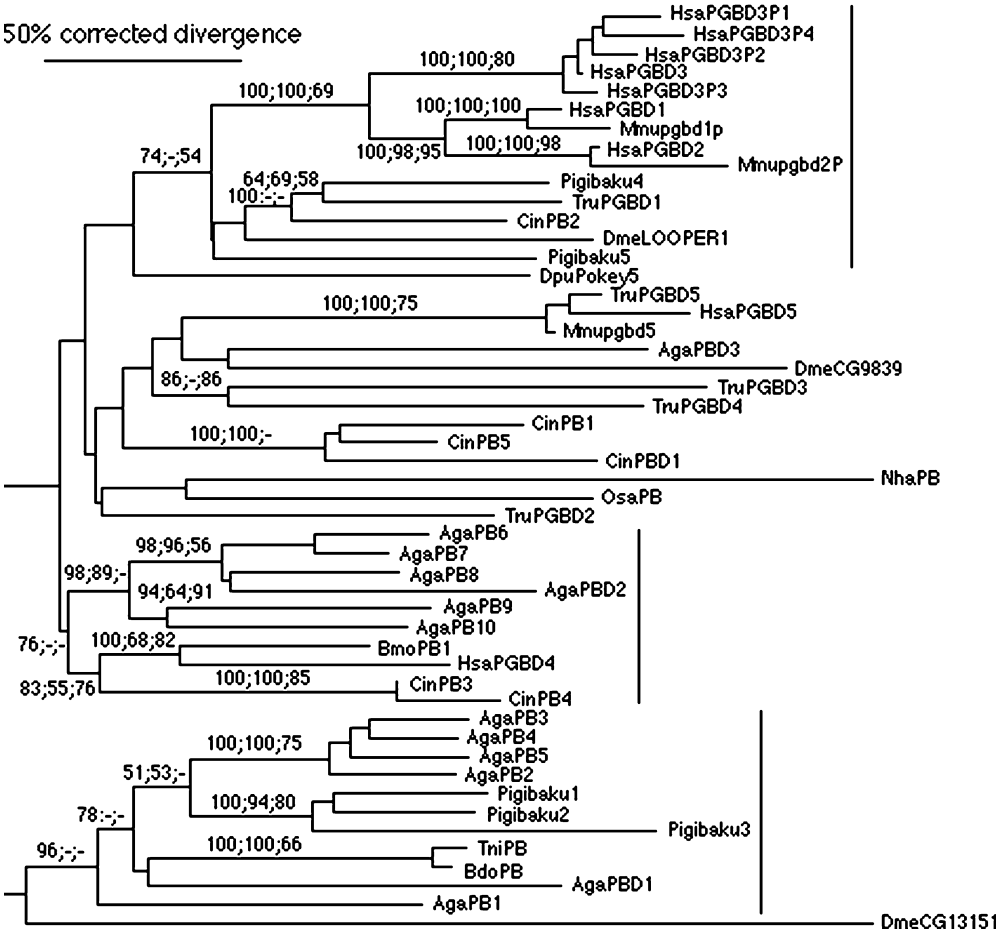

Fig. 1 Phylogenetic tree of *piggyBac*-like sequences. This minimum evolution tree is based on amino acid distances corrected with the maximum-likelihood model in TREE-PUZZLE v5.0 (Schmidt et al. 2002) using default settings, except that the BLOSUM 62 matrix was employed. It was constructed using the Heuristic Search option of PAUP*4.0b10 (Swofford 2002). Support for branches from three additional analyses is shown; 1000 replications of uncorrected distance and parsimony analyses, and maximum likelihood values from 10,000 quartet-puzzling steps in TREE-PUZZLE. The tree is rooted at the midpoint in the absence of a suitable outgroup. Abbreviations: Aga, *Anopheles gambiae*; Bdo, *Bactrocera dorsalis*; Bmo, *Bombyx mori*; Cin, *Ciona intestinalis*; Dme, *Drosophila melanogaster*; Dpu, *Daphnia pulicaria*; Hsa, *Homo sapiens;* Mmu, *Mus musculus*; Nha, *Nectria haematococca*; Osa, *Oryza sativa*; Tni, *Trichoplusia ni*; Tru, *Takifugu rubripes*; PB, *piggyBac*-like transposase (the *pigibaku* name for the *Takifugu rubripes* elements is retained); PGBD, *piggyBac* -derived gene/protein (PBGD is used for the human and other vertebrate genes because PBD was already employed in the HUGO nomenclature database)

