Published in final edited form as:
Genomics. 2012 September ; 100(3): 157–161. doi:10.1016/j.ygeno.2012.06.002.

ERE Database: a database of genomic maps and biological properties of endogenous retroviral elements in the C57BL/6J mouse genome

Damian Kao, **Karen Hsu**, **Sophia Chiu**, **Vince Tu**, **Alex Chew**, **Kang-Hoon Lee**, **Young-Kwan Lee**, **Deug-Nam Kwon**, **David G. Greenhalgh**, and **Kiho Cho**[*]
Burn Research, Shriners Hospitals for Children Northern California and Department of Surgery, University of California, Davis, Sacramento, CA 95817

Abstract

Endogenous retroviral elements (EREs), a family of transposable elements, constitute a substantial fraction of mammalian genomes. It is expected that profiles of the ERE sequences and their genomic locations are unique for each individual. Comprehensive characterization of the EREs' genomic locations and their biological properties is essential for understanding their roles in the pathophysiology of the host. In this study, we identified and mapped putative EREs (a total of 111 endogenous retroviruses [ERVs] and 488 solo long terminal repeats [sLTRs]) within the C57BL/6J mouse genome. The biological properties of individual ERE isolates (both ERVs and sLTRs) were then characterized in the following aspects: transcription potential, tropism trait, coding potential, recombination event, integration age, and primer binding site for replication. In addition, a suite of database management system programs was developed to organize and update the data acquired from current and future studies and to make the data accessible via internet.

Keywords

mouse genome; database; murine lukemia virus; endogenous retrovirus; retroelement; solo long terminal repeat

1. Introduction

Transposable elements, which constitute a significant fraction of the genome of various organisms, can be classified into elements with RNA intermediates (retroelements) and those without (DNA transposons). Generated by genomic insertions from ancient retroviral infections of germline cells, endogenous retroviral elements (EREs) are transposable elements with RNA intermediates [1, 2]. The EREs are passed from the host to its progeny generations following Mendelian inheritance.

[*] Corresponding author, Burn Research Shriners Hospitals for Children Northern California and Department of Surgery, University of California, Davis, 2425 Stockton Blvd., Sacramento, CA 95817, 916-453-2284 (Tel), 916-453-2288 (Fax), kcho@ucdavis.edu.

Conflict of interest
There is no conflict of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

EREs are found at a high percentage in the genomes of all vertebrates [3]. The universal prevalence of EREs in vertebrates suggests that EREs may play an important role in their evolutionary development, maintenance, and pathophysiology [4, 5]. Although some EREs, primarily due to mutations and recombination events, are defective for coding potential or lack transcription regulatory elements essential for expression and replication, a significant number of replication-competent EREs exist [6, 7].

There are multiple groups of EREs in the mouse genome [8, 9]. The two most widely studied EREs are the murine leukemia virus (MuLV)-type and the mouse mammary tumor virus-type. In this first phase of the project, we identified and mapped MuLV-type EREs, both endogenous retroviruses (ERVs) and solo long terminal repeats (sLTRs), in the C57BL/6J mouse genome and characterized their biological properties. The data obtained from this study were organized into a newly-developed database management system (DBMS) and the resulting database, called “ERE Database,” is available for download to the public through a dedicated website. The eventual goal of this project is to establish a comprehensive database of EREs in both human and murine genomes.

2. Results and Discussion

2.1. DBMS design

The suite of DBMS programs designed for the ERE data consists of a database editor for updating new entries into the database and a database viewer for accessing the information. Both programs were developed using Microsoft Visual Studios 2005 and written in Visual Basic.NET. Two custom class libraries were also created for the software: a database adapter class library to standardize and simplify inserting or retrieving information to or from the database and a database objects class library to define custom data structures used in the program. Microsoft Access was the format used to organize the database. Another format, such as SQLite, a software library that implements a server-less database engine, may be used in the future if the data size becomes too large.

The database editor allows users to create new strain and/or species databases, add new entries or edit previously inserted entries. It also allows users to configure the probe tree and customize the data display in the database viewer program. The editor also permits the user to add different data types, such as image or spreadsheet.

The database viewer program retrieves entries from the database and displays them (Figure 2). The user can choose either to browse, by probe or retroviral element, or to search directly for the data. To browse by probes, a homology tree of the chosen probe set is displayed and the probe name can be selected. To browse by retroviral elements, the user can choose EREs based on their chromosomal locations. The search menus allow users to query for probes or EREs by probe name, ERE name, chromosome, ERE type (ERV or sLTR), strand, and location. The data viewing interface was designed so that multiple entries can be viewed at the same time. A series of tabs allows rapid switching between different data entries.

