DNA Transposons and the Evolution of Eukaryotic Genomes

Cédric Feschotte and Ellen J. Pritham

Department of Biology, University of Texas, Arlington, Texas 76019; email: cedric@uta.edu

Key Words
transposable elements, transposase, molecular domestication, chromosomal rearrangements

Abstract
Transposable elements are mobile genetic units that exhibit broad diversity in their structure and transposition mechanisms. Transposable elements occupy a large fraction of many eukaryotic genomes and their movement and accumulation represent a major force shaping the genes and genomes of almost all organisms. This review focuses on DNA-mediated or class 2 transposons and emphasizes how this class of elements is distinguished from other types of mobile elements in terms of their structure, amplification dynamics, and genomic effect. We provide an up-to-date outlook on the diversity and taxonomic distribution of all major types of DNA transposons in eukaryotes, including Helitrons and Mavericks. We discuss some of the evolutionary forces that influence their maintenance and diversification in various genomic environments. Finally, we highlight how the distinctive biological features of DNA transposons have contributed to shape genome architecture and led to the emergence of genetic innovations in different eukaryotic lineages.
INTRODUCTION

Dazzling advances in molecular biology, genetics, and genomics have allowed scientists to understand in great detail many aspects of transposable element (TE) biology. Significant discoveries at the interface of these fields have provided new insight into transposition mechanisms, allowed the identification of new TEs and the broadening of their taxonomic distribution, revealed relationships between TEs and viruses, and uncovered the means by which TE movement can be controlled epigenetically by their host. Coupled to these new discoveries is a greater understanding of the extent to which TEs influence the structure and dynamics of the genomes they inhabit. The focus of this review is on one specific class of TEs, the class 2 or DNA transposons. We begin by presenting key features of the structure and life cycle of these elements, with an emphasis on the factors that govern their maintenance and propagation within the genome and throughout the eukaryotic tree of life. We then shift our focus to the repercussions of DNA transposon movement and amplification on the genome, including large-scale structural changes and epigenetic modifications, and the contribution of elements of this type to the generation of allelic diversity, new genes, and biological innovations.

EVOLUTIONARY DYNAMICS OF DNA TRANSPOSONS

Classification and Distribution of DNA Transposons

Class 2 transposable elements (TEs) or DNA transposons are mobile DNA that move utilizing a single- or double-stranded DNA intermediate (35). Eukaryotic DNA transposons can be divided into three major subclasses: (i) those that excise as double-stranded DNA and reinsert elsewhere in the genome, i.e., the classic “cut-and-paste” transposons (35); (ii) those that utilize a mechanism probably related to rolling-circle replication, Helitrons (91); and (iii) Mavericks, whose mechanism of transposition is not yet well understood, but that likely replicate using a self-encoded DNA polymerase (94, 160). Both Helitrons and Mavericks most likely rely on distinct transposition mechanisms involving the displacement and replication of a single-stranded DNA intermediate, respectively. Thus these elements probably transpose through a replicative, copy-and-paste process.

All cut-and-paste transposons are characterized by a transposase encoded by autonomous copies and, with few exceptions, by the presence of terminal inverted repeats (TIRs). Helitrons have no TIRs, but rather short conserved terminal motifs and autonomous copies encode a Rep/Helicase (91, 158). Mavericks, also known as Pointons, are very large transposons with long TIRs and coding capacity for multiple proteins, most of which are related to double-stranded DNA viruses, including a B-type DNA polymerase (52, 94, 160).

To date, ten superfamilies of cut-and-paste DNA transposons are recognized (Table 1). Elements belong to the same superfamily when they can be linked to transposases that are significantly related in sequence. Typically, transposases from the same superfamily can be confidently aligned in their core catalytic region and a monophyletic ancestry can be inferred from phylogenetic analysis (22, 164). In some cases, such as Tcl/mariner, the superfamily can be further divided into monophyletic groups that deeply diverged in eukaryotic evolution (155, 164). Two superfamilies (CACTA and PIF/Harbinger) are characterized by the presence of a second transposon-encoded protein required for transposition (Table 1).

The explosion of sequence data in the databases over the past decade has fueled the discovery of large numbers of elements in a wide range of organisms. These discoveries have yielded several new insights into the distribution and broad evolutionary history of eukaryotic DNA transposons. First, the taxonomic distribution of superfamilies initially
Table 1  Classification and characteristics of eukaryotic DNA transposons

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Related IS</th>
<th>TSD</th>
<th>Length$^1$ (kb)</th>
<th>TIRs$^1$ (bp)</th>
<th>Terminal motif (5'$-3'$)</th>
<th>TPase$^1$ (aa)</th>
<th>Catalytic motif</th>
<th>DNA-binding motif</th>
<th>Additional proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tc1/mariner</td>
<td>IS630</td>
<td>TA</td>
<td>1.2–5.0</td>
<td>17–1100</td>
<td>Variable</td>
<td>300–550</td>
<td>DD(30–41)D/E</td>
<td>HTH (cro/paired)</td>
<td></td>
</tr>
<tr>
<td>hAT</td>
<td>nd</td>
<td>8 bp</td>
<td>2.5–5</td>
<td>10–25</td>
<td>YARNG</td>
<td>600–850</td>
<td>D(68)D(324)E$^2$</td>
<td>ZnF (BED)</td>
<td></td>
</tr>
<tr>
<td>MuDR/Foldback</td>
<td>IS256</td>
<td>7–10 bp</td>
<td>1.3–7.4</td>
<td>0-sev. Kb</td>
<td>Variable</td>
<td>450–850</td>
<td>DD(~110)E</td>
<td>ZnF (WRKY/GCM1)</td>
<td></td>
</tr>
<tr>
<td>CACTA</td>
<td>nd</td>
<td>2/3 bp</td>
<td>4.5–15</td>
<td>10–54</td>
<td>CMCWR</td>
<td>500–1200</td>
<td>nd</td>
<td>nd</td>
<td>TNPA (DNA-binding protein)</td>
</tr>
<tr>
<td>PiggyBac</td>
<td>IS1380</td>
<td>TTAA</td>
<td>2.3–6.3</td>
<td>12–19</td>
<td>CCYT</td>
<td>550–700</td>
<td>DDE?</td>
<td>nd</td>
<td></td>
</tr>
<tr>
<td>Merlin</td>
<td>IS1016</td>
<td>8/9 bp</td>
<td>1.4–3.5</td>
<td>21–462</td>
<td>GGNRM</td>
<td>270–330</td>
<td>DD(36–38)E</td>
<td>nd</td>
<td></td>
</tr>
<tr>
<td>Transib</td>
<td>nd</td>
<td>5 bp</td>
<td>3–4</td>
<td>9–60</td>
<td>CACWATG</td>
<td>650–700</td>
<td>DD(206–214)E</td>
<td>nd</td>
<td></td>
</tr>
<tr>
<td>Banshee</td>
<td>IS481</td>
<td>4/15 bp</td>
<td>3–5</td>
<td>41–950</td>
<td>TGT</td>
<td>300–400$^4$</td>
<td>DD(34)E</td>
<td>HTH</td>
<td></td>
</tr>
<tr>
<td>Helitron</td>
<td>IS91</td>
<td>none</td>
<td>5.5–17</td>
<td>none</td>
<td>5’T-C...CTAR-3’</td>
<td>1400–3000$^5$</td>
<td>HHYY (“REP motif”)</td>
<td>ZnF-like</td>
<td>RPA (in plants)</td>
</tr>
</tbody>
</table>

1Refers to a potentially complete, autonomous element.
2Motif in Hermes TPase.
3Motif in Drosophila P element TPase.
4RVE integrase-like.
5REP-Helicase.
nd = not determined.
believed to be restricted to a few related taxa has been significantly expanded to cover several eukaryotic kingdoms or supergroups (41, 68, 170) (e.g., P element, CACTA, Piggy-Bac; see Table 1 and Figure 1). Second, links have been established between superfamilies that were previously separated [e.g., union of MuDR and Foldback (C. Marquez, E.J.P., unpublished)]. Finally, novel superfamilies have been recognized (e.g., PIF/Harbinger, Merlin, Transib, Banshee) (48, 90, 93, 206; A. Barrie, J. Pham, E.J.P., unpublished) and two distinct subclasses of DNA transposons have been identified (Helitrons and Mavericks).

A superimposition of the distribution of each cut-and-paste DNA transposon...
superfamily on the most current representation of the eukaryotic tree of life (97) reveals that 8 of the 10 superfamilies are represented in two or more eukaryotic supergroups (Figure 1). Given that there is no convincing evidence for horizontal transfer of DNA transposons between eukaryotic supergroups, this distribution suggests that most superfamilies were already differentiated in the eukaryotic ancestor. Furthermore, alliances with prokaryotic insertion sequence families can be drawn for six of the ten eukaryotic superfamilies (Table 1), suggesting that the divergence of most superfamilies may even predate the split of eukaryotes and prokaryotes. Finally, Helitrons and Mavericks are also distributed across multiple eukaryotic supergroups (Figure 1). These data underscore the extremely ancient roots of the major types of DNA transposons and their remarkable persistence over evolutionary time.

Differential Success of DNA Transposons among Species

Eukaryotic species show enormous variation in the amount of TEs occupying their genomes (1, 111). It is now well established that these variations largely account for the wide differences in genome size observed among eukaryotes, and even between closely related species (64, 100). Retrotransposons seem to be major players in promoting rapid increase, and perhaps also decrease, in the genome size of multicellular eukaryotes (7, 10, 124, 128, 150, 171, 185). This is best exemplified by studies of maize and of the rice Oryza australiensis, showing that massive bursts of LTR retrotransposon amplification caused a concomitant doubling of the genome independently in the lineages of these two species (152, 169).

