Journal of Molecular Evolution
Volume 63, Number 5 / November, 2006
http://dx.doi.org/10.1007/s00239-005-0301-2

Local Mutagenic Impact of Insertions of LTR Retrotransposons on the Mouse Genome

Erick Desmarais[1,*], Khalid Belkhir[1], John Carlos Garza[2] and François Bonhomme[1]

(1) Laboratoire Génome, Populations, Interactions, Adaptation, UMR5171 CNRS-IFREMER, Université Montpellier II, CC-G3 Montpellier Place E. Bataillon, 34095, France
(2) Southwest Fisheries Science Center, 110 Shaffer Road Santa Cruz, CA 95060, USA

*: Corresponding author : desmarais@univ-montp2.fr

Abstract:

Solitary LTR loci are the predominant form of LTR retrotransposons in most eukaryotic genomes. They originate from recombination between the two LTRs of an ancestral retrovirus and are therefore incapable of transposition. Despite this inactivity, they appear to have a substantial impact on the host genome. Here we use the murine RMER10 LTR family as an example to describe how such elements can reshape regions of the genome through multiple mutations on an evolutionary time scale. Specifically, we use phylogenetic analysis of multiple copies of RMER10 in rodent species, as well as comparisons of orthologous pairs in mouse and rat, to argue that insertions of members of this family have locally induced the emergence of tandem repeat loci as well as many indels. Analysis of structural aspects of these sequences (secondary structures and transcription factors signals) may explain why RMER10 can become endogenous "mutagenic" factors through induction of replication fork blockages and/or error-prone repair of aberrant DNA structures. This hypothesis is also consistent with features of other interspersed repeated elements.

Keywords: Mouse genome - LTR retrotransposons - Simple tandem repeats - RMER10

Introduction

Insertion of transposable elements (TE) in a genome may have immediate harmful consequences that can be lethal for the cell, or even the entire organism when in the germline. In certain cases, however, the effect may become apparent later on with the onset of pathology after the birth of the mutated individual (Kazazian 1998; Deininger and Batzer 1999, 2002; Prak and Kazazian 2000; Batzer and Deininger 2002; Deininger et al. 2003). Such major effects are usually the result of interruption of a coding region or shuffling of gene arrangement through ectopic recombination. However, the insertion of a TE in a non-coding region may also have a local effect, sometimes dramatic, on the region of insertion. Remarkable examples of this include some of the most unstable loci in the human and mouse genomes, which are repetitive regions that have emerged from copies of interspersed elements: human minisatellite MS32 in LTR10a, a repetitive element derived from the long terminal repeat (LTR) of an endogenous retrovirus-like sequence (HERV-I) (Armour et al. 1989); MSY in MLT1c, a Mammalian-apparent LTR retrotransposon (MaLR) (Smit 1993); the large mouse short tandem repeat (STR or microsatellite) loci Ms6-hm and Hm-2, which are also found in MaLRs (MTc and Orr1 respectively) (Kelly et al. 1989; Kelly 1994) and MMS10 in a B1 element (Bois et al. 1998). Another example of the possible effect of TE insertion on DNA sequences in the region of insertion comes from the general observation of associations between interspersed elements and STR; microsatellite loci have frequently been reported to be associated with SINEs or LINEs in genomic library screening experiments (Armour et al. 1989; Zuliani and Hobbs 1990; Gastier et al. 1995; Yandava et al. 1997). Surveys of genomic sequence databases tend to confirm this association, at least for a large proportion of such repeat loci (Kaukinen and Varvio 1992; Arcot et al. 1995; Jurka and Pethiyagoda 1995; Nadir et al. 1996; Ovchinnikov et al. 2001). To explain this association, it has been proposed that STRs are derived from poly-A tails integrated into genomic sequence during TE retrotransposition. If this is the case, then interspersed repeated elements might be the source of many short tandemly repeated sequences (Levinson and Gutman 1987; Arcot et al. 1995). In dipterans, a retro-transposon called Mini-me contains two proto-microsatellite regions in its sequence, a TA_{4-6} and a cryptic tetranucleotide repeat, that have given rise to a large proportion of the tetranucleotide STR loci found in Drosophila species (Wilder and Hollocher 2001). The high redundancy (many STR arrays share the same flanking sequences)

observed in lepidopteran genomic libraries enriched for microsatellite markers has been attributed to the reiterated derivation of STR loci from this Mini-me element (Zhang 2004).

Nevertheless, several questions remain unanswered: Why don't all poly-A tails degenerate into STR loci or all TE integration sites give rise to tandem repeats? How can repetitive sequences emerge from TEs that do not have a poly-A tail? How is the series of mutations that leads to an STR locus initiated? Perhaps most fundamentally, what are the unique characteristics, if any, of the loci where such an accumulation of evolutionary events occurs? The slipped strand mispairing model of mutation (SSM) can explain most changes in the number of repeats in an already established STR locus (Levinson and Gutman 1987). Moreover, experimental data (Sia et al. 1997) and surveys of genomic databases suggest the existence of a size threshold for a locus to experience SSM (Bell and Jurka 1997; Rose and Falush 1998; Zhu et al. 2000; Dieringer and Schlotterer 2003), suggesting that another model of mutation could exist for the shorter loci. Thus, the SSM mechanism does not explain convincingly the *de novo* emergence of repetitive sequences and the underlying causal factors in their origin remain generally unknown. To address these questions, we have analyzed the consequences of integration of copies of RMER10, an LTR-retrotransposon family, in the murine genome and how these sequences have evolved since they spread through the genome.

RMER10 is a rodent-specific family of MaLR with two described sub-families, 10A and 10B (RepBase, Jurka et al. 2005); the 10B sub-family differs from 10A primarily by a 16bp deletion in the 5' half and by a conserved and specific 3' end. Like many other MaLRs, RMER10 probably originated from the LTR part of an endogenous retrovirus (Smit 1993) of which it maintains the classical structure consisting of 3 well-defined regions, U3, R and U5 (Figure 1). Specific DNA Pol.II transcription signals (the TATA box and the Polyadenylation signal) originally present in functional LTRs are often still identifiable but, because the rest of the provirus is lacking, they are considered as stationary and inactive, or at least non-autonomous. They do not possess a poly-A tail since they are integrated not as retrotranscribed mRNA but as part of a provirus with the complete retroviral genome.

In a previous study, we showed that an RMER10 containing the STR marker locus *D1mit29* has experienced complex evolution during radiation of the genus *Mus* (Garza and Desmarais 2000). This study revealed a high degree of variability in both structure and sequence of the STR locus among *Mus* species, as well as the presence of many insertion/deletion (indels) variants in the flanking sequences. Together, these observations

suggest that RMER10 may be a good candidate to study the mechanisms and causes of STR emergence and instability.

Here, we evaluate potential mechanisms of origin for such STR loci through a quasi-exhaustive analysis of the copies of RMER10 present in the mouse genome. All identified copies were assessed for the presence of tandem repeats, which revealed a greater than expected association of these sequences with STR loci. The ancestral states of the RMER10 subfamilies were then reconstructed to look for the presence of special features that might favour the emergence of tandem repeats in copies that stem from them. We detected the combined presence of transcription factor signals and palindromes that could destabilize the DNA double-helix through formation of abnormal secondary structures. Such structures might then induce either a replication fork arrest or simply trigger repair pathways that can account for the high sequence variability of the RMER10 copies and the gain/loss of many microsatellite repeats.

Methods.

Retrieval of mouse RMER10 sequences from genomic databases

We retrieved all the sequences referenced as RMER10 in the mouse genome using the ENSEMBL server (Hubbard et al. 2002; Clamp et al. 2003) with the Mouse Genome Assembly database v19.30. RMER10 sequences were extracted, along with 50 bp of flanking sequence on each side of the target sequence, and analysed successively by CENSOR (Jurka et al. 1996) and REPEATMASKER (Smit et al. 1996-2004; http://www.repeatmasker.org/) to confirm their sub-family assignment and also to remove sequences annotated or aligned with low confidence (usually very short stretches of nucleotides).

We found that many single RMER10 elements have been broken up into 2 or more sections by insertion of exogenous DNA or the emergence of tandem repeats, giving rise to separately referenced sequences in the ENSEMBL database. To account for the existence of these split elements, we looked for all sets of compatible and complementary adjacent segments able to compose the longest possible single RMER10 copy within a 1Kb window. We then aligned all such sequences with the consensus to confirm that they indeed originally belonged to the same copy of RMER10 and, when they did, we treated the entire genomic region encompassing them as a unique copy of RMER10.

