GENOME-WIDE EXPRESSION PROFILES OF ENDOGENOUS RETROVIRUSES IN LYMPHOID TISSUES AND THEIR BIOLOGICAL PROPERTIES

Young-Kwan Lee[a], **Alex Chew**[a], **Ho Phan**, **David G. Greenhalgh**, and **Kiho Cho**
Burn Research, Shriners Hospitals for Children and Department of Surgery, University of California, Davis, Sacramento, CA 95817

Abstract

Endogenous retroviruses (ERVs) constitute approximately 8−10% of the human and mouse genome. Some autoimmune diseases are attributed to the altered expression of ERVs. In this study, we examined the ERV expression profiles in lymphoid tissues and analyzed their biological properties. Tissues (spleen, thymus, and lymph nodes [axillary, inguinal, and mesenteric]) from C57BL/6J mice were analyzed for differential murine ERV (MuERV) expression by RT-PCR examination of polymorphic U3 sequences. Each tissue had a unique profile of MuERV expression. A genomic map identifying 60 putative MuERVs was established using 22 unique U3s as probes and their biological properties (primer binding site, coding potential, transcription regulatory element, tropism, recombination event, and integration age) were characterized. Interestingly, 12 putative MuERVs retained intact coding potentials for all three polypeptides essential for virus assembly and replication. We suggest that MuERV expression is differentially regulated in conjunction with the transcriptional environment of individual lymphoid tissues.

Keywords

endogenous retrovirus; spleen; thymus; lymph node; coding potential; transcription; genome

Introduction

Infection of germline cells with retroviruses leads to their permanent colonization into the germline genome. These germline-integrated retroviruses are transmitted vertically to the offspring in a Mendelian fashion and are called endogenous retroviruses (ERVs), in contrast to exogenous retroviruses, which are acquired from external surroundings and transmitted horizontally. ERVs represent collections of retroviral proviruses introduced into the germline and accumulated throughout an entire set of generations.

ERVs and other forms of transposable repetitive elements make up ~45% of the human and mouse genomes (Urnovitz and Murphy, 1996). Among them, ~8% of the human genome and ~10% of the mouse genome consist of ERVs (Griffiths, 2001; Lander et al., 2001; Waterston

et al., 2002). Although the majority of ERVs are defective primarily due to deletions and point mutations leading to a disruption of the coding potential of the *gag*, *pol*, and/or *env* genes, some are characterized to be biologically active.

Most ERVs have the potential to modulate the expression of host genes near the integration site, for example, the transcriptional control of the amylase gene in the salivary gland (Samuelson, Phillips, and Swanberg, 1996). The HTDV (human teratocarcinoma-derived virus), a member of the human ERV (HERV)-K family, produces type C retroviral particles and retains the coding potential for retroviral polypeptides essential for virus assembly and replication (Knossl, Lower, and Lower, 1999; Lower et al., 1993). The HTDV U3 promoter is highly specific for the transcriptional environment in testicular tumor cells (Lower, Lower, and Kurth, 1996).

Furthermore, the pathogenic processes of certain autoimmune diseases, such as systemic lupus erythematosus, insulin-dependent diabetes mellitus, and multiple sclerosis, have been associated with altered expression of ERVs (Conrad et al., 1997; Deas et al., 1998; Wang et al., 2001). In particular, expression of the HERV-W encoded envelope glycoprotein, called syncytin, was increased in the brain of multiple sclerosis patients (Antony et al., 2004). Syncytin expression in astrocytes was responsible for neuroinflammation, resulting in demyelination and the death of oligodendrocytes. The proinflammatory properties of syncytin demonstrated a unique role of an ERV protein in a pathologic process in humans. In addition, recent studies from our laboratory provide evidence that burn-elicited stress signals can alter the transcriptional activities of certain murine ERVs (MuERVs) in distant organs of mice (Aziz, Hanna, and Jolicoeur, 1989; Cho, Adamson, and Greenhalgh, 2002; Cho and Greenhalgh, 2003). Interestingly, some of these burn-elicited MuERVs are structurally similar to the MAIDS (murine acquired immunodeficiency syndrome)-inducing virus.

It is likely that the unique transcriptional environment in each cell or tissue type directly influences the MuERV expression profile. The genome-wide distribution of MuERVs and their differential expression in various types of tissues and cells may be directly networked to a vast range of signaling events controlling normal physiologic as well as pathologic processes. Two main types of MuERVs, murine leukemia virus (MuLV) and mouse mammary tumor virus (MMTV), are reported to be involved in various pathophysiologic processes involving immune organs (Choi et al., 1987; King and Corley, 1990). In this study, we examined the MuERV expression profiles in various lymphoid tissues of mice and characterized their biological properties.

Results

Differential expression of MuERVs among various lymphoid tissues

We postulate that MuERV expression profiles vary depending on tissue type mainly due to the unique composition of the transcriptional environment within the diverse cell populations comprising each tissue. In this study, the MuERV expression profile in five different lymphoid tissues (spleen, thymus, and lymph nodes [axillary, inguinal, and mesenteric]) were examined in female C57BL/6J mice. The NCBI (National Center for Biotechnology Information) mouse genome database, which is derived from the C57BL/6J strain, was used for comparative analysis. RT-PCR analyses of the MuERV expression profiles were performed by amplifying the polymorphic U3 region within the 3' long terminal repeats (LTRs). Electrophoretic analysis of amplified U3 products revealed a unique MuERV expression profile within each lymphoid tissue examined (Figure 1). There were variations in the length (ranging from ~470 bp to ~750 bp) as well as in the intensity of amplified MuERV U3s. Interestingly, the genomic MuERV profile (size and intensity of amplified U3s) was substantially different from the expression profiles of all lymphoid tissues examined. Only four different MuERV U3

fragments (labeled as I, II, III, and IV), which were present in all five tissues, were selected for a direct comparison of differential expression levels followed by downstream characterization of biological properties. Further investigation into the other MuERV U3 fragments, which were presumed to be derived from certain MuERVs expressed only in some tissues examined, may provide additional information in regard to differential expression of MuERVs in lymphoid tissues.

Multiple alignment and phylogenetic analyses of differentially expressed MuERV U3 sequences

Cloning and sequencing analyses of four different U3 fragments (I, II, III, and IV) from all five lymphoid tissues yielded 43 MuERV U3 sequences which represent at least two sequences from each fragment (Figure 2A). These U3 sequences were subjected to multiple alignment and phylogenetic analyses. Since the amplified U3 regions include ~117 bp upstream and downstream of the exact U3 sequence, these additional nucleotides were trimmed prior to the alignment analyses. The initial multiple alignment analysis yielded 22 unique U3s (24 U3s are shown to include at least one sequence from each fragment) (Figure 2B). The U3 sequences were apparently grouped into six main branches coinciding with the differences in their sizes (ranging from 346 bp to 601 bp) (Figure 2A). In addition, the Th(IV)-1 U3 sequence formed a unique branch, which is consistent with its unique size of 392 bp and absence of a direct repeat (1/1*) (Figure 2B) although it had high sequence similarities with the branches of Th (III)-1 and mLN(IV)-2. Multiple alignment of the 24 unique U3 sequences revealed that both the 5'-end and 3'-end were well-conserved and the middle region, which includes a single 190 bp insertion, was hypervariable (Figure 2B). It has been established that the cellular tropism of MuERVs can be determined by surveying specific sequence features (e.g., direct repeat, unique region) within the U3 promoter (Tomonaga and Coffin, 1998; Tomonaga and Coffin, 1999). Putative cellular tropisms of the unique U3 sequences were determined by the examination of four direct repeat regions (1/1*, 4/4*, 5/5*, and 6/6*), a single 190 bp insertion, and two unique regions (2 and 3) (Figure 2B and Table 1). There were 15 xenotropic and nine polytropic/modified polytropic U3 sequences.

Profile of transcription regulatory elements on individual MuERV U3 promoters.

