An Ancient Retrovirus-like Element Contains Hot Spots for SINE Insertion

Michael A. Cantrell, Brian J. Filanoski,[1] Angela R. Ingermann,[2] Katherine Olsson, Nicole DiLuglio, Zach Lister and Holly A. Wichman

Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844

Manuscript received August 10, 2000
Accepted for publication February 28, 2001

ABSTRACT

Vertebrate retrotransposons have been used extensively for phylogenetic analyses and studies of molecular evolution. Information can be obtained from specific inserts either by comparing sequence differences that have accumulated over time in orthologous copies of that insert or by determining the presence or absence of that specific element at a particular site. The presence of specific copies has been deemed to be an essentially homoplasy-free phylogenetic character because the probability of multiple independent insertions into any one site has been believed to be nil. *Mys* elements are a type of LTR-containing retrotransposon present in Sigmodontine rodents. In this study we have shown that one particular insert, *mys*-9, is an extremely old insert present in multiple species of the genus Peromyscus. We have found that different copies of this insert show a surprising range of sizes, due primarily to a continuing series of SINE (*s*hort *in*terspersed *e*lement) insertions into this locus. We have identified two hot spots for SINE insertion within *mys*-9 and at each hot spot have found that two independent SINE insertions have occurred at identical sites. These results have major repercussions for phylogenetic analyses based on SINE insertions, indicating the need for caution when one concludes that the existence of a SINE at a specific locus in multiple individuals is indicative of common ancestry. Although independent insertions at the same locus may be rare, SINE insertions are not homoplasy-free phylogenetic markers.

RETROTRANSPOSONS are transposable elements that produce additional copies for insertion into new genomic sites by reverse transcription of an RNA intermediate. Although some retrotransposon insertions alter gene expression, most of those maintained in mammalian genomes have presumably inserted into noncoding regions and are selectively neutral. Such insertions may therefore be ideal sequences both for estimating organismal phylogenies and for study of neutral evolution at the molecular level.

Vertebrate retrotransposons present at specific loci have been deemed to be essentially homoplasy-free phylogenetic characters because the probability of insertion occurring more than once at any single site has been presumed to be vanishingly low (Batzer and Deininger 1991; Batzer *et al.* 1994; Takahashi *et al.* 1998; Nikaido *et al.* 1999). The ancestral state at any site, absence of the insert, is also believed to be unambiguous because precise excision of these retrotransposons has not been seen. In particular, the presence or absence of SINE (*s*hort *in*terspersed *e*lement; see Deininger 1989 for a general review) insertions is easy to assay, and thus SINEs have found increasing use as phylogenetic markers at many taxonomic levels. Human-specific SINE (*Alu*) insertions that are not fixed within the species have been used to support the hypothesis of a recent African origin for modern humans (Batzer and Deininger 1991; Batzer *et al.* 1994). Examples of use of SINEs as markers at higher taxonomic levels also include studies of the relationships among salmonid fishes (Murata *et al.* 1993; Hamada *et al.* 1998) and work showing that hippopotamuses are the closest extant relatives to the whales (Shimamura *et al.* 1997; Nikaido *et al.* 1999).

The presence or absence of retrovirus-like insertions has also been used as a taxonomic marker. Although they are longer and thus not as easy to assay as SINE insertions, retrovirus-like elements have the advantage of a built-in molecular clock—the paired LTRs (*l*ong *t*erminal *r*epeats). Retrovirus-like elements replicate by the same mechanism as retroviruses, and their LTRs are expected to be identical at the time of insertion. After insertion the elements accumulate changes at the neutral rate, and the divergence between the paired LTRs can thus be used to estimate the relative time since insertion of that element. Previously, we examined three loci in white-footed mice (*Peromyscus leucopus*) from multiple geographic locations for presence or absence of the retrovirus-like retrotransposon, *mys*, and determined the relative time since insertion of each of those elements (Sawby and Wichman 1997). We found that the two loci containing the most recent insertions based on differences between the paired LTRs had a *mys* element

Corresponding author: Holly A. Wichman, Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844-3051. E-mail: hwichman@uidaho.edu

[1] *Present address:* Molecular Probes, Inc., P.O. Box 22010, Eugene, OR 97402-0469.

[2] *Present address:* Department of Pediatrics, NRC5, Oregon Health Sciences University, Portland, Oregon 97201.

in only 1 of the 28 mice examined, while the locus with the greatest difference between LTRs contained a *mys* element in 9 individuals limited to the northern part of the species range.

In this study we find that a *mys* insertion, *mys*-9, originally found in *P. leucopus* (Wichman *et al.* 1985), appears, on the basis of both a 20.3% uncorrected sequence difference between its LTRs and its presence in multiple species of Peromyscus, to be ancient. Phylogenetic analysis of 13 orthologous copies of this element is consistent with the accepted species phylogeny. We see a surprising range of *mys*-9 allele sizes at this locus caused by a large number of SINE insertions. Within this locus we find two incidents of independent, multiple SINE insertion events at identical sites. These results have major repercussions for phylogenetic analyses based on SINE insertions, indicating the need for caution before interpreting shared SINE insertions as incontrovertible evidence of common ancestry.

