D68–D73 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 18 October 2016*
doi: 10.1093/nar/gkw925

L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes

Tobias Penzkofer[1], Marten Jäger[2], Marek Figlerowicz[3], Richard Badge[4], Stefan Mundlos[2], Peter N. Robinson[2,5] and Tomasz Zemojtel[2,3,*]

[1]Department of Radiology, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany, [2]Institut für Medizinische Genetik und Humangenetik, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany, [3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, 60-569 Poznan, Poland, [4]Department of Genetics, University of Leicester, Leicester, LE1 7RH, UK and [5]The Jackson Laboratory for Genomic medicine, 10 Discovery Drive, Farmington, CT 06032, USA

Received August 13, 2016; Revised September 28, 2016; Accepted October 05, 2016

ABSTRACT

LINE-1 (L1) insertions comprise as much as 17% of the human genome sequence, and similar proportions have been recorded for other mammalian species. Given the established role of L1 retrotransposons in shaping mammalian genomes, it becomes an important task to track and annotate the sources of this activity: full length elements, able to encode the *cis* and *trans* acting components of the retrotransposition machinery. The L1Base database (http://l1base.charite.de) contains annotated full-length sequences of LINE-1 transposons including putatively active L1s. For the new version of L1Base, a LINE-1 annotation tool, L1Xplorer, has been used to mine potentially active L1 retrotransposons from the reference genome sequences of 17 mammals. The current release of the human genome, GRCh38, contains 146 putatively active L1 elements or full length intact L1 elements (FLIs). The newest versions of the mouse, GRCm38 and the rat, Rnor_6.0, genomes contain 2811 and 492 FLIs, respectively. Most likely reflecting the current level of completeness of the genome project, the latest reference sequence of the common chimpanzee genome, PT 2.19, only contains 19 FLIs. Of note, the current assemblies of the dog, CF 3.1 and the sheep, OA 3.1, genomes contain 264 and 598 FLIs, respectively. Further developments in the new version of L1Base include an updated website with implementation of modern web server technologies. including a more responsive design for an improved user experience, as well as the addition of data sharing capabilities for L1Xplorer annotation.

INTRODUCTION

Long interspersed elements class 1 (LINE-1s, L1s) are the active autonomous non-LTR retrotransposons in mammalian genomes. L1s are present in a large number of copies, resulting in them making up about 17% of the human genome (1). Due to the ability of L1s to 'copy' and 'paste' themselves into multiple genomic locations they have had a significant impact on genome and organismal evolution (2). In addition to insertional mutagenesis, because of features such as an anti-sense promoter located in the human and mouse L1s, which can drive the transcription of adjacent genes (3,4) L1 elements can interfere with genomic content and functionality in numerous ways; transcriptional disruption (5,6) alternative splicing and exonization (4,7,8), as well as the creation of processed pseudogenes (9), and mobilization of the high copy number *Alu* SINE sequence family (10,11) that themselves contain gene regulatory elements (12–14). While direct disruption of exons by L1 insertion is a well established cause of human genetic disease (15), recently it has been shown that intronic insertions of nearly full-length L1 elements can result in recessive genetic diseases, when inherited from both parental alleles (16). This observation emphasizes the need for current annotation of these structural variants, which are often assumed to be benign.

As LINE-1s have been active in mammalian genomes since before the mammalian radiation (17–19), the human genome contains 16 distinct L1 families (L1PA16–L1PA1), which have gradually evolved from the mammalian radiation to the present day (17,18,20). Interestingly, only members of the young Ta, ACA/G (L1PA1) and pre-Ta, ACG/G subfamilies, defined by the shared sequence variants (SSVs) AC[A/G] and G in their 3′ UTRs, display retrotransposition activity *in vivo* (21–23).

