Published in final edited form as:

Chromosome Res. 2012 October ; 20(7): 859–874. doi:10.1007/s10577-012-9322-z.

Identification of a group of *Mus dunni* endogenous virus-like endogenous retroviruses from the C57BL/6J mouse genome: proviral genomes, strain distribution, expression characteristics, and genomic integration profile

Kang-Hoon Lee, Ri-Na You¹, David G. Greenhalgh, and Kiho Cho*

Department of Surgery, University of California, Davis and Shriners Hospitals for Children Northern California, Sacramento, CA 95817

¹Division of Electrical Engineering, School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea 305-701

Abstract

About 10 % of the mouse genome is occupied by sequences associated with endogenous retroviruses (ERVs). However, a comprehensive profile of the mouse ERVs and related elements has not been established yet. In this study, we identified a group of ERVs from the mouse genome and characterized their biological properties. Using a custom ERV mining protocol, 191 ERVs (159 loci reported previously and 32 new loci), tentatively named *Mus dunni* endogenous virus (MDEV)-like ERVs (MDL-ERVs), were mapped on the C57BL/6J mouse genome. Seven of them retained putative full coding potentials for three retroviral polypeptides (*gag*, *pol*, and *env*). Among the 57 mouse strains examined, all but the *Mus pahari/Ei* strain had PCR amplicons corresponding to a conserved MDL-ERV region. Interestingly, the *Mus caroli/Ei* amplicon was somewhat larger than the others, coinciding with a substantial phylogenetic distance between the MDL-ERV populations of *Mus caroli/Ei* and C57BL/6J strains. MDLERVs were highly expressed in the lung, spleen, and thymus of C57BL/6J mice compared to the brain, heart, kidney, and liver. Seven MDL-ERVs were mapped in the introns of six annotated genes. Of interest, some MDL-ERVs were mapped periodically on three clusters in the chromosome X. The finding that these MDL-ERVs were one of several types of retroelements, which form mosaic-repeat units of tandem arrays, suggests that the formation of the mosaic-repeat unit preceded the tandem arrangement event. Further studies are warranted to understand the biological roles of MDL-ERVs in both normal and pathologic conditions.

Introduction

A substantial fraction of mammalian genomes are populated with endogenous retrovirus (ERV) sequences, which are “fossils” of ancient colonization of the germline genome by exogenous retroviruses (Rowe and Pincus 1972; Lander et al. 2001; Waterston et al. 2002). The ERV sequences are reported to constitute ~10 % of the mouse genome, and murine ERVs were classified into three distinctive classes primarily based on the phylogenetic relatedness of their reverse transcriptase domains (McCarthy and McDonald 2004). The Class I ERV sequences, which include murine leukemia virus (MLV)-type ERVs, murine retrovirus-related DNA sequences (MuRRSs), GLNs (associated with “glutamine tRNA

*Corresponding author, Kiho Cho, DVM, PhD, Department of Surgery, University of California, Davis and Shriners Hospitals for Children Northern California, 2425 Stockton Blvd., Sacramento, CA 95817, Tel: 916-453-2284, Fax: 916-453-2288, kcho@ucdavis.edu.

primer-binding site”), and retrovirus-like 30 (VL30) elements, make up ~0.7 % of the mouse genome (Lander et al. 2001; Stocking and Kozak 2008). The members of the Class I MLV-type ERV clade are closely related to the gammaretrovirus genus of exogenous retroviruses which are represented by type C MLV. The Class II ERV sequences occupy ~3 % of the mouse genome and there are three main clades: mouse mammary tumor virus (MMTV)-type ERVs, early transposons (ETNs), and intracisternal A-type particles (IAPs). The last class (Class III) is composed of murine endogenous retrovirus (MuERV)-Ls and mammalian apparent long terminal repeat retrotransposons (MaLRs), and its members comprise ~5.4 % of the mouse genome. The clades of the ERV elements (*e.g.*, VL30s, ETNs, and MaLRs), which lack coding potentials for the reverse transcriptase domains, are determined based on each clade’s presumptive ancestral ERV(s) (Stocking and Kozak 2008).

The vast majority of ERVs are presumed to be defective in regard to their coding potential and promoter activity (Tonjes et al. 1999; Belshaw et al. 2005). Previous reports have suggested that ERVs and their gene products participate in a range of pathobiologic processes in humans and animals, such as placental morphogenesis, neuronal degeneration, response to injury, and inflammation (Harris 1998; Christensen 2005; Lee et al. 2007).

Miller *et al.* isolated an endogenous retrovirus from the cells of *Mus dunni* Asian wild mice during a marker rescue assay as a part of human gene transfer experiments, and named it *Mus dunni* endogenous virus (MDEV) (Miller et al. 1996). Although MDEV is determined to be transcriptionally dormant in the *Mus dunni* cells in culture, activation of MDEV by treatment with hydrocortisone or 5-iodo-2'-deoxyuridine results in the production of virus particles, which are capable of infecting a spectrum of cell types derived from different species (Miller et al. 1996; Bonham et al. 1997). It was reported that one to two copies of the MDEV provirus reside in the genome of *Mus dunni* mice, and none of the laboratory mouse genomes carry the MDEV locus (Bonham et al. 1997; Wolgamot et al. 1998). Whereas the sequences of the MDEV’s long terminal repeats (LTRs), harboring six 80-nucleotide repeats in the U3 regions, and non-coding regions are highly homologous to those from the non-autonomous retrotransposons, called VL30 elements, its coding regions share a substantial level of sequence homology with those of gibbon ape leukemia virus (GALV), a member of the MLV clade of Class I (Howk et al. 1978; Wolgamot et al. 1998). Besmer *et al.* initially identified VL30 elements as contaminants of MLV virions with an acquired capability for cell-to-cell transmission while they appeared to be incapable of coding functional proteins essential for encapsulation (Besmer et al. 1979; Adams et al. 1988).

