Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 1443–1448, February 1996
Evolution

Tiggers and other DNA transposon fossils in the human genome

(interspersed repeats/*pogo*/*mariner*/Tc1/centromere protein CENP-B)

ARIAN F. A. SMIT* AND ARTHUR D. RIGGS

Department of Biology, Beckman Research Institute of the City of Hope, 1450 East Duarte Road, Duarte, CA 91010

Communicated by Maynard V. Olson, University of Washington, Seattle, WA, October 24, 1995

ABSTRACT We report several classes of human interspersed repeats that resemble fossils of DNA transposons, elements that move by excision and reintegration in the genome, whereas previously characterized mammalian repeats all appear to have accumulated by retrotransposition, which involves an RNA intermediate. The human genome contains at least 14 families and >100,000 degenerate copies of short (180–1200 bp) elements that have 14- to 25-bp terminal inverted repeats and are flanked by either 8 bp or TA target site duplications. We describe two ancient 2.5-kb elements with coding capacity, *Tigger1* and *-2*, that closely resemble *pogo*, a DNA transposon in *Drosophila*, and probably were responsible for the distribution of some of the short elements. The deduced *pogo* and *Tigger* proteins are related to products of five DNA transposons found in fungi and nematodes, and more distantly, to the Tc1 and *mariner* transposases. They also are very similar to the major mammalian centromere protein CENP-B, suggesting that this may have a transposase origin. We further identified relatively low-copy-number *mariner* elements in both human and sheep DNA. These belong to two subfamilies previously identified in insect genomes, suggesting lateral transfer between diverse species.

A large fraction of the human genome is composed of interspersed repetitive sequences that by and large represent inactivated copies (fossils) of transposable elements. Our haploid genome contains (*i*) more than a million short interspersed repetitive DNA elements (SINEs), 100- to 300-bp elements that originated from structural RNA pseudogenes (1, 2), (*ii*) several hundred thousand long interspersed DNA elements (LINEs), elements up to 7 kb long without long terminal repeats (LTRs) (3, 4), (*iii*) more than 100,000 MaLRs, 2- to 3-kb elements with LTRs (5, 6), and (*iv*) thousands of endogenous retroviral sequences (7). The latter two usually are found as solitary LTRs, probably through internal recombination. Only retroviruses and LINEs have coding capacity for a reverse transcriptase, but all elements are thought to have spread by retroposition, a process that involves reverse transcription of an intermediate RNA product.

No mammalian interspersed repeats have yet been described that resemble fossils of DNA or class II transposons, which move by excision and reintegration into the genome, without an RNA intermediate. DNA transposons are characterized by terminal inverted repeats (TIRs) of a length (10–500 bp) not found in known retroposons. Autonomous elements code for a transposase that binds specifically to the TIRs and catalyzes the cutting and pasting of the element (8, 9). Integration results in a short constant-length duplication of the target site, visible as direct repeats flanking the element. DNA transposons have been classified based on similarity in target site duplication, TIRs, and transposase sequence. In eukaryotes, the best studied groups are the Tc1/*mariner* and *Ac*/*hobo* elements, which duplicate 2 bp (TA) and 8 bp upon integration, respectively (9, 10). The Tc1/*mariner* transposases are related to transposases of prokaryotic elements and together they form the IS*630*-Tc1 family (11, 12).

DNA transposition, by itself not replicative, can result in duplication of the element if it moves from a replicated to a still nonreplicated part of the genome (13) or if the gap resulting from the excision is repaired using as a template the sister chromatid or homologous chromosome that still contains the element (14). Indeed, elements with TIRs account for significant fractions of the genomes of, for example, *Xenopus laevis* (15, 16) and *Zea mays* (17, 18). Because transcription and translation are uncoupled, in eukaryotes class II transposition necessarily results from transactivation. Thus, mobility may require little more than conservation of the TIRs and nonautonomous elements are as likely to be transposed as autonomous elements. Nonautonomous elements may be mutated (often internally deleted) coding elements or arise from unrelated sequences incidentally flanked by functional TIRs. As an example of the latter, *Ds1* in *Z. mays* is mobilized by the transposase of the *Ac* element with which it shares only the terminal 11 bp (17).

By analysis of 40 published fragments of human medium reiterated frequency repeats (MERs) (19–22), we found that, although the most abundant are part of SINEs, LINEs, or LTR elements (refs. 2, 4, and 6 and unpublished results), 13 are part of short MERs with TIRs and other characteristics of (nonautonomous) DNA transposon fossils. By looking for sources of transposase responsible for the accumulation of these MERs, we found, interspersed in human DNA, fossils of *mariners* and of two elements, named *Tiggers*, related to the *Drosophila* transposon *pogo*. *Tiggers* probably were responsible for the spread of some of the MERs since they share TIRs and target duplication sites. Similarities between the putative transposases and other shared features suggest that *pogo* and *Tiggers* belong to the Tc1/*mariner* family of DNA transposons.