It has been suggested that the typical human genome contains between 80–100 retrotransposition-active LINE-1s (24), which is in line with the number of ∼145 putatively

*To whom correspondence should be addressed. Tel: +49 30 450 566 306; Fax: +49 30 450 569 915; Email: tomasz.zemojtel@charite.de

Table 1. Number of annotated LINE-1 Elements in the 2007 and the current 2016 releases of reference genomes

Species	FLI-L1s		ORF2-L1s		FLnI-L1s	
Year	**2007**	**2016**	**2007**	**2016**	**2007**	**2016**
Primates						
Homo sapiens*	145	146	103	107	11 653	13 418
Pongo abelii (Orangutan)	-	20	-	22	-	7601
Pan troglodytes (Chimp)	-	19	-	24	-	12 218
Macaca mulatta (Macaque)	-	20	-	21	-	6605
Gorilla gorilla (Gorilla)	-	0	-	1	-	7074
Callithrix jacchus (Marmorset)	-	6	-	4	-	8251
Chlorocebus sabaeus (Vervet)	-	6	-	1		7841
Papio anubis (Baboon)	-	12	-	32	-	22 389
Rodents						
Mus musculus** (Mouse)	2382	2811	466	563	13 692	14 076
Rattus norvegicus*** (Rat)	377	492	183	292	5236	10 073#
Carnivora						
Canis familaris (Dog)	-	264	-	57	-	9653
Felis catus (Cat)	-	1	-	0	-	8075
Bovidae						
Ovis aries (Sheep)	-	598	-	181	-	3551
Bos taurus (Cow)	-	16	-	20	-	2989
Others						
Equus caballus (Horse)	-	72	-	37	-	6766
Sus scrofa (Pig)	-	0	-	0	-	4495
Oryctolagus cuniculus (Rabbit)	-	0	-	0	-	4296

*NCBI36 versus GRCh38, **NCBIm35 versus GRCm38, ***RGSC 3.4 versus Rnor_6.0, # due to updated Repbase consensus sequences for rat, the length threshold for Rnor_6.0 FLnI was lowered from 6000 nt to 5500 nt.

active L1s residing in the reference genome (25). More recently, it has been also suggested that highly active or 'hot' LINE-1 elements are more frequent in the human population than previously assumed, but are under ascertained due to presence/absence polymorphism (21). Ongoing and future studies will likely continue to report novel human-specific polymorphic LINE-1s and database resources, like euL1db (26), will aid greatly in collecting these L1 insertions. Such resources will reveal the frequency and distribution of L1 insertions in increasing detail, but due to limitations of the underlying technology most often used (short read NGS) lack detailed annotation.

To gain more insight into a repertoire of active LINE-1 elements in mammalian genomes we employed L1Xplorer (25), to catalog and most importantly, characterize with respect to functionality and phylogeny, two types of potentially active L1 insertions: (i) full length L1s (FLI-L1s) with intact ORF1 and ORF2 (ii) L1s with intact ORF2 but disrupted ORF1 (ORF2-L1s) that may still be able to drive Alu elements (10). We also annotated full length non-intact insertions, which constitute evolutionary fossils in mammalian genomes. The current version of L1Base (25), in addition to updated annotations for the human, mouse and rat genomes, also annotates seven other primate genomes, including the common chimpanzee genome and another seven mammalian genomes (e.g. cow, horse, dog).