2.2. Database design

The database file is in a Microsoft Access format. Each strain/species has its own database file containing 32 relational tables. The main two tables are the probe and ERE tables; since each probe may have multiple ERE targets and each ERE may have multiple probes aligned, they are joined by a junction table to form a many-to-many relationship (Figure 3). Multiple sub-tables are related to the two main tables to store information regarding the probe or ERE. The database tables were designed with a degree of built-in redundancy to accommodate future data type additions.

The DBMS was created as a de-centralized and downloadable database system for several reasons. The size of the data is compact enough to download in its entirety to the user's personal computer so that the information can be accessed without an internet connection. Having the entire set of data on the user's computer also allows faster and easier access to the information. An interactive user interface, such as the one used in this DMBS, is also more difficult in a web browser based environment since interactive web techniques, such as Asynchronous Java Script and XML, are not advanced enough to encompass the necessary level of complexity for the database.

3. Conclusions

The current ERE Database presents a set of data obtained from surveying the entire C57BL/6J mouse genome *in silico*, using a library of unique ERE-mining probes cloned from C57BL/6J mouse genomic DNA. Future studies will yield additional datasets and the information will be periodically updated into the database. Currently, the ERE Database is designed as a de-centralized and downloadable system. As the dataset size and complexity increases and if more frequent updates are needed, the database may be converted to a centralized, web-based system. The goal of this project is to establish a comprehensive database of murine and human EREs in regard to their sequences, genomic locations and neighboring host genes, and biological properties. This information may eventually illuminate the role of EREs in evolutionary development, maintenance, and pathophysiology of the host organisms.

The ERE Database is accessible at http://eredatabase.ucdmc.ucdavis.edu/.

4. Material and Methods

4.1. Establishment of a library of probes for ERE-mining

Genomic DNA was obtained from the C57BL/6J mouse. Five different primer sets: ERV-U2 and ERV-U1, MUP-1A and MUP-2A, MUP-1A and ERV-U1, MUP-1A and MV-2D, and MUP-1A and MV-2E were used for PCR amplification of the polymorphic U3 promoter sequences of ERVs from the murine genome. The primer sequences were as follows: ERV-U1: 5'-CGG GCG ACT CAG TCT ATC GG-3'; ERV-U2: 5'-CAG TAT CAC CAA CTC AAA TC-3'; MUP-1A: 5'-GAC CCC ACC ATA AGG CTT AG-3'; MUP-2A: 5'-CTC AGT CTA TCG GAG GAC-3'; MV-2D: 5'-CTC AGT CTG TCG GAG GAC TG-3'; MV-2E: 5'-CGG ATG TAA TCA GCA AGA GGC-3'. The amplified U3 sequences were then cloned into the pGEMT-Easy vector (Promega, Madison, WI) and a total of 297 plasmid DNAs containing the U3 inserts was sequenced. The pGEMT-Easy sequence and the excess sequences, resulting from primer design/location, were removed to identify only the U3 sequences.

The resulting U3 sequences were analyzed for uniqueness using the VectorNTI's AlignX program (Invitrogen, Carlsbad, CA) which utilizes the ClustalW algorithm. Unique U3 sequences were identified using a multi-step strategy. The initial alignment of the U3 sequences was based on the primer sets associated with the U3; the potential unique sequences were determined by analyzing the branching pattern of the phylogenetic tree generated by the AlignX program (Invitrogen), which utilizes the neighbor joining method. The potential unique sequences resulting from the initial alignment were then aligned against each other and the resulting phylogenetic tree revealed five groups based on the branching pattern. Then individual U3 sequences were analyzed for mutations by comparison to the generated consensus sequence; they were then subgrouped based on the number of mutations found (0-2 nucleotides, 3–4 nucleotides, and 5 or more nucleotides). A representative U3 sequence was then chosen from each of the subgroups and each representative sequence was then aligned against the others. A phylogenetic tree of the final 66 unique U3 sequences was generated using the MEGA4 program (www.megasoftware.net) through bootstrapping with 100 replications (Figure 1). The final 66 unique U3 sequences were designated as ERE-mining probes and used for a subsequent survey of the C57BL/6J mouse genome to identify and map EREs, both ERVs and sLTRs.