METHODS

Our analysis is based on derivation of repeat consensus sequences from multiple alignments as described (6). A consensus approximates the original sequence of a transposable element, since the vast majority of its interspersed copies have no genomic function and mutations have accumulated randomly and at a neutral rate. The average divergence of the copies from this consensus roughly reflects the time elapsed since transposition. For calculation of percentage similarity or divergence each insertion or deletion is considered one mismatch.

Database searches were performed with BLAST (23) and the program IFIND in the IntelliGenetics package (24) by using

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: LTR, long terminal repeat; TIR, terminal inverted repeat; MER, medium reiterated frequency sequence; LINE and SINE, long and short interspersed repetitive DNA elements, respectively.

*To whom reprint requests should be addressed at: Department of Molecular Biotechnology, University of Washington, Box 352145, Seattle, WA 98195.

Table 1. DNA transposon-like elements in the human genome

Name	Length, bp	Target site	TIR	Divergence, %	No. in databases	No. in genome
Ac (maize)	4560	8 bp	CAGGGATGAAAA			
1723 (frog)	≈8000	8 bp	TAGGGATGTAGCGAACGT			
MER1 (a, b)	337/527	ATCTARAN	CAGgGGTCCCCAACC	7–20	45	7,000
MER30	230	NTYTANAN	CAGGggTGTCCAAtC	7–17	27	5,000
MER3	209	YTCTAGAG	CaGCGCTGTCCAATA	10–30	58	11,000
MER33	324	NTCTAGAN	CaGCGtTGTCCAATA	17–26	46	8,000
MER5 (a, b)	178/189	NTCTARAN	CAGTGGTTCTCAAA	16–35	250	50,000
MER20	218	NTYTANRN	CAGTGGTTCTCAACC	16–29	83	16,000
MER45	190	8 bp	CAGGgCCGGCTtCAT	18–27	27	5,000
Human *mariner*	1276	TA	TTAGGTTGGTGCAAAAGTAAT...(30 bp)	ND	6	1,000
Made1	80	TA	TTAGGTTGGTGCAAAAGTAAT...(37 bp)	8–21	38	8,000
Tc1 (nematode)	1611	TA	CAGTGCTGGCCAAAAAGATATCCACTTT			
pogo (fruit fly)	2121	TA	CAGT-ATAATTCGCTTAGCTGCATCGA			
Tigger1	2417	TA	CAGGCATACCTCGtttTATTGcG	13–26	69	3,000
Tigger2	2708	TA	CAGTTGACCCTTGAACAACaCGGG	13–20	20	1,000
MER28	434	TA	CAGTTGACCCTTGAACAACaCGGG	13–20	33	5,000
MER8	239	TA	CAGTTGACCCTTGAACAACACGGG	19–27	13	3,000
MER2	345	TA	CAGTCGtCCCTCgGTATCCGTGGG	14–26	53	9,000
MER44 (a–c)	333–726	TA	CAGTAGTCCCCCCTTATCCGCGG	14–24	29	4,000
MER46	234	TA	CAGGTTGAG3CCCTtATCCgAAA	18–27	30	5,000
MER6	862	TA	CAGcAgGTCCTCgaaTAACGcCGTT	17–21	8	1,000
MER7 (a, b)	335/1205	TA	CAGTCATGCGtcGCtTAACGACG	12–21	62	8,000
Total					897	150,000

The information relates to consensus sequences. Features of a few DNA transposons in other organisms are included for comparison. Size variants are indicated by lowercase type and multiple length entries. MER7b incorporates MER17 (20) and MER29 (21), and *Tigger1* includes MER37 (22, 27). The palindromic target site duplications could be distinguished from the TIRs, since many copies were found inserted in other interspersed repeats with known consensus sequences, enabling us to infer the original target site (data not shown). Many of the TIRs are imperfect; unmatched bases are in lowercase type. A gap and deletion were introduced in the TIRs of *pogo* and MER46 to expose the similarities with the other TIRs. Except for MER1, -6, -7, -30, and the *mariners*, all elements were found in some nonprimate mammalian sequence entries. This and the high divergence of the copies from their consensus sequence (divergence) suggest a mesozoic origin for most elements. The number in the databases is the number of elements found in all nonredundant human sequences in GenBank release 86 by using IFIND (24). Since this database largely consists of mRNA sequences while interspersed repeats mostly are confined to noncoding DNA, a better estimate for the total number of repeats in the genome may be derived from their presence in a subset of large (>20 kb) human genomic sequences. We found 196 DNA transposon fossils covering 43 kb of the 4 × 10^6 bp of such sequences currently in the database, which extrapolates to a total of about 150,000 elements constituting 1% of our DNA. The estimates in the last column are based on a total number of 150,000 and the relative frequency of repeat families in the total human database, adjusted as described above.

default settings. Both XNU and SEG filters were used in all BLASTP, BLASTX, and TBLASTN searches. We performed multiple protein alignments using CLUSTALW (25) with the slow/accurate settings and default parameters. Construction and use of profiles were with various Genetics Computer Group (Madison, WI) programs (version 8) (26).