It is expected that the 3' U3 sequences are almost identical to the 5' U3 sequences, the latter which serve as promoters. To examine the transcription potential of the 22 unique MuERV U3 sequences (3' U3s), the putative transcription regulatory elements were mapped on each U3 (Table 2). Only the transcription regulatory elements with a core similarity (compared to conserved sequences) of greater than 90% were selected. This mapping study yielded 71 transcription regulatory elements, which included binding sites for the glucocorticoid receptor and NF-κB. Although some elements were shared by the majority of the U3 sequences examined, other elements were unique to certain U3 promoters. For instance, a glucocorticoid response element (GRE), binding site for glucocorticoid receptor (highlighted with *), was present only in the iLN(II)-2 U3 promoter. A binding site for signal transducers and activators of transcription 3 (STAT3) (highlighted with *) was mapped only in the Th(IV)-1 U3 promoter. It will be of interest to test whether the putative GRE responds to stimulation by glucocorticoids *in vitro* and *in vivo*. On the other hand, a binding site for NF-κB (highlighted with *), a key transcription factor involved in inflammatory response, was identified in five different U3 promoters (Th(IV)-1, Th(I)-2, iLN(I)-1, Sp(I)-2, and aLN(I)-1). The unique profiles of transcription regulatory elements on each U3 promoter suggest their differential transcription potentials in a given transcriptional environment.

Genomic localization of putative MuERVs using the U3 sequences as a probe and characterization of their biological properties

Genomic localization of putative MuERVs—In order to investigate the genome-wide distribution of MuERVs harboring the U3 sequences identified in various lymphoid tissues, the NCBI mouse genome database was probed with individual U3 sequences. Subsequently, the putative MuERVs with a homology of greater than 98% relative to the U3 probe and approximately 5 kb to 9 kb in size were selected for further analyses (Table 3). The defective MuERVs with substantial deletions in *pol* and *env* genes tend to be approximately 5 kb and the full-length MuERVs are about 9 kb. Genomic probing using the 22 unique U3 sequences revealed 60 defective or full-length MuERVs dispersed throughout the entire genome in both strands except for on chromosomes 17 and Y. Chromosomal location, including proviral size and strand orientation, of these putative MuERVs are summarized in Table 3.

Open reading frame (ORF)—The structures of the putative MuERVs identified in this study were determined by comparison to full-length murine leukemia virus (MuLV) reference sequences (GenBank accession nos. AF033811 and DQ241301). ORF analyses revealed that 12 of the 60 putative MuERVs were full-length and had intact coding potentials for all three retroviral polypeptides (*gag*, *pol*, and *env*) (highlighted in gray, Table 3). Six putative MuERVs (highlighted in yellow) had an insertion of a nucleotide (cytosine) within various poly C regions resulting in a defective *gag* polypeptide due to an insertional frame shift. The locations (nucleotide number from the 5'-end of the provirus) of the insertions within these MuERVs are indicated in parentheses: Th(I)-2a 99% (1,700), Th(I)-2b 98% (1,701), aLN(II)-1a 99% (1,738), Th(II)-1a 100% (1,657), aLN(II)-1p 99% (1,738), and aLN(II)-1h 99% (1,657). On the other hand, putative MuERV Th(I)-2g 99% had a deletion of a cytosine at 1,657 resulting in a frameshift of the *gag* polypeptide. It is possible that the insertion or deletion of a single cytosine in the poly C regions may be attributable to a sequencing error in the NCBI database. In addition, "ATG→GTG" mutations were observed in 26 putative MuERVs (highlighted in blue, Table 3) at the first start codon (ATG) of the reverse transcriptase (RT) region within the *pol* gene, resulting in a loss of three amino acids from the start of the RT protein. The putative MuERV Th(II)-1r 99% contains a green box under the *gag* column because the *gag* ORF was deemed complete, despite the one amino acid deletion within the p12 region.

Primer binding site (PBS)—Analyses of PBSs for all 60 putative MuERVs revealed that 59 proviruses contained a glutamine (Q) tRNA binding site, while the Th(IV)-1a 100% putative MuERV contained a proline (P) tRNA binding site (Table 3).

Recombination event and integration age—Only 6 out of 60 putative MuERVs had LTR mutations (5' LTR sequence compared to 3' LTR sequence) ranging from 0.1349% to 0.2869% (Table 3). The putative iLN(III)-1a 98% MuERV had a deletion of 39 nucleotides in the middle of the 5' LTR in comparison to its 3' counterpart and the mutation rate was not calculated. The integration age of these MuERVs was calculated (ranging from 1.038 million years [MYr] to 2.2072 MYr) based on a formula of "0.13% mutation rate between two flanking LTRs/one MYr" (Sverdlov, 2000). To determine whether there were genetic rearrangements by recombination during the lifespan of the putative MuERVs, we examined the presence or absence of a direct repeat sequence flanking both ends of the proviral sequences. Direct repeats are formed during initial integration of proviruses and any downstream recombination events result in two unique sequences instead. There were 10 putative MuERVs with unique sequences flanking the integration sites, suggesting recombination events, and the rest had direct repeats of four nucleotides (49 putative MuERVs) or 12 nucleotides (one putative MuERV: aLN(II)-1h 99%) (Table 3).

Tropism analysis by restriction fragment length polymorphism (RFLP)—Cellular tropisms of 12 full-length putative MuERVs with intact coding potentials were determined by *in silico* RFLP using three restriction enzymes (*BamH*I, *EcoR*I, and *Hind*III) (Stoye and Coffin, 1988). It revealed six polytropic and six modified polytropic putative MuERVs (Figure 3). It needs to be noted that although the putative tropism traits described in this study are likely to be correct, it may be necessary to determine each MuERV's tropism by an *in vitro* infection study.

Discussion

The findings from this study demonstrated that certain MuERVs were actively transcribed and their expression profiles were specific for each lymphoid tissue examined. Further studies may be necessary to determine the MuERV expression profile in various subsets of cells within each lymphoid tissue. Individual lymphoid tissues and cell types are presumed to have unique characteristics specific to their transcriptional environment, such as transcription factor pool (Schon et al., 2001; van Opijnen et al., 2004). Considering the tissue and cell type-specific transcription factor pool, in conjunction with the unique set of transcription regulatory elements within each MuERV U3 promoter, the findings of differential MuERV expression in various lymphoid tissues in this study are readily predictable. Moreover, the transcription factor profile within each tissue or cell type is affected by its own pathophysiologic status and the surrounding environment, such as systemic immune modulation and carcinogenesis (Boral, Okenquist, and Lenz, 1989).

Due to the polymorphic nature of the MuERV U3 sequences, each U3 promoter is likely to harbor a unique transcription potential, primarily determined by its profile of binding sites for transcription factors. It suggests that changes in the transcription environment due to stress (e.g., injury, infection) will lead to differential expression of certain MuERVs in different tissues. The viral gene products and viral replication itself associated with the altered MuERV expression may participate in a range of pathophysiologic activities in a tissue- and cell type-specific manner.

In silico cloning of the putative MuERVs in the genome of the C57BL/6J strain using the U3 sequences as a probe allowed us to characterize their biological properties such as coding potential, replication competency, and tropism. The results from this study demonstrated that some of the putative MuERVs are full-length with intact ORFs for the *gag*, *pol*, and *env* genes and are presumably replication-competent. Further functional characterization of these putative MuERVs is essential to understanding their roles in normal physiology and pathology of individual lymphoid tissues and cells. In particular, the putative MuERVs retaining intact coding potentials for *gag*, *pol*, and/or *env* may be cloned from the C57BL/6J genome into an expression vector for functional analyses *in vitro* (e.g., cell tropism, viral gene expression, cytokine production, cytotoxicity) and *in vivo* (e.g., infection of virions into mice followed by examination of their effects on the immune system). In addition, the biological properties of individual gene products, primarily *gag* and *env* proteins, from each putative MuERV may be investigated *in vitro* using overexpression as well as knock-out protocols.

Retroviral U3 promoters are capable of controlling the expression of host genes adjacent to the integration sites through certain transcription regulatory elements, such as enhancers and negative regulatory elements (Trusko, Hoffman, and George, 1989; Yu et al., 2005). It will be of interest to investigate whether the U3 promoters of putative MuERVs identified in this study play a role in the transcriptional activities of neighboring host genes.

It may be reasonable to postulate that MuERVs, distributed throughout the genome, play crucial and differential roles in normal physiologic and pathologic processes in various lymphoid

tissues and cells. The specificity of MuERV expression is primarily dependent upon the polymorphic U3 promoter sequences and the specific transcriptional environment in each tissue/cell type. Understanding the biological characteristics of tissue-specific MuERVs and their relationship with neighboring genes will broaden insight into their roles in a range of pathophysiologic events pertaining to individual lymphoid tissues and cells.