4.2. Identification and mapping of EREs

Using the National Center for Biotechnology Information (NCBI)'s Megablast program, each unique U3 sequence was employed to probe the C57BL/6J mouse genome (NCBI database [Build 36.1]) for putative MuLV-type ERVs and sLTRs with the following parameters: "limit by entrez query" was set to "NC_000067:NC_000087," "filters" was set to "none," "word size" was set to "64," and the "percent identity, match, mismatch scores" was set to "80 (higher than 80 % identity), 2, −3." The resulting hits were then identified as putative ERV proviruses if two long terminal repeats (LTRs) were found within 12 kb of each other. The putative ERV proviral sequences, as well as an additional 50 bp upstream of the 5' LTR and 50 bp downstream of the 3' LTR, were cloned *in silico* from the NCBI database for further functional characterization. sLTRs were defined as LTRs lacking any neighboring LTRs within the designated 12 kb. A total of 8,986 LTR hits were generated from the survey using the 66 U3 probes, and the setting of "80 % sequence identity" yielded 710 LTR hits which resulted in 111 ERVs (two LTR hits/ERV) and 488 sLTRs (one LTR hit/sLTR).

4.3. Tropism trait analysis

The potential tropism traits of the 66 U3 probes were analyzed by comparing their sequence characteristics of direct repeats, a 190 bp insertion, and the unique regions in comparison to the reference sequences reported by Tomonaga et al. [10]. Five direct repeats, one 190 bp insertion, and one unique region were identified within the U3 probe sequences analyzed.

4.4. Profiling of transcription regulatory elements within U3 probes/promoters

Each U3 probe/promoter sequence was analyzed for the presence of potential transcription regulatory elements using the MatInspector program 8.0 (Genomatix, Munich, Germany) [11]. The parameters were set to the vertebrate matrix group with a core similarity of 0.90, resulting in a 10 % or fewer mismatches within the most conserved bases, and the matrix similarity was set to "optimized" to reduce false positives.

4.5. Determination of recombination event and integration age

The 5' and 3' LTRs of each ERV were aligned against each other to determine the age of the initial proviral integration into the germline. Integration age was calculated based on the formula of "every 0.13 % mutation rate between the 5' and 3' LTRs equals to 1 million years" [12, 13]. Recombination events for ERVs were determined by examining the direct repeat sequences of 4–12 bp upstream of the 5' LTR and downstream of the 3' LTR. The absence of direct repeats on these two sites next to the ERV indicates that a potential recombinant event occurred.

4.6. Open reading frame (ORF) analysis of ERVs and sLTRs

Individual ERV sequences were subjected to ORF analysis for *gag*, *pol* and *env* polypeptides using the VectorNTI program (Invitrogen) by comparing the putative ERV sequences against selected reference MuLV sequences (GenBank accession numbers: AF033811, J02255, DQ241301, and S80082) which are capable of encoding intact polypeptides. In addition, analysis for potential viral and non-viral polypeptides located within 10 kb downstream of sLTRs was performed by using the VectorNTI program

(Invitrogen) under the criteria that only ORFs of 450 bp or larger in size were selected. The resulting ORFs were then blasted using the blastp program of NCBI, and retroviral proteins matching mouse, rat, pig, or human sequence were noted. In the absence of retroviral proteins, the best matching protein within the above described organisms was noted.

4.7. Determination of primer binding site (PBS)

The PBS for individual ERVs identified in this study was determined by analyzing a stretch of 18 bp located immediately downstream of the 5’ LTR region and comparing this region to conserved PBS sequences [14].

Acknowledgments

This study was supported by grants from Shriners of North America (No. 86800) and National Institute of General Medical Sciences (R01GM071360).

Abbreviations

DBMS	database management system
ERE	endogenous retroviral element
ERV	endogenous retrovirus
LTR	long terminal repeat
MuLV	murine leukemia virus
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PBS	primer binding site
sLTR	solo long terminal repeat

