Mol Gen Genet (1996) 252:761–766

SHORT COMMUNICATION

Hugh M. Robertson

Members of the pogo superfamily of DNA-mediated transposons in the human genome

Received: 8 March 1996 / Accepted: 24 June 1996

Abstract A new superfamily of transposons from fungi, nematodes, and flies related to the pogo element of *Drosophila melanogaster* was recognized that represents a branch of the extended superfamily of transposase and integrase proteins sharing a common D,D35E catalytic domain. Searches of human sequences in the public databases for similarity to this domain revealed at least two members of this new superfamily, with many highly mutated copies, in the human genome. A full-length consensus was constructed for one of them, which includes the MER37 medium reiteration frequency sequence recognized previously, from 343 human sequence accessions (261 of which are unique). Most of these were Expressed Sequence Tags, some were Sequence-Tagged Sites, and a few are from long genomic sequences. The 2417 bp consensus has the hallmarks of a pogo superfamily transposon, including 12 bp inverted terminal repeats, and encodes two long open reading frames. The first ORF encodes a polypeptide with 42% amino acid sequence identity to pogo in the D,D35E region. The second element shows 49% amino acid sequence identity with the first, and 40% with pogo in this region. These elements coincide with those described recently as Tigger1 and Tigger2, respectively. These transposons appear to have been active 80-90 Myr ago in the genome of an early primate or primate ancestor.

Key words Transposable element · Human genome · Tigger · MER37 · pogo

Communicated by D. J. Finnegan

H. M. Robertson
Department of Entomology, University of Illinois at Urbana-Champaign, 505 S. Goodwin, Urbana, IL 61801, USA

Introduction

Transposable elements that move via a cut-and-paste mechanism, also known as DNA-mediated or Class II transposons (Finnegan 1989), are known in almost all well-studied organisms (see Berg and Howe 1989). Until recently mammals were a major exception; however, two distinct mariner family transposons have now been discovered in the human genome by several groups (Auge-Gouillou et al. 1995; Morgan 1995; Oosumi et al. 1995; Reiter et al. 1996; Smit and Riggs 1996; Robertson et al. 1996). The individual copies of these transposons are highly mutated from their presumed ancestrally active sequences, differing from their consensus sequences by 10–20% in DNA sequence, and usually having suffered multiple insertions and deletions. In each case, a few full-length genomic sequences are known, but most of the available sequences are from various human Expressed Sequence Tag (EST) projects (see Boguski 1995) and human genome mapping projects employing Sequenced-Tagged Sites (STSs) (e.g. Green et al. 1991), as well as PCR fragments generated from internal sequences (Auge-Gouillou et al. 1995; H. M. Robertson and K. L. Zumpano, unpublished data). These "molecular fossils" are thought to represent an element that was once active in the genome of an early primate.

In an effort to discover the transposase source responsible for generation of numerous short inverted repeat pairs in the human genome, Smit and Riggs (1996) also similarly constructed consensus sequences for two transposons in the human genome related to the pogo element of *Drosophila melanogaster*, called Tigger1 and Tigger2. Here I describe an independent derivation of consensus sequences for these elements, report the possible existence of a third rare Tigger element, and date the origin of Tigger1 in the primate genome.

The pogo transposable element was discovered as an insertion in the *white* eye color gene of *D. melanogaster*

(Tudor et al. 1992), and all strains examined, but not other drosophilid species, have copies of this transposon (Boussy et al. 1993). A putatively full-length element of 2120 bp called pogoR11 was sequenced, revealing 21 bp inverted terminal repeats and two ORFs separated by a short intron. When spliced, the mRNA encodes a 499 amino acid protein presumed to be the pogo transposase. The only amino acid sequence similarity these authors could detect in the public databases was an enigmatic similarity to a major human centromere-associated antigen (CENP-B) (Earnshaw et al. 1987). Since then an equally enigmatic similarity to the *jerky* gene of mice, which is involved in epileptic seizures, has been recognized (Toth et al. 1995). A more distant similarity was recognized by Doak et al. (1994) in their description of an extended superfamily of transposases that encode a D,D35E putative catalytic domain with similarity to the catalytic domain of retroviral and retrotransposon integrases. This extended superfamily includes the mariner and Tc1 families (see Robertson 1995). They noted that the pogo transposase had a similar domain, but because the final conserved glutamic acid was not present, they did not include it in the superfamily.