DNA transposons may also contribute substantially to genome expansion. An estimated 65% of the genome of the single-celled eukaryote Trichomonas vaginalis, which was recently sequenced, is made of repetitive DNA (23). Virtually all TEs that have been recognized in this genome are DNA transposons (94, 160, 174; E.J.P, unpublished). In fact, only a handful of retrotransposon-related proteins are recognizable in the genome (23) and it is not yet clear whether they actually belong to mobile elements (E.J.P, unpublished). Recent studies indicate that genome expansion in this species can be largely accounted for by the massive amplification of Maverick transposons (160). There are an estimated 3000 Maverick copies per haploid genome and considering an average size of these elements in T. vaginalis of 15 to 20 kb, it can be inferred that these transposons occupy a stunning ~60 Mb of the ~160 Mb genome, that is ~37% of the genomic space.

Tremendous variation also exists among species in the relative abundance of DNA transposons and retrotransposons, regardless of their sheer numbers (Figure 2). For example, DNA transposons seem completely absent from the genomes of budding and fission yeasts, although different families of LTR retrotransposons have survived in both species. Yet DNA transposons are common in filamentous fungi and occur occasionally in other yeasts, such as Candida albicans (36). Thus, two independent extinction events of DNA transposons occurred in the lineages leading to Saccharomyces cerevisiae and Schizosaccharomyces pombe.

The human TE landscape is clearly dominated by retrotransposons (Figure 2), mostly LINEs and associated SINEs (111). Nonetheless, human DNA transposons are highly diversified (120 families falling into 5 superfamilies) and they are numerically abundant (111, 151). With 300,000 copies, the human genome contains about 15 times more DNA elements than the DNA transposon-rich genome of Caenorhabditis elegans and 40 times more than Drosophila melanogaster (Table 1 and data from the UCSC Genome Browser). In addition, nearly 100,000 DNA transposon copies from 40 families and 4 different superfamilies integrated during the primate radiation (151). None of these elements, however, appears to have survived a seemingly

Horizontal transfer: the transmission of genetic material between the genomes of two individuals (that may belong to different species) by nonvertical inheritance

Insertion sequence: prokaryotic mobile element that resembles eukaryotic DNA transposons in their structure and transposition mechanism
The relative amount of retrotransposons and DNA transposons in diverse eukaryotic genomes. The graph shows the contribution of DNA transposons and retrotransposons in percentage relative to the total number of transposable elements in each species. The data were compiled from papers reporting draft genome sequences (references available upon request) and from the Repeatmasker output tables available at the UCSC Genome Browser (http://genome.ucsc.edu) or from the following sources: *E. histolytica* and *E. invadens* (159); *T. vaginalis*: E. Pritham, unpublished data. Species abbreviations: Sc: *Saccharomyces cerevisiae*; Sp: *Schizosaccharomyces pombe*; Hs: *Homo sapiens*; Mm: *Mus musculus*; Os: *Oryza sativa*; Ce: *Caenorhabditis elegans*; Dm: *Drosophila melanogaster*; Ag: *Anopheles gambiae*, malaria mosquito; Aa: *Aedes aegypti*, yellow fever mosquito; Eh: *Entamoeba histolytica*; Ei: *Entamoeba invadens*; Tv: *Trichomonas vaginalis*.

general extinction event of DNA transposons that occurred about 40 My (million years) in an anthropoid primate ancestor. The picture emerging from the initial analyses of the mouse, rat, and dog genome sequences is strikingly similar, with no evidence for the activity of DNA transposons during the past 40–50 My (59, 118, 192; J. Pace & C.F., unpublished). At first sight, these data suggest an intriguing scenario whereby DNA transposons went extinct independently in different mammalian lineages around the same evolutionary time (Eocene, 35–55 My) and have not been maintained or reintroduced into these lineages since this epoch.

Does it mean that all mammals are now refractory to the propagation of DNA transposons? The answer, which came unexpectedly from the genome of the little brown bat, *Myotis lucifugus*, is no. With a haploid genome size of ~2,000 Mb, *M. lucifugus* is one of the smallest mammalian genomes, but it harbors a surprisingly diverse collection of DNA transposons that is also distinct from other mammalian genomes examined. In particular, the genome is packed with *Helitrons* (at least 3% of the genome) (158), whereas none are recognizable in any of 22 other placental species (including two other bat species) for which a substantial amount of genomic sequences is now available. In contrast to other mammals so far examined, the recent data point to a continuous colonization of the vesper bat genome(s) by various DNA transposon families (158, 161; D. Ray, J. Smith, H.J.T. Pagan, E.J.P., C.F., N.L. Craig, submitted). Several waves of amplification of different families have succeeded over the past 40 My. Moreover, the invasion seems to be ongoing because there is mounting evidence that some *hAT* and *piggyBac* families are still active in natural populations of *Myotis* (161; D. Ray, J. Smith, H.J.T. Pagan, E.J.P., C.F., N.L. Craig, submitted).

Hence, sharp variation in the success of DNA transposons may exist even between closely related species. This variation is also illustrated by a comparative study of TE
composition in the genomes of four species of Entamoeba, a single-celled eukaryote distantly related to animals and fungi (159). The four Entamoeba species all have relatively small genomes estimated to be about 20 Mb, but their TE composition varies dramatically. The genomes of E. invadens and E. moshkovskii host many families of DNA transposons from four different superfamilies and few retrotransposons (159), whereas the genomes of E. histolytica and E. dispar contain virtually no DNA transposons but instead were colonized by several lineages of non-LTR retrotransposons (5). The genomes of Entamoeba, despite harboring completely different TE complements, are composed of the same relative proportion of TEs (5%–7%), and all four genomes contain recently active elements (5, 159). Thus these genomes seem to be similarly constrained in size, but retrotransposons and DNA transposons have experienced differential success.

Population Dynamics of DNA Transposons Within Genomes: The MITE Paradox

DNA transposons are typically grouped into families. In principle, members of the same family are all descended from a common autonomous ancestor copy, which transposed and generated copies of itself in the process. Because most DNA transposons move through a nonreplicative mechanism, these elements increase their copy numbers through indirect mechanisms that rely on the host machinery (35). The first mechanism invokes the transposition of the element during DNA replication from a newly replicated chromatid to an unreplicated site. The transposon is thereby effectively replicated twice, leading to a net gain of one transposon copy. This behavior has been documented for the maize Ac and Spm elements (106). For Ac, the timing of transposition during DNA replication is explained by the preferential binding of Ac transposase to hemimethylated binding sites (166). The second mechanism draws on the repair of the double-strand break left by excision of the element. If the element is present on the homologous chromosome, gap repair via homologous recombination results in the reintroduction of the transposon at the donor site. If transposition occurs during the S phase of the cell cycle, the sister chromatid may also be used as the template for gap repair, resulting in the restoration of the excised element. Gap repair has been demonstrated to be the mechanism by which P elements rapidly increase their copy number in D. melanogaster (46). This process operates for other transposons in various species and gives rise to various internal deletion derivatives as a result of abortion, slippage, or template switching during gap repair (46, 77, 154, 168).

Because the terminal sequences of DNA transposons are often the only requirement for transposase recognition (35), internally deleted or rearranged nonautonomous elements may still transpose by using enzymes encoded elsewhere in the genome by an autonomous copy. The frequent emergence of nonautonomous derivatives coupled to the apparent lack of cis-preference of eukaryotic transposases poses a major hurdle for the successful propagation of an autonomous element (70). Indeed, unless there exists a mechanism to prevent the formation of nonfunctional copies upon gap repair [see the possible case of Tam3 in snapdragon (198)], it can be predicted that autonomous copies will be rapidly out-numbered by nonautonomous copies (70, 112). As copy number increases, the entire family potentially faces two constraints: (i) titration of the transposase by binding to multiple nonautonomous copies and (ii) an increased chance to trigger host- or self-induced repression mechanisms, such as RNA interference (RNAi) (1, 69, 112, 173, 177). Both constraints would eventually prevent the autonomous element from replicating, leading to its elimination or inactivation from the population and to the extinction of the entire family.

Considering this disastrous sequence of events, also referred to as vertical inactivation.
Vertical diversification: the emergence (speciation) of a slightly different variant of an autonomous transposon from another element, either within the same species or during the radiation of a species, giving rise to a new transposon family (69), the amplification of nonautonomous copies could be viewed as a death sentence for DNA transposons. Yet we observe a paradoxical situation where the genomes that harbor the most diverse and the highest density of DNA transposons (i.e., rice, nematodes, human) are also the ones filled with the largest amount of miniature inverted-repeat transposable elements (MITEs) (56, 85, 151). Why MITEs are so prevalent and how DNA transposons can be so successfully maintained and propagated in this context are the subjects of the next section.

