GENOME-WIDE DISTRIBUTION AND LOCALIZATION OF PUTATIVE FUNCTIONAL HUMAN LINE-1 RETROTRANSPOSONS

Christine Steinhoff [1,3], Wolfgang A. Schulz [2]

[1] Computational Molecular Biology,
Max-Planck-Institute for Molecular Genetics, Berlin, Germany
[2] Urologische Klinik, Biologisch-Medizinisches Forschungszentrum,
Heinrich-Heine-Universität Düsseldorf, Germany

Keywords: LINE, repetitive elements, functional retrotransposon, genome-wide localization

[3] Corresponding author:
Max Planck Institute for Molecular Genetics
Dept Computational Molecular Biology
Ihnestraße 73; D-14195 Berlin; Germany
Email: christine.steinhoff@molgen.mpg.de
Phone: +49 30 8413 1171
Fax: +49 30 8413 1152

Abstract. Three human LINE families comprise 20.4% of the human genome. LINE-1 sequences with 55 subfamilies account for 16.9% and contain all retrotransposons for which autonomous retrotransposition has been documented although most L1 elements are non-functional. While it is known that there are ∼7000 elements in the human genome, the number and distribution of autonomously active LINE-1 elements are less certain. We scanned the draft sequence of the human genome for the essential functional parts, viz. promoter, ORF1 and ORF2. These fragments were assembled by allowing gaps of varying sizes between promoter and ORF1 or between ORF1 and ORF2. This procedure reduces the number of potentially active LINE-1 elements from overall searches (∼7000) to 177 potentially autonomously active elements including previously described functional LINE-1 elements. Intact elements are apparently stochastically distributed in the genome, with the potential exception of the X chromosome. Unexpectedly, plots of gap sizes between promoter and ORF1 and ORF1 and between ORF2 revealed that while the distribution of intact LINE-1 parts is also random, their distance is not. This list of candidates of autonomously active LINE -1 elements and their exact position within the human genome provides a basis for functional analyses of retrotransposition.

1. Introduction

Transposable elements are classified into elements transposing via DNA intermediates (transposons) or RNA intermediates (retroelements), respectively. The most prevalent classes of transposable elements in the human genome are all retroelements, long interspersed elements (LINEs) represented predominantly by LINE-1 (L1) sequences, small interspersed elements (SINEs) represented predominantly by ALU sequences, LTR retrotransposons, which resemble retroviruses, represented primarily by human endogenous retrovirus sequences (HERVs). The latter two classes are unlikely to include autonomously active elements in humans while recent transposition of LINEs has been documented in a number of cases, in disease (1), (2), (3), (4) as well as in cell culture (5), (6). The published draft of the human genome (7) predicted 45% of the human DNA sequence to consist of transposable element sequences and a fraction of 16.9% to consist of LINE-1 elements (7). LINE-1 amplification during human evolution has been examined in detail recently (8), (9), (10). It appears that after the divergence of humans from their closest relatives the LINE-1 family has further expanded in the genome. LINE elements are believed to be present in all mammals and have a great impact on mammalian genomes, but may also impinge on their regulation (11) as follows. First, LINEs contain their own retrotransposition machinery which is thought to also enable Alu transposition (for example SINEs) and the creation of processed pseudogenes. In fact, a considerable number of SINEs survive by exploiting the LINE retrotranspositional mechanism (7), (12). Second, LINE-1 retrotransposons are capable of transducing neighboring sequences. Sequences including coding exons adjacent to active LINE elements can be shuffled to new sites (4). Third, it has been proposed that LINE-1 elements significantly influence the regulation of surrounding genes (13). Fourth, in a number of cases LINE elements have been associated with human diseases. An ancient retrotransposition underlies Fukuyama muscular dystrophy (14). Recent transpositions include insertion of an L1H element into the gene encoding factor VIII in two independent patients (1), an L1H insertion into the MYC gene in a breast adenocarcinoma (2), and the disruption of the APC gene in a colon cancer (3). Furthermore, transcription of LINE elements is strongly activated in teratocarcinomas (15), (16) and to a lower degree in other tumors (17), (18) and may contribute to genomic instability. This activation is likely facilitated by hypomethylation of the promoter region of active LINE elements frequent in human cancers (19).

The distribution of LINE elements within the human genome has been

considered in various publications (for an overview see: (7), (20). It appears that LINE elements are rather evenly distributed on different chromosomes with the prominent exception of the X chromosome. Three types of LINEs are found in the human genome, but it is believed that only LINE-1 elements are still active (7). Within this family approximately 55 closely related subfamilies can be distinguished. Overall, 516,000 LINE copies have been reported corresponding to a fraction of 16.9% of the draft genome sequence (7). In the ensembl database (21) 1,265,498 LINEs are presently annotated (v.7.29 a.3; July, 12^{th} 2002), among them 846,411 LINE-1, of which 6,761 are apparently full-length (i.e. $>$ 6000 bps). Intact full-length elements contain an internal 5 promoter sequence, and two open reading frames, ORF1 encoding an RNA-binding protein possibly required for translation and ORF2 encoding the reverse transcriptase (RT) and endonuclease essential for retrotransposition. However, even most full-length elements are not capable of autonomous retrotransposition due to internal mutations, although fragmented elements can occasionally transpose through trans-complementation by autonomous active elements (22).

