RESEARCH ARTICLE

Recent integrations of mammalian *Hmg* retropseudogenes

EILLEN TECLE, LEANN ZIELINSKI and DAVID H. KASS*

Department of Biology, Eastern Michigan University, Ypsilanti, MI 48197, USA

Abstract

We propose that select retropseudogenes of the high mobility group nonhistone chromosomal protein genes have recently integrated into mammalian genomes on the basis of the high sequence identity of the copies to the cDNA sequences derived from the original genes. These include the *Hmg1* gene family in mice and the *Hmgn2* family in humans. We investigated orthologous loci of several strains and species of *Mus* for presence or absence of apparently young *Hmg1* retropseudogenes. Three of four analysed elements were specific to *Mus musculus*, two of which were not fixed, indicative of recent evolutionary origins. Additionally, we datamined a presumptive subfamily (*Hmgz*) of mouse *Hmg1*, but only identified one true element in the GenBank database, which is not consistent with a separate subfamily status. Two of four analysed *Hmgn2* retropseudogenes were specific for the human genome, whereas a third was identified in human, chimpanzee and gorilla genomes, and a fourth additionally found in orangutan but absent in African green monkey. Flanking target-site duplications were consistent with LINE integration sites supporting LINE machinery for their mechanism of amplification. The human *Hmgn2* retropseudogenes were full length, whereas the mouse *Hmg1* elements were either full length or 3′-truncated at specific positions, most plausibly the result of use of alternative polyadenylation sites. The nature of their recent amplification success in relation to other retropseudogenes is unclear, although availability of a large number of transcripts during gametogenesis may be a reason. It is apparent that retropseudogenes continue to shape mammalian genomes, and may provide insight into the process of retrotransposition, as well as offer potential use as phylogenetic markers.

[Tecle E., Zielinski L. and Kass D. H. 2006 Recent integrations of mammalian *Hmg* retropseudogenes. *J. Genet.*, **85**, 179–185]

Introduction

Retropseudogenes are a class of nonfunctional duplicated genes derived via an RNA intermediate. They are defined by sequence similarity to a known gene, by the lack of introns, by having polyadenylated tails, and by being flanked by small direct repeats that are products of the integration events. The human genome is estimated to contain 10,000 – 33,000 retropseudogenes (Goncalves *et al*. 2000; Harrison *et al*. 2002), whereas the mouse genome has been estimated to contain over 7000 (Waterston *et al*. 2002). In contrast there exist approximately 2.8 million retrotransposon sequences in mouse (Waterston *et al*. 2002) and human genomes (Lander *et al*. 2001), with several identified as recent species-specific integrations. Therefore it is known that these elements, including SINEs (short interspersed nucleotide elements) and LINEs (long interspersed nucleotide elements), continue to shape mammalian genomes, although retropseudogenes have been studied to a much lesser extent with regard to their genomic impact. Devor and Moffat-Wilson (2003) analysed five orthologous retropseudogenes among various primates and identified a relatively recent integration present only within human, chimpanzee and gorilla genomes. Sage *et al*. (1997) identified recent retrotranspositions of the *Sycp1* meiotic gene within the *Mus* genus. In this study we identified evolutionarily recent integrations of retropseudogenes derived from nonhistone chromosomal high mobility group (HMG) protein genes in both primates and rodents.

The HMG proteins consist of a few families of genes with a relatively large number of retrotransposed pseudogene copies (RPCs) (Strichman-Almashanu *et al*. 2003). These authors identified and analysed 219 *Hmg* RPCs in humans. These RPCs exhibited the hallmarks of retrotransposition, including A-tails and target-site duplications. Additionally, the TT/AAAA consensus was the preferential integration

*For correspondence. E-mail: dkass@emich.edu.

