ELSEVIER

Gene 205 (1997) 203–217

GENE
AN INTERNATIONAL JOURNAL ON
GENES AND GENOMES

# Molecular evolution of an ancient *mariner* transposon, *Hsmar1*, in the human genome

Hugh M. Robertson *, Karen L. Zumpano

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

Accepted 22 July 1997

**Abstract**

A confident consensus sequence for *Hsmar1*, the first *mariner* transposon recognized in the human genome, was generated using three genomic and 15 cDNA sequences. It is thought to represent the ancestrally active copy that invaded an early primate genome. The consensus is 1287 base pairs (bp) long, has 30 bp perfect inverted terminal repeats (ITRs), and encodes a 343 amino acid (aa) *mariner* transposase. Each copy has diverged from the consensus largely independently of the others and mostly neutrally, and most are now defective. They differ from the consensus by an average of 7.8% in DNA sequence and 7.5 indels per kilobase, both of which values indicate that the copies were formed about 50 Myr ago. On average, only 20% of the 73 surmised CpG hypermutable sites in the consensus remain. A remarkable exception to this loss of functionality is revealed by a set of ten cDNA clones derived from a particular genomic copy that has diverged only 2.4% from the consensus, retained 54% of its hypermutable CpG pairs, and which has a full-length transposase open reading frame. The complete sequence of one of these cDNAs (NIB1543) indicates that the transposase gene of this copy may have been conserved because it is spliced to a human cellular gene encoding a SET domain protein. A specific PCR assay was used to reveal the presence of *Hsmar1* copies in all primates examined representing all major lineages, but not in close relatives of primates. PCR fragments cloned and sequenced from a representative sample of primates confirmed that *Hsmar1* copies are present in all major lineages, and also revealed another cecropia subfamily *mariner* in prosimians only, and a third highly divergent *mariner* present in the greater slow loris *Nycticebus coucang*. There are about 200 copies of *Hsmar1* in the human genome, as well as ±2400 copies of a derived 80 bp paired ITR structure and ±4600 copies of solo ITRs. Thus, this transposon had a considerable insertional mutagenic effect on past primate genomes. © 1997 Elsevier Science B.V.

*Keywords:* D,D35E motif; Gibbon; Monkey; Tarsier; Prosimian; SET domain

## 1. Introduction

The *mariner* family of Class II or DNA-mediated transposable elements (Hartl, 1989; Robertson, 1995) is extremely widespread among invertebrates. Members have been detected in the genomes of representatives of most major insect orders (Robertson, 1993; Robertson et al., 1997) and several other arthropods (Robertson and MacLeod, 1993; Jeyaprakash and Hoy, 1995), as well as nematodes, flatworms, and hydras (Sedensky et al., 1994; Garcia-Fernàndez et al., 1995; Robertson, 1997). Recently, this distribution has been extended to mammals with the discovery of two quite different *mariners* in the human genome. The first *mariner* is a member of the cecropia subfamily (see Robertson and MacLeod, 1993 for definitions of *mariner* subfamilies) and was discovered independently by several groups in two ways. Auge-Gouillou et al. (1995) sequenced PCR fragments amplified using degenerate primers designed to conserved regions of *mariner* transposase genes (Robertson, 1993), not only from the human genome but also from sheep and cow DNA. Morgan (1995) reported a preliminary consensus sequence for the encoded transposase of this element based on a preliminary consensus nucleotide sequence of genomic sequences in human accessions in the public databases, as well as Expressed Sequence Tags (ESTs) from human genome sequencing projects (primarily Hillier et al., 1996). He also described a consensus for a very short

* Corresponding author. Tel.: +1 217 3330489; Fax: +1 217 2443499; e-mail: hughrobe@uiuc.edu

Abbreviations: aa, amino acid(s); bp, base pair(s); EST, Expressed Sequence Tag; HGSC, human genome sequencing center; *Hsmar1*, *Homo sapiens mariner* 1; indels, insertions/deletions; ITR, inverted terminal repeat; STS, Sequence Tagged Site.

*PII* S0378-1119(97)00472-1

derivative of this *mariner*, consisting of paired inverted terminal repeats separated by a few additional base pairs, of which there are thousands of copies in the human genome. These inverted terminal repeat pairs were also detected by Oosumi et al. (1995) in a systematic screen of human sequence accessions for such derivatives of transposons. Smit and Riggs (1996) also describe deriving a preliminary consensus for this *mariner* from public database accessions, having similarly first detected these inverted terminal repeat pairs. We first detected a copy of this *mariner* in EST 01059 from one of the first EST projects (Adams et al., 1992). Completion of the sequence of this cDNA clone HHCPC29 revealed that it represented three-quarters of a cecropia subfamily *mariner*. Subsequent amplifications using our degenerate PCR primers designed to conserved regions of *mariner* transposase genes produced an appropriate size PCR product that had not previously been detected (Robertson, 1993). Sequencing of a cloned sample of these products confirmed the identity of this *mariner*, and it was similarly independently discovered by A.R. Lohe and D.L. Hartl (personal communication; Hartl, 1996). The second human *mariner* is a basal lineage of the irritans subfamily, and was discovered by Oosumi et al. (1995), Reiter et al. (1996), and Kiyosawa and Chance (1996). We have reported a preliminary consensus of the available genomic and EST sequences that encodes a transposase with 38% amino acid identity to that from an irritans subfamily *mariner* from the green lacewing *Chrysoperla plorabunda* (Robertson et al., 1996).

These two *mariners* are clearly ancient components of the human genome because each copy differs from their respective preliminary consensus by 8–15% DNA divergence. In addition, a particular inverted terminal repeat pair of the cecropia subfamily element in the first intron of the interphotoreceptor retinoid-binding protein (IRBP) gene has been sequenced from a variety of New World monkeys for a molecular phylogeny study (Schneider et al., 1993), and it is apparently homologous to the copy in the human genome (Morgan, 1995). Furthermore, the amplification by Auge-Gouillou et al. (1995) of comparable PCR products from sheep and cow DNA indicates that this *mariner* might have entered mammalian genomes either a very long time ago or on multiple occasions. This antiquity makes generating a confident consensus sequence for these *mariners* difficult, but also provides an opportunity to investigate the evolutionary history of these elements in primate genomes. As the first representatives of Class II DNA-mediated transposons recognized in mammalian genomes they provide a contrast with the intensively studied Class I RNA-mediated transposons such as the *Alu* and *LINE-1* elements (see Smit, 1996). In an accompanying paper we analyze the molecular evolution of the second human *mariner*, *Hsmar2*, (Robertson and Martos, 1997), and two similarly ancient members of the *pogo* family of transposons called *Tigger 1* and *2* have also recently been discovered in human sequence accessions (Smit and Riggs, 1996; Robertson, 1996). Here we describe the construction of a confident consensus sequence for the cecropia subfamily *mariner*, which we propose should be named *Hsmar1* (*Homo sapiens mariner 1*), because it was the first discovered (Auge-Gouillou et al., 1995; Morgan, 1995; Hartl, 1996; Robertson et al., 1996) (see Robertson and Asplund, 1996 for proposed naming convention for *mariners*). This consensus is presumed to represent the original active copy of this element that we estimate invaded early primate genomes approximately 50 Myr ago. We trace the subsequent evolution of this *mariner* through the primates, and describe two additional *mariners* discovered in other primates. A particular copy of *Hsmar1* in the human genome is remarkably conserved, perhaps because it has become part of a functional human gene.

## 2. Materials and methods

Fifteen cDNA clones identified from the dbEST database with greater than 80% DNA sequence identity to a preliminary consensus of *Hsmar1* were obtained. Thirteen from the WashU/Merck EST project (Hillier et al., 1996) were kindly supplied by the IMAGE consortium. The HHCPC29 cDNA clone (Adams et al., 1992) was obtained from the American Type Culture Collection and the NIB1543 cDNA clone (Berry et al., 1995) was kindly supplied by T.J. Stevens and J.M. Sikela. These cDNA clones were completely sequenced using custom primers, and have been deposited in GenBank (accession numbers U35429 and U80763–U80776). They were aligned by eye with three genomic sequences in GenBank, a full-length 1275 bp copy in the 685 kb sequence of the $\beta$ T cell receptor locus (Rowen et al., 1996; base pairs 42971–45196 of GenBank accession U66061), a truncated 970 bp copy in the 200 kb intron 1b of the *ABL* oncogene (Chissoe et al., 1995; complement of base pairs 26904–27873 of GenBank accession U07562), and a 435 bp STS sequence from Goold et al. (1993) which contains 208 bp of the 5′ end of a *Hsmar1* copy (STS-4-161; base pairs 227–435 of dbSTS accession M95263).

A consensus sequence was constructed by majority rule using the Match First capability of MacClade 3.0 (Maddison and Maddison, 1993). The frequencies of synonymous (Ks) and non-synonymous (Ka) changes from this consensus were calculated for each copy after removing all indels between each sequence and the consensus, while maintaining an aligned appropriate reading frame by removing entire codons where necessary. These rates were calculated with the Macintosh program KsKaCalc (H. Akashi, personal communication) which computes them following Nei and Gojobori (1986).

Genomic DNA samples from primates representing all the major lineages, as well as some close relatives of primates, were kindly supplied by three colleagues or were obtained locally (Table 1). A specific PCR assay was employed to determine whether they carried cecropia subfamily *mariners* similar to *Hsmar1* and closely related *mariners* in a beetle and a hydra (see Results) using primers designed to stretches of amino acids that are nearly unique to these three *mariners* in our data set of cecropia subfamily *mariner* PCR fragments (Robertson and MacLeod, 1993; Robertson et al., 1997). The primers are partially degenerate to accommodate only those codons present in these three *mariners*. They are MAR-137F (5′-CGTCRTTTTGARGTRTCNTC) designed to the amino acids RRFEVSS and MAR-202R (5′-GTYTCRCCMGGATTCAGRAA) designed to the amino acids FLNPGET (see Fig. 2).