MATERIALS AND METHODS

Tissues and DNA: *P. leucopus* (Georgia: TK24940); *P. maniculatus* (Mexico: TK27653, Iowa: TK25398, California: TK13404, Maine: TK29798); *P. difficilus* (TK32541); *P. truei* (TK21858); and *P. crinitus* (TK26309) tissues were from The Museum, Texas Tech University. *P. leucopus* (Massachusetts: H408) tissue was from Harvard University. *P. maniculatus* (New Mexico: GK362) was from Texas A&M, and *P. leucopus* (Texas: 20-3143234 and Connecticut) was from Wesleyan University. DNA was prepared from tissue by the method of Longmire *et al.* (1988). *mys*-9 was initially shown to be contained within a clone isolated from a genomic library of a *P. leucopus* individual (Wichman *et al.* 1985). We sequenced 6 kb of this clone, including the entire *mys*-9 element and part of the single-copy flanking regions. We then designed PCR primers specific for the single-copy flanking sequences: M9-17 (5′ side of *mys*-9, CTCATTCCCAGAAACCTACATGCTAA) and M9-16 (3′ side of *mys*-9, ACTACAAAGATAAGGAGCCTAGCTGAGTG). PCR amplification was carried out with the extra long PCR kit (PE Applied Biosystems, Foster City, CA) using 30 pmol of each of these primers, 100 ng of genomic DNA, and 1.4 mm magnesium acetate in a total volume of 50 μl. Thermal cycling was performed in a GeneAmp PCR System 9600 (PE Applied Biosystems) with the following parameters: hot start (addition of DNA and polymerase at 65°) and initial denaturation for 2 min at 94°, followed by 20 cycles of 15 sec at 94°, 30 sec at 58°, 4 min at 70°, followed by 10 cycles with the same parameters but with the elongation step extended by an additional 15 sec at each succeeding cycle, followed by a final elongation for 10 min at 72°. The annealing temperature was reduced to 55° for species outside of the *P. leucopus* group/*P. maniculatus* group clade. Sizes of amplification products were determined on 0.6–0.7% agarose gels. Size standards were mixed with samples in many cases. All cloning, restriction analyses, and Southern hybridizations were done by standard techniques (Ausubel *et al.* 1989).

Sequencing and sequence analysis: Clones were either manually sequenced as previously described (Casavant *et al.* 1996) or were sequenced using a Licor 4000L or ABI-377 automated sequencer. Sequences were aligned and analyzed using GeneWorks (Intelligenetics, Mountain View, CA), and MegAlign 3.1.7 (DNAStar, Madison, WI). Alignments were refined by hand. All phylogenetic analyses were performed using PAUP*, version 4.0b4 (Swofford 2000). Corrected pairwise sequence differences (changes per 100 bp) were determined using the Hasegawa *et al.* (1985; HKY85) substitution model, which was found to be the most statistically defensible by hierarchical likelihood ratio tests (Sullivan *et al.* 1997; Sullivan and Swofford 1997). The phylogenetic analysis shown in Figure 1 was carried out using maximum likelihood under the HKY model of evolution (Hasegawa *et al.* 1985). The tree shown in Figure 2 was constructed using parsimony analysis and 1000 bootstrap replicates in a heuristic search.

RESULTS

***Mys*-9 is an ancient *mys* insert that contains multiple indels:** We initially characterized the *mys*-9 element from *P. leucopus* because differences between its restriction map and the restriction maps of other *mys* elements (Wichman *et al.* 1985) suggested that it represented a substantially different type of *mys* element. Sequence analysis of its left and right LTRs showed that the left edge of the left LTR had suffered either a large rearrangement or a deletion. Alignment of the remaining 209 bp of that LTR with the corresponding portion of the right LTR revealed a sequence difference between the two LTRs of 20.3% (excluding gaps), giving rise to a corrected pairwise distance of 24.8 changes/100 bp (Table 1). Because the two LTRs of a retrovirus-like element are identical at the time of insertion, each LTR should have diverged by approximately half that amount, or 12.4 changes/100 bp, from the original ancestral sequence since the insertion of *mys* at this locus. If the neutral mutation rate in Peromyscus is similar to the estimated rate of ~1 change/100 bp per million years for other rodents (She *et al.* 1990 and references therein), this would suggest that the element inserted into this locus roughly 12.4 million years ago.

We sequenced the entire *mys*-9 element (3444 bp) and compared its sequence to the previously characterized *mys*-1 element, whose LTRs differ at only 2 out of 344 bp. A surprising number of indels (insertions or deletions) were seen. Target site duplications at the borders of the four larger indels lead us to suggest that they are insertions into *mys*-9. These are a B1 SINE, a B2 SINE, an ID SINE, and an unknown insert of SINE size. Three additional indels >13 bp and of unknown origin were also seen.

***Mys*-9 is widely distributed:** Multiple animals were tested for presence of the *mys*-9 element by amplification with PCR primers specific for single-copy regions flanking *mys*-9. PCR amplification would give rise to a product of ~4 kb if the *mys*-9 element were present at this locus (a filled site) and would produce a product of ~0.3 kb if the *mys*-9 element were not present at the locus (an empty site). Initial examination of ~32 *P. leucopus* and four *P. maniculatus* individuals from across the range of those species showed the *mys*-9 locus to be occupied in every case in which a PCR product was detected, with

TABLE 1

Corrected pairwise differences between LTRs and between *mys*-9 alleles

	criY	criZ	dif	tru	manMX	manCA	manIA	manIA	manME	leuGA	leuMA	leuTX	leuCT
	3.78	3.78	3.50	3.48	3.85	8.1	4.10	3.75	3.93	3.70	3.82	3.94	3.75
criY3.78	*24.1*												
criZ3.78	0.58	*23.3*											
dif3.50	4.15	4.03	*25.5*										
tru3.48	4.15	4.03	1.14	*25.5*									
manMX3.85	5.42	5.29	4.09	4.57	*21.8*								
manCA8.1	5.45	5.32	4.22	4.58	1.26	*21.8*							
manIA4.10	5.58	5.45	4.35	4.71	2.32	2.44	*23.7*						
manIA3.75	5.45	5.32	4.47	4.83	0.57	1.37	2.68	*23.1*					
manME3.93	5.95	5.82	4.83	5.20	0.91	1.72	3.03	1.03	*21.8*				
leuGA3.70	5.60	5.47	4.34	4.85	2.88	3.02	3.12	3.26	3.63	*25.2*			
leuMA3.82	5.05	4.92	3.97	4.46	2.31	2.68	2.92	2.67	3.03	0.81	*25.7*		
leuTX3.94	5.16	5.03	4.08	4.57	2.42	2.79	3.03	2.79	3.14	0.93	0.34	*25.1*	
leuCT3.75	5.15	5.02	4.07	4.56	2.42	2.78	3.03	2.78	3.14	0.93	0.34	0.45	*24.8*