Keywords. retropseudogene; retrotransposition; transposable element; nonhistone chromosomal high mobility group protein.

site, as implicated also for LINEs, human *Alu*, and rodent ID elements (Jurka 1997). The *Hmgn2* (*Hmg17*) gene family appeared to be the most prolific source of HMG RPCs in human (Srikantha *et al*. 1987; Strichman-Almashanu *et al*. 2003), and exhibits the highest sequence identity to the mRNA sequence (Strichman-Almashanu *et al*. 2003). Therefore, we chose to study this group of RPCs for recent activity in primate genomes. Additionally, to better appreciate the recent contributions of retropseudogenes within mammalian genomes we analysed mouse *Hmg1* RPCs, which have been shown to have a very high degree of sequence similarity to the corresponding mouse cDNA sequence (Ferrari *et al*. 1994), indicative of recent evolutionary origins. Ferrari *et al*. (1994) also identified a subset of related RPCs suggesting a source gene (*Hmgz*) potentially derived from *Hmg1*. Therefore we included a datamining investigation of these elements in this study.

By analysing evolutionarily young *Hmg* RPCs we propose that retropseudogenes continue to shape mammalian genomes, that these elements may have more than one source gene, and that there is uncharted potential use as phylogenetic tools as their presence demonstrates identity by descent. We also show that recent integrations of retropseudogenes can be identified by investigating elements of high sequence identity in relation to the source (mRNA). This includes presence/absence of the retropseudogene within a species and among closely related species. Additionally, we analyse the target-site duplications (TSDs) to potentially further verify their use of LINEs for retrotransposition.

Materials and methods

DNA samples

DNA from the following *Mus* species and strains were purchased from The Jackson Laboratory: *M. musculus musculus* (SKIVE/Ei), *M. m. molossinus* (MOLF/Ei), *M. m. castaneus* (CASA/Rk), *M. hortulanus* (PANCEVO/Ei) and *M. spretus* (SPRET/Ei). DNA from *M. musculus* strains ICR, BALB/c, and C57BL/6 was kindly provided by Dr Astrid Engel from Tulane University Medical Center. DNA samples from chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), African green monkey (*Cercopithecus aethiops*) and humans were available from previous studies (Kass *et al*. 1995). Additional human DNA samples were kindly provided by Dr Mark Batzer of Louisiana State University.

Identification of loci containing young HMG retropseudogenes

Highly conserved mouse *Hmg1*-related genomic sequences were selected by a bias screening of the GenBank database using a BLAST search with a 2231-nucleotide *Hmg1* cDNA query sequence (Z11997; Yotov and St-Arnaud 1992). Human *Hmgn2*-related genomic sequences were selected with the 1198-bp cDNA query sequence NM_005517. BLAST searches were also performed with *Hmgz* sequences (HMG1-R-135 (620099) and HMG1-R-87 (620107)).

PCR, cloning and DNA sequencing

Sequences were analysed for integration sites and target-site duplications, as well as flanking sequences. Primers were designed that flank the retropseudogene sequences to interrogate orthologous loci of different species for presence or absence of the integrated element. MacVector 7.0 (Oxford Molecular Group) was used to determine effectiveness of the primers for PCR amplification, and BLAST searches were performed to verify lack of homology of the primers to repetitive DNA. Integration times were estimated using the Jukes–Cantor correction for multiple mutations at a single site (Jukes and Cantor 1969) and a neutral mutation rate of 4.6×10^{-9} (Li *et al*. 1981), then contrasted with the divergence times of species that share the integrations in relation to those lacking it. Representative presence and absence PCR fragments were either cloned into the pGEMTEasy cloning vector (Promega) or directly sequenced using the ABI Prism 310 automated sequencer to verify amplification of the correct locus.

PCR amplifications were performed in 25-μl volumes containing 1× Failsafe buffer (Epicentre), which includes the dNTPs, 0.20 mM of each primer (table 1), 1 U Failsafe DNA polymerase (Epicentre) and 50 ng DNA. Reactions were performed using an MJ thermal cycler with various annealing temperatures (NN) for the different loci (table 1) under the following conditions: 94°C for 2 min, 1×; 94°C for 1 min, NN°C for 1 min, 72°C for 2 min, 30×; 72°C for 5 min. Amplification products were analysed by 1% agarose gel electrophoresis.