It has been estimated that only around 60-100 LINE-1s per diploid genome are active (23), (7). A very recent search for Ta elements which are considered the most active family, was based on a 19 nt consensus sequence and yielded 124 elements while only for 40 elements intact ORF1 was found (10). However, this search would miss intact elements that do not exhibit the consensus in the 3‘ UTR whose functional relevance is uncertain. However, we demonstrate that for the vast majority of 177 outative active elements significant matches for the 19 nucleotide consensus sequence in the 3‘ UTR can be found. Simple BLAST searches using the full length sequence on the other hand may miss elements with gaps in non-essential regions, while identifying those with inactivating point mutations. We therefore chose a new approach. First, we derived a consensus sequence from 18 LINE-1 elements for which at least RT activity has been demonstrated or that have transposed very recently (24), (25), (23), (26). We searched the human genome for putative active LINE-1 elements by searching each part, promoter, ORF1 and ORF2 separately with very strict settings and merging of the parts allowing variable gap lengths between them. In this fashion, we restricted the list of around 7000 full-length elements to 177 LINE-1 elements which are likely to retain retrotransposition potential. Interestingly, their distribution relative to the full set suggests that the initial set of LINEs was more evenly distributed, and subsequently altered by spread, deletion and insertion. This evolutionary mechanism appears to have included a certain extent of specificity

which led to specific chromosomes (and regions) such as X containing a very high percentage of LINE-1 sequences.

2. Results

2.1. **Comparison of published full length LINE-1 elements and derivation of a representative consensus sequence.** To derive a consensus sequence for potentially active LINE-1 elements in the human genome, we selected 18 full-length active LINE-1 sequences from the literature (Table 1). Of these, two inserted into the β-globin gene and retinitis pigmentosa-2 gene, resp. (25), (24). The others, including L1.3 (accession number L19088), L1.4 (L19092), L1.19 (U93568), L1.20 (U93569) and L1.39 (U93574) were shown to encode at least active RT and/or to be capable of retrotransposition in HeLa cells (23). We used ClustalW to align the sequences and to derive a consensus sequence (27). The alignment of these 18 full length elements showed at least 98% similarity and no mismatch was longer than 2 bps, which indicates very close relationship. All differences between individual sequences towards the consensus sequence were due to point mutations, but not to gaps, insertions, deletions or inversions. Overall, all sequences showed at most 2% dissimilarity to the derived consensus sequence (Table 1). We therefore assume that all active full-length LINE-1 retrotransposons are closely related and show only small mutational differences within promoter, ORF1 and ORF2, which can be accounted for by allowing point mutations in genome-wide searches. The consensus sequence is available at: `http://www.molgen.mpg.de/~steinhof/LINE`.

2.2. **Databank Search.** To determine the distribution of active LINE-1 elements in the human genome, the following strategy was used. We first searched for the three essential functional parts, i.e. promoter, ORF1 and ORF2, separately with very strict settings only allowing point mutations but no deletions or insertions using BLAST. In the following we define gap length to be the DNA sequence not examined by BLAST search between promoter and ORF1 and between ORF1 and ORF2 with respect to the definition of promoter, ORF1 and ORF2 given below. We performed BLAST searches along the draft of the human sequence (goldenPath version 28^{th}June 2002). From the output we selected only full length matches, allowing 1% discrepancy in length. Merging these parts using variable gap lengths between them gives a view of the distribution of LINE-1 elements depending on the gap length. This describes a LINE-1 element dependent on the length x of the non-aligned gap between promoter and ORF1 and the length y of the non-aligned gap between ORF1 and ORF2. Of these, only those

with appropriate gap lengths remain candidates for ability to transpose autonomously.

The definition of the essential promoter region was based on the report by Swergold et al. (28), describing a sequence of 661 bp in the 5‘ UTR region as essential for full promoter activity of the LINE element which is in accord with our own studies of the L1.2B promoter (unpublished data). The corresponding region from the consensus sequence was therefore defined as "promoter". Based on the literature, we defined ORF1 to comprise bps 913-1927 and ORF2 bps 1991-5818 in the consensus sequence. With this selection for each matched part or the full length elements, we examined the following parameters: (A) We looked for correlation between the total number of elements found on each chromosome and (i) the length of the chromosome, (ii) known CpG islands, (iii) number of known genes, (iv) number of known ALU sequences and (v) number of annotated LINE sequences. Annotations of known genes, ALU sequences and LINE sequences were obtained from the ensembl annotation (21). The results of the correlation study are illustrated in figure 1. We considered the distribution of the elements within each chromosome and relative to CpG islands (Figure 1). Distributions of neither promoter nor ORF1 nor ORF2 sequences correlated with those of CpG islands (B), genes (C) or ALUs (D), while they correlated well with chromosomal length (A) and LINE sequences annotated in the ensembl database (E-G). In particular, correlation of the essential parts with annotated full length LINE-1 elements was high and stronger than towards annotated LINE-1 of any length or LINEs of all families overall, as would be expected (F-G).

2.3. **Potentially active LINE-1 elements in the human genome.**

In order to find potential autonomously active elements we assembled the fragments from the retrieval described above. The resulting elements should be members of the LINE-1 class with high retrotransposition potential. Thus we searched for successive fragments in the order promoter-ORF1-ORF2 on both strands as a function of gap lengths (in the sense of gap length definition given above) between promoter and ORF1 and between ORF1 and ORF2. We examined the gap lengths in 50 bps steps up to a gap length of 1000 bps between promoter and ORF1 and up to 500 bps between ORF1 and ORF2. For comparison, the consensus sequence shows 252 bps between promoter and ORF1 and 63 bps between ORF1 and ORF2. The number of elements depending on the gap width are displayed as 3-D plots in Figure 2. Here we displayed the number of elements found on each human chromosome as a function of the gap length between promoter and ORF1

and between ORF1 and ORF2. Thus, gap lengths in 50 bps between promoter and ORF1 are displayed on the x-axis, gap lengths between ORF1 and ORF2 on the y-axis and the number of elements we find for the respective values on the x- and y-axis on the z-axis.