The numbers in italics on the diagonal are the corrected pairwise sequence differences (changes per 100 bp) between the common parts of the left and right LTRs (209-bp alignment). All other numbers are the corrected pairwise differences between alleles, using assembled sequences that contain the flanks of each allele out to the PCR primer binding sites plus each of the LTRs (903-bp alignment). The allele names shown in the top row give the species from which the allele was obtained (first three letters), a designation in capital letters of either the allele (Y or Z) or the location from which the mouse was obtained (MX, Mexico; CA, California; IA, Iowa; ME, Maine; GA, Georgia; MA, Massachusetts; TX, Texas; CT, Connecticut), and the size of the amplified allele in kilobases (*e.g.*, 3.78). The criY3.78 and criZ3.78 alleles came from the same individual, as did the manIA4.10 and manIA3.75 alleles. GenBank accession nos. for the sequences in the order shown in this table are AY017268–AY017293. There are two accession numbers assigned to each allele, the first for the left flank and left LTR, the second for the right LTR and right flank.

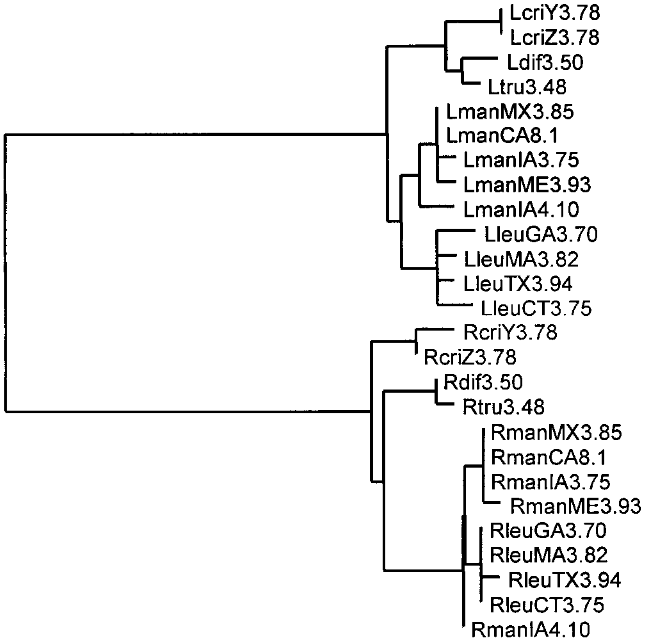

FIGURE 1.—Phylogenetic analysis of LTRs from *mys*-9 alleles. L designates the left LTR of each indicated allele, and R designates the right LTR. Analysis was based on the 209-bp region common to the left and right LTRs of each element. The phylogram is drawn with midpoint rooting.

no examples of empty sites. The *mys*-9 locus was also found to be filled in single specimens of *P. difficilis, P. truei,* and *P. crinitus*. Attempts to amplify the *mys*-9 locus in more distantly related species have produced no PCR products, possibly due to divergence of the primer binding sites. Nonetheless, amplification in the above species suggests that the insertion is ancient since it appears to predate the divergence of those species. These results extend our previous observations. We now see a distinct correlation between LTR divergence and increased distribution among four *mys* elements. *Mys*-9 appears to be widely distributed in Peromyscus and fixed in at least *P. leucopus* and *P. maniculatus,* while *mys*-6, which shows a divergence of 4.8 changes/100 bp between its LTRs, is found in many *P. maniculatus* and *P. leucopus* individuals, but is not fixed. *mys*-7 and *mys*-1, respectively, show divergences of 1.8 and 0.6 changes/100 bp between their LTRs and have been found only in single individuals (Sawby and Wichman 1997).

The *mys*-9 locus shows a wide range of allele sizes: In the process of scoring different animals for presence of the *mys*-9 element, we found the site to be highly polymorphic with respect to size. Elements are commonly seen with sizes varying from 3.5 to 4.1 kb, but one element has a size of 8.1 kb. Many individuals are also found to exhibit two element sizes, suggesting heterozygosity with respect to allele size at this locus. A Southern blot of genomic DNA from a number of individuals probed with single-copy DNA immediately adjacent to the *mys*-9 insert showed that those size variants occurred at a single locus and were not a PCR artifact or the result of gene duplication. This range of allele sizes raised a number of questions. What accounts for the variation in allele size? How are the individual alleles related to each other? Do specific size classes represent more closely related alleles, or has the same allele size evolved multiple times?

Relationships between *mys*-9 alleles are clarified: Thirteen alleles were selected for further analysis. These alleles were chosen to represent all species from which we amplified filled sites, and multiple alleles were chosen from *P. leucopus* and *P. maniculatus* to represent a wide geographic and allele size range. The LTRs and flanking regions were sequenced for each allele. Alignment of these regions showed that all of the alleles contained the rearrangement of the left LTR described above. Table 1 is a distance matrix of all 13 *mys*-9 alleles. Corrected pairwise distances between the common regions of the left and right LTRs of each allele (209 bp) are shown in italics on the diagonal. Corrected pairwise distances between alleles based on the sequence of both LTRs and the single-copy flanking region (903 bp) are shown below the diagonal.