By surveying the DNA sequence database derived from the 129S6/SvEvTac mouse strain, Bromham *et al.* identified multiple copies of a murine type C retrovirus in the *Mus musculus* mouse genome, named *Mus musculus* endogenous retrovirus (MmERV), which is presumed to be a recombinant formed with the VL30 LTRs and MLV coding regions (Simpson et al. 1997; Bromham et al. 2001). In addition, an examination of the expressed sequence tags derived from the *Mus musculus* strains revealed that MmERVs are expressed in a broad range of mouse tissues (Bromham et al. 2001). An alignment analysis with the conserved coding regions of 10 mammalian type C leukemia retroviruses demonstrated that MmERVs share the highest sequence homology (~90 %) with MDEV (Bromham et al. 2001). Recently, a study by Elfaitouri *et al.* identified 300 genomic loci of gammaretrovirus like-ERVs in the mm8 mouse genome assembly; however, those ERVs have not been fully characterized yet (Elfaitouri et al. 2011).

Our previous study identified a unique 40-nucleotide probe which retains very low homology with both MLV-type and MMTV-type ERVs. This probe was used to map a group of ERV loci, which share a high sequence homology with MDEV and MmERVs, within the C57BL/6J mouse genome (Kao et al. 2012; unpublished data). We then

investigated the viral genome structure, strain distribution, and expression characteristics of these ERVs as well as their genomic integration profiles.

Materials and Methods

Four-step *in silico* ERV mining

ERVs were mapped from the C57BL/6J mouse genome using the National Center for Biotechnology Information (NCBI) database (Build 37.2) and a four-step *in silico* mining protocol. For the first step, the mouse genome was surveyed using a 40-nucleotide probe, identified in our previous study (Kao et al. 2012; unpublished data), and the BLASTn program, which resulted in the mining of a defective ERV of 8.623 Kb. Second, a mouse genome survey using the entire 8.623 Kb sequence of the defective ERV as a probe and BLASTn yielded seven ERV loci from which seven intact open reading frames (ORFs) (two *gag*, two *pol*, and three *env*), were retrieved. Third, using the amino acid sequences of the ORFs as a probe and BLASTp, the mouse database was surveyed to identify un-annotated ORFs, which are highly homologous to the individual probes. Among the hits from this survey, two were traced to the ERV loci, including one previously unmapped locus in which the coding potentials for all three polypeptides are intact. Lastly, the entire sequence of the newly identified full-length ERV served as a final probe to identify a group of related ERVs using BLASTn.

Animal experiment

For the MDL-ERV expression analyses, female C57BL/6J mice (~12 weeks old) were purchased from the Jackson Laboratory (Bar Harbor, ME) and housed at the Center for Laboratory Animal Science of the University of California, Davis (UC Davis) according to the guidelines of the National Institutes of Health. The UC Davis Animal Use and Care Administrative Advisory Committee approved the experimental protocol. Three mice were sacrificed by CO₂ inhalation without any other treatment, and normal tissues (brain, heart, kidney, liver, lung, spleen, and thymus) were collected immediately followed by snap-freezing.

Total RNA isolation and RT-PCR

Total RNAs were isolated from the tissues using an RNeasy Mini kit (Qiagen, Valencia, CA). cDNAs were synthesized from 100 ng of total RNA from each tissue using the QuantiTect Reverse Transcription kit (Qiagen). MDL-ERVs were amplified with the primers Udr1A (5'-AGGACCTTTAAT TAC ACTTAT G-3') and Udr2A (5'-CAC CTA GTG GTG TAG AGG AGA-3'), and GAPDH with GAPDH-rt1-For (5'-TGA CCA CAG TCC ATG CCA TC-3') and GAPDH-rt1-Rev (5'-GAC GGA CAC ATT GGG GGT AG-3'). Real-time RT-PCR was performed using the Brilliant SYBR Green QPCR kit from Agilent Technologies (LaJolla, CA), and the results were analyzed using the MxPro (v.4.10) software (Agilent Technologies).

PCR amplification of MDL-ERVs from mouse genomic DNAs

Genomic DNAs from 57 different mouse strains were purchased from the Jackson Laboratory. MDL-ERV sequences were amplified from 100 ng of each genomic DNA with the same primer set as with the RT-PCR analysis using Qiagen Taq polymerase. The PCR conditions were as follows: 95°C for 30 sec, 58 °C for 1 min, and 72 °C for 1 min for 33 cycles.

Cloning of MDL-ERV PCR products and sequencing

MDL-ERV PCR products from the genomic DNAs of the C57BL/6J and *Mus caroli/EiJ* mice were cloned into a pGEM-T Easy vector (Promega, Madison, WI), and sequencing was performed at Functional Biosciences (Madison, WI). Sequences were analyzed using the EditSeq program (DNASTAR, Madison, WI).

Profiling of transcription regulatory elements

The U3 sequences of the individual MDL-ERVs isolated from the C57BL/6J and *Mus caroli/EiJ* mice were surveyed for transcription regulatory elements using MatInspector (v. 8.04) (Genomatix, Munich, Germany). The survey was performed within the vertebrate matrix group and hits (only on the plus strand) with a core similarity higher than 0.95 were selected.

Phylogenetic analyses

Sequence diversity of two different MDL-ERV populations (cloned from the genomic DNAs of the C57BL/6J and *Mus caroli/EiJ* mice) was analyzed using the MEGA4 (Build 4028) program (The Biodesign Institute, Tempe, AZ) (Tamura et al. 2007). The phylogenetic relatedness of MDL-ERV sequences (both cloned from genomic DNAs and identified during the *in silico* survey) was determined using the bootstrap and neighbor-joining method of MEGA4. The following reference sequences were used for the phylogenetic analyses: MuLV (U13766.1), feline leukemia virus (FeLV) (M93134.1), rat leukemia virus (RLV) (M77194.1), MMTV (AF228550.1), MDEV (AF053745.1), GALV (M26927.1), *Phascolarctos cinereus* retrovirus (KoRV) (AF151794.2), porcine endogenous retrovirus (PERV) (HM159246), *Sus scrofa* endogenous retrovirus (SERV) (AF038600.1), GLN-1 (AC136922: 167843-176257), IAPE (AC123738.8: c152862-161181), and MmERV (AC005743: 112341-121005).

