Table 1 **Human cDNAs supporting the different transcript variants of *SETMAR**

GenBank Acc. Nb	Tissue	match over #bp (% id.)	supported aceview transcript[#]
AF054989	infant brain	2066 bp (100%)	SETMAR.a
AA323982	cerebellum brain	289 bp (97%)	SETMAR.a
BF743200	breast	476 bp (97%)	SETMAR.a
BP374382	umbilical cord	582 bp (100%)	SETMAR.a
BP377589	uterus	581 bp (100%)	SETMAR.a
BP378351	uterus	581 bp (100%)	SETMAR.a
BP379044	uterus	452 bp (83%)	SETMAR.a
BX388110	B CELLS (RAMOS CELL LINE)	911 bp (100%)	SETMAR.a
BX460192	Fetal Brain brain	921 bp (93%)	SETMAR.a
F13556	total brain brain	304 bp (100%)	SETMAR.a
AL515729	neuroblastoma cells brain neuroblastoma	1007 bp (93%)	SETMAR.a
AL528596	neuroblastoma cells brain Neuroblastoma	994 bp (92%)	SETMAR.a
BF237728	leiomyosarcoma cell line uterus	515 bp (45%)	SETMAR.a
BG286864	transitional cell papilloma, cell line bladder	681 bp (72%)	SETMAR.a
BI545504	hippocampus brain	599 bp (99%)	SETMAR.a
AW962016	nd	637 bp (92%)	SETMAR.a
AV713754	dendritic cells, mature	617 bp (100%)	SETMAR.a
T64992	Liver and Spleen	560 bp (99%)	SETMAR.a
N62795	multiple sclerosis lesions	578 bp (99%)	SETMAR.a
N74675	fetal lung	338 bp (99%)	SETMAR.a
W05064	fetal lung	281 bp (99%)	SETMAR.a
W93323	heart fetal	51 bp (100%)	SETMAR.a
W93433	heart fetal	346 bp (99%)	SETMAR.a
AA152184	uterus	467 bp (99%)	SETMAR.a
AA535990	colon	533 bp (99%)	SETMAR.a
AA505765	breast	524 bp (99%)	SETMAR.a
AA244414	prostate epithelial cells	375 bp (98%)	SETMAR.a
AA779676	total fetus, 8-9 weeks old	595 bp (99%)	SETMAR.a
AA872060	germinal center B cell	613 bp (100%)	SETMAR.a
AA953286	2 pooled tumors (clear cell type) kidney	375 bp (100%)	SETMAR.a
AI128753	heart	477 bp (100%)	SETMAR.a
AI223375	testis	446 bp (100%)	SETMAR.a
AI301364	placenta	428 bp (98%)	SETMAR.a
AI635647	well-differentiated endometrialadenocarcinoma (uterus)	723 bp (100%)	SETMAR.a
AW055198	glioblastoma (pooled) brain	653 bp (100%)	SETMAR.a
AW250939	small cell carcinoma lung	322 bp (81%)	SETMAR.a
CF141437	prostate	605 bp (99%)	SETMAR.a
CB045016	juvenile granulosa tumor	427 bp (100%)	SETMAR.a
CB045017	juvenile granulosa tumor	657 bp (99%)	SETMAR.a
BF939387	ovary, fibrotheoma	382 bp (99%)	SETMAR.a
BF194883	kidney	226 bp (100%)	SETMAR.a
BQ020736	Metastatic Chondrosarcoma lung	532 bp (99%)	SETMAR.a
BM750512	Ascites stomach	115 bp (100%)	SETMAR.a
T16164	infant brain	348 bp (100%)	SETMAR.a
T16607	infant brain	245 bp (100%)	SETMAR.a
BM970828	lung	600 bp (100%)	SETMAR.a
CD367822	Alveolar Macrophage,lung	206 bp (100%)	SETMAR.a
CD366647	Alveolar Macrophage,lung	596 bp (100%)	SETMAR.a
AA234618	Pooled human melanocyte, fetal heart, and pregnant uterus	438 bp (100%)	SETMAR.a, SETMAR.b
AA236558	Pooled human melanocyte, fetal heart, and pregnant uterus	338 bp (100%)	SETMAR.a, SETMAR.b
BG393984	testis,embryonal carcinoma	839 bp (78%)	SETMAR.a, SETMAR.b
BC008931	small cell carcinoma lung	1310 bp (100%)	SETMAR.b
AW248177	small cell carcinoma lung	556 bp (100%)	SETMAR.b
BG284487	adenocarcinoma, cell line prostate	691 bp (100%)	SETMAR.b
BM677483	fetal eye eye	628 bp (100%)	SETMAR.b
BM723028	fetal eye eye	705 bp (100%)	SETMAR.b
CA313648	Human Lung Epithelial cells	693 bp (100%)	SETMAR.c
BC011635	Uterus, leiomyosarcoma	1653 bp (100%)	SETMAR.d
CR600492	Fetal Brain	1651 bp (100%)	SETMAR.d
AW160952	frontal lobe brain	649 bp (100%)	SETMAR.d
AW162556	frontal lobe brain	589 bp (100%)	SETMAR.d
AI547025	prostate	349 bp (56%)	SETMAR.d
BX451998	Fetal Brain brain	924 bp (100%)	SETMAR.d
BX451999	Fetal Brain brain	926 bp (98%)	SETMAR.d
AA972798	pooled	288 bp (100%)	SETMAR.d
BE890511	melanotic melanoma skin	557 bp (100%)	SETMAR.d
AW583371	Islets of Langerhans pancreas	294 bp (100%)	SETMAR.d
AW583590	ling, small cell carcinoma	247 bp (100%)	SETMAR.d
BU164805	retinoblastoma eye	865 bp (96%)	SETMAR.d
BQ650846	hepatocellular carcinoma, cell line liver	914 bp (99%)	SETMAR.d
BQ651111	hepatocellular carcinoma, cell line liver	863 bp (91%)	SETMAR.d
W27060	retina eye	474 bp (81%)	SETMAR.d
AW583427	Islets of Langerhans pancreas	467 bp (100%)	SETMAR.d