Table 1 shows that divergence between alleles ranges from 0.34 to 5.8 changes/100 bp, which suggests that some shared a common ancestor as recently as 160,000 years ago and the most divergent as long as 2.9 million years ago. As expected, divergence tends to be low within species and greatest between the most divergent species. However, some *P. maniculatus* alleles are as divergent from each other as they are from some *P. leucopus* alleles, while the *P. truei* and *P. difficilus* alleles are more similar to each other than are most of the *P. maniculatus* alleles. Interestingly, the two *P. maniculatus* Iowa alleles, taken from the same mouse, are the most divergent of the *P. maniculatus* alleles. The divergence between LTRs of each element, which shows an average corrected distance of 24 changes/100 bp, suggests that *mys*-9 inserted about 12 million years ago, indicating that the locus itself is much older than the alleles we have examined and is likely to be present in even more distant species. The relative age of the alleles compared to the locus was confirmed by maximum likelihood analysis of the 209 bp of common sequence between the left and right LTRs of each element (Figure 1). The long branch separating the left and right LTRs confirms the age of the locus, while the tight clustering of each LTR is generally consistent with the recent divergence of the alleles. There is not complete concordance between the phylogenies of the left and right LTRs, but this is not surprising since the region examined is quite small.

The relationships among the alleles were further examined by phylogenetic analysis that included both LTRs and the flanking regions (903 bp). A consensus parsimony tree from 1000 bootstrap replicates is shown in Figure 2. Bootstrap values are shown below each

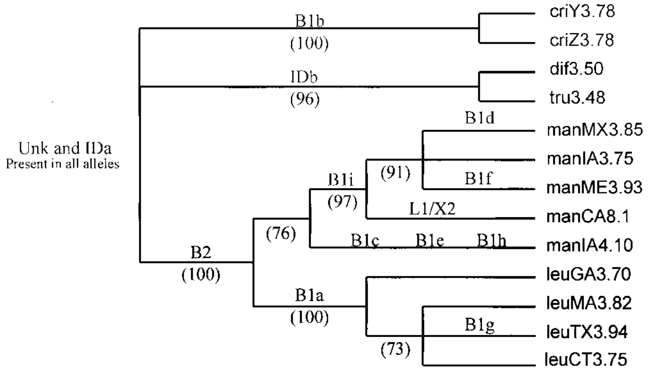

FIGURE 2.—Unrooted consensus tree derived from total sequence set overlaid with SINE insertions. Parsimony analysis was based on 903 bp that included both LTRs and flanking regions for each element. Nodes with >70% support are shown. The bootstrap support is shown below the node and insertions are shown above the node.

branch. This analysis confirms the within-species clustering of alleles. Furthermore, the allele tree agrees with the well-corroborated relationships among these species in that *P. maniculatus* and *P. leucopus* form a well-supported clade (*e.g.*, CARLETON 1989).

Most of allele size variation can be explained by insertion of a large number of SINEs into *mys*-9: We further dissected each of the 13 cloned *mys*-9 alleles by a combination of restriction mapping, hybridization with SINE probes, and sequencing of selected regions. We were surprised to find that these alleles contain a minimum of 14 insertions >100 bp, including nine B1 elements, two ID elements, one B2 element, an insert of unknown origin the size of a SINE, and a 4-kb insert containing the 5′ end of a LINE-1 element followed by nonrepetitive DNA. The 4-kb insert and at least one copy of each of the SINE inserts were sequenced. We overlaid these insertions onto the phylogenetic tree shown in Figure 2 to determine if they could be placed in a manner that would explain most of the size variation seen. With one exception (B1a/B1i, discussed and resolved below), each of the insertions could be placed on a branch such that it was present in all alleles derived from that branch. This large number of insertion events explains the majority of the size variation seen in the *mys*-9 alleles and shows that size of the *mys*-9 locus would not be a useful marker for phylogenetic studies. Congruence of the SINE insert data with both the observed allele sizes and the tree based on sequence data further suggests that none of the SINEs found here has been deleted in any of the *mys*-9 alleles.

The minimum and maximum ages of a number of the SINEs shown in Figure 2 can be estimated using the sequence differences shown in Table 1, the position of the element in the tree in Figure 2, and the assumption of a mutation rate of 1 change/100 bp per million years. For example, all of the *P. leucopus* alleles have been shown to contain the B1a insert. The average difference between the leuGA3.70 allele and the other *P. leucopus* alleles, which form a polytomy, is 0.89 changes/100 bp. These alleles have therefore diverged from a common ancestor by 0.44 changes/100 bp, indicating that the B1a element probably inserted more than 0.44 million years ago. A probable maximum age of 1.4 million years can be assigned to the B1a insertion because the average difference between any *P. leucopus* alleles and any *P. maniculatus* alleles (all contained in the clade whose branch immediately predates the B1a insert) is 2.87 changes/100 bp.

Two separate regions show SINE inserts that have used identical target DNA nick sites: During Southern blot analysis it appeared that a number of the *P. leucopus* and *P. maniculatus* alleles were phylogenetically united by a single B1 insertion event (B1a) because they all appeared to contain the same left flank sequence and B1 sequence, with only minor variations attributable to random mutation after insertion. However, phylogenetic analysis of the LTRs and flanking regions suggested that *mys*-9 elements containing this B1 insert do not form a monophyletic group. Further sequencing of these SINEs and their flanks in four of the *P. maniculatus* alleles and two of the *P. leucopus* alleles showed that the *P. maniculatus* B1 insertion is distinct from the B1 insertion found in the *P. leucopus* alleles; this new insertion is designated B1i. This SINE inserted by 3′ nicking at exactly the same nucleotide but by use of a 5′ nick site 3–6 bp upstream of the B1a nick site (see Figure 3). The ambiguity of 3–6 bp is an unresolvable consequence of a 3-bp mononucleotide A repeat in *mys*-9 at those positions. However, these data strongly support the view that B1a and B1i are independent insertions into the same site in the *mys*-9 locus.