Results

Identification and sequence attributes of 'young' Hmg retropseudogenes

Presumably young *Hmg* retropseudogenes were ascertained from the best sequence matches to the query sequences of *Hmgn2* mRNA (NM_005517), *Hmg1* mRNA (NM_010439), and *Hmgz* genomic sequences (HM1-R-135 (620099) and HMG1-R-159 (620106)) using BLAST searches. The analysed retropseudogenes (accession numbers provided in tables 1 and 2) ranged from 95% to 98% identity to the cDNA query sequences (table 2), indicative of fairly recent origins.

Although sequences of the young retropseudogenes were conserved, the sizes varied greatly among mouse *Hmg* elements (note size variations between presence/absence forms in table 1). All the human *Hmgn2* retropseudogenes and the mouse MM*Hmg*-A were full-length integrations, whereas retropseudogenes MM*Hmg*-B, MM*Hmg*-C, MM*Hmg*-D, MM*Hmg*-E and MM*Hmg*-F were truncated at their 3′ ends. MM*Hmg*-B, MM*Hmg*-D and MM*Hmg*-F were truncated at nearly equivalent positions. MM*Hmg*-C and MM*Hmg*-E were truncated further upstream, differing by six nucleotides.

Table 1. *Hmg* retropseudogene loci analysed in this study. Locus names were assigned arbitrarily with MM designating mouse sequences and HS designating human sequences. GenBank database accession numbers for the sequences are indicated, as well as forward (F) and reverse (R) primers and annealing conditions for PCR. Expected sizes of PCR products for the presence and absence of the retropseudogenes are also indicated, the difference indicates the size of the retropseudogene integration.

Locus	Accession number	Primers (5′–3′)	Annealing conditions	Expected +insert	Expected −insert
MM*Hmg*-A	25896322	F: caaagcaaaactaagctcccatcc	55°C	2507	283
		R: gacatgctaacaaagggttgcc			
MM*Hmg*-B	47076230	F: caatagtgtctgggttgttggc	55°C	1645	375
		R: ggattggggtgggtgtttatg			
MM*Hmg*-C	20068524	F: cactgattcctaaagtgtgagtgcg	60°C	1225	355
		R: gacattcaatctgctctgtc			
MM*Hmg*-E	25013379	F: acacacacgagcacacacatgc	55°C	1088	216
		R: tcagaaggtaagagccctggttg			
HS*Hmg*-A	28951175	F: ccacttaatttcgcagtctgtcg	58°C	1607	373
		R: gaaaagctggcacacactgaaac			
HS*Hmg*-B	15375165	F: ccatctgtgtgacctcgcag	55°C	1841	577
		R: taatagtggatctaatcaatga			
HS*Hmg*-C	19424606	F: gagcccaggagctcaggttc	55°C	1519	105
		R: aagcactgcagcctcaatac			
HS*Hmg*-D	12657326	F: aaacagagagattttgtcatactaagtca	55°C	1539	365
		R: ttacctaacacataaaagat			

Table 2. Analysed human (HS) and mouse (MM) *Hmg* retropseudogenes with high sequence homology to the respective mRNA query sequence, with chromosomal locations and estimated integrations times. Accession numbers for MM*Hmg*-D and MM*Hmg*-F are 25189035 and 24962919, respectively; other accession numbers are provided in table 1.