These plots show some features which are unexpected if one assumes that LINE-1 elements and fragments are essentially randomly spread in the human genome. First, there is a small plateau at gap lengths of 250-300 bps (promoter-ORF1) and 100 bps (ORF1-ORF2) for almost all chromosomes. Above that gap length an abrupt increase in the number of assembled elements is detected. Assuming that functional elements show a gap length of 250-300 bps between promoter and ORF1 and $\sim$ 100bps between ORF1 and ORF2 the low first plateau comprises all functional elements. If we further assume that there is a random event of spreading, truncating and deleting LINE elements at random sites we would expect that the 3-D function depending on both parameters of gap length displayed on the x- and y-axis to be continuously increasing at almost constant slope. In fact, there are plateaus at very specific gap lengths sizes which are similar for all chromosomes. Secondly, after this first increase in the number of assembled elements, some chromosomes, such as the smaller chromosomes 16, 17, 19, 20, 21, 22 and Y, show almost no further increase. This finding is not compatible with random deletion of individual parts of LINE-1 elements on these chromosomes. Finally, while after the first step the number of elements continues to grow steadily with increasing gap length on many chromosomes, e.g. chromosomes 4 and 6, some display further discontinuities, e.g. chromosome 13.

From the 3-D plots (Figure 2) it is clear that the gap width settings of 300 and 100 bps, resp., yield a discrete group of elements not within the range of random assembly of element parts. Obviously, this setting still overestimates the true number of active LINE-1 elements but ought to comprise all in the available human sequence. The distribution of elements extracted from this search along the length of each chromosome is displayed in Figure 3a. Interestingly, no candidates were identified on chromosomes 21 and Y. As in the searches using full-length sequences (7), our approach yielded more then twice as many potentially active LINE-1 elements on chromosome X (11.2 elements per 108 bps) than on the average autosome (mean 5.1 elements per 108 bps). Because of the high variance (mean 5.1 per 108 bps; standard deviation (SD) 3.0), this value lies within a range of 2 standard deviations around the mean. This is graphically shown in figure 3b.

A complete list of all 177 elements including their chromosomal position is available at: `http://www.molgen.mpg.de/~steinhof/LINE/`.

As a further check of the procedure we searched for elements of which the chromosomal position is documented in the list of putative active elements and confirmed the positions for those with accession numbers U09116; U93563; U93565 and U93572. Finally, we checked from each chromosome (apart from chromosome 21 and Y) a subset of the putative active elements found in this study for their annotation in ensembl. In fact, in all cases the respective sequences were annotated as L1-elements of full length. A summary of this search is displayed in table 2. All putative active elements showed at least a pairwise similarity of 88% using the ClustalW algorithm. Dissimilarities were mainly due to sequencing gaps in the human draft sequence. For all elements we found in this study we searched for the 19 nt consensus sequence (published in (10)) in the 3‘ UTR characteristic of the Ta subfamily. In fact 24.3% showed perfect matches and 53.7% showed only mismatches for either the last nucleotide or the last three nucleotides (3‘ end). Interestingly, almost all mismatches concerning the last nucleotide at the 3‘ end were due to a A to G change while for mismatches concerning the three nucleotides we regularly found GAG instead of ACA. Furthermore, 10.7% showed one further mismatche and for and for 8.5% we found 1-3 additional single nucleotide mismatches. Only 2.8% contained multiple mismatches an two elements displayed unsequenced stretches at the positions of the consensus sequence. The results from the alignment are available at: `http://www.molgen.mpg.de/~steinhof/LINE/`.

3. Discussion

In previous searches for LINE-1 sequences in the human genome (7), (29), full-length elements were examined after selection for elements comprising around 6 kb with a consensus sequence similar to the one used here. However, to obtain an indication on the potential functionality of these elements, differences in their sequence must be weighted according to the sites where they appear. Evidently, deletions or point mutations within the promoter, ORF1 or ORF2 sequence will have a larger impact on functionality than those in connecting sequences or the 3’ UTR. Searching for the essential parts first and assembling them in a second step allows much stricter BLAST settings and a better selection for potentially active elements. This method leads to the restriction of $\sim$ 7000 full length LINE-1s (7) to 177 putative functional elements which comprise all those already described as functional in the literature.

Approximately half of the elements in our list belong to the Ta subfamily. A large group of our elements differ by a common exchange either of

the nucleotide at the very 3‘ end or in the three nucleotides at the very 3‘ end. Thus, there is a significant group of full-length elements with intact ORFs and promoter not fitting the Ta consensus. The compilation obtained here can now be used in molecular assays to better define the actual requirements for function. For instance, it is not exactly clear which gap sizes are compatible with retrotransposon function. In order to get all putative functional elements while minimizing the number of false negatives we allowed gaps of 300 bps between promoter and ORF1 and 100 bps between ORF1 and ORF2. There may therefore be some false positives in this collection. This is difficult to ascertain, because the effect of gap size is not known but can now be studied on this set of elements. An overview of all putative functional elements is available at: `http://www.molgen.mpg.de/~steinhof/LINE`.