Orthologous sequences from the rat genome were recovered from the Compara database on the ENSEMBL server and used to assess the presence of STR loci (see below).

Sequences from the locus *D1mit29* were determined in several rodent species – *Apodemus sylvaticus*, *Mus cookii* and *Mus plathythrix* – using the conditions previously described by Garza and Desmarais (2000), but PCR amplifications were performed at 56°C and contained 2.5% formamide. In *A. sylvaticus*, the locus was visible with ethidium bromide staining and was directly sequenced. For the two other species, a radioactive PCR was first performed and the bands were then extracted from the dried gel, re-amplified and directly sequenced (Desmarais et al. 1998).

Tandem repeat search

Tandem Repeat Finder (Benson 1999) was used to detect the presence of tandemly-repeated nucleotides (Tautz et al. 1986) in the collected genomic fragments. This program can detect any type of tandem repeat element without *a priori* information on its sequence, length or conservation of its repeat unit. However, in order to reduce the number of positive hits, the program parameters were set with maximum period size (roughly the size of the repeat unit) of 10 and minimum alignment score between two adjacent copies of 40. This means that the repeat units can be partially degenerated and the repeat motif size can range from 1 to 10. We further discarded the loci that have less than three copies of the repeat. This strategy means that the detected tandem repeat regions we found include substantially more loci than the simple repeat regions annotated in ENSEMBL that are restricted to "canonical" patterns of repeated nucleotides.

Sequence alignment

Automated multiple alignment procedures were unsuccessful due to the very large number of insertions, deletions, and truncations present in the different RMER10 copies and their extreme divergence in the 3' region. Therefore, alignments were performed for only the 5' part of the elements, the U3 region of the LTR, which was the only region that could be aligned confidently between the RMER10A and RMER10B subfamilies. Pairwise alignments of each sequence with the RMER10 consensus sequences (Jurka et al. 2005) were then used to guide the construction of multi-sequence alignments of each sub-family using the GeneDoc program (Nicholas and Nicholas 1997; www.psc.edu/biomed/genedoc).

Phylogeny estimation and ancestral sequence reconstruction.

Phylogenies of several subsets of the aligned sequences were estimated using the maximum likelihood procedure implemented in the PHYML program (Guindon and Gascuel 2003). Several models of evolution were tested by varying the gamma parameter for the distribution of rate variation among sites, and the one with the highest likelihood was retained. A first phylogeny (not shown) was built with a data set of 207 mouse and rat sequences of only the U3 part of the RMER10, containing 87 pairs of putative orthologous sequences and 33 other RMER10 copies. This was to check the validity of the alignment and whether phylogenetic analysis can confidently be performed using only the 5' parts of the RMER10 elements, in spite of their reduced size and high level of variation. A second tree was then constructed using 3 members of each sub-family from each mouse chromosome, whenever possible, and according to the completeness of their U3 sequence (but not their homology with the consensus sequence). This tree included a total of 182 sequences. Ancestral sequences were then reconstructed with the program DNAML from the Phylip package ver.3.61 (Felsenstein 2004; http://evolution.genetics.washington.edu/phylip.html) using the <User-defined tree> option.

Structure computations

Secondary structures and the corresponding free energies (ΔG) of the 210 nucleotides of the U3 region at 38°C were estimated using the mFOLD program for DNA folding (Zucker 2003) and manipulated with RNADRAW (Matzura and Wennborg 1996). The results obtained for consensus and reconstructed ancestral sequences were compared to those of a set of 100 random sequences of exactly the same size and base composition (Option scramble in RNADRAW), a set of 172 actual RMER10 sequences, and one of 1362 genomic sequences flanking the 5' or 3' side of RMER10 sites on chromosomes 1 and X. Due to the substantial number of indels in some genomic fragments, the ΔG was weighted by the size of the fragment.

Detection of transcription factors.

Putative transcription factor recognition sites were identified with the program AliBaba2 (Grabe 2002) on the Biobase server (http://www.gene-regulation.com/pub/programs.html) using matrices from the TRANSFAC database v6.0 (Matys et al. 2003). The minimum homology with the TRANSFAC matrices was set to 80%.

Results.

Association of RMER10 copies and tandem repeats

RMER10 fragment retrieval.

Extraction from the ENSEMBL database of all fragments annotated as RMER10A or RMER10B gave a total of 7085 sequences: 3915 from the 10A and 3170 from the 10B subfamilies. After filtering and re-assembling fragments that were separately annotated in the database but belonged to the same RMER10 copy, 5858 fragments remained, with 3322 RMER10A and 2336 RMER10B copies. In addition, 193 elements, including *D1mit29*, with a chimeric structure were identified. This structure probably resulted from the recombination between a 10A and a 10B element since it combines a 10A-specific U3 region, recognized by the diagnostic 16 bp indel, and a 10B-specific R region, characterized by the sequence around the polyadenylation signal. This specialized structure was confirmed by analysing the 5' and 3' parts of the copies separately with CENSOR and REPEATMASKER software. We therefore named these chimerically structured fragments as RMER10AB. We also found 7 elements with the reciprocal BA configuration (i.e. U3 region of a 10B with an R region of a 10A).

Nature and structure of associated tandem repeat loci

Of the 5858 fragments analysed with the TRF program, 1495 (25.5%) contained at least one STR (the total number of repeat arrays detected was 1638) and in 72% of these fragments the STR was located inside the RMER10 element itself, with the others found in the immediately flanking 50 bp. Strikingly, the repeat arrays were not homogenously distributed over the whole span of the RMER10. The vast majority of them were found in the same R region of the LTR, where they are also localized in the locus *D1mit29* (Garza and Desmarais 2000).

The parameters used with TRF set a minimum locus size of 20 bp and allowed the detection of loci with less than 6 repeats, a minimum number commonly accepted for a locus to be considered as variable (Weber 1990; Messier et al. 1996). However, we identified few short loci (Table 1) and some of the shorter ones are also obviously part of a larger, repeated locus that stretches out of the fragment examined. Most of the other short loci found had an

imperfect basic motif with a large repeat length (>6 bp) containing shorter cryptic repeats that are very likely to behave like perfect repetitions. We choose to include loci with repeat motifs up to 10 bp, even though microsatellites have been defined by some authors as loci with repeats <6 bp, as differences in mutational mechanisms have not been demonstrated between repeats in this size range (Sia et al. 1997). However, these loci are rare and become very infrequent as the size of the repeat motif increases (Table 1). We also found a locus with a 49 bp motif repeated four times. It was not used in further analyses, but it was nonetheless interesting because of its size, indicating that duplication events can encompass large stretches of nucleotides, its location in the R region, and its structure, formed by nested palindromic sequences.

The size of the identified STR loci ranged from 20, the TRF minimum threshold, for poly-A or poly-T tracts, to 458 bp for a composite locus formed by CTTTC and CT repeats. As TRF allows the presence of divergent repeat copies in a locus, some of them were degenerated and may not represent STR loci *sensu stricto*, nor would they likely have been recognized as such by other methods. Nevertheless, there were very few such degenerated STR, while 41% of the loci had 100% of the repeats matching the basic motif and more than half had a matching score between adjacent repeat units of at least 95% overall (for example, one TT present in a stretch of 21 CT).

The orientation of the RMER10 copy allowed the distinction between complementary strand patterns that revealed a dissymmetry for the base composition of the repeated motifs: 872 loci (55%) did not contain any A and 579 (35%) were composed exclusively of C and/or T. For example, we found 295 CT vs. only 106 AG.

Association of RMER10 elements with STR loci

The proportion of genomic fragments that are associated with a tandem repeat is variable among the sub-families of RMER10 (Table 2). We compared the percentage of RMER10-containing fragments where an STR locus was also found to that observed in randomly drawn genomic fragments, which assumes an even distribution of STR loci across the genome (null hypothesis). One hundred batches of 2400 randomly drawn mouse genomic fragments, 425bp in length (number and average size of the RMER10A copies, the sub-family with the largest members) and of the same chromosomal distribution, were scanned for STR using the same TRF parameters. The values observed with all RMER10 sub-families fell outside the distribution of values obtained under the null hypothesis of random association (mean: 15.28). Larger fragments are more likely to contain a repeat region simply because of

their size and to get an association rate equivalent to that of RMER10B, we had to extend the size of the random fragments to 1025 bp. Thus we can conclude that RMER10 are significantly more associated with STR than sequences of the same size randomly drawn from the genome.