The interspersed repeats discussed in this article are significantly (15 to >50%) diverged from other copies of the same element and their copy number in the genome is difficult to determine by hybridization experiments. More reliable copy numbers can be calculated by extrapolation from the number of matches in the databases. Fragments of longer interspersed elements are more likely to be incidentally present in the databases than shorter elements with the same genomic copy number (the repeat size range is 80–2700 bp), especially since the average length of human database entries (GenBank release 86, December 1994) was only 536 bp. To adjust for the repeat length, we used the following formula for our extrapolations:

$$\text{copy no.} = \frac{\text{no. in database} \times \text{genome size}}{\text{database size} + (\text{length element} - 60\text{ bp}) \times \text{no. of database entries.}}$$

The 60-bp factor reflects that the repeat needs to overlap the database entry by usually at least 30 bp (on either side) to be detected by the search program.

RESULTS AND DISCUSSION

Abundant Human Interspersed Repeats with TIRs. By construction of full-length consensus sequences incorporating 40 published MER fragments (19–22), we found that 13 of these belong to 11 MERs with TIRs typical for elements transposed by excision and reintegration (Table 1).[†] We report three additional MERs (MER44–46), which we discovered as inserts in LINE1 and MaLR elements, that also contain TIRs.

The MER1, -3, -5, -20, -30, and -33 consensus sequences have similar 14- or 15-bp TIRs, are flanked by 8-bp direct repeats in the genome, and share a palindromic preferred target site (NTCTAGAN) (Table 1). In structure, duplication size, and TIR sequence, these abundant repeats resemble fossils of nonautonomous members of the *Ac/hobo* DNA transposon group, like *Ds* (Table 1). Similar features suggested a relationship between the maize *Ac*, snapdragon Tam3, and *Drosophila hobo* elements (28), which was later confirmed (10) by homology of their products. MER45 also has a 15-bp TIR and duplicates 8 bp upon insertion, but both TIR and target sequence differ from that of the "MER1 group." MER2, -6, -7, -8, -28, -44, and -46, forming the "MER2 group," have similar 23- to 25-bp TIRs and are flanked by TA dimers (Table 1),

[†] All consensus sequences described in this manuscript have been deposited in the human repetitive sequence reference database maintained by Jerzy Jurka and A.F.A.S., which is available by autonomous FTP at ncbi.nlm.nih.gov in the repository/repbase/REF subdirectory.

features characteristic for the Tc1/*mariner* family of DNA transposons and *pogo* in *Drosophila* (29).

Several more observations suggest that these MERs have accumulated by DNA transposition rather than retroposition. (*i*) They lack clear regulatory sequences that identify short retroposed elements, like the RNA polymerase III promoter boxes in SINEs and the polyadenylylation signal in solitary LTRs. (*ii*) Like many DNA transposons, these MERs often have internal inverted repeat structures not requiring T·G base pairing. For example, MER5 is an almost perfect 178- or 189-bp palindrome, thereby resembling the *tourist* and *stowaway* elements in plants (18, 30) and the short version of the Tc4 element in *Caenorhabditis elegans* (31, 32). (*iii*) Most retroposed interspersed repeat families are readily divided into a series of gradually more degenerate subfamilies based on multiple shared diagnostic mutations (1, 4, 6). Although most of the MERs appear with many copies in the databases and show a wide range in divergence from the consensus sequence (Table 1), there is no indication for such subfamilies. Instead, the MER1, MER7, and MER44 length variants (Table 1) differ by internal deletions alone, reminiscent of heterogeneous length DNA transposons in other organisms (17, 31, 33). The retroposon subfamilies are thought to reflect their origin from one or a few evolving source genes, possibly since almost all transposed copies lack or soon lose (retro)transcriptional competence necessary for transposition (for review, see ref. 34). DNA transposition is not expected to lead to such subfamilies, since most transposed copies could remain mobile if only the TIRs are essential for transposition.

Three of the MERs show unusually high sequence similarities to (repetitive) sequences in other vertebrate genomes, possibly reflecting a relationship through horizontal transfer rather than germ-line transmission. We found that both terminal 70 bp of MER46 are 75% similar to those of the abundant 335-bp *Xenopus* interspersed repeat JH12 (35). Base pairs 85–170 of MER6 are 95% conserved in our consensus sequences for two previously unreported repeats, one in bony fish (e.g., GenBank accession no. M89643, bp 408–789) and another in cartilaginous fish (e.g., GenBank accession no. X56517, bp 2397–2480). Finally, similarity to the first 100 bp of MER30 (85%) has been reported in *X. laevis* DNA (36).

Search for a Transposase Source. The large number of apparent nonautonomous DNA transposon fossils in the human genome (some 150,000, see Table 1) implies an old source of transposases, likely in the form of autonomous elements. Considering their similarities (Table 1), *Ac/hobo*- and Tc1/*mariner*-like elements may have been responsible for the spread of the MER1 and MER2 groups, respectively. Therefore, we searched the conceptually translated DNA sequence databases with DNA transposase sequences and their conserved domains by using TBLASTN (23).