Sequence analysis of the IDb insert from the dif3.50 allele and B1b insert from the criY3.78 allele showed the even more surprising result that the two independent events giving rise to these SINE insertions used identical nick sites at both the 5′ and the 3′ ends of their target sites. Figure 3 shows that these two independent events would have been indistinguishable had it not been that the insertions were of two different types of SINEs.

The *mys*-9 locus contains two hot spots for SINE insertion: When the location and orientation of each of the major insertions into these 13 *mys*-9 alleles were determined, the picture shown in Figure 4 emerged. Within the *mys*-9 locus are two regions that are hot spots for insertion of SINEs and possibly of other elements. The first region, which includes the B1a and B1i SINEs, has undergone at least six independent insertion events within a space of 248 bp, with five of those insertions into a region of 83 bp centered around base pair 1096 of the *mys*-9 locus. The second region, which includes the identically positioned IDb and B1b elements, also has undergone at least six insert events localized to a 150-bp region centered around base pair 1975.

Why are the above two regions insertional hot spots?

```
B1a Flanks:      CATGTAAAGATggatattac-GCC~AAAA-ggatattacACAGAGAATCT

B1i Flanks:      CATGTaaagatggatattat-TCC~AAAA-aaagattggatattacACAG

B1c Flanks:      ACAAAATAAAAGGTCATaag-GCC~ATAA-aagACACATTTGCAAACCTG

B1g Flanks:      AAGTaaaacatttcaaacaa-GCC~AAAA-aaaacatttcaaacaaATTG

B1d Flanks:      TATAAAAaataagccatgtc-GCC~TAAT-aataagccatgtaAAGATGG

B1e Flanks:      SATTaagatacatcaggttt-GCC~AAAA-aagatacatcaggtttGACC

B1h Flanks:      CAGTTaagaatcccatagac-GCC~AAAA-aagaatcccataaacAACAC

B1f Flanks:      ATGATagaataaaagggtag-GCC~AAAA-agaataaaagggtagATTGC

B1b Flanks:      CTTTaaaagtaggttcagcg-GCC~AAAA-aaaaataggttcagcaATCT


IDb Flanks:      CTTTaaaagtaggttcagta-GGG~AAAG-aaaagtaggttcagcaATCT

IDa Flanks:      CAAATTTAAAGTCaattttg-GGG~TAAT-aattttaTACTGTATATATA


B2 Flanks:       TTTTcaaaaaggaaggtttt-GGG~AAAA-caaaaaggaaggttttAACT


L1/X2 FLanks:    TCTTCCTaaaagaagagggg-GAG~AAAA-aaaaaaagaggggAATATGA


Unknown Insert's Flanks:
                 AAGCAATAGATTCacaataa-GAC~AGAG-acagtaaAGGTAGTATTAAA
```

FIGURE 3.—Sequence flanking insertions into *mys*-9. The first and last 20 bases of sequence in each line is the sequence surrounding the designated inserts. The three bases immediately after the first dash in each line are the first three bases of the insert, and the four bases immediately before the second dash are the last four bases of the insert. ~ denotes insert sequence that is not shown. Target site duplications are in lowercase. Pyrimidines immediately 5′ of the target site duplication that are believed to be important for SINE insertion mediated by LINE-1 reverse transcriptase/endonuclease are boxed. Purine stretches believed to be important for the same process are underlined in the left target site duplication. In some cases of differences between the target site duplications, it has been possible to identify the ancestral state (shown in boldface) by comparison with other *mys*-9 alleles.

LINEs and SINEs appear to preferentially insert into regions of high AT content (FURANO *et al.* 1986; DEININGER 1989; HUTCHISON III *et al.* 1989). Plots of AT content throughout the *mys*-9 locus show the two insertional hot spots to be of a relatively high percentage of AT but not the most AT-rich regions in the locus. Jurka has suggested that kinkable DNA may serve as a preferential substrate for the LINE-1 endonuclease likely to be involved in insertion of at least some retrotransposons (JURKA *et al.* 1998). We surveyed the *mys*-9 locus to determine the density of the kinkable dinucleotides TG, TA, and CA but found no correlation between high densities of these dinucleotides and position of inserts.

Finally, we examined the sequence flanking each of these inserts (Figure 3) to search for motifs important for insertion. Target site duplications are associated with every insert and are shown in lowercase in Figure 3. The great majority of these target site duplications have a number of characteristics that are consistent with previous suggestions that insertion of B1, B2, and ID ele-

