An Ancient RNase H1 Splice Junction Mutant Preserved in a 19-Million-Year-Old Genetic Fossil in Ape Genomes

E. J. Devor and K. Moffat-Wilson

From Molecular Genetics and Bioinformatics, Integrated DNA Technologies, Coralville, IA 52241 (Devor and Moffat-Wilson) and Department of Anthropology, University of Kansas, Lawrence, KS 66045 (Devor).

Address correspondence to Eric J. Devor, Molecular Genetics and Bioinformatics, Integrated DNA Technologies, 1710 Commercial Park, Coralville, IA 52241, or e-mail: rdevor@idtdna.com.

A retroprocessed pseudogene (retropseudogene) descended from the gene encoding ribonuclease (RNase) H1 has been found in ape genomes that preserves a splice junction mutation event that altered the carboxyl-terminal end of the enzyme. The GT → GC transition mutant at the 5' splice junction of RNase H1 exon 7/intron 7 led to the absence of exon 8 and more than 1 kb of intron 7 sequence being substituted. Comparison of source gene and pseudogene sequences indicates that the retrotranscription event occurred 19 million years ago. Present in these sequences is an in-frame stop and several available polyadenylation signals, suggesting that the mutant allele could have been translated. At the present time, the genetic fossil is the only evidence that the mutation ever occurred, and thus represents an archival marker of an ancient genetic event in primate evolution.

Ribonuclease (RNase) H's are ubiquitously expressed enzymes that specifically degrade the RNA moiety of RNA–DNA duplexes. These enzymes are found in prokaryotes and eukaryotes as well as in retroviruses (Grouch and Toulme 1998). Though they vary in size and molecular weight from species to species, RNase H's fall into two major classes based on activity and cofactor requirements. RNase H2 has a molecular weight between 68,000 and 90,000 Da, is activated in the presence of Mg2+ and Mn2+, and is not inhibited by sulfhydryl reagents such as N-ethylmaleimide (NEM). RNase H1 is smaller, with molecular weights between 35,000 and 45,000 Da, is activated by Mg2+ but inhibited by Mn2+, and is inactivated by sulfhydryl reagents (Eder and Walder 1991). While the biological roles of mammalian RNase H's are not completely understood, it has been suggested that RNase H2 is involved in DNA transcription and RNase H1 is involved in DNA replication (Wu et al. 1999). The gene encoding human RNase H2 (RNASEH2) is located on chromosome 19p13 and the gene encoding human RNase H1 (RNASEH1) is located on chromosome 2p25.

Retroprocessed pseudogenes, or retropseudogenes, arise through the processes of reverse transcription of mature mRNAs and subsequent retrotransposition into new locales in the genome (Vanin 1985). This process results in retropseudogenes displaying a number of distinctive characteristics. Among these are that they are intronless, contain no upstream regulatory sequences, retain vestiges of a 3' poly-A tract, and are flanked by direct repeats generated from sequence at the insertion site (Mighell et al. 2000). Once thought to be mere genetic curiosities, retropseudogenes have been found in the genomes of bacteria, plants, insects, and vertebrates (Mighell et al. 2000). They are most common in the genomes of mammals. Gonclaves et al. (2000) estimate that as many as 23,000 to 33,000 such loci exist in the human genome. Recently Harrison et al. (2002) suggested that this number might actually be fewer than 20,000. Given that the average size of a retropseudogene is 1 kb (Devor and Moffat-Wilson 2003), this lowered estimate still represents from 0.5% to 1% of the human genome.

Human RNASEH1 has three retropseudogenes: RNASEH1P1 on chromosome 17p11.2, RNASEH1P2 on chromosome 1q32, and RNASEH1P3 also on chromosome 17p11.2. RNASEH1P1 and RNASEH1P3 are nearly identical and each is located in a copy of a 24 kb duplication, dup(17) (p11.2p11.2), that lies in a well-known unstable region of chromosome 17 (Park et al. 2002; Stankiewicz et al. 2001). The sequence of RNASEH1P1 shows that the first seven exons of RNASEH1 are present as expected in a retropseudogene, but that exon 8 is missing. In place of exon 8 is 1070 bp of RNASEH1 intron 7 sequence. The
absence of exon 8 sequence in RNASEH1P1 and the continuation of the RNA transcript into intron 7 is due to a T → C transition in the intron 7 5’ splice junction site (GT → GC). In addition, RNASEH1 intron 7 and pseudogene sequence contain an in-frame stop codon 36 bp 3’ of the splice junction and eight potential polyadenylation signals further downstream. Therefore the pseudogene has not only preserved evidence of an ancient mutation but also raises the possibility that the allelic form of RNASEH1 could have been transcribed.