Mechanism and Consequences of MITE Amplification

MITEs are short transposons (100–600 bp) that are distinguished from other nonautonomous elements by high copy numbers and length homogeneity (19, 56, 205). The structural homogeneity of MITE families indicates that they arose by amplification of a single or a few progenitor copies (49, 50). Presumably, the progenitor copy arises by deletion of a larger transposon during gap repair. Yet it is sometimes difficult, if at all possible, to directly connect a given MITE family with an autonomous transposon present within the same genome (53, 207). In many cases, sequence similarity between MITEs and the closest autonomous element is restricted to the TIRs (53, 149). Two hypotheses can be put forward to explain this paradox. First, some MITEs may arise de novo from the fortuitous juxtaposition of solo TIRs or sequences resembling the TIRs of an autonomous transposon (127, 183). A second possibility is that MITE progenitors are the relics of the past invasion of transposons whose autonomous copies have been erased or have not reached fixation within the population (53). In many cases, sequence similarity between MITEs and the closest autonomous element is restricted to the TIRs (53, 149). Two hypotheses can be put forward to explain this paradox. First, some MITEs may arise de novo from the fortuitous juxtaposition of solo TIRs or sequences resembling the TIRs of an autonomous transposon (127, 183). A second possibility is that MITE progenitors are the relics of the past invasion of transposons whose autonomous copies have been erased or have not reached fixation within the population (53). In many cases, sequence similarity between MITEs and the closest autonomous element is restricted to the TIRs (53, 149). Two hypotheses can be put forward to explain this paradox. First, some MITEs may arise de novo from the fortuitous juxtaposition of solo TIRs or sequences resembling the TIRs of an autonomous transposon (127, 183). A second possibility is that MITE progenitors are the relics of the past invasion of transposons whose autonomous copies have been erased or have not reached fixation within the population (53).

The accumulation of MITE families over time creates a reservoir of elements ready for accidental cross-activation by newly emerged autonomous transposons, triggering new waves of MITE amplification (53). This scenario is supported by studies of rice mariner-like transposons and their related Stowaway MITEs, which reveal that currently active transposases can bind to the TIRs of a wide diversity of distantly related MITEs represented by thousands of copies within the same genome (51).

How then do DNA transposons replicate given such strong competition? One explanation is that MITE amplification might pass under the radar of the host defense system, either because the transposons are too small or because they fail to trigger the trans-silencing of the autonomous transposon providing the source of transposase (53). One mechanism of defense evasion may occur as a result of the absence of homology between MITEs and the transposase gene or its promoter region. In this model, the lack of shared sequence similarity allows the continuous expression of the transposase source, which serves to propagate the MITEs, as well as the autonomous transposon but at a lower frequency. The potential problem of titration of transposase molecules by binding to many illegitimate targets remains (38, 69), perhaps representing a major selective force favoring the emergence of transposon variants that minimize cross-interaction with MITEs present in the same genome (38, 51, 110, 126). Thus the presence of MITEs could actually benefit the long-term evolution of DNA transposons by driving their vertical diversification.

The recent isolation of active MITE-transposase systems (84, 103, 198a) has allowed most of these hypotheses to be tested in the laboratory and also in the context of natural populations. The most promising model is the mPing/Pong system of rice. mPing was identified as the first actively transposing MITE in any organism (84, 103). mPing transposition has been observed in vivo in response to various stress conditions and correlated with the coactivation of Pong, a distantly related autonomous transposon of the PIF/Harbinger superfamily (84, 116, 172). Recently, it was also shown that Pong and Ping-encoded proteins are necessary and sufficient to mobilize mPing in transgenic Arabidopsis.
plants (198b). Finally, evidence was gathered that mPing copy number has recently exploded in the field and reached approximately 1,000 copies in some cultivated rice strains (142). This situation offers an unprecedented opportunity to comprehend how MITEs attain such high copy numbers without killing the host or silencing their autonomous transposon partner.

**Horizontal Transmission and Vertical Diversification**

Even in the absence of MITE amplification, the vertical inactivation theory predicts that DNA transposons would ultimately go extinct unless autonomous elements can be periodically reintroduced in a genome that has not been previously exposed to the proliferation of the same element (69). The best way to achieve this is by horizontal introduction of an autonomous element to a new species (or population). Clear cases of horizontal transfer (HT) of DNA transposons have been documented, especially for Tc1/mariner and P elements among insect species (39, 164, 175). Recently, a possible HT of a MULE transposon between plants was reported (42). Thus, it is believed that all DNA transposons rely heavily on HT for their propagation and maintenance throughout evolution (69, 164).

Support for the notion that DNA transposons are well adapted to HT comes from *in vitro* experiments, which showed that, for all transposon systems so far examined, transposase is the only protein needed for transposition [for review, see (35)]. Consistent with the apparent lack of requirement for host-specific factors, most active transposons isolated from one species are readily functional in a wide range of heterologous species [for review, see (135, 148, 155)].

Important gaps remain in our understanding of the evolutionary dynamics of DNA transposons. Recent large-scale phylogenetic analyses of DNA transposon populations within species and in closely related species indicate that HT cannot account for the diversity and multiplicity of DNA transposons coexisting within a single genome. For example, phylogenetic analysis of 68 distinct mariner-like transposase sequences from 25 grass species revealed no instances of HT, but is consistent with vertical transmission and continuous diversification of multiple lineages of transposases during grass evolution (55). Likewise, distant Entamoeba species shared deeply diverged lineages of transposases, indicative of their presence in the common ancestor of the species followed by their vertical diversification (159). These data point to the existence of mechanisms allowing DNA transposons to rapidly diversify within species. Rapid diversification would limit the chances for cross-interactions between related copies and promote the speciation of new active families (1, 51, 110). One possible opportunity for diversification is during gap repair following transposon excision. The capture of filler DNA sequences at double-strand breaks owing to template switching and other aberrant repair events has been documented in various organisms (61, 117, 193, 199). These processes can readily explain the capture of new internal sequences by transposons (72, 83, 143, 168). Likewise, a frequent exchange of sequence information was recorded between actively transposing Tc1 elements dispersed in the genome of *C. elegans*, suggesting that gap repair processes following excision may accelerate the evolution of the elements (57). Such mechanisms could account for the great sequence variation observed in the sub-terminal regions of transposons that share otherwise highly conserved transposase genes (53, 207).

**IMPACT OF DNA TRANSPOSONS ON GENOME EVOLUTION**

Like other transposable elements, DNA transposons have the potential to influence the evolutionary trajectory of their host in three distinct ways: (i) via alterations of gene function through insertion; (ii) through the
induction of chromosomal rearrangements; (iii) as a source of coding and noncoding material that allows for the emergence of genetic novelty (such as new genes and regulatory sequences). DNA transposons have properties distinct from those of retrotransposons that uniquely affect the means and propensity for participation in each of these mechanisms. Here we review how the properties of DNA transposons contribute to the generation of allelic diversity in natural populations, shape the genomic and epigenetic landscape of their hosts, and contribute to the creation of new genes.

**Generation of Allelic Diversity through Insertion and Excision of DNA Transposons**

Like other TEs, DNA transposons are potent insertional mutagens. The insertion of DNA transposons may affect host gene expression in myriad ways, the phenotypic consequences of which were richly illustrated by the molecular characterization of a plethora of TE-induced mutations during the first decades of TE research (40, 49, 101, 194). The most straightforward outcome of TE insertion is the disruption of the coding sequences of a gene inhibiting the production of viable gene product. However, TE insertion, for example within promoters, introns, and untranslated regions, can directly trigger the full gambit of phenotypes, ranging from subtle and epigenetic regulatory perturbations to the complete loss of gene function (101, 194).

Unlike the majority of retrotransposons, many cut-and-paste transposons exhibit a marked preference for insertion into or within the vicinity of genes, a property that has allowed their development into powerful gene-tagging tools routinely used by geneticists (9, 179, 187). P elements in *Drosophila* (179), *Mutator* elements in maize (43), and the *Tc3* element in nematodes (163) have all been shown to have a bias for insertion into genic neighborhoods. Additionally, in both plant and animal genomes MITEs are typically found in low-copy-number genomic regions and gene-rich environments (19, 56, 205). A breakthrough study of a recent MITE explosion in rice demonstrated for the first time that this pattern of insertion, at least for *mPing*, was primarily due to targeting rather than the result of selection (142). The genic proximity of DNA transposon insertions confers on them a significant potential for generating allelic diversity in natural populations. In addition, we propose that genic proximity also facilitates the co-option of DNA transposons for gene regulation (see below).

Another important property of DNA transposon-mediated insertional mutagenesis is the ability of DNA transposons, unlike retrotransposons, to subsequently undergo spontaneous excisions (194). Therefore DNA transposons frequently generate unstable mutations with reversible phenotypes. Excisions are often imperfect, leaving behind a transposon footprint and/or altering the flanking host DNA (e.g., 103, 154, 194). The nature of these changes has been determined through the examination of the sites of DNA transposon excision and includes small deletions, inversions, as well as the introduction of random filler DNA. Multiple alleles with an array of phenotypic consequences have been identified in fungi, plants, and animals (29, 60, 101, 193, 195). A striking example was recently reported involving a member of the *hAT* superfamily, *Tol2*, in the medaka fish (105). In an inbred line, a wide range of pigmentation phenotypes could be recovered, ranging from albino to wild type through partially pigmented patterns. Closer molecular examination revealed that individuals homozygous for a *Tol2* insertion in the promoter region of a pigmentation gene exhibit complete albino phenotypes. Perfect *Tol2* excision accounted for wild-type individuals, and imprecise excisions gave rise to new alleles with different footprints and various heritable pigmentation phenotypes. The phenotypic mutation rate induced by *Tol2* excision at this locus was as high as 2% per gamete, representing a 1000-fold increase from
the spontaneous mutation rates previously determined for this species (105).

