Genes & Genomics (2011) 33: 679-684
DOI 10.1007/s13258-011-0103-4

RESEARCH ARTICLE

Development of GEBRET: a web-based analysis tool for retroelements in primate genomes

Hong-Seok Ha · Woo-Keun Chung · Kung Ahn · Jin-Han Bae · Sang-Je Park · Jae-Woo Moon · Kyu-Hwi Nam · Kyudong Han · Hwan-Gue Cho · Heui-Soo Kim

Received: 19 May 2011 / Accepted: 05 July 2011 / Published online: 10 December 2011

Abstract
Retroelements play important roles in primate evolution. Specifically, human endogenous retroviruses (HERVs) and *Alu* elements are primate-specific retroelements. In addition, SVA elements belong to the youngest family of hominid non-long terminal repeat (LTR) retrotransposons. Retroelements can affect adjacent gene expression, supplying cis-regulatory elements, splice sites, and poly-A signals. We developed a database, GEnome-wide Browser for RETroelement (GEBRET, http://neobio.cs.pusan.ac.kr/~gebre/), for comparing the distribution of primate-specific retroelements and adjacent genes. GEBRET database components include 47,381 HERVs, 53,924 *Alu*s and 4639 SVAs in five primate genomes of human, chimpanzee, orangutan, rhesus macaque, and marmoset. Host genes located upstream of a retroelement were also visualized and classified as five categories (0.0, 0.5, 1.0, 2.0, and 3.0Kb). Our results suggest that retroelements preferentially integrate into the distal promoter region relative to the core promoter region. GEBRET database is designed to investigate the distribution of retroelements (HERVs, *Alu*s and SVAs) in the primate genomes that have been sequenced. Our software will be useful in the field to study the impact of retroelements on primate genome evolution.

Keywords Retroelements; Primates; Visualization; Evolution; Bioinformatics

Introduction

Retroelements occupy ~44% of the human genome (Lander et al., 2001). While *Alu* elements and long interspersed element-1 (LINE-1 or L1) are the most abundant retroelements in the human genome, human endogenous retroviruses (HERVs) account for 1%~8% of the genome sequences (Belshaw et al., 2004). In addition to the major retroelement families, SVA (short interspersed element, variable number of tandem repeat, and *Alu*) is present in the hominoid species. Although they are less ubiquitous than other retroelements, SVA elements have been identified along the primate lineage.

HERVs are remnants of ancient retroviral infections in germline cells during early primate evolution. Like ERVs, they were inherited in a Mendelian manner. Although the majority of HERVs have lost their function by nucleotide substitutions and deletion events, a few HERVs seem to have retained their functions. Actually, it was reported that some HERVs have positive effects on hosts (e.g., Syncitin) (Mi et al., 2000), and can be associated with several human pathologies including cancers and autoimmune diseases (Voisset et al., 2008).

Alu and SVA elements belong to non-long terminal repeat (non-LTR) retrotransposons and they are called non-autonomous elements. *Alu* elements account for ~11% of the human genome. Among them, less than 0.5% are polymorphic (Roy-Engel et al., 2001). The majority of human *Alu* elements are shared in non-human primates, but ~7,000 *Alu* insertions are unique to humans (Chimpanzee Sequencing and Analysis Consortium 2005). *Alu*-mediated recombination occasionally induces deletion or duplication of host DNA sequences (Batzer

H.-S. Ha and W.-K. Chung contributed equally to this work.

H.-S. Ha · K. Ahn · J.-H. Bae · S.-J. Park · J.-W. Moon · K.-H. Nam · H.-S. Kim (✉)
Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea
e-mail: khs307@pusan.ac.kr

W.-K. Chung · H.-G. Cho (✉)
Department of Computer Science and Engineering, Pusan National University, Busan 609-735, Korea
e-mail: hgcho@pusan.ac.kr

K. Han
Department of Nanobiomedical Science & WCU Research Center, Dankook University, Cheonan 330-714, Korea

and Deininger, 2002; Sen et al., 2006; Han et al., 2007). Thus, *Alu* elements have been linked to various human genetic disorders. SVA elements are hominid-specific retrotransposons. SVA elements can also cause several genetic diseases such as Fukuyama-type congenital muscular dystrophy (Kobayashi et al., 1998). In addition, SVA elements could produce 3' flanking sequence transductions like LINE-1 retroelements (Xing et al., 2006).