References

1. Deininger PL, Batzer MA. Mammalian retroelements. Genome Res. 2002; 12:1455–1465. [PubMed: 12368238]
2. Urnovitz HB, Murphy WH. Human endogenous retroviruses: nature, occurrence, and clinical implications in human disease. Clin Microbiol Rev. 1996; 9:72–99. [PubMed: 8665478]
3. Bannert N, Kurth R. Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A. 2004; 101(Suppl 2):14572–14579. [PubMed: 15310846]
4. Kowalski PE, Freeman JD, Mager DL. Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes. Genomics. 1999; 57:371–379. [PubMed: 10329003]
5. Mager DL, Hunter DG, Schertzer M, Freeman JD. Endogenous retroviruses provide the primary polyadenylation signal for two new human genes (HHLA2 and HHLA3). Genomics. 1999; 59:255–263. [PubMed: 10444326]
6. 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; 7:1088–1095. [PubMed: 15452578]
7. Lee YK, Chew A, Phan H, Greenhalgh DG, Cho K. Genome-wide expression profiles of endogenous retroviruses in lymphoid tissues and their biological properties. Virology. 2008; 373:263–273. [PubMed: 18187179]
8. Boeke, JDaSJP. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin, JM.; Hughes, SH.; Varmus, HE., editors. Retroviruses. Cold Spring Harbor Press; Cold Spring Harbor: 1997. p. 343-435.

9. Keshet E, Schiff R, Itin A. Mouse retrotransposons: a cellular reservoir of long terminal repeat (LTR) elements with diverse transcriptional specificities. Adv Cancer Res. 1991; 56:215–251. [PubMed: 1851374]

10. 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; 73:4327–4340. [PubMed: 10196331]

11. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics. 2005; 21:2933–2942. [PubMed: 15860560]

12. Lebedev YB, Belonovitch OS, Zybrova NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdlov ED. Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene. 2000; 247:265–277. [PubMed: 10773466]

13. Dangel AW, Baker BJ, Mendoza AR, Yu CY. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. Immunogenetics. 1995; 42:41–52. [PubMed: 7797267]

14. 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; 254:10979–10985. [PubMed: 115865]

Highlights

- Putative EREs were identified and mapped in the C57BL/6J genome.
- The biological properties of each ERE were characterized.
- A database management system was developed using the data obtained from this study.

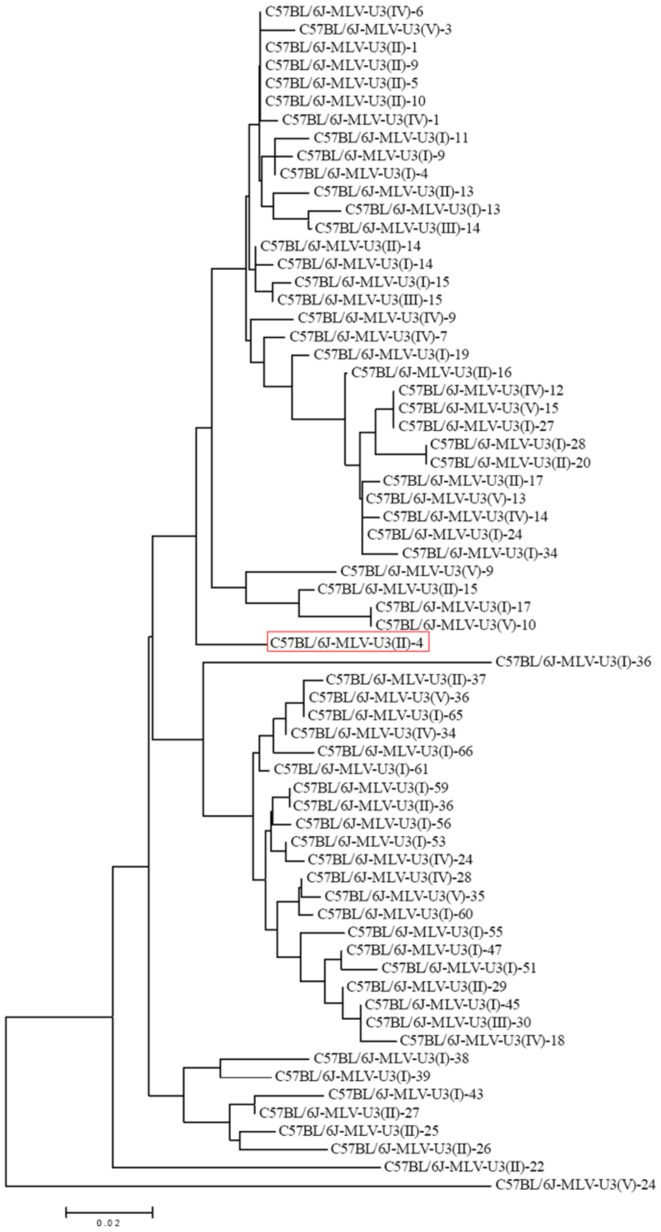
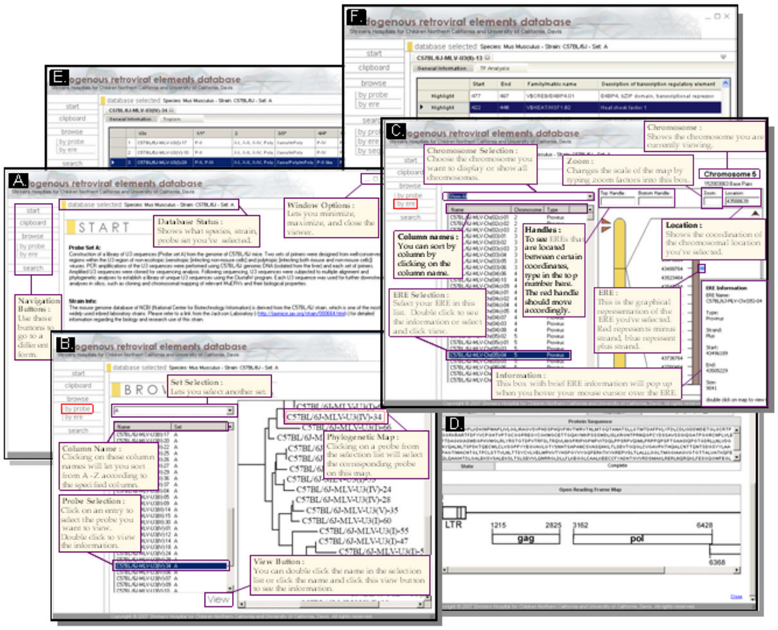