Mapping of annotated genes neighboring MDL-ERV loci

The respective regions spanning 100 Kb upstream and downstream of each MDL-ERV locus were surveyed for the annotated genes using the NCBI mouse genome database. In addition, all the intragenic MDL-ERV loci were characterized further in regard to their specific locations in the genes and detailed line maps were drawn for each annotated gene.

Identification of tandem arrays and mosaic-patterned repeat units

To determine the existence of tandem arrays of the candidate repeat units, a two-sequence alignment was performed using the NCBI bl2seq program with varying word sizes (28, 64, and 128) and the results were presented as a dot-matrix plot. When two similar repeat units were compared, both the two-sequence alignment and self-alignment of the individual repeat units were performed with a word size of seven. The profiles/mosaic patterns of the repetitive elements within the repeat units were obtained using RepeatMasker (open-3.3.0) (Institute for Systems Biology, Seattle, WA) (Smit et al. 2004).

Measurement of a relative distance within a population of MDL-ERV sequences

The sequence information relevant to the 27 MDL-ERV loci (~8–9 Kb in size) on chromosome X from different mouse strains was retrieved from the Ensembl database (release 64-Sep 2011) (Flicek et al. 2011). The relative distance within a pool of MDL-ERV sequences was measured using MEGA4 (The Biodesign Institute).

Counting of transition and transversion events

The proviral sequences of 165 MDL-ERVs (~8–9 Kb in size) were obtained from the NCBI C57BL/6J genome database (Build 37.2) and subjected to an alignment analysis followed by counting the transition and transversion events using MEGA4.

Results

Identification of a group of endogenous retroviral loci from the C57BL/6J mouse genome

A four-step *in silico* ERV mining protocol was developed to identify new ERVs from the C57BL/6J mouse genome using a probe sequence which shares a very low homology to the envelope region of MLV-type murine ERVs (MLV-ERVs). A group of 191 ERV loci, which include seven full-length ERVs capable of encoding the intact retroviral polypeptides of *gag*, *pol*, and *env*, were identified throughout the genome except in chromosome Y (Supplementary Tables S1 and S4). Each chromosome had a varying number of loci, ranging from two to 39. Among the 191 ERV loci, 159 loci have been reported in previous studies and 32 are newly identified in this study (Bromham et al. 2001; Elfaitouri et al. 2011). The 159 previously-reported loci are crossreferenced in Supplementary Table S1. Interestingly, 17 of the 32 new ERV loci were mapped on chromosome X (Supplementary Table S4). The phylogenetic tree in Figure 1 was established with the LTR sequences of the 191 ERV loci in which only the chromosomal identification of each locus is indicated. The branching patterns of 12 of the 17 new ERV loci on chromosome X indicate that their LTR sequences are closely related to each other in comparison to the other 179 loci. The vast majority (184 of 191) of the ERVs retained partial or completely defective coding potentials. Initial alignment analyses using the NCBI database demonstrated that these ERVs are highly homologous to MDEV and MmERV, and distantly related to MLV-ERV; primarily due to the lack of a well-defined classification scheme for these ERV loci (both previously reported and new ones from this study), these ERV loci were tentatively called MDEV-like ERVs (MDLERVs), which is a more descriptive name to characterize the data in this report (Wolgamot et al. 1998; Bromham et al. 2001).

Comparison of seven putative full-length MDL-ERVs

An alignment analysis of the seven putative full-length MDL-ERVs identified a number of mismatches throughout their proviral sequences (Figure 2A). The two LTRs within each MDL-ERV were almost identical, suggesting that they were not recombinants of two ERVs with different LTRs. Four of them (MDL-ERV₄₋₈, MDL-ERV_{X-30}, MDL-ERV₁₈₋₄, and MDL-ERV₁₁₋₁) had the tRNA^{Pro} primer binding site (PBS), while the others (MDL-ERV₈₋₃, MDL-ERV₂₋₂, and MDL-ERV_{X-37}) shared the tRNA^{Gly} PBS (Figure 2A). In addition, a region containing variable copies of a 25~30-nucleotide repeat unit was present between the 3'-end of the 5'-LTR and the start of the *gag* coding sequence. A pair of well-conserved splicing signals (donor and acceptor) for the generation of the *env* transcripts of these MDL-ERVs was identified. Phylogenetic analyses of the seven putative full-length MDL-ERVs, in comparison to 12 reference retroviral sequences (both proviruses and polypeptides) derived from various species, revealed that these MDL-ERVs share a unique branch with MmERV which is located immediately next to the MDEV branch (Figure 2B).

Genomic presence of the MDL-ERV loci in the laboratory- and wild-derived mouse strains

Individual species or strains of mouse have different types and/or copy numbers of ERVs in their genomes (Benit et al. 1999). To determine whether the MDL-ERV loci are present in the genomes of 57 laboratory- and wild-derived mouse strains, a primer set was designed from a conserved region in the seven full-length MDL-ERVs, spanning from the 3'-end of the *env* coding sequence to the 3' U3 sequence. It needs to be noted that this primer set is

capable of amplifying the other MDL-ERVs. All but the *Mus pahari/Ei* strain had a PCR amplicon, presumed to be derived from the MDL-ERV loci (Figure 3A). Interestingly, the PCR amplicon from the *Mus caroli/EiJ* strain was determined to be somewhat larger than the expected MDLERV region. The branching patterns of the *Mus caroli* and *Mus pahari* species on the phylogenetic tree indicate that they are evolutionarily more ancient than the others examined in this study (Boursot et al. 1993; Suzuki et al. 2004).