Retroelements have influenced the evolution of primate genomes through transposition, translocation, and recombination (Baban et al., 1996; Goodchild et al., 1995; Jamain et al., 2001; Belshaw et al., 2007). Furthermore, previous studies have shown that retroelements can affect adjacent gene expression, supplying tissue-specific enhancers (Ting et al., 1992; Ling et al., 2002), promoters (Medstrand et al., 2001), splice donors (Kato et al., 1987), acceptor sites (Feuchter-Murthy et al., 1993), and poly-A signals (Juang et al., 1994). We developed GEBRET (Genome-wide Browser for Retroelement) into a web-accessible engine to use conveniently when studying retroelements which affect gene expression in primate genomes.

Materials and Methods

Searching for target retroelements from BLAT results

The nucleotide sequences of HERVs, *Alu*s and SVAs were downloaded from Repbase Update (www.girinst.org). HARLEQUIN, PRIMA4, MER21I, MER41I, and MER57I sequences were also downloaded from Repbase Update because they belong to primate-specific LTRs. Indeed, the HERV16, HERVL_40, HERVL68, HERVL74 subfamilies are not primate-specific. Nevertheless, we included them into our database in order to analyze entire HERVs that affect primate genomes. Only the internal sequences of HERVs are recognized as HERVs in current databases. Thus, we also investigated the LTR sequence for each HERV subfamily and supplemented our database with the LTR sequence (LTR16 for HERV16, LTR17 for HERV17, LTR23 for HERV23, LTR4 for HERV3, MER4I for HERV39, LTR12 for HERV9, LTR2 for HERVE, LTR21 for HERVFH21I, LTR7 for HERVH, MER48 for HERVH48I, LTR10B for HERVI, LTR5 for HERVK, MER11D for HERVK11DI, MER11A for HERVK11I, LTR13 for HERVK13I, LTR22C0 for HERVK22I, MER9 for HERVK9I, LTR14 for HERVKC4, LTR6A and LTR6B for HERVS71). Using the BLAT program, we constructed each retroelement library (http://genome.ucsc.edu/cgi-bin/hgBlat) for each species of human, chimpanzee, orangutan, rhesus macaque, and marmoset genomes. To identify all retroelements including full-length and partial sequences, we utilized all BLAT results into our database. As shown in Fig. 1C, the GEBRET database includes 47,381 HERVs, 53,924 *Alu*s and 4639 SVAs from the five fully-sequenced primate genomes. Since low-score data can result in false outputs,

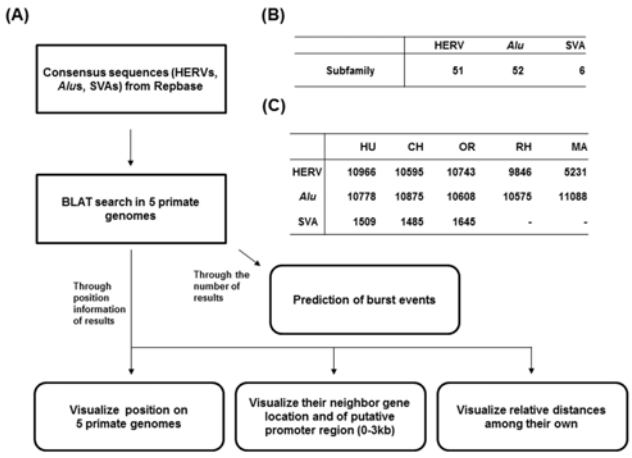

Figure 1. Development strategy and datasets. (A) Overview of the data integration process for developing the GEBRET database. (B) The total number of subfamilies used in the GEBRET database (C) The amount of BLAT data.

GEBRET uses a score filter to increase the accuracy of outputs. Especially, in the case of SVA elements, we recommend a threshold score greater than 400.

Identification of retroelements-adjacent genes

As mentioned before, our database includes the five primate genome datasets including human (hg18), chimpanzee (panTro2), orangutan (ponAbe2), rhesus macaque (rheMac2), and marmoset (calJac3). Each genome sequence was obtained from the UCSC genome database (International Human Genome Sequencing Consortium 2004; Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). To identify retroelement-adjacent genes, we compared an mRNA database with the BLAT database (Kim et al., 2009).For each selected retroelement, GEBRET could identify adjacent genes based on the distance between the retroelement and the adjacent genes. We divided the distance into five categories (0.0, 0.5, 1.0, 2.0, and 3.0Kb; host genes located upstream of a retroelement).