Reference	cDNA divergence	Chromosome	Integration (years)
HS*Hmgn2*-A	98.4%	16	1.7×10^6
HS*Hmgn2*-B	97.5%	5	2.7×10^6
HS*Hmgn2*-C	97.5%	15	2.7×10^6
HS*Hmgn2*-D	97%	14	3.3×10^6
MM*Hmg*-A	95%	11	5.6×10^6
MM*Hmg*-B	98%	X	2.2×10^6
MM*Hmg*-C	98%	11	2.2×10^6
MM*Hmg*-D	97%	19	3.3×10^6
MM*Hmg*-E	97%	3	3.3×10^6
MM*Hmg*-F	98%	16	2.2×10^6

These results contrast with LINEs, which are routinely truncated at various 5′ ends. However, the most plausible explanation would be use of alternative polyA signals within the 3′-UTR (each retropseudogene contains the entire coding sequence). MM*Hmg*-B, MM*Hmg*-D and MM*Hmg*-F extend seven nucleotides downstream from an AGTAAA sequence which has been implicated as an alternative splice site (Beaudoing *et al*. 2000; Tian *et al*. 2005). MM*Hmg*-C and MM*Hmg*-E extend 13 nucleotides downstream from an AAGAAA site which has also been implicated as a potential alternative polyA site (Anand *et al*. 1997; Beaudoing *et al*. 2000; Tian *et al*. 2005) although this sequence has also been shown to inactivate polyadenylation potential in relation to the AAUAAA standard (Wilusz and Shenk 1988). These results indicate that much of the 3′-UTR is not critical for 'hijacking' LINE machinery for retrotransposition.

Analysis of Hmg integration sites

The consensus LINE integration site is TTAAAA (Jurka 1997), although other variants are common, including TTAAGA, TTAGAA, TTGAAA, TTAAAG and CTAAAA (Jurka 1997). Therefore we analysed the integration sites of the *Hmg* retropseudogenes (figure 1) to assess their potential use of LINEs for the retrotransposition process. Of the four analysed human elements HS*Hmgn2*-A had an integration site match exactly to TTAAAA, HS*Hmgn2*-D was an exact match to one of the five common variants, HS*Hmgn2*-B had one nucleotide variant from one of the

common variants, and HS*Hmgn2*-C apparently integrated into the A-tail of an older *Alu* element (which corresponds to a two-nucleotide variant from the consensus integration site). This suggests that the HS*Hmgn2* retropseudogenes have utilized LINE machinery for retrotransposition, and the integration of retrotransposons potentially provides new integration sites. Additionally, Devor and Moffat-Wilson (2003)

Figure 1. Analysis of integration sites of human (HS) and mouse (MM) *Hmg* retropseudogenes. Bold letters indicate potential LINE-based endonuclease recognition sites. Uppercase letters correspond to target-site duplications. Italicized letters correspond to the presumptive non-L1 endonclease recognition site of AAG.

analysed retropseudogene integration sites and found that 33 out of 60 had an AnG (primarily AAG) integration site. Two of the human retropseudogene integrations have an AAG, which (although a small sample size) corresponds to the roughly 50% determined by Devor and Moffat-Wilson (2003). Of the six mouse *Hmg1* retropseudogene integration sites analysed MM*Hmg*-E had an exact match to the consensus, MM*Hmg*-F and MM*Hmg*-D matched common variants, MM*Hmg*-B had a single nucleotide variant in relation to the consensus, MM*Hmg*-A had a single nucleotide difference in relation to one of the common variants, and MM*Hmg*-C varied from the consensus by two nucleotides. The overall closeness of sequence identity to the LINE integration sites additionally supports use of LINE machinery by the mouse retropseudogenes. Additionally, three of the six had the AAG site, again (albeit a small sample size) corresponding to the roughly 50% presented by Devor and Moffat-Wilson (2003). However, since AAG is present in two of the common LINE integration variants, it is difficult to assess importance of this sequence for retrotransposition in the mouse, although it was proposed to be a site for a non-L1 DNA endonuclease (Devor and Moffat-Wilson 2003). The conservation between upstream and downstream TSDs, for the six mouse *Hmg* integrations, provides further support to the proposal of recent integration events. Only one TSD variant was found among the human integrants, although it was difficult to assess the HS*Hmgn2*-C integration within the *Alu* A-tail.