Searches with a variety of *Ac/hobo* transposases revealed only one human sequence potentially derived from a *hobo*-like transposon; translation of bp 94–318 of expressed sequence tag y172a04 (GenBank accession no. H13305) reveals similarity to a conserved C-terminal region of *Ac/hobo* transposases that is essential for transposase activity (10). The best matches were with *hobo* transposase-like proteins in *C. elegans*, CEK09A11_1 ($P_{\rm BLASTX} = 2.8 \times 10^{-5}$) and CELC10A4_7 ($P = 0.0039$) (37) and with the Hermes (MDOHETR_1, $P = 0.00057$) (38) and *hobo* transposases (DROHFL1, $P = 0.044$) (10) in insects. However, we found no other copies of this sequence in the databases, and its origin and relationship to the MER1 group are unclear.

Tc1-Like Elements in Frogs and *mariners* in Mammals. To detect potential sources for the MER2 group mobility, we performed TBLASTN searches with multiple Tc1/*mariner* family transposases. Tc1 elements are widespread in metazoans, including fish (11), but have not yet been described in tetrapods. We found no mammalian Tc1-like sequences but did encounter two elements in *Xenopus* (GenBank accession nos. X71067, bp 15346–16922, and Z34530, bp 1036–2471), with highest matches to *Caenorhabditis* Tc1 ($P_{\rm BLASTX} = 8.1 \times 10^{-15}$) and salmon Tc1 elements (39) ($P_{\rm BLASTX} = 1.8 \times 10^{-38}$), respectively (see ref. 40 for details). Considering the relatively small size of the *Xenopus* database, Tc1-like elements probably are quite abundant in the *Xenopus* genome.

TBLASTN searches with four artificial sequences containing the conserved residues of four mariner subfamilies identified in insects (41) revealed two types of *mariner*-like elements in the mammalian genome (unpublished results). One full-length (1274 bp) element, a member of the Cecropia (moth) group, is located in the human T-cell receptor β locus (GenBank accession no. L36092, bp 495294–497519). It has integrated in a LINE1 element, thereby revealing its exact termini and the TA duplication site. We found only five more fragments of the human *mariner* in GenBank release 86, indicating that it is a relatively low copy number repeat. However, this database also contained 37 copies of an 80-bp palindromic element resembling a *mariner* with all but the terminal 40 bp at each site deleted. We name these Made1, for mariner-dependent (or derived) element 1.

Another mutated but full-length *mariner*, belonging to the Mellifera (honeybee) subfamily (41), resides in the 3′ untranslated region of the sheep prion mRNA (GenBank accession no. M31313, bp 2670–3864) ($P_{\rm TBLASTN} = 1.0 \times 10^{-21}$). The presence in mammals of two subfamilies identified in insects and the fact that the *mariner* in the human T-cell receptor β locus is 74% similar to the partial (451 bp) DNA sequence of a *mariner* in a beetle genome (*Carpelimus* sp.) (GenBank accession no. U04455) strongly suggest lateral transfer of these elements between diverse species. Horizontal transfer has been invoked previously to explain the distribution of mariners in insects (41).

***pogo*-Related Elements in the Human Genome.** The MER2 group has quite different TIRs than the human *mariner* and probably did not use its transposase for mobilization. However, we derived a consensus sequence for another uncharacterized repetitive element that resembles an autonomous element with TIRs similar to those of the MER2 group (see Table 1). The 2417-bp consensus sequence contains two long open reading frames, one of which is 1335 bp and encodes a product closely related to the putative transposase of the *Drosophila pogo* element (29) ($P_{\rm TBLASTN} = 2.1 \times 10^{-40}$) (Fig. 1). We name this element, which incorporates MER37 (22, 27), *Tigger1*, as it represents a mammalian *pogo* (44). Like *Tigger1, pogo* has two long open reading frames, which, as indicated by cDNA analysis (29), are joined by splicing before translation.

By using the *Tigger1* product as a query in TBLASTN searches, we found fragments of a related less common human interspersed repeat (*Tigger2*) that could be pieced together to form a 2708-bp consensus sequence. The *Tigger1* and *Tigger2* products are 48% identical, whereas their DNA sequences, aligned as guided by their products, are only 54% similar in the coding region. Base pairs 1–59 and 2333–2708 of *Tigger2* match MER28. Thus, MER28 represents a simple *Tigger2* internal deletion product but is much more common than the full-length element (see Table 1). Some other *Tigger2* sequences share a deletion between bp 765 and 2385. This pattern is very similar to that of *pogo* in the *Drosophila* genome, which has many copies of a 190-bp internal deletion product, 10–15 copies of an approximately 1.3-kb element, and only a few full-length (2.1 kb) *pogo* sequences (29). In contrast, *Tigger1* seems primarily represented by full-length elements; only two copies of a 365-bp internal deletion product were found in the databases. The 5′ 60 bp of the otherwise dissimilar MER8 are almost identical to those of *Tigger2* and MER28, and its distribution possibly was dependent on the *Tigger2* transposase. Other members of the MER2 group may be internal deletion products of *Tigger*-like elements.