apparently functional genes in this genome that were apparently derived from *piggyBac*-like transposons. All of these are unique sequences in the *Takifugu* genome, and so do not appear to be single potentially active copies of *piggyBac*-like transposons. They are most likely to be bona-fide domesticated genes. One of these, *TruPBD1*, clusters with the *Pigibaku4* sequence (Fig. 1), but is so divergent that it must have been derived not from a particular copy of *Pigibaku4*, but from a related transposon whose other copies have long since been deleted from this compact genome. *TruPGBD2–4* are independent lineages within the family, while, as noted above, *TruPGBD5* is the ortholog of the mouse and human genes of the same name (based on the long, unique, and strongly supported branch leading to the three *PGBD5* proteins in Fig. 1, and on microsynteny of the divergently transcribed *COG2* gene which is 250 kb upstream in the human genome but just 3 kb away in *Takifugu*). This encodes a highly divergent protein in which only the N-terminal half aligns well with the other *piggyBac*-like sequences, and as in the case of *HsaPGBD1*, the C-terminal region has been lost. This pattern of sequence evolution resembles that of the gene for the mammalian centromere binding protein CENPB, which is one of ten *Tigger* -transposase-derived genes in the human genome (Kipling and Warburton 1997; International Human Genome Sequencing Consortium 2001), but has retained sequence conservation of only the N-terminal DNA-binding domain. *TruPGBD1–3* have single long ORFs encoding proteins of 593, 653, and 680 amino acids, respectively, while *TruPGBD4* appears to have an intron inserted near the N-terminus and an appropriate upstream exon has not been identified. No mammalian orthologs are apparent for *TruPGBD1–4*, so either these transposases were only domesticated in teleost fish, or they were lost from the mammalian lineage.

piggyBac- like sequences in the *An. gambiae* genome

A preliminary analysis of the *An. gambiae* mosquito genome (Holt et al. 2002) listed six copies of a *piggyBac*-like transposon. We have identified ten distinct *piggyBac*-like transposon families in the *An. gambiae* genome (*AgaPB1–10*). Furthermore, we identified three potential domesticated genes (*AgaPBD1–3*), each encoded by a single long ORF. The *piggyBac*-like transposons are present in multiple copies in the genome. *AgaPB1* represents a potentially mobile transposon with identifiable inverted repeat termini (Genbank TPA Accession No. BK 001270). Four of the transposon sequences (*AgaPB2–5*) belong to a well defined clade, while five of the transposon sequences (*AgaPB6–10*), together with the potential domesticated gene *AgaPBD2*, belong to a separate, distinct clade (Fig. 1). The individual copies of the transposons *AgaPB6–10* appear to be defective in their coding regions.

There is a 5′ EST for *AgaPBD1* in GenBank (BM642323), and it reveals a short intron with well-predicted splice sites in the 5′ UTR. *AgaPBD2* has a similar well-predicted short intron in its 5′ UTR, although such an intron is not easily identified for *AgaPBD3*. The presence of such 5′ UTR introns might be important for the expression of these domesticated genes (Zieler and Huynh 2002). Such an intron also appears in the potential 5′ UTR region immediately upstream of the coding region of the canonical *T. ni piggyBac* (donor splice sites at positions 145 or 179 and acceptor at 305 in GenBank J04364.2), although it appears not to be spliced out in *piggyBac* transcripts (Cary et al. 1989), so these 5′ UTR introns might have come from the original *piggyBac*-like transposon rather than being secondarily inserted.