The generation of new alleles and the creation of novel regulatory circuits are major forces underlying the diversification of species (14, 24, 104, 196). As DNA transposon excision can rapidly generate allelic diversity, many subtle adaptive modifications of gene and promoter sequences could conceivably have involved insertion/excision of DNA transposons, but unless the transposon is caught in the act, these would prove difficult to demonstrate (15, 60, 105). Thus, the broad range of alterations and phenotypes caused by transposon excision in the lab may just represent the tip of the iceberg of what has actually occurred in nature.

**TE-Mediated Epigenetic Effects on Gene Expression**

McClintock first made the observation that maize transposons could influence nearby gene expression in a heritable fashion, and therefore designated them as controlling elements (131, 132). She also realized that the regulatory influence of transposons was reversible independent of their movement, alternating phases of quiescence and reactivation. Based on these results and on a number of intricate experiments, she put forward the visionary hypothesis that the regulatory influence of transposons on nearby genes was epigenetic in nature and could be modulated by changes in the environment (133). Although this model was largely overlooked at the time, the explosion of epigenetic research over the past decade has revived these ideas and validated several aspects of the model (120, 178). There is now clear evidence that DNA transposons play a major role in attracting the machinery responsible for formation and maintenance of heterochromatin comes from the comparative analysis of two large duplicated regions of *Arabidopsis* chromosome 4 using tiling microarrays (119). One region is a heterochromatic knob replete with repetitive sequences, including a high density of CACTA and MULEs, conspicuously enriched in CpG and H3K9 methylation, whereas the other region is euchromatic, almost completely free of TEs, hypomethylated, and enriched in H3K4 methylation. The heterochromatic transposons are also associated with matching siRNAs. The epigenetic marks of heterochromatin were essentially erased in plants mutant for *DDM1*, a chromatin-remodeling factor essential for the silencing of CACTA and MULEs in *Arabidopsis* (178). In the *ddm1* background, the transposons become awakened as a result of the loss of transcriptional silencing. Several examples of silenced transposons inserted in the proximal promoter regions were also found to provoke the transcriptional silencing of the adjacent gene, and both transposons and the associated genes were transcriptionally reactivated in the *ddm1* mutant (119, 137, 176). A tight association between the silencing status of a MULE transposon and a nearby gene was also previously reported in maize (6). These data are consistent with McClintock’s hypothesis of transposons acting as controlling elements of gene expression (131, 132). In addition, studies in *Drosophila of Hoppel* (71), a member of the P-element superfamily, indicate that these mechanisms are not restricted to plants, but also operate in animals and frequently implicate DNA transposons.

Together these data converge toward a model whereby DNA transposons (and
other TEs) act as moving targets for local heterochromatin formation as a byproduct of their structure (TIRs) and/or simply their repetitive nature (65, 178). Together with other sequence elements such as boundary or insulator elements, which may also be derived from repeats (145, 184), and their associated trans-acting siRNAs, transposons actively participate in a partitioning of the genome into chromosomal domains with distinct epigenetic marks and transcriptional activity. These marks are inheritable and normally stable, but they may be subject to dynamic changes in response to environmental cues and genetic stress, such as interspecific hybridization or polyploidization (2, 31, 95, 146). These events may in turn trigger further movement and amplification of TEs, provoking a structural and epigenetic reshuffling of the genome and offering an opportunity for natural selection to establish new chromosomal domains and regulatory circuits. This scenario essentially corroborates McClintock's genomic shock theory (133).

Large-Scale Chromosomal Rearrangements

There is a rich record implicating DNA transposons in the induction of large-scale chromosomal rearrangements in plants and animals. The transposition mechanisms of these elements, which involve multiple double-strand breaks and repair events, predispose them to actively participate in these processes. Among the initial examples reported, the Foldback (FB) elements of Drosophila stand out because of the high frequency and the amplitude of the provoked rearrangements (30). The unusually high frequency of interelement recombination between FB elements strongly suggests participation of transposase cleavage activities at the termini of one or both of the elements (63, 114). The large size (often over 10 kb) and complex inverted repeat structure of FB might also be factors contributing to the recurrent involvement of these elements in rearrangements (30).

Recently, FB elements once again took the front stage in TE-induced chromosomal rearrangements. However, this was the first time that rearrangements similar to those observed in the lab were identified in natural populations and linked to events with evolutionary consequences. In a series of elegant studies, the group of Alfredo Ruiz demonstrated that ectopic recombination between oppositely oriented FB-like transposons generated two independent chromosomal inversions in Drosophila buzzatii (21, 26). These inversions are geographically widespread and polymorphic in natural populations, which strongly suggests that they are selectively advantageous (8). In each case, the inversion breakpoints occur within genomic hotspots that are highly variable in sequence and structure between populations (20, 26). The inversion breakpoints are characterized by complex nesting of DNA transposons (mostly of the FB and hAT superfamilies), but strikingly, no recognizable retrotransposons. Once again, the frequency of the rearrangements and the prevalence of transposons point to a transposase-triggered mechanism rather than passive ectopic recombination events.

Transposase-induced rearrangements have long been recognized as a particular class of recombination events with a strong potential for restructuring the chromosome (63, 114). These events, termed alternative transposition processes, occur typically when the termini from separate transposon copies synapse together and engage in a complete or partial cut-and-paste reaction. Depending on the orientation of the termini used for the reaction and on the chromosomal location of the elements, alternative transpositions can lead to various outcomes, including chromosomal inversions, duplications, and deletions of over 100 kb (63, 114, 157, 203). Translocations can also occur if the insertion site is on a different chromosome from that of the two elements involved in alternative transposition. The molecular ontology of each type of rearrangement has
been reviewed in detail elsewhere (63). Most of the possible outcomes have been recovered experimentally in diverse model organisms, such as snapdragon, maize, Drosophila, and Fusarium oxysporum, and with members of various DNA transposon superfamilies, such as hAT, P element, Tcl/mariner, and PIF/Harbinger (47, 79, 122, 189, 203).

Local hopping is a property of many DNA transposons that may augment their propensity to create local genomic rearrangements (63). Local hopping is the preference of an element to transpose to a linked chromosomal location, a behavior exhibited by transposons from different superfamilies (66, 98, 106, 139, 182). This and other targeting activities tend to create genomic clusters of related elements (78, 86), which would further enhance the probability for alternative transposition events.

Although many of the chromosomal rearrangements observed in the laboratory would be deleterious in nature, some may even occasionally bring a selective advantage to individuals carrying them, e.g., the D. buzzatii inversions (8, 21, 26). Chromosomal rearrangements and gene relocation events have been linked to speciation events (129, 144, 147). A recent study of alternative transposition pathways using reversed Ac element termini in maize showed that these events can mediate exon shuffling and create new chimeric functional genes (204). The mechanism is analogous to V(D)J recombination, a process that generates endless combinations of antibody genes in the immune system of jawed vertebrates. As discussed below in more details, this parallel makes sense in light of growing evidence that the V(D)J recombination system is actually derived from immobilized DNA transposons.

**Involvement of DNA Transposons in Gene Transduction, Duplication, and Exon Shuffling**

The capture of host genes as part of mobile elements was first discovered in the context of cellular oncogenes transduced by retroviruses (180). Non-LTR retrotransposons are also capable of transduction of adjacent host sequences, specifically the L1 family and related genomic parasites in human (138). Given the abundance of retrotransposons and other retroviral-like elements in some eukaryotic genomes, one might expect this process to be an evolutionarily potent mechanism for the duplication and movement of host genes. However, very few examples of host gene transduction by retrotransposition have been reported (45, 167, 197).

In contrast to these isolated examples, recent studies have shown that several types of DNA transposons have transduced hundreds to thousands of gene fragments in grass genomes. MULEs have been long suspected of capturing and carrying host gene fragments (121). The recent availability of the rice genome sequence has allowed a first quantitative appreciation of the extent of this phenomenon. Jiang et al. (83) identified over 3000 so-called PACK-MULEs containing fragments from more than 1000 cellular genes. Remarkably, about one fifth of the identified PACK-MULEs had captured exons from multiple loci, and some elements had effectively assembled chimeric genes representing novel exon combinations producing processed transcripts in planta. Although it remains to be shown whether rice PACK-MULEs have given rise to new genes with cellular function (89), the study clearly established the tremendous potential of PACK-MULEs for gene shuffling and duplication. Moreover, the tendency of MULEs to capture host sequences is not restricted to rice, but also occurred at appreciable frequency in dicot plants (75, 76). A recent example of PACK-MULE-mediated gene duplication in Arabidopsis shows that the mechanism can give rise to genes retaining functional coding capacity and likely novel function (75). Many more examples will likely soon be identified in other plant species and perhaps in other eukaryotes, given the widespread occurrence of MULEs (Table 1).
The mechanism by which PACK-MULEs capture host gene fragments is not understood. It is conceivable that it involves template switching and other aberrant events during the gap repair mechanism that follows transposon excision. Similar events of DNA capture have been reported during the repair of DSB left by excision of *Drosophila* P elements and maize *Ac/Ds* elements (61, 143, 168). Hence, not just MULEs, but other cut-and-paste DNA transposons are expected to be prone to capture. Indeed, several lines of evidence indicate that plant CACTA elements frequently transduce host sequences (96, 200). However, *Helitrons*, with their distinct mechanism of amplification, may raise the bar even higher in their ability to shuffle and duplicate host sequences (140).