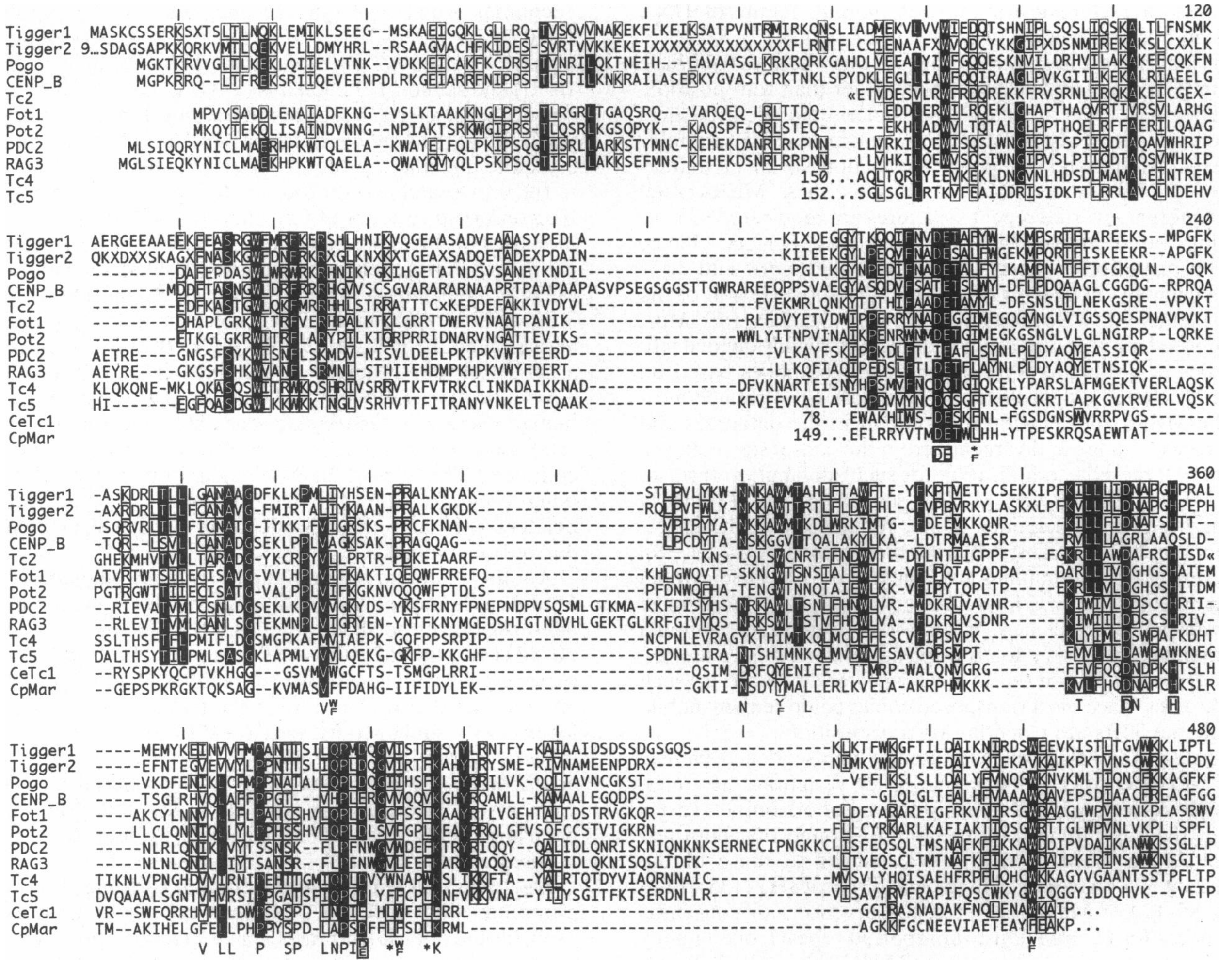

FIG. 1. Alignment of the *Tigger* transposases and related proteins, constructed with the program CLUSTALW (25). Conserved residues present in at least 7 of the 11 proteins are in white type on a black background; other conserved residues are boxed. Dashes indicate gaps introduced for the alignment. Excluded from the figure are the nonconserved C-terminal ends of these proteins and the dissimilar N-terminal 150 residues of Tc4 and Tc5. The central domains of the lacewing *mariner* (CpMar) and *C. elegans* Tc1 (CeTc1) transposases are aligned with the *pogo*-like proteins, by using CLUSTAW to align the multiple alignments of the *pogo* group and IS*630*-Tc1 group (12). Residues conserved in the Tc1-like and *mariner* transposases (12) are printed underneath the alignment (* = I/L/M/V); residues invariable in the IS*630*-Tc1 family are boxed. The CENP-B sequence is human and differs from the murine protein only outside regions that are conserved within the *pogo* family (42). The *Tigger* proteins contain ambiguous residues, indicated with an X, carried over from ambiguities in the consensus DNA sequences. The other transposase sequences are derived from insertion elements, which are not necessarily autonomous and may contain mutations in the coding region. For example, guided by similarity to the other *pogo*-like transposases, we deduced part of the Tc2 product from the DNA sequence (GenBank accession no. L00665, bp 581–700, 745–991, and 1061–1339). The translated region contains two stop codons, one of which (TGA) replaces a TGG tryptophan codon in most other proteins (position 323), suggesting that the published Tc2 sequence (43) represents a nonautonomous element. GenBank accession numbers and *P* values in *Tigger1* or *-2* TBLASTN ($P_{T1/2}$) or BLASTP ($P_{P1/2}$) searches for each protein are as follows: *pogo* (X59837; $P_{T1} = 2.1 \times 10^{-40}$), CENP_B (X55039; $P_{P2} = 1.2 \times 10^{-12}$), Tc2 (X59156; $P_{T2} = 0.3$), Fot1 (X70186; $P_{P2} = 5.9 \times 10^{-6}$), Pot2 (Z33638; $P_{P2} = 8.6 \times 10^{-5}$), PDC2 (X65608 and L19880; $P_{P1} = 9.7 \times 10^{-9}$), RAG3 (X70186; $P_{P1} = 4.4 \times 10^{-9}$), Tc4 (L00665; $P_{P1} = 0.26$), Tc5 (Z35400; $P_{P2} = 0.032$), CpMar (L06041), and CeTc1 (X01005). The highest *P* value for a nonrelated protein in any of these searches was 0.26.