The chromosomal localization of the resulting putative functionally active elements shown in figure 3a suggest that their distribution is random. Thus, there is no positive or negative correlation with the distribution of either CpG islands or annotated genes that would indicate a requirement for chromosomal environment to maintain function over evolutionary times, in accord with (30). The chromosomal environment may still restrict the actual retrotransposition function, e.g. by influencing promoter DNA methylation (unpublished data).

As in previous searches (31), (29), (7), there are indications in the present investigation that more intact LINE-1 elements as well as element parts are present on the X chromosome, although this enrichment is within the range of 2 standard deviations. This distribution may reflect evolutionary mechanisms. Likely, an initial overall equal distribution has been destroyed by events specific for sequence or chromatin structure conditional for either insertion, recombination or deletion events. The lack of full-length LINE-1 elements on the smallest chromosomes 21 and Y, on the other hand, could well be due to chance, i.e. a low frequency would be expected for chromosomes of this size.

An unexpected finding revealed by the procedure applied in this study is the appearance of plateaus in the function displaying the number of elements depending on the length of the gaps between promoter and ORF1 as well as between ORF1 and ORF2, instead of the expected continuous distribution. The non-stochastic increase in the number of assembled elements does not fit the assumption of random deletion or integration of LINE-1 elements. For almost all chromosomes but chromosome 21 and Y, a unique increase in the number of assembled elements occurs at 250-300 bps (promoter-ORF1) and 100 bps (ORF1-ORF2). These plateaus do not occur on chromosome 21 and Y, where no active element could be detected. Here, the threshold is 1000 bps

(promoter-ORF1) and 300 bps (ORF1-ORF2) for chromosome 21 and 700 bps (promoter-ORF1) and 900 bps (ORF1-ORF2) for chromosome Y. Obviously, these large gaps make it unlikely that the successive promoter, ORF1 and ORF2 sequences are part of the same element. This first plateau may reflect the border line between autonomously active elements and non-autonomous parts. Furthermore, on several chromosomes further increases in the number of assembled elements above the first threshold are also not stochastic. Overall, these findings suggest a specificity in the mechanism by which clusters of LINE-1 sequences are created or destroyed leading to an overall depletion in active elements.

4. Materials and Methods

4.1. **Derivation of a Consensus Sequence.** Sequences of 18 LINE elements were downloaded from GenBank (32) according to the published accession numbers (Table 1). The consensus sequence was extracted using GAP v4.6 (Staden package (version 4.4), (33)). Alignments for comparison with the extracted consensus sequence were performed using ClustalW (27). The consensus sequence is available at: `http://www.molgen.mpg.de/~steinhof/LINE`

4.2. **Database Search.** In this study the NCBI assembled human sequence from April 05th, 2002 GenBank (goldenPath version 28^{th} June 2002) was used. For the parts: part 1: -193/+661 (promoter); part 2: 913-1927 (ORF1); part 3: 1991-5818 (ORF2) of the consensus sequence separate BLAST searches against the human genome were performed using the following settings: expectation value: 0.01, cost to open a gap: 20, and cost to extend a gap: 10^4. Fragments of lengths: promoter $>$ 654 nt; ORF1 $>$ 1003 nt and ORF2 $>$ 3788 nt were filtered and subjected to further analysis. These lengths correspond to full length of either of the parts 1, 2 or 3 with up to 1% variation in length.

4.3. **Extraction of putative functional elements by assembly.** Assembly of the parts 1, 2 and 3 of the elements with their localization was examined using MATLAB (Mathworks, Inc., Version 6.0.0.88 Release 12, Sept 2000). For this purpose, the localization of each fragment obtained from the BLAST search was used and parts were assembled according to their gap lengths between part 1 and part 2 or between part 2 and part 3 on both strands. Analysis of the distribution of sequence parts, graphical features, localization of fragments on the chromosomes, analysis of gap lengths and statistical analyses were also programmed in MATLAB. The algorithm is available upon request.

5. Acknowledgements

We thank Antje Krause for fruitful discussion and critical reading of the manuscript.

6. Figure legends

Figure 1 Correlation of fragments of LINE-1 consensus sequence parts found by BLAST searches relative to chromosomal length (A), CpG islands (B), annotated genes (C), annotated ALUs (D), annotated LINEs (E), annotated full-length LINE-1 (F). Information about CpG islands, genes, ALUs and LINEs as well as their localization in the human genome were obtained from the ensembl database. For each chromosome, chromosome length, number of CpG islands, number of annotated genes, annotated LINEs, annotated LINE-1, full length LINE-1s were plotted vs. the number of either promoter, ORF1 or ORF2 parts found by BLAST search. Correlation (r) was calculated $(Cov(x_{ij})/\sqrt{(}Cov(X_{ii})Cov(X_{jj})))$, where (x_{ij}) is the matrix of number of promoter, ORF1, ORF2 or full length element vs. either of the variables CpG islands, annotated genes, annotated ALUs, annotated LINEs, annotated LINE-1, full length LINE-1.

Figure 2 Display of the number of assembled elements (z-axis) depending on gap widths between promoter and ORF1 (x-axis) or ORF1 and ORF2 (y-axis). Gap lengths increase in steps of 50 bps.