piggyBac- like sequences in other eukaryotic genomes

In light of the relatively large numbers of *piggyBac*-like transposons and derived genes found in these three genomes, we examined all other publicly available sequences. No *piggyBac*-like sequences were identified from the genomes of any prokaryotic organisms, nor were any found in the apicomplexans and kinetoplastids. Sequences similar to *piggyBac* were, however, identified from the ascomycete fungus *Nectria haematococca* and in the oomycete fungi *Phytophthora sojae* and *Phytophthora infestans*, but not in *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*. Among plant genomes, *piggyBac*-like sequences were not found in the dicot *Arabidopsis thaliana*, but at least one *piggyBac*-like sequence is present in the monocot *Oryza sativa* (rice). Neither the *Nectria* nor the *Oryza* sequences have full-length ORFs, yet they are effectively unique in these genomes, so it is unclear what they represent. We were also unable to identify *piggyBac*-like sequences in the completely sequenced genomes of the microsporidian *Encephalitozoon cuniculi* and the nematode *Caenorhabditis elegans* (there are internally deleted MITEs with apparent TTAA-specific insertion deposited in RepBase v7.7 as being *piggyBac*-like in *C. elegans*, so there might once have been *piggyBac* transposons in this genome too).

Many novel *piggyBac*-like sequences were identified in other metazoans (Table 2). *piggyBac*-like sequences were also identified in the *D. melanogaster* genome. The annotated genes CG9839 and CG13151 (Berkeley Drosophila Genome Project) encode highly divergent members of the family that we interpret as *piggyBac* - derived genes. In addition, Jerzy Jurka has deposited a consensus sequence for LOOPER1_DM in RepBase, which we corrected and extended to the C-terminus of a *piggyBac* family transposase, based on overlaps with three additional genome scaffolds. Two *piggyBac*-like sequences were identified in ESTs from the silkmoth *B. mori* (one nearly complete transposase gene based on overlap of three ESTs is included in our analyses), while another was recognized in a BAC end sequence from the sea urchin *Stongylocentrotus purpuratus*. Five long sequences of defective transposons present in a few copies each were assembled from the *C. intestinalis* draft genome sequence (Dehal et al. 2002), together with an apparently domesticated sequence (*CinPBD1*). In addition to the human, mouse, and pufferfish genomes, all vertebrates for which significant amounts of genomic DNA sequence are available appear to contain *piggyBac*-like sequences, including the zebrafish *Danio rerio* (both Jurka's Repbase annotations and Genbank sequences), the amphibian *Xenopus laevis*, and various mammals (*Rattus rattus*, *Bos taurus*, and *Canis familiaris*). Sequencing of all of these genomes should be completed within the next few years, and each will probably be found to contain multiple *piggyBac*-like

Table 2 Putative novel *piggyBac*-like sequences in the genomes of various other organisms

Elementa	Organism	Genbank gi Number[a]
BmoPB 1	*Bombyx mori*	4159281 + 4159430
BmoPB 2	*Bombyx mori*	4160760
TnvPB	*Tetraodon nigriviridis*	8024937
XlaPB 1	*Xenopus laevis*	17503319
XlaPB 2	*Xenopus laevis*	17503624
DrePB 1	*Danio rerio*	20192993
DrePB 2	*Danio rerio*	16097302
DrePB 3	*Danio rerio*	20193075
DrePB 4	*Danio rerio*	16099643
DrePB 5	*Danio rerio*	17034088
BtaPB 1	*Bos taurus*	10759709
BtaPB 2	*Bos taurus*	7035014
CfaPB 1	*Canis familiaris*	18823727
CfaPB 2	*Canis familiaris*	18817197
RatPB	*Rattus spp.*	8081135
PsoPB	*Phytophthora sojae*	9836175
PinPB	*Phytophthora infestans*	10230913
SpuPB	*Strongylocentrotus purpuratus*	8373111

[a]With the exception of BmoPB1, these sequences were not included in phylogenetic analyses, except *BmoPB1*.
[b]The Genbank gi number refers to the unique Gene Identifier number of the Genbank entry

sequences, both transposons or their remnants and domesticated genes.