BI549653	hippocampus brain	779 bp (99%)	SETMAR.f
BI546231	hippocampus brain	709 bp (83%)	SETMAR.g
BG505051	testis,embryonal carcinoma	701 bp (100%)	SETMAR.h
CD696799	nasopharynx	447 bp (93%)	SETMAR.i
AL703233	human skeletal muscle	668 bp (100%)	SETMAR.j
BP378735	uterus	550 bp (100%)	SETMAR.k
BG399243	kidney	641 bp (100%)	SETMAR.k
BG479260	placenta, choriocarcinoma	690 bp (100%)	SETMAR.k

* The data was obtained from the AceView database at http://www.ncbi.nih.gov/IEB/Research/Acembly/

Refers to the different transcript variants annotated in the Aceview database. All transcripts are predicted to initiate within the same 5' region upstream of the SET-coding sequence, but varies in their splicing pattern and 3' termination. Three transcript variants (a, b and d) are supported by multiple cDNA clones from various tissues, thus we suspect these to be the most biologically relevant transcripts. Transcript a corresponds to the full-length version of SETMAR (SET=exons 1-2, plus MAR=exon 3) which potentially encodes the complete fusion protein of 671 aa. This transcript is the most represented in the human cDNA database, with 48 supporting clones from 18 different normal and/or cancerous tissues. Transcript d is the second most abundant transcript (supported by 14 cDNA cloned from 10 different tissues). It corresponds to the first two exons of SETMAR and terminate within the intron 2 of SETMAR. Thus it potentially encodes the complete SET region alone. This transcript may still encode a functional protein homolog of the ancestral SET gene. The stop codon of the ancestral SET gene was lost (see Fig. 2), but another stop codon located 153 bp downstream of the original one may terminate the protein. Hence, the creation of SETMAR might not have completely 'displaced' the function of the ancestral SET-only protein.