We then analysed the sequences immediately flanking each RMER10-containing fragment (both 5' and 3' over the same length, but beyond the 50bp initially analyzed) to evaluate whether these LTRs are situated in chromosomal regions particularly rich in STR loci. In all cases we found an association rate of the flanking sequences higher than in the randomly drawn 425bp genomic fragments (Table 2). However, many STR found in the flanking sequences are, in fact, extensions of those found in the RMER10 region. When these loci are removed, the association values for flanking sequences of all but the 10AB subfamily drop down closer to values similar to those of random fragments but remain significantly higher (outside the distribution of values obtained for random fragments), except for the 5' flanking the 10B (p-value=0.4). The values for the sequences flanking the B and AB sub-families are lower than that of the LTR fragments themselves, whereas those for the A subfamily were not. Because many members of this sub-family are apparently truncated, and thus lacking the R region, we separately reanalyzed only the 2147 nearly full-length (larger than 300 bp) copies from the A group and found an association rate of 24.64, which become significantly higher than the rate of association in the corresponding flanking regions (p-value< 0.0001) as it is in the RMER10B sub-family (p-value< 0.0001) but not in the RMER10AB. For this latter sub-family, both the low number of members and a high rate of association of flanking sequences explain this lack of significance.

Finally, we assessed the extent of the area of higher STR density around RMER10 fragments. We worked directly on the annotated chromosomes from the NCBI Mouse Genome Assembly m36 and analyzed the genomic neighbourhood of RMER10 for the presence of STR. We cut the flanking regions of RMER10 in slices of 265 pb, the mean size of the RMER10 annotated fragments, and summed the fragments containing STRs in each of these windows. The high STR density is restricted to the vicinity of the RMER10 copies since it is nearly reduced by a factor 2 at a distance of only 265bp away from the ends of the RMER (not shown). This confirmed that RMER10 are not integrated in genomic region with a higher density of STRs.

The association between RMER10 fragments and tandem repeats could be due to at least two different phenomena: first, multiplication of an ancestral copy that originally

contained an STR; second, multiple, independent appearances of tandem repeats which arise preferentially at nearly the same place in the different copies of RMER10. This latter, less-parsimonious hypothesis implies that there are factors in RMER10 sequences that favour the emergence of STR loci. To evaluate the respective contribution of both modes of STR appearance, we used phylogenetic reconstructions to reconstitute the ancestral sequences at the base of each sub-family of RMER10. We also tried to evaluate the time when the STR appeared, by first comparing orthologous loci between rat and mouse (one identical initial state for two independently evolving loci) and then analyzing the relationship between the age of RMER10 copies, as assessed by their level of divergence from the ancestral sequence and the association with STR.

Analysis of orthologous loci

As exemplified by the locus *D1mit29*, an STR can be "transmitted" from one ancestor to its descendants with rearrangements that give the locus a different repeat motif and structure. At this locus, *M. musculus* possesses a $(CA)_n$ while *Rattus norvegicus* possesses a $(CT)_n$ repeat. A simple substitution (A<->T) can not explain this difference. In most species in the genus *Mus*, this locus is a compound STR composed of $(CT)_n$ followed by $(CA)_n$. The appearance of the simple repeat structure found in *M. musculus* was achieved through a deletion of the $(CT)_n$ block (Garza and Desmarais 2000). Moreover, in other murine species that branch phylogenetically between rat and mouse, at least two additional STR repeat motifs/arrays are found: *M. platythrix* possesses a $(GA)_n$ motif adjacent to the compound $(CT)_n(CA)_n$ and in *A. sylvaticus* a $(CTAT)_n$ repeat region is present. This diversity of repeat motifs and structure shows first that this locus is frequently rearranged and, second, clearly demonstrates the difficulty in predicting the sequence or repeat elements of this locus in one species through comparison with even closely related species.

In spite of such potential complex evolution, examination of pairs of orthologous loci can be informative on the state (presence of an STR vs. absence) of the sequence of the common ancestor of the two species. We aligned 201 pairs of orthologous RMER10 loci extracted from the rat and mouse databases (Table 3) and evaluated them for the presence of tandem repeats. In a majority of pairs (55%), no STR was present in either the rat or mouse genome. STR were found in only one species in 31% of the pairs and only 12% of both species' sequences were found to have an STR. Moreover, for these latter, only 11 pairs contain similar STR, that is loci with the same localization of the repeat region in the RMER10 but not necessarily the same basic repeat motif. Such a low proportion of loci that

share the same STR indicates that very few pairs have a common ancestor that contained a STR prior to the mouse/rat divergence, or that if they did, this ancestral STR was eliminated in at least one of the lineages.

Alignment

From the preliminary sequence alignments, it is apparent that the RMER10 elements have been subject to many modifications, especially indels, throughout their entire length. Through comparison with consensus sequences, it is evident that the two ends of deleted regions often consist of short direct repeats of 3-4 nucleotides. A good example of this is the 16bp deletion that is found only in the 10B sub-family and that is bordered on both sides by an identical CTG sequence. Several duplications involving larger sequences (up to 100 bp) were also observed and, again, were frequently flanked by short repetitive stretches.

However these mutations did not affect homogeneously the sequences of the RMER10 and their alignment with the consensus sequences revealed a striking feature: the relative conservation of the 5' section, the LTR U3 region, and the contrasting high degeneration of the 3' section, the LTR R region (Figure 1). This latter region is also the site of most tandem repeats and, even when no STR is present, contains many other substitutions and indels. The boundary between the two differentially conserved regions is well defined and corresponds to the expected junction between the U3 and R regions in a functional LTR, which is the site of origination of transcription.

The high sequence divergence of the R regions made it impossible to confidently align them and the majority of analyses are therefore based on only the 'conserved' 5' U3 region of the sequences. A phylogenetic analysis of rat/mouse orthologs (not shown) based only on these U3 regions showed that all but 8 of 87 orthologous pairs are monophyletic in the topology, which indicates that the alignment is correct and these reduced parts of the RMER10 still contain sufficient phylogenetic information for analysis. Of the eight pairs which do not cluster together, five include at least one sequence with > 20% of the region deleted, according to the alignment, and one pair is probably not formed by orthologous sequences (rat sequence is a 10B where it should be 10A). The sequences determined through amplification with the *D1mit29* primers all cluster together (Figure 2) with the exception of the sequence in *M. cookii*, which is probably paralogous.

Reconstruction and analysis of ancestral sequences

Phylogenetic analysis of 182 sequences found three main groups that correspond primarily to sequences previously assigned to the 10A, 10B and 10AB sub-families, using the diagnostic 16 bp deletion and poly-A signal region (Figure 2). It is notable that this phylogenetic analysis does not use the 16 bp deletion or the poly-A signal region, indicating the robustness of the diagnostic criteria used to differentiate the sub-families of RMER10. The RMER10B and 10AB sub-families are less differentiated from each other than from the 10A sub-family and the boundary between them is not well defined by a long internal branch. The most recent common ancestor of all 10A sequences (RMER10A-Anc) was easily identified, whereas those of the 10B and 10AB sub-families were not. For this reason, we choose to use two distinct nodes on either side of the central trifurcation (RMER10-Anc) to represent the ancestors of the 10B (RMER10B-Anc) and 10AB (RMER10AB-Anc) groups. The inferred ancestral 10B sequence is identical to the 10B consensus but the 10A one differs from the 10A consensus by 6 bp.

Structural analysis

Analysis of the *D1mit29* sequences revealed the presence of palindromic sequences prone to the formation of hairpin structures, which could explain some of the indel events observed in the locus for some species (Garza and Desmarais 2000). In addition, retroviral LTRs contain many functional secondary structures (Berkhout 1996; Berkhout and van Wamel 2000). This prompted a search for such structures that could potentially shed light on the evolutionary processes acting on RMER10 elements.

A detailed examination of the U3 region sequence confirms the existence of many palindromic or quasi-palindromic motifs, sometimes interleaved, as well as a mirror sequence in the RMER10 elements (Figure 3A). Three main palindrome domains were detected in all the sequences analysed. We used the program mFOLD to detect all potential secondary structures of both strands of the consensus, ancestral and *D1mit29* sequences and to evaluate the stability of such structures by calculating their free energy (ΔG). In the global folding of the entire U3 region, the hairpins formed by the palindromes of the three domains are embedded in larger structures (Figure 3B) where they generally make up the tips of longer stems. While the D1 structures are relatively constant in their position and shape, the D2 can give rise to shifted variations, sometimes inside a single sequence, because of the existence of several C triplets that can match with the G triplet present at the beginning of the D2 domain.