Results and Discussion

Comparative analysis of retroelements from different primate genomes

GEBRET is accessible at http://neobio.cs.pusan.ac.kr/~gebre/. Java Runtime Environment software ver. 1.6 or newer is required to use GEBRET. This program is designed to select a specific retroelement and genome on the main page. The main screen contains five sub-windows and each sub-window represents each primate genome. The web-based genome browser operates on JavaServer Faces technology. The web

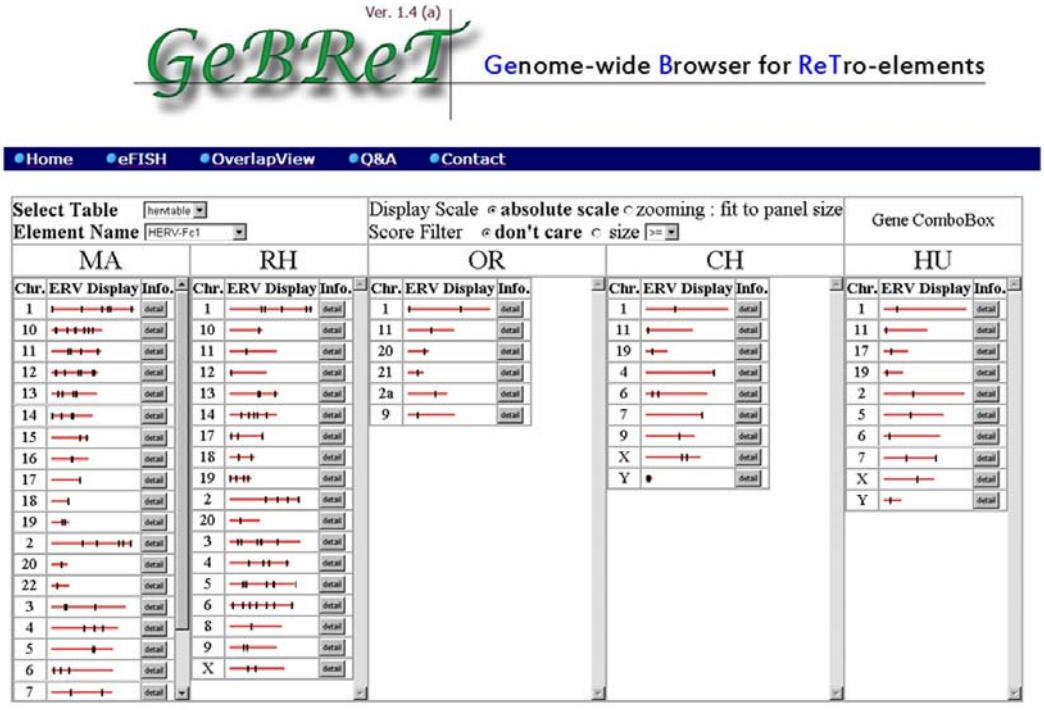

Figure 2. GEBRET web interface 1: distribution of retroelements. The main page shows five sub-windows visualizing the distribution of retroelements. For example, the distribution form of HERV-Fc1 over five primate genomes are compared, when users choose element name of HERV-Fc1. The "detail" icons are linked sub-window visualizing retroelement-adjacent genes.

interface allows users to access the database content via a repeat name and score filter. Retroelements are easily selected by scrolling a combo box containing all retroelement names. By selecting a score, users would be able to control specificity of outputs. Thus, this program is useful for analytical studies of retroelements. Figure. 2 shows the outputs of HERV-Fc1 over five primates.

Analysis of the distance between a retroelement and its adjacent gene

Previous studies on the relation of a retroelement and its adjacent gene have focused on transcript variants generated by alternative promoters (Medstrand et al., 2001), splice sites (Kato et al., 1987; Feuchter-Murthy et al., 1993) and poly-A signal (Juang et al., 1994). However, they overlooked the impact of an enhancer on gene expression because they are lo-