Since there was a size difference in the MDL-ERV PCR amplicons between *Mus caroli/EiJ* and the rest of mouse strains, we then compared the putative MDL-ERV populations in the respective PCR amplicons of the *Mus caroli/EiJ* and C57BL/6J strains. Thirty one unique clones were isolated and sequenced from each set of the MDL-ERV amplicons from the C57BL/6J and *Mus caroli/EiJ* strains. A phylogenetic analysis of all 62 MDL-ERV clones established four major branches, two from each strain, indicating each branch was formed with the sequences derived only from either the C57BL/6J or *Mus caroli/EiJ* strain (Figure 3B). One of the two *Mus caroli/EiJ* branches had the MDEV reference and seven other sequences distantly related to the rest of the *Mus caroli/EiJ* sequences. On the other hand, the MmERV reference was found in a branch consisting of nine MDL-ERVs derived from the C57BL/6J strain. In addition, the diversity level of the MDL-ERV population within each strain was determined by a pair-wise comparison at each nucleotide position in all 31 clones, and it was found that the MDL-ERV population in the *Mus caroli/EiJ* strain was more diverse than one in the C57BL/6J strain (Figure 3B).

Differential profiles of transcription regulatory elements in four MDL-ERV groups found in the C57BL/6J and *Mus caroli/EiJ* mice

The 5' U3 regions of ERV proviral sequences serve as a promoter for viral expression (Coffin et al. 1997). To investigate the transcription potential of the four MDL-ERV branches/groups (Figure 3B) which were identified from the C57BL/6J and *Mus caroli/EiJ* strains, representative U3 sequences (one from each group) were surveyed to obtain a profile of the transcription regulatory elements (TREs) (Supplementary Table S2). Some TREs, such as the binding site for signal transducer and activator, were shared by all four groups plus the MDEV U3 sequence, whereas others were present (*e.g.*, nuclear factor kappa B in *Mus caroli/EiJ*) or absent (*e.g.*, octamer binding protein in C57BL/6J) only in a specific strain or in MDEV. The findings from this survey suggest that the four putative MDL-ERV groups retain different transcriptional potentials.

Tissue-specific expression of MDL-ERVs in C57BL/6J mice

A previous study reported that MDEV is not expressed in cultured *Mus dunni* cells (Miller et al. 1996). However, it has been reported that some other MuERV families, such as MLV type-ERVs and MMTV type-ERVs, were expressed in mice under both normal and stress conditions in a tissue specific manner (Lee et al. 2007; Cho et al. 2008; Lee et al. 2011). In this experiment, the expression of MDL-ERVs was examined in seven different tissues (brain, heart, kidney, liver, lung, spleen, and thymus) of 12 week-old C57BL/6J female mice. The seven tissues were selected primarily to test whether there are tissue-specific expression patterns of MDL-ERVs, if any. The results from real-time PCR analyses revealed that the expression of MDL-ERVs is tissue-specific; there was substantially higher expression in the lung, spleen, and thymus compared to the brain, heart, kidney, and liver (Figure 4).

Mapping of annotated genes neighboring the MDL-ERV loci on the C57BL/6J mouse genome

To determine whether the MDL-ERV loci are near coding genes, the sequence 100 Kb upstream and downstream of each MDL-ERV locus was surveyed for any annotated genes

using the NCBI C57BL/6J mouse genome database. A total of 218 annotated genes were mapped with 80 of the 191 MDL-ERV loci (Supplementary Table S3). Seven of the 80 MDL-ERV loci, including two full-length MDL-ERVs, were integrated into the introns of genes, including the variable region of T cell receptor- α cluster (Figure 5). One full-length and one defective MDLERV were found in the minus strand of the introns of the Large gene locus. Further investigation is needed to understand the potential interaction between the MDL-ERVs and neighboring genes in regard to their transcriptional activities.

Identification of tandem arrays of mosaics consisting of MDL-ERVs and other repetitive elements (REs)

It was observed that there are three distinct clusters of MDL-ERVs in chromosome X (Supplementary Table S1 and Figure 6A). The first cluster (cluster-A) of approximately 1.97 Mb (3,009,213 – 4,982,843) consisted of nine MDL-ERV loci, with each locus ranging from 6,028 bp to 8,846 bp (MDL-ERV_{X-1} through MDL-ERV_{X-9}). Seven MDL-ERV loci of 8,746 bp ~8,815 bp in size (MDL-ERV_{X-12} through MDL-ERV_{X-18}) formed the second cluster (cluster-B) (27,064,840 – 30,475,289) of approximately 3.41 Mb. The third cluster (cluster-C) (30,256,273 – 32,551,630) of 2.29 Mb contained eight MDL-ERV loci, with each locus ranging from 6,143 to 12,900 bp (MDL-ERV_{X-19} through MDL-ERV_{X-26}). Within these clusters, especially in cluster-B, it was evident that MDL-ERVs appear periodically (Figure 6). This finding led us to speculate that these MDL-ERVs constitute a portion of the large repeat unit in tandem arrays. Initially, we defined the entire sequence from the 5'-end of MDL-ERV_{X-14} to one nucleotide upstream of MDL-ERV_{X-15} (28,072,807 – 28,421,386) as a reference repeat unit of cluster-B (UB1). To visually determine and confirm whether cluster-B is organized into a tandem array of the reference unit UB1, a dot-matrix plot was drawn for the entire cluster-B sequence using the reference unit UB1 sequence as a probe. This cluster-B dot-matrix plot, which was drawn with 64 words, revealed a tandem array of UB1 (Figure 6A). Another dot-matrix analysis with the same two sequences (cluster-B and UB1) using 128 words further confirmed that cluster-B is a tandem array of a ~348.58 Kb repeat unit. Similar to the process used with cluster-B, cluster-A and cluster-C were probed with another putative reference unit (named UA-UC) spanning from the 5'-end of MDL-ERV_{X-6} to one nucleotide upstream of MDL-ERV_{X-7} (3,978,603 – 4,085,746) to determine whether these clusters are organized as a tandem array. The dot-matrix plots of cluster-A and cluster-C revealed evident, although imperfect, footprints of tandem arrays of the reference unit UA-UC.