Ever since the discovery of *Helitrons* in eukaryotic genomes, their potential to act as exon shuffling machines (54) was apparent from the observation that some plant elements had seemingly captured one or multiple RPA-like proteins from the host genome to serve their own propagation (91). These proteins are involved in rolling-circle replication of other mobile elements, but normally are encoded by the host. The likelihood for *Helitrons* to capture genes useful for transposition implies that transduction events, regardless of the mechanism, must be extremely frequent (54). Preliminary evidence that this is indeed the case came from the isolation of the first *Helitrons* from maize (108, 109). These elements were large in size (up to 17.7 kb) and were packed with fragments of seemingly unrelated genes. Most of the gene fragments were pseudogenes in various states of decomposition, and they had apparently been captured progressively from different genomic loci in the maize genome (67). Nonetheless, these PACK-*Helitrons* had clearly been recently active, as judged by their absence at orthologous position in other maize inbred lines (108, 109).

Only more recently has it become clear that the first identified maize *Helitrons* represent only the tip of the iceberg. Elegant whole-genome analyses of gene content polymorphism between two inbred maize lines revealed ~10,000 large DNA insertions disrupting colinearity between the two lines (141). Eight of nine insertions molecularly characterized were found to be typical insertions of nonautonomous *Helitrons* replete with host gene fragments. It was shown that these elements and their internal gene fragments are frequently transcribed and that they transpose replicatively, peppering the genome with pieces of genes, while capturing additional gene fragments in the process (18, 67). The extrapolation of these findings to the whole maize genome revealed an unprecedented image of genome plasticity. Furthermore, if the captured fragments are indeed transcribed as was reported in the study, this could potentially create havoc considering the potential collision and interference of gene expression among the captured gene fragments and their parental copy. There must exist some mechanisms, most likely epigenetic, to keep this transcriptional burden under control.

Is the amplitude of *Helitron*-mediated transductions unique to the maize genome? *Helitrons* and many other TEs have clearly been unleashed recently in maize and are probably still in an epoch of massive expansion in this species. However, it should be kept in mind that vast numbers of *Helitrons* have colonized the genome of a broad range of animals, including worms, mosquitoes, sea urchin, zebrafish, or bats (91, 92, 156, 158). Thus, there is no reason to assume that *Helitron*-mediated transduction events would be restricted to the maize genome. In fact, an instance of *Helitron*-mediated exon transduction and its subsequent amplification to ~1000 copies has been identified in the genome of the bat *M. lucifugus* (158). A more comprehensive assessment of the extent of this phenomenon in the bat genome is under way and should reveal whether this mechanism has also contributed to mammalian genome evolution.

In summary, it is becoming increasingly clear that DNA transposon-mediated transduction has been a significant mechanism...
contributing to the structural evolution of the genome. In fact, the maintenance of captured RPA sequences in plant Helitrons also illustrates the other side of the coin, namely that DNA transposons can take advantage of this mechanism for their own, typically modular, evolution. Likewise, it is tempting to speculate that the murdB gene that is unique to the maize MuDR element originates from a host gene fortuitously captured. Hence, there seems to be a continuous flux of sequences from the host to the DNA transposons. As we describe in the final section of this review, the flux is reversible: DNA transposons can also donate sequences to their host.

Molecular Domestication of DNA Transposons

One of the most direct contributions of TEs to host genome evolution is as a source of raw material that can be used for the assembly of new genes and functions (12, 16, 101, 102, 125, 186). TEs have numerous properties that predispose them for molecular domestication (134) or exaptation (17) by the genome for host function. For example, the palindromic structure of some MITEs may predispose them to evolve into microRNA genes (153). In this section, we focus on a particular category of exaptation events that seem to regularly implicate DNA transposons: the donation of protein-coding sequences to assemble new host genes.

Estimations of the rate at which TE-encoded proteins have been domesticated throughout evolution are necessarily conservative owing to our limited ability to recognize relationships between host genes and TEs. Indeed, many events are likely to have been erased through evolutionary time or they are so ancient that it cannot be inferred whether the TE gave rise to the host gene or vice versa [e.g., telomerase (44)]. In addition, TE genes and host genes cannot easily be distinguished in those genomes where large amounts of related and recently active TEs occur. In these genomes, very recent events of domestication will be difficult to detect.

Different studies aimed at systematically identifying TE-derived genes have used different criteria. Some were purposely very stringent (202), whereas others were perhaps too pliant and likely yielded many false positives (11). Estimates from analyses of the human genome range widely from a few dozens to thousands (11, 13, 62, 111, 202). The reality is probably somewhere in between these estimates. Regardless of the exact count, all the studies point to a similar pattern whereby DNA transposons contribute to a proportionally large number of TE-derived genes relative to their abundance in the genome.

We have adopted a relatively conservative approach and list in Table 2 only examples of DNA transposon-derived genes in animal, fungi and plant species that have received extensive support for their transposon origin and functionality. Specifically, these genes fulfill at least three of the following criteria:

1. Absence of flanking transposon hallmarks (such as TIRs) and no evidence for recent mobility;
2. Phylogenetic placement of the encoded protein within a cluster of transposon-encoded proteins;
3. Intact coding capacity and evolution under functional constraints (as opposed to TE coding regions, which typically evolve under neutral evolution);
4. Detection of intact orthologs in syntenic genomic regions of distantly related species (TE genes are not expected to be maintained intact for extended period of time at orthologous positions between two distantly related species, such as human and mouse, for example);
5. Evidence of transcription (in contrast, TE genes are often transcriptionally silenced);
6. Genetic evidence for a critical biological function in vivo.

Most of the genes listed in Table 2 encode transposase-related proteins, since this is the only protein encoded by most DNA transposase. Exceptions include the c-integrases of mammals, which appear to...
Table 2  Transposase-derived genes and their functions

<table>
<thead>
<tr>
<th>Related TE Superfamily/ subgroup</th>
<th>Gene ID</th>
<th>Full name</th>
<th>Original species/distribution</th>
<th>Functions and activities</th>
<th>Protein domains derived from Tpase</th>
<th>Other domains fused</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tc1/mariner/pogo</td>
<td>CENP-B</td>
<td>Centromere protein B</td>
<td>H. sapiens/Mammals</td>
<td>Centromeric chromatin assembly, binds CENP-B box in alphoid satellite</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>(148a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>JRK</td>
<td>Jerky</td>
<td>H. sapiens/Mammals</td>
<td>Probable translational regulator in neurons (mutant mouse epileptic), DNA- and RNA-binding activity</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>(124a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>JRKL</td>
<td>Jerky-like</td>
<td>H. sapiens/Mammals</td>
<td>Unknown</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>TIGD1</td>
<td>Tigger transposable element derived 1</td>
<td>H. sapiens/Primates?</td>
<td>Unknown</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>TIGD2, 3, 5–7</td>
<td>Tigger transposable element derived 2, 3, 5–7</td>
<td>H. sapiens/Mammals</td>
<td>Unknown</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>TIGD4</td>
<td>Tigger transposable element derived 4</td>
<td>H. sapiens/Amniotes</td>
<td>Unknown</td>
<td>DBD (CENPB) + core</td>
<td></td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/Tc2</td>
<td>POGK</td>
<td>pogo transposable element with KRAB domain</td>
<td>H. sapiens/Mammals</td>
<td>KRAB domain typically functions in transcriptional repression</td>
<td>DBD (CENPB) + core</td>
<td>KRAB</td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/Tc2</td>
<td>POGZ</td>
<td>pogo transposable element with ZNF domain</td>
<td>H. sapiens/Vertebrates</td>
<td>Unknown</td>
<td>DBD (CENPB) + core</td>
<td>ZnF</td>
<td>a</td>
</tr>
<tr>
<td>Tc1/mariner/ mariner</td>
<td>SETMAR</td>
<td>SET domain and mariner transposase fusion gene</td>
<td>H. sapiens/Anthropoid Primates</td>
<td>Binds DNA specifically, methylates histone H3 at K36 and facilitates DSB repair</td>
<td>DBD (HTH) + core</td>
<td>SET</td>
<td>(33, 113)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>rib</td>
<td>ribbon</td>
<td>D. melanogaster</td>
<td>Promotes epithelial cell migration and morphogenesis</td>
<td>DBD (HTHpsq)</td>
<td>BTB</td>
<td>(172a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>pfk</td>
<td>piefke</td>
<td>D. melanogaster</td>
<td>Binds polytene chromosome, nuclear protein present in larval salivary glands and ovaries</td>
<td>DBD (HTHpsq × 3)</td>
<td>BTB</td>
<td>(172b)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>psq</td>
<td>pipsqueak</td>
<td>D. melanogaster</td>
<td>Developmental regulator with pleiotropic functions during oogenesis, embryonic pattern formation, and adult development, binds GAGAG consensus motif, involved in formation of repressive chromatin</td>
<td>DBD (HTH&lt;sub&gt;psq&lt;/sub&gt; x 4)</td>
<td>BTB</td>
<td>(172b)</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>bab1</td>
<td>bric-a-brac1</td>
<td>D. melanogaster</td>
<td>Homeotic and morphogenetic regulator in development of ovaries, appendages and abdomen, binds to A/T-rich regions with TA or TAA repeats</td>
<td>DBD (HTH&lt;sub&gt;psq&lt;/sub&gt;)</td>
<td>BTB</td>
<td>(125a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>bab2</td>
<td>bric-a-brac2</td>
<td>D. melanogaster</td>
<td>Synergistic, distinct and redundant functions with Bab1 during imaginal development</td>
<td>DBD (HTH&lt;sub&gt;psq&lt;/sub&gt;)</td>
<td>BTB</td>
<td>(125a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>BtbVII</td>
<td>BTB-protein-VII</td>
<td>D. melanogaster</td>
<td>Unknown</td>
<td>DBD (HTH&lt;sub&gt;psq&lt;/sub&gt;)</td>
<td>BTB</td>
<td>(172b)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>Eip93F</td>
<td>Drosophila cell death protein E93</td>
<td>D. melanogaster</td>
<td>Directs steroid-triggered programmed cell death</td>
<td>DBD (HTH&lt;sub&gt;psq&lt;/sub&gt;)</td>
<td>BTB</td>
<td>(172b)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>Abp1</td>
<td>(ars)-binding protein 1</td>
<td>S. pombe / Schizosaccharomyces</td>
<td>Chromosome segregation and centromeric heterochromatin assembly, binds outer repeat of centromere, also required for efficient DNA replication through interaction with Cdk23</td>
<td>DBD (CENPB) + core</td>
<td>-</td>
<td>(124b, 142a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>Chh1</td>
<td>CENP-B homolog 1</td>
<td>S. pombe / Schizosaccharomyces</td>
<td>Chromosome segregation and centromeric heterochromatin assembly, binds outer repeat of centromere</td>
<td>DBD (CENPB) + core</td>
<td>-</td>
<td>(142a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>Chh2</td>
<td>CENP-B homolog 2</td>
<td>S. pombe / Schizosaccharomyces</td>
<td>Chromosome segregation, binds inner core region of centromere</td>
<td>DBD (CENPB) + core</td>
<td>-</td>
<td>(142a)</td>
</tr>
<tr>
<td>Tc1/mariner/pogo</td>
<td>Pdc2</td>
<td>Pyruvate decarboxylase 2</td>
<td>S. cerevisiae / Saccharomyces</td>
<td>Transcription activator of pyruvate decarboxylase and thiamin metabolism</td>
<td>DBD (CENPB) + core</td>
<td>-</td>
<td>(137a)</td>
</tr>
</tbody>
</table>