Analysis of orthologous loci

Orthologous loci for four *Hmgn2* retropseudogenes were analysed among various primates to further assess integration times (table 3). Two of the retropseudogenes (HS*Hmgn2*-A and HS*Hmgn2*-B) were only present within the human genome, indicating an integration that postdates the split between humans and chimpanzees (figure 2). HS*Hmgn2*-C was present in human, chimp and gorilla, whereas HS*Hmgn2*-D was present in all the hominids including orangutan. The estimated integration times of the four analysed retropseudogenes based on hominid phylogeny are shown in figure 2. The estimated times for HS*Hmgn2*-A and HS*Hmgn2*-B are consistent with the calculation in table 2. However, HS*Hmgn2*-C and HS*Hmgn2*-D are apparently both older than originally calculated. The correspondence of 'young' *Hmg* retropseudogenes with known phylogeny of hominids demonstrates potential use as a phylogenetic tool.

Recent integrations may not be fixed within a species yielding presence/absence variants as observed for both *Alu* (Batzer and Deininger 2002) and LINE (Sheen *et al*. 2000) retrotransposons in humans. Since the HS*Hmgn2*-A and HS*Hmgn2*-B retropseudogenes were human specific, we investigated several samples (80 and 59, respectively) from four ethnic groups (African American,

Table 3. Analysis of orthologous loci among humans and nonhuman primates. Loci refer to designated *Hmg* retropseudogenes (see table 1). A + indicates the presence of the retropseudogene integrations, and a – refers to the absence among the various species. A blank space means either no product obtained or not tested.

Locus	Human	Chimpanzee	Gorilla	Orangutan	African green monkey
HS*Hmg*-A	+	–	–	–	–
HS*Hmg*-B	+	–			
HS*Hmg*-C	+	+	+	–	
HS*Hmg*-D	+	+	+	+	–

Table 4. Interrogation of various loci for the presence (+) and absence (−) forms of *Hmg* retropseudogene loci among strains and subspecies of *Mus musculus*, and related *Mus* species. These include *Mus musculus musculus* strains BALB/c. C57BL/6 (C57), ICR Swiss (ICR), *M. molossinus*, (Mmm), *M. m. castaneus* (Mmc), *M. m. domesticus* (Mmd), *M. spretus* (Ms) and *M. hortulanus* (Mh). The *Hmg* loci accession numbers are found in table 1.

Locus	Balb/c	C57	ICR	Mmm	Mmc	Mmd	Ms	Mh
MM*Hmg*-A	−	+	+	−	+	−	−	−
MM*Hmg*-B	+	+	+	+	+	+	−	−
MM*Hmg*-C	−	+	+/−	+	+	+	−	−
MM*Hmg*-E	+	+	+	+	+	+	+	+