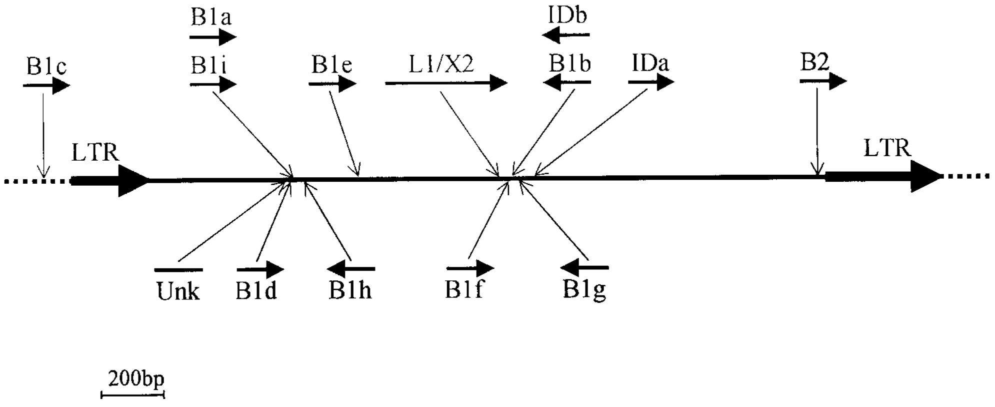

FIGURE 4.—Location of inserts >50 bp found in *mys*-9 alleles. A *mys*-9 allele with no major inserts is shown, including single-copy flanking sequence (dotted lines) through the regions complementary to the M9-17 and M9-16 primers used for amplification of the locus. The arrows associated with each SINE or LINE-1 indicate the orientation of that insert. The LINE-1 insert contains the 5′ end of a LINE-1, including most of its open reading frame-1 (ORF-1), followed by a region containing nonrepetitive DNA that includes a portion showing homology with a human X2 box repressor cDNA.

ments is aided by LINE-1 reverse transcriptases (Jurka and Klonowski 1996; Jurka 1997). The initial nick site has been proposed to occur immediately 5′ of a purine residue, which is in most cases an A residue (Feng *et al.* 1996; Jurka and Klonowski 1996). By the orientation shown in Figure 3, this would correspond to nicking in the bottom strand, with an associated pyrmidine (usually a T residue) in the top strand immediately 5′ of the left target site duplication. Thirteen of the 14 inserts examined here have a pyrimidine immediately 5′ of the target site duplication and 11 of those 13 have a T at this position. The target site duplication for the single exception, B1d, would actually extend an additional four A's in the 5′ direction, if the four TAA trinucleotide repeats seen at the 3′ end of B1d were produced by repeat expansion after insertion of the element. Such a scenario would give B1d the common T nucleotide immediately 5′ of the target site duplication, resulting in a pyrimidine at the expected position in all 14 of the inserts. Examination of the dinucleotides cleaved by a presumed initial nick under this model reveals that 11 of the 14 inserts contain kinkable DNA (TG, TA, or CA) at this position. The 5′ bases in the majority of the target site duplications contain a polypurine tract (underlined in Figure 3), which has been proposed to increase the probability of nicking by the LINE-1 reverse transcriptase (Feng *et al.* 1996). More specifically, this tract in many cases contains the poly(A) tract believed to be part of the recognition sequence TTAAAA (Jurka and Klonowski 1996). The only striking feature found in the right flank of the target site duplication is an A immediately 3′ of the duplication in 12 of the 14 inserts. These results suggest that the majority, or all, of these inserts have transposed with the help of LINE-1 reverse transcriptase, but that sequence context alone is not sufficient to explain why insertions have occurred so commonly in the two regions described.

DISCUSSION

The distribution of the *mys*-9 element among Peromyscus species and the divergence of its LTRs indicate that it is extremely old, having inserted approximately 12 million years ago. Since paleontological evidence suggests that the Peromyscine radiation occurred around or before 6.5 million years ago (Catzeflis *et al.* 1992), it may well be that *mys*-9 inserted around, or before, the time of that radiation and well before the divergence within the genus Peromyscus.

Multiple *mys*-9 alleles have existed within *P. maniculatus* for extended periods, showing differences as high as 3.03 changes/100 bp (Table 1), which would suggest these alleles may have coexisted within the species for roughly 1.5 million years. Even within a single animal, the mouse from Iowa, coexisting alleles show divergence from each other of 2.68 changes/100 bp. Perhaps this greater divergence seen among the *P. maniculatus* alleles than among the *P. leucopus* alleles reflects the greater geographical distribution of *P. maniculatus* (King 1968). Phylogenetic trees based on sequence data also show that allele size is not a valid phylogenetic character at the species level. Alleles from distantly related species can have nearly identical sizes while closely related alleles can have different sizes.

The surprising range of allele sizes seen among these 13 *mys*-9 alleles led to our discovery that this locus contains two distinct regions that are hot spots for insertion. In the more 5′ of these two regions, six insertion events have occurred within 248 bp, with five of those insertions occurring in an area of only 83 bp. In the 3′ region, six insertions are localized to a 150-bp area. The question of whether the entire *mys*-9 locus represents an overall hot spot for insertion is a bit harder to address in this study because the 9 alleles originating from *P. leucopus* and *P. maniculatus* were selected for further study from 37 animals on the basis of both the geographic locations of the individual animals and the range of allele sizes. Other loci in rodents and humans have been noted as hot spots for SINE and LINE-1 insertion (Qin *et al.* 1991; Wells and Bains 1991), but those loci have not collected insertions within such small regions. In this study we have been able to peer into discrete windows of time by analyzing alleles at a presumably unselected locus (*mys*-9) of known age from both a range of species and from multiple individuals within single species.

Why are these two *mys*-9 regions insertional hot spots? It has been recognized for some time that, although many mammalian retroelements are dispersed throughout the genomes of mammals, insertion is not completely random (Baker and Wichman 1990; Sandmeyer *et al.* 1990; Wichman *et al.* 1992; Craig 1997). While the inserts into *mys*-9 tended to group at regions of higher AT, those regions with the highest AT content are not the regions showing greatest insertion. In light of evidence that B1s, B2s, and ID elements may insert at sites of nicking in kinkable DNA (Jurka *et al.* 1998), we determined the density of kinkable dinucleotides throughout *mys*-9 but found no evidence for preferential insertion at regions containing a high density of kinkable DNA sites. Higher-order structures such as chromatin configuration are believed to play a role in insertion of some retroelements with increased incidence of insertion seen into DNA located on nucleosomes and oriented such that the major groove is open to the surface (Pryciak and Varmus 1992; Muller and Varmus 1994). Perhaps the same is true of SINE and LINE-1 insertion.