Figure 1.

Phylogenetic tree of a library 66 unique ERE probes cloned from C57BL/6J mouse genomic DNA for ERE mining *in silico*. Red box indicates selected probe.



Window Options : Lets you minimize, maximize, and close the viewer.

Database Status : Shows what species, strain, probe set you've selected.

Navigation Buttons : Use these buttons to go to a different form.

Set Selection : Lets you select another set.

Column Name : Clicking on these column names will let you sort from A - Z according to the specified column.

Probe Selection : Click on an entry to select the probe you want to view. Double click to view the information.

Phylogenetic Map : Clicking on a probe from the selection list will select the corresponding probe on this map.

View Button : You can double click the name in the selection list or click the name and click this view button to see the information.

Chromosome Selection : Choose the chromosome you want to display or show all chromosomes.

Chromosome : Shows the chromosome you are currently viewing.

Zoom : Changes the scale of the map by typing zoom factors into this box.

Column names : You can sort by column by clicking on the column name.

Handles : To see EREs that are located between certain coordinates, type in the top number here. The red handle should move accordingly.

Location : Shows the coordination of the chromosomal location you've selected.

ERE Selection : Select your ERE in this list. Double click to see the information or select and click view.

ERE : This is the graphical representation of the ERE you've selected. Red represents minus strand, blue represent plus strand.

Information : This box with brief ERE information will pop up when you hover your mouse cursor over the ERE.

Figure 2.

Screen shots (with tutorial comments) of the ERE Database. A) Main page, B) Browse by probe, C) Browse by ERE, D) Coding potential, E) Tropism, and F) Transcription regulatory element binding site.

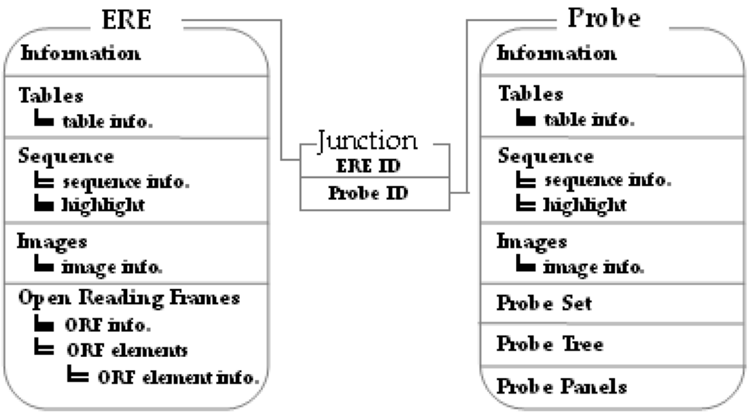


Figure 3.

Simplified schema of the ERE Database showing the two main tables (ERE and Probe) and the junctional relationship table. ORF elements (*gag, pol,* and *env* polypeptides).