Within cluster-B, UB1 and an additional ~348.58 Kb repeat unit containing MDL-ERV_{X-15} (UB2) were selected for further analyses in regard to their structural details using dot-matrix plots, which are formed by the self-alignment of each repeat unit and display occurrences of REs in an unbiased manner (Figure 6B). Although the overall dot-matrix plot patterns were similar between the repeat units UB1 and UB2, the regions that were different were examined further. To identify the highly polymorphic regions, a dot-matrix plot was drawn by an alignment of the two repeat units. Instead of a perfect diagonal line, which indicates two identical sequences, three major discontinuous and/or shifted regions were observed (Figure 6B). The dot-matrix plot patterns of these repeat units reflect a dense RE (both characterized and uncharacterized) population. Subsequently, the profile of characterized REs within each repeat unit was identified (Figure 6C). The repeat units UB1 and UB2 were compared by linear mapping of annotated REs in each unit to determine whether they share a similar structural pattern. It turned out that each repeat unit was a mosaic of ERVs, LINEs, SINEs, various tandems, and other REs, and as expected, both repeat units shared a well-conserved mosaic pattern. However, the three discontinuous and/or shifted regions in the dot-matrix plot retained a high level of polymorphisms in their mosaic patterns. In addition, several deletion and/or insertion points were observed when the two mosaic patterns were

compared. These findings suggest that a mosaic of various types of REs, including MDL-ERVs and other retroelements, was initially formed then followed by the generation of tandem arrays and insertional and/or deletional events, which may be linked to DNA break and repair in these regions.

Characteristics of the MDL-ERVs embedded in mosaic repeat units

To determine whether the MDL-ERVs, which were a component of the mosaic-patterned repeat units, are present in chromosomes other than X, the entire proviral sequences of the MDL-ERVs (a total of 165), ~8–9 Kb in size, were subjected to an alignment and phylogenetic analyses (Supplementary Table S4 and Figure 7A). Only the group of MDL-ERVs of ~8–9 Kb was selected for this analysis since various deletional and/or insertional events in the smaller MDL-ERVs may hinder proper calculation of phylogenetic relatedness if included. Eight MDL-ERVs from chromosome X, including the seven (MDL-ERV_{X-12} through MDL-ERV_{X-18}) found in cluster-B, formed a unique branch without any MDL-ERVs from the other chromosomes. Its branching point was the second most distant in the tree. All 27 MDL-ERVs (~8–9 Kb in size; five new and 22 previously annotated ones) from chromosome X were aligned separately to examine their relationships (Supplementary Table S4). They were segregated into two main branches, and similar to the results from the 165 sequence analysis above, the same group of eight MDL-ERVs was also sequestered into a unique downstream branch, which had the furthestmost branch point (Figure 7B-top). It turned out that the eighth locus (MDL-ERV_{X-26}) was also a component of a mosaic-patterned repeat unit, which is highly similar to the repeat units of the tandem array containing MDL-ERV_{X-12} through MDL-ERV_{X-18} (Figure 7B-bottom right panel).

The finding that the eight MDL-ERVs of chromosome X, which are subunits of the mosaic-patterned repeat units, share a high level of sequence homology (greater than 99.7 %) led us to investigate whether the MDL-ERVs residing in these mosaic repeat units are less prone to mutation compared to the other MDL-ERV loci in chromosome X. The Ensembl database was surveyed for the sequence information of all 39 MDL-ERV loci on chromosome X from different mouse strains (Supplementary Table S4). There was somewhat limited sequence information for these loci in the database, and a total of 66 sequences (363 bp-long for each locus: 574 to 936 of MDL-ERV_{XD37} as a reference) were obtained from five different mouse strains. The MDL-ERV sequences from some loci, which were derived only from the C57BL/6J strain, were not analyzed. The diversity potential of the MDL-ERV population at each locus was measured by calculating the relative distances among the MDL-ERV sequences from various mouse strains. Interestingly, different mouse strains share relatively high sequence homology in some MDL-ERV loci (named stable loci), but not in the others (named dynamic loci) (Figure 7C–D). The eight MDL-ERVs (MDL-ERV_{X-12} through MDL-ERV_{X-18}, and MDL-ERV_{X-26}), which were components of the mosaic-patterned repeat units, were among the stable loci.

Missense mutations are more frequently associated with transition (purine to purine and pyrimidine to pyrimidine) than transversion (purine to pyrimidine and vice versa) (Topal and Fresco 1976; Topal and Fresco 1976; Rosenberg et al. 2003). In other words, transversion occurs more rarely than transition. In this experiment, we investigated whether the stable status of the eight MDL-ERV loci within the mosaic-patterned repeat units is associated with their relatively low transition/transversion ratio. For all 157 MDL-ERVs (except the eight MDL-ERVs in the cluster) of ~8–9 Kb in size, the overall transition/transversion bias rate was 3.506, while the bias rate for the eight MDL-ERVs (MDL-ERV_{X-12} through MDL-ERV_{X-18}, and MDL-ERV_{X-26}) was 1.02 (Table 1). This finding that the missense mutational events among the eight MDL-ERVs are less biased toward transition compared to the average of all the other 157 MDL-ERVs may coincide with a relatively low mutation rate at these loci.

Discussion

Using a 40-nucleotide probe and a four-step ERV mining protocol, a group of 191 MDL-ERVs, which include 159 ERV loci reported from previous studies, was identified in this study (Bromham et al. 2001; Elfaitouri et al. 2011). Overall, these MDL-ERVs share a high level of sequence homology with MmERVs and MDEVs (Wolgamot et al. 1998; Bromham et al. 2001). The previous reports that MDEV is not present in the genome of laboratory mouse strains suggest that the MDL-ERVs, which were identified from the C57BL/6J mouse genome, are members of a large family of ERVs including MDEVs and MmERVs. It needs to be noted that the MmERV sequences were cloned *in silico* from the genome database of the 129S6/SvEvTac mouse strain (Bromham et al. 2001). The biological properties of MmERVs, at least one of which retains full coding potential, are largely uncharacterized, while only a few studies have investigated the viral genomic organization, strain distribution, and expression profile of MDEVs (Bonham et al. 1997; Wolgamot and Miller 1999).