To amplify diverse *mariner* fragments for cloning, the MAR-124F and MAR-276R degenerate PCR primers of Robertson (1993) (see Fig. 2 for locations in *Hsmar1*) were used to amplify the central half of the transposase gene of *mariners* from genomic DNA of the primates indicated with an asterisk in Table 1. The PCR products were cloned into plasmids and a sample sequenced as in Robertson (1993) (GenBank accession numbers U94620–U94695). The *Hsmar1* DNA sequences were aligned by eye with each other, the PCR fragments of Auge-Gouillou et al. (1995), and the equivalent regions of the cDNA clones and genomic copies, using the consensus sequence as a guide. Several additional PCR fragments from various prosimians did not closely resemble the *Hsmar1* consensus, but instead grouped phylogenetically with another lineage of the cecropia subfamily, so these were aligned and analyzed separately (GenBank accessions U91394–U91409).

Phylogenetic analysis was performed using maximum parsimony as implemented by PAUP v3.1.1 for the Mac (Swofford, 1993) using the Heuristic algorithm with random addition of sequences, tree-bisection-and-reconnection branch swapping, and 100 iterations. For the *Hsmar1* PCR fragment dataset of 66 sequences, 100 bootstrap replications were performed, with 1000 replications for the prosimian-only dataset of 16 fragments.

## 3. Results

### 3.1. A confident consensus sequence for Hsmar1 from cDNA clones and genomic copies

Sequencing of the 15 distinct cDNA clones available revealed that they derive from mRNAs that were illegitimate transcripts beginning within the *mariner* copy at sites other than the likely promoter region of this

Table 1
Mammal species examined for mariner transposable elements using PCR assays

| Species | Common name | Group | Source[a] |
|---|---|---|---|
| *Homo sapiens** | Human | Great Ape | |
| *Pan troglodytes* | Chimpanzee | Great Ape | NM |
| *Gorilla gorilla* | Gorilla | Great Ape | NM |
| *Pongo pygmaeus* | Orangutan | Great Ape | NM |
| *Hylobates lar** | White-handed Gibbon | Lesser Ape | LC |
| *Cercopithecus aethiops** | African Green Monkey | Old World Monkey | NM |
| *Macaca fascicularis** | Crab-eating Macaque | Old World Monkey | NM |
| *Macaca arctoides* | Stump-tailed Macaque | Old World Monkey | NM |
| *Macaca mulatta* | Rhesus Monkey | Old World Monkey | NM |
| *Ateles geoffroyi** | Spider Monkey | New World Monkey | LC |
| *Callicebus moloch** | Dusky Titi | New World Monkey | LC |
| *Pithecia irrorata* | Bald-faced Saki | New World Monkey | MG |
| *Saguinus bicolor* | Pied Tamarin | New World Monkey | MG |
| *Tarsius syrichta** | Philippine Tarsier | Tarsier | MG |
| *Lemur macaco** | Black Lemur | Prosimian | LC |
| *Galago senegalensis** | Senegal Bushbaby | Prosimian | NM |
| *Propithecus virreauxi** | Virreaux's Sifaka | Prosimian | MG |
| *Nycticebus coucang** | Greater Slow Loris | Prosimian | MG |
| *Cynocephalus variegatus* | Cobego (Flying Lemur) | Dermoptera | MG |
| *Tupaia glis* | Common Tree Shrew | Scandentia | MG |
| *Rattus rattus* | Rat | Rodentia | |
| *Mus domesticus* | Mouse | Rodentia | |
| *Oryctolagus cuniculus* | European Rabbit | Lagomorpha | |

PCR products were cloned and sequenced from those indicated with an asterisk.
[a]DNA samples kindly provided by the following: MG, Morris Goodman (Wayne State University); LC, Leona Chemnick (Zoological Society of San Diego); NM, Nobuyo Maeda (University of North Carolina).

*mariner*, or were readthrough transcripts that originated in flanking DNA outside of the *mariner* copy. As expected if these are illegitimate and readthrough transcripts, they were approximately evenly split in orientation with respect to the *mariner* copy (nine forward and six reverse). Fig. 1 is a schematic representation of these cDNAs, as well as the three genomic copies from human accessions identified in GenBank.

None of the nine forward orientation transcripts appear to have originated from the *mariner* promoter, but rather originate elsewhere within the *mariner* sequences or in the 5′ flanking DNA. Four of these nine cDNA clones (42944, NIB1543, 122887, and 249649) are similar in that transcription began in 5′ flanking DNA and the *mariner* sequences begin at the same base pair of the consensus sequence, position 176, three bases before the ATG start codon. It seems rather unlikely that all four of the genomic copies from which these cDNAs arose would have had their 5′ ends deleted at exactly the same position, instead we propose that this region of the *mariner* sequence is a cryptic 3′ acceptor splice site, which has been used in conjunction with a 5′ donor splice site in the flanking DNA. The sequence of this putative cryptic 3′ splice site in the consensus sequence is TTTATTTTAG/AC, with the proposed splice site marked by the backslash, and resembles the human consensus of (TC)nNCTAG/G (Mount, 1982). Clone 251800 is possibly also spliced using a slightly downstream 3′ acceptor site. It is possible that the large internal deletion in clone 127694 is also a spliced cryptic intron. Characterization of the genomic copies from which these transcripts originate will allow determination of their true 5′ ends. Of the nine forward transcripts, three appear to have been polyadenylated in response to the *mariner* polyadenylation AATAAA signal that overlaps the stop codon in the consensus (clones NIB1543, 251800, and 244965). Another three terminate within the *mariner* sequence, apparently in response to novel AATAAA sequences within these highly mutated copies (clones 122887, 249649, and 140214), while three more end in 3′ flanking DNA sequences (clones 42944, 127694, and 25267), in at least two cases perhaps because the *mariner* polyadenylation signal is mutated. The six reverse orientation transcripts originate once again at a variety of points in the 3′ flanking DNA or within the *mariner*. Clones 11681, 53125, and 278137

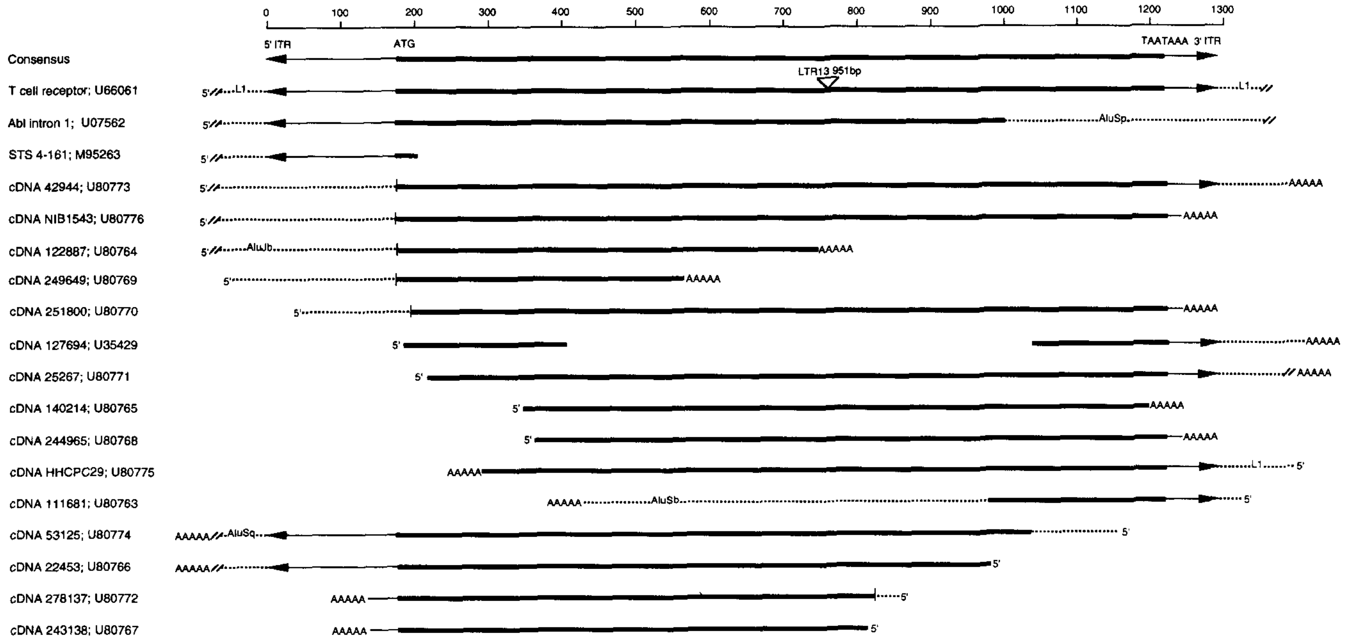

Fig. 1. Schematic representation of the genomic DNA segments and cDNA clones used to construct the *Hsmar1* consensus. Arrows indicate the inverted terminal repeats; thin lines the 5′ and 3′ non-coding regions; thick lines the transposase coding region; dotted lines the flanking DNA; and 5′ ends and poly-A tails are indicated. Positions of the 951 bp insertion in the T cell receptor copy and the 635 bp deletion in the 127694 clone are shown (the many small insertions and deletions are not shown). Flanking DNA is identified where possible. GenBank accession numbers are given after the clone or sequence names. Six of the cDNA clones represent several additional clones in dbEST that appear to derive from the same genomic copies (most ESTs are from Hillier et al. (1996); GenBank accession numbers are in parentheses). cDNA 22453 represents 33319 (R19145 and R44810), 50847 (H17928 and H17929), 35579 (R45968), and HSC3AG011 (F12491 and F10112); cDNA 53125 represents 38405 (R35606); cDNA 251800 represents 261141 (H98194 and H98195) and 250712 (H98547 and H98570); cDNA 25267 represents 134768 (32196), 137168 (R36390), 266261 (N35570), 376844 (AA046989), 548105 (AA081958 and AA082640), EST84985 (T35429), HSC27C011 (Z44695), and HUMHBC4527 (D82296); cDNA 140214 represents 293693 (N69677 and N94285); and cDNA NIB1543 represents c-3ma06 (F13556), 23658 (R38169 and T77093), 66857 (T64992), 278419 (N62795), 298675 (N74675 and W05064), 357123 (W93433 and W93323), 359595 (AA010900), 504614 (AA152184 and AA150133), and IB3615 (T16164).

may again be spliced to flanking DNA. These transcripts were again polyadenylated at a variety of positions within and beyond the *mariner* sequences.