The *piggyBac*-like transposase protein family

Alignment of fifty full-length or nearly full-length *piggyBac* -transposase-like amino acid sequences (see Electronic Supplementary Material for FASTA, CLUSTALX, and NEXUS/PAUP files), allows recognition of distinctive features of this novel protein family. The N-terminal regions (positions 1–129 of *T. ni piggyBac* transposase) are not highly conserved and, by analogy with other transposases, might be involved in binding of the inverted terminal repeats (ITRs). The ITRs are known only from *T. ni piggyBac*, *Daphnia Pokey*, *Takifugu Pigibaku1, Ciona CinPB3* and *Anopheles AgaPB1*, and several MITEs; they are short (13–16 bp) and highly divergent in sequence.

The conserved core region of the transposase alignment, from positions 130–522 in the *T. ni piggyBac* transposase, contains several highly conserved blocks of amino acids. We presume that this is the catalytic domain of these transposases and derived proteins. Like many other transposases it contains about twelve highly conserved aspartic acid (D) and glutamic acid (E) residues. The spacing of these, and the conserved residues between them, do not readily correspond to known transposase catalytic motifs such as the widespread DDE domain, and extended iterations of PSI-BLASTP searches did not yield any similarities outside of the superfamily. Nevertheless, we note that two of these fifty proteins, TrubPBD1 and 2, which are not particularly closely related to each other (Fig. 1), yield weak matches ($P = 0.001$ and 0.002, respectively) in NCBI Conserved Domain Searches to the Pfam domain pfam01609, Transposase_11, Transposase DDE domain, which is for the bacterial *IS*4/5 transposase family. Matches are only found in the DD portion of the domain, and, if correct, suggest that the corresponding aspartic acids in the *T. ni piggyBac* transposase are D268 and D346. These are two of the most highly conserved aspartic acids in the entire alignment, they are in a roughly appropriate location in the middle of the conserved core region, and, intriguingly, they are followed by a glutamic acid and an asparagine, respectively. This is reminiscent of the situation in several other DDE transposase domains, for example in the *Tc1/mariner* superfamily (Doak et al. 1994; Robertson and Lampe 1995). There is no highly conserved aspartic or glutamic acid residue that is readily recognizable as the third residue in this triad; however, most of these proteins do have such a residue within 100 amino acids of the second D, and D447 in the *T. ni piggyBac* transposase is our best candidate for this third residue. Whether *piggyBac* transposases do indeed contain a DDE-like motif is a question that will probably require 3-D structures to resolve. If the *piggyBac* transposase family indeed shares 3-D structure with the DDE megafamily of transposases and integrases, for example by comparison to the *Tn*5 transposase (Davies et al. 1999), this would indicate both that this protein structure is even more widespread than already appreciated, and that it can diverge even more radically in primary amino acid sequence. If not, then the similarity of the conserved aspartic acids might be a convergent feature of D/E-rich DNA-manipulating protein folds.

The C-terminal region (*T. ni piggyBac* transposase positions 523–594) is again highly variable; however, it contains several cysteines (seven in *T. ni piggyBac*) with somewhat-conserved spacing, but without matches to any of the many known C-rich domains. The PSORT webserver (http://psort.nibb.ac.jp/) predicts a bipartite nuclear localization signal in this region for several of these proteins (KKRTYCTYCPSKIRRKAN for *T. ni piggyBac*), although others seem to have nuclear localization signals elsewhere in the protein; this type of variation that is not uncommon in transposases. Recently, Beall et al. (2002) showed that the *P* element transposase in *Drosophila* is in part regulated via phosphorylation at several sites recognized by serine/threonine-kinases of the ATM family: (S/T)Q or Q(S/T). The *piggyBac*-like transposase sequences contain many such sites—although they are seldom highly conserved throughout the family; thus, it is not clear how important they might be.