A close examination of the dot-matrix plots of the individual repeat units provides some evidence suggesting that the repeat units are a palindrome containing two smaller palindromes. Further structural analyses of these repeat units are needed to explain how the mosaic-patterned palindromes were constructed. Among the eight mosaic-patterned repeat units identified in cluster-B and -C of chromosome X, seven repeat units were arranged into a tandem array with the fifth unit in an opposite orientation to the others. The eighth repeat unit was located ~1,969 Kb downstream of the tandem array. Interestingly, the eight MDL-ERV sequences residing in these repeat units were phylogenetically grouped into two branches: MDL-ERV_{X-12, 14, 16, and 18} and MDL-ERV_{X-13, 15, 17, and 26} (Figure 6). One possible explanation is that two slightly different mosaic-patterned repeat units, each of which contains a MDL-ERV from either group, were formed independently followed by recombination to form a double array (tandemization), which was then subjected to duplication events leading to a tandem array (Figure 8). It is unclear why the eighth repeat unit, which contains MDL-ERV_{X-26} and is located on the same strand as the fifth repeat unit, is distantly separated from the tandem array on cluster-B.

Among the 27 MDL-ERV loci in chromosome X (~8–9 Kb in size; five new and 22 previously annotated ones), three loci (MDL-ERV_{X-1}, MDL-ERV_{X-10}, and MDL-ERV_{X-11}) were more polymorphic among different mouse strains compared to the others, especially the ones in the tandem array (Figure 7D, Supplementary Table S4). The MDL-ERV sequences in the tandem array were well conserved and the overall transition rates for both the purines and pyrimidines of these sequences were substantially lower in comparison to the average rate for the other MDL-ERV loci throughout the mouse genome (Table 1).

Further studies are needed to understand the impact of the expression of MDL-ERVs under both normal and pathologic conditions. It is possible that certain MDL-ERVs exert their biological functions via the production of virus particles and/or specific viral proteins, such as *gag* and/or *env* polypeptides. Additional investigation is warranted to understand the underlying mechanisms which control the differential stability of the MDL-ERV loci in conjunction with the structural configuration of the tandem array of chromosome X.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by grants from Shriners of North America (No. 86800 to KC, No. 84302 to KHL [postdoctoral fellowship]) and the National Institutes of Health (R01 GM071360 to KC).

References

- Adams SE, Rathjen PD, Stanway CA, et al. Complete nucleotide sequence of a mouse VL30 retroelement. *Mol Cell Biol*. 1988; 8:2989–2998. [PubMed: 2850474]
- Belshaw R, Dawson AL, Woolven-Allen J, et al. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERVK(HML2): implications for present-day activity. *J Virol*. 2005; 79:12507–12514. [PubMed: 16160178]
- Benit L, Lallemand JB, Casella JF, Philippe H, Heidmann T. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J Virol*. 1999; 73:3301–3308. [PubMed: 10074184]
- Besmer P, Olshevsky U, Baltimore D, Dolberg D, Fan H. Virus-like 30S RNA in mouse cells. *J Virol*. 1979; 29:1168–1176. [PubMed: 221671]
- Bonham L, Wolgamot G, Miller AD. Molecular cloning of *Mus dunni* endogenous virus: an unusual retrovirus in a new murine viral interference group with a wide host range. *J Virol*. 1997; 71:4663–4670. [PubMed: 9151860]
- Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. The evolution of house mice. *Annual Review of Ecology and Systematics*. 1993:119–152.
- Bromham L, Clark F, McKee JJ. Discovery of a novel murine type C retrovirus by data mining. *J Virol*. 2001; 75:3053–3057. [PubMed: 11222735]
- Cho K, Lee YK, Greenhalgh DG. Endogenous retroviruses in systemic response to stress signals. *Shock*. 2008; 30(2):105–116. [PubMed: 18317406]
- Christensen T. Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. *Rev Med Virol*. 2005; 15:179–211. [PubMed: 15782388]
- Coffin, JM.; Hughes, SH.; Varmus, HE. *Retroviruses*. Cold Spring Harbor Laboratory Press; 1997.
- Elfaitouri A, Shao X, Mattsson Ulfstedt J, et al. Murine gammaretrovirus group G3 was not found in Swedish patients with myalgic encephalomyelitis/chronic fatigue syndrome and fibromyalgia. *PLoS One*. 2011; 6:e24602. [PubMed: 22022360]
- Flicek P, Amode MR, Barrell D, et al. Ensembl 2011. *Nucleic Acids Res*. 2011; 39:D800–D806. [PubMed: 21045057]
- Harris JR. Placental endogenous retrovirus (ERV): structural, functional, and evolutionary significance. *Bioessays*. 1998; 20:307–316. [PubMed: 9619102]
- Howk RS, Troxler DH, Lowy D, Duesberg PH, Scolnick EM. Identification of a 30S RNA with properties of a defective type C virus in murine cells. *J Virol*. 1978; 25:115–123. [PubMed: 202730]
- Kao D, Hsu K, Chiu S, et al. ERE database: A database of genomic maps and biological properties of endogenous retroviral elements in the C57BL/6J mouse genome. *Genomics*. 2012; 100:157–161. [PubMed: 22691267]
- Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409:860–921. [PubMed: 11237011]
- Lee KH, Horiuchi M, Itoh T, Greenhalgh DG, Cho K. Cerebellum-specific and age-dependent expression of an endogenous retrovirus with intact coding potential. *Retrovirology*. 2011; 8:82. [PubMed: 21992658]
- Lee YK, Chew A, Fitzsimon L, et al. Genome-wide changes in expression profile of murine endogenous retroviruses (MuERVs) in distant organs after burn injury. *BMC Genomics*. 2007; 8:440. [PubMed: 18045489]
- McCarthy EM, McDonald JF. Long terminal repeat retrotransposons of *Mus musculus*. *Genome Biol*. 2004; 5:R14. [PubMed: 15003117]
- Miller AD, Bonham L, Alfano J, et al. A novel murine retrovirus identified during testing for helper virus in human gene transfer trials. *J Virol*. 1996; 70:1804–1809. [PubMed: 8627704]