Together with the three available genomic sequences, these 15 cDNA sequences provide an overlapping set of sequences for the entire transposon between 5 and 15 deep. At most positions a single base is the clear majority, which is enough sequence information to generate the confident majority rule consensus sequence shown in Fig. 2 (GenBank accession number U52077), with two exceptions. First, positions 557–564 are shown in the consensus as CAAAAAAA, but this exact sequence is present in only one of the 14 sequences covering this region (clone NIB1543). In the other clones it varies from three adenines to a string of 28 thymines and adenines. Five clones have a string of thymines in this region, and only 3 of the 14 sequences have the cytosine at position 557. Nevertheless, this sequence is chosen for the consensus sequence that might represent an ancestrally active *mariner* on several grounds. First, the length of this sequence is the marginal majority rule and maintains an open reading frame, consistent with an ancestrally active sequence encoding a functional transposase. Second, this sequence is represented in the PCR fragments described below, and CAAAAAAA is the marginal majority rule of the 53 unique clones available, including those of Auge-Gouillou et al. (1995). Third, this sequence is present in the closely related *mariners* from a staphylinid beetle *Carpelimus* sp. (Robertson and MacLeod, 1993) and *Hydra littoralis* (Robertson, 1997), sequences confirmed by cloning and sequencing of genomic copies from these invertebrates (K.K.O. Walden, D.J. Lampe and H.M. Robertson, unpublished results). Such high mutability of polyadenine stretches is usually ascribed to slippage of DNA polymerase (e.g., Schlötterer and Tautz, 1992).

The second exception to the clear majority rule consensus was that at 73 places in the sequence the dinucleotide pairs CG, CA, and TG were approximately equally represented, with a few additional mutations as well. If a strict majority rule consensus was employed these dinucleotide pairs would be any of these three at different positions, and sometimes ties for the transition pairs C/T and G/A were obtained at particular positions. Thirty-three of these dinucleotide pairs are in the PCR fragments and are again extremely variable in that data set. There are two reasons for believing that in the ancestrally active sequence these dinucleotide pairs were all CG, and they are shown as such in lowercase letters in the consensus sequence in Fig. 2. First, this phenomenon has been previously observed for the similarly ancient and highly mutated copies of *Alus* and other

```
TTAGGTTGGTGCAAAAGTAATTGcgGTTTTTGCATTGTTGGAATTTGCcgTTTGATATTGGAATACATTCTTAAATAAATGTGGTTATGTTATACATCAT  100
TTTAATGcgCATTTCTcgCTTTAcgTTTTTTTGCTAATGACTTATTACTTGCTGTTTATTTTATGTTTATTTTAGACTATGGAAATGATGTTAGACAAAA  200
                                                                               M  E  M  M  L  D  K     7
AGCAAATTcgAGcgATTTTCTTATTcgAGTTCAAAATGGGTcgTAAAGCAGcgGAGACAACTcgCAACATCAACAAcgCATTTGGCCCAGGAACTGCTAA  300
K  Q  I  R  A  I  F  L  F  E  F  K  M  G  R  K  A  A  E  T  T  R  N  I  N  N  A  F  G  P  G  T  A  N   41
cgAAcgTACAGTGCAGTGGTGGTTCAAGAAGTTTTGCAAAGGAGAcgAGAGCCTTGAAGATGAGGAGcgTAGTGGCcgGCCATcgGAAGTTGACAAcgAC  400
 E  R  T  V  Q  W  W  F  K  K  F  C  K  G  D  E  S  L  E  D  E  E  R  S  G  R  P  S  E  V  D  N  D     74
CAATTGAGAGCAATCATcgAAGCTGATCCTCTTACAACTAcgcgAGAAGTTGCcgAAGAACTCAAcgTcgACCATTCTAcgGTcgTTcgGCATTTGAAGC  500
 Q  L  R  A  I  I  E  A  D  P  L  T  T  T  R  E  V  A  E  E  L  N  V  D  H  S  T  V  V  R  H  L  K    107
AAATTGGAAAGGTGAAAAAGCTcgATAAGTGGGTGCCTCATGAGCTGAGcgAAAATCAAAAAAATcgTcgTTTTGAAGTGTcgTCTTCTCTTATTCTAcg  600
Q  I  G  K  V  K  K  L  D  K  W  V  P  H  E  L  S  E  N  Q  K  N  R  R  F  E  V  S  S  S  L  I  L  R  141
CAACAACAAcgAACCATTTCTcgATcgGATTGTGAcgTGcgAcgAAAAGTGGATTTTATAcgACAACcgGcgAcgACCAGCTCAGTGGTTGGACcgAGAA  700
 N  N  N  E  P  F  L  D  R  I  V  T  C  D  E  K  W  I  L  Y  D  N  R  R  R  P  A  Q  W  L  D  R  E    174
GAAGCTCCAAAGCACTTCCCAAAGCCAAACTTGCACCAAAAAAAGGTCATGGTCACTGTTTGGTGGTCTGCTGCcgGTCTGATCCACTACAGCTTTCTGA  800
 E  A  P  K  H  F  P  K  P  N  L  H  Q  K  K  V  M  V  T  V  W  W  S  A  A  G  L  I  H  Y  S  F  L    207
ATCCcgGcgAAACCATTACATCTGAGAAGTATGCTCAGCAAATcgATGAGATGCACcgAAAACTGCAAcgCCTGCAGCcgGCATTGGTCAACAGAAAGGG  900
N  P  G  E  T  I  T  S  E  K  Y  A  Q  Q  I  D  E  M  H  R  K  L  Q  R  L  Q  P  A  L  V  N  R  K  G  241
CCCAATTCTTCTCCAcgACAAcgCCcgACcgCAcgTcgCACAACCAAcgCTTCAAAAGTTGAAcgAATTGGGCTAcgAAGTTTTGCCTCATCcgCCATAT 1000
 P  I  L  L  H  D  N  A  R  P  H  V  A  Q  P  T  L  Q  K  L  N  E  L  G  Y  E  V  L  P  H  P  P  Y    274
TCACCTGACCTCTcgCCAACcgACTACCACTTCTTCAAGCATCTcgACAACTTTTTGCAGGGAAAAcgCTTCCACAACCAGCAGGATGCAGAAAATGCTT 1100
 S  P  D  L  S  P  T  D  Y  H  F  F  K  H  L  D  N  F  L  Q  G  K  R  F  H  N  Q  Q  D  A  E  N  A    307
TCCAAGAGTTcgTcgAATCCcgAAGCAcgGATTTTTATGCTACAGGAATAAACAAACTTATTTCTcgTTGGCAAAAATGTGTTGATTGTAATGGTTCCTA 1200
F  Q  E  F  V  E  S  R  S  T  D  F  Y  A  T  G  I  N  K  L  I  S  R  W  Q  K  C  V  D  C  N  G  S  Y  341
TTTTGATTAATAAAGATGTGTTTGAGCCTAGTTATAATGATTTAAAATTCAcgGTCCAAAACcgCAATTACTTTTGCACCAACCTAA              1287
 F  D  *                                                                                               343
```

Fig. 2. The consensus DNA sequence of *Hsmar1* and conceptual translation of the encoded transposase. The 73 hypermutable CG dinucleotides are shown in lower case. Various features are highlighted by underlining: the 30 bp inverted terminal repeats, the polyadenylation site that overlaps the stop codon, and the amino acids to which the partially degenerate specific PCR primers MAR-137F and MAR-202R were designed. The aspartic acid (D) residues of the D,D34D catalytic domain and the amino acids to which the fully degenerate general PCR primers MAR-124F and MAR-276R were designed are indicated by double underlining. Nucleotide and amino acid numbers are shown on the right side. These sequences are available as GenBank accession number U52077.

transposable elements in the human genome (e.g., Britten et al., 1988). This hypermutability is ascribed to the methylation of the C of each CpG dinucleotide after insertion of the individual transposon copy, and such methylated cytosines are subject to deamination to thymine. Such conversion on one strand leads to CA, on the other to TG, and subsequent mutations lead to infrequent other combinations. In support of this interpretation, there are no CpG dinucleotide pairs in the consensus that have not suffered this hypermutability, while there are 91 CA and 85 TG dinucleotides that are not hypermutable. Second, at 52 positions these CG dinucleotides are present in either or both of the closely related *Carpelimus* sp. and *Hydra littoralis mariner* genomic sequences (K.K.O. Walden, D.J. Lampe and H.M. Robertson, unpublished results), indicating that they were probably shared with a common ancestor of these three *mariners*, and have not been subject to such hypermutability in the hydra and beetle genomes.

The consensus sequence of *Hsmar1* is 1287 bp long with 30 bp perfect inverted terminal repeats and encodes a 343 amino acid *mariner* transposase (Fig. 2; GenBank accession number U52077). The transposase has the major features of a *mariner* transposase (see Robertson, 1995), in particular the D,D34D putative catalytic domain (Doak et al., 1994). This consensus sequence for *Hsmar1* is a considerable improvement over the preliminary consensus sequences that could be built from comparison of the genomic and EST sequences alone. For example, Morgan (1995) provided a conceptual translation for such a preliminary consensus, which he called HuMLE, which is missing two amino acids from the N-terminus, one of the three serines at positions 135–137, and differs at another 20 amino acid positions. A similar preliminary consensus sequence for the *Hsmar2* element (Robertson et al., 1996) is similarly improved by completion of multiple genomic and cDNA sequences (Robertson and Martos, 1997).