Figure 3 Distribution of potentially active LINE-1 elements
(a) Genomic localization of potentially active LINE-1 elements with gap lengths of up to 300 bps between promoter and ORF1 and up to 100 bps between ORF1 and ORF2 and localization of annotated CpG islands. For each chromosome (C) the upper row indicates CpG islands, the lower localization of potentially active LINE-1 elements. Arrows mark the positions of four representative elements described in the literature: LRE2 (Chromosome 1) (4), L1.6 (Chromosome X) (23), L1.12 (Chromosome 18) citeSassaman1997, L1.25 (Chromosome 1) (23).
(b) Plot of chromosome length versus number of putative active elements found by allowing gap widths of 300 bp between promoter and ORF1 and up to 100 bp between ORF1 and ORF2. Human chromosomes are marked by "C" followed by chromosomal number or X, Y resp.

7. TABLES

Table1: Result from the alignment of each of the indicated sequences with the consensus sequence.

Nr	Accession Number Name	Chromo-some Number	5´UTR	ORF1	ORF2 % match (# mismatch)*	Reference
1	AF148856 L1 RP	-	99% (3)	99% (3)	99% (8)	[18]
2	AF149422 L1 b-thal	-	99% (3)	98% (6)	98% (19)	[19]
3	U93562 L1.5	11	98% (7)	98% (6)	98% (24)	[20] [16]
4	U93571 L1.24	12	98% (7)	98% (12)	98% (29)	[20] [16]
5	U93573 L1.33	20	99% (1)	99% (3)	98% (12)	[20] [16]
6	L19088 L1.3	14	99% (4)	99% (4)	98% (10)	[20]
7	L19092 L1.4	9	99% (5)	99% (3)	98% (13)	[20]
8	U09116 LRE2	1	99% (3)	98% (9)	98% (17)	[4]
9	U93563 L1.6	X	98% (10)	99% (5)	98% (27)	[16]
10	U93564 L1.8	14	99% (1)	99% (1)	98% (22)	[16]
11	U93565 L1.12	18	99% (6)	99% (5)	98% (16)	[16]
12	U93566 L1.14	X	99% (2)	99% (4)	98% (18)	[16]
13	U93567 L1.15	5	99% (3)	99% (3)	99% (9)	[16]
14	U93568 L1.19	7	98% (10)	99% (2)	98% (18)	[16]
15	U93569 L1.20	20	99% (4)	99% (5)	98% (20)	[16]
16	U93570 L1.21	n.d.	98% (8)	99% (5)	98% (34)	[16]
17	U93572 L1.25	n.d.	98% (11)	98% (6)	98% (24)	[16]
18	U93574 L1.39	14	99% (5)	99% (2)	98% (20)	[16]

* The consensus sequence was generated by comparing the sequence of the 18 full length elements shown. For each sequence the similarity to the consensus is given in percentage of matched base pairs within 5´UTR, ORF1 and ORF2 each. Numbers of mismatches are shown in brackets.

Table 2: List of putative autonomously active LINE-1 elements for which site was verified using ensembl (http://www.ensembl.org)

Chr*	Nr†	prom start‡	pred. start‡	Length[1]	Name[2]
1	1.1	119841703	119841701	6029	L1HS
1	1.2	245708125	245708125	6031	L1HS
1	1.3	114453178	114453178	6011	L1HS
1	1.4	72477906	72477903	6032	L1PA2
1	1.5	72103442	72103403	6044	L1HS
2	2.1	165900157	165900157	6036	L1HS
2	2.2	195958931	195958928	6031	L1HS
2	2.3	156156620	156156621	6032	L1PA2
2	2.4	164397636	164397637	6029	L1PA2
2	2.5	157310690	157310687	6032	L1PA2
3	3.1	105262191	105262191	6025	L1HS
3	3.2	154773080	154773077	6024	L1HS
3	3.3	78664337	78664337	6028	L1PA2
3	3.4	105712753	105712718	6031	L1HS
4	4.1	80553571	80553571	6029	L1HS
4	4.2	59682318	59682318	6030	L1HS
4	4.3	79197313	79197311	6033	L1HS
4	4.4	93056714	93056711	6202	L1HS
4	4.5	14618743	14618742	6030	L1HS
5	5.1	110261109	110261109	6024	L1HS
5	5.2	39966007	39966007	6019	L1PA2
5	5.3	73745410	73745409	6030	L1PA2
5	5.4	101124265	101124262	6029	L1HS
5	5.5	151525245	151525245	6001	L1HS
6	6.1	19822661	19822661	6026	L1HS
6	6.2	121254168	121254168	6009	L1HS
6	6.3	83990010	83990007	6029	L1HS
6	6.4	117258608	117258608	6032	L1HS
6	6.5	116116511	116116509	6028	L1PA2
6	6.6	104772934	104772934	6031	L1PA2
7	7.1	30121341	30121338	6032	L1HS
7	7.2	15900428	15900426	6029	L1PA2
7	7.3	22210401	22210398	6035	L1PA3
8	8.1	58900807	58900806	6027	L1PA2
8	8.2	91049016	91049016	6019	L1PA2
8	8.3	88101397	88101397	6032	L1HS
8	8.4	97316885	97316892	6023	L1PA2
8	8.5	87588499	87588497	6032	L1PA2
9	9.1	104351577	104351574	6032	L1HS
9	9.2	83751577	83751577	6030	L1HS
9	9.3	89273441	ME[3]	ME[3]	L1HS
9	9.4	68083515	68083516	6030	L1HS
10	10.1	86072224	86072224	6032	L1HS
10	10.2	5334889	5334883	6031	L1HS
10	10.3	65495501	65495282	5976	L1PA3
11	11.1	87362537	87362534	6032	L1HS
11	11.2	62693555	62693555	6029	L1PA2
11	11.3	96100343	96100342	6031	L1PA2
11	11.4	97499769	97499698	6030	L1HS