More striking is the D3 domain that can lead to interleaved alternate structures involving the two strands (Figure 3C) and to the pairing of bases with two or more other nucleotides, perhaps through oscillating hydrogen-bonding interactions (Weitzmann et al. 1997, 1998).

We then extended this analysis to the genomic RMER10 copies, through measure of their global ΔG at 38°C, to assess their potential to adopt secondary structures (Figure 4). The ΔG values for RMER10-containing or flanking sequences do not differ significantly from that of random sequences. However, all the reconstructed or consensus sequences have a ΔG that are among the lowest observed and these sequences could therefore clearly generate structures with high stability. Overall, the ancestral RMER10 sequence had a very high stability, surpassed only by two individual RMER10 fragments and five other genomic sequences. Examination of these five genomic sequences revealed the presence of an LTR retrotransposon in all but one, which instead contained a TGG repeat.

Discussion

Association of RMER10 with tandem repeats

We found a greater rate of association of RMER10 elements with tandem repeats than other genomic fragments of the same size randomly drawn from the genome or drawn from the extended genomic sequences containing the LTR regions. While many of these repetitive regions are not perfect repeat loci, but degenerate imperfect loci, the observed difference is still notable since the same algorithm and parameters were used to assess the presence of tandem repeats in both RMER10 elements and other fragments. Moreover, this association is not due to a preferential insertion of RMER10 in genomic regions that previously contained STR since most of the STR observed are found inside the LTR itself. There are two salient hypotheses that can explain this observation: first, that an ancestral retrovirus bearing an STR in its LTR (a future RMER10 element) spread through the murine genome, but the STR has been lost in some copies (the proliferation hypothesis); second, that the observed prevalence of STR loci is the result of many independent appearances of such repetitive sequences in integrated RMER10 copies (the multiple, independent appearance hypothesis).

Proliferation vs. multiple appearance hypotheses

The presence of a large, variable repetitive region at the beginning of the R region of the LTR of an exogenous retrovirus is unlikely, since this corresponds to the site of

transcription initiation and forms the 5' leader region of the retrovirus RNA. This part of the viral genome is essential for the virus to be dimerized and ultimately packaged in the capsid (Alford et al. 1991; Berkhout 1996; Paillart et al. 1996). It is therefore subject to structural, and hence sequence, constraints that make the presence of STR regions improbable. However, such repetitive regions can be present in an endogenous retrovirus that does not need packaging or may appear *de novo* either during reverse transcription of a provirus, thus producing an inactive genomic copy, or in an already inactive genomic LTR that, for example, results from excision of the rest of the proviral genome.

In the latter cases, the proliferation hypothesis can be rejected since such a defective provirus or LTR cannot by itself be replicated and retrotransposed to another genomic location, except in the case of a functional retrovirus capturing a defective one in its capsid. This situation is not likely to occur since experimental work tends to show that such viral RNA dimers can not be properly processed to complete the virus cycle (Mikkelsen et al. 2000). Moreover, *trans* mobilization (i.e. capture of exogenous RNA by the retrotransposition machinery of an element) is highly disadvantageous for LTR retrotransposons and L1 LINEs (Wei et al. 2001; Dewannieux et al. 2004).

However, the presence of STR in LTR of a functional endogenous retrovirus has been documented, as exemplified by the case of the Intracisternal A Particles (IAP, Christy et al. 1985). Some IAP copies are still active in the mouse genome (Dewannieux et al. 2004), in spite of the presence of short STR in the R region, which is the same location where they are found in RMER10 elements. When the master copy of a transposable element family contains a tandem repeat region, the proliferation scenario can easily explain the association with STR loci, but most of the members of the family should then also contain an STR locus. However, this is not observed for IAP LTRs, as a survey of the mouse genome (not shown) reveals that the family of IAP-related LTR most frequently associated with STR, IAPLTR2, does not contain a tandem repeat in 32% of individual elements. For the RMER10 elements, between 68% to 80% of copies do not have STR loci. Under the proliferation hypothesis, one must therefore invoke an unlikely number of multiple independent losses of STR to reconcile the observed association rates with the expectation that most elements initially had such repeats.

Analysis of orthologous pairs of RMER10 elements from mouse and rat provided additional insight into the question of STR origins. The majority of pairs had no STR present (Table 3), strongly indicating that no repeat region was present in the common ancestral

sequence. In 64 cases, an STR is missing in one of the two loci and in only 25 pairs STR can be found in both species.

The most parsimonious explanation for the presence of an STR in both loci of a pair is that the ancestor possessed the repetitive region, regardless of the time when the ancestor acquired the repeats (i.e. during RNA reverse transcription or later, during DNA replication), but only a small fraction of pairs had an STR present in both species. In addition, most of these pairs do not have the same repeat motif and/or the STR region is not found in the exact same location, suggesting a distinct origin of the repeat region. When an STR is missing in one of the two loci then this means that it has been either lost or acquired once after speciation, although without additional outgroup information these two possibilities can not be distinguished. In either case, this reveals a high degree of instability of the RMER10 DNA sequences as they have undergone many reiterated appearance/loss of STR loci since the separation of the mouse and rat lineages.

To evaluate when in the transposable element life cycle STR appear in RMER10 elements, we compared the age of RMER10 copies with their STR association rate: if STR appears in the RMER10 elements after they are integrated, then old copies should have an association rate higher than young copies, since they have had more time to acquire repetitive regions. If STR are present before integration of RMER10, then the association rate of young copies must be of at least the same order as the rate of older copies. Thus, we approximated the ages of about 1600 full-length RMER10 elements using the homology scores of their U3 region with the consensus sequences. RMER10 copies were then binned into estimated age classes and plotted against the association rate with STR loci (Figure 5). This shows that the older copies of RMER10 are less associated with STR than more recently arisen ones (R^2=0.42, p-value= 0.044). If we consider the 10A and 10B subfamilies separately, this trend is much more pronounced in RMER10A (R^2=0.75, n=1114, p-value= 0.0012) than in RMER10B elements (R^2=0.39, n=418, p-value: 0.1805). This tends to indicate that STR are present before the integration of RMER10, arise during the integration process or possibly early after it, but that they are then lost afterwards, leading to a reduced association rate of older copies.

This analysis suggests a way that STR may be generated in these elements. It may be that STR loci are not present in the retroviral genome, but instead are repeatedly produced during reverse transcription of the proviral messenger. This would explain why there is no absolute association of RMER10 (and IAP) with STR, since mutation is a stochastic process.

However, the similar localization of the tandem repeat regions in the R region of both RMER10 and IAP, two unrelated elements, and the elevated level of STR elimination observed for old copies still require explanation.

One possibility that would explain all of the different observations is that the sequence of RMER10, and also of IAP, contains intrinsic factors that perturb their correct replication and induce repeated mutation at nearly the same place.

Secondary structures as potential mutagenic factors

The presence of special secondary structures in the U3 region of an LTR is not surprising, since such conformations are relatively common and are often associated with biological functions, such as the promotion of transcription (Catasti et al. 1999) or the export of viral RNA from the nucleus to the cytoplasm (K-RRE in the HERV-K virus; Yang et al. 2000). The many potential secondary structures (hairpins and others) found with the RMER10 sequences are likely the remnants of these functional elements. However, they may also provide clues about the processes that have led to the emergence of tandem repeats in the R region of the RMER10 elements. Indeed, strong secondary structures are well recognized as obstacles for both transcription and replication and are easily formed on RNA molecules such as the retroviral transcript. If the reverse transcriptase encounters such a difficulty in the U3 of the LTR, it seems likely that template translocation errors would accumulate and could lead to the formation of repeated nucleotides in the R region. Such errors can happen with DNA templates as well. The formation of stable hairpins, or more complex structures, upstream of the replication fork has been shown to hinder the unwinding of the double helix and cause polymerase arrest (Kang et al. 1995; Weitzmann et al. 1997; Hyrien 2000). In addition, it has been shown that when DNA polymerase pauses or is stalled, this can induce many replication errors or recombination, at least in prokaryotes (Cox et al. 2000; Michel et al. 2001). Even simply slowing down the progression of the replication complex can be sufficient to trigger a rescue mechanism (Nyberg et al. 2002) with low fidelity polymerase (Bebenek et al. 2003; Ramadan et al. 2004).