Relation of *Tigger* and *pogo* Products to Other Transposases. *pogo* has been considered a DNA transposon because it has TIRs, although its product could not be related to known transposases (29). We report here that the *pogo* and *Tigger* products have similarity to the products of two apparent DNA transposons in fungal genomes, *Fot1* and *Pot2* (45, 46), and three elements in *C. elegans*, Tc2, Tc4, and Tc5 (31, 43, 47) (Fig. 1), none of which had been related to other proteins before (Figs. 1 and 2). Furthermore, many of the most conserved residues in the central domain of these *pogo*-like transposases are also conserved in Tc1/*mariner* transposases (Fig. 1). The region concerned contains the "D35E motif" (12), which was originally identified in retroviral integrases and bacterial IS transposases and is thought to form (part of) the catalytic site (48). TA target site duplication has been suggested to be a common property of the IS*630*–Tc1 transposons (12), a feature shared by most *pogo*-like transposons. Tc4 and Tc5 target a TNA site (31, 47) but encode products that lack two residues (positions 212 and 356 in Fig. 1) that are invariable in all other *pogo*-like and IS*630*–Tc1 transposases. Based on the protein similarities, the conservation of the D35E motif, the TA target site duplication, and the TIR structure (Table 1), we propose that *pogo*-like elements represent members of the IS*630*–Tc1 family of DNA transposons, closer related to the Tc1/*mariner* branch than to the prokaryotic elements.

Relation of *Tigger* and *pogo* Products to Nontransposases. Three proteins in the Fig. 1 alignment are not associated with transposons; PDC2 (49) and RAG3 (GenBank accession no. X70186) are fungal transcription factors and CENP-B is a mammalian centromere protein (50). The close similarity of the predicted *pogo* product to CENP-B has been reported (29). CENP-B specifically binds to a 17-bp sequence in α-satellite

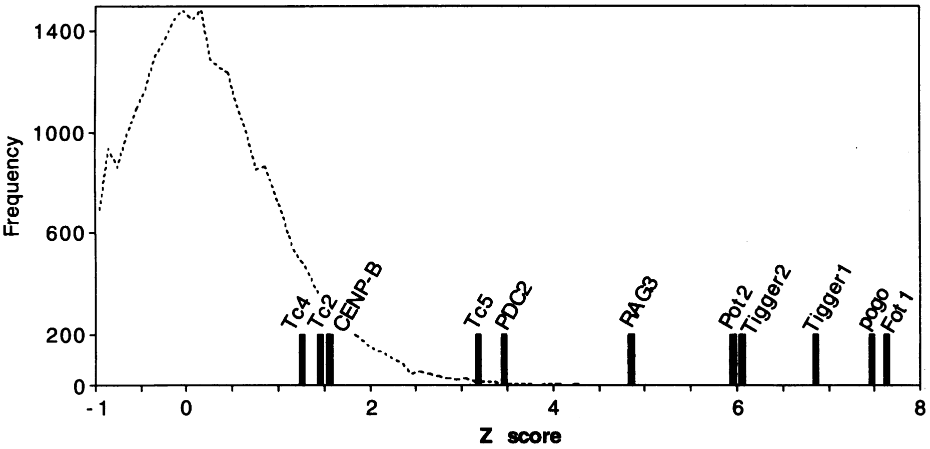

FIG. 2. Detection of *pogo* and related proteins with a profile of IS*630*-Tc1 family transposases. The profile was created from the 200-residue alignment in figure 3 of Doak *et al.* (12) expanded with impala (S75106), *C. elegans mariner* (U29380 cds3), and planarian *mariner* (X71979) transposase sequences. We searched the Swiss-Prot database augmented with the ≈180-residue central domain sequences of the *pogo*-like proteins. Scores of all database entries are indicated with a dotted curve. Excluding IS*630*-Tc1 family members, the top six scores are by the *pogo*-like proteins. Doak *et al.* (12) noticed that *pogo* scored high with their IS*630*-Tc1 D35E profile but were unable to align it with the Tc1/*mariner* transposases. Program settings for PROFILEMAKE were default and for PROFILESEARCH were gap penalty of 3, gap extension penalty of 0.2, and minimum protein length of 150 amino acids.