### 3.2. *Molecular evolution of* Hsmar1 *copies*

The molecular evolution of these *mariner* copies can be examined by comparison with this confident consensus. With one exception discussed below and excluded from this analysis (cDNA clone NIB1543), they differ from the consensus in very similar ways that allow inference of the approximate timing of when this original active version existed. First, several copies appear to have suffered major insertions. The only full-length sequence available is the genomic copy in the T cell receptor locus (Rowen et al., 1996), and as described by Morgan (1995), it has a 951 bp insertion of an endogenous retroviral LTR13 (RepeatMasker; http://ftp.genome.washington.edu/cgi-bin/RepeatMasker; A.F.A. Smit and P. Green, unpublished). The only other long genomic copy available is from the 200 kb intron 1b of the *ABL* oncogene (Chissoe et al., 1995), and it is truncated at position 992, perhaps by the insertion of a Sp subfamily *Alu* (Kapitonov and Jurka, 1996) (the available intron sequence continues for another 27 kb with no indication of the remaining 3′ sequences, so it is unlikely that the rest of this copy is present somewhere nearby). Unfortunately it is not easy to determine if the *mariner* sequences in the cDNA clones are an accurate reflection of the genomic copies they represent, especially given the likelihood that some are spliced to flanking DNA. Clone 127694 has a particularly long internal deletion of 635 bp, however, it is unclear whether this might be a cryptic intron.

Second, the average number of indels per kb is 7.5 ± 3.6 (100 in a total of 13 311 bp also equals 7.5), and deletions are 2.6-times more common than insertions. Although many of these indels are single base pairs, some are longer and commonly occurred between short direct repeats of at least three bases or in strings of adenines or thymines. The vast majority of these indels are unique to particular copies (one deletion of three bases is shared by four clones, and another by two clones) indicating that these copies have been evolving independently for a long time. Saitou and Ueda (1994) have established a rate of 0.15/kb/Myr for accumulations of insertions and deletions (excluding *Alu*s and other transposons) in non-coding nucleotide sequences of primates (this rate is the lowest of those they determined and is used because it applies to their oldest comparisons involving New World monkeys). The pattern of indels in these *mariner* sequences is similar to those they examined with respect to sizes and relative frequencies of insertions and deletions, and it also seems reasonable to assume that these *mariner* copies have been evolving neutrally in a fashion similar to the pseudogenes they examined (see below). An indel rate of 0.15/kb/Myr suggests that these *mariner* copies were formed approximately 50 Myr ago.

Third, the rate of neutral evolution across primates has been estimated to be about 0.16% per Myr (e.g., Britten, 1994), measured as the percent divergence excluding CpG hypermutable sites and treating indels as single events. The average divergence of these *Hsmar1* copies from the consensus by this measure is 7.8 ± 0.9%, yielding an age estimate of 49 Myr. When CpG hypermutable sites are included this measure increases dramatically to 11.8 ± 0.9%.

Fourth, as expected from the apparent age of these inserts, the proportion of CpG dinucleotides remaining is fairly low, averaging 19.7 ± 9.3% with a range from 0 to 33% (one of the short sequences, the STS.4-161, retained none of its five CpG pairs). The same proportion of 19.7% is obtained when calculated across all 18 sequences (161 of 819 CpG pairs remaining). As noted above, most changes were to CA or TG, but sometimes additional mutations had occurred.

The mode of evolution of these copies seems clear from the above results, that is, they appear to be evolving neutrally as pseudogenes, having long since lost coding function due to various indels and/or nucleotide substitutions. This inference was tested by comparing the rates of substitution at synonymous (Ks) vs non-synonymous (Ka) positions in the transposase coding regions, comparing each copy with the consensus sequence, and removing all indels relevant to each pairing to maintain aligned reading frames (computation of all pairwise comparisons between copies is excessive and probably not more informative than comparison of each with the consensus, given the great age of these copies). If they are evolving neutrally, then the Ks/Ka ratio should be approximately one, whereas most normal coding genes evolving with some conservation of amino acid sequence have ratios much higher than one, especially when compared across such long time frames. The average Ks value for the 16 sequences was 0.21 ± 0.04, while the average Ka value was about half that at 0.11 ± 0.01. The average Ks/Ka ratio was 2.0, ranging from 1.4 to 3.3. This ratio suggests that some conservation of the coding capacity of these individual copies occurred, perhaps during the initial phase of copy number expansion, however, most of their history has been dominated by neutral evolution given their current clear pseudogene status.

### *3.3. The exceptional NIB1543 cDNA may represent a chimeric functional cellular gene*

The exceptional NIB1543 clone is from a forward transcript that appears to have been spliced across the 5′ end of this *mariner* copy. The splice is inferred to occur three bases before the putative *mariner* start codon and the transcript apparently terminated in response to the *mariner* polyadenylation signal. It therefore includes the entire *Hsmar1* coding region (Fig. 1). In every respect the *mariner* copy from which this transcript originated is significantly closer to the consensus sequence than any of the others. It has suffered just one single base deletion in the 3′ untranslated region, it differs from the consensus by just 2.4% (4.9% including CpG positions), it retains 54% of its CpG pairs, and its Ks/Ka ratio relative to the consensus is 5.8. All of these features suggest that this copy has somehow been remarkably conserved, indeed it retains the full-length open reading frame of the consensus. Its encoded transposase differs from the consensus at 19 positions (5.5%), none of which are known to be crucial to transposase function, for example, the D,D34D residues or their neighbors (Lohe et al., 1997).

A possible explanation for this conservation is provided by the 5′ sequence of the 2063 bp NIB1543 cDNA clone. This 1004 bp sequence has a possible initiation Met codon after 23 bases with a context compatible with vertebrate translation start sites (e.g., Cavener and Ray, 1991; Kozak, 1996), and the first 140 bp may be part of a CpG island with 70% GC content, 15 CpG pairs and three *Hpa*II (CCGG) sites, features characteristic of the promoter and first exons of 56% of human genes, although tissue-specific genes commonly have internal CpG islands (e.g., Larsen et al., 1992). It encodes a long open reading frame that is spliced in frame to the codon before the *mariner* ORF, suggesting that the coding region of this *mariner* copy has become part of a functional cellular gene. The 327 amino acid sequence encoded by the 5′ half of the NIB1543 transcript has 28% amino acid identity to the C-terminus of the human G9a gene product (Milner and Campbell, 1993) and 31% identity to the C-terminal region of the human gene No. KIAA0067 (Nomura et al., 1994) (Fig. 3). In addition, most of this sequence corresponds to the SET domain of amino acid similarity shared by the C-termini of the *Suppressor of variegation 3–9*, *Enhancer of zeste*, and *trithorax* proteins of *Drosophila melanogaster* (Jones and Gelbart, 1993; Tschiersch et al., 1994). The function of the SET domain has not been established, but it is implicated in the protein–protein and chromosomal interactions of these *Drosophila* proteins that influence transcription of multiple other genes. It is present in the mammalian homologs of *Enhancer of zeste* and *trithorax*, and additional proteins with SET domains include three hypothetical *C. elegans* gene products, the *Drosophila Ashl* gene, and a yeast gene (Stassen et al., 1995). Most of these proteins, including *Enhancer of zeste* and *Suppressor of variegation 3–9*, also have cysteine-rich regions immediately N-terminal of the SET domain, although these are difficult to align with the seven shared cysteines in Fig. 3. If NIB1543 is indeed a full-length cDNA, then the SET domain of this protein is in the N-terminal domain, in contrast to its generally C-terminal or occasionally internal location in the other proteins (Stassen et al., 1995). The largest set of cDNA clones, ten in total, is from this *Hsmar1* copy (see Fig. 1 legend), and at least two others, 23658 and c-3ma06 have the same 5′ end as NIB1543. Clones IB3615 and 278419 might be similarly long, while five more are shorter cDNAs originating within the *Hsmar1* sequence (clones 504614, 357123, 66857, 298675, and 359595). The NIB1543, IB3615, 23658, and c-3ma06 clones all derive from the same infant brain cDNA library (Hillier et al., 1996).