Materials and Methods

Animals

Female C57BL/6J mice from Jackson Laboratories (Bar Harbor, ME) were housed according to the guidelines of the National Institutes of Health. The Animal Use and Care Administrative Advisory Committee of the University of California, Davis, approved the experimental protocol. Three mice were sacrificed by cervical dislocation for lymphoid tissue collection without any pretreatments.

RT-PCR analysis

Total RNA isolation and cDNA synthesis were performed based on protocols described previously (Cho et al., 2000). Briefly, total RNA was extracted using a RNeasy kit (Qiagen, Valencia, CA). Total RNA (100 ng) from each tissue sample was subjected to reverse transcription. The sequence of the oligo-dT primer was as follows: 5'-GGC CAC GCG TCG ACT AGT ACT TTT TTT TTT TTT TTT T-3'. Primers, ERV-U1 (5'-CGG GCG ACT CAG TCT ATC GG-3') and ERV-U2 (5'-CAG TAT CAC CAA CTC AAA TC-3') were used to amplify the U3 regions of nonecotropic MuERVs. The primers for *β*-actin amplification were 5'-CCA ACT GGG ACG TGG AA-3' and 5'-GTA GAT GGG CAC AGT GTG GG-3'. The comparability between samples was determined by both the electrophoresis of an equal amount (500 ng) of total RNA and the RT-PCR amplification of *β*-actin from each sample.

Cloning of MuERV U3 sequences

PCR products were gel purified using the QIAquick Gel Extraction kit (Qiagen) and cloned into the pGEMT-Easy vector (Promega, Madison, WI). Plasmid DNAs for sequencing analysis were prepared using a miniplasmid kit from Qiagen. Sequencing was performed at Davis Sequencing Inc. (Davis, CA).

Multiple alignment and phylogenetic tree analysis

The resulting 43 MuERV U3 sequences were aligned using Vector NTI Advance 10 (Invitrogen, Carlsbad, CA) program to identify unique U3 sequences. A phylogentic tree was obtained using the neighbor-joining protocol within the MEGA3 program (Kumar, Tamura, and Nei, 2004; Saitou and Nei, 1987). Bootstrap evaluation of the branching pattern was performed with 100 replications.

Tropism analysis

The putative tropism of unique U3 sequences was determined by comparison to the reference sequences (direct repeats, unique region) first reported by Tomonaga *et al.* (Tomonaga and Coffin, 1998; Tomonaga and Coffin, 1999). A total of four direct repeats (1/1*, 4/4*, 5/5*, and 6/6*), a single 190 bp insertion, and two unique sequences (2 and 3) were utilized for the tropism analysis.

Analysis of transcription regulatory elements

The 22 unique U3 sequences were analyzed for transcription regulatory elements using the MatInspector program (Genomatix, Munich, Germany). The core similarity was set to 0.9 and

the matrix similarity was optimized within the vertebrate matrix group. The U3 sequences are organized in Table 2 to match the order within the multiple alignment and phylogenetic tree.

In silico mapping and cloning of putative MuERVs

Putative MuERV sequences were identified by probing the NCBI mouse genome database using the 22 unique U3 sequences with the NCBI Megablast program. The key reference for identifying viral sequences of interest was a 5−9 kb sequence region flanked by LTRs, which harbor U3 sequences. The following parameters were used with the Megablast program: in the options section "NC_000067:NC_000087" was entered into the "limited by entrez query" field. The filters were set to none while the percent identity, match, mismatch scores were set to 98, 1, and -3, respectively.

MuERV provirus nomenclature

The putative MuERV proviruses were named after the U3 probes/sequences used in Megablast (e.g., Th(I)-2), followed by the percentage homology of the blast hits (e.g., Th(I)-2 99%). The letters following the U3 probe/sequence (e.g., Th(I)-2a 99%) are to differentiate multiple hits found using the U3 probe with similar blast percentage hits.

ORF analysis

The ORFs of each putative MuERV were analyzed using the ORF search feature within Vector NTI (Invitrogen). The parameters were set at a minimum of 50 codons with "ATG" as the start codon, and each candidate ORF was translated. The amino acid translations were then compared using Vector NTI AlignX (Invitrogen) to the following MuLV references retrieved from NCBI (GenBank accession nos. M17327, AY219567.2, AA037285, DQ241301, and AF033811). The criteria for defining the intactness of each proviral gene depended on the p12 region of *gag*, RT (reverse transcriptase) of *pol* and SU (surface domain) of *env*. Proviral genes were deemed intact (+) if the aforementioned sequences were intact and the remaining amino acid sequence of each respective gene matched one of the reference sequences, while allowing for missense mutations. Proviral genes were classified as partial (p) if the defining sequences were intact but the remaining proviral gene sequences were defective. Defective (−) proviral gene sequences contained amino acid deletions and premature stop codons in addition to improper start codons leading to defective defining coding sequences.

Integration age and recombination event

5' and 3' LTR sequences of the putative MuERVs were compared using the Vector NTI AlignX program (Invitrogen). The integration age was calculated based on a formula of "0.13% mutation rate between two flanking LTRs per one MYr". In case there is only a single nucleotide difference between two flanking LTRs, its integration age is recorded as less than the estimated age in consideration of potential error during cloning and sequencing. To examine the genomic rearrangement between MuERVs as well as with other parts of the genome, a short stretch of sequences (4 bp to 12 bp) flanking each MuERV was surveyed for a direct repeat, which is formed during the initial proviral integration.

PBS analysis

A stretch of 18 bp, located immediately downstream of the 5' U5 region, was examined to determine PBSs for the putative MuERVs. The conserved PBS sequences for tRNA[Proline (P)] and tRNA[Glutamine (Q)] were used as references (Harada, Peters, and Dahlberg, 1979; Nikbakht et al., 1985).

RFLP tropism analysis of full-length MuERVs

Tropism of the putative full-length MuERVs was determined by *in silico* RFLP analysis using three restriction enzymes, *BamH*I, *Eco*RI, and *Hin*dIII, using Vector NTI (Invitrogen). The RFLP data of each putative MuERV were compared to the reference profiles for each tropism (ecotropic, xenotropic, polytropic, and modified polytropic) reported previously (Stoye and Coffin, 1988).

Acknowledgement

References

Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C. Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination. Nat Neurosci 2004;7(10):1088–95. [PubMed: 15452578]

Aziz DC, Hanna Z, Jolicoeur P. Severe immunodeficiency disease induced by a defective murine leukaemia virus. Nature 1989;338(6215):505–8. [PubMed: 2538760]

Boral AL, Okenquist SA, Lenz J. Identification of the SL3−3 virus enhancer core as a T-lymphoma cell-specific element. J Virol 1989;63(1):76–84. [PubMed: 2535754]

Cho K, Adamson LK, Greenhalgh DG. Induction of murine AIDS virus-related sequences after burn injury. J Surg Res 2002;104(1):53–62. [PubMed: 11971678]

Cho K, Greenhalgh D. Injury-associated induction of two novel and replication-defective murine retroviral RNAs in the liver of mice. Virus Res 2003;93(2):189–98. [PubMed: 12782367]

Cho K, Zipkin RI, Adamson LK, McMurtry AL, Griffey SM, Greenhalgh DG. Differential regulation of c-jun expression in liver and lung of mice after thermal injury. Shock 2000;14(2):182–6. [PubMed: 10947164]

Choi YW, Henrard D, Lee I, Ross SR. The mouse mammary tumor virus long terminal repeat directs expression in epithelial and lymphoid cells of different tissues in transgenic mice. J Virol 1987;61(10): 3013–9. [PubMed: 3041021]

Conrad B, Weissmahr RN, Boni J, Arcari R, Schupbach J, Mach B. A human endogenous retroviral superantigen as candidate autoimmune gene in type I diabetes. Cell 1997;90(2):303–13. [PubMed: 9244304]

Deas JE, Liu LG, Thompson JJ, Sander DM, Soble SS, Garry RF, Gallaher WR. Reactivity of sera from systemic lupus erythematosus and Sjogren's syndrome patients with peptides derived from human immunodeficiency virus p24 capsid antigen. Clin Diagn Lab Immunol 1998;5(2):181–5. [PubMed: 9521140]