Phylogenetic analyses of the *piggyBac*-like sequences

Phylogenetic analyses of these proteins reveal three major clades of transposons and domesticated genes with some bootstrap support, albeit only from distance-based analyses (Fig. 1). Each of these three clades comprises a mixture of insect and vertebrate sequences, and overall the *piggyBac* phylogeny appears to deviate from the phylogeny of the species. The current limited data do not allow us to differentiate between lineage sorting, sampling bias, and horizontal transfer, except in the clearcut case of horizontal transfer implied by the high degree of similarity between the original moth *T. ni piggyBac* and the sequences from the Oriental fruit fly *Bactrocera* (Handler and McCombs 2000), although the well supported grouping of a sequence from the silkmoth *B. mori* with *HsaPGBD4* is also intriguing. Most of the remaining sequences are apparent domesticated genes, which are commonly highly divergent—for example, the apparent DDD residues are often altered. That so many independent domestications of *piggyBac* transposases have occurred in essentially all genome lineages examined suggests that their DNA-binding and/or manipulating capabilities have repeatedly been turned to advantage by the host, a phenomenon that is also shown by the diversity of transposons that have been domesticated in the human genome alone (International Human Genome Sequencing Consortium 2001, Table 13).

piggyBac-like sequences are widespread and potentially useful

This study looked at the distribution of *piggyBac*-like sequences in the currently available published genomes

and other sequence information in public databases. The distribution of *piggyBac*-like sequences in various organisms reveals that they are widespread and ancient transposons, copies of which have commonly been domesticated in various genomes. Other TTAA-specific transposons that may belong to the *piggyBac* family include: the *tagalong* element from *T. ni*, which also inserts within a TTAA target site, but all examined copies appear to have a large internal deletion in the region that could potentially encode a transposase (Wang and Fraser 1993); the *Tx* elements that have internal deletions with similar termini, from the genome of *Xenopus* frogs (Carroll et al. 1989); and SINES with TTAA insertion site specificity (Unsal and Morgan 1995). However, due to the lack of information regarding the transposases of these elements, these elements can not be definitely classified as belonging to the *piggyBac* family of transposable elements. The mobility of the original *T. ni piggyBac* element in various insects suggests that *piggyBac* family transposons might prove to be useful genetic tools in organisms other than insects. We are currently isolating an intact *piggyBac* element from *An. gambiae* (*AgaPB1*) to test its mobility in various organisms.

Acknowledgements The work described here was supported by NIH grants GM58826 (HMR) and P01AI45123 (P.I. Frank Collins), NIH cooperative agreement U01AI48846 (P.I. Frank Collins) and NIH cooperative agreement U01AI50687-01 (Large scale sequencing and assembly of the *An. gambiae* genome; P.I. Robert A. Holt, Celera Genomics).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402

Aparicio S, et al (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Beall EL, Mahoney MB, Rio DC (2002) Identification and analysis of a hyperactive mutant form of Drosophila *P* -element transposase. Genetics 162:217–227

Carroll D, Knutson DS, Garrett JE (1989) Transposable elements in *Xenopus* species. In: Berg DE, Howe MM (eds) Mobile DNA. ASM Press, Washington, D.C, pp 567–574

Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ (1989) Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon *IFP2* insertions within the FP-locus of nuclear polyhedrosis viruses. Virology 172:156–169

Craig NL, Craigie RC, Gellert M, Lambowitz AM (2002) Mobile DNA II. ASM Press, Washington D.C.