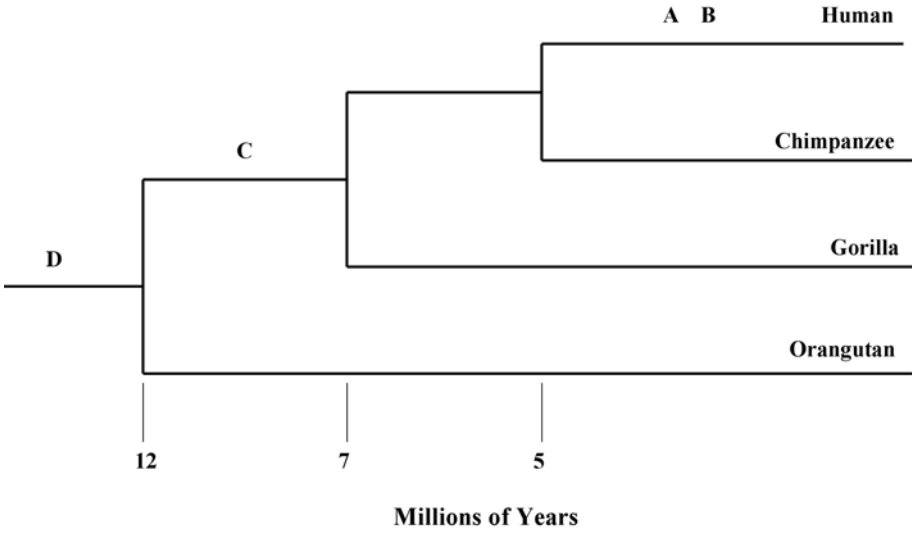

Figure 2. Estimated integration times of the four analysed retropseudogenes (HS*Hmg*-A, HS*Hmg*-B, HS*Hmg*-C, HS*Hmg*-D) based on Great Ape phylogeny.

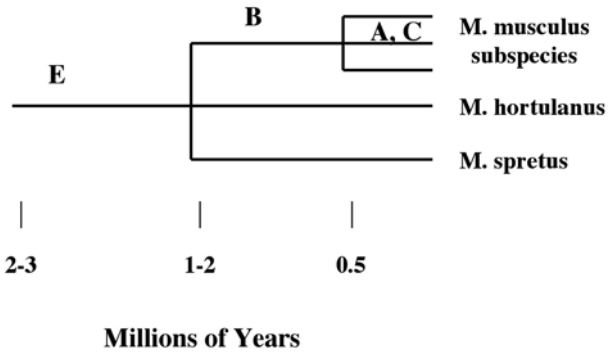

Figure 3. Estimated integration times of four analysed retropseudogenes (MM*Hmg*-A, MM*Hmg*-B, MM*Hmg*-C, MM*Hmg*-E) based on mouse phylogeny.

European Caucasian, Asian, South American) to identify a potential absence variant. All the analysed samples were homozygous for the presence form, suggesting that both of these retropseudogenes are fixed in the human genome.

Orthologous loci for four of the *Hmg1* retropseudogenes in *Mus* were analysed among several species, as well as subspecies and strains of *M. musculus* (table 4). Three of the four were present only within *M. musculus* (absent in the related species *M. spretus* and *M. hortulanus*). Additionally, two of the three *M. musculus*-specific retropseudogenes were not fixed in *M. musculus* as the absence form was also identified. The MM*Hmg1*-A retropseudogene was absent in the BALB/c strain of mouse and *M. m. molossinus*. The MM*Hmg*-C retropseudogene was also absent in BALB/c whereas the outbred ICR Swiss mouse was polymorphic for the presence and absence forms. MM*Hmg*-E was observed in all the analysed *Mus* species, suggesting an integration time of at least 2–3 million years (figure 3). The *M. musculus*-specific integrations suggest an age within the past 1–2 million years and the lack of fixation for two of these elements suggests even more recent events. Therefore retropseudogenes continue to shape the mouse genome.

Assessment of additional source genes

To assess whether the retropseudogenes were propagated from a single source gene we analysed the translated sequences for common variants that are unique from the

amino acid sequence derived from the cDNA. *Hmgn2*-A and *Hmgn2*-B had in common one unique amino acid variant, and *Hmgn2*-A and *Hmgn2*-C also had in common a unique variant. It is possible that these could have occurred by independent mutations, as additional common variants from the cDNA would promote a more convincing argument for multiple source genes. Analysis of the six translated mouse *Hmg1* retropseudogenes demonstrated no common amino acid variants from the translated cDNA *Hmg1* sequence, indicative of this being the lone source. Additionally, we analysed the nucleotide sequences from the six retropseudogenes and although we identified a few common variants from the cDNA sequence (*Hmg*-A and *Hmg*-E had two, *Hmg*-C and *Hmg*-E one, *Hmg*-A and *Hmg*-C one, *Hmg*-B and *Hmg*-C two, *Hmg*-E and *Hmg*-F one, *Hmg*-A and *Hmg*-B one) there was no consistency of two retropseudogenes having several variants in common. This is consistent with only a single source gene available for the generation of the mouse *Hmg1* retropseudogenes.