Griffiths DJ. Endogenous retroviruses in the human genome sequence. Genome Biol 2001;2 (6):REVIEWS1017. [PubMed: 11423012]

Harada F, Peters GG, Dahlberg JE. The primer tRNA for Moloney murine leukemia virus DNA synthesis. Nucleotide sequence and aminoacylation of tRNAPro. J Biol Chem 1979;254(21):10979–85. [PubMed: 115865]

King LB, Corley RB. Lipopolysaccharide and dexamethasone induce mouse mammary tumor proviral gene expression and differentiation in B lymphocytes through distinct regulatory pathways. Mol Cell Biol 1990;10(8):4211–20. [PubMed: 2164635]

Knossl M, Lower R, Lower J. Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. J Virol 1999;73(2):1254–61. [PubMed: 9882329]

Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 2004;5(2):150–63. [PubMed: 15260895]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A,

Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowski J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860–921. [PubMed: 11237011]

Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, Kurth R. Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci U S A 1993;90(10):4480–4. [PubMed: 8506289]

Lower R, Lower J, Kurth R. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A 1996;93(11):5177–84. [PubMed: 8643549]

Nikbakht KN, Ou CY, Boone LR, Glover PL, Yang WK. Nucleotide sequence analysis of endogenous murine leukemia virus-related proviral clones reveals primer-binding sites for glutamine tRNA. J Virol 1985;54(3):889–93. [PubMed: 2987543]

Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4(4):406–25. [PubMed: 3447015]

Samuelson LC, Phillips RS, Swanberg LJ. Amylase gene structures in primates: retroposon insertions and promoter evolution. Mol Biol Evol 1996;13(6):767–79. [PubMed: 8754213]

Schon U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mosch C. Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats. Virology 2001;279(1):280–91. [PubMed: 11145909]

Stoye JP, Coffin JM. Polymorphism of murine endogenous proviruses revealed by using virus class-specific oligonucleotide probes. J Virol 1988;62(1):168–75. [PubMed: 2824845]

Sverdlov ED. Retroviruses and primate evolution. Bioessays 2000;22(2):161–71. [PubMed: 10655035]

Tomonaga K, Coffin JM. Structure and distribution of endogenous nonecotropic murine leukemia viruses in wild mice. J Virol 1998;72(10):8289–300. [PubMed: 9733873]

Tomonaga K, Coffin JM. Structures of endogenous nonecotropic murine leukemia virus (MLV) long terminal repeats in wild mice: implication for evolution of MLVs. J Virol 1999;73(5):4327–40. [PubMed: 10196331]

Trusko SP, Hoffman EK, George DL. Transcriptional activation of cKi-ras proto-oncogene resulting from retroviral promoter insertion. Nucleic Acids Res 1989;17(22):9259–65. [PubMed: 2555789]

Urnovitz HB, Murphy WH. Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev 1996;9(1):72–99. [PubMed: 8665478]

van Opijnen T, Jeeninga RE, Boerlijst MC, Pollakis GP, Zetterberg V, Salminen M, Berkhout B. Human immunodeficiency virus type 1 subtypes have a distinct long terminal repeat that determines the replication rate in a host-cell-specific manner. J Virol 2004;78(7):3675–83. [PubMed: 15016888]

Wang Y, Pelisson I, Melana SM, Go V, Holland JF, Pogo BG. MMTV-like env gene sequences in human breast cancer. Arch Virol 2001;146(1):171–80. [PubMed: 11266212]

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420(6915):520–62. [PubMed: 12466850]