DNA and is thought to have a central function in the assembly of centromere structures (51).

The region of similarity of *Tigger1, Tigger2, pogo*, and CENP-B contains the DNA binding domain of CENP-B and the catalytic (D35E) domain of the transposases (Fig. 3). Given the similarity to CENP-B, *Tigger* and *pogo* transposases, like most transposases (54), probably bind DNA via their N-terminal domain. Some of the invariable D35E motif residues of *pogo*- and Tc1-related transposases are mutated in CENP-B, RAG3, and PDC2 (Fig. 1). This is not surprising, since these proteins are not thought to have transposase activity. The antiquity of the D35E motif, present in both retrotransposal integrases and DNA transposases, suggests that the transposase function is ancestral in this family of proteins and that CENP-B, despite its high conservation in mammals (42), is derived from a *pogo*-like transposase rather than vice versa. This could be an ancient example of the acquisition or exaptation of a cellular function by a transposable element.

In summary, we have provided evidence that sequences derived from DNA transposons are quite abundant in the human genome and make up at least 1% of our total DNA. The presence of the cut and paste activity of DNA transposases during mammalian evolution may have supplied the mammalian genome with a heretofore unrecognized source of evolutionary flexibility.

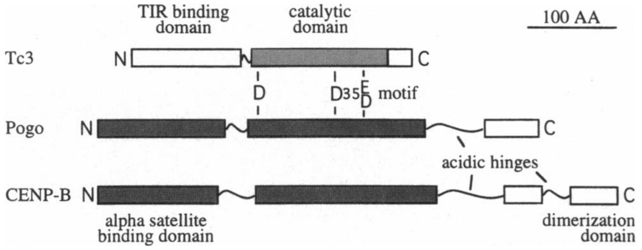

FIG. 3. Putative domain structures of Tc1 and *pogo* transposases and CENP-B. Homologous domains are shaded. The (unrelated) N-terminal domains of both CENP-B (51–53) and Tc1-like transposases (54, 55) contain specific DNA-binding activity. The central domain of IS*630*-Tc1 transposases contains the D35E motif that is essential for transpositional activity of, at least, Tc3 (9). Independent of DNA binding, CENP-B forms a homodimer through the C-terminal 60 amino acids (51) and possibly through the central domain as well (53). The C-terminal domain seems not conserved among *pogo, Tiggers*, and CENP-B, but it is consistently joined to the rest by a region rich in acidic residues.

Note Added in Proof. While this paper was in press, one of the two mammalian *mariners* described here, the *Cecropia* type element, was reported by three other groups (56–58). These authors (56, 57) also identified mammalian *mariners* belonging to a third subfamily, the horn fly group (41), further supporting our argument for their presence in the mammalian genome by horizontal transfer.

1. Deininger, P. L. & Batzer, M. A. (1993) in *Evolutionary Biology*, eds. Hecht, M. K., MacIntyre, R. J. & Clegg, M. T. (Plenum, New York), Vol. 27, pp. 157–196.
2. Smit, A. F. A. & Riggs, A. D. (1995) *Nucleic Acids Res.* **23,** 98–102.
3. Hutchison, C. A., III, Hardies, S. C., Loeb, D. D., Shehee, W. R. & Edgell, M. H. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 593–617.
4. Smit, A. F. A., Tóth, G., Riggs, A. D. & Jurka, J. (1995) *J. Mol. Biol.* **246,** 401–417.
5. Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schindler, C. W., Rush, M. G., Kadyk, L. & Leinwand, L. (1985) *Nature (London)* **316,** 359–361.
6. Smit, A. F. A. (1993) *Nucleic Acids Res.* **21,** 1863–1872.
7. Wilkinson, D. A., Mager, D. L. & Leong, J. C. (1994) in *The Retroviridae*, ed. Levy, J. A. (Plenum, New York), Vol. 3, pp. 465–535.
8. Mizuuchi, K. (1992) *Annu. Rev. Biochem.* **61,** 1011–1051.
9. Van Luenen, H. G. A. M., Colloms, S. D. & Plasterk, R. H. A. (1994) *Cell* **79,** 293–301.
10. Calvi, B. R., Hong, T. J., Findley, S. D. & Gelbart, W. M. (1991) *Cell* **66,** 465–471.
11. Henikoff, S. (1992) *New Biol.* **4,** 382–388.
12. Doak, T. G., Doerder, F. P., Jahn, C. L. & Herrick, G. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 942–946.
13. Chen, J., Greenblatt, I. M. & Dellaporta, S. L. (1992) *Genetics* **130,** 665–676.
14. Engels, W. R., Johnson Schlitz, D. M., Eggleston, W. B. & Sved, J. (1990) *Cell* **62,** 515–525.
15. Carroll, D., Knutzon, D. S. & Garrett, J. E. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 567–574.
16. Ünsal, K. & Morgan, G. T. (1995) *J. Mol. Biol.* **248,** 812–823.
17. Fedoroff, N. V. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington, DC), pp. 375–411.
18. Bureau, T. E. & Wessler, S. R. (1992) *Plant Cell* **4,** 1283–1294.
19. Jurka, J. (1990) *Nucleic Acids Res.* **18,** 137–141.
20. Kaplan, D. J., Jurka, J., Solus, J. F. & Duncan, C. H. (1991) *Nucleic Acids Res.* **19,** 4731–4738.