While abnormal DNA structures can potentially be adopted by either complementary single strand of an RMER10 element when they are separated, it is unclear whether they actually take on these conformations *in vivo*. However, the ratio between the stabilities of the normal double-helix and the aberrant DNA structures provides some insight into this question. In palindromic sequences, the DNA strands can adopt complex secondary

structures, even at physiological conditions, particularly when supercoiled DNA is formed upstream of the replication fork (Schroth and Ho 1995). In some cases, hairpins and tetraplexes can also be formed under physiological conditions (Chen et al. 1995; Weitzmann et al. 1997; Catasti et al. 1999) with sequences particularly rich in GGG/CCC triplets such as in the D3 domain (Figure 3C). In addition, many transcription factor signals are found in close proximity to the sequences that can form these structures. The fixation of such regulatory proteins and the initiation of transcription can provide the energy necessary for the transition from the double-helix conformation to an intra-strand secondary structure by creating negative super-coiled DNA (Wada and Suyama 1986; Wells et al. 1988; Schroth and Ho 1995). In RMER10, the abundance of SP1 sites, involved in the early phase of transcription initiation, may induce the recruitment and fixation of proteins on DNA even if the process aborts and is not followed by actual transcription. Moreover, the analysis of the time of emergence of STR shows that STR appeared either on the RNA or shortly following insertion of the RMER10 element, at which time the transcription factor fixation sites must still have been intact and fully efficient at recruiting protein. It therefore seems likely that at least some palindromic sequences in RMER10 DNA were able to form intra-strand secondary structures under physiological conditions.

If such aberrant secondary structures do form, then they should tend to be eliminated through progressive modification (most frequently truncation) since they are bound to be poorly replicated/repaired and, therefore, only imperfectly transmitted to the next generation. Such a purifying process should leave signatures or traces in the sequences where it occurs. The numerous indels observed when aligning RMER10 elements, including in the relatively conserved U3 region, may be such signatures of past events. For example, the alignment of the 182 nearly complete sequences used in the reconstruction of the ancestral sequences contains 503 nucleotides (Figure 1) whereas the longest consensus sequence is only 210 bp. Most of these indels affect the ability to form hairpins as revealed by their global weighted ΔG that is higher than almost all of the ancestral sequences (Figure 4; see also *M. caroli* deletion in Figure 3). This result is consistent with a general trend towards the elimination from the genome of sequences that may form potentially aberrant DNA structures, as a result of the mutagenic properties of structures such as these that are biased toward the deletion of hairpins. The reconstructed ancestral RMER10 sequences are the most representative of the original LTR and its functional elements and also have a lower ΔG than almost all contemporary RMER10 elements and other genomic fragments (except several that contain

another LTR retrotransposon or G-rich repeated sequence). Moreover, the RMER10 group that has the lowest ΔG, 10B, is also the most strongly associated with STR loci. It also seems to be the youngest because of the greater sequence homogeneity of its members. The degeneration of the secondary structures of the U3 region, inferred by the numerous indels observed there, may therefore be related to the loss of STR by older copies of RMER10.

A possible scenario

The initial seed for the emergence of STR, or short repetitive stretches may be the reverse transcription of an endogenous retroviral RNA as for the still active copies of IAP. The complex secondary structures that can be formed by the RNA in the U3 region combined with potential problems in the re-priming of the reverse transcription to complete the cDNA synthesis, can lead to polymerase stuttering. Once the RMER10 element is integrated, both the binding of transcription factors and the formation of hairpin structures can then interfere with replication or trigger a repair pathway. DNA binding by protein factors can be a direct cause of polymerase arrest (Rothstein et al. 2000), but it may also simply serve as a destabilising agent that allows the palindrome sequences to form hairpins. The intervention of the repair enzyme may then induce a pause of the polymerase complex followed by its dissociation from the DNA strands (Viguera et al. 2001). This can induce a template shift in the R region of the RMER10 elements where the replication fork is stalled. Low complexity repetitive sequence may then appear if not already present, and become the target of SSM mutation. This process acting on an RMER10 that was integrated with an STR seed will generate variation in size between LTRs and copies of RMER10. Afterwards, instability in this genomic region can remain high after the insertion of the LTR until the remnant transcriptional activity and/or hairpin formation are eliminated through deletion.

Although indels in the U3 region can be the result of interference between secondary structures and replication of the RMER10 element, such events might also simply be caused by the presence of tandem repeats, since they can induce bending of DNA or formation of cruciform structures (Wells et al. 1988; Bolshoy et al. 1991; Pearson and Sinden 1998). These aberrant conformations could be propagated to the immediate flanking region, provoking replication errors there. However, the presence of many indels in RMER10 elements that do not contain STRs indicates that this is not a general explanation of such indels. Once tandem repeats have appeared, they may however further contribute to global instability of the region and be involved in large deletions encompassing the flanking areas.

The direct implication of transposable elements in the appearance of STR loci has been previously described for two somewhat exceptional repeat loci, Ms6-hm and Hm2, which are also found in MaLRs (such as RMER10). These loci are among the most variable and largest STR loci in mouse and were initially classified as minisatellites (Kelly et al. 1989). The tandem repeat in Ms6-hm can fold into an even more perfect tetraplex than RMER10 and this structure has been proposed to explain the instability of this mouse locus, through the obstruction of replication (Weitzmann et al. 1998). Two other highly unstable minisatellite loci also emerged from LTR transposons, MSY and MS32. In the well-studied case of MS32, a triplex DNA structure can be formed with a mirror sequence found upstream of the tandem repeat and variant alleles with a disrupted mirror sequence show a dramatic reduction in the instability of the locus (Monckton et al. 1994).

Although there is no sequence homology between IAP-LTR and RMER10 elements, other than for short signal sequences such as the TATA box, there is a striking similarity in the location of STR loci in their R region. A more detailed analysis revealed that IAP-LTR families can also potentially form strong secondary structures with stems and loops, one of them harbouring a SP1 site similar to that found in RMER10 elements. In addition, the still-active IAP-LTR2 family is associated with STR in 68% of the copies examined. These elements may be in a phase where transcriptional activity of the LTR still interferes with genome stability.

For non-LTR retrotransposons, the high frequency of association of repeat sequences with the *Alu* and B1 elements prompts evaluation of the presence of special features, even though secondary structure has not been implicated in repeat evolution for these loci. The mere presence of a poly-A tail is not sufficient to explain the appearance of all STRs, as not all the repeat regions appear in the poly-A stretches (Nadir et al. 1996). Moreover, the emergence of an STR locus requires a series of replication and repair errors; the first is probably a slippage event that can be induced by a pause in the progression of the polymerase (Levinson and Gutman 1987). However, in *Alu* and B1 elements, as in LTRs, either transcription signals and/or potential secondary structures are present. Indeed, these elements appear to have maintained the folding ability of the RNA from which they originated (Sinnett et al. 1991; Labuda et al. 1991; Labuda and Zietkiewicz 1994; Quentin 1994). This indicates that the same scenario as described above may be involved in the emergence of tandem repeats in these SINEs.

We therefore hypothesize that the appearance of new tandem repeat loci is greatly favoured by the combined presence of destabilizing factors and aberrant folding or bending of DNA strands. These factors cause the replication fork to stall and a low complexity sequence is generated (if not already pre-existing) that can then further evolve into a tandem repeat locus. This repeat locus then acquires its own dynamics and can independently generate replication problems that lead to its expansion/contraction.

ACKNOWLEDGMENTS. The authors wish to thank Dr. N Gilbert for helpful discussions about this work.

FIGURE LEGENDS

Figure 1: General organization and comparison of 3 genomic copies each of LTR elements assigned to either the RMER10A or RMER10B sub-families, with the consensus sequences of each sub-family also shown. *D1mit29* is the sequence of the *M. m. domesticus* locus found in the ENSEMBL database. The shading is proportional to the similarity between sequences. The three primary regions of classical LTR elements, U3, R and U5, are present, as are the TATA box and poly adenylation signal. The junction between U3 and R regions (about 15 bp after the TATA) corresponds to the transition between the highly conserved 5' and the highly variable 3' parts of these sequences. The 16bp deletion in the U3 region combined with the sequence surrounding the poly-A signals in the R region can be used diagnostically to distinguish RMER10A and 10B.