11	11.5	95200195	95200123	6035	L1HS
12	12.1	90234332	90234330	6032	L1PA2
12	12.2	92100083	92100083	6025	L1PA2
12	12.3	62083874	62083871	6029	L1PA2
12	12.4	116660787	116660715	5850	L1HS
12	12.5	78692466	78692591	6057	L1PA2
13	13.1	29863060	29863058	6031	L1HS
13	13.2	40999345	40999342	6032	L1PA2
13	13.3	80587150	80586893	6031	L1PA2
14	14.1	68519050	68519044	6032	L1HS
14	14.2	46885143	46885143	6018	L1PA2
14	14.3	28208705	28208633	6031	L1PA2
14	14.4	77740107	77739850	6033	L1PA3
15	15.1	84189324	84189324	6029	L1HS
15	15.2	90661119	90661119	6029	L1PA2
15	15.3	78121091	78121087	6032	L1PA2
15	15.4	31831188	31831181	6030	L1PA3
15	15.5	53975285	53975068	6030	L1PA2
16	16.1	44490368	44490368	6028	L1HS
16	16.2	18014706	18014706	6020	L1HS
16	16.3	36109048	36109048	6026	L1HS
16	16.4	74519813	74519728	6044	L1HS
17	17.1	63816393	63816383	6016	L1HS
17	17.2	9870580	9870577	6016	L1HS
17	17.3	70697534	70697534	6011	L1HS
17	17.4	68210666	68210594	6031	L1HS
17	17.5	58951457	58951454	6023	L1PA2
18	18.1	45489985	45489987	6029	L1HS
18	18.2	73278584	73278584	6032	L1HS
18	18.3	54957741	54957784	6032	L1PA2
18	18.4	32820300	32820307	6030	L1PA2
19	19.1	38942715	38942713	6031	L1PA2
20	20.1	11601485	11601485	6025	L1HS
20	20.2	51817685	51817613	6035	L1PA2
22	22.1	25755371	25755368	6032	L1HS
X	X.1	53556713	53556713	6032	L1HS
X	X.2	68139155	68139155	5654	L1P1
X	X.3	141261544	141261544	6033	L1HS
X	X.4	68414002	68414002	6022	L1HS
X	X.5	124297077	124297076	6030	L1HS

* Chromosome

† Numbering index

‡ Genomic localization according to the human draft sequence from 28th June 2002 with starting nucleotide in bps from p to q end of the chromosome as predicted in this study and ‡ genomic localization as predicted in ensembl (http://www.ensembl.org).

1 Length and **2** name of the predicted element according to the human draft sequence from 28th June 2002.

3 Multiple elements were predicted in this region.

Table 3: Summary of 177 putative autonomously active LINE-1 elements.

Chromosome	# putative functional *	# putative functional LINE-1/10^8 bps †
1	12	4.86
2	10	4.15
3	11	5.64
4	17	8.85
5	15	8.29
6	14	8.22
7	4	2.54
8	13	9.04
9	5	3.78
10	3	2.23
11	11	8.00
12	9	6.85
13	5	4.41
14	5	4.79
15	6	6.05
16	5	6.12
17	6	7.50
18	4	5.16
19	1	1.67
20	2	3.18
21	0	0
22	2	4.19
X	17	11.39
Y	0	0
Total ‡	177	5.29 ± 2.90

*Total numbers of putative functional elements on each chromosome found in our study

† Number of putative functional elements per 10^8 bps

‡ Total number of elements, mean number of elements adjusted to genome length and standard deviation (SD)