- Rosenberg MS, Subramanian S, Kumar S. Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol.* 2003; 20:988–993. [PubMed: 12716982]
- Rowe WP, Pincus T. Quantitative studies of naturally occurring murine leukemia virus infection of AKR mice. *J Exp Med.* 1972; 135:429–436. [PubMed: 4334098]
- Simpson EM, Linder CC, Sargent EE, et al. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nat Genet.* 1997; 16:19–27. [PubMed: 9140391]
- Smit, A.; Hubley, R.; Green, P. RepeatMasker Open-3.0. 2004. <<http://www.repeatmasker.org>>
- Stocking C, Kozak CA. Murine endogenous retroviruses. *Cell Mol Life Sci.* 2008; 65:3383–3398. [PubMed: 18818872]
- Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. *Mol Phylogenet Evol.* 2004; 33:626–646. [PubMed: 15522792]
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 2007; 24:1596–1599. [PubMed: 17488738]
- Tonjes RR, Czauderna F, Kurth R. Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type. *KJ Virol.* 1999; 73:9187–9195.
- Topal MD, Fresco JR. Base pairing and fidelity in codon-anticodon interaction. *Nature.* 1976a; 263:289–293. [PubMed: 958483]
- Topal MD, Fresco JR. Complementary base pairing and the origin of substitution mutations. *Nature.* 1976b; 263:285–289. [PubMed: 958482]
- Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002; 420:520–562. [PubMed: 12466850]
- Wolgast G, Bonham L, Miller AD. Sequence analysis of *Mus dunni* endogenous virus reveals a hybrid VL30/gibbon ape leukemia virus-like structure and a distinct envelope. *J Virol.* 1998; 72:7459–7466. [PubMed: 9696842]
- Wolgast G, Miller AD. Replication of *Mus dunni* endogenous retrovirus depends on promoter activation followed by enhancer multimerization. *J Virol.* 1999; 73:9803–9809. [PubMed: 10559291]

List of Abbreviations

ERV	endogenous retrovirus
ETn	early transposon
FeLV	feline leukemia virus
GALV	gibbon ape leukemia virus
IAP	intracisternal A-type particle
KoRV	koala retrovirus
LINE	long interspersed nuclear element
LTR	long terminal repeat
MaLR	mammalian apparent long terminal repeat retrotransposon
MDEV	<i>Mus dunni</i> endogenous virus
MDL-ERV	<i>Mus dunni</i> endogenous virus-like endogenous retrovirus
MLV	murine leukemia virus
MmERV	<i>Mus musculus</i> retrovirus
MMTV	mouse mammary tumor virus
MuERV	murine endogenous retrovirus

MuRRS	murine retrovirus-related DNA sequence
NCBI	national center for biotechnology information
ORF	open reading frame
PBS	primer binding site
PERV	porcine endogenous retrovirus
RLV	rat leukemia virus
SERV	<i>Sus scrofa</i> endogenous retrovirus
SINE	short interspersed nuclear element
TRE	transcription regulatory element
UA	unit of cluster-A
UB	unit of cluster-B
UC	unit of cluster-C
VL30	retrovirus-like 30

Figure 1. Phylogenetic relatedness of the LTR sequences of the 191 MDL-ERVs from the C57BL/6J mouse genome

The results from an alignment analysis of the LTR sequences of the 191 MDL-ERVs are presented as a phylogenetic tree. Different shapes, gray shades, and orientations (only the triangles) identify specific chromosomes from which the individual MDL-ERVs are mapped. Thick branch lines identify seven putative full-length MDL-ERVs.

*MmERVs, +new ERVs.

Figure 2. Comparison of seven MDL-ERVs with full coding potential for *gag*, *pol*, and *env* polypeptides

A. Alignment of seven full-length MDL-ERVs and their proviral structures. Polymorphic alignment positions among the MDL-ERVs are indicated with vertical lines. A set of key sequence features are indicated: open reading frame, LTR (long terminal repeat), U3 (unique region 3), U5 (unique region 5), R (repeat), PBS (primer binding site), MA (matrix protein), CA (capsid protein), RT (reverse transcriptase), RNaseH, IN (integrase), SU (envelopesurface protein), TM (envelope-transmembrane protein), PPT (polypurine tract), polyA (polyadenylation signal), SD (splice donor), and SA (splice acceptor). Nucleotide sequences of two PBSs (tRNA^{Pro} and tRNA^{Gly}) and a repeat unit are presented. **B.** Phylogenetic analyses of the seven full-length MDL-ERVs with reference retrovirus sequences. The seven full-length MDL-ERVs, both their proviral sequences and amino acid sequences of three polypeptides (*gag*, *pol*, and *env*), were subjected to phylogenetic analyses with 12 reference retroviruses. The information regarding the references can be found in the Materials and Methods section.

\$watermark-text \$watermark-text \$watermark-text

Figure 3. Presence of MDL-ERV loci in the genomes of various mouse strains, both laboratory- and wild-derived

A. Evidence for the presence of the MDL-ERV sequences in the genomes of 56 mouse strains. A PCR survey for the presence of the MDL-ERV sequences in the genome of 57 mouse strains using a primer set (UDr1A and UDr2A) revealed that all but *Mus pahari/Ei* strain, were positive. *indicates two mouse strains (C57BL/6J and *Mus caroli/EiJ*), which were subjected to sequence analyses. **B.** Phylogenetic relatedness of the MDL-ERV sequences isolated from the C57BL/6J and *Mus caroli/EiJ* mouse strains. A phylogenetic analysis of the MDL-ERV sequences isolated from the C57BL/6J and *Mus caroli/EiJ* mice resulted in four distinct branches (two unique branches from each strain). MDEV was closer

to one of the *Mus caroli/EiJ* branches and MmERV was found in one of the C57BL/6J branches. ^ddiversity value.