Table 2: PAML results

Rooting strategy	*Hsmar1* consensus	*Mmmar1* consensus	No consensus
1) K_A/K_S ANALYSES			
Entire transposase sequence (1029 bp)			
Free K_A/K_S ratio vs. single K_A/K_S ratio	$P = 0.162$	$P = 0.118$	$P = 0.053$
Single K_A/K_S ratio estimate	0.311	0.387	0.289
Single K_A/K_S ratio vs. $K_A/K_S = 1$ (neutrality)	$P < 0.000001$	$P < 0.000001$	$P < 0.000001$
5' half (positions 1 to 513)			
Free K_A/K_S ratio vs. single K_A/K_S ratio	$P = 0.504$	$P = 0.144$	$P = 0.489$
Single K_A/K_S ratio estimate	0.106	0.202	0.084
Single K_A/K_S ratio vs. $K_A/K_S = 1$ (neutrality)	$P < 0.000001$	$P < 0.000001$	$P < 0.000001$
3' half (positions 514 to 1029)			
Free K_A/K_S ratio vs. single K_A/K_S ratio	$P = 0.087$	$P = 0.086$	$P = 0.064$
Single K_A/K_S ratio estimate	0.695	0.673	0.721
Single K_A/K_S ratio vs. $K_A/K_S = 1$ (neutrality)	$P = 0.163$	$P = 0.111$	$P = 0.282$
2) SITE SELECTION ANALYSES			
Model M1a (no positive selection) vs. model M2a (positive selection)	$P = 0.833$	$P = 0.457$	$P = 0.354$

Table 3: PCR conditions (Ann. Temp., annealing temperature).

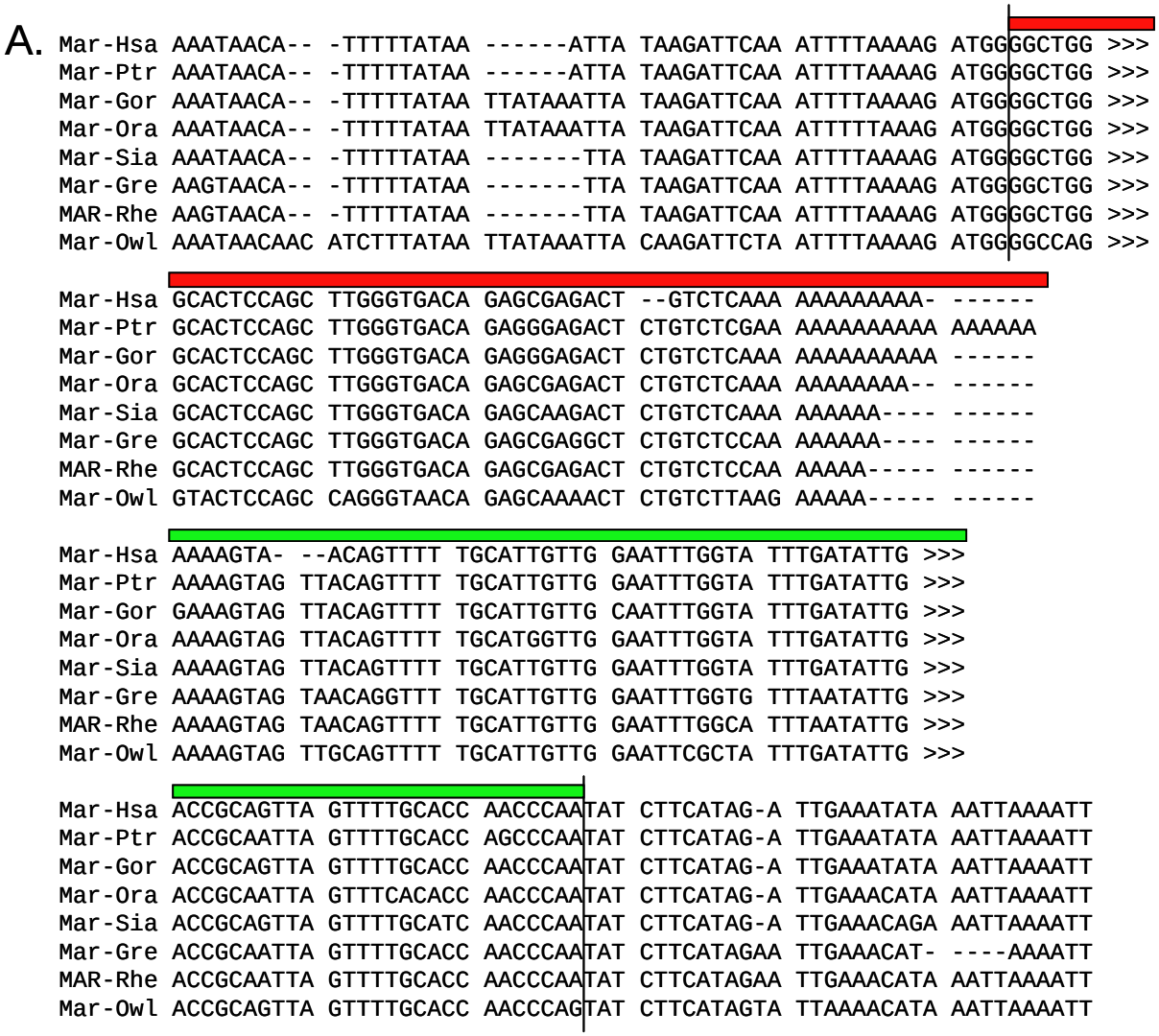
Species name	Scientific name	SET exon 2			MAR region	
		Primer pair	Ann. Temp.	Number of cycles	Primer pair	Ann. Temp.
Human	*Homo sapiens*	-	-	-	F/R	56 ˚C
Chimpanzee	*Pan troglodytes*	Fint/R	60 ˚C	25	F/R	54 ˚C
Gorilla	*Gorilla gorilla*	Fint/R	60 ˚C	25	F/R	60 ˚C
Orangutan	*Pongo pygmaeus*	Fint/R	60 ˚C	25	F/R	60 ˚C
Siamang	*Hylobates syndactylus*	Fint/R	60 ˚C	25	F/R	60 ˚C
African Green Monkey	*Chlorocebus aethiops*	Fint/R	60 ˚C	25	F/R	52 ˚C
Owl Monkey	*Aotus trivirgatus*	Fint/R	60 ˚C	25	F2/R	52 ˚C
Tarsier	*Tarsius syrichta*	Fint/Rev6	58 ˚C	35	F/R2	46 ˚C
Galago	*Galago senegalensis*	Fint/Rev6	58 ˚C	35	-	-