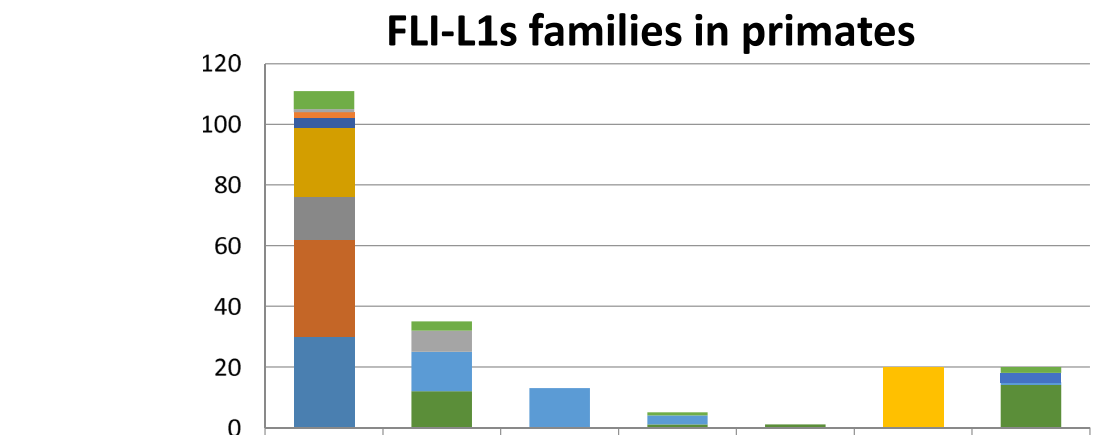
DATABASE

Data acquisition

The most recent available (Ensembl Version 84) releases of reference genome assemblies for species already annotated in L1Base: *Homo sapiens, Mus musculus, Rattus norvegicus,* as well as newly added species: *Pan troglodytes, Macaca mulatta, Pongo abelii, Gorilla gorilla* and *Bos Taurus* (PT 2.19, MMUL 1.0, PPYG2, gorGor3.1, UMD3.1) were downloaded from the Ensembl internet resource (27) along with the MySQL annotation databases containing the current genomic and RepeatMasker (http://www.repeatmasker.org, (28)) annotations for the respective genome versions. A standalone version of L1Xplorer was executed to generate the L1Base annotations for these genomes. As described previously (25) L1Xplorer is a set of Perl applications performing (i) a BLAST search using a known L1 template for a given species (i.e. Homo sapiens L1.2, GenBank: AH005269.2) or predefined coordinates from genomic RepeatMasker annotations (ii) extraction of the corresponding genomic regions and detection of ORF1 and ORF2 via TFASTX (29) and (iii) annotation of a number of features specific to L1 lineages, family classifications and predicted or experimentally characterized features which have been implicated in L1 biology. The L1Xplorer outputs, along with the previous contents of L1Base were stored in a web-accessible interface.

Database description

L1Base contains three different LINE-1 categories: (i) full-length L1s with intact ORF1 and ORF2 (FLI-L1s), (ii) L1s with intact ORF2 but disrupted ORF1 (ORF2-L1s) and full length non-intact L1s (FLnI-L1s). The sequences of full length non-intact L1s (FLnI-L1) were extracted from the Ensembl RepatMasker annotations with the following thresholds: (i) Mus musculus: 5000 bp, (ii) Rattus norvegicus: 5500 bp and (iii) for Homo sapiens, Pan troglodytes, Macaca mulatta, Pongo Abelii, Gorilla gorilla, Callithrix jacchus, Papio anubis, Felis catus, Canis familiaris, Bos taurus, Ovis aries, Sus scrofa and Orycolagus cuniculus the threshold of 4500 bp was used. For Equus caballus and Chlorocebus sabaeus, the 4500 bp threshold for En-



Species-Repeat Masker	HS-L1HS	HS-L1PA2	PT-L1Pt	PT-L1PA2	PT-L1PA3	MM-L1PA5	PA-L1PA3
others	6	3		1			2
AAGA-Ta-0							3
GAGA-L1PA5						20	
ACGA-non-canonical	1	7					
ACGA-Ta-0	2						
GAGA-non-canonical		13	13	3			1
GAGA-Ta-0		12		1	1		14
ACGG-non-canonical	3						
ACGG-Ta-0	23						
ACAG-Ta-1nd	14						
ACAG-Ta-1d	32						
ACAG-Ta-0	30						

Figure 1. Annotation of FLI-L1s elements present in four primate genomes. HS: *Homo sapiens*, PT: *Pan troglodytes*. MM: *Macaca mulatta*, PA: *Pongo abelii*, Repeatmasker annotations: L1PA2, L1Pt, L1PA2,3,5. The first column lists annotations obtained by employing L1Xplorer.

sembl RepeatMasker annotations resulted in an insignificant number of FLnIs (Equcab: 0, Chlsab: 101) and thus we used the L1Xplorer tool to extract FLnI-L1s from these two genomes. Currently, genomes of 17 species are annotated: eight primates, two rodents and seven other mammals (Table 1). It can be observed that in accordance with the maturity of the reference genome sequence, the number of annotated FLI-L1s, ORF1-L1s and FLnIs increases, often dramatically (Table 1). Interestingly, the current reference genomes of the dog and the sheep contain as many as 264 and 598 FLIs, respectively. The latter is suggestive of the high rate of L1 activity in these two species.

User interface

The entry point into the L1Base website is the pull down menu for species selection. Upon species selection, the user is brought to the query form, where queries involving multiple criteria such as chromosomal localization, integrity of ORFs or conserved motifs can be executed.