21. Jurka, J., Kaplan, D. J., Duncan, C. H., Walichiewicz, J., Milosavljevic, A., Murali, G. & Solus, J. F. (1993) *Nucleic Acids Res.* **21,** 1273–1279.
22. Iris, F. J., Bougueleret, L., Prieur, S., Caterina, D., Primas, G., *et al.* (1993) *Nat. Genet.* **3,** 137–145.
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
24. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
25. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
26. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
27. Lutfalla, G., McInnis, M. G., Antonarakis, S. E. & Uze, G. (1995) *J. Mol. Evol.* **41,** 338–344.
28. Streck, R. D., MacGaffey, J. E. & Beckendorf, S. K. (1986) *EMBO J.* **5,** 3615–3623.
29. Tudor, M., Lobocka, M., Goodell, M., Pettitt, J. & O'Hare, K. (1992) *Mol. Gen. Genet.* **232,** 126–134.
30. Bureau, T. E. & Wessler, S. R. (1994) *Plant Cell* **6,** 907–916.
31. Li, W. & Shaw, J. E. (1993) *Nucleic Acids Res.* **21,** 59–67.
32. Yuan, J. Y., Finney, M., Tsung, N. & Horvitz, H. R. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 3334–3338.
33. O'Hare, K. & Rubin, G. M. (1983) *Cell* **34,** 25–35.
34. Deininger, P. L., Batzer, M. A., Hutchison, C. A. & Edgell, M. H. (1992) *Trends Genet.* **8,** 307–311.
35. Deen, P. M., Terwel, D., Bussemakers, M. J., Roubos, E. W. & Martens, G. J. (1991) *Eur. J. Biochem.* **201,** 129–137.
36. Koike, T., Inohara, N., Sato, I., Tamada, T., Kagawa, Y. & Ohta, S. (1994) *Biochem. Biophys. Res. Commun.* **202,** 225–233.
37. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al.* (1994) *Nature (London)* **368,** 32–38.
38. Warren, W. D., Atkinson, P. W. & O'Brochta, D. A. (1994) *Genet. Res.* **64,** 87–97.
39. Radice, A. D., Bugaj, B., Fitch, D. H. & Emmons, S. W. (1994) *Mol. Gen. Genet.* **244,** 606–612.
40. Smit, A. F. A. (1995) Dissertation (Univ. of Southern Calif., Los Angeles).
41. Robertson, H. M. (1993) *Nature (London)* **362,** 241–245.
42. Sullivan, K. F. & Glass, C. A. (1991) *Chromosoma* **100,** 360–370.
43. Ruvolo, V., Hill, J. E. & Levitt, A. (1992) *DNA Cell Biol.* **11,** 111–122.
44. Milne, A. A. (1928) *The House at Pooh Corner* (Sutton, New York).
45. Daboussi, M. J., Langin, T. & Brygoo, Y. (1992) *Mol. Gen. Genet.* **232,** 12–16.
46. Kachroo, P., Leong, S. A. & Chattoo, B. B. (1994) *Mol. Gen. Genet.* **245,** 339–348.
47. Collins, J. J. & Anderson, P. (1994) *Genetics* **137,** 771–781.
48. Polard, P. & Chandler, M. (1995) *Mol. Microbiol.* **15,** 13–23.
49. Hohmann, S. (1993) *Mol. Gen. Genet.* **241,** 657–666.
50. Earnshaw, W. C., Sullivan, K. F., Machlin, P. S., Cooke, C. A., Kaiser, D. A., Pollard, T. D., Rothfield, N. F. & Cleveland, D. W. (1987) *J. Cell. Biol.* **104,** 817–829.
51. Kitagawa, K., Masumoto, H., Ikeda, M. & Okazaki, T. (1995) *Mol. Cell. Biol.* **15,** 1602–1612.
52. Pluta, A. F., Saitoh, N., Goldberg, I. & Earnshaw, W. C. (1992) *J. Cell Biol.* **116,** 1081–1093.
53. Sugimoto, K., Hagishita, Y. & Himeno, M. (1994) *J. Biol. Chem.* **269,** 24271–24276.
54. Colloms, S. D., Van Luenen, H. G. & Plasterk, R. H. (1994) *Nucleic Acids Res.* **22,** 5548–5554.
55. Vos, J. C. & Plasterk, R. H. A. (1994) *EMBO J.* **13,** 6125–6132.
56. Auge-Gouillou, C., Bigot, Y., Pollet, N., Hamelin, M. H., Meunier-Rotival, M. & Periquet, G. (1995) *FEBS Lett.* **368,** 541–546.
57. Oosumi, T., Belknap, W. R. & Garlick, B. (1995) *Nature (London)* **378,** 672.
58. Morgan, G. T. (1995) *J. Mol. Biol.* **254,** in press.