Table 4: List of PCR primers

Locus	Primer name	Primer sequence (5' > 3')
MAR region	F	TGCCATATTTTTGAGAATGTTGA
	R	TCCGTGGAATCTACTTCAAATG
	F2	ATTTTTGAGAATGTTGACACTTC
	R2	CATTCTTAATAAAGTCCGTGGA
	Fint	TGCAGTGGTGGTTCAAGAAG
	Rint	TTGTAAGGGGATCAGCTTCG
SET exon 2	Fint	GATGATTCCTGTCCGAATTGA
	R	TAGCCAACCACTGTCTTCCA
	Rev6	GCAAGCAGGGTTTMGGTATG

Figure 4: Alignment of the orthologous *MAR* region in nine primates and dog, in which the *Hsmar1* transposon is present (A) or absent (B) and alignment of the orthologous 3' end of the second exon of the *SET* gene in nine primate species (C).

Hsa, human; Ptr, common chimpanzee; Gor, gorilla; Ora, orangutan; Sia, siamang; Gre, African green monkey; Rhe, rhesus macaque; Owl, owl monkey; Tar, tarsier; -, sequence gap; >>>, a portion of the sequence is not shown. The colored boxes above delimited alignment portions indicate the location of the *Alu*Sx insertion (red), the *Hsmar1* transposon insertion (green), the 3' end of SET exon 2 (orange), the 27-bp deletion in anthropoid primates relative to tarsier (blue) and the 77-bp exonized sequence linking the former SET exon 2 to the donor splice site of the second intron of *SETMAR* (yellow). The location of the SET stop codon in tarsier is shown in bold and the start of intron 2 of *SETMAR* at dinucleotide GT is underlined in anthropoid primates.



B.

```
MAR-Hsa CACTTC---T TTTGAGAGAA AACATTTAAA AATATACTTC CACTTGACTA TTATCCCATG
MAR-Tar CACGTCTATT T--------- ----TTTAAA AATATACTTC TGCTTGACTA TTATCCTATG
MAR-Dog CACTTCTGTT TTTGAGAACA AACATTAAAA GATATA---- TATTTTTTTA TTATCT----

MAR-Hsa ATAAAATAAC AT-----TTT TAT------A AATTATAAG- -ATTCAAATT TTAAAAGATG
MAR-Tar GTAAAATAAT ATAAAAGTTT TAT---TACA AATTACAAG- -ATTTGAATT TTTTAGAGTG
MAR-Dog -TAAAATA-T ATATAAACTT TACAACTAAA AAGTACTGCA AACTCGAAGT TTAAAAGGTG

MAR-Hsa G----GGCTG GGTGCAGTGG CTCACACCTG TAATCCCAGC ACTTTGGGAG GCTGAGGCAG >>>
MAR-Tar GCATG----- ---------- ---------- ---------- ---------- ---------- >>>
MAR-Dog GCATA----- ---------- ---------- ---------- ---------- ---------- >>>

MAR-Hsa AGCCTAGTTA TAATGATTTA AAATTCACAG TCCAAAACCG CAGTTAGTTT TGCACCAACC
MAR-Tar ---------- ---------- ---------- ---------- ---------- ----------
MAR-Dog ---------- ---------- ---------- ---------- ---------- ----------

MAR-Hsa CAATATCT-- -TCATAG-AT TGAAATATAA ATTAAAATTG CATTTGAAGT AGAT
MAR-Tar ---TATCTAG CTCATGGAAT TGAAACA--A ATAAAAATTG GCTTTAAAGT AAAT
MAR-Dog ---TATCT-- -TCATGAAAT TAAAACATAA A----AATTG GCTTTGAAGT CAAT
```

