D498–D500 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki044

L1Base: from functional annotation to prediction of active LINE-1 elements

Tobias Penzkofer[2], Thomas Dandekar[2] and Tomasz Zemojtel[1,2,*]

[1]Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics,
Ihnestrasse 73, D-14195 Berlin, Germany and [2]Department of Bioinformatics, University of Wuerzburg,
Am Hubland, D-97074 Wuerzburg, Germany

Received August 13, 2004; Revised and Accepted September 28, 2004

ABSTRACT

L1Base is a dedicated database containing putatively active LINE-1 (L1) insertions residing in human and rodent genomes that are as follows: (i) intact in the two open reading frames (ORFs), full-length L1s (FLI-L1s) and (ii) intact ORF2 but disrupted ORF1 (ORF2-L1s). In addition, due to their regulatory potential, the full-length (>6000 bp) non-intact L1s (FLnI-L1s) were also included in the database. Application of a novel annotation methodology, L1Xplorer, allowed in-depth annotation of functional sequence features important for L1 activity, such as transcription factor binding sites and amino acid residues. The L1Base is available online at http://l1base.molgen.mpg.de. In addition, the data stored in the database can be accessed from the Ensembl web browser via a DAS service (l1das.molgen.mpg.de:8080/das).

INTRODUCTION

Long interspersed elements (LINE-1, L1s) are the only active autonomous retrotransposons (1) in mammals, covering as much as 18% of their genomes. L1s' activity results in a great repertoire of actions, such as gene disruption (2), transcriptional regulation (3), alternative splicing (4), creation of exons and gene coding regions (5) and amplification of the processed pseudogenes and the *Alu* SINE family (6–8).

The full-length mammalian L1 is ∼6000 bp long and is composed of the 5′-untranslated region (5′-UTR) bearing an internal promoter, two open reading frames (ORF1 and ORF2) separated by intergenic region and 3′-UTR containing a poly(A) tail (1). The ORF1 is a non-sequence-specific RNA binding protein (1) and the ORF2 harbors three domains involved in L1 retrotransposition activity: endonuclease, reverse transcriptase and a 3′ terminal zinc finger-like domain (1).

The great majority of L1 insertions are truncated in the 5′ regions and/or contains various insertions/deletions. Still, it is the full length, intact in the two ORFs, 5′-UTR-located internal promoter and 3′-UTR regions LINE-1s (FLI-L1s), which are the most likely to display the retrotransposition activity. Interestingly, it has been recently proposed that L1 insertions containing a disrupted ORF1 gene but an intact ORF2 (ORF2-L1s) may be competent for the mobilization of *Alu* sequences (7).

Another class of L1s contributing to the genomic content and functionality are the retrotransposition-inactive, full length, non-intact due to multiple mutations LINE-1s (FLnI-L1s). A population of those may have retained an ability to be expressed and, although at a low frequency (9), could be retrotransposed by the proteins encoded by retrotransposition-active FLI-L1s. One of their regulatory potentials is embraced within the 5′-UTR located antisense promoter, which, when intact, may be capable of guiding the expression of many genes (10).

It, therefore, becomes an important task to identify and functionally characterize the FLI-L1s, ORF2-L1s and FLnI-L1s residing in mammalian genomes. This task can only be accomplished, when detailed analyses of conservation of the sites known to be important for L1 activity can be executed on the genomic scale. With this motivation we built the L1Base.

DATABASE

Data acquisition

To identify and functionally annotate the two types of putatively active L1 insertions residing in mammalian genomes, FLI-L1s and ORF2-L1s, we created and utilized the novel annotation methodology, L1Xplorer. Briefly, the L1Xplorer is a suite of perl scripts, which are designed to detect L1 insertions either by performing genomic BLAST searches (11) with the L1 template sequence as a query (i.e. *Homo sapiens* L1.2, gi: M80343) or analyzing the Repeatmasker annotation [provided by Ensembl (12)]. During a series of

[*]To whom correspondence should be addressed. Tel: +49 30 8413 1147; Fax: +49 30 8413 1152; Email: zemojtel@molgen.mpg.de

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use permissions, please contact journals.permissions@oupjournals.org.

© 2005, the authors

Table 1. Statistics on the current content of the L1Base

Species	Ensembl version	No. of FLnI-L1s	No. of FLI-L1s	No. of ORF2-L1s	No. of features annotated[a]
H.sapiens	v23.35	6389	151	99	46
M.musculus	v24.33	5713	1501	261	20
R.norvegicus	v23.3c	5236	372	163	16

[a] Refer to the Supplementary Table 1 for details.

tests, we established that the sensitivity of BLAST searches (BLAST parameters: -p blastall –f F, *E*-value threshold of E−10) is sufficient, when mining for L1s harboring the intact ORFs. After extraction of genomic region corresponding to L1 insertion, L1Xplorer detects the two LINE-1 ORFs, checks on their intactness and recognizes a number of experimentally characterized features important for activity on the LINE-1 sequence (such as the transcription factor binding sites) using HMM-profiles (HMMER versions 1.8.4 and 2.3.2) (13), TFASTX program of the FASTA suite (14) and ClustalW (15) alignments. In addition, it carries out family classifications based on diagnostic residues, located in the 5′- and 3′-UTR. Supplementary Table 1 lists the recognized sites including specific classifications for human, mouse and rat L1 elements.