### *3.4.* Hsmar1 *and other* mariners *in primates*

The age of the copies of *Hsmar1*, and the consequent inferred age of the ancestral active copy, indicated that *Hsmar1* copies should be present in the genomes of most primates (see Martin, 1993; Kay et al., 1997), so we examined representatives of all of the major primate lineages and some related mammals (Table 1) using a

```
Shared                      Dva G En P            P    Y           Id        C Cv        C C
NIB1543    1-MAEFKEKPEAPTEQLDVACGQENLPVGAWPPGAA---PAPFQYTPDHVVGPGADIDPTQITFPGCICVKTPCLPGTCSCLRHGENYDDNS
G9a      701-GVGNRAIRTEKIICRDVARGYENVPIPCVNGVDGEPCPEDYKYISENCETSTMNIDRNITHLQHCTCVD-DCSSSNCLCGQLSIRRWYDK
KIAA0067 666-VDRKFQPYKPFYYILDITYGKEDVPLSCVNEIDTTP-PPQVAYSKERIPGKGVFINTGPEFLVGCDCKDGCRDKSKCACHQLTIQATACT

Shared         r                     fECN  C C    C NRvVQ G q   Q fkT kkGWG R L  IpkG FvCeYaG      e
NIB1543      CLRDIGSGGKYAE---------PVFECNVLCRCSDH-CRNRVVQKGLQFHFQVFKTHKKGWGLRTLEFIPKGRFVCEYAGEVLGFSEVQR
G9a          DGRLLQEFNKIEPP--------LIFECNQACSCWRN-CKNRVVQSGIKVRLQLYRTAKMGWGVRALQTIPQGTFICEYVGELISDAEADV
KIAA0067     PGGQINPNSGYQYKRLEECLPTGVYECNKRCKCDPNMCTNRLVQHGLQVRLQLFKTQNKGWGIRCLDDIAKGSFVCIYAGKILTDDFADK

Shared       r       d                          D  y GNiqRflNHsC PN1    V         p  A Fa kdI   eEL  DY
NIB1543      RIHLQTKSDSNYIIAIREHV-----YNGQVMETFVDPTYIGNIGRFLNHSCEPNLLMIPVRIDSM---VPKLALFAAKDIVPEEELSYDY
G9a          REDDSYLFDLDNKDGE---------------VYCIDARYYGNISRFINHLCDPNIIPVRVFMLHQDLRFPRIAFFSSRDIRTGEELGFDY
KIAA0067     EGLEMGDEYFANLDHIESVE-336-YDGEESCYIIDAKLEGNLGRYLNHSCSPNLFVQNVFVDTHDLRFPWVAFFASKRIRAGTELTWDY

Shared         r                     C CG   C   1
NIB1543      SGRYLNLTVSASKERLDHGKLRKPCYCGAKSCTAFLPFDSSLYCPVEKSNISCGNEKEPSMCGSAPSVFPSCKRLTLEt-328mariner
G9a          GDRFWDIKS-----------KYFTCQCGSEKCKHSAEAIALEQSRLARLDPHPELLPELGSLPPVNT-1001
KIAA0067     NYEVGSVEG-----------KELLCCCGAIECRGRLL-1291
```

Fig. 3. Alignment of the amino acid sequence encoded by the 5′ end of cDNA NIB1543 with the C-terminal sequences of two other human proteins. Amino acids shared by all three sequences are shown in upper case letters above the alignment, and those shared by NIB1543 and one of the other proteins are shown in lower case. The SET domain is indicated by underlining of these shared amino acids. The amino acid numbers of these sequences are indicated at the start and end. The single threonine encoded by the NIB1543 cDNA clone from the *Hsmar1* sequence before the Met initiation codon of the *Hsmar1* transposase is indicated in lower case. Note that the KIAA0067 sequence has a 336 amino acid insertion within the SET domain (this insert has no significant similarity to any other sequences in GenBank).

specific PCR assay based on two primers designed to the original PCR fragments from human, staphylinid beetle, and hydra DNA (see Materials and Methods). All the primate species examined tested positive, but none of the relatives did.

To examine these primate *mariner* copies further, the original fully degenerate PCR primers described in Robertson (1993) were employed to amplify the central 450 bp region of these and other *mariners* from the primates indicated with an asterisk in Table 1. This strategy was employed for several reasons. First, at the time we did not have sufficient confident sequence information to design primers to amplify a larger section of this *mariner.* Second, this region has been used successfully to analyze relationships of a large variety of *mariners*, and provides an adequate sequence sample of the entire element (Robertson, 1993, 1997; Robertson and MacLeod, 1993; Robertson et al., 1997). Third, we wanted to keep open the possibility of obtaining sequences of additional *mariners* that might be in primate genomes.

We obtained sequences of at least two different PCR fragment clones from each of the primates indicated with asterisks in Table 1. All but *Galago senegalensis* yielded at least one PCR fragment derived from a *Hsmar1* copy. In general, these diverged from the consensus in similar ways to the human cDNA clones and genomic copies. There was no evidence among these PCR fragments of copies closely related to the exceptional NIB1543 clone. These results confirm the specific PCR assay that copies of *Hsmar1* are present in the genomes of primates representing all of the major lineages. Phylogenetic analysis of these primate *Hsmar1* PCR sequences, together with the human, sheep, and cow PCR fragments of Auge-Gouillou et al. (1995), and the equivalent regions of the cDNA clones and genomic copies, was performed on the aligned DNA sequences using maximum parsimony. The tree was rooted using six sequences, three PCR fragments each from a staphylinid beetle *Carpelimus* sp. (Robertson and MacLeod, 1993) and *Hydra littoralis* (Robertson, 1997), with which the *Hsmar1* consensus shares about 85% DNA and encoded amino acid identity. An arbitrary representative of the three equally most parsimonious trees obtained is shown in Fig. 4. The copies from particular species do not cluster together, but rather are spread throughout the tree. This conclusion is supported by the bootstrap analysis, which provided strong support for sister relationships of only three pairs of sequences (Fig. 4). The *Cercopithecus aethiops* 12 and *Homo sapiens* cDNA 25267 clones and the *Hylobates lar* 3 and *Homo sapiens* 8 PCR clones clustered together (these pairs also share several unique indels). These relationships might best be explained as orthologous comparisons between the same copy, separated only by the divergence of Old World Monkeys and Lesser Apes, respectively, from humans. The confident clustering of two *Tarsius syrichta* PCR clones (numbers 4 and 9) might be explained by a subsequent duplication of a genomic region in this lineage. As expected, the branch to the exceptional cDNA NIB1543 copy is short in comparison with the long branches leading to most other copies.

In addition to PCR fragments from *Hsmar1* copies, we obtained sequences representative of two additional *mariners* in the genomes of some of the prosimians. One of these is another member of the cecropia subfamily and multiple clones representing it were obtained from

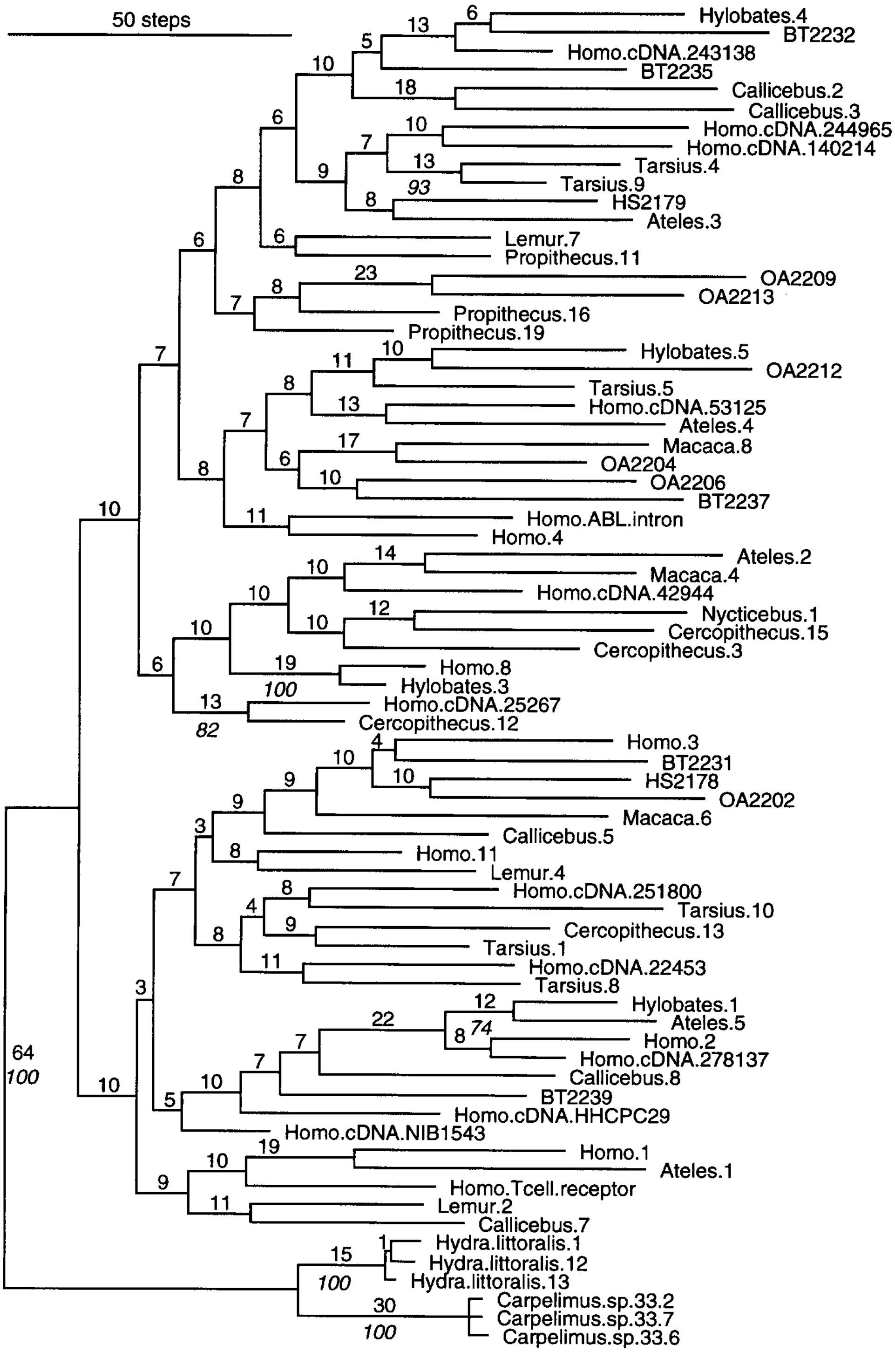

Fig. 4. Phylogenetic relationships of *Hsmar1* PCR fragments, cDNA clones, and genomic copies from humans with the PCR fragments from other primates. The staphylinid beetle *Carpelimus* sp. and *Hydra littoralis* sequences were defined as the outgroup. This is an arbitrary representative of the three equally most parsimonious trees of 3135 steps with a consistency index of 0.31. These three trees differ only in the relationships of the *Carpelimus* sp. outgroup. Branch lengths in number of nucleotide changes are shown for branches supporting nodes. Bootstrap percentages are shown below branches supporting nodes present in more than 50% of the 100 replications. Only two of the five human PCR clones from Auge-Gouillou et al. (1995) (2178 and 2179) are included, because the other three are essentially identical to our PCR fragments or cDNA clones (2173 = PCR fragment 11; 2217 = PCR fragment 4; 2177 = cDNA 140214).

each of the four prosimians examined. These copies differ from their consensus in similar ways to the *Hsmar1* copies, thus they have suffered an average of 10.3 indels per kilobase, differ from the consensus by an average of 11%, and retain an average of 16.3% of their CpG pairs (the 451 bp consensus has 18 CpG pairs). They represent a distinct lineage of the cecropia subfamily most closely related to sequences from the house ant *Tapinoma sessile* (Robertson and MacLeod, 1993; Robertson et al., 1997), with which their consensus shares 73% DNA and 77% encoded amino acid identity. These *mariners* presumably derive from an invasion of a common ancestor of the prosimian clade shortly after its separation from the lineage that led to the higher primates. Phylogenetic analysis using the house ant clones to root the tree and 1000 bootstrap replications again indicates little structure to the relationships of these prosimian *mariner* sequences (Fig. 5).