\$watermark-text

\$watermark-text

\$watermark-text

Figure 4. Differential expression of MDL-ERVs in C57BL/6J mice in a tissue-specific manner MDL-ERVs were relatively highly expressed in the lung (LU), spleen (SP), and thymus (TH) of the C57BL/6J mice compared to the other tissues (brain [BR], heart [HE], kidney [KI], and liver [LI]) examined.

Figure 5. Annotated genes near the MDL-ERV loci on the C57BL/6J mouse genome
Among the annotated genes mapped near (100 kb upstream and downstream) the MDL-ERV loci, six had one or two MDL-ERVs integrated into their introns. Two MDL-ERVs, including one full-length, were integrated into the minus strand of two different introns of the Large gene. Gray shaded polypeptide regions indicate a full-coding potential for *gag*, *pol*, and/or *env* genes. Vertical lines drawn on each annotated gene represent exons. The transcription start site and orientation of each annotated gene are indicated with a vertical bar and arrow.

Figure 6. Identification of tandem arrays of mosaic-patterned repeat units containing MDL-ERVs

A. Survey for the existence of tandem arrays of repeat units containing MDL-ERVs by dot-matrix plotting. The candidate repeat units, which were identified and defined based on the periodic appearance of MDL-ERVs of two different sizes (~6.2 Kb or ~8.8 Kb) in three clusters (A, B, and C) of chromosome X, were aligned against the respective clusters (cluster-A and -C: candidate unit with an ~6.2 Kb MDL-ERV; cluster-B: candidate unit with an ~8.8 Kb MDL-ERV). Different word sizes (WS28, WS64, or WS128) were used for the alignment analyses. Dotted circles (blue: ~8.8 Kb virus and red: ~6.2 Kb virus) indicate the MDL-ERV positions. Two repeat units of cluster-B, which were subjected to further

comparison analyses, are indicated as UB1 and UB2. **B.** Similar profiles of repetitive elements (REs) residing in the UB1 and UB2 repeat units. The self-alignment dot-matrix plot patterns of the UB1 (left) and UB2 (right) repeat units were similar and appeared to be a palindrome containing two smaller palindromes within it. The middle dot-matrix plot, which was drawn based on an alignment of UB1 and UB2 repeat units, showed they are related but also identified key polymorphic regions (dotted squares with *). **C.** Mosaic-patterned repeat units containing MDL-ERVs and other types of retroelements. Both UB1 and UB2 repeat units shared the mosaic pattern, which consists of MDL-ERVs, SINEs, LINEs, and other REs. The three polymorphic regions (* in panel B) are indicated with red dotted lines outlining opaque images. In addition, potential insertion (*i*) and deletion (Δ) points are indicated in the UB2 mosaic, and the corresponding insertion points (V) are marked in the UB1 mosaic. Numbers on the left side indicate the sequence position (in Kb) within the individual cluster regions.

Figure 7. Characteristics of the MDL-ERVs embedded in mosaic-patterned repeat units
A. Phylogenetic relatedness of the proviral sequences of the 165 MDL-ERVs, which are ~8–9 Kb in size. A total of eight MDL-ERVs (MDL-ERV_{X-12} through MDL-ERV_{X-18}, and MDL-ERV_{X-26}) from chromosome X formed a unique branch. These MDL-ERVs are indicated with a closed circle in panels A, B, and C. **B.** Identification of a repeat unit distant from the tandem array. Phylogenetic analysis of the LTR sequences of the MDL-ERVs (~8–9 Kb in size) from chromosome X showed a unique branch with the same MDL-ERV loci as in panel A. An alignment analysis with the UB1 repeat unit followed by dot-matrix plotting confirmed that the ~350 Kb region containing MDL-ERV_{X-26} is highly similar to the repeat

units of the tandem array in cluster-B. **C.** Genomic locus-specific diversity of the MDL-ERV sequences among different mouse strains. Within the 363 nucleotide-long sequences derived from the individual loci examined (a total of 39), different mouse strains shared high sequence homology in some MDL-ERV loci (highlighted with red [100 %] and purple [not 100 %, but statistically not different from red]), but not in the others (highlighted in blue [statistically different from red]). **D.** The 39 MDL-ERV loci are marked with the corresponding colors (red, purple, or blue) on a chromosome X ideogram (modified from Ensembl database).

\$watermark-text

\$watermark-text

\$watermark-text

Figure 8. Schematic illustration of a hypothetical model describing the formation of a tandem array of mosaic-patterned repeat units

Tandemization of two mosaic-patterned repeat units followed by duplication of the tandemized unit. Two independent mosaic-patterned repeat units, each of which contains a MDL-ERV from either group (red or blue), are tandemized by recombination, resulting in a double array. Subsequently, the double array is subjected to duplication events leading to a tandem array.

Table 1
Patterns of nucleotide substitution: MDL-ERVs in cluster-B versus 157 MDL-ERVs of ~8–9 Kb in size

Transition and transversion rates in the MDL-ERV proviral sequences were measured using MEGA4. Numbers in bold indicate transitional substitutions. Ts (transition), Tv (transversion), k (transition/transversion rate ratio). R represents overall transition/transversion bias, with A representing the number of adenosines, G the number of guanosines, T the number of thymidines, and C the number of cytidines. ($R = [A * G * k_{\text{purines}} + T * C * k_{\text{pyrimidines}}] / [(A + G) * (T + C)]$).

MDL-ERVs	Cluster-B				Non-Cluster-B			
Nucleotide	A	G	T	C	A	G	T	C
A	-	9.39	5.8	6.32	-	19.95	2.69	2.9
G	10.08	-	5.8	6.32	20.13	-	2.69	2.9
T	6.45	6.01	-	16.35	2.8	2.77	-	19.52
C	6.45	6.01	15.01	-	2.8	2.77	18.11	-
Bias	Purine		Pyrimidine		Purine		Pyrimidine	
Ts/Tv	1.56		2.59		7.2		6.74	
R	1.02				3.51			