Yu X, Zhu X, Pi W, Ling J, Ko L, Takeda Y, Tuan D. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. J Biol Chem 2005;280(42):35184–94. [PubMed: 16105833]
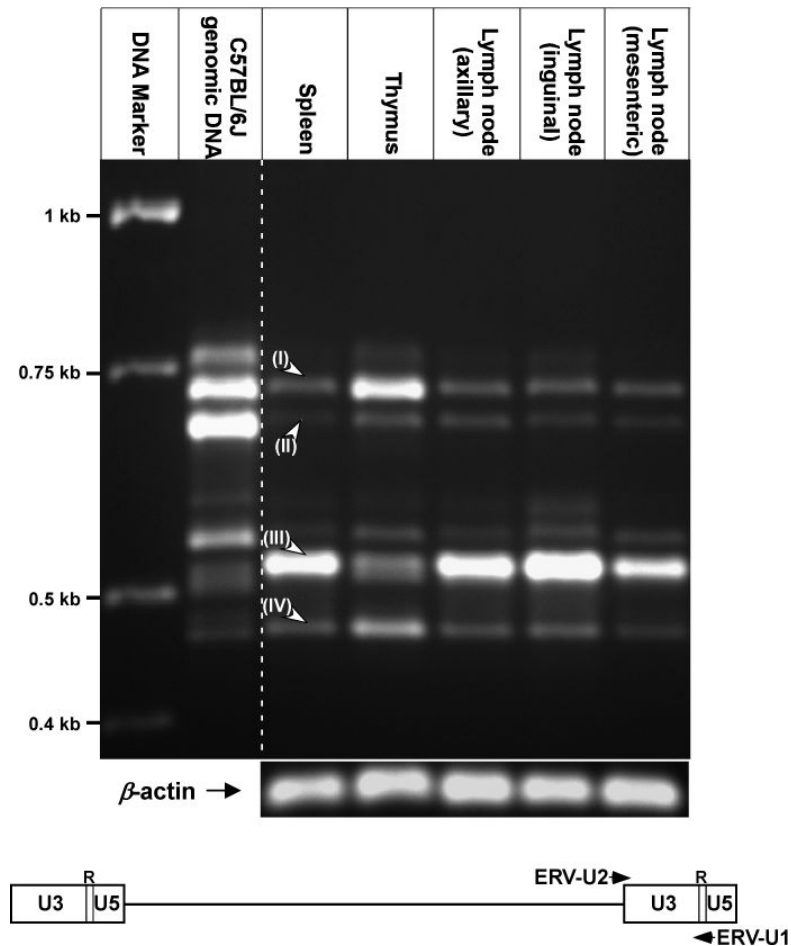
Figure 1.
RT-PCR analyses of the MuERV expression profiles in various lymphoid tissues. Five different lymphoid tissues (spleen, thymus, lymph nodes [axillary, inguinal, and mesenteric]) from C57BL/6J mice were subjected to RT-PCR analysis of MuERV expression using the primer set (ERV-U1 and ERV-U2). Tissue-specific MuERV expression profile is evident. Four different fragments from each tissue, which were subjected to further analyses, are indicated (I, II, III, and IV). The genomic MuERV U3 profile serves as a reference. Bottom panel depicts locations of primers on a schematic MuERV.
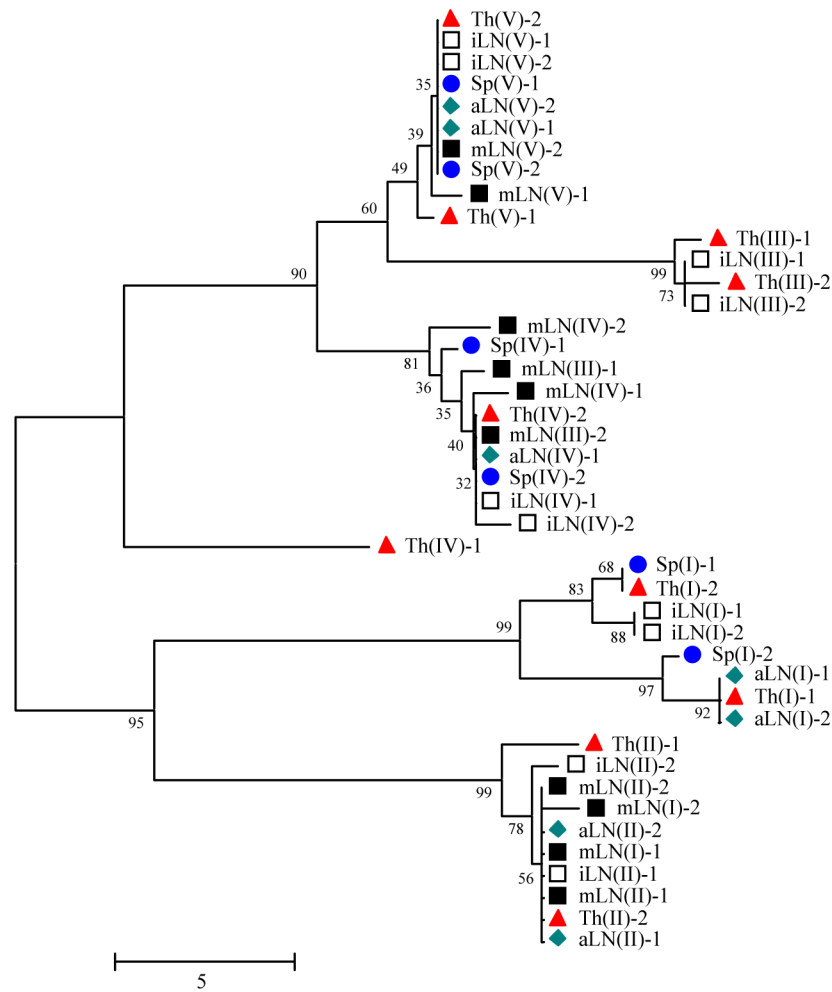
Figure 2.
A. Phylogenetic analysis of differentially expressed MuERV U3 sequences. A phylogenetic tree of differentially expressed MuERV U3 sequences was established. The confidence values at the branch nodes represent the percentage support of a particular branching. Various colored shapes represent different lymphoid tissues. Sp (spleen), Th (thymus), aLN (axillary lymph node), iLN (inguinal lymph node), mLN (mesenteric lymph node). **B.** Multiple alignment analysis of differentially expressed MuERV U3 sequences. Differentially expressed MuERV U3 sequences were subjected to a multiple alignment analysis and yielded 22 unique U3 sequences. Yellow (100 % homology), white (low homology), blue (conserved sequences), and dash (absence of sequences). The locations of the direct repeats (1/1*, 4/4*, 5/5*, and 6/6*), unique regions (2 and 3), and TATA box are indicated in dotted boxes. An insertion of 190 bp is identified in the middle.
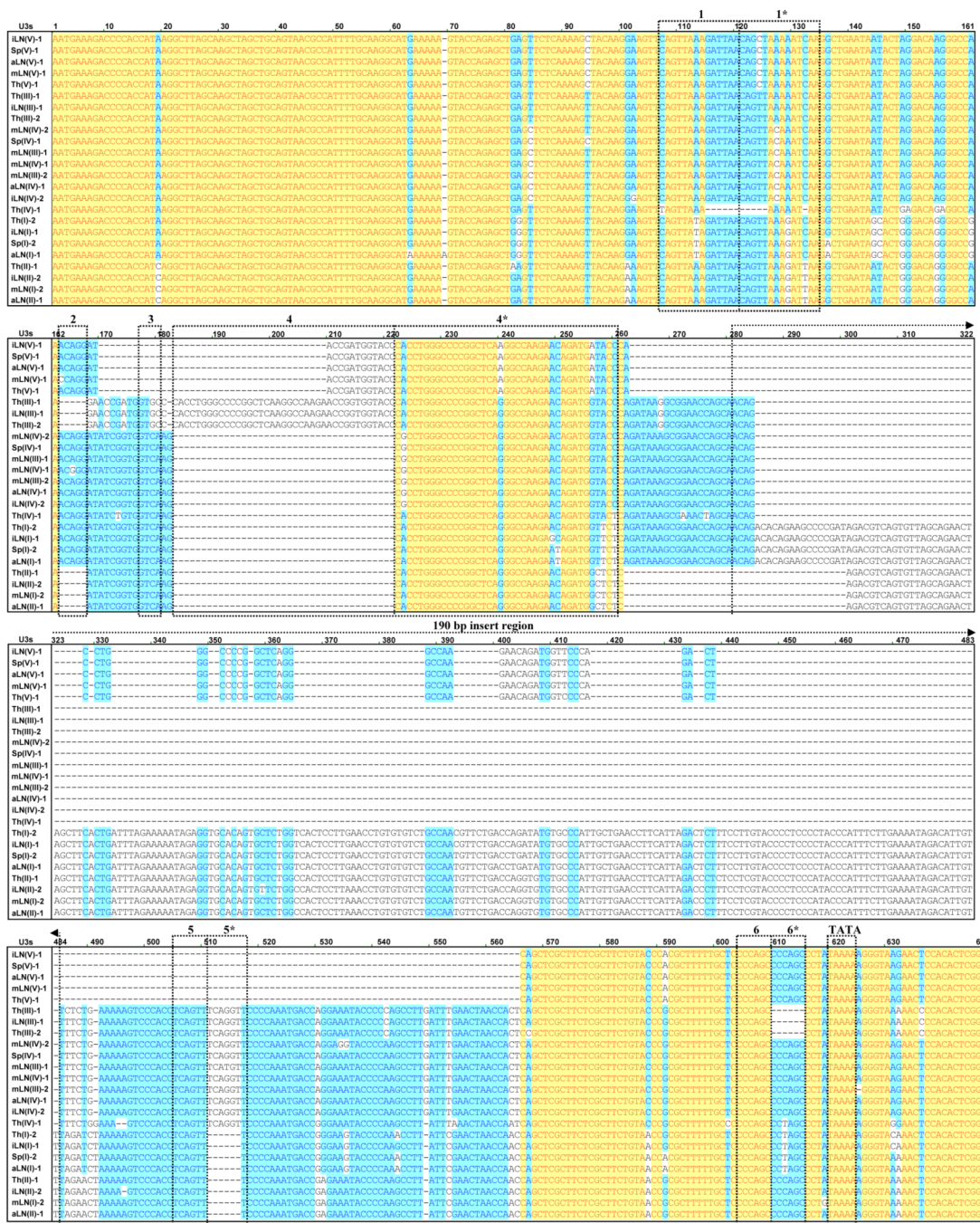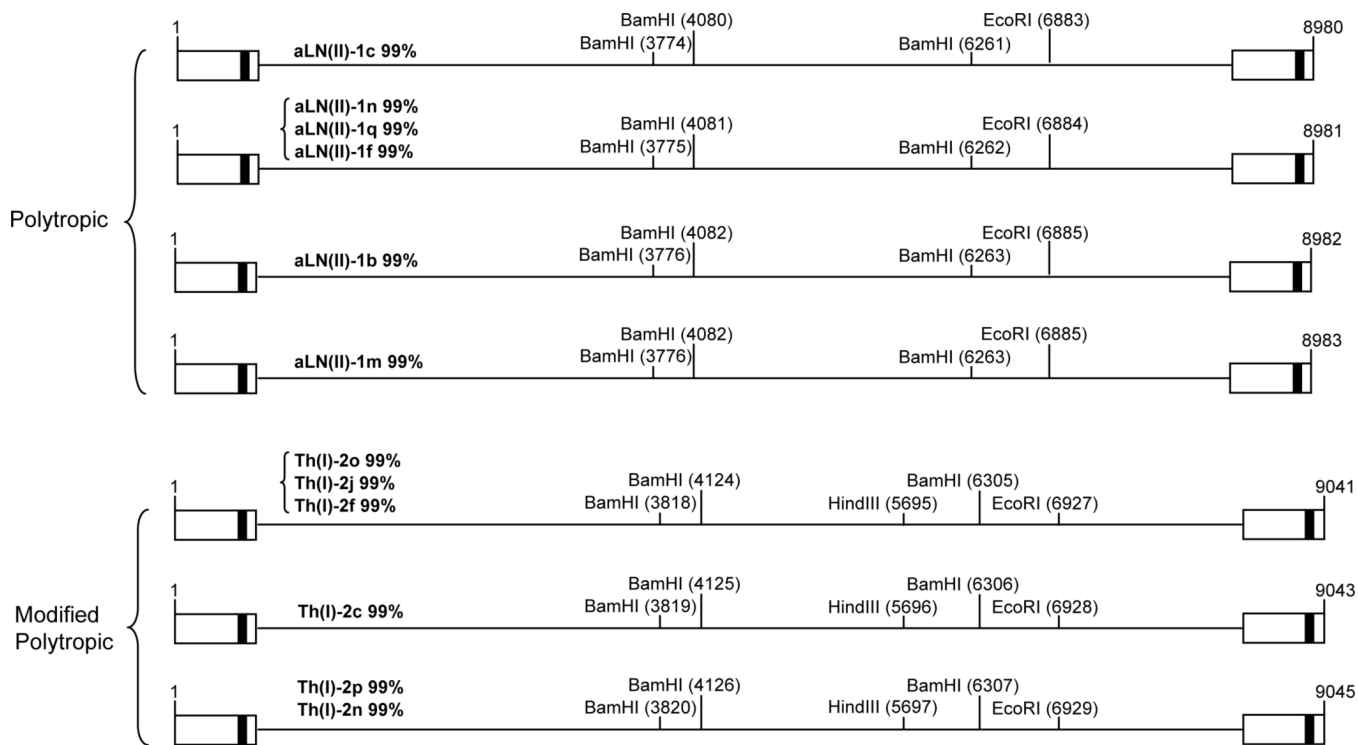
Figure 3.
RFLP tropism analysis of the putative full-length MuERVs with intact coding potentials. Cellular tropisms of 12 putative full-length MuERVs were determined by *in silico* RFLP analysis using three restriction enzymes (*BamH*I, *EcoR*I, and *Hin*dIII). Each restriction site is indicated in parenthesis on schematic MuERVs.