Figure1

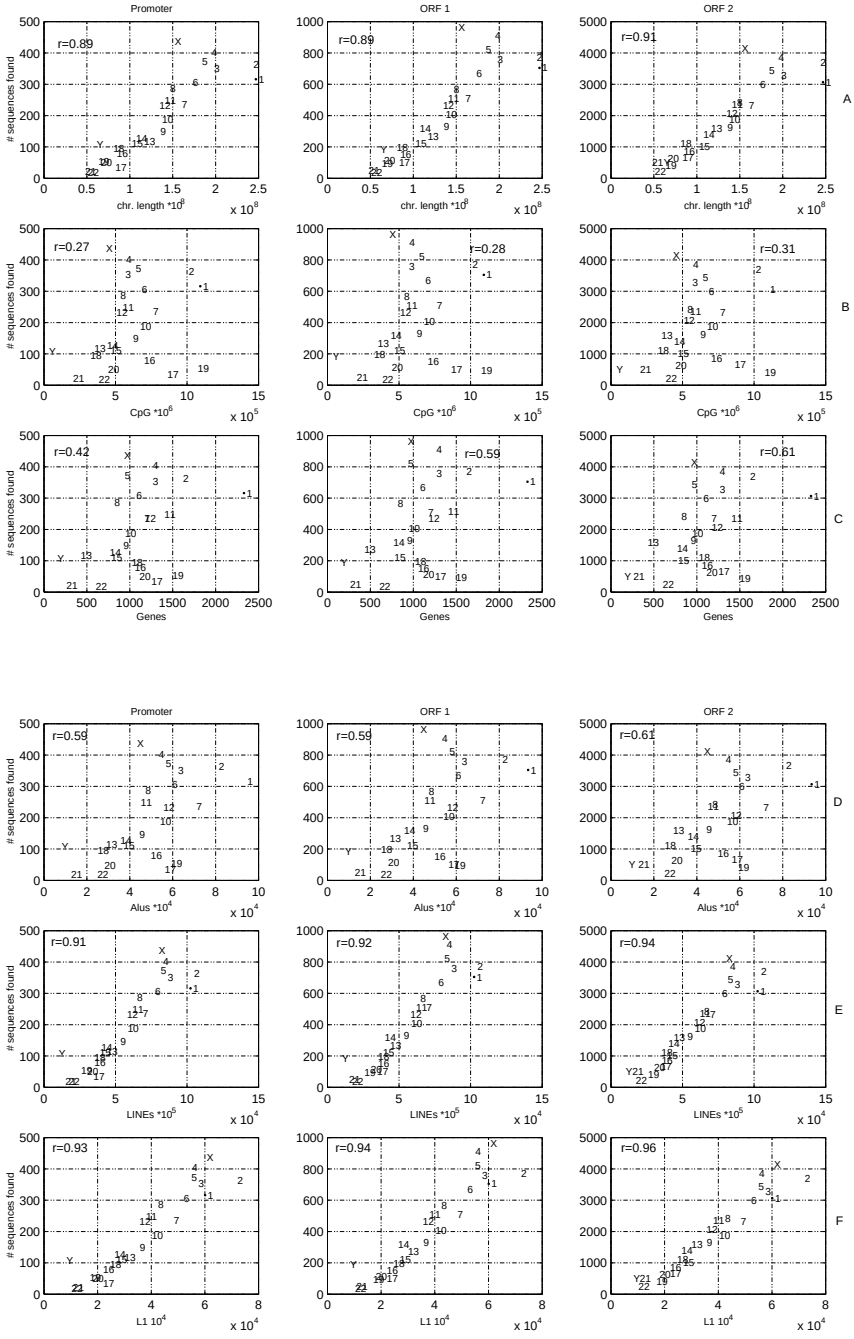

Promoter
ORF 1
ORF 2
r=0.89
r=0.89
r=0.91
A
sequences found
chr. length *10^8
x 10^8
r=0.27
r=0.28
r=0.31
B
CpG *10^6
x 10^5
r=0.42
r=0.59
r=0.61
C
Genes
Promoter
ORF 1
ORF 2
r=0.59
r=0.59
r=0.61
D
Alus *10^4
x 10^4
r=0.91
r=0.92
r=0.94
E
LINEs *10^5
x 10^4
r=0.93
r=0.94
r=0.96
F
L1 10^4
x 10^4