(Continued)
Table 2 (Continued)

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Full name</th>
<th>Original species/distribution</th>
<th>Functions and activities</th>
<th>Protein domains derived from Tpase</th>
<th>Other domains fused</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>piggyBac</td>
<td>PGBD1–5</td>
<td>PiggyBac-derived 1–5</td>
<td></td>
<td></td>
<td></td>
<td>(170)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H. sapiens/Mammals-Primates</td>
<td>Unknown</td>
<td>DBD?</td>
<td>+ core</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transib</td>
<td>RAG1</td>
<td>recombinational-activating gene 1</td>
<td>Interacts with RAG1 to catalyze V(D)J recombination in immune B and T cells</td>
<td>DBD?</td>
<td>+ core</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>H. sapiens/Jawed vertebrates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>far-red-impaired response protein 1, far-red elongated hypocotyl 3</td>
<td>Transcriptional activator binding upstream of FHY1 gene and triggering signaling cascade for far-red light sensing</td>
<td>DBD (WRKY)</td>
<td>+ core</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutator/Foldback/MULE</td>
<td>FRS1–11</td>
<td>FAR1-related sequences 1–11</td>
<td>Unknown</td>
<td>DBD (WRKY)</td>
<td>+ core</td>
<td>(115)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. thaliana/Angiosperms</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutator/Foldback/MULE</td>
<td>MUG1</td>
<td></td>
<td></td>
<td>DBD (WRKY)</td>
<td>+ core</td>
<td>(115)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A. thaliana/Angiosperms</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutator/Foldback/MULE</td>
<td>Aft1, Rcs1, Rbf1</td>
<td>activator of ferrous transport 1, 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S. cerevisiae/Saccharomycetales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutator/Foldback/MULE</td>
<td>DREF</td>
<td>DNA replication-related element-binding factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D. melanogaster/Drosophila</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hAT</td>
<td>ZBED1 [H. sapiens, D. melanogaster]</td>
<td>Human homolog of DRE, Zinc finger-BED domain containing protein 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hAT</td>
<td>ZBED4 [KIAA0637]</td>
<td>Zinc finger-BED domain containing protein 4</td>
<td></td>
<td></td>
<td></td>
<td>(197a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hAT</td>
<td>ZBED8 [421477]</td>
<td>Zinc finger-BED domain containing protein 4</td>
<td></td>
<td></td>
<td></td>
<td>(197a)</td>
</tr>
<tr>
<td>hAT</td>
<td>BEAF-32</td>
<td>boundary element-associated factor of 32 kDa</td>
<td>D. melanogaster/ Drosophilidae</td>
<td>Binds scs and other chromatin boundary elements and hundreds of sites on polyene chromosome, homodimerizes through C-term domain</td>
<td>DBD (BED) + hATC?</td>
<td>(207a)</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>---------------------------------------------</td>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------</td>
<td>--------</td>
</tr>
<tr>
<td>hAT/Charlie1</td>
<td>Buster1 (ZBED5)</td>
<td>Zinc finger BED domain containing protein 5</td>
<td>H. sapiens/Mammals</td>
<td>Translational repressor modulating interferon-γ-induced apoptosis, fused with part of eIF4G2 protein</td>
<td>DBD (BED) + core + hATC</td>
<td>(178a)</td>
</tr>
<tr>
<td>hAT</td>
<td>Daysleeper</td>
<td>Daysleeper</td>
<td>A. thaliana/Unknown</td>
<td>Essential for plant development, binds to motif upstream of Ku70 repair gene, likely transcription factor</td>
<td>DBD (BED) + core + hATC</td>
<td>(18a)</td>
</tr>
<tr>
<td>hAT/Charlie8</td>
<td>GTF2IRD2</td>
<td>GTF2I repeat domain containing 2</td>
<td>H. sapiens/Mammals</td>
<td>Fused to GTF2I domain of TFII-I transcription factor family with essential function in vertebrate development</td>
<td>DBD (BED) + core</td>
<td>(181a)</td>
</tr>
<tr>
<td>hAT (+ P element?)</td>
<td>LIN-15B</td>
<td>abnormal cell LImage family member 15B</td>
<td>C. elegans/Unknown</td>
<td>Key developmental regulator, interacts genetically with C. elegans homolog of mammalian retinoblastoma, inhibits G1-S cell-cycle transition</td>
<td>DBD (BED) + core + THAP</td>
<td>(27a)</td>
</tr>
<tr>
<td>hAT (+ P element?)</td>
<td>GON-14</td>
<td>gonadogenesis deficient, hnt15b family member</td>
<td>C. elegans/Unknown</td>
<td>Required for gonadogenesis and probably a pleiotropic transcriptional regulator of development (mutant shows gonad, vulval, growth, and cell division defects)</td>
<td>DBD (THAP) + hAT core + hATC</td>
<td>(27a)</td>
</tr>
<tr>
<td>P element (+ hAT?)</td>
<td>THAP0 (p52rPIK)</td>
<td>interferon-induced protein kinase-interacting protein, death-associated protein DAP4</td>
<td>H. sapiens/Vertebrates</td>
<td>Upstream regulator of interferon-induced translational repressor PKR by interaction with and inhibition of p58IPK</td>
<td>DBD (THAP) + hATC</td>
<td>(178a)</td>
</tr>
</tbody>
</table>

(Continued)
Table 2  (Continued)