Results and discussion

Searches of the public databases using TBLASTN (Altschul et al. 1990) revealed that the D,D35E region of the pogo transposase has significant similarity to the transposases of several other transposons. The most similar are three elements from fungi, the Fot1 element of *Fusarium oxysporum* (Daboussi et al. 1992), the Pot2 element of *Magnaporthe grisea* (Kachroo et al. 1994), and the Fcc1 element of *Cochliobolus carbonum* (Panaccione et al. 1996). A large fragment of a closely related element is present in the 3′ region of the *agdA* gene of *Aspergillus oryzae* (Minetoki et al. 1995). The transposases of these elements are approximately the same length as that of pogo and they share 21–26% amino acid sequence similarity in the central conserved and alignable D,D35E region (the N and C termini are difficult to align with those of pogo) (see Fig. 1). Three transposons from the nematode *Caenorhabditis elegans* are more distantly related to pogo. The Tc4 (Yuan et al. 1991; Li and Shaw 1993) and Tc5 (Collins and Anderson 1994; EMBL Z35400) elements encode transposases with 19 and 14% identity to the D,D35E region of pogo, respectively, and again the N- and C-termini are quite divergent and the N-termini are considerably longer. Finally, the sequenced copy of the Tc2 element (Ruvolo et al. 1992) encodes similarity to the region surrounding the first Asp residue, but because the published copy has suffered many mutations, it is difficult to include it in comparisons. All of these transposons share another feature that supports their relationship to the D,D35E superfamily: the duplication of a target TA upon insertion. They can all be considered members of the pogo superfamily of transposons, named for the original pogo element of *D. melanogaster*, with homology and probably similar catalytic activity to the extended D,D35E superfamily of transposases and integrases (see Craig 1995; Grindley and Leschziner 1995). Three families can be recognized within this pogo superfamily, those similar to the Tc4, Fot1, and pogo transposases.

A search of the public databases with the D,D35E region of the pogo transposase using TBLASTN detected several human sequences encoding significantly similar amino acid sequences. Database searches with these sequences in turn, using BLASTN, revealed an overlapping series of sequences that extended in each direction. Repeated searches with the ends of each newly constructed consensus of these sequences extended the consensus to 2417 bp before similarity of the sequences ended. Included in these sequences are a set recognized as a medium reiteration frequency sequence called MER37 in the June 1994 Release 2.1 of the Reference Collection of Human Repetitive Elements (see Jurka et al. 1992). The sequences were aligned by eye in the editor of PAUP v3.1.1 for the Macintosh (Swofford 1993) and a preliminary majority rule consensus constructed by eye. This consensus was searched against the public databases to recover sequences missed in the initial compilation, and then evaluated and corrected using the "match first" capability of MacClade v3.0 for the Macintosh (Maddison and Maddison 1993). This consensus sequence is available from GenBank as accession number U49973. It is one base pair shorter than the Tigger1 consensus of Smit and Riggs (1996) and differs by 2% nucleotide divergence. These differences are due to uncertainties in the construction of the consensus described below, and will soon be resolvable with the acquisition of more genomic sequences by the Human Genome Sequencing Project.

Altogether, 343 human sequence accessions had significant similarity, generally over 75% nucleotide sequence identity, to the consensus. The majority of these were ESTs, most of which were from the WashU/Merck I.M.A.G.E. consortium (see Boguski 1995), some were STSs, and a few were from long genomic sequences (see below). Some of the ESTs were essentially identical to each other, presumably from equivalent cDNA clones from the same genomic copy, however none matched the genomic sequences. The remainder appear to represent independent copies of this transposon; however, the ESTs and STSs are generally less than 500 bp in length, so it is possible that some derive from different regions of the same transposon copy. Altogether 261 unique sequences were used to build the consensus.

The majority rule consensus was generally staightforward to derive because the minimum coverage for

any region of the transposon was 25 sequences, however there were two difficulties. First, near the 5′ and 3′ ends (positions 55–60 and 2249–2256, respectively) there are runs of thymines that vary in number from 4–12 and from 2–10, respectively, in the various sequences; thus, a confident consensus could not be determined for these runs. These regions are not in the ORFs so they do not affect the coding capacity of the element, only its length. Second, at about 100 positions in the sequence the transition pairs C/T and A/G were equally common or nearly so, usually as a CA/TG doublet, so it was not possible to establish a confident consensus. This phenomenon has been observed before in neutrally evolving human transposon sequences [e.g. Britten et al. (1988) for Alu elements], and is ascribed to the rapid evolution of a CG pair to CA or TG by deamination of methylated cytosine after insertion of the transposon copy. Unfortunately there is no clear criterion for considering such a site to have originally been CG, so a strict majority rule consensus was employed, with only ties being decided in favor of C or G.