Analysis of the target site duplications flanking each of the inserts in this region reveals a number of characteristics supporting suggestions that insertion of B1, B2, and ID elements is aided by LINE-1 reverse transcriptase/endonuclease (Jurka and Klonowski 1996; Jurka 1997). Depending on how the sequences are interpreted, either 13 of the 14, or all 14 of the inserts, have a pyrmidine just 5′ of the left target site duplication. Eleven of the 14 inserts contain kinkable DNA at this position. The 5′ bases in the majority of the target site duplications contain a polypurine tract, which has been proposed to increase the probability of nicking by the LINE-1 reverse transcriptase/endonuclease. There is a real possibility that all of these inserts have gained admission to the region by use of the LINE-1 reverse transcriptase. Interestingly, the insert of approximately 4 kb present in the manCA8.1 allele is a segment that appears to contain the 5′ end of a LINE-1 followed by nonrepetitive DNA. We have shown by PCR that this arrangement of the 5′ end of LINE-1 followed by the same nonrepetitive sequence is present in the genomes of other *P. maniculatus* that do not have the CA8.1 allele. A probable explanation is therefore that this insertion occurred as a result of transcription of the segment from a promoter in a defective LINE-1 followed by the relatively unusual aid of another LINE-1 functioning *in trans* (Esnault *et al.* 2000; Furano 2000).

Transposons have long been known to insert into genomes with a wide range of specificities, ranging from nearly random insertion to extremely sequence-specific insertion mechanisms such as those used by the insect retrotransposons R1 and R2 that limit insertion almost exclusively to the 28S rRNA genes (Xiong and Eickbush 1988; Baker and Wichman 1990; Sandmeyer *et al.* 1990; Wichman *et al.* 1992; Craig 1997). Even though vertebrate SINEs and LINEs have been known to exhibit some site specificity, their widespread dispersal throughout vertebrate genomes has led to the belief that their specificity is low and that the probability of multiple independent insertions at identical sites is nil (*e.g.*, so low that a SINE insertion can be considered to be an essentially homoplasy-free phylogenetic character) (Batzer and Deininger 1991; Batzer *et al.* 1994; Takahashi *et al.* 1998; Nikaido *et al.* 1999). Perhaps the most important information arising from this study comes from our observation of two cases in which different SINE insertion events have occurred at identical sites. These are the first reported cases of multiple SINE inserts at identical locations in Sigmodontine rodents. In the course of analyzing LINE-1 insertions into the β-globin cluster in mice, Burton *et al.* identified a position at which either gene conversion or same-site LINE-1 transposition had occurred, but the sequences obtained in that situation did not allow them to determine which of the alternative scenarios was more probable (Burton *et al.* 1991). Slattery and co-workers have recently found what appear to be two independent insertions of the same tRNA-derived SINE element at the same site in members of the cat family (*Felis silvestris* and *Lynx rufus*; Slattery *et al.* 2000). A clear example of two independent B1 inserts into the same site in the genus Mus has recently been found by Kass and co-workers (2000). In the case of the B1a and B1i elements we suspected that the two inserts represented independent events only because a single insert at that location was incongruent with the phylogeny of the *mys*-9 alleles. We were able to distinguish the two events by virtue of a 3- to 6-bp difference in the left border of the target site duplications. In the case of the IDb and B1b elements both the insert site and the target site duplications were identical. If the same type of SINE had inserted each time in this latter case, we would have interpreted the two events as a single one. Thus the estimate of two same-site insertions formally represents the minimum number in this data set.

Although same-site insertions are probably rare, these results suggest that SINEs exhibit a greater specificity for insertion at specific sites than previously recognized, to the extent that multiple identical insertions can indeed occur at single sites. The presence of a retrotransposon at a single locus in multiple taxa remains an extremely powerful phylogenetic marker, but caution is required before concluding that the existence of a particular SINE at a particular locus in multiple individuals is indicative of common ancestry (Hillis 1999). Such caution is particularly warranted in cases where a single insertion event is the sole support for a specific phylogenetic hypothesis.

We acknowledge the use of the University of Idaho Automated DNA Sequencing Facility. We thank Drs. Robert Baker, Kimberly Nelson, and Ira Greenbaum for donating tissues for this study and Jack Sullivan for helpful discussions regarding the phylogenetic analysis. This work was supported by National Institutes of Health grant GM-38737 to H.A.W.

LITERATURE CITED

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman *et al.* (Editors), 1989 *Current Protocols in Molecular Biology*. Green Publishing/Wiley-Interscience, New York.

Baker, R. J., and H. A. Wichman, 1990 Retrotransposon mys is concentrated on the sex chromosomes: implications for copy number containment. Evolution **44:** 2083–2088.

Batzer, M. A., and P. L. Deininger, 1991 A human-specific subfamily of Alu sequences. Genomics **9:** 481–487.

Batzer, M. A., M. Stoneking, M. Alegria-Hartman, H. Bazan, D. H. Kass *et al.*, 1994 African origin of human-specific polymorphic *Alu* insertions. Proc. Natl. Acad. Sci. USA **91:** 12288–12292.

Burton, F. H., D. D. Loeb, M. H. Edgell and C. A. Hutchison, 1991 L1 gene conversion or same-site transposition. Mol. Biol. Evol. **8:** 609–619.

Carleton, M. D., 1989 Systematics and evolution, pp. 7–141 in *Advances in the Study of Peromyscus (Rodentia)*, edited by G. L. Kirkland, Jr. and J. N. Layne. Texas Tech University Press, Lubbock, TX.