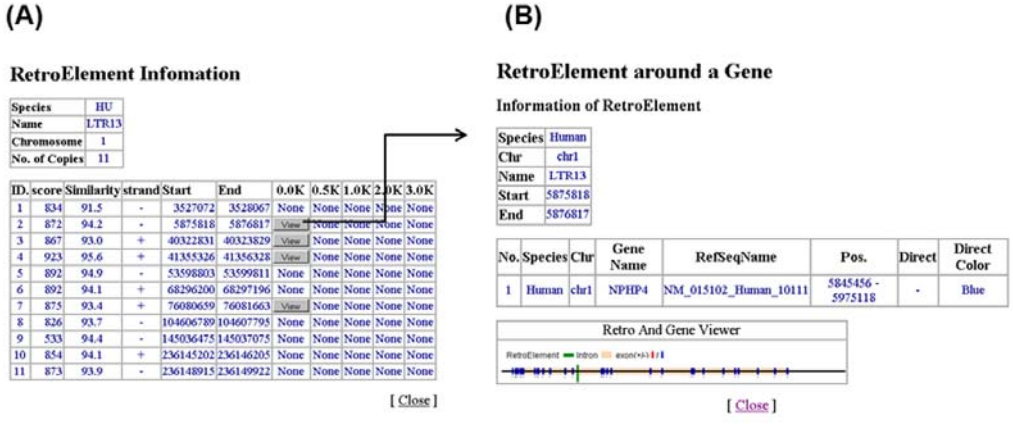

Figure 3. GEBRET web interface 2: retroelement-adjacent genes. (A) Schema representing the five categories derived according to the distance between a retroelement and the adjacent genes. If adjacent genes exist in five categories, new "view" icons appear in the sub-window. (B) The "view" icon links to another sub-window, which enables detailed depiction of the structures between retroelements and their adjacent genes.

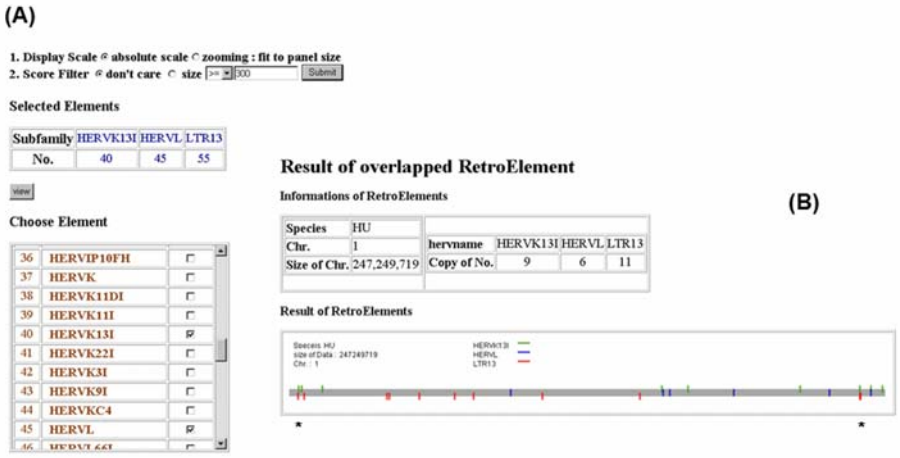

Figure 4. GEBRET web interface 3: multiple retroelements. (A) The users can choose different retroelements in specific chromosome and species (B) When we chose three elements; HERVK13I, HERVL, and LTR13, Nine HERVK13Is, six HERVLs, and ten LTR13s were detected on human chromosome 1 (threshold >300). Asterisks indicate that LTR13 sequences seem to be components of a HERVK13I element.

cated kilobases away from genes (Ting et al., 1992; Ling et al., 2002). To open the field to further studies of enhancers generated by retroelements, we analyze the distance between retroelements and adjacent genes from 0 to 3kb. The distance is divided into five categories (0.0, 0.5, 1.0, 2.0 and 3.0Kb; Fig. 3). Category 0.0Kb means that a retroelement is located in the intragenic region, categorie 0.5, 1.0, 2.0 and 3.0Kb indicate that a retroelement may reside on a promoter or an enhancer. The "detail" icon which is located on the main screen (Fig. 2) is linked with a list of retroelements on a chromosome. Thus, by clicking the "detail", users could access the distribution of retroelements. If any gene exists within the distance categories, "view" icons will appear on the window (Fig. 3A). By clicking the "view" icon, users are able to approach detailed gene information such as gene structure in the window (Fig. 3B). It was reported that the human genome contains 20,000–25,000 protein-coding genes (International Human Genome Sequencing Consortium, 2004). Alternative splicing, alternative promoter, epigenetic regulation and post-translational modifications significantly contribute to human proteome dynamics (Ewing et al., 2000). Actually, more than 75% of human genes use alternative promoters (Davuluri et al., 2008). Thus, more than 90,000 human proteins are estimated in humans (Modrek and Lee, 2002; Woodley and Valcarcel, 2002). Retroelement insertion can provide an alternative promoter to host genes (Landry et al., 2003) and they can be utilized in exonic/intronic sequences (Corvelo et al., 2008). Therefore, we suggest that the physical distance between retroelements and their adjacent genes could be a key factor in studying the impact of retroelements on gene regulation and expression.