The second new primate *mariner* is represented by two clones from the prosimian *Nycticebus coucang* (slow loris) that are quite distinct from all *mariners* found to date both in encoded amino acid sequence and in length (GenBank accession numbers U94660 and U94661). They therefore apparently represent a new minor basal subfamily of *mariners* (see Robertson et al., 1997). Like the *Hsmar1* and prosimian cecropia subfamily *mariner* clones, they require several frameshifts each for translation of an alignable reading frame and therefore must derive from similarly highly defective genomic copies.

### 3.5. *Paired and solo ITRs derived from* Hsmar1 *in the human genome*

Another feature of *Hsmar1* is the presence of large numbers of paired inverted terminal repeats (ITRs) and solo ITRs in the human genome. Discovery of the former led Oosumi et al. (1995) and Smit and Riggs (1996) to these elements, and Morgan (1995) described them in detail. These authors gave the paired ITRs different names (hum7, Made1, and Mrs, respectively). The hum7 sequence reported by Oosumi et al. (1995) is an incomplete version of 28 bp. Morgan (1995) derived a more confident consensus for the paired ITRs (Mrs) of 80 bp, consisting of 37 bp ITRs separated by six unique base pairs, and the Smit and Riggs (1996) Made1 consensus is the same. Our more recent searches of human sequence accessions in the public databases with this 80 bp consensus confirm it. The first 30 bp of this ITR are identical to the *Hsmar1* ITR, the next three base pairs are shared with the 5′ end of *Hsmar1*, while the last four base pairs of the ITR and the six unique base pairs are of unclear origin (Fig. 6a). Presumably, as proposed by these authors, these paired ITRs derive from a particular original internally deleted *Hsmar1* copy (or from an inverted duplication of a solo ITR derived from *Hsmar1*) that was subsequently copied many times around the genome by *Hsmar1* transposase. This event must have happened soon after the horizontal transfer of *Hsmar1* into an ancestral primate genome, because the consensus of the paired and solo ITRs is identical to the *Hsmar1* consensus ITR for the first 30 bp. The frequency of *mariner* transposition in vitro is strongly size-dependent (Lampe et al., 1996; D.J. Lampe, T. Grant, and H.M. Robertson, unpublished data), so the small size of the paired ITRs would predispose them to enhanced rates of transposition. Most of the solo ITRs appear to have been derived from these paired ITRs, rather than directly from *Hsmar1* copies, because they have at least part of the last 3 bp and the unique 6 bp. They may have resulted

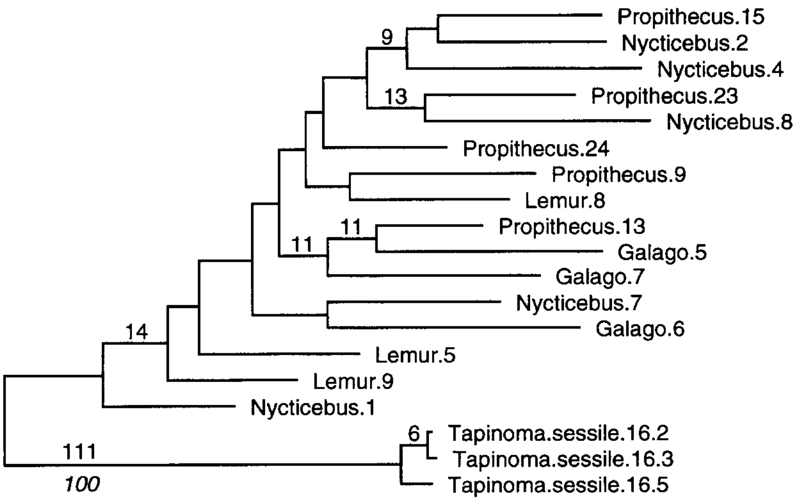

Fig. 5. Phylogenetic relationships of the prosimian cecropia subfamily *mariner* clones, rooted with *mariner* sequences from the house ant *Tapinoma sessile*. This is an arbitrary representative of the 14 equally most parsimonious trees of 896 steps with a consistency index of 0.65. Branch lengths in number of nucleotide changes are shown for those branches supporting nodes present in all 14 trees using the semi-strict consensus option of PAUP v3.1.1. Only the distinction of the prosimian and house ant sequences had bootstrap support (100%) from 1000 replications.

```
a.
Hsmar1 ITR                                  TTAGGTTGGTGCAAAAGTAATTGcgGTTTT
Hsmar1 5' end                               TTAGGTTGGTGCAAAAGTAATTGcgGTTTTTGCattgttggaa
Hsmar1 paired ITR                           TTAGGTTGGTGCAAAAGTAATTGcgGTTTTTGCCATTgaaagt
b.
Human osteopontin  AAAATAAATGGTCCTATTTATTATTAGGTTGCTGCAAAAGTAATTGTGGTTTTTATTAGTCCTAAAGTGTCAGCTTCCTTTTGCCAAA
Pig osteopontin    AAGATAAATGGTCCTATTTAT--------------------------------TATTAGTCCTAAGGAGTCAGTTCCTATTTACCAAA
c.
Human mRNA U09197  GTGCATACTTGTAATCCGAGCTATTAGGTTGGTGCAAAAGTAACTGCAGTACTTGGGAGGCTGAGGCTCGAGAATCGCTTGAATGCAG
Sx subfamily Alu   GCGCGCGCCTGTAATCCCAGC----------------------------TACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGG
```

Fig. 6. Terminal sequences of *Hsmar1* and paired and solo ITR derivatives. **(a)** Alignment of the 30 bp *Hsmar1* consensus ITR sequences with those of the *Hsmar1* 5′ end and the paired ITR consensus. The additional 4 bp of the 37 bp paired ITR are underlined and the six unique central base pairs of the 80 bp paired ITR consensus are underlined and in lowercase. A hypermutable CpG pair in the ITRs is shown in lowercase. **(b)** Sequence of the 5′ region of the human osteopontin gene and pig homolog, showing insertion of a solo ITR in the human sequence. The human sequence is the complement of positions 774–861 of GenBank accession D14813 (Hijiya et al., 1994), while the pig sequence is the complement of positions 811–866 of GenBank accession M84121 (Zhang et al., 1992). The additional TA in the human sequence is presumably the flanking TA duplication typically induced upon insertion of a *mariner* transposon. **(c)** Insertion site of a solo ITR in a Sx subfamily *Alu* element at the 5′ end of a 744 bp mRNA (GenBank accession U09197; Burn et al., 1994). The Sx subfamily *Alu* consensus is positions 150–209.

individually from deletion of one of the 37 bp ITRs with variable precision by host factors or from incomplete double-strand gap repair of excision sites of the paired ITRs (see Gloor et al., 1991).

The copy number of the paired ITRs in the human genome was estimated by Morgan (1995) to be 1000–5000 and by Smit and Riggs (1996) to be 8000 (their estimates of *Hsmar1* copy number were 100–300 and 1000, respectively). With the generation of large amounts of human DNA sequence in a relatively random fashion from cosmids and bacterial artificial chromosomes by several human genome sequencing centers, it is now possible to obtain more accurate estimates. Searches of the 'finished' and 'unfinished' human sequence databases of the Sanger and Washington University Human Genome Sequencing Centers using their World Wide Web BLASTN servers revealed the presence of 24 paired ITRs and 46 solo ITRs in a total of approximately 30 Mbp of sequence. This is approximately 1% of the estimated 3 billion base pairs in the human genome, so the estimated copy numbers are 2400 for the paired ITRs and 4600 for the solo ITRs. Their unfinished sequence databases also contain two apparently full-length genomic copies of *Hsmar1*, yielding a preliminary copy number estimate of 200, in agreement with Morgan (1995). The presence of so many ITRs in the human genome (approximately 10 000 total) might mean that any available transposase would be titrated out by these, rather than bind both ITRs of any particular copy to transposase it, thus significantly quenching the rate of transposition of intact elements. It will be of particular interest to determine whether the transposase domain encoded by the exceptional NIB1543 copy binds these ITRs.

As noted by Morgan (1995), the presence or absence of particular paired ITRs in various genomic locations supports the apparent age of *Hsmar1*. He described two instances where copies are present in human gene sequences (the prolactin and interphotoreceptor retinoid-binding protein (IRBP) genes) but absent from the bovine homologs. In addition, he noted that a paired ITR is present in identical positions in the first intron of the IRBP gene in both humans and all New World Monkeys examined by Schneider et al. (1993). This evidence suggests that paired ITRs, and therefore perhaps *Hsmar1*, were not present in the common ancestor of artiodactyls and primates, but were in the common ancestor of humans and New World Monkeys, placing their origin between 100 and 40 Myr ago. A similar situation holds for a particular solo ITR, which is approximately 2500 bp upstream of the TATA box of the human osteopontin gene, but is absent from the same region of the pig homolog (Fig. 6b). Conversely, a particular solo ITR has inserted in the middle of an *Alu* element of the Sx subfamily, indicating that movement of these ITRs and perhaps *Hsmar1* copies was ongoing about 37 Myr ago, since the Sx subfamily is thought to have expanded about this time (Britten, 1994; Kapitonov and Jurka, 1996). Study of the distribution of particular copies of *Hsmar1* and paired and solo ITRs in primates should allow determination of when they were first and last actively transposing.