A subfamily of mouse *Hmg* retropseudogenes, referred to as *Hmgz*, was proposed by Ferrari *et al*. (1994) on the basis of the isolation of five related sequences from a mouse genomic DNA library. When translated to amino acids these elements had the same pattern of 10 deviations (at scattered positions) from the cDNA sequence. Additionally, these elements demonstrated higher sequence identity to each other than to the four *Hmg1* elements isolated by this group. These authors proposed that the source gene (referred to as *Hmgz*) is either derived from or shares a common ancestor with the *Hmg1* gene. We therefore compared the amino acid positions that are diagnostic for *Hmgz* and found none of these in the *Hmg1* retropseudogenes we analysed. We then screened the GenBank database for *Hmgz* sequences using the nucleotide sequences from two of the five *Hmgz* retropseudogenes. The BLAST search identified the five sequences (nucleotide length of 190) of Ferrari *et al*. (1994) but these contained no flanking sequences for further analysis. Only one genomic clone (*Hmgz1*: accession number 2765280) was found to have greater than 91% sequence identity to the *Hmgz* query sequence. We therefore analysed the translated amino acid sequence of this retropseudogene, and it had nine of the *Hmgz* amino acid diagnostics, potentially implicating this as a true z element. Two of the next highest scoring BLAST-identified elements had only two (accession number: 28850184) or one (accession number: 45120417) *Hmgz* diagnostic, suggesting either intermediates or chance mutations, and lacking support for confirmation as true *Hmgz* retropseudogenes. The paucity of identifiable *Hmgz* elements complicates promoting the status of *Hmgz* as a genuine retropseudogene subfamily. Additionally, we analysed variations in codon positions between the *Hmgz1* consensus and *Hmg1* cDNA and found five second-codon position variations and three third-position variations. The greater non-third-position variants suggest that the presumptive *Hmgz* source would not be a protein-encoding gene. This is further supported by the failure to identify a presumptive source gene using a BLASTp search queried with the translated Z sequence.

Discussion

We analysed *Hmg* RPCs from both humans and mice and found these to be recently active (based on species specificity), apparently to be utilizing LINE machinery, and to be derived from single source genes. Additionally, we provide a mechanism for identifying recent integrations of retropseudogenes that provide potential use as a phylogenetic tool. These markers provide advantages such as ease of use, identity by descent (as opposed to state), including the knowledge that the absence form is the ancestral allele, as well as the absence of homoplasy. Retrotransposons have been successfully used as genetic markers, although there is a greater background of older elements to sift through (Roy *et al*. 1999) and the numerous integration events of retrotransposons seemingly spawn rare instances of parallel integrations (Kass *et al*. 2000).

It is unclear what features contribute to the relatively successful propagation of *Hmg* retropseudogenes. On the basis of our findings, we can rule out much of the 3′-UTR as critical for utilizing LINE machinery. Additionally, the finding that there is apparently no 'trans' preference for retrotransposition of polymerase-II-transcribed sequences by LINEs (Esnault *et al*. 2000), suggests that availability of the transcripts may be of greater consequence. This is consistent with the findings of Sage *et al*. (1997). They reported recent retrotranspositions of *Sycp1* in *Mus*, and proposed that since retrotransposed meiotic genes are immediately inherited this lends support to providing an efficient mechanism of genome remodelling. Since HMG proteins are expressed during gametogenesis (Bucci *et al*. 1984), it can be reasoned that simply the availability of *Hmg* transcripts during gametogenesis can allow for occasional heritable retrotransposition within the mammalian genome. Currently we are analysing retrotransposition of *Hmg* sequences in a cell culture assay to further assess the importance of sequence for utilizing LINE machinery.