<table>
<thead>
<tr>
<th>Related TE Superfamily/subgroup</th>
<th>Gene ID</th>
<th>Full name</th>
<th>Original species/distribution</th>
<th>Functions and activities</th>
<th>Protein domains derived from Tpase</th>
<th>Other domains fused</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>P element</td>
<td>THAP1</td>
<td>nuclear proapoptotic factor THAP1</td>
<td>H. sapiens/Vertebrates</td>
<td>Nuclear proapoptotic factor, binds DNA specifically and regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes</td>
<td>DBD (THAP)</td>
<td></td>
<td>(28a)</td>
</tr>
<tr>
<td>P element</td>
<td>THAP7</td>
<td>Thanatos-Associated Protein 7</td>
<td>H. sapiens/Vertebrates</td>
<td>Transcriptional repressor, binds histone H4 tail and recruits HDAC3 and NcoR to specific DNA sites, associates with template activating factor-IIbeta and inhibits H3 acetylation</td>
<td>DBD (THAP)</td>
<td></td>
<td>(126b)</td>
</tr>
<tr>
<td>P element</td>
<td>THAP2–5, 8, 10, 11</td>
<td>THAP-containing proteins</td>
<td>H. sapiens/ Mammals-Vertebrates</td>
<td>Unknown</td>
<td>DBD (THAP)</td>
<td></td>
<td>(28a)</td>
</tr>
<tr>
<td>P element</td>
<td>HIM-17</td>
<td>High Incidence of Males 7</td>
<td>C. elegans</td>
<td>Chromatin-associated protein required for initiation of meiotic recombination and chromosome segregation</td>
<td>DBD (THAP × 6)</td>
<td>coiled-coil</td>
<td>(161c)</td>
</tr>
<tr>
<td>P element</td>
<td>LIN-36</td>
<td>abnormal cell LINeage family member 36</td>
<td>C. elegans</td>
<td>Functions in vulval development as an inhibitor of the G1-to-S phase cell-cycle transition, regulates cell proliferation</td>
<td>DBD (THAP)</td>
<td></td>
<td>(10a)</td>
</tr>
<tr>
<td>P element</td>
<td>CDC-14B</td>
<td>cell-cycle regulator tyrosine phosphatase, isoform B</td>
<td>C. elegans</td>
<td>Cell cycle G1/S inhibitor, required for genome stability</td>
<td>DBD (THAP × 2)</td>
<td>cdc14</td>
<td>(28a)</td>
</tr>
<tr>
<td>P element</td>
<td>CTB-1</td>
<td>homolog of CtBP transcriptional corepressor</td>
<td>C. elegans</td>
<td>Vertebrate homolog CtBP is a global transcription corepressor critical for development and oncogenesis</td>
<td>DBD (THAP)</td>
<td>NAD&lt;sub&gt;b&lt;/sub&gt;</td>
<td>(28a)</td>
</tr>
<tr>
<td>Gene Name</td>
<td>Genus</td>
<td>Species</td>
<td>Transposon Type</td>
<td>Function</td>
<td>Core Domain</td>
<td>References</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>---------</td>
<td>-----------------</td>
<td>----------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>P element-</td>
<td>THAP9</td>
<td>H. sapiens</td>
<td>DBD (THAP)</td>
<td>+ partial core</td>
<td>(84)</td>
</tr>
<tr>
<td>P element-</td>
<td>P element-</td>
<td>homologous</td>
<td>gene</td>
<td>Polycomb (PcG) group proteins</td>
<td>+ partial core</td>
<td>(142)</td>
<td></td>
</tr>
<tr>
<td>Mammals-Amniotes</td>
<td>Mammals-Amniotes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>THAP–E2F6 fusion</td>
<td>THAP and cell-cycle transcription factor E2F6</td>
<td>D. rerio</td>
<td>DBD (THAP)</td>
<td>+ partial core</td>
<td>(84)</td>
</tr>
<tr>
<td>Fish</td>
<td>Fish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amphibians</td>
<td>Amphibians</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>PI-F/Harbinger</td>
<td>HARBI1, nuclear apoptosis-inducing factor 1</td>
<td>H. sapiens</td>
<td>DBD (HTH)</td>
<td>+ partial core</td>
<td>(126a)</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>Vertebrates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>P-neogene</td>
<td>P-tsa casi stationary truncated</td>
<td>D. tsacasi</td>
<td>DBD (THAP)</td>
<td>+ partial core</td>
<td>(160a)</td>
</tr>
<tr>
<td>D. tsacasi</td>
<td>D. tsacasi</td>
<td></td>
<td></td>
<td>Unknown, but binds DNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>P-neogene</td>
<td>P-neo G and A type</td>
<td>D. subobscura</td>
<td>DBD (THAP)</td>
<td>+ partial core</td>
<td>(134)</td>
</tr>
<tr>
<td>D. subobscura</td>
<td>D. subobscura</td>
<td></td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P element</td>
<td>P element</td>
<td>P-neogene</td>
<td>P-neo montium</td>
<td>D. montium</td>
<td>DBD (THAP)</td>
<td>+ partial core</td>
<td>(162a)</td>
</tr>
<tr>
<td>D. montium</td>
<td>D. montium</td>
<td></td>
<td></td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PI-F/Harbinger</td>
<td>PI-F/Harbinger</td>
<td>DPLG1–7</td>
<td>Drosophila PIF-like gene 1–7</td>
<td>Drosophila</td>
<td>DBD (HTH)</td>
<td>+ partial core</td>
<td>(26a)</td>
</tr>
<tr>
<td>PI-F/Harbinger</td>
<td>PI-F/Harbinger</td>
<td>DPM7</td>
<td>Drosophila PIF MAD-like gene</td>
<td>Drosophila</td>
<td>DBD (HTH)</td>
<td>+ partial core</td>
<td>(26a)</td>
</tr>
<tr>
<td>Maverick?</td>
<td>Maverick?</td>
<td>KRBA, ZNF452</td>
<td>c-integrases</td>
<td>H. sapiens</td>
<td>RVE core</td>
<td>+ partial core</td>
<td>(57a)</td>
</tr>
</tbody>
</table>

DBD = DNA-binding domain.
Core = transposase catalytic core domain.

---

2. L. Sinzelle & Z. Ivics, personal communication.
derive from a Maverick transposon (52, 57a), and a MADF domain-containing protein in Drosophila that was domesticated from the accessory protein of a PIF-like transposon (26a).

In addition, several proteins listed in Table 2 are chimeric proteins that result from the fusion of transposase-derived domains to domains of other origins. This process is consistent with the modular evolution of proteins in general and the concept of evolutionary tinkering introduced by François Jacob (82).

Only a small fraction of the genes listed in Table 2 have been studied functionally. Thus, in most cases, the functional contribution of the transposase domain(s) to the corresponding protein remains a matter of speculation. However, one can draw several predictions based on the functional analyses of related transposases. All eukaryotic transposases that have been biochemically characterized possess two functionally separable domains: a N-terminal region that binds to the ends of the cognate transposons (generally the TIRs) and a central or C-terminal core region that catalyzes the cleavage and transfer steps of the transposition reaction (35). Any of these activities can be potentially co-opted to serve cellular function(s) and, as we outline below, there is now evidence that these activities have been differentially retained in different transposase-derived proteins. Nonetheless, a recurrent theme is the recycling of transposase DNA-binding domain (DBD) to build transcription factors (Table 2).

As long-term genomic residents coevolving with their host, transposases are expected to have developed a number of interactions with host proteins, even though these interactions may not be strictly required for transposition (107). For example, the Sleeping Beauty transposase interacts directly with the Ku70 repair protein, the DNA-bending high-mobility group protein HMGB1 and the transcription factor Miz-1 (81, 188, 201). Each of these proteins has a large number of interacting partners, and interaction with the Sleeping Beauty transposase may influence and modulate their cellular function. Similarly, the pogo transposase of D. melanogaster interacts with PCNA, the proliferating cell nuclear antigen, a key protein for DNA replication and repair (191). A functional PCNA-binding motif is also present in Tigger1, the human relative of pogo transposase (191), and a similar motif is present at a comparable position in the Arabidopsis pogo-like transposase Lemi1 (50), suggesting that PCNA interaction with pogo-like transposases is evolutionary conserved.

The association of transposons with DNA repair and replication factors appears as a recurrent theme. It is easy to conceive how this association could benefit both transposon and host. In turn, these interactions may predispose the transposase to domestication and be preserved in transposase-derived proteins. This might explain why several transposase-derived proteins appear to be involved in cell cycle control, recombination, and other aspects of chromosome dynamics (Table 2).

Below we recount three tales of transposase domestication and discuss the evolutionary consequences of these innovations.

The origin of the adaptive immune system of jawed vertebrates. V(D)J recombination is the process by which a virtually infinite population of distinct antibodies can be generated in B and T lymphocytes. The acquisition of V(D)J recombination is often regarded as a key step in the evolution of the adaptive immune system of jawed vertebrates (32, 88). The two essential components of V(D)J recombination are (i) the RAG1 and RAG2 proteins, which interact to form the recombinase responsible for the joining and transfer activities; and (ii) the recombination signal sequences (RSS) flanking the V (variable), D (diversity), and J (joining) segments, which define the specific sequences bound, cleaved, and joined by the RAG1/2 protein complex (58). The analogy of the process of V(D)J recombination to a transposition reaction is striking. RAG1/2 can catalyze transposition of a DNA segment flanked by RSS in vitro (3, 74) and in vivo in yeast (28) and mammalian cells (27, 162). Additionally, it has been
observed that several eukaryotic transposases utilize a cleavage chemistry similar to that seen in V(D)J recombination (37, 208). However, until recently no transposase directly related to RAG1/2 had been identified.

Evidence of this relationship came from the discovery that RAG1, which provides the catalytic core for the reaction, is closely related in sequence to transposases encoded by Transib elements, a group of DNA transposons recently identified in the genomes of diverse invertebrates (93). Additional support for the relationship came from comparisons of the structure of the RSS to the TIRs of Transib transposons and the conservation of spatial and sequence characteristics [the so-called 12/23 rule; (58)]. Finally, Transib elements provoke a 5-bp TSD upon transposition, as do most cut-and-paste reactions mediated by RAG1 in vitro (3, 74). Together the data leave little doubt that V(D)J recombination is the product of a fortuitous event of DNA transposon domestication, an event that may be viewed as a crucial step in vertebrate evolution.

Light-sensing in plants and the FAR1/ FHY3 family of transcription factors. As sessile organisms, higher plants have evolved a network of photoreceptors to sense light changes in the environment (130). Among the photoreceptors, the phytochrome A (phyA) pathway has been extensively characterized. Photoactivation leads to the conversion of phyA into an active form allowing its import to the nucleus from the cytoplasm (190). Once in the nucleus, phyA is thought to directly activate a set of transcription factors, which in turn induce a molecular cascade resulting in light-mediated photomorphogenic responses (87, 181, 190). PhyA accumulation in the nucleus is dependent on the presence of two homologous proteins, FHY1 and FHL (R. Lin, L. Ding, C. Casola, D. Ripoll, F. Nagy, C.F., H. Wang, submitted). Unexpectedly, it turns out that FHY3 and FAR1 are members of an ancient gene family that is related to MULE transposases (80, 115). Evolutionary analyses indicate that the entire FHY3/FAR1 family is most likely derived from a single domestication event of a MULE transposase at the dawn of angiosperm evolution (R. Lin, L. Ding, C. Casola, D. Ripoll, F. Nagy, C.F., H. Wang, submitted). This domestication event would coincide with the origin of FHY1 and FHL and with the early evolution of the phyA pathway (130).