## 4. Discussion

Our consensus sequence for this *Homo sapiens mariner 1 (Hsmar1)* transposon in the human genome is almost completely confident, with only a few ambiguous positions. The ambiguities involve a string of adenines, regions known to be subject to high rates of mutation (e.g., Schlötterer and Tautz, 1992), and the inferred presence of 73 CpG hypermutable pairs in the ancestral sequence of this transposon. The 15 cDNA clones containing regions of *Hsmar1* copies all appear to represent illegitimate or read-through transcripts. Some of them appear to have undergone splicing using cryptic 3′ acceptor splice sites in these *Hsmar1* copies, however, the remaining sequences may be accurate reflections of the genomic *Hsmar1* copy sequences from which they derive. Comparisons of these sequences with the consensus indicate that they are very old components of the human genome. Calculations based on their DNA divergence from the consensus and their accumulation of indels relative to the consensus suggest that these copies were formed around 50 Myr ago.

We examined representatives of all the major primate lineages using a PCR assay in an effort to pinpoint when this transposon entered primate genomes, and found sequences representative of it in most of them, including two prosimians (the most basal extant lineage of primates). The phylogenetic relationships of these sequences revealed no grouping by species of origin. There are at least two possible explanations for these observations, depending on the age of primates. If we consider primates to be around 55-Myr-old based on oldest fossil evidence (e.g., Kay et al., 1997), then given uncertainties in the aging of the human *Hsmar1* copies, it might be that a single initial invasion of a primate ancestor by a *Hsmar1* copy led to all the copies we recovered by PCR, which would explain why their relationships reveal no clustering by species. Another estimate that would support this explanation is the average rate of 0.21 for synonymous substitutions in each human copy relative to the consensus. This value exceeds the value of 0.18 for all of primate evolution inferred by Bulmer et al. (1991), again suggesting that these copies predate primate evolution.

Given the sampling bias of fossil evidence it seems likely, however, that primates are older than 55 Myr, perhaps as old as 80 Myr (Martin, 1993). Other recent reports that mammal ordinal lineages are of this or older age support this contention (Hedges et al., 1996; Archibald, 1996). In this case our aging of *Hsmar1* copies in the human genome to 50 Myr would suggest that *Hsmar1* entered existing primate lineages. In that case the lack of species clustering of sequences in Fig. 4 must be explained by an independent but reasonably closely timed entry into the two major primate lineages, Strepsirhini and Haplorhini, or by a horizontal transfer of a *Hsmar1* copy into one lineage soon after initial invasion of the other lineage. This kind of explanation might also fit the observation by Auge-Gouillou et al. (1995) of *Hsmar1* sequences in cow and sheep genomes (these sequences similarly do not cluster by species in Fig. 4). Given that the primate-artiodactyl split probably occurred at least 80 Myr and perhaps as long as 100 Myr ago (Hedges et al., 1996; Archibald, 1996), their presence in artiodactyls must also have involved contemporaneous invasions of this lineage. Consistent with this interpretation is our failure to detect *Hsmar1* copies in the genomes of likely primate ancestors, such as Scandentia, Dermoptera, and Rodentia.

Whatever the evolutionary history of this *mariner* within mammal genomes, it was apparently gained by a horizontal transfer from another animal, presumably as a single active copy that invaded a mammal genome. The closest relatives of *Hsmar1* in our database of *mariner* PCR fragments are elements from a staphylinid beetle of the genus *Carpelimus* (Robertson and MacLeod, 1993) and a hydra species *H. littoralis* (Robertson, 1997). These invertebrate *mariners* share 92% DNA identity with each other, and they share 85% DNA identity with *Hsmar1*. This level of identity of DNA sequences in animals from three different phyla at least 600 Myr old can only be explained in terms of horizontal transfers of these *mariners* (Robertson, 1997), however, the sources of these transfers cannot be resolved. First, the *Hsmar1* transfer occurred long ago, so it will be difficult to relate *Hsmar1* directly to any other *mariner*. Secondly, we have only examined about 700 animal species (Robertson and MacLeod, 1993; Robertson, 1997; Robertson et al., 1997), so closer relatives could well be found in other species. Third, the beetle and hydra *mariners* appear to be recent invaders of their genomes (in each case the three copies obtained differ from their consensus by less than 1% DNA divergence), so it is unlikely that they are directly related to *Hsmar1*, most copies of which were probably defective within a few million years of their entry into ancestral primate genomes. Fourth, a possible clue is provided by the presence of so many CpG base pairs, as many as might be expected by chance, which might indicate that the source was not another mammal or other animal with extensive CpG methylation, which would have led to loss of CpG pairs.

Similar conclusions apply to the distinct cecropia subfamily *mariner* sequences found only in prosimian genomes (Fig. 5). If anything, these are older than *Hsmar1* in this lineage, although faster rates of evolution in Strepsirrhini may explain their somewhat higher levels of divergence from their consensus sequence. Again, the closest known relative of this *mariner* is in an insect, in this case the house ant *Tapinoma sessile*, and the high frequency of CpG pairs (18 in the 451 bp consensus sequence) again indicates that this *mariner* originated from a non-mammalian host. The third primate *mariner* described here from the prosimian *Nycticebus coucang* has no known close relatives. It might represent a subfamily of *mariners* that has sequences that are not easily amplified with these primers, or that is common in animals we have not yet surveyed. Finally, no sequences representative of the second ancient *mariner* in the human genome, *Hsmar2*, were obtained from any primate in this survey, most likely because the forward primer MAR-124F was designed from the amino acid sequence WVPHEL and is, therefore, unlikely to anneal to *Hsmar2* copies, the consensus sequence of which encodes WVPKPL (Robertson et al., 1996; Robertson and Martos, 1997).

A remarkable exception to the above discussion is the *mariner* copy represented by the cDNA clone NIB1543. Although included in construction of the *Hsmar1* consensus, this copy is significantly different from all other genomic, cDNA, and PCR copies of *Hsmar1* obtained from all primates in that it is much less diverged in all respects (2.4% DNA divergence, only one single base pair deletion, 54% of CpG pairs retained, and a Ks/Ka ratio relative to the *Hsmar1* consensus of 5.8). Most significantly, this copy has a full-length transposase open reading frame. The 5′ 1004 bp of this cDNA encodes 327 amino acids, part of which is a SET domain, and is spliced in frame to the *Hsmar1* transposase open reading. Therefore, this NIB1543 cDNA appears to encode a chimeric functional cellular protein. The presence of multiple cDNAs derived from this gene in dbEST, at least two of which are apparently also full-length, suggests that it is a functional human gene. Evolution of transposons into genes with possible cellular roles has been observed before, for example, a P element derivative appears to have formed a functional gene in *Drosophila guanche* and *D. subobscura* (Miller et al., 1995). Alternative explanations for the conservation of this *Hsmar1* copy include that it represents an extreme of the distribution of *Hsmar1* sequence conservation or that it represents a second horizontal transfer of a closely related element. Preliminary results of sequencing of many additional genomic *Hsmar1* copies do not support either of these alternatives, in that all are highly defective and none are extremely similar in sequence to the NIB1543 transposase gene (K.K O. Walden, D.J. Lampe and H.M. Robertson, unpublished results). Nevertheless, considerable additional characterization is needed to confirm that this exceptional *Hsmar1* copy has indeed become part of a functional human gene. It will be of particular interest to determine whether the *mariner* transposase domain of this chimeric protein retains some or all of the biological functions of a *mariner* transposase (see Lampe et al., 1996), in particular whether it can bind the thousands of paired and/or solo ITRs distributed around the genome. The SET domain is shared by diverse proteins involved in chromosomal interactions and maintenance or inhibition of gene expression (Jones and Gelbart, 1993; Tschiersch et al., 1994; Stassen et al., 1995), so perhaps the addition of a partially functional *mariner* transposase domain would allow such a chromosomally located protein to manipulate DNA.

## 5. Conclusions

(1) The molecular evolution of *Hsmar1* in the primate genome is quite different from that of RNA-mediated or Class I transposons such as *LINE* and *Alu* elements (see Smit, 1996). There is no evidence for lineages of related elements. Instead, all copies appear to have been formed at around the same time, presumably within a short time of the horizontal transfer of a single active copy of *Hsmar1* into the genome of an ancestral primate, with possible additional horizontal transfers among primates and other mammals.

(2) Each copy appears to have been evolving largely independently and neutrally since then. Today almost all copies in the human and other primate genomes are therefore inactive defective molecular fossils of this early transpositional burst, along with the thousands of paired and solo ITRs derived early on in this process.

(3) It seems unlikely that any active copies of *Hsmar1* remain among the approximately 200 present in the human genome, or other primate genomes, however the exceptional copy represented by the NIB1543 cDNA clone might still encode a *mariner* transposase with some activities, albeit now apparently fused with a SET domain protein.

## Acknowledgement

We thank T.J. Stevens and J.M. Sikela for the NIB1543 clone, the IMAGE consortium for the other cDNA clones, Leona Chemnick, Morris Goodman, and Nobuyo Maeda for genomic DNA samples, Michelle Asplund and Kimberly Walden for technical assistance, and Christina Nordholm for comments on the manuscript. This work was supported by NSF grant MCB 93-17586.

## References

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., Venter, J.C., 1992. Sequence identification of 2375 human brain genes. Nature 355, 632–634.