Table 1

Tropism analysis of 22 unique MuERV U3 sequences isolated from five different lymphoid tissues. Dot indicates absence of a homology to reference sequences. Underscore represents the reference type closest to the individual U3 sequences examined. * Direct repeat region.

| U3s	Direct Repeat/Unique Region						Tropism
	1/1*	2	3	4/4*	5/5*	6/6*	
iLN(V)-1	X-II, X-III	X-I,X-II, X-IV, P-II, P-III, P-V	.	X-III	.	X-I, X-II, X-III, X-IV, Poly	X-III
Sp(V)-1	X-II, X-III	X-I,X-II, X-IV, P-II, P-III, P-V	.	X-III	.	X-I, X-II, X-III, X-IV, Poly	X-III
aLN(V)-1	X-II, X-III	X-I,X-II, X-IV, P-II, P-III, P-V	.	X-III	.	X-I, X-II, X-III, X-IV, Poly	X-III
mLN(V)-1	X-II, X-III		.	X-III		X-I, X-II, X-III, X-IV, Poly	X-III
Th(V)-1	X-II, X-III	X-I,X-II, X-IV, P-II, P-III, P-V	.	X-III	.	X-I, X-II, X-III, X-IV, Poly	X-III
Th(III)-1	X-II, X-III	X-III, P-I, P-IV		X-II, X-IV	X-II, X-III, X-IV, P-IV	X-I, X-II, X-III	X-II
iLN(III)-1	X-II, X-III	X-III, P-I, P-IV	.	X-II, X-IV	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III	X-II
Th(III)-2	X-II, X-III	X-III, P-I, P-IV	.	X-II, X-IV	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III	X-II
mLN(IV)-2	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
Sp(IV)-1	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
mLN(III)-1	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
mLN(IV)-1	X-III	.	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
mLN(III)-2	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
aLN(IV)-1	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
iLN(IV)-2	X-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	X-II	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	X-II
Th(IV)-1	.	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	P-IV	X-II, X-III, X-IV, P-I, P-IV	X-I, X-II, X-III, X-IV, Poly	P-IV
Th(I)-2	P-II,P-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	P-II	P-I, P-II, P-III	P-II	P-II
iLN(I)-1	P-II,P-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	P-II	P-I, P-II, P-III	P-II	P-II
Sp(I)-2	P-II,P-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	P-II	P-I, P-II, P-III	P-II	P-II
aLN(I)-1	P-II,P-III	X-I,X-II, X-IV, P-II, P-III, P-V	Xeno/Poly/mPoly	P-II	P-I, P-II, P-III	P-II	P-II
Th(II)-1	P-I	X-III, P-I, P-IV	Xeno/Poly/mPoly	P-I	P-I, P-II, P-III	X-I, X-II, X-III, X-IV, Poly	P-I
iLN(II)-2	P-I	X-III, P-I, P-IV	Xeno/Poly/mPoly	P-I	P-I, P-II, P-III	X-I, X-II, X-III, X-IV, Poly	P-I
mLN(I)-2	P-I	X-III, P-I, P-IV	Xeno/Poly/mPoly	P-I	P-I, P-II, P-III	X-I, X-II, X-III, X-IV, Poly	P-I
aLN(II)-1	P-I	X-III, P-I, P-IV	Xeno/Poly/mPoly	P-I	P-I, P-II, P-III	X-I, X-II, X-III, X-IV, Poly	P-I

Table 2

Transcription regulatory elements in 22 unique MuERV U3 promoter sequences isolated from five different lymphoid tissues. The number in the box indicates frequency of a transcriptional regulatory element within a given U3 promoter sequence, and the gray box indicates no occurrence of the specific transcriptional regulatory element.

Transcription Regulatory Element columns (left to right):

- AML1/CBFA2 Runt domain binding site
- AREB6 (Atp1a1 regulatory element binding factor 6)
- Barbiturate-inducible element
- B-cell-specific activator protein
- CCAAT/enhancer binding protein
- Cell cycle gene homology region (CHR)
- Cell cycle-dependent element, (CDE)
- Cellular and viral TATA box elements
- c-Ets-2 binding site
- c-Myb
- Complex of Lmo2 bound to Tal-1, E2A proteins, and GATA-1, half-site 2
- Downstream regulatory element-antagonist modulator (DREM)
- E2F
- E4BP4, bZIP domain, transcriptional repressor
- Ecotropic viral integration site 1 encoded factor
- Elk-1
- Estrogen related receptor
- Fork head related activator-2 (FOXF2)
- GATA binding factor
- GATA-binding factor 1
- GATA-binding factor 3
- Glucocorticoid receptor*
- Growth factor independence 1
- Gut-enriched Krueppel-like factor
- Heat shock factor 1
- Heat shock factor 2
- Hepatic nuclear factor 1
- HMGAs (HMGA1, HMGA2)
- Hox-1.3 vertebrate homeobox protein
- Hypoxia induced factor-1 (HIF-1)
- Interferon regulatory factor (IRF)-related protein
- Kruppel-like zinc finger protein 219
- LBP-1c (leader-binding protein-1c)
- LYL1-E12 heterodimeric complex
- Mammalian C-type LTR TATA box
- Member of the RSRF protein family
- MOK-2 protein
- Muscle TATA box
- Myelin transcription factor 1-like, neuronal C2HC zinc finger factor 1
- Myocyte-specific enhancer factor 2
- MyT1 zinc finger transcription factor
- Neuron-restrictive silencer factor
- NF-kappaB*
- NMP4 (nuclear matrix protein 4) / CIZ (Cas-interacting zinc finger protein)
- Nuclear factor 1
- Nuclear factor Y (Y-box binding factor)
- Nuclear receptor (Nr2e3)
- Nuclear respiratory factor 1 (NRF1)
- PAR-type chicken vitellogenin promoter-binding protein
- PBX - HOXA9 binding site
- PPAR/RXR heterodimers
- Promyelocytic leukemia zink finger (TF with nine Krueppel-like zink fingers)
- PTF1 binding sites
- Ras-responsive element binding protein 1
- Retinoic acid receptor DR5 site
- Signal transducer and activator of transcription 3*
- Spi-B transcription factor (Spi-1/PU.1 related)
- SF1 steroidogenic factor 1
- Tal-1beta/E47 heterodimer
- Tal-1beta/HEB heterodimer
- Tal-1beta/ITF-2 heterodimer
- TCF11/MafG heterodimers
- Thing1/E47 heterodimer
- Transcriptional repressor (ZNF202)
- v-Myb, AMV v-myb
- v-Myb, variant of AMV v-myb
- Winged helix protein
- Yin and Yang 1 repressor sites
- Zebrafish PAX2 paired domain protein
- Zinc finger transcription factor RU49
- ZNF143

MuERV U3 Promoter rows:

- Sp(V)-1 (346 bp)
- mLN(V)-1 (346 bp)
- Th(V)-1 (346 bp)
- Th(III)-1 (433 bp)
- iLN(III)-1 (433 bp)
- Th(III)-2 (433 bp)
- mLN(IV)-2 (406 bp)
- Sp(IV)-1 (406 bp)
- mLN(III)-1 (406 bp)
- mLN(IV)-1 (406 bp)
- mLN(III)-2 (405 bp)
- aLN(IV)-1 (406 bp)
- iLN(IV)-1 (406 bp)
- Th(IV)-1 (392 bp)
- Th(I)-2 (600 bp)
- iLN(I)-1 (600 bp)
- Sp(I)-2 (600 bp)
- aLN(I)-1 (601 bp)
- Th(I)-1 (556 bp)
- iLN(II)-1 (555 bp)
- mLN(I)-2 (556 bp)
- aLN(II)-1 (556 bp)

Table 3

In silico cloning and biological properties of putative MuERVs differentially expressed in lymphoid tissues. Gray indicates putative putative full-length MuERVs. Two putative MuERVs mapped by only one U3 probe are highlighted in purple. Absence of direct repeat sequences is indicated with "/" between two sequences. Yellow indicates putative MuERVs with a cytosine insertion. Blue indicates putative MuERVs with ATG→GTG mutation at the RT start site. Green indicates that the *gag* was considered intact despite one amino acid deletion within p12. PBS (primer binding site): P (tRNA^{proline}) and Q (tRNA^{glutamine}). ND (not determined). Coding potential: intact (+), partial (P), and defective (−). NA (lacking a proper start codon).

Putative MuERV	Source U3	Ch	Location & Orientation	Size (bp)	PBS	Direct Repeat	Integration Age (Mutation Rate)	Open Reading Frame		
								gag (a.a.)	*pol* (a.a.)	*env* (a.a.)
aLN(II)-1d 100%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	1	NC_000067.4 184093795–184102775 (+)	8981	Q	TTTG	<1.1036(<0.1435%)	(+) 565	(p) 1072	(+) 641
aLN(I)-1d 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	1	NC_000067.4 193634228–193643208 (+)	8981	Q	GTTG	<1.1036(<0.1435%)	(+) 536	(p) 1028	(-) 196
mLN(I)-2 99%	mLN(I)-2, aLN(II)-1, iLN (II)-2, Th(II)-1	1	NC_000067.4 166060482–166066227 (+)	5746	Q	CAAG	<1.1020(0.1433%)	(+) 536	(-) NA	(-) NA
aLN(I)-1a 100%	aLN(I)-1, Sp(I)-2, Th(I)-2, iLN (I)-1, iLN(I)-2	1	NC_000067.4 133470113–133479166 (+)	9054	Q	ACAC	2.0567(0.2673%)	(-) 85	(+) 1088	(+) 632
aLN(II)-1n 99%	aLN(I)-1, Th(II)-1, iLN(II)-2, mLN(I)-2	2	NC_000068.5 57038778–57029798 (-)	8981	Q	GTGT	<1.1036(<0.1435%)	(+) 536	(+) 1091	(+) 641
Th(I)-2a 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN (I)-1, iLN(I)-2	2	NC_000068.5 15947290–15938249 (-)	9042	Q	ATTG	<1.0381(<0.1350%)	(-) 162	(+) 1088	(+) 632
Th(I)-2b 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN (I)-1, iLN(I)-2	3	NC_000069.4 67184007–671191723 (+)	7717	Q	ACTT	<1.0367(<0.1348%)	(+) 536	(-) 70	(-) 390
Th(I)-2c 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN (I)-1, iLN(I)-2	3	NC_000069.4 152260526–152269568 (+)	9043	Q	ATGT	<1.0367(<0.1348%)	(+) 536	(+) 1088	(+) 632
aLN(II)-1d 98%	aLN(I)-1, iLN(II)-2, mLN (I)-2, Th(II)-1	4	NC_000070.4 101370070–101361089 (-)	8982	Q	CACC	<1.1036(<0.1435%)	(+) 536	(p) 886	(+) 641
aLN(II)-1m 99%	aLN(I)-1, iLN(II)-2, mLN (I)-2, Th(II)-1	4	NC_000070.4 107652948–107643966 (-)	8983	Q	AAAC	1.102(0.1432%)	(+) 536	(+) 1091	(+) 641
Th(I)-2b 98%	Th(I)-2, iLN(I)-1,iLN(I)-2	4	NC_000070.4 33126327–33118967 (-)	7362	Q	CCAA/CCTG	ND	(-) 162	(p) 886	(-) 19
Th(I)-2d 99%	Th(I)-2, Sp(I)-2, iLN(I)-1, iLN (I)-2, aLN(I)-1	4	NC_000070.4 15169957–15163352 (-)	6606	Q	CAGG	1.038(0.1349%)	(-) 34	(-) 189	(-) 19
aLN(IV)-1a 100%	aLN(IV)-1, Sp(IV)-1, mLN (IV)-1, mLN(IV)-2, mLN (III)-2, mLN(III)-1	4	NC_000070.4 133431467–133436778 (+)	5312	Q	AACA	<1.4063(0.1828%)	(+) 536	(-) NA	(-) 48
Sp(V)-1b 100%	Sp(V)-1, Th(V)-1, mLN(V)-1	4	NC_000070.4 132368033–132373699 (+)	5667	Q	CCTT	<1.5795(0.2053%)	(-) 56	(-) 63	(-) NA
Th(IV)-1a 100%	Th(IV)-1	5	NC_000071.4 23221069–23212403 (-)	8667	P	CTGG	<1.4432(<0.1876%)	(+) 537	(-) 1081	(+) 645
aLN(II)-1a 100%	aLN(I)-1, iLN(II)-2, Th(II)-1, mLN(I)-2	5	NC_000071.4 144923366–144932346 (+)	8981	Q	AGGG	<1.1036(0.1435%)	(-) 34	(+) 1091	(+) 641
aLN(II)-1a 99%	aLN(I)-1, iLN(II)-2, mLN (I)-2, Th(II)-1	5	NC_000071.4 44422608–44431589 (+)	8982	Q	CCAC	<1.1036(0.1435%)	(-) 189	(+) 1091	(+) 641
aLN(II)-1b 99%	aLN(I)-1, iLN(II)-2, mLN (I)-2, Th(II)-3	5	NC_000071.4 78005291–78014272 (+)	8982	Q	CACC	<1.1036(0.1435%)	(+) 536	(+) 1091	(+) 641
Th(I)-2d 98%	Th(I)-2, ILN(I)-1,iLN(I)-2	5	NC_000071.4 122453464–122460825 (+)	7362	Q	GATG	<1.0381(0.1350%)	(-) 34	(p) 722	(-) 19
Th(I)-2e 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN (I)-1, iLN(I)-2	5	NC_000071.4 24740764–24749735 (+)	8972	Q	ATAC	<1.0381(0.1350%)	(+) 536	(p) 1025	(p) 452
Th(I)-2f 99%	Th(I)-2, Sp(I)-2, iLN(I)-1, iLN (I)-2	5	NC_000071.4 43496189–43505229 (+)	9041	Q	ATAT/TTAT	ND	(+) 536	(+) 1088	(+) 632
Th(I)-2g 99%	Th(I)-2, iLN(I)-1, iLN(I)-2, Sp (I)-2, aLN(I)-1	5	NC_000071.4 110148316–110140851 (-)	7466	Q	ATAG	<1.0381(0.1350%)	(p) 143	(p) 1025	(-) NA
Th(I)-2k 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN (I)-1, iLN(I)-2	6	NC_000072.4 73223659–73216301 (-)	7359	Q	ACAA/ACAC	ND	(+) 536	(-) 244	(-) 19
aLN(II)-1i 99%	aLN(II)-1, iLN(II)-2, mLN (I)-2, Th(II)-1	7	NC_000073.4 6682978–6673997 (-)	8982	Q	ATGA	2.2072(0.2869%)	(p) 469	(-) 104	(-) NA