Data export/sharing

In order to support the user tracks feature, as implemented by different genome browser sites, L1Base now supports a .BED file export for every database provided, thus replacing the outdated DAS protocol. A link to the respective .BED data source is provided by the main L1Base database list. These .BED annotations can be easily included in custom UCSC genome browser views, to enable integration with other annotation data (https://genome.ucsc.edu/goldenpath/help/customTrack.html) (30).

L1Xplorer, the annotation tool provided with L1Base, has been improved to support data sharing between users by implementing a share link feature available at the top of a webpage with specific L1 annotation.

DATABASE ACCESS

L1base has been moved from http://l1base.molgen.mpg.de to http://l1base.charite.de.

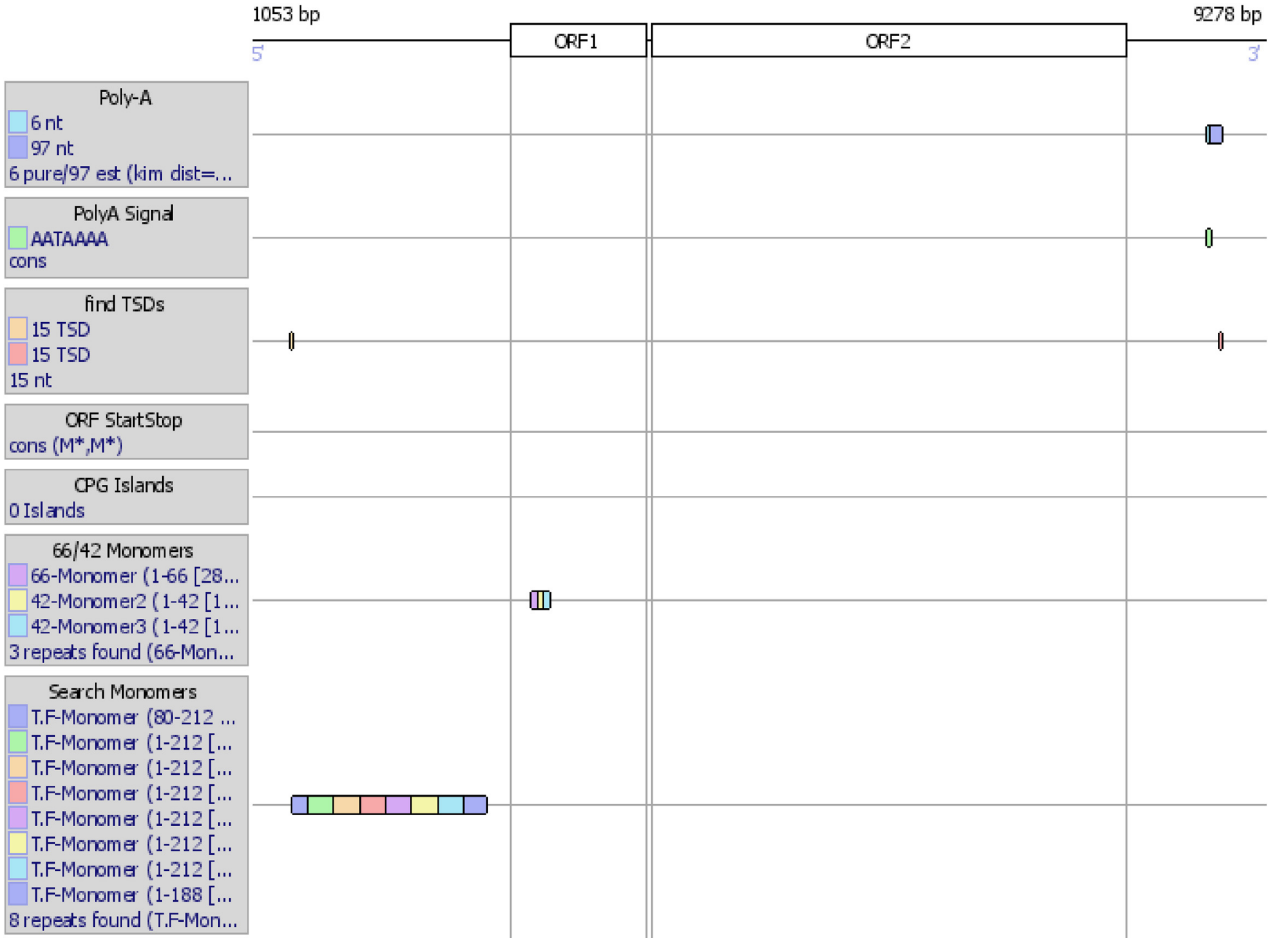


Figure 2. L1Xplorer-based annotation of the L1spa element.