Davies DR, Braam LM, Reznikoff WS, Rayment I (1999) The three-dimensional structure of a Tn*5* transposase-related protein determined to 2.9-Å resolution. J Biol Chem 274: 11904–11913

Dehal P, et al (2002) The draft genome sequence of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167

Doak TG, Doerder FP, Jahn CL, Herrick G (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

Fraser MJ (2000) The TTAA-specific family of transposable elements: identification, functional characterization, and utility for transformation of insects. In: Handler AH, James AA (eds) Insect transgenesis. CRC Press, Boca Raton, pp 249–270

Fraser MJ, Cary L, Boonvisudhi K, Wang HG (1995) Assay for movement of lepidopteran transposon *IFP2* in insect cells using a baculovirus genome as a target DNA. Virology 211:397–407

Fraser MJ, Ciszczon T, Elick T, Bauser C (1996) Precise excision of TTAA-specific lepidopteran transposons *piggyBac* (*IFP2*) and *tagalong* (*TFP3*) from the baculovirus genome in cell lines from two species of Lepidoptera. Insect Mol Biol 5:141–151

Gardner MJ, (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419:498–511

Grossman GL, Rafferty CS, Clayton JR, Stevens TK, Mukabayire O, Benedict MQ (2001) Germline transformation of the malaria vector, *Anopheles gambiae* , with the *piggyBac* transposable element. Insect Mol Biol 10:597–604

Handler AM (2002) Use of the *piggyBac* transposon for germ-line transformation of insects. Insect Biochem Mol Biol 32:1211–1220

Handler AM, Harrell RA 2nd (1999) Germline transformation of *Drosophila melanogaster* with the *piggyBac* transposon vector. Insect Mol Biol 8:449–457

Handler AM, McCombs SD (2000) The *piggyBac* transposon mediates germ-line transformation in the Oriental fruit fly and closely related elements exist in its genome. Insect Mol Biol 9:605–612

Holt RA, et al (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149

Horn C, Wimmer EA (2000) A versatile vector set for animal transgenesis. Dev Genes Evol 210:630–637

Horn C, Offen N, Nystedt S, Hacker U, Wimmer EA (2003) *piggyBac* -based insertional mutagenesis and enhancer detection as a tool for functional insect genomics. Genetics 163:647–661

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

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

Kipling D, Warburton PE (1997) Centromeres, *CENP-B* and *Tigger* too. Trends Genet 13:141–145

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392:917–920

Lobo NF, Hua-Van A, Li X, Nolen BM, Fraser MJ (2002) Germ line transformation of the yellow fever mosquito, *Aedes aegypti* , mediated by transpositional insertion of a *piggyBac* vector. Insect Mol Biol 11:133–139

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Nolan T, Bower TM, Brown AE, Crisanti A, Catteruccia F (2002) *piggyBac* -mediated germline transformation of the malaria mosquito *Anopheles stephensi* using the red fluorescent protein dsRED as a selectable marker. J Biol Chem 277:8759–8762

Peloquin JJ, Thibault ST, Staten R, Miller TA (2000) Germ-line transformation of pink bollworm (Lepidoptera: Gelechiidae) mediated by the *piggyBac* transposable element. Insect Mol Biol 9:323–333

Penton EH, Sullender BW, Crease TJ (2002) *Pokey*, a new DNA transposon in *Daphnia* (Cladocera: Crustacea). J Mol Evol 55:664–673

Robertson HM (2002) Evolution of DNA transposons in eukaryotes. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. ASM Press, Washington, D.C., pp 1093–1110

Robertson HM, Lampe DJ (1995) Recent horizontal transfer of a *mariner* transposable element among and between Diptera and Neuroptera. Mol Biol Evol 12:850–862

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663

Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods. Sinauer Press, New York

Tamura T, et al (2000) Germline transformation of the silkworm *Bombyx mori* L. using a *piggyBac* transposon-derived vector. Nat Biotechnol 18:81–84

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Unsal K, Morgan GT (1995) A novel group of families of short interspersed repetitive elements (SINEs) in *Xenopus*: evidence of a specific target site for DNA-mediated transposition of inverted-repeat SINEs. J Mol Biol 248:812–823

Wang HG, Fraser MJ (1993) TTAA serves as the target site for *TFP3* lepidopteran transposon insertions in both nuclear polyhedrosis virus and *Trichoplusia ni* genomes. Insect Mol Biol 1:109–116

Zieler H, Huynh CQ (2002) Intron-dependent stimulation of marker gene expression in cultured insect cells. Insect Mol Biol 11:87–95