Casavant, N. C., A. N. Sherman and H. A. Wichman, 1996 Two persistent LINE-1 lineages in Peromyscus have unequal rates of evolution. Genetics **142:** 1289–1298.

Catzeflis, F. M., J. P. Aguilar and J. J. Jaeger, 1992 Muroid rodents: phylogeny and evolution. Tree **7:** 122–126.

Craig, N. L., 1997 Target site selection in transposition. Annu. Rev. Biochem. **66:** 437–474.

Deininger, P. L., 1989 SINEs: short interspersed repeated DNA elements in higher eucaryotes, pp. 619–636 in *Mobile DNA*, edited by D. E. Berg and M. M. Howe. American Society of Microbiology, Washington, DC.

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

Feng, Q., J. V. Moran, H. H. Kazazian, Jr. and J. D. Boeke, 1996

Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87:** 905–916.

Furano, A. V., 2000 The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog. Nucleic Acid Res. Mol. Biol. **64:** 255–294.

Furano, A. V., C. C. Somerville, P. N. Tsichlis and E. D'Ambrosio, 1986 Target sites for the transposition of rat long interspersed repeated DNA elements (LINEs) are not random. Nucleic Acids Res. **14:** 3717–3727.

Hamada, M., N. Takasaki, J. D. Reist, A. L. DeCicco, A. Goto *et al.*, 1998 Detection of the ongoing sorting of ancestrally polymorphic SINEs toward fixation or loss in populations of two species of charr during speciation. Genetics **150:** 301–311.

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Hillis, D. M., 1999 SINEs of the perfect character. Proc. Natl. Acad. Sci. USA **96:** 9979–9981.

Hutchison, C. A., III, S. C. Hardies, D. D. Loeb, W. R. Shehee and M. H. Edgell, 1989 LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome, pp. 593–617 in *Mobile DNA*, edited by D. E. Berg and M. M. Howe. American Society for Microbiology, Washington, DC.

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

Jurka, J., and P. Klonowski, 1996 Integration of retroposable elements in mammals: selection of target sites. J. Mol. Evol. **43:** 685–689.

Jurka, J., P. Klonowski and E. N. Trifonov, 1998 Mammalian retroposons integrate at kinkable DNA sites. J. Biomol. Struct. Dyn. **15:** 717–721.

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

King, J. A. (Editor), 1968 *Biology of Peromyscus (Rodentia)*. American Society of Mammalogists, Bowling Green, KY.

Longmire, J. L., A. K. Lewis, N. C. Brown, J. M. Buchingham, L. M. Clark *et al.*, 1988 Isolation and molecular characterization of a highly polymorphic centromeric tandem repeat in the family Falconidae. Genomics **2:** 14–24.

Muller, H. P., and H. E. Varmus, 1994 DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. EMBO J. **13:** 4704–4714.

Murata, S., N. Takasaki, M. Saitoh and N. Okada, 1993 Determination of the phylogenetic relationships among Pacific salmonids using short interspersed elements (SINES) as temporal landmarks of evolution. Proc. Natl. Acad. Sci. USA **90:** 6995–6999.

Nikaido, M., A. P. Rooney and N. Okada, 1999 Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proc. Natl. Acad. Sci. USA **96:** 10261–10266.

Pryciak, P. M., and H. E. Varmus, 1992 Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell **69:** 769–780.

Qin, Z. H., I. Schuller, G. Richter, T. Diamantstein and T. Blankenstein, 1991 The Interleukin-6 gene locus seems to be a preferred target site for retrotransposon integration. Immunogenetics **33:** 260–266.

Sandmeyer, S. B., L. J. Hansen and D. L. Chalker, 1990 Integration specificity of retrotransposons and retroviruses. Annu. Rev. Genet. **24:** 491–518.

Sawby, R., and H. A. Wichman, 1997 Analysis of orthologous retrovirus-like elements in the white-footed mouse, Peromyscus leucopus. J. Mol. Evol. **44:** 74–80.

She, J. X., F. Bonhomme, P. Boursot, L. Thaler and F. M. Catzeflis, 1990 Molecular phylogenies in the genus *Mus*: comparative analysis of electrophoretic, scnDNA hybridization and mtDNA RFLP data. Biol. J. Linn. Soc. **41:** 83–103.

Shimamura, M., H. Yasue, K. Ohshima, H. Abe, H. Kato *et al.*, 1997 Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature **388:** 666–670.

Slattery, J. P., W. J. Murphy and S. J. O'Brien, 2000 Patterns of diversity among SINE elements isolated from three Y-chromosome genes in carnivores. Mol. Biol. Evol. **17:** 825–829.

Sullivan, J., and D. L. Swofford, 1997 Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. **4:** 77–86.

Sullivan, J., J. A. Markert and C. W. Kilpatrick, 1997 Phylogeography and molecular systematics of the Peromyscus aztecus species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. **46:** 426–440.

Swofford, D. L., 2000 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Asssociates, Sunderland, MA.

Takahashi, K., Y. Terai, M. Nishida and N. Okada, 1998 A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. Mol. Biol. Evol. **15:** 391–407.

Wells, D., and W. Bains, 1991 Characterization of an unusual human histone H3.3 pseudogene. DNA Seq. **2:** 125–127.

Wichman, H. A., S. S. Potter and D. S. Pine, 1985 Mys, a family of mammalian transposable elements isolated by phylogenetic screening. Nature **317:** 77–81.

Wichman, H. A., R. A. Van Den Bussche, M. J. Hamilton and R. J. Baker, 1992 Transposable elements and the evolution of genome organization in mammals. Genetics **86:** 287–293.

Xiong, Y., and T. H. Eickbush, 1988 The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Mol. Cell. Biol. **8:** 114–123.

Communicating editor: W. F. Eanes