Archibald, J.D., 1996. Fossil evidence for a late Cretaceous origin of hoofed- mammals. Science 272, 1150–1153.

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 Lett. 368, 541–546.

Berry, R., Stevens, T.J., Walter, N.A., Wilcox, A.S., Rubano, T., Hopkins, J.A., Weber, J., Goold, R., Soares, M.B., Sikela, J.M., 1995. Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. Nature Genet. 10, 415–423.

Britten, R.J., 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.

Britten, R.J., Baron, W.F., Stout, D.B., Davidson, E.H., 1988. Sources and evolution of human *Alu* repeated sequences. Proc. Natl. Acad. Sci. USA 85, 4770–4774.

Bulmer, M., Wolfe, K.H., Sharp, P.M., 1991. Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA 88, 5974–5978.

Burn, T.C., Petrovick, M.S., Hohaus, S., Rollins, B.J., Tenen, D.G., 1994. Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. Blood 84, 2776–2783.

Cavener, D.R., Ray, S.C., 1991. Eukaryotic start and stop translation sites. Nucleic Acids Res. 19, 3185–3192.

Chissoe, S.L., Bodenteich, A., Wang, Y.-F., Wang, Y.-P., Burian, D., Clifton, S.W. et al., 1995. Sequence and analysis of the human *ABL* gene, the *BCR* gene, and regions involved in the Philadelphia chromosomal translocation. Genomics 27, 67–82.

Doak, T.G., Doerder, F.P., Jahn, C.L., 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.

Garcia-Fernàndez, J., Bayascas-Ramírez, J.R., Marfany, G., Muñoz-Mármol, A.M., Casali, A., Baguñà, J., Saló, E., 1995. High copy number of highly similar *mariner*-like transposons in planarian (Platyhelminthe): evidence for a trans-phyla horizontal transfer. Mol. Biol. Evol. 12, 421–431.

Gloor, G.B., Nassif, N.A., Johnson-Schlitz, D.M., Preston, C.R., Engels, W.R., 1991. Targeted gene replacement in Drosophila via P element-induced gap repair. Science 253, 1110–1117.

Goold, R.D., diSibio, G., Xu, H., Lang, D.B., Dadgar, J., Magrane, G., Dugaiczyk, A., Smith, K.A., Cox, D.R., Masters, S.B., Myers, R.M., 1993. The development of sequence-tagged sites for human chromosome 4. Hum. Mol. Genet. 2, 1271–1288.

Hartl, D.L., 1989. Transposable element *mariner* in *Drosophila* species. In: Berg, D.E., Howe, M.M., (Eds.), Mobile DNA. American Society for Microbiology, Washington, DC, pp. 5531–5536.

Hartl, D.L., 1996. The most unkindest cut of all. Nature Genet. 12, 227–229.

Hedges, S.B., Parker, P.H., Sibley, C.G., Kumar, S., 1996. Continental breakup and the ordinal diversification of birds and mammals. Nature 381, 226–229.

Hijiya, N., Setoguchi, M., Matsuura, K., Higuchi, Y., Akizuki, S., Yamamoto, S., 1994. Cloning and characterization of the human osteopontin gene and its promoter. Biochem. J. 303, 255–262.

Hillier, L., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S. et al., 1996. Generation and analysis of 280 000 human expressed sequence tags. Genome Res. 6, 807–828.

Jeyaprakash, A., Hoy, M.A., 1995. Complete sequence of a *mariner* transposable element from the predatory mite *Metaseiulus occidentalis* isolated by an inverse PCR approach. Insect Mol. Biol. 4, 31–39.

Jones, R.S., Gelbart, W.M., 1993. The Drosophila *Polycomb*-group gene *Enhancer of zeste* contains a region with sequence similarity to *trithorax*. Mol. Cell. Biol. 13, 6357–6366.

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

Kay, R.F., Ross, C., Williams, B.A., 1997. Anthropoid origins. Science 275, 797–804.

Kiyosawa, H., Chance, P.F., 1996. Primate origin of the CMT1A-REP repeat and analysis of a putative transposon-associated recombinational hotspot. Human Mol. Gen. 5, 745–753.

Kozak, M., 1996. Interpreting cDNA sequences: some insights from studies on translation. Mamm. Genome 7, 563–574.

Lampe, D.J., Churchill, M.E.A., Robertson, H.M., 1996. A purified *mariner* transposase is sufficient to mediate transposition in vitro. EMBO J. 15, 5470–5479.

Larsen, F., Gundersen, G., Lopez, R., Prydz, H., 1992. CpG islands as gene markers in the human genome. Genomics 13, 1095–1107.

Lohe, A.R., de Aguiar, D., Hartl, D.L., 1997. Mutations in the *mariner* transposase: the D,D(35)E consensus sequence is nonfunctional. Proc. Natl. Acad. Sci USA 94, 1293–1297.

Maddison, W.P., Maddison, D.R., 1993. MacClade, Analysis of Phylogeny and Character Evolution, version 3. Sinauer Associates, Sunderland, Massachusetts, USA.

Martin, R.D., 1993. Primate origins: plugging the gaps. Nature 363, 223–234.

Miller, W.J., Paricio, N., Hagemann, S., Martinez-Sebastian, M.J., Pinsker, W., de Frutos, R., 1995. Structure and expression of clustered P element homologues in *Drosophila subobscura* and *Drosophila guanche*. Gene 156, 167–174.

Milner, C.M., Campbell, R.D., 1993. The G9a gene in the human major histocompatibility complex encodes a novel protein containing ankyrin-like repeats. Biochem. J. 290, 811–818.

Morgan, G.T., 1995. Identification in the human genome of mobile elements spread by DNA-mediated transposition. J. Mol. Biol. 254, 1–5.

Mount, S.M., 1982. A catalogue of splice junction sequences. Nucleic Acids Res. 10, 459–472.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nomura, N., Nagase, T., Miyajima, N., Sazuka, T., Tanaka, A., Sato, S., Seki, N., Kawarabayasi, Y., Ishikawa, K., Tabata, S., 1994. Prediction of the coding sequences of unidentified human genes. II. The coding sequences of 40 new genes (KIAA0041–KIAA0080) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res. 1, 223–229.

Oosumi, T., Belknap, W.R., Garlick, B., 1995. *Mariner* transposons in humans. Nature 378, 672

Reiter, L.T., Murakami, T., Koeuth, T., Pentao, L., Muzny, D.M., Gibbs, R.A., Lupski, J.R., 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a *mariner* transposon-like element. Nature. Genet. 12, 288–297.

Robertson, H.M., 1993. The *mariner* transposable element is widespread in insects. Nature 362, 241–245.

Robertson, H.M., 1995. The *Tc1-mariner* superfamily of transposons in animals. J. Insect Physiol. 41, 99–105.

Robertson, H.M., 1996. Members of the pogo superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. 252, 761–766.

Robertson, H.M., 1997. Multiple *mariner* transposons in flatworms and hydras are related to those of insects. J. Hered. 88, 195–201.

Robertson, H.M., Asplund, M.L., 1996. *Bmmar1*: a basal lineage of the *mariner* family of transposable elements in the silkworm moth *Bombyx mori*. Insect Biochem. Mol. Biol. 26, 945–954.

Robertson, H.M., MacLeod, E.G., 1993. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol. Biol. 2, 125–139.

Robertson, H.M., Martos, R., 1997. Molecular evolution of the second ancient human *mariner* transposon, *Hsmar2*, illustrates patterns of neutral evolution in the human genome lineage. Gene 205, 219–228.

Robertson, H.M., Soto-Adames, F.N., Walden, K.K.O., Avancini, R.M.P., Lampe, D.J., 1997. The *mariner* transposons of animals: horizontally jumping genes. In: Syvanen, M., Kado, C. (Eds.), Horizontal Gene Transfer. Chapman and Hall, London, in press.

Robertson, H.M., Zumpano, K.L., Lohe, A.R., Hartl, D.L., 1996. Reconstructing the ancient *mariners* of humans. Nature Genet. 12, 360–361.

Rowen, L., Koop, B.F., Hood, L., 1996. The complete 685-kilobase DNA sequence of the human $\beta$ T cell receptor locus. Science 272, 1755–1762.

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

Schlötterer, C., Tautz, D., 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 20, 211–215.

Schneider, H., Schneider, M.P.C., Sampaio, I., Harada, M.L., Stanhope, M., Czelusniak, J., Goodman, M., 1993. Molecular phylogeny of the New World Monkeys (Platyrrhini, Primates). Mol. Phylo. Evol. 2, 225–242.

Sedensky, M.M., Hudson, S.J., Everson, B., Morgan, P.G., 1994. Identification of a *mariner*-like repetitive sequence in *C. elegans*. Nucleic Acids Res. 22, 1719–1723.

Smit, A.F.A., 1996. The origin of interspersed repeats in the human genome. Cur. Opin. Genet. Dev. 6, 743–748.

Smit, A.F.A., Riggs, A.D., 1996. *Tiggers* and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA 93, 1443–1448.

Stassen, M.J., Bailey, D., Nelson, S., Chinwalla, V., Harte, P.J., 1995. The Drosophila trithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins. Mech. Dev. 52, 209–223.

Swofford, D.L., 1993. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.1. Smithsonian Institution, Washington, DC.

Tschiersch, B., Hofmann, A., Krauss, V., Dorn, R., Korge, G., Reuter, G., 1994. The protein encoded by the *Drosophila* position-effect variegation suppressor gene *Su(var)3-9* combines domains of antagonistic regulators of homeotic gene complexes. EMBO J. 13, 3822–3831.

Zhang, Q., Wrana, J.L., Sodek, J., Characterization of the promoter region of the porcine *opn* (osteopontin secreted phosphoprotein 1) gene: identification of positive and negative regulatory elements and a ‘silent’ second promoter. 1992. Eur. J. Biochem. 207, 649–659.