Figure 2

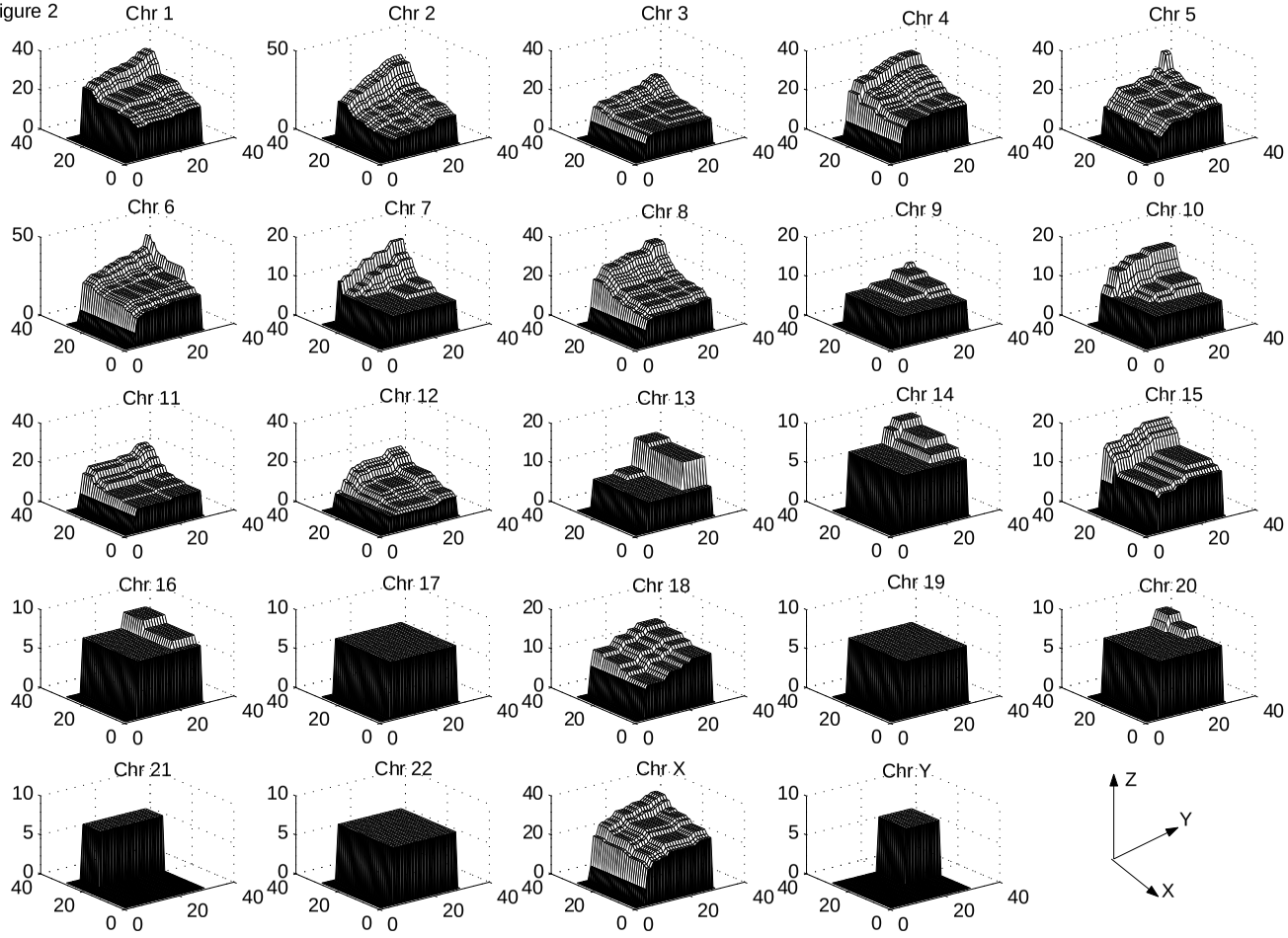

Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr X
Chr Y
Z
Y
X

Figure 3a

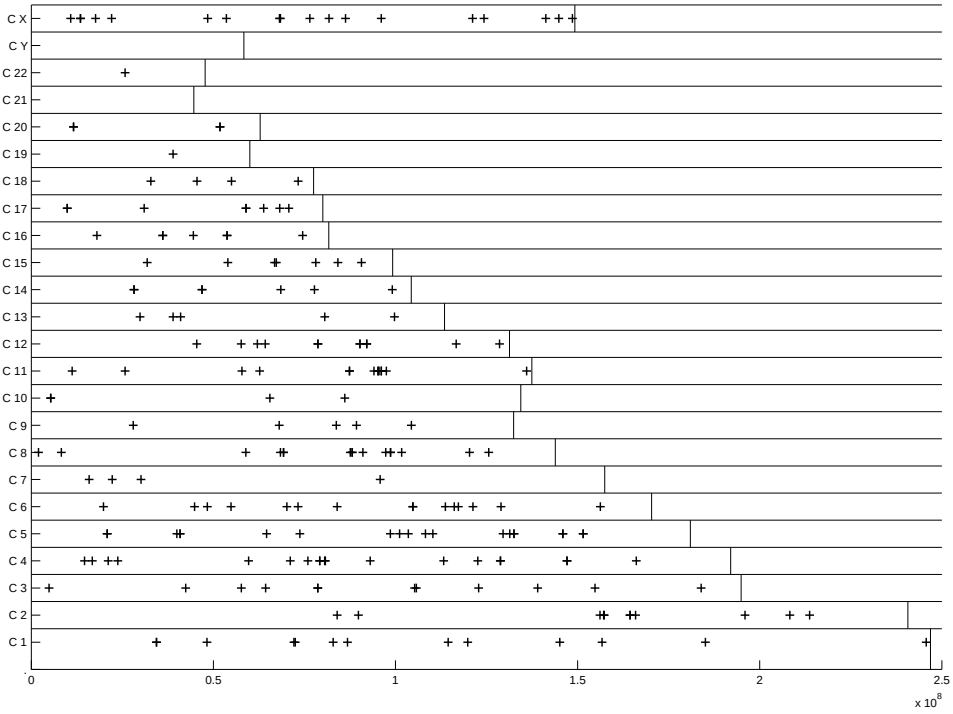
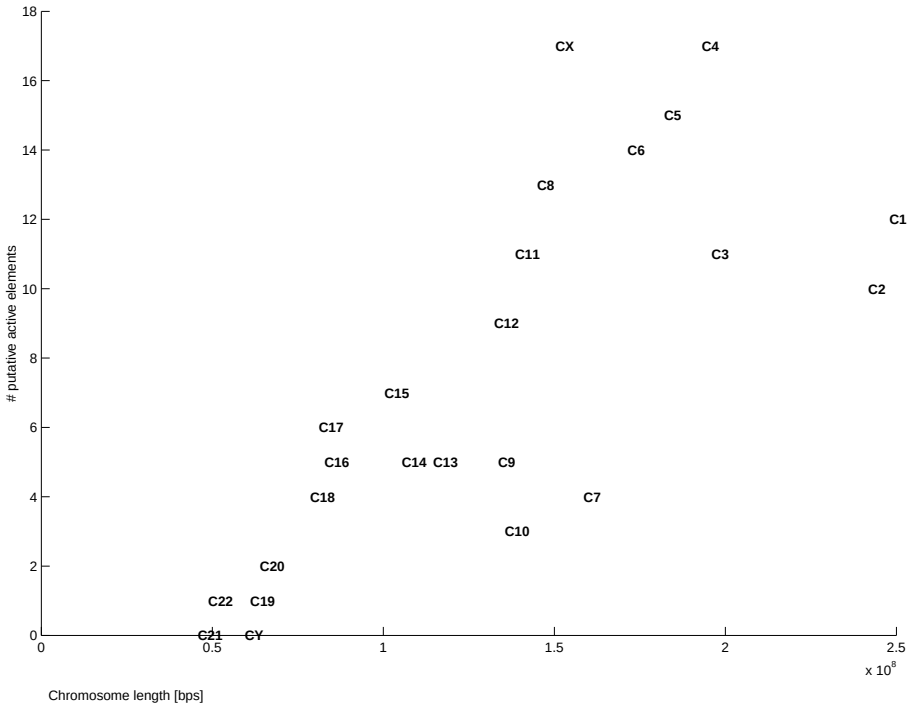


Figure