The consensus has 12 bp perfect inverted terminal repeats (ITRs) (CAGGCATACCTC), which resemble those of some other pogo superfamily elements, as well as Tc1 family elements, in starting with CAG. The consensus 2 base pairs flanking these repeats in the 40 sequences at the 5′ end and the 33 sequences at the 3′ end are TA, again in keeping with this element being a member of the pogo superfamily and D,D35E catalytic superfamily. Beyond this apparent duplication of a TA target site there is very little similarity in the flanking DNA sequences that would indicate a larger target sequence.

The consensus has two long ORFs, ORF1 and ORF2, corresponding to positions 425–1786 and 1790–2203, encoding 454 and 138 amino acids, respectively. These ORFs are separated by a single in-frame stop codon that is encoded by 14 of the 25 sequences at this position (the remaining 11 sequences have various mutations from this consensus, rather than a single alternative codon that might represent an ancestrally active sequence). Smit and Riggs (1996) suggest that the ORFs might be joined in a mature mRNA by splicing of an intron containing this single stop codon, as are the two ORFs of the pogo element, however this does not appear to be the case. First, there is no indication of consensus 5′ or 3′ splice sites in the surrounding DNA sequence. Second, many of the sequences crossing this region are from cDNAs from which any intron should have been spliced (even if these cDNAs do not represent authentic Tigger1 transcripts, as is likely given the mutated condition of these elements, but rather illegitimate or read-through transcripts). Sequences characteristic of promoters are not obvious in the 400 bp preceding ORF1, however ORF2 is followed by a consensus polyadenylation site (AATAAA) at positions 2218–2223, 11 bp after the stop codon at positions 2204–2206. It is therefore unclear how these ORFs are expressed. The amino acid sequence encoded by ORF2 has no significant similarity to any peptide in the public databases.

ORF1 encodes a polypeptide with 42% identity to pogo transposase in the central conserved D,D35E domain (see Fig. 1) and is alignable with pogo transposase for most of its length. This element is therefore considered to be a member of the pogo family of transposons. There is a second member of the family in the human genome represented by 36 sequences (27 unique), which when translated yields a reading frame encoding a product with 49% identity to Tigger1 and 40% to pogo in the central conserved region (Fig. 1). The 5′ and 3′ regions of this transposon could not be identified because the consensus is based on so few sequences that it was not possible to extend it; however, because Smit and Riggs (1996) discovered these transposons by virtue of internally truncated versions in the human genome, they were able to extend their consensus of Tigger2 to 2708 bp. This consensus must be considered preliminary however, because it does not encode a full-length ORF1 equivalent to that of Tigger1. A third Tigger transposon might be represented by the sequence of WashU/Merck EST clone 126906 (GenBank R07817), but no overlapping clones to construct a consensus were found in GenBank.

The number of copies of these transposons in the human genome is difficult to determine; however, the observation that none of the ESTs or STSs come from eight long genomic copies indicates that the copy number of Tigger1 is high. As noted above, the copy number of Tigger2 is approximately four-fold lower (the equivalent region of Tigger1, from approximately bp 700–1900, is represented by 131 unique sequences), while the third possible element must be present in just a few copies.

These copies of the elements appear to be "molecular fossils", copies of an originally active transposon that have subsequently accumulated many mutations. Thus, each sequence differs from the Tigger1 consensus by 10–25% DNA divergence, and similar divergences apply to Tigger2. Each sequence also differs from the consensus by many small insertions and deletions. In addition, the long genomic sequences reveal the effects of more severe mutations including large insertions, commonly of Alu elements, large internal deletions, and apparent truncations of termini. For example, the most intact genomic copy is in intron 11 of the NFκB1 gene for the p50 and p105 subunits of transcription factor NFκB (bp 5828–8193 of EMBL Z47735). This 2230 bp copy has 19 small deletions and 7 small insertions relative to the consensus, a large deletion of 104 bp from consensus positions 573–676, a 136 bp insertion at position 1086 that is part of an Alu element, and otherwise differs from the consensus by 11.5% DNA divergence. A similarly intact 2287 bp copy is in the second intron of the CRFB4 cytokine receptor gene (bp 9696–6782 of GenBankU08988; Lutfalla et al. 1995),

