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, Kiho Cho *  

Burn Research, Shriners Hospitals for Children Northern California, Sacramento, CA 95817, USA  
Department of Surgery, University of California, Davis, Sacramento, CA 95817, USA

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. 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

* Corresponding author at: Department of Surgery, University of California, Davis, 2425 Stockton Blvd., Sacramento, CA 95817, USA. Fax: +1 916 453 2288.  
E-mail address: kcho@ucdavis.edu (K. Cho).
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 (Fig. 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,

**Fig. 1.** Phylogenetic tree of a library of 66 unique ERE probes cloned from C57BL/6J mouse genomic DNA for ERE mining in silico. Red box indicates selected probe.
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 (Fig. 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/.

Fig. 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. (F) Transcription regulatory element binding site.
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 GCC 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 CCT 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 160 probes were selected. 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).

4.2. 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 using the ClustalW algorithm. Unique U3 sequences were then 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 identiﬁed 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 (Fig. 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. 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.

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

Individual ERV sequences were subjected to ORF analysis for \textit{gag}, \textit{pol} and \textit{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 poly-peptides. In addition, analysis for potential viral and non-viral poly-peptides 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 retro-viral 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].

Conflict of interest

There is no conflict of interest.

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBMS</td>
<td>database management system</td>
</tr>
<tr>
<td>ERE</td>
<td>endogenous retroviral element</td>
</tr>
<tr>
<td>ERV</td>
<td>endogenous retrovirus</td>
</tr>
<tr>
<td>LTR</td>
<td>long terminal repeat</td>
</tr>
<tr>
<td>MuLV</td>
<td>murine leukemia virus</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PBS</td>
<td>primer binding site</td>
</tr>
<tr>
<td>sLTR</td>
<td>solo long terminal repeat</td>
</tr>
</tbody>
</table>

Acknowledgments

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

References