FAR1 and FHY3 have a specific DBD located in the N-terminal region of the protein. This region is conserved in the transposases encoded by the most closely related MULEs found in modern plant genomes (4). It is tempting to speculate that the binding sites of FHY3/FAR1 are themselves derived from the TIRs of ancient MULE transposons that integrated upstream of the target genes regulated by FHY3 and FAR1, including FHY1 and possibly other targets (80). In this model, not only the transposase but also its unlinked binding sites, dispersed in the genome as the result of the past propagation of MULEs, could have been co-domesticated to establish a regulatory network (see below). Finally, note that FHY3 possesses intrinsic transcriptional activation ability that is separable from its DNA-binding activity (R. Lin, L. Ding, C. Casola, D. Ripoll, F. Nagy, C.F., H. Wang, submitted). This activity requires residues located within the predicted catalytic domain of the ancestral transposase that are also conserved in distant MULE transposases. This observation indicates that many MULE transposases might have intrinsic transcription factor activity, and it would explain why several MULE transposases seem to have been domesticated repeatedly during eukaryotic evolution (4, 34).

The primate SETMAR fusion gene. The two examples described above show that transposase domestication events have been
instrumental in the emergence of key innovations both in vertebrates and flowering plants, respectively. In both cases, it seems that not only the transposase but also sequences present at unlinked transposon copies were co-domesticated. We hypothesize that the fundamental property of transposase molecules to recognize and act in trans on multiple DNA elements dispersed throughout the genome is a major factor contributing to their recurrent domestication in eukaryotes. Indeed, transposase domestication can be instantly accompanied by the selective recruitment of a ready-to-use network of binding sites in the genome (Figure 3).

To test this model and better understand the early steps of transposase domestication, it is necessary to study relatively recent exaptation events, where the transposase and its associated binding sites would be still readily recognizable as being derived from the same transposon family. We believe that the example of SETMAR described in the next section provides the ideal system to study the early steps of this model.

SETMAR is a human gene first identified by Hugh Robertson as a particular copy of the Hsmar1 family, one of two mariner-like families found in the human genome (165). SETMAR originates from the transcriptional fusion of a SET domain to a mariner transposase. The function of the SETMAR protein is currently unknown, but in vitro experiments have shown that the SET portion of SETMAR has specific histone methyltransferase activity for lysine 36 of histone H3 (113). The function of this epigenetic mark is not well understood in mammals, but studies in yeast indicate that it may act as a repressive chromatin mark to prevent...
spurious intragenic transcription (25, 99). In addition, overexpression of SETMAR in human cells facilitates DNA repair via the non-homologous end-joining pathway (113). However, the contribution of the transposase domain to this activity and to the function of SETMAR remains unclear.

In order to gain further insights, comparative genome sequencing was used to trace the origin of SETMAR and delineate the steps leading to the fusion of SET and MAR domains (33). The results show that SETMAR has emerged between 58 and 40 Mya in an anthropoid primate ancestor, through an intricate process involving transposition of a 

\[ \text{mariner} \]

transposon downstream of a pre-existing gene encoding a stand-alone SET domain, followed by genomic deletion of intervening DNA and creation of a new intron. The transposase region of SETMAR has been subject to strong evolutionary constraint in all extant major lineages of anthropoid primates, suggesting that the addition of a transposase domain to the pre-existing SET domain led to the advent of a beneficial new function in primates. The signature of purifying selection has been particularly intense on the N-terminal region of the transposase containing the predicted DBD, whereas the catalytic domain appears to evolve essentially neutrally (33).

Consistent with these predictions, biochemical studies indicate that SETMAR is deficient for cleavage at the 3' ends of the element (123, 136), but has retained the ability to bind specifically DNA through its N-terminal DBD, assemble a paired-ends complex, and inflict single-strand nicks on adjacent DNA (33, 123, 136). Furthermore, SETMAR has retained strong specificity for binding to a 19-bp site located within the TIR of the related Himar1 or MADE1 transposons.

Figure 3

Model for the assembly of a regulatory network by domestication of a transposase and its binding sites.

(a) Initial transposase domestication event. A family of DNA transposon is shown with autonomous and nonautonomous copies dispersed in the genome. Each TIR (black arrowhead) contains a binding site for a transposase encoded by autonomous copies (pink/yellow boxes). Flanking host genes are shown as grey boxes. One of the transposase genes (yellow box) is recruited. In this example, recruitment is promoted by transcriptional fusion of the transposase to a flanking host gene (blue box) encoding a regulatory domain, leading to the expression of a fusion protein with transposase (yellow) and regulatory domains (blue). This is similar to the emergence of SETMAR, which arose by fusion of a 

\[ \text{mariner} \]

transposase with an adjacent gene encoding a SET domain. Note, however, that transposase domestication does not need to involve fusion with another domain, particularly if the transposase itself possesses regulatory activity, as demonstrated for 

\[ \text{FHY3} \]

, a transcription factor in Arabidopsis entirely derived from a 

\[ \text{Mutator} \]

transposase. (b) Immediate consequences of transposase domestication. The translational fusion immediately allows the regulatory domain to be tethered to all the sites in the genome recognized by the transposase, i.e., the TIRs of all the transposon copies previously dispersed in the genome. Depending on the genomic environment of the transposons, binding of the fusion protein might have various effects on the expression of the surrounding genes: activation, repression, or no effect. These effects are symbolized by the blue arrow acting on the adjacent gene. (c) Binding sites selection. Natural selection will retain interactions that provide an immediate benefit to the host and will eliminate deleterious interactions. Site elimination (red cross) may occur through substitutions or deletion driven by positive selection. Sites that are selectively neutral (with no positive or negative impact on adjacent genes) are expected to evolve neutrally and most will eventually disappear. Mobility of the transposons at this stage (if it persists) might accelerate the shaping of the network through transposon excision events and/or fixation of new advantageous insertions. (d) A regulatory network is born. The end result is the assembly of a regulatory network, where the domesticated transposase and a subset of its ancestral binding sites conferring beneficial interactions are evolving under purifying selection, while the rest of the transposons are eroded by mutations. Note that the system also provides an intuitive opportunity for the establishment of a feedback loop "F" (positive or negative) through domestication of binding sites that were originally linked to the domesticated transposase.
The binding site is dispersed in over 1500 conserved copies throughout the human genome and nearly all of these sites map within the TIRs of the related transposons. Together these data support a model whereby the specific DNA-binding activity of the transposase region now provides a means to target the SET domain to specific sites within the human genome (33) (Figure 3). For this model to be validated, it will be necessary to pinpoint the DNA targets of SETMAR and determine the effect of tethering the protein to specific chromosomal sites.

### SUMMARY POINTS

1. The great diversity of DNA transposons can be organized into 3 major subclasses: cut-and-paste transposons, with ten major superfamilies; rolling-circle transposons (Helitrons); and self-replicating transposons (Mavericks).

2. Almost all subclasses and superfamilies are represented in a wide range of eukaryotes, including various protozoans. Thus, DNA transposons diversified very early in evolution and have been maintained in all major branches of the eukaryotic tree of life.

3. Vast variations occur among species in the level of amplification of their DNA transposon populations. Different amplification of DNA transposons among species may or may not translate into substantial differences in genome size, but probably reflect a complex combination of intrinsic (host- or self-mediated) and extrinsic (environmental, ecological) factors modulating the activity and retention of transposon activity over evolutionary time.

4. The evolutionary success and astonishing diversity of eukaryotic DNA transposons offer an intriguing paradox because their amplification dynamics seems to represent an evolutionary dead end favoring the proliferation of non-autonomous derivatives (MITEs) to the detriment of the autonomous copies. We propose a more subtle vision whereby the accidental amplification of MITEs passes under the radar of the host defense system and drives the diversification of autonomous copies.

5. Like other TEs, DNA transposons play a significant role in shaping eukaryotic genomes, but they possess specific features that enhance or accentuate some of their influence on the host. These features include their capacity to excise imprecisely, jump locally, inflict multiple double-strand breaks, and undergo alternative transposition.

6. DNA transposons have been a recurrent source of coding sequences for the emergence of new genes. We propose that this is a pervasive pathway to create a genetic network as the recruitment of transposase DNA-binding domains opens the door for selection to instantly retain a set of unlinked binding sites previously dispersed in the genome and/or co-opt their interactions with host proteins.

### DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.
ACKNOWLEDGMENTS

We are grateful to Sue Wessier for continuous encouragement and inspiration prior to and throughout the preparation of this review. We thank members of our labs and of the Genome Biology Group at UT Arlington, Dominique Anxolabéhére (Institut Jacques Monod Paris, France), Alfredo Ruiz (Universitat Autònoma de Barcelona, Spain), and Xiaoyu Zhang (UCLA), for comments on the manuscript and helpful discussions. We are also thankful to Zoltán Ivics and Ludivine Sinzelle (Max Delbruck Center, Berlin, Germany) for communicating data prior to publication. This research is supported by start-up funds from UT Arlington and by the National Institute of Health grant R01GM77582-01 to C.F. and grant R01AI068908-01 to J. Kissinger, University of Georgia, with subcontract to E.J.P. and C.F.

LITERATURE CITED

16. Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–34
26a. Casola C, Lawing AM, Betran E, Feschotte C. 2007. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* In press


77. Hsia AP, Schnable PS. 1996. DNA sequence analyses support the role of interrupted gap repair in the origin of internal deletions of the maize transposon, *MuDR*. *Genetics* 142:603–18
93. Kapitonov VV, Jurka J. 2005. RAG1 Core and V(D)J recombination signal sequences were derived from transib transposons. PLoS Biol. 3:e181
110. Lampe DJ, Walden KK, Robertson HM. 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. 18:954–61


180. Stehelin D, Varmus HE, Bishop JM, Vogt PK. 1976. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature 260:170–73