Figure 2: Maximum likelihood phylogeny of RMER10 fragments. Phylogeny of the most complete copies of RMER10 elements of the three different sub-families, as well as the *D1mit29* sequences. The unshaded sequence names are those that do not cluster in the sub-family to which they were assigned on the basis of other diagnostic features. The branching position of the consensus sequences are indicated in ovals. The position of internal nodes used as "ancestral" sequences are labelled R10, R10A, R10B and R10AB.

Figure 3: Analysis of the potential secondary structures that can be formed by RMER10 sequences. (A) Alignment of the U3 regions of consensus, ancestral and actual RMER10 (*D1mit29*) sequences. The three main domains (D1, D2, D3) where very stable hairpin structures may form are boxed. Palindromes are presented as inward arrows below the sequences and mirror sequences as outward arrows. Thinner lines are used to indicate a G-T pairing. The putative transcription factor binding sites are symbolized with ellipses (light: SP1 site, dark: Oct-1, NF-1 and TBP from 5' to 3'). (B) mFOLD structure of the entire U3 region of the RMER10-Anc sequence. (C) Potential 2D "bi-cycle" structure of the double-strand region from position 140 to 187 of the alignment (numbering in bp from the beginning of the sequence). Each base on the two strands can be the subject of multiple intra and inter strand pairings. The fixation sites of transcription factors are shaded.

Figure 4: Distribution of Delta (Δ) G values of actual and reconstructed sequences (ΔG weighted by the length of the sequences). Smaller ΔG values indicate greater stability. Row -

1: Randomized sequences from RMER10-Anc; Row 0: from left to right - RMER10-Anc, RMER10B-Anc, RMER10AB-Anc, RMER10A-Anc, RMER10A-cons; Row 1: RMER10 genomic fragments used in figure 3; Rows 2-3: 3'and 5' flanking sequences, respectively, from RMER10 fragments on chromosome 1; Rows 4-5: 3'and 5' flanking sequences, respectively, from RMER10 fragments on the X chromosome.

Figure 5: Comparison of estimated age of RMER10 copies with their observed association rate with STR. Classes of RMER10 copies (all sub-families) were formed according to their ages estimated using their Smith-Waterman homology scores (calculated with RepeatMasker) with their consensus sequences (X-axis, age inversely proportional to score) and then compared with their observed association rate.

Figure 6: Possible scenario for the generation of tandem repeats in RMER10 copies. (A) Linear representation of the genomic fragment with the recognition sites of several transcription factors. (B) The fixation of transcription factors (and/or the supercoiling due to the progression of the replicative fork) destabilizes the double-helix. The single strands can then adopt aberrant conformations that induce a polymerase pause in the R region that then causes replication errors.

TABLES

Table 1: Analysis of the repeat motifs and loci identified

Motif size	**Mono**	**Di**	**Tri**	**Tetra**	**Penta**	**Hexa**	**Hepta**	**Octa**	**Nona**	**Deca**	**Total**
Loci	81	637	146	288	197	142	47	57	29	14	1638
Motifs	4	5	16	25	22	23	11	13	6	4	129
95% homology	61	375	55	173	117	50	9	6	2	3	851
Max length (copies)	56	122.5	68.7	72.5	83	58	27.7	39.1	11.6	25.5	
< 6 copies (truncated)	0	0	0	25 (7)	65 (12)	47 (2)	22 (3)	31 (0)	23 (2)	9 (0)	222 (26)
	1491						147				1638

Table 2. Association rate of the different RMER10 sub-families with tandem repeats.

	RMER10A	RMER10B	RMER10AB	RMER10BA
LTR element	20.2*	32.83	28.5	28.5
Random 425 bp	15.28	15.28	15.28	15.28
5' flanking raw/corrected	21.54/18.8	18.16/15.49	18.48/18.48	ND
3' flanking raw/corrected	20.87/17.8	19.73/16.8	23.50/21.8	ND

* 24.64 for 2147 copies > 300 bp

Table 3. Presence of STR in orthologous pairs of rat and mouse RMER10

No STR	Mouse only	Rat only	Both (different/similar)	Total
112	29	35	25 (14/11)	201

REFERENCES

Alford RL, Honda S, Lawrence CB, Belmont JW (1991) RNA secondary structure analysis of the packaging signal for Moloney murine leukemia virus. Virology 183:611-9

Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) *Alu* repeats: a source for the genesis of primate microsatellites. Genomics 29:136-44

Armour JA, Wong Z, Wilson V, Royle NJ, Jeffreys AJ (1989) Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements. Nucleic Acids Res 17:4925-35

Batzer MA, Deininger PL (2002) *Alu* repeats and human genomic diversity. Nat Rev Genet 3:370-9

Bebenek K, Garcia-Diaz M, Blanco L, Kunkel TA (2003) The frameshift infidelity of human DNA polymerase lambda. Implications for function. J Biol Chem 278:34685-90

Bell GI, Jurka J (1997) The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J Mol Evol 44:414-21

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573-80

Berkhout B (1996) Structure and function of the human immunodeficiency virus leader RNA. Prog Nucleic Acid Res Mol Biol 54:1-34

Berkhout B, van Wamel JL (2000) The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. RNA 6:282-95

Bois P, Williamson J, Brown J, Dubrova YE, Jeffreys AJ (1998) A novel unstable mouse VNTR family expanded from SINE B1 elements. Genomics 49:122-8

Bolshoy A, McNamara P, Harrington RE, Trifonov EN (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. Proc Natl Acad Sci USA 88:2312-6

Catasti P, Chen X, Mariappan SV, Bradbury EM, Gupta G (1999) DNA repeats in the human genome. Genetica 106:15-36

Chen X, Mariappan SV, Catasti P, Ratliff R, Moyzis RK, Laayoun A, Smith SS, Bradbury EM, Gupta G (1995) Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. Proc Natl Acad Sci USA 92:5199-203

Christy RJ, Brown AR, Gourlie BB, Huang RC (1985) Nucleotide sequences of murine intracisternal A-particle gene LTRs have extensive variability within the R region. Nucleic Acids Res 13:289-302

Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E (2003) Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res 31:38-42

Cox MM, Goodman M, Kreuzer KN, Sherratt DJ, Sandler SJ, Marians KJ (2000) The importance of repairing stalled replication forks. Nature 404:37-41

Deininger PL, Batzer MA (1999) *Alu* repeats and human disease. Mol Genet Metab 67:183-93

Deininger PL, Batzer MA (2002) Mammalian retroelements. Genome Res 12:1455-65

Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr. (2003) Mobile elements and mammalian genome evolution. Curr Opin Genet Dev 13:651-8

Desmarais E, Lanneluc I, Lagnel J (1998) Direct amplification of length polymorphisms (DALP), or how to get and characterize new genetic markers in many species. Nucleic Acids Res 26:1458-65

Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T (2004) Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat Genet 36:534-9

Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13:2242-51

Garza JC, Desmarais E (2000) Derivation of a simple microsatellite locus from a compound ancestor in the genus *Mus*. Mamm Genome 11:1117-22.

Gastier JM, Pulido JC, Sunden S, Brody T, Buetow KH, Murray JC, Weber JL, Hudson TJ, Sheffield VC, Duyk GM (1995) Survey of trinucleotide repeats in the human genome: assessment of their utility as genetic markers. Hum Mol Genet 4:1829-36

Grabe N (2002) AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol 2:S1-15

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M (2002) The Ensembl genome database project. Nucleic Acids Res 30:38-41

Hyrien O (2000) Mechanisms and consequences of replication fork arrest. Biochimie 82:5-17

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462-7

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. Computers and Chemistry 20:119-122

Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 40:120-6

Kang S, Ohshima K, Shimizu M, Amirhaeri S, Wells RD (1995) Pausing of DNA synthesis *in vitro* at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. J Biol Chem 270:27014-21

Kaukinen J, Varvio SL (1992) Artiodactyl retroposons: association with microsatellites and use in SINEmorph detection by PCR. Nucleic Acids Res 20:2955-8

Kazazian HH, Jr. (1998) Mobile elements and disease. Curr Opin Genet Dev 8:343-50

Kelly R, Bulfield G, Collick A, Gibbs M, Jeffreys AJ (1989) Characterization of a highly unstable mouse minisatellite locus: evidence for somatic mutation during early development. Genomics 5:844-56

Kelly RG (1994) Similar origins of two mouse minisatellites within transposon-like LTRs. Genomics 24:509-15

Labuda D, Sinnett D, Richer C, Deragon JM, Striker G (1991) Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol 32:405-14

Labuda D, Zietkiewicz E (1994) Evolution of secondary structure in the family of 7SL-like RNAs. J Mol Evol 39:506-18.