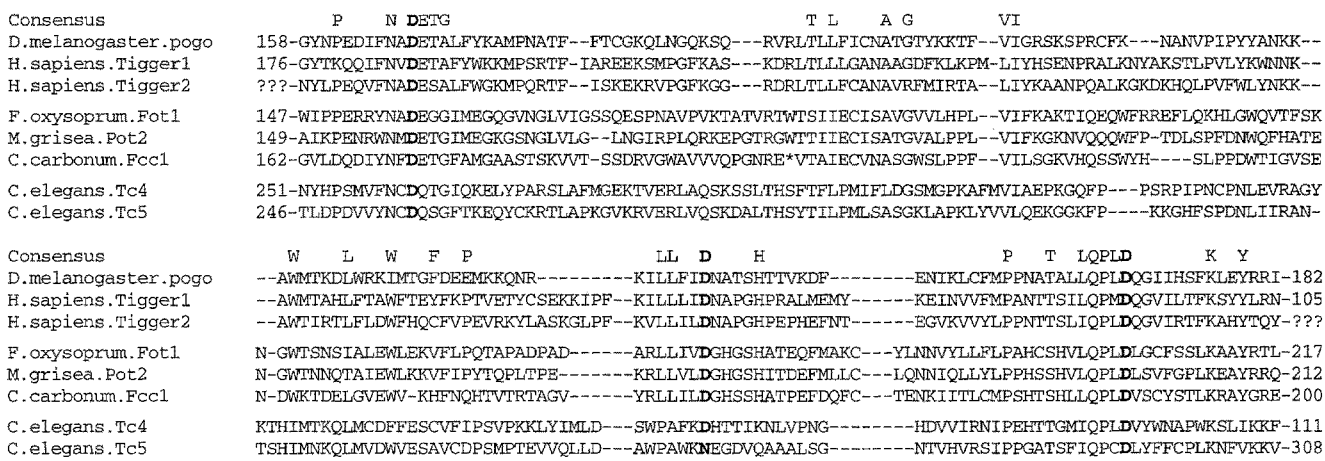

Fig. 1 Alignment of the central conserved region of pogo superfamily transposases, equivalent to the D,D35E transposase/integrase catalytic domain. The pogo, Fot1, and Tc4 family sequences are separated for clarity. Consensus amino acids are shown for positions shared by the majority of members of at least two of the families. The Asp residues of the D,D35E-equivalent domain are highlighted in *boldface* [see Smit and Riggs (1996) for alignment with *mariner* and *Tc1* transposases]. The number of amino acids on either side of this region is indicated for each transposase, except the second human pogo family element for which the ends are not known. The *asterisk* in the Fcc1 sequence indicates an in-frame stop codon in the copy sequenced

most of which had been recognized as a MER37 repeat. This copy has 15 small deletions and 8 small insertions relative to the consensus, a deletion of 100 bp at consensus positions 380–480, two Alu insertions of 311 and 316 bp at positions 812 and 1472 respectively, and otherwise differs from the consensus by 10.9%.

Recognition of this consensus sequence allows one to account for some other genomic sequences. For example, the first 168 bp of the *Blym-1* transforming gene (GenBank K01884; Diamond et al. 1984) corresponds to consensus positions 1793–2000, and includes the putative 5′ regulatory regions and the first exon of this gene. This may be an example of recruitment of transposon sequences for functional roles in the genome. A quite different fate befell another Tigger1 copy, which was the substrate into which an approximately 1200 bp retropseudogene of calmodulin (CaMII-ψ2) inserted (GenBank X52995; Koller et al. 1991). The approximately 1000 bp of sequenced flanking DNA does not reach to the 5′ and 3′ ends of the consensus, but differs from it by 15.7%, 10 deletions, and 2 insertions.