Database description

The database contains sequences along with annotations produced by the L1Xplorer for the three classes of L1 elements residing in the human, mouse and rat genomes: (i) identified by the L1Xplorer putatively active FLI-L1s, (ii) ORF2-L1s and (iii) identified by applying RepeatMasker full length (>6000 bp), and classified by L1Xplorer as non-intact, retrotransposition-inactive, L1s (FL-nIL1s).

The functional annotation of the LINE-1 loci produced by the L1Xplorer was further complemented by SNPs (dbSNP) (16), repeat [RepeatMasker (17)] and coding genes annotation, as available in the recent version of Ensembl (12) (*H.sapiens* Ensembl v23.34e; *Mus musculus*, Ensembl v24.33; and *Rattus norvegicus* Ensembl v23.3c). The Table 1 contains a summary of the current L1Base content.

User interface

L1Base can be searched via the MySQL-driven query system by using criteria, such as conservation of the functional sites important for activity (for details see the Supplementary Table 1), chromosomal localization and families.

A user can take advantage of MySQL regular expressions and Boolean AND/OR operators to compose complex queries. The database can also be searched by executing Blastn-based (11) queries with a user-specified L1 sequence. A detailed display mode (DDM), activated each time a user points to any particular result of a query, allows for an easy identification of all annotated features on LINE-1 sequence via a graphical interface utilizing color-coding schemes.

Data export

The results of queries can be exported in the comma separate value (CSV), Fasta and GeneBank formats. While in the DDM, the database entry can be exported to fasta and tinyseq-xml formats.

Links to other resources

Each entry of the database is html cross-linked to the Ensembl genome web browser (12).

Database access

The L1Base is freely available through http://l1base.molgen.mpg.de. In addition, the annotation data stored in the L1Base can be accessed from the Ensembl genome web browser via a DAS service (l1das.molgen.mpg.de:8080/das).

Case study

Mouse LINE-1s are characterized by the presence of a variable number of 200 bp long repeats called monomers in the 5′-UTR region. It has been shown that monomers possess promoter activity and that increasing their number increases the level of transcription (18,19). Therefore, using the number of the monomers as a criteria, we searched the L1Base for putatively highly expressed full-length, intact L1s (FLI-L1s) belonging to the young G(F) subfamily (20) (query executed using regular expression: ‘G.F-monomer*’, with the ‘Search Monomers’ option selected). As a result, L1s with the following L1Base IDs (FLI-L1 Database): 692, 636 and 113 were identified as the top three hits, with the ID 692 having as many as 13 monomers.

Perspectives

LINE elements are autonomously active and cover a substantial fraction (up to 18%) of mammalian genomes. Given their extent, a precise and full annotation of their specific features and subclasses is helpful for a better understanding of mammalian genomes, their evolution and encoded activities (e.g. promoter activity in the case study). We plan on extending our annotation to other mammalian genomes, exploiting the ongoing progress of sequencing projects (by including i.e. chimpanzee and cat genomes). Meanwhile, the database will be improved with respect to the total number of features annotated on rodent L1 sequences. Finally, we aim for automatic updates of the L1Base.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

T.Z. thanks Prof. Martin Vingron for helpful discussions and support. This work has been a part of BioSapiens project funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’, contract number LHSG-CT-2003-503265.

REFERENCES

1. Furano,A.V. (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 255–294.

2. Kazazian,H.H.,Jr, Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) *Haemophilia* A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
3. Han,J.S., Szak,S.T. and Boeke,J.D. (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature*, **429**, 268–274.
4. Kondo-Iida,E., Kobayashi,K., Watanabe,M., Sasaki,J., Kumagai,T., Koide,H., Saito,K., Osawa,M., Nakamura,Y. and Toda,T. (1999) Novel mutations and genotype–phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). *Hum. Mol. Genet.*, **8**, 2303–2309.
5. Nekrutenko,A. and Li,W.H. (2001) Transposable elements are found in a large number of human protein-coding genes *Trends Genet.*, **17**, 619–621.
6. Kazazian,H.H.,Jr (2000) Genetics. L1 retrotransposons shape the mammalian genome. *Science*, **289**, 1152–1153.
7. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.*, **35**, 41–48.
8. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.*, **24**, 363–367.
9. Wei,W., Gilbert,N., Ooi,S.L., Lawler,J.F., Ostertag,E.M., Kazazian,H.H., Boeke,J.D. and Moran,J.V. (2001) Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell. Biol.*, **21**, 1429–1439.
10. Nigumann,P., Redik,K., Matlik,K. and Speek,M. (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, **79**, 628–634.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
13. Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 114–120.
14. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
15. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
16. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
17. Smit,A.F., Toth,G., Riggs,A.D. and Jurka,J. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.*, **246**, 401–417.
18. Severynse,D.M., Hutchison,C.A.,III and Edgell,M.H. (1992) Identification of transcriptional regulatory activity within the 5′ A-type monomer sequence of the mouse LINE-1 retroposon. *Mamm. Genome*, **2**, 41–50.
19. DeBerardinis,R.J. and Kazazian,H.H.,Jr (1999) Analysis of the promoter from an expanding mouse retrotransposon subfamily. *Genomics*, **56**, 317–323.
20. Goodier,J.L., Ostertag,E.M., Du,K. and Kazazian,H.H.,Jr (2001) A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.*, **11**, 1677–1685.