Levinson G, Gutman GA (1987) Slipped-strand mispairing : a major mechanism for DNA sequence evolution. Molecular Biology and Evolution 4:203-221

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374-8

Matzura O, Wennborg A (1996) RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. Comput Appl Biosci 12:247-9

Messier W, Li SH, Stewart CB (1996) The birth of microsatellites [letter]. Nature 381:483

Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, Bidnenko V (2001) Rescue of arrested replication forks by homologous recombination. Proc Natl Acad Sci USA 98:8181-8

Mikkelsen JG, Lund AH, Duch M, Pedersen FS (2000) Mutations of the kissing-loop dimerization sequence influence the site specificity of murine leukemia virus recombination in vivo. J Virol 74:600-10

Monckton DG, Neumann R, Guram T, Fretwell N, Tamaki K, MacLeod A, Jeffreys AJ (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nat Genet 8:162-70

Mul YM, Verrijzer CP, van der Vliet PC (1990) Transcription factors NFI and NFIII/oct-1 function independently, employing different mechanisms to enhance adenovirus DNA replication. J Virol 64:5510-8

Nadir E, Margalit H, Gallily T, Ben-Sasson SA (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proc Natl Acad Sci USA 93:6470-5

Nyberg KA, Michelson RJ, Putnam CW, Weinert TA (2002) Toward maintaining the genome: DNA damage and replication checkpoints. Annu Rev Genet 36:617-56

Ovchinnikov I, Troxel AB, Swergold GD (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res 11:2050-8.

Paillart JC, Marquet R, Skripkin E, Ehresmann C, Ehresmann B (1996) Dimerization of retroviral genomic RNAs: structural and functional implications. Biochimie 78:639-53

Pearson CE, Sinden RR (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. Curr Opin Struct Biol 8:321-30

Prak ET, Kazazian HH, Jr. (2000) Mobile elements and the human genome. Nat Rev Genet 1:134-44

Quentin Y (1994) Emergence of master sequences in families of retroposons derived from 7sl RNA. Genetica 93:203-15

Ramadan K, Shevelev IV, Maga G, Hubscher U (2004) De novo DNA synthesis by human DNA polymerase lambda, DNA polymerase mu and terminal deoxyribonucleotidyl transferase. J Mol Biol 339:395-404

Rose O, Falush D (1998) A threshold size for microsatellite expansion. Mol Biol Evol 15:613-5.

Rothstein R, Michel B, Gangloff S (2000) Replication fork pausing and recombination or "gimme a break". Genes Dev 14:1-10

Schroth GP, Ho PS (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. Nucleic Acids Res 23:1977-83

Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD (1997) Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol 17:2851-8

Sinnett D, Richer C, Deragon JM, Labuda D (1991) *Alu* RNA secondary structure consists of two independent 7 SL RNA-like folding units. J Biol Chem 266:8675-8

Smit AF (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res 21:1863-72
Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652-6
Viguera E, Canceill D, Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. Embo J 20:2587-95
Wada A, Suyama A (1986) Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog Biophys Mol Biol 47:113-57
Weber JL (1990) Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7:524-30
Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21:1429-39
Weitzmann MN, Woodford KJ, Usdin K (1997) DNA secondary structures and the evolution of hypervariable tandem arrays. J Biol Chem 272:9517-23
Weitzmann MN, Woodford KJ, Usdin K (1998) The mouse Ms6-hm hypervariable microsatellite forms a hairpin and two unusual tetraplexes. J Biol Chem 273:30742-9
Wells RD, Collier DA, Hanvey JC, Shimizu M, Wohlrab F (1988) The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. Faseb J 2:2939-49
Wilder J, Hollocher H (2001) Mobile elements and the genesis of microsatellites in dipterans. Mol Biol Evol 18:384-92.
Yandava CN, Gastier JM, Pulido JC, Brody T, Sheffield V, Murray J, Buetow K, Duyk GM (1997) Characterization of *Alu* repeats that are associated with trinucleotide and tetranucleotide repeat microsatellites. Genome Res 7:716-24
Yang J, Bogerd H, Le SY, Cullen BR (2000) The human endogenous retrovirus K Rev response element coincides with a predicted RNA folding region. RNA-a Publication of the RNA Society 6:1551-1564
Zhang D-X (2004) Lepidopteran microsatellite DNA: redundant but promising. Trends in Ecology & Evolution 19:507-509
Zhu Y, Strassmann JE, Queller DC (2000) Insertions, substitutions, and the origin of microsatellites. Genet Res 76:227-36.
Zuliani G, Hobbs HH (1990) A high frequency of length polymorphisms in repeated sequences adjacent to *Alu* sequences. Am J Hum Genet 46:963-9

Figure 1

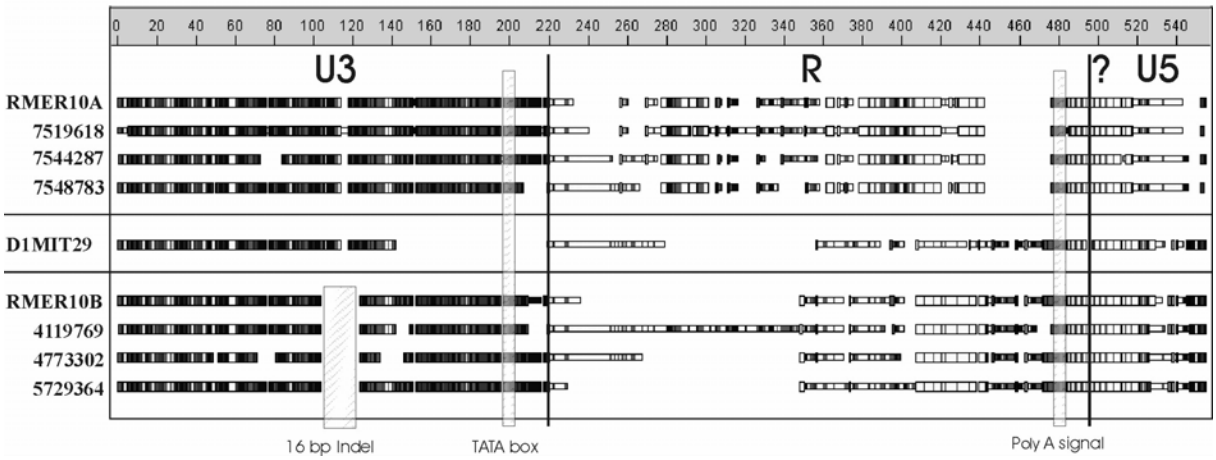

0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480 500 520 540
U3
R
? U5
RMER10A
7519618
7544287
7548783
D1MIT29
RMER10B
4119769
4773302
5729364
16 bp Indel
TATA box
Poly A signal

Figure 2

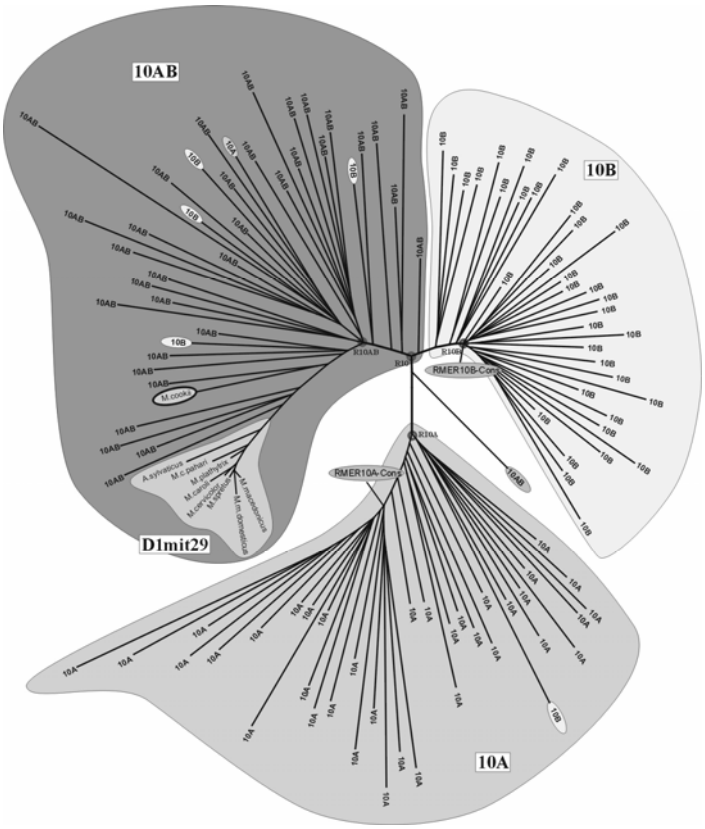

10AB
10B
10A
D1mit29
RMER10A-Cons
RMER10B-Cons
M.cooki
A.sylvaticus
M.c.pahari
M.caroli
M.cervicolor
M.platythrix
M.m.castaneus
M.m.domesticus