Putative MuERV	Source U3	Ch	Location & Orientation	Size (bp)	PBS	Direct Repeat	Integration Age (Mutation Rate)	Open Reading Frame		
								gag (a.a.)	pol (a.a.)	env (a.a.)
aLN(II)-1j 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	7	NC_000073.4 29323967–29332665 (+)	8699	Q	ATTT	<1.1036(<0.1435%)	(+) 536	(+) 1091	(−) 99
aLN(II)-1k 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	7	NC_000073.4 30397400–30406380 (+)	8981	Q	CATG	<1.1036(<0.1435%)	(+) 536	(p) 1072	(−) NA
aLN(II)-1l 99%	aLN(II)-1, mLN(I)-2, Th(II)-1, iLN(II)-2	7	NC_000073.4 116223258–116228931 (+)	5674	Q	CTAA	<1.1020(0.1433%)	(p) 509	(−) NA	(−) 72
Th(I)-2j 99%	Th(I)-2, Sp(I)-2, iLN(I)-1, iLN(I)-2, aLN(I)-1	7	NC_000073.4 64005512–64014552 (+)	9041	Q	CCTG	<1.0381(0.1350%)	(+) 536	(+) 1091	(+) 632
iLN(III)-1c 98%	iLN(III)-1, Th(III)-1, Th(III)-2	8	NC_000074.4 44819280–44810547 (−)	8734	Q	GGTC/GTCT	ND	(+) 536	(p) 1084	(+) 643
iLN(III)-1d 98%	iLN(III)-1, Th(III)-1, Th(III)-2	8	NC_000074.4 88022347–88015960 (−)	6388	Q	GTAT	<1.3401(0.1742%)	(+) 536	(−) NA	(−) NA
Th(I)-2i 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	8	NC_000074.4 126050806–126058167 (+)	7362	Q	GGAA/GGTG	ND	(+) 536	(−) 637	(−) 19
Sp(V)-1a 100%	Sp(V)-1, Th(V)-1, mLN(IV)-1	8	NC_000074.4 93776396–93782063 (+)	5668	Q	ATAT	<1.5795(0.2053%)	(−) 56	(−) 63	(−) NA
Th(II)-1a 100%	Th(II)-1, aLN(II)-1, iLN(II)-2, mLN(I)-2	8	NC_000074.4 122783017–122775542 (−)	7476	Q	AGGT	<1.1036(<0.1435%)	(−) 162	(−) NA	(+) 641
iLN(III)-1a 98%	iLN(III)-1	9	NC_000075.4 62237180–62245938 (+)	8759	Q	CTGG	ND	(+) 536	(p) 1084	(+) 641
iLN(III)-1b 99%	iLN(III)-1, Th(III)-1, Th(III)-2	9	NC_000075.4 41663263–41656480 (−)	6784	Q	GGAA/GGGG	ND	(−) NA	(p) 1084	(+) 644
aLN(II)-1p 99%	aLN(II)-1, iLN(II)-2, Th(II)-1, mLN(I)-2	10	NC_000076.4 8241458–8234325 (−)	7134	Q	GTGC	<1.1036(<0.1435%)	(−) 189	(−) 182	(−) 36
aLN(II)-1q 99%	aLN(II)-1, Th(II)-1, iLN(II)-2, mLN(I)-2	10	NC_000076.4 41125063–41134043 (+)	8981	Q	GATG	<1.1036(<0.1435%)	(+) 536	(+) 1091	(+) 641
Th(I)-2l 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	10	NC_000076.4 4628826–4619788 (−)	9039	Q	ACAG/TTAG	ND	(−) 162	(−) 179	(−) 45
Th(I)-2p 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	10	NC_000076.4 22423694–22432738 (+)	9045	Q	CTGC/TTGC	ND	(+) 536	(+) 1088	(+) 632
aLN(II)-1c 100%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	11	NC_000077.4 6658394–6649425 (−)	8970	Q	GTTC	<1.1036(<0.1435%)	(+) 536	(−) 182	(+) 641
aLN(II)-1c 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	11	NC_000077.4 8820301–8829280 (+)	8980	Q	ATAG	<1.1036(<0.1435%)	(+) 536	(+) 1091	(+) 641
Th(I)-2m 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	11	NC_000077.4 60402203–60394844 (−)	7360	Q	ACAC	<1.0381(0.1350%)	(−) 85	(−) NA	(−) 19
Th(I)-2n 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	11	NC_000077.4 76365003–76374047 (+)	9045	Q	AGGG/TTGG	ND	(+) 536	(+) 1088	(+) 632
Th(I)-2o 99%	Th(I)-2, Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	11	NC_000077.4 86698511–86707551 (+)	9041	Q	ACAC	<1.0381(0.1350%)	(+) 536	(+) 1091	(+) 632
Sp(I)-2a 98%	Sp(I)-2, aLN(I)-1, iLN(I)-1, iLN(I)-2	11	NC_000077.4 102899753–102908794 (+)	9042	Q	GAAAC	1.038(0.1349%)	(p) 235	(p) 988	(+) 632
Th(II)-1r 99%	Th(II)-1, aLN(II)-1, iLN(II)-2, mLN(I)-2	11	NC_000077.4 88713056–88706270 (−)	6787	Q	GGAG	<1.1036(<0.1435%)	(+) 535	(−) 358	(+) 641
aLN(II)-1t 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	12	NC_000078.4 20945382–20951142 (+)	5761	Q	CTGT	<1.1036(<0.1435%)	(−) 205	(−) NA	(−) NA
Th(I)-2h 99%	Th(I)-2, Sp(I)-2, iLN(I)-1, iLN(I)-2, aLN(I)-1	12	NC_000078.4 70465411–70474451 (+)	9041	Q	AGAC	<1.0381(0.1350%)	(p) 235	(+) 1091	(+) 632
aLN(II)-1o 99%	aLN(II)-1, iLN(II)-2, Th(II)-1, mLN(I)-2	13	NC_000079.4 100095412–100086432 (−)	8981	Q	GTGG	<1.1036(<0.1435%)	(+) 536	(p) 1069	(+) 641
Th(I)-2a 100%	Th(I)-2, iLN(I)-1, iLN(I)-2, Sp(I)-2, aLN(I)-1	13	NC_000079.4 21819875–21827661 (+)	7787	Q	CTAC	<1.0381(0.1350%)	(+) 536	(+) 1088	(−) 160
aLN(IV)-1a 99%	aLN(IV)-1, Sp(IV)-1, mLN(IV)-1, mLN(IV)-2, mLN(III)-1	13	NC_000079.4 68392677–68383991 (−)	8687	Q	GTAC	<1.4063(0.1828%)	(+) 536	(p) 1084	(+) 644
iLN(III)-1 99%	iLN(III)-1, Th(III)-1, Th(III)-2	14	NC_000080.4 53484571–53492393 (+)	7823	Q	GTAT	1.3401(0.1742%)	(−) NA	(p) 1084	(+) 643
aLN(II)-1h 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	15	NC_000081.4 76394257–76387072 (−)	7186	Q	AAACAAACAAAC	<1.1036(<0.1435%)	(−) 162	(−) NA	(+) 641

Putative MuERV	Source U3	Ch	Location & Orientation	Size (bp)	PBS	Direct Repeat	Integration Age (Mutation Rate)	Open Reading Frame		
								gag (a.a.)	*pol* (a.a.)	*env* (a.a.)
aLN(II)-1c 98%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	16	NC_000082.4 76026704–76017723 (–)	8982	Q	CTGG	<1.1036(<0.1435%)	(+) 536	(p) 886	(–) 45
aLN(II)-1g 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	16	NC_000082.4 93594977–93585997 (–)	8981	Q	GCCT	<1.1036(<0.1435%)	(–) 34	(+) 1091	(+) 641
aLN(II)-1b 100%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	18	NC_000084.4 82827979–82836858 (+)	8880	Q	CCTG	<1.1036(<0.1435%)	(–) 164	(+) 1091	(+) 641
iLN(II)-1b 98%	iLN(II)-1, Th(II)-1, Th(III)-1, Th(III)-2	19	NC_000085.4 60988696–60979970 (–)	8727	Q	CTTG	<1.3401(0.1742%)	(+) 536	(p) 1084	(+) 640
aLN(II)-1s 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	19	NC_000085.4 38440166–38447227 (+)	7062	Q	CTTC	<1.1036(<0.1435%)	(–) 34	(p) 883	(–) NA
aLN(II)-1e 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	X	NC_000086.5 14631057–14639937 (+)	8981	Q	ATAC	<1.1036(<0.1435%)	(+) 536	(p) 998	(+) 641
aLN(II)-1f 99%	aLN(II)-1, iLN(II)-2, mLN(I)-2, Th(II)-1	X	NC_000086.5 51501607–51510587 (+)	8981	Q	TGAA/GAGT	ND	(+) 536	(+) 1091	(+) 641