Acknowledgements

This work was supported by an NIH AREA award GM062828 to D.H.K., and an Eastern Michigan University Graduate Studies and Research Support Fund award.

References

Anand S., Batista F. D., Tkach T., Efremov D. G. and Burrone O. R. 1997 Multiple transcripts of the murine immunoglobulin epsilon membrane locus are generated by alternative splicing and differential usage of two polyadenylation sites. *Mol. Immunol.* **34**, 175–183.

Batzer M. A. and Deininger P. L. 2002 Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379.

Beaudoing E., Freier S., Wyatt J. R., Claverie J. M. and Gautheret D. 2000 Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001–1010.

Bucci L. R., Brock W. A., Goldknopf I. L. and Meistrich M. L. 1984 Characterization of high mobility group protein levels during spermatogenesis in the rat. *J. Biol. Chem.* **259**, 8840–8846.

Devor E. J. and Moffat-Wilson K. A. 2003 Molecular and temporal characteristics of human retropseudogenes. *Hum. Biol.* **75**, 661–672.

Esnault C., Maestre J. and Heidmann T. 2000 Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367.

Ferrari S., Ronfani L., Calogero S. and Bianchi M. E. 1994 The mouse gene coding for high mobility group 1 protein (HMG1). *J. Biol. Chem.* **269**, 28803–28808.

Goncalves I., Duret L. and Mouchiroud D. 2000 Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10**, 672–678.

Harrison P. M., Hegyi H., Balasubramanian S., Luscombe N. M., Bertone P., Echols N. *et al*. 2002 Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280.

Jukes T. H. and Cantor C. R. 1969 Evolution of protein molecules. In *Evolution of protein molecules* (ed. H. N. Munro), pp. 21–132. Academic Press, New York.

Jurka J. 1997 Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. USA* **94**, 1872–1877.

Kass D. H., Batzer M. A. and Deininger P. L. 1995 Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol. Cell. Biol.* **15**, 19–25.

Kass D. H., Raynor M. E. and Williams T. M. 2000 Evolutionary history of B1 retroposons in the genus *Mus*. *J. Mol. Evol.* **51**, 256–264.

Lander E. S., Linton L. M., Birren B., Nussbaum C., Zody M. C., Baldwin J. *et al*. 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Li W. H., Gojobori T. and Nei M. 1981 Pseudogenes as a paradigm of neutral evolution. *Nature* **292**, 237–239.

Roy A. M., Carroll M. L., Kass D. H., Nguyen S. V., Salem A. H., Batzer M. A. *et al*. 1999 Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**, 149–161.

Sage J., Yuan L., Martin L., Mattei M. G., Guenet J. L., Liu J. G. *et al*. 1997 The Sycp1 loci of the mouse genome: successive retropositions of a meiotic gene during the recent evolution of the genus. *Genomics* **44**, 118–126.

Sheen F. M., Sherry S. T., Risch G. M., Robichaux M., Nasidze I., Stoneking M. *et al*. 2000 Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res.* **10**, 1496–1508.

Srikantha T., Landsman D. and Bustin M. 1987 Retropseudogenes for human chromosomal protein HMG-17. *J. Mol. Biol.* **197**, 405–413.

Strichman-Almashanu L. Z., Bustin M. and Landsman D. 2003 Retroposed copies of the HMG genes: a window to genome dynamics. *Genome Res.* **13**, 800–812.

Tian B., Hu J., Zhang H. and Lutz C. S. 2005 A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucl. Acids Res*. **33**, 201–212.

Waterston R. H., Lindblad-Toh K., Birney E., Rogers J., Abril J. F., Agarwal P. *et al*. 2002 Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Wilusz J. and Shenk T. 1988 A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. *Cell* **52**, 221–228.

Yotov W. V. and St-Arnaud R. 1992 Nucleotide sequence of a mouse cDNA encoding the nonhistone chromosomal high mobility group protein-1 (HMG1). *Nucl. Acids Res.* **20**, 3516.

Received 7 December 2005