The distribution of retroelements on a chromosome

Many retroelements reside on a chromosome. Their position and density on a chromosome are important for studying genomic rearrangements during primate evolution. By using the function of "OverlapView", we are able to get the copy number of a specific retroelement subfamily and their physical distribution on each chromosome. As shown in Fig. 4A, we can choose retroelement subfamilies of interest and the output of "OverlapView" to gather useful information as mentioned above. One full-length retroelement which contains many nucleotide substitutions can be recognized as several partial retroelements by searching programs currently in use. In the case where the retroelement is a member of a HERV, this error could be somewhat corrected by using the "OverlapView". The output of the "OverlapView" shows the positions of HERVK13I and LTR13 which is the LTR of HERV13I in Fig. 4B. Using their position information, we can recognize that the retroelements are not partial elements but one full-length HERVK13I. In addition, we suggest that this GEBRET is a helpful tool to detect retroelement-mediated recombination, which generates chimeric retroelements. As reported in recent studies of *Alu* recombination-mediated deletions (Sen et al., 2006; Han et al., 2007), the chimeric *Alu* elements are created by overlapping between two adjacent *Alu* elements.

Use of GEBRET as a FISH simulation tool

Fluorescence in situ hybridization (FISH) is a cytogenetic technique used to detect the presence or absence of specific DNA fragments on chromosomes. BLAST is a bioinformatics tool using principles similar to FISH. Recently, eFISH (FISH simulation tool, http://projects.tcag.ca/efish/) was developed. It uses BLAST results of the DNA sequences as probes. GEBRET could be used as a eFISH by connecting a BLAT (BLAST-like alignment tool) program instead of BLAST.

To evaluate usefulness of GEBRET as the eFISH, we com-

pared an eFISH result obtained from GEBRET with an experimental FISH data set of HERV-W which was recently reported (Kim et al., 2008). HERV-17 (HERV-W) subfamily is one of the most ubiquitous HERV elements and has been identified through successive overlapping cDNA clones from multiple sclerosis patients and human placental tissues (Blond et al., 1999; Komurian-Pradel et al., 1999). As shown in Supplementary Figure 1, 72% of the experimental HERV-W FISH data is consistent with *in silico* analysis of GEBRET. Therefore, this program can be utilized as a recently updated eFISH version to rapidly illustrate the distribution of retroelements on each chromosome.

The retroelements including HERVs, *Alu*s and SVAs have played a dynamic role in the evolution of primate genomes. To carry out a comparative analysis of human, chimpanzee, orangutan, rhesus macaque, and marmoset genomes, we developed the GEBRET web-based tool. This tool visualizes the physical distribution of the retroelements and their adjacent genes. In addition, this study showed that GEBRET can be used as an eFISH. Thus, we believe that this software will aid studies aimed at investigating the impact of retroelements on primate genome evolution.

Acknowledgements We are grateful for funding provided by a Korea Science and Engineering Foundation grant (MOST; No. R01-2007- 000-20035-0).

References

Baban S, Freeman JD and Mager DL (1996) Transcripts from a novel human KRAB zinc finger gene contain spliced Alu and endogenous retroviral segments. Genomics 33: 463-472.

Batzer MA and Deininger PL (2002) Alu repeats and human genomic diversity. Nat. Rev. Genet. 3: 370-379.

Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A and Tristem M (2004) Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U.S.A. 101: 4894-4899.

Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A and Tristem M (2007) Rate of recombinational deletion among human endogenous retroviruses. J. Virol. 81: 9437-9442.

Blond JL, Besème F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B and Mallet F (1999) Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family J. Virol. 73: 1175-1185.