The high level of divergence of the individual copies from an ancestrally active sequence indicates that they were formed many millions of years ago when this transposon was last active. An approximate dating can be obtained by three measures. First, assuming a neutral rate of evolution in primates of 0.16% per Myr (e.g. Britten 1994), the average divergence of all 261 unique Tigger1 sequences from the consensus of 14.8% (14.3% for the eight genomic copies) indicates that the copies were inserted approximately 90 Myr ago. Second, a rate for accumulation of insertions and deletions (indels) (excluding Alu and other transposons) in noncoding nucleotide sequence of primates has been determined by Saitou and Ueda (1994) as 0.15/kb/Myr for older comparisons among primates. The numbers and kinds of indels seen in the long genomic Tigger1 copies are similar to those they examined, in that single base pair events predominate, and deletions are approximately twice as common as insertions. Assuming again that since their origin these Tigger1 copies have been evolving neutrally at a similar rate to the non-coding pseudogene sequences they examined, the number of indels per kb of Tigger1 genomic sequences is 13 (126 in a total of 9800 bp), indicating that these copies were formed around 86 Myr ago. Third, the ages of different subfamilies of Alu elements have been determined (Kapitonov and Jurka 1996), so insertions of members of particular Alu subfamilies set lower limits on the ages of the Tigger1 copies. Most of the Alu elements in the genomic Tigger1 copies are from various S subfamilies, including the older Sp and Sq subfamilies that inserted approximately 40 Myr ago. However, the first MER37 sequence recognized was a 119 bp region in the middle of a large cluster of Alu elements near the *Bat2* gene (Iris et al. 1993; EMBL Z15025), and comparison with the Tigger1 consensus shows that this is in fact a 400 bp region of the 5′ end of a Tigger1 (the rest may be in an adjacent unsequenced cosmid). Five of seven Alu sequences inserted into this Tigger1 fragment are from the J subfamilies, which inserted about 80 Myr ago and differ from their consensus by 13% on average (Kapitonov and Jurka 1996). They are therefore slightly younger than Tigger1 copies, again consistent with the origin of Tigger1 copies between 80 and 90 Myr ago. This dating would predict that Tigger1 copies will be found in the genomes of all primates, including prosimians, and perhaps other closely related mammals (Martin 1993).

Searches of the public databases for further examples of Class II DNA-mediated transposons in the human

genome have been negative. These include the Tc1 family which is extremely widely represented in animals (see Robertson 1995), the hAT superfamily from flies and plants (see Calvi et al. 1991), the P family from flies (see Clark et al. 1994), and the TTAA-specific elements of Lepidoptera (see Cary et al. 1989). In the case of the Tc1 family, a PCR assay designed to detect family members was also negative for human DNA (Avancini et al. 1996), supporting the contention that there are probably no members of this family in our genome. At the current rate of acquisition of human genome sequences it should soon become clear whether the two mariner and two pogo family elements discovered so far are the only DNA-mediated transposons in the human genome.

Acknowledgements I thank Felipe Soto-Adames for comments on the manuscript. This work was supported by NSF grant MCB-9317586.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Auge-Gouillou C, Bigot Y, Pollet N, Hamelin M-H, Meunier-Rotival M, Periquet G (1995) Human and other mammalian genomes contain transposons of the *mariner* family. FEBS Letters 368:541–546

Avancini RMP, Walden KKO, Robertson HM (1996) The genomes of most animals have multiple members of the Tc1 family of transposable elements. Genetica, in press

Britten RJ (1994) Evidence that most human *Alu* sequences were inserted in a process that ceased about 30 million years ago. Proc Natl Acad Sci USA 91:6148–6150

Berg DE, Howe MH (1989) Mobile DNA. American Society for Microbiology, Washington DC

Boguski MS (1995) The turning point in genome research. Trends Biochem Sci 20:295–296

Boussy IA, Charles L, Hamelin MH, Periquet G, Shapiro DY (1993) The occurrence of the transposable element pogo in *Drosophila melanogaster*. Genetia 88:1–10

Britten RJ, Baron WF, Stout DB, Davidson EH (1988) Sources and evolution of human *Alu* repeated sequences. Proc Natl Acad Sci USA 85:4770–4774

Calvi BR, Hong TJ, Findley SD, Gelbart WM (1991) Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. Cell 66:465–471

Cary LC, Goebel M, Corsaro BG, Wang H-G, 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

Clark JB, Maddison WP, Kidwell MG (1994) Phylogenetic analysis supports horizontal transfer of *P* transposable elements. Mol Biol Evol 11:40–50

Collins JJ, Anderson P (1994) The Tc5 family of transposable elements in *Caenorhabditis elegans*. Genetics 137:771–781

Craig NL (1995) Unity in transposition reactions. Science 270:253–254

Daboussi MJ, Langin T, Brygoo Y (1992) Fot1, a new family of fungal transposable elements. Mol Gen Genet 232:12–16

Diamond A, Devine JM, Cooper GM (1984) Nucleotide sequence of a human *Blym* transforming gene activated in a Burkitt's lymphoma. Science 225:516–519