C.

```
SETex2-Tar   ATCAATTAGG CCGTTTCCTG GCCTCTCTCA GTGGGTGGGA ATGAGGGGGA ACTAAGCTTA
SETex2-Owl   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Rhe   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Gre   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Sia   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Ora   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Gor   ATCAGGT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG
SETex2-Ptr   ATCAGTT--- ---------- ---------- ----GTGGAC ATGAGAAGGA ACCCAGCATG
SETex2-Hsa   ATCAGTT--- ---------- ---------- ----GTGGAA ATGAGAAGGA ACCCAGCATG

SETex2-Tar   AGTGGCTCAG CCACTTCTGC CTTCCTCTCT TGCAAGCAGT TTATGCTCGA GGTAAGTCTG
SETex2-Owl   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Rhe   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Gre   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Sia   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Ora   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Gor   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Ptr   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
SETex2-Hsa   TGTGGCTCAG CCCCTTCTGT GTTCCCCTCC TGCAAGCGAT TGACCCTTGA GGTGAGTCTG
```

Figure 5: Phylogenetic tree of 205 human *Hsmar1* transposons and the *MAR* region in eight primates.

The tree was reconstructed using the neighbor-joining algorithm and the Kimura 2-parameter nucleotide substitution model, based on the DNA sequence corresponding to the transposase gene. The tree is rooted with the consensus sequence of the mouse *Mmmar1 mariner*-like transposon. Unlabelled terminal branches correspond to *Hsmar1* transposons. The monophyletic clade of *MAR* sequences is highlighted in gray (Hsa, human; Ptr, common chimpanzee; Gor, gorilla; Ora, orangutan; Sia, siamang; Gre, African green monkey; Rhe, rhesus macaque; Owl, owl monkey). The tree is parted in two halves, at the level of the two horizontal bars interrupting deep branches of the tree.