Materials and Methods

RNASEH1P1 and RNASEH1P3 were detected via a BLAST search of the human genome using the human RNase H1 cDNA sequence (GenBank accession no. AF048995). The complete sequence of RNASEH1P1 is contained in chromosome 17 contig NT_030843 and RNASEH1P3 is contained in chromosome 17 contig NT_010718. Pseudogene sequence was used to design polymerase chain reaction (PCR) primers that flank the splice junction mutation site. Primer sequences are exon 7 forward, 5’-GGGGAAAGGAGGATCAGAATACAAAG-3’, and intron 7 reverse, 5’-TGGAAGATCTTTATATGTTTCAATTAACCA-3’. Primer sequences were chosen and melting temperature and secondary structure analyses were carried out using the BioTools primer design and analysis software available on-line at Integrated DNA Technologies (www.idtdna.com). Primers were synthesized at Integrated DNA Technologies using standard phosphoramidite chemistries. PCR amplifications were performed on genomic DNAs using standard conditions (68°C for 5 min, 94°C for 30 s, 56°C for 30 s, 72°C for 45 s, and 72°C for 7 min). In all cases the expected 179 bp amplicon was obtained.

PCR amplicons were cloned into the pCR II-TOPO vector (Invitrogen) following manufacturer’s protocols. All clones were sequenced in both directions on an Applied Biosystems Model 310 Automated Fluorescence DNA Sequencer.

Primate genomic DNAs used were obtained from a number of sources. Chimpanzee (*Pan troglodytes*) and olive baboon (*Papio anubis*) were provided by the San Diego Zoo Center for Reproduction of Endangered Species (CRES); Assamese macaque (*Macaca assamensis*) and vervet monkey (*Chlorocebus aethiops*) were provided by the Sochi Institute for Medical Primatology in Russia.

Results

The chromosome 17 RNASEH1 retropseudogene, RNASEH1P1, presents with all of the canonical features of a retropseudogene except that all of exon 8 and the 3’ untranslated region is missing. Following the end of exon 7 is 1070 bp of RNASEH1 intron 7. This intron 7 continuation ends at an 11 base sequence we have identified as the insertion target site duplication (TSD) (5’-TTTTGTCTATT; 3’-TTTATGTCTATT). Alignment of RNASEH1P1 sequence with RNASEH1 exons 1–7 plus the 1070 bp intron 7 continuation yields a 5.4% nucleotide sequence divergence (P = 0.054). Included in this sequence divergence of RNASEH1P1 from RNASEH1 are 93 single base substitutions (68 transitions and 25 transversions). In addition, indels in exons 1–7 of RNASEH1P1 lead to the creation of numerous stop codons. Thus, as with the vast majority of retropseudogenes, this locus would not have been functional. Applying the expression T = K/2r, where K is the Jukes-Cantor correction for multiple mutations at the same site, K = -0.75ln(1-4/3P) and the mutation rate, r, is taken to be 1.5 × 10^-9/site/year (Li 1997; Sakoyama et al. 1987), RNASEH1P1 was retrotranscribed 19.0 million years ago.

PCR primers were designed that amplify a 179 bp amplicon from both RNASEH1 and RNASEH1P1. The primers flank the T → C transition site at the exon 7/intron 7 splice junction. This transition alters the recognition site of the restriction endonuclease HphI (GGTTGA → GGGGA) such that RNASEH1 amplicons will be cleaved with this enzyme and pseudogene amplicons will not. PCR amplification of genomic DNA from human (n = 7), chimpanzee (n = 2), western lowland gorilla (n = 1), orangutan (n = 2), siamang (n = 1), olive baboon (n = 13), vervet monkey (n = 12), and Assamese macaque (n = 6) yielded 179 bp amplicons in all cases. Among these, only orangutan amplicons could not be cleaved with HphI. Human, chimp, gorilla, and siamang amplicons yielded both cleaved and uncleaved fragments, while baboon, vervet monkey, and macaque presented only fully uncleaved fragments. Amplicons were cloned into the pCR II-TOPO vector (Invitrogen) and several clones from each species sequenced. Amplicon sequences from all eight species are presented in Figure 1A. Amplicon sequences show that only ape genomes contain the chromosome 17 retropseudogene. This result is consistent with the 19-million-year age estimate for retrotranscription (cf., Friedberg and Rhoads 2000; Nei and Glazko 2002; Szalay and Delson 1979). The failure of HphI to cleave any orangutan amplicon is explained by an A → C transversion mutation in the HphI site in orangutan RNASEH1. Also seen in these sequence data is the in-frame stop codon (TGA) in the intron 7 sequence of both RNASEH1 and the pseudogene. This is immediately followed (3 bp) by a minor polyadenylation signal (AA-TAAG). Seven additional potential polyadenylation signals are located downstream (Figure 1B). Among these additional potential sites is a canonical AATAAA 528 bp downstream, two AATAAA sequences 74 bp and 136 bp downstream, and a rare signal (ATTACA) 1019 bp downstream from the stop codon. The AATAAA site and one of the two ATTAAA sites are found within the Alu elements that were inserted in intron 7 long before the T → C mutation at the splice junction (cf., Kapitonov and Jurka 1996). Only one of the potential polyadenylation signals is located near the 3’ TSD (ATTACA, 5 bp upstream).
Figure 1. (A) Comparative RNASEH1 and pseudogene amplicon sequences for eight primate species. Conserved nucleotides are indicated by dots and the preserved splice junction mutation in the pseudogene is bold. Also shown are the in-frame stop and the proximal polyadenylation signal. Genomic sequences from NT_030834 and NT_010718 containing the two RNASEH1 pseudogenes designated as RNASEH1P1 and RNASEH1P3 show no differences between the two in this region. (B) A map of RNASEH1P1 from exon 7 to the 3' TSD showing the location of all potential polyadenylation signals and both Alu elements.
RESOLVING RNASEH1P1 AND RNASEH1P3