Brookfield JF (2001) Selection on Alu sequences? Curr. Biol. 11: R900-1.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87

Corvelo A and Eyras E (2008) Exon creation and establishment in human genes. Genome Biol. 9: R141.

Davuluri RV, Suzuki Y, Sugano S, Plass C and Huang TH (2008) The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 24: 167-177.

Ewing B and Green P (2000) Analysis of expressed sequence tags indicates 35,000 human genes. Nat. Genet. 25: 232-234.

Feuchter-Murthy AE, Freeman JD and Mager DL (1993) Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene. Nucleic Acids Res. 21: 135-143.

Goodchild NL, Freeman JD and Mager DL (1995) Spliced HERV-H endogenous retroviral sequences in human genomic DNA: evidence for amplification via retrotransposition. Virology 206: 164-173.

Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, Liang P and Batzer MA (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet. 3: 1939-1949.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

Jamain S, Girondot M, Leroy P, Clergue M, Quach H, Fellous M and Bourgeron T (2001) Transduction of the human gene FAM8A1 by endogenous retrovirus during primate evolution. Genomics 78: 38-45.

Juang SH, Huang J, Li Y, Salas PJ, Fregien N, Carraway CA and Carraway KL (1994) Molecular cloning and sequencing of a 58-kDa membrane- and microfilament-associated protein from ascites tumor cell microvilli with sequence similarities to retroviral Gag proteins. J. Biol. Chem. 269: 15067-15075.

Kato N, Pfeifer-Ohlsson S, Kato M, Larsson E, Rydnert J, Ohlsson R and Cohen M (1987) Tissue-specific expression of human provirus ERV3 mRNA in human placenta: two of the three ERV3 mRNAs contain human cellular sequences. J. Virol. 61: 2182-2191.

Kim DS, Cho CY, Huh JW, Kim HS and Cho HG (2009) EVOG: a database for evolutionary analysis of overlapping genes. Nucleic Acids Res. 37: D698-702.

Kim HS, Kim DS, Huh JW, Yi JM, Lee JR, Hirai H and Cho HG (2008) Chromosomal distribution, genomic feature and functional implication of the HERV-W family in humans. Kor. J. Genet. 30: 63-72.

Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, et al. (1998) An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature 394: 388-392.

Komurian-Pradel F, Paranhos-Baccala G, Bedin F, Ounanian-Paraz A, Sodoyer M, Ott C, Rajoharison A, Garcia E, Mallet F, Mandrand B, et al. (1999) Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology 260: 1-9.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. (2001) International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Landry JR, Mager DL and Wilhelm BT (2003) Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. 19: 640-648.

Ling J, Pi W, Bollag R, Zeng S, Keskintepe M, Saliman H, Krantz S, Whitney B and Tuan D (2002) The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. J. Virol. 76: 2410-2423.

Medstrand P, Landry JR and Mager DL (2001) Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. J. Biol. Chem. 276: 1896-1903.

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, et al. (2000) Syncytin is a

captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785-789.
Mighell AJ, Markham AF and Robinson PA (1997) Alu sequences. FEBS Lett. 417: 1-5.
Modrek B and Lee C (2002) A genomic view of alternative splicing. Nat. Genet. 30: 13-19.
Ostertag EM, Goodier JL, Zhang Y and Kazazian HH Jr (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. 73: 1444-1451.
Rhesus Macaque Genome Sequencing and Analysis Consortium (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316: 222-234.
Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, Batzer MA and Deininger PL (2001) Alu insertion polymorphisms for the study of human genomic diversity. Genetics 159: 279-290.
Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P and Batzer MA (2006) Human genomic deletions mediated by recombination between Alu elements. Am. J. Hum. Genet. 79: 41-53.
Smit AF, Tóth G, Riggs AD and Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. 246: 401-417.
Ting CN, Rosenberg MP, Snow CM, Samuelson LC and Meisler MH (1992) Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. 6: 1457-1465.
Voisset C, Weiss RA and Griffiths DJ (2008) Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease. Microbiol. Mol. Biol. Rev. 72: 157-196.
Woodley L and Valcárcel J (2002) Regulation of alternative pre-mRNA splicing. Brief. Funct. Genomic. Proteomic. 1: 266-277.
Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL and Batzer MA (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc. Natl. Acad. Sci. U.S.A. 103: 17608-17613.
Zhang J, Feuk L, Duggan GE, Khaja R and Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. 115: 205-214.