Case study 1: Identification of the full length intact L1s (FLI-L1s) in four primate genomes

The current release of the human genome, GRCh38, contains 146 FLI-L1s (Table 1). The L1Xplorer-based family classifications (see Figure 1) revealed that 76 (∼52%) of them belong to the *Homo sapiens*-specific young Ta (diagnostic nucleotides ACA/G) family (31). Interestingly, a substantial number, 25 (∼17%), of the FLI-L1s found belong to the older L1PA2 (diagnostic nucleotides GAG/A) element class, which amplified during the period of primate radiation (31).

Likely reflecting the current stage of genome sequencing and assembly, we discovered only ∼20 FLIs in the 3 primate genomes: chimpanzee, orangutan and rhesus macaque. L1Xplorer annotations, as they are stored in L1Base, compared to RepeatMasker annotations, allow further phylogenetic subcategorization (Figure 1). For example, the human FLI-L1s annotated by RepeatMasker simply as L1HS, can be subdivided by L1Xplorer into more than seven subcategories based on the annotated features (Figure 1).

Case study 2: Annotation of active mouse L1spa element

The insertion of LINE-1 element (GenBank: AF016099.1), L1spa, into intron 6 of *Glrb* has been associated with the spastic mouse phenotype (32–34). In order to annotate it, we executed L1Xplorer with the advanced option: extend locus by 2000 nt. Since L1Xplorer detects the 5′ UTR monomers (including A-Monomer I, A-Monomer II, A-Monomer III, A-Monomer IV, A-Monomer V, A-Monomer VI, T.F-Monomer, F-Monomer, G.F-Monomer) (32), this analysis revealed that the L1spa element belongs to the T.F family (Figure 2). In detail, L1Xplorer analysis showed that the sequence of the first detected T.F monomer is missing the last 24 nt, the next six T.F monomers are of full length and the last one is missing the first 79 nt, as compared to the template of the T.F monomer. The L1spa element has one copy of a 66 bp repeat, and two copies of a 42 bp repeat, in the length polymorphism region of ORF1 (35,36). The most similar FLI mouse sequence to the L1spa element resides on chr.12: 91946103-91935703 (GRCm38, L1Base ID: 2590), as identified by executing the Blast search function of L1Base. The L1Base ID: 2590 entry shows 99.95% identity, and this element differs by only 4 nt (7498/7502 nt), from L1spa. We might reasonably conclude that this element is the most likely progenitor of the L1spa insertion, highlighting the utility of detailed annotation in exposing LINE-1 biology.

ADDITIONAL FEATURES

In order to implement more modern web technologies and to improve accessibility, the L1Base website code was updated to support responsive design elements. In brief, the static HTML4 technology was replaced by HTML5/Javascript using state-of-the-art libraries (Bootstrap, JQuery). This ensures seamless adaptation to different device classes (tablet, personal computer) and screen resolutions, as well as improving the overall user experience.

PERSPECTIVES

With the development of sequencing technologies enabling long-read generation (37), and the concomitant increasing inclusion of long dispersed repeats, like L1 elements, in assemblies of individual genomes, polymorphic L1 transposon insertions will become much better represented in publically available databases. L1Base will aim to catalog the ongoing expansion of LINE-1 variation, with a focus on the putatively active L1 repertoire in personal genomes, offering the potential for data and annotation resources that more realistically represent the true extent and diversity of active L1elements segregating in populations.