Mmmar1
Mar-Owl
Mar-Gre
Mar-Rhe
Mar-Sla
Mar-Ora
Mar-Gor
Mar-Ptr
Mar-Hsa
0.05

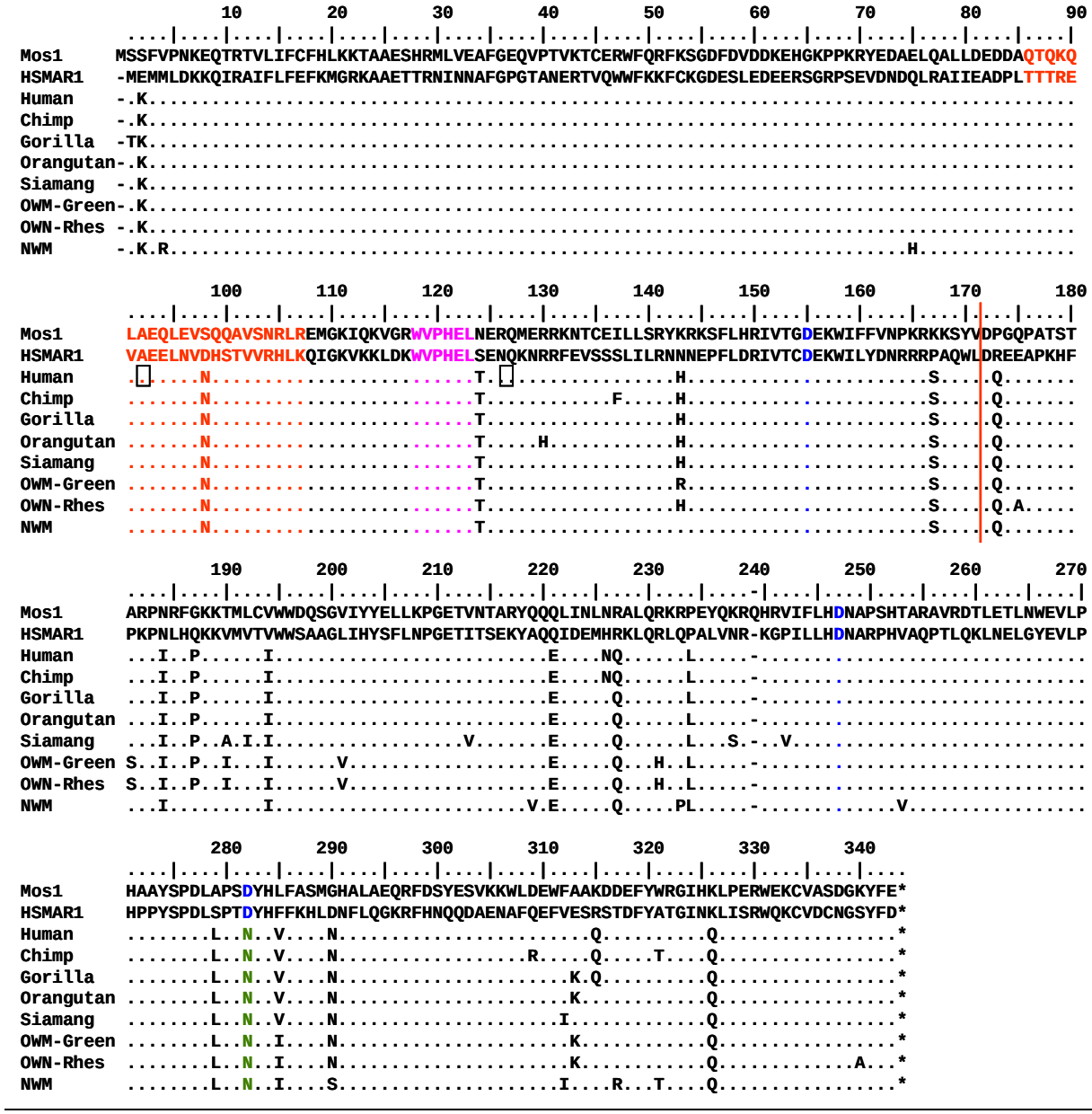
Figure 6: A cryptic acceptor splice site in *Hsmar1* allows the recurrent transcriptional capture of *Hsmar1* transposase sequences by splicing.

Alignment of the intron/exon junction of four different human genomic loci producing a chimeric transcript involving the junction of an *Hsmar1*-unrelated exon to an *Hsmar1* transposase sequence located dowstream. The conserved position of the intron/exon 3' junction (indicated by "^") highlights the fact that the acceptor splice site is located at the exact same position in the four different *Hsmar1* sequences. The first line is the consensus *Hsmar1* sequence. The second sequence is from the *SETMAR* locus (intron/exon junction supported by multiple human EST, e.g. U80776). The accession numbers for EST supporting the three other junctions are given. The putative start codon of the *Hsmar1* transposase gene is boxed.

```
        1         11        21        31        41        50
        |         |         |         |         |         |
HSMAR1: TTAGGTTGGTGCAAAAGTAATTGCGGTTTTTGCATTGTTGGAATTTGCCG

        51        61        71        81        91        100
        |         |         |         |         |         |
HSMAR1: TTTGATATTGGAATACATTCTTAAATAAATGTGGTTATGTTATACATCAT

        101       111       121       131       141       150
        |         |         |         |         |         |
HSMAR1: TTTAATGCGCATTTCTCGCTTTACGTTTTTTTGCTAATGACTTATTACTT

        151       161       171      181       191       200
        |         |         |    ^    |         |         |
HSMAR1: GCTGTTTATTTTATGTTTATTTTAG^ACTATGGAAATGATGTTAGACAAAA
SETMAR: GCTGTTTA-----TGTTTATTTTAG^ACTATGAAAATG...
U80773: GCTGTTTATTTTGTGTTTATTTTAG^ATGATGGAAATG...
U80769: GCTGTTCATTTCATG----TTTTAG^ACTATGAAAATG...
U80764: GCTGTTTTGTTTATGTTTATTTTAG^ACTATGGAAATG...
```

Figure 7: Protein sequence alignment of the active *mos1* transposase from *Drosophila mauritiana* (GenBank X78906), *Hsmar1* consensus transposase (AAC52010) and MAR region from 8 primate species.