Doak TG, Doerder FP, Jahn CL, Herrick G (1994) A proposed superfamily of transposase-related genes: new members in transposon-like elements of ciliated protozoa and a common "D35E" motif. Proc Natl Acad Sci USA 91:942–946

Earnshaw WC, Sullivan KF, Machlin PS, Cooke CA, Kaiser DA, Pollard TD, Rothfield NF, Cleveland DW (1987) Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. J Cell Biol 104:817–829

Finnegan DJ (1989) Transposable elements. Curr Opin Genet Dev 2:861–867

Green ED, Mohr RM, Idol JR, Jones M, Buckingham JM, Deaven LL, Moyzis RK, Olson MV (1991) Systematic generation of sequence-tagged sites for physical mapping of human chromosomes: application to the mapping of human chromosome 7 using yeast artificial chromosomes. Genomics 11:548–564

Grindley ND, Leschziner AE (1995) DNA transposition: from a black box to a color monitor. Cell 83:1063–1066

Iris FJM, Bougueleret L, Prieur S, Caterina D, Primas G, Perrot V, Jurka J, Rodriguez-Tome P, Claverie JM, Dausset J, Cohen D (1993) Dense *Alu* clustering and a potential new member of the *NFκB* family within a 90 kilobase HLA class III segment. Nature Genet 3:137–145

Jurka J, Walichiewicz J, Milosavljevic A (1992) Prototypic sequences for human repetitive DNA. J Mol Evol 35:286–291

Kachroo P, Leong SA, Chattoo BB (1994) Pot2, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol Gen Genet 245:339–348

Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42:59–65

Koller M, Baumer A, Strehler EE (1991) Characterization of two novel human retropseudogenes related to the calmodulin-encoding gene, *CaMII*. Gene 97:245–251

Li W, Shaw JE (1993) A variant Tc4-transposable element in the nematode *C. elegans* could encode a novel protein. Nucleic Acids Res 21:59–67

Lutfalla G, McInnis MG, Antonarakis SE, Uze G (1995) Structure of the human CRFB4 gene: comparison with its IFNAR neighbor. J Mol Evol 41:338–344

Maddison WP, Maddison DR (1993) MacClade, Analysis of Phylogeny and Character Evolution, version 3. Sinauer Associates, Switzerland, Mass.

Martin RD (1993) Primate origins: plugging the gaps. Nature 363:223–234

Minetoki T, Gomi K, Kitamoto K, Kumagai C, Tamura G (1995) Nucleotide sequence and expression of alpha-glucosidase-encoding gene (*agdA*) from *Aspergillus oryzae*. Biosci Biotechnol Biochem 59:1516–1521

Morgan GT (1995) Identification in the human genome of mobile elements spread by DNA-mediated transposition. J Mol Biol 254:1–5

Oosumi T, Belknap WR, Garlick B (1995) *Mariner* transposons in humans. Nature 378:672

Panaccione DG, Pitkin JW, Walton JD, Annis SL (1996) Transposon-like sequences at the *TOX2* locus of the plant-pathogenic fungus *Cochliobolus carbonum*. Gene, in press

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a *mariner* transposon-like element. Nature Genet 12:288–297

Robertson HM (1995) The *Tc1-mariner* superfamily of transposons in animals. J Insect Physiol 41:99–105

Robertson HM, Zumpano KL, Lohe AR, Hartl DL (1996) Reconstructing the ancient *mariners* of humans. Nature Genet 12:360–361

Ruvolo V, Hill JE, Levitt A (1992) The Tc2 transposon of *Caenorhabditis elegans* has the structure of a self-regulated element. DNA Cell Biol 11:111–122

Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol Biol Evol 11:504–512

Smit AFA, Riggs AD (1996) Tiggers and other DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93:1443–1448

Swofford DL (1993) PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.1, Smithsonian Institution, Washington DC

Toth M, Grimsby J, Buzsaki G, Donovan GP (1995) Epileptic seizures caused by inactivation of a novel gene, *jerky*, related to centromere binding protein-B in transgenic mice. Nature Genet 11:71–75

Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K (1992) The pogo transposable element family of *Drosophila melanogaster*. Mol Gen Genet 232:126–134

Yuan J, Finney M, Tsung N, Horvitz HR (1991) Tc4, a *Caenorhabditis elegans* transposable element with an unusual foldback structure. Proc Natl Acad Sci USA 88:3334–3338