HERVd: database of human endogenous retroviruses

Jan Pačes, Adam Pavlíček and Václav Pačes*

Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, CZ-16637 Prague, Czech Republic

Received August 14, 2001; Accepted October 12, 2001

ABSTRACT

The human endogenous retroviruses database (HERVd) is maintained at the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, and is accessible via the World Wide Web at http://herv.img.cas.cz. The HERVd provides complex information on and analysis of retroviral elements found in the human genome. It can be used for searches of individual HERV families, identification of HERV parts, graphical output of HERV structures, comparison of HERVs and identification of retrovirus integration sites.

INTRODUCTION

Analysis of the human genome revealed that some 45% of it consists of various kinds of transposable elements. Around 8% of the human DNA is derived from retrovirus-like elements (1–3). They originate from ancient retroviral infections or are relics of retroviral transposomal activity in the germ-line cells. Human endogenous retroviruses (HERVs) comprise a part of these elements. They have undergone substantial changes such as mutations of all kinds, deletions and insertions of other transposons, recombinations and mini- and micro-satellite expansion. This is why it is often difficult to identify individual retroviral genes and other retroviral DNA regions.

HERVs are classified according to several criteria (4) that are, however, mostly artificial, especially in view of rearrangements and mutations changing the original retroviral DNA sequences.

HERVs became extremely useful tools in studying evolution and the plasticity of primate genomes (5). Some of them acquired important functions. For example, HERV long terminal repeats (LTRs) serve as transcription regulators, alternative promoters and polyadenylation signals for several cellular genes (6,7). Another function proposed for HERVs is in determining resistance to viral infections. The best example proposed for the HERV function is that the product of the HERV-W envelope gene on chromosome 7 is the human syncytin, which is a protein involved in the placenta formation (8,9). HERVs are likely to be cofactors in several diseases (10), such as multiple sclerosis (11), schizophrenia (12) and cancer (13).

It is important to analyze the structure and distribution of HERVs and compare them with situations in other genomes, e.g. those of primates and the mouse. These comparisons may help to find genomic elements that contribute to developmental phenotypical differences between humans and other organisms. Precise localization and characterization of insertion sites may be useful for designing retroviral vectors for gene therapy.

DATABASE

Construction of the database

The source for the database construction was output of the human genome project in NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch) and GoldenPath (http://genome.ucsc.edu). Only DNA sequences mapped at particular locations on individual chromosomes were included in the search. A total of 2935 Mb were analyzed.

Classification of HERV families is based on RepBase (1,14,15) that is designed specifically for repetitive DNA elements. This database contains, under different names, all recently described HERV families such as HERV-W, HERV-F and HERV-S, and also new families not yet reported.

The excellent RepeatMasker program of A. Smit and P. Green (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) was used for the primary search. This program is currently the standard program for identification of repetitive DNA elements in vertebrates, primarily in humans. RepeatMasker uses RepBase for retroelement identification. RepeatMasker is a useful tool for retroelements identification because it filters out other repeats containing integrases (LINEs).

Due to massive changes of individual HERVs by mutations, recombinations, expansion of satellites, etc., classification of various HERV portions within related families is difficult. We developed a defragmentation algorithm that makes it possible to define more precisely HERV families and to ascribe individual members to their families. This algorithm is based on the creation of fragment clusters of related families in a particular chromosomal location. Parts of each HERV are identified by HMMERs (16) corresponding to their families. Building of these HMMERs for all HERV families is reiterated to get greater precision.

With this tool, organization of HERVs in the particular chromosomal location was assessed. Then the flanking DNA regions are characterized. Whenever possible the target site duplication (TSD) was identified and its characteristics assessed. This made it possible to decide whether a retrovirus or processed pseudogene generated by LINE1 machinery was identified. The result of this search is the retroviral residue with its chromosomal location and characterization of its insertion site.

Programs used in this work were: RepBase version 6.1 (http://www.girinst.org), RepeatMasker-09/20/2000 (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker), Cross_match

*To whom correspondence should be addressed. Tel: +420 2 20183541; Fax: +420 2 24311019; Email: vpaces@img.cas.cz version 0.990329 and HMMER version 2.1.1. (http://hmmer.wustl.edu).

Database description

The database is accessible without restriction for academic research purposes at http://herv.img.cas.cz. It is based on repetitive elements and HERV portions identified by the RepeatMasker and processed by the algorithm described above. The current database version contains only families colinear with the typical retroviral genome (LTR-gag-pol(-env)-LTR). Non-autonomous families such as MaLR and MER4-group have not been included so far. The total number of HERV families is 39, with 90 corresponding LTRs. This comprises 78.9 Mb (2.7% of the genome). Out of this, 52.7 Mb (1.8% of the genome) are LTRs and 26.2 Mb (0.9% of the genome) are internal HERV sequences.

The database can be searched by HERV families, chromosomal locations and several other features. The graphic output of all HERVs and their portions is available.

The database will be free for downloading after the structure of human genome data is stabilized. Meanwhile, the database will be improved in accordance with the progress of human genome analysis and requests from the scientific community.

SUPPLEMENTARY MATERAL

An example of the HERV family analysis from the HERVd is available as Supplementary Material at NAR Online. This example shows the numerical and graphical output of one member of the HERV-H family.

ACKNOWLEDGEMENTS

This work was supported by the Center for Integrated Genomics and with grant no. M023 from the Grant Agency of the Czech Republic.

REFERENCES

1. Smit,A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, **9**, 657–663.
2. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Venter,J.D., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. and Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. Wilkinson,D. (1994) Human endogenous retroviruses. In Levy,J.A. (ed.), *The Retoviriadae.* Plenum Press, New York, NY, Vol. **3**, pp. 465–553.
5. Sverdlov,E.D. (2000) Retroviruses and primate evolution. *Bioessays*, **22**, 161–171.
6. Brosius,J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.
7. Schon,U., Seifarth,W., Baust,C., Hohenadl,C., Erfle,V. and Leib-Mosch,C. (2001) Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats. *Virology*, **279**, 280–291.
8. Mi,S., Lee,X., Li,X., Veldman,G.M., Finnerty,H., Racie,L., LaVallie,E., Tang,X.Y., Edouard,P., Howes,S., Keith,J.C.,Jr and McCoy,J.M. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, **403**, 785–789.
9. Blond,J.L., Lavillette,D., Cheynet,V., Bouton,O., Oriol,G., Chapel-Fernandes,S., Mandrand,B., Mallet,F. and Cosset,F.L. (2000) An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J. Virol.*, **74**, 3321–3329.
10. Lower,R. (1999) The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends Microbiol.*, **7**, 350–356.
11. Perron,H., Garson,J.A., Bedin,F., Beseme,F., Paranhos-Baccala,G., Komurian-Pradel,F., Mallet,F., Tuke,P.W., Voisset,C. and Blond,J.L. (1997) Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. *Proc. Natl Acad. Sci. USA*, **94**, 7583–7588.
12. Karlsson,H., Bachmann,X., Silke,B., Johannes,S., Justin,McA., Torrey,E.F. and Yolken,R.H. (2001) Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc. Natl Acad. Sci. USA*, **98**, 4634–4639.
13. Griffiths,D.J. (2001) Endogenous retroviruses in the human genome sequence. *Genome Biol.*, **2**, 1017.1–1017.3.
14. Jurka,J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, **8**, 333–337.
15. Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **9**, 418–420.
16. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.