See legend to Fig. 2 for species names and abbreviations. The catalytic DD34D triad is shown in blue, while the N residue replacing the last D in MAR is in green. The predicted HTH motif is shown in red. The motif was predicted with 100% probability in all sequences using the *npsa* prediction server (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hth.html). Note that the mutation D98N in MAR sequences does not seem to alter the potential of this region to form a HTH structure. In fact, the motif is predicted with a higher score of 6.40 (vs. 5.63 in the *Hsmar1* consensus). The boxed residues in the human MAR sequence mark the last residue of the deletion peptides MAR-N92 and MAR-N126, respectively (see Fig. 3). The WVPHEL motif, highly conserved in *mariner* transposases and involved in the ability of the *Mos1* transposase to assemble organized complexes of DNA with transposase tetramers (*5*) is shown in purple. Note that the motif is conserved in MAR and included within the MAR-N126 peptide, in agreement with the formation of at least two protein-DNA complexes with different electrophoretic mobility in EMSA experiments (Fig. 3c). The vertical red line represents the separation between the two halves of the proteins for which K_A/K_S analyses were conducted independently (see text).


10 20 30 40 50 60 70 80 90
Mos1 MSSFVPNKEQTRTVLIFCFHLKKTAAESHRMLVEAFGEQVPTVKTCERWFQRFKSGDFDVDDKEHGKPPKRYEDAELQALLDEDDAQTQKQ
HSMAR1 -MEMMLDKKQIRAIFLFEFKMGRKAAETTRNINNAFGPGTANERTVQWWFKKFCKGDESLEDEERSGRPSEVDNDQLRAIIEADPLTTTRE
Human -.K
Chimp -.K
Gorilla -TK
Orangutan-.K
Siamang -.K
OWM-Green-.K
OWN-Rhes -.K
NWM -.K.R
100 110 120 130 140 150 160 170 180
Mos1 LAEQLEVSQQAVSNRLREMGKIQKVGRWVPHELNERQMERRKNTCEILLSRYKRKSFLHRIVTGDEKWIFFVNPKRKKSYVDPGQPATST
HSMAR1 VAEELNVDHSTVVRHLKQIGKVKKLDKWVPHELSENQKNRRFEVSSSLILRNNNEPFLDRIVTCDEKWILYDNRRRPAQWLDREEAPKHF
Human N T H S Q
Chimp N T F H S Q
Gorilla N T H S Q
Orangutan N T H H S Q
Siamang N T H S Q
OWM-Green N T R S Q
OWN-Rhes N T H S Q A
NWM N T S Q
190 200 210 220 230 240 250 260 270
Mos1 ARPNRFGKKTMLCVWWDQSGVIYYELLKPGETVNTARYQQQLINLNRALQRKRPEYQKRQHRVIFLHDNAPSHTARAVRDTLETLNWEVLP
HSMAR1 PKPNLHQKKVMVTVWWSAAGLIHYSFLNPGETITSEKYAQQIDEMHRKLQRLQPALVNR-KGPILLHDNARPHVAQPTLQKLNELGYEVLP
Human I..P..... I E NQ L -
Chimp I..P..... I E NQ L -
Gorilla I..P..... I E Q L -
Orangutan I..P..... I E Q L -
Siamang I..P..A.I.I V E Q L..S.-..V
OWM-Green S..I..P..I..I.....V E Q..H..L -
OWN-Rhes S..I..P..I..I.....V E Q..H..L -
NWM I..... I V.E Q PL -.....V
280 290 300 310 320 330 340
Mos1 HAAYSPDLAPSDYHLFASMGHALAEQRFDSYESVKKWLDEWFAAKDDEFYWRGIHKLPERWEKCVASDGKYFE*
HSMAR1 HPPYSPDLSPTDYHFFKHLDNFLQGKRFHNQQDAENAFQEFVESRSTDFYATGINKLISRWQKCVDCNGSYFD*
Human L..N..V....N Q Q *
Chimp L..N..V....N R Q T Q *
Gorilla L..N..V....N K.Q Q *
Orangutan L..N..V....N K Q *
Siamang L..N..V....N I Q *
OWM-Green L..N..I....N K Q *
OWN-Rhes L..N..I....N K Q A *
NWM L..N..I....S I R T Q *