FUNDING

Funding for open access charge: Institutional support.
Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Deininger,P.L., Moran,J.V., Batzer,M.A. and Kazazian,H.H. Jr (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.*, **13**, 651–658.
3. Speek,M. (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.*, **21**, 1973–1985.
4. Zemojtel,T., Penzkofer,T., Schultz,J., Dandekar,T., Badge,R. and Vingron,M. (2007) Exonization of active mouse L1s: a driver of transcriptome evolution? *BMC Genomics*, **8**, 392.
5. 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.
6. 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.
7. 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.
8. 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.
9. Zemojtel,T., Duchniewicz,M., Zhang,Z., Paluch,T., Luz,H., Penzkofer,T., Scheele,J.S. and Zwartkruis,F.J. (2010) Retrotransposition and mutation events yield Rap1 GTPases with differential signalling capacity. *BMC Evol. Biol.*, **10**, 55.
10. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.*, **35**, 41–48.
11. Kazazian,H.H. Jr (2000) Genetics. L1 retrotransposons shape the mammalian genome. *Science*, **289**, 1152–1153.
12. Zemojtel,T., Kielbasa,S.M., Arndt,P.F., Behrens,S., Bourque,G. and Vingron,M. (2011) CpG deamination creates transcription factor-binding sites with high efficiency. *Genome Biol. Evol.*, **3**, 1304–1311.
13. Zemojtel,T., Kielbasa,S.M., Arndt,P.F., Chung,H.R. and Vingron,M. (2009) Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends Genet.*, **25**, 63–66.
14. Polak,P. and Domany,E. (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics*, **7**, 133.
15. Beck,C.R., Garcia-Perez,J.L., Badge,R.M. and Moran,J.V. (2011) LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.*, **12**, 187–215.
16. Kagawa,T., Oka,A., Kobayashi,Y., Hiasa,Y., Kitamura,T., Sakugawa,H., Adachi,Y., Anzai,K., Tsuruya,K., Arase,Y. *et al.* (2015) Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. *Hum. Mutat.*, **36**, 327–332.
17. Furano,A.V. (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 255–294.
18. 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.
19. Burton,F.H., Loeb,D.D., Voliva,C.F., Martin,S.L., Edgell,M.H. and Hutchison,C.A. 3rd (1986) Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. *J. Mol. Biol.*, **187**, 291–304.
20. Boissinot,S. and Furano,A.V. (2001) Adaptive evolution in LINE-1 retrotransposons. *Mol. Biol. Evol.*, **18**, 2186–2194.
21. Beck,C.R., Collier,P., Macfarlane,C., Malig,M., Kidd,J.M., Eichler,E.E., Badge,R.M. and Moran,J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, **141**, 1159–1170.
22. Moran,J.V. (1999) Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica*, **107**, 39–51.
23. Skowronski,J., Fanning,T.G. and Singer,M.F. (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.*, **8**, 1385–1397.
24. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H. Jr (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 5280–5285.
25. Penzkofer,T., Dandekar,T. and Zemojtel,T. (2005) L1Base: from functional annotation to prediction of active LINE-1 elements. *Nucleic Acids Res.*, **33**, D498–D500.
26. Mir,A.A., Philippe,C. and Cristofari,G. (2015) euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.*, **43**, D43–D47.
27. Aken,B.L., Ayling,S., Barrell,D., Clarke,L., Curwen,V., Fairley,S., Fernandez Banet,J., Billis,K., Garcia Giron,C., Hourlier,T. *et al.* (2016) The Ensembl gene annotation system. *Database (Oxford)*, doi:10.1093/database/baw093.
28. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.*, **110**, 462–467.
29. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
30. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
31. Ovchinnikov,I., Rubin,A. and Swergold,G.D. (2002) Tracing the LINEs of human evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 10522–10527.
32. Naas,T.P., DeBerardinis,R.J., Moran,J.V., Ostertag,E.M., Kingsmore,S.F., Seldin,M.F., Hayashizaki,Y., Martin,S.L. and Kazazian,H.H. (1998) An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J.*, **17**, 590–597.
33. Mulhardt,C., Fischer,M., Gass,P., Simon-Chazottes,D., Guenet,J.L., Kuhse,J., Betz,H. and Becker,C.M. (1994) The spastic mouse: aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. *Neuron*, **13**, 1003–1015.

34. Kingsmore,S.F., Giros,B., Suh,D., Bieniarz,M., Caron,M.G. and Seldin,M.F. (1994) Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. *Nat. Genet.*, **7**, 136–141.
35. 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.
36. Schichman,S.A., Adey,N.B., Edgell,M.H. and Hutchison,C.A. 3rd (1993) L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution. *Mol. Biol. Evol.*, **10**, 552–570.
37. Chaisson,M.J., Huddleston,J., Dennis,M.Y., Sudmant,P.H., Malig,M., Hormozdiari,F., Antonacci,F., Surti,U., Sandstrom,R., Boitano,M. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608–611.