As noted, the processed pseudogenes RNASEH1P1 and RNASEH1P3 are nearly identical. Moreover, the genomic clones in which each resides, AC090774 and AC098850 for RNASEH1P1 and RNASEH1P3, respectively, show that they are part of a 24 kb duplication. PCR amplification and direct sequencing of genomic DNA from human, chimpanzee, gorilla, orangutan, and siamang indicate that the proximal copy is present in all five species, while the distal copy is found only in human, chimpanzee, and gorilla genomes. Direct sequence comparison of RNASEH1P1 and RNASEH1P3 yields an estimated separation time of 8.4 million years. This estimate is confirmed throughout the duplication. Thus we suggest that the copy located centromeric to the proximal Smith-Magenis syndrome repeat (SMR-REP), adjacent to the low-copy repeat known as LCR17pD, is the original 19-million-year-old retroposed copy and that the more distal copy that lies telomeric to the distal SMS-repeat was duplicated approximately 10 million years later (Figure 2). For the sake of consistency with the designations assigned by the HUGO Gene Nomenclature Database (www.gene.ucl.ac.uk/nomenclature/), we refer to the original as RNASEH1P1 and the duplicate copy as RNASEH1P3.

DISCUSSION

Retropseudogenes are reverse transcribed from mature mRNAs using an endogenous reverse transcriptase (Maestre et al. 1995) and then mobilized for retroposition by L1 retroposon-encoded enzymes acting in trans (Cost and Boeke 1998; Esnault et al. 2000; Szak et al. 2001). The origin of the RNASEH1P1 pseudogene is most likely to have been either a point mutation in RNASEH1, in which the mutated allele was reverse transcribed, or a misincorporation of C for U during transcription of the wild-type form. In either case, the only evidence of this event at the present time is its preservation in the genetic fossil. The intron 7 sequence continuation resulting from the splice junction mutation in RNASEH1 contains an in-frame stop and several potential polyadenylation signals in the 3' untranslated region. These features suggest the possibility that the mutant form of RNase H1 could have been translated into a viable enzyme. Some support for this suggestion can be seen in extant alternate RNASEH1 transcripts.

At the present time NCBI ACCEMBLY lists 12 alternately spliced transcripts of human RNASEH1. Many of these transcripts are substantially altered throughout their length. Two, designated RNASEH1a and RNASEH1c, display altered carboxy-terminal ends only. Transcript 1c is spliced at the exon 7/intron 7 junction to a frame-shifted 7 amino acid sequence contained within exon 8, while transcript 1a replaces the final six amino acids of exon 8 with an 18 amino acid sequence located in a ninth exon 13.5 kb further downstream. In addition, transcript 1a utilizes a GC-AG splice, known to be the major alternative splice sequence (Thanaraj and Clark 2001). While the mutation preserved in RNASEH1P1 is also GC, the pseudogene sequence shows that it was not processed.

Similar to transcripts 1a and 1c, transcripts of the mutated allele of RNASEH1, if translated, would give rise to a protein identical to the wild type except at the carboxy-terminal end. Wild-type exon 8 encodes the 28 amino acid sequence MHVPGHSFIGNEADRLAREGAKQSED, whereas the splice junction mutant would encode the amino acid sequence ASILNVHLWLL. Transcript 1c ends with AMKBLTD, while transcript 1a is MHVPGHSFIGNEADRLAREGHHYKLL1YGCFVKEEIT. Functional studies of human RNase H1 by Wu et al. (1999, 2000) indicate that amino acids critical for both substrate binding and catalytic activity are encoded in exons 4–7. These residues are present in transcripts 1a and 1c and would be in the mutant RNase H1 as well. Thus the function of the various real or potential carboxy-terminal ends in RNase H1 remains unknown.

We conclude that a wild-type and a mutant RNASEH1 were present in the ape genome 19 million years ago, but at the present time, no evidence exists that the mutant allele remains in extant hominoid genomes. Indeed, there is only the retropseudogene RNASEH1P1 to document that it ever existed. The presence of an in-frame stop codon and several potential polyadenylation signals following the splice junction mutation, as well as evidence of extant alternately spliced human RNase H1 involving substituted carboxy termini, suggest the possibility that the mutant enzyme could have been viable. However, in the absence of a functional analysis of the putative mutant protein, this final point remains only a possibility.

REFERENCES


Received May 8, 2003
Accepted January 15, 2004

Corresponding Editor: Stephen J. O’Brien