Figure 3

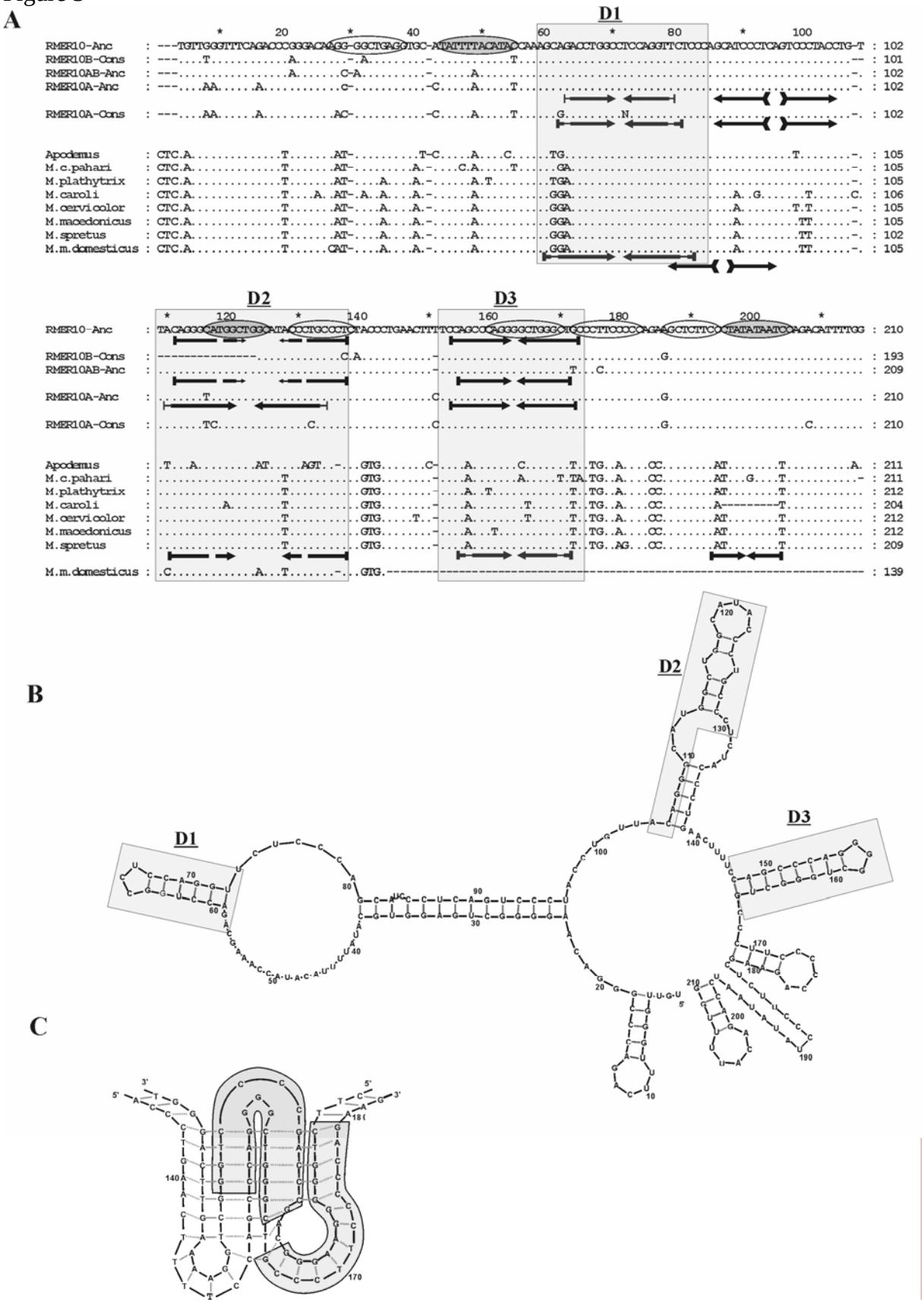

A
D1
D2
D3
RMER10-Anc
RMER10B-Cons
RMER10AB-Anc
RMER10A-Anc
RMER10A-Cons
Apodemus
M.c.pahari
M.plathytrix
M.caroli
M.cervicolor
M.macedonicus
M.spretus
M.m.domesticus
B
C

Figure 4

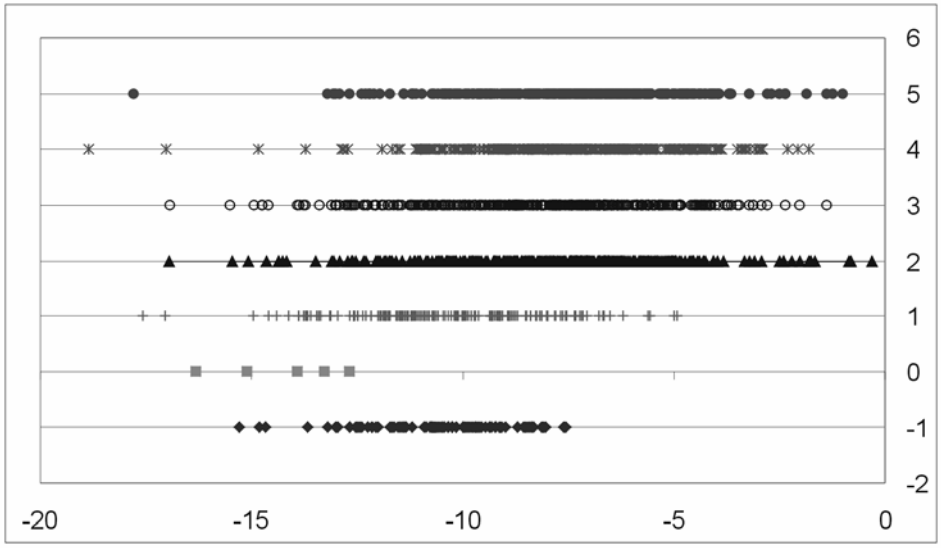
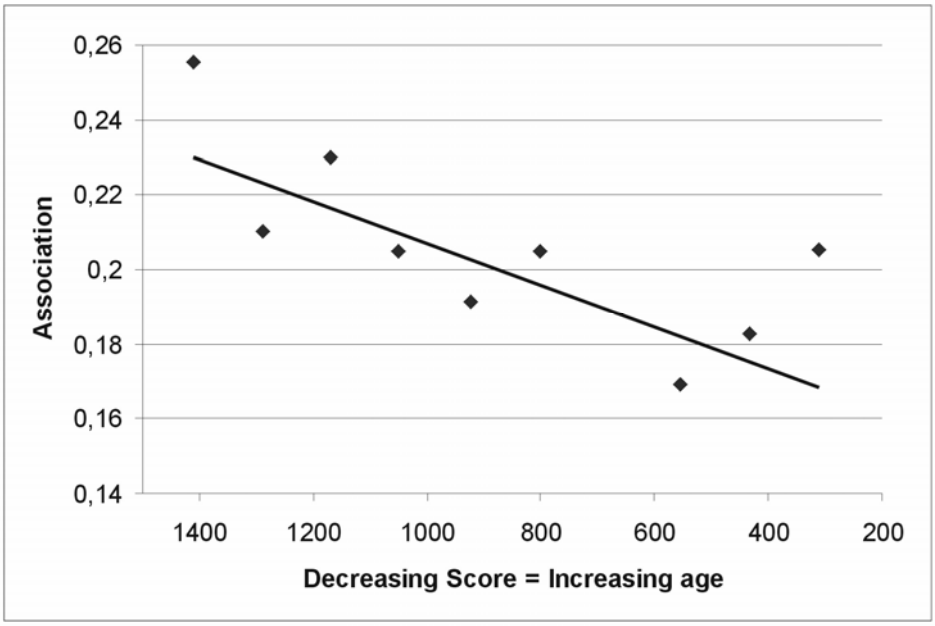


Figure 5


0,26
0,24
0,22
0,2
0,18
0,16
0,14
Association
1400
1200
1000
800
600
400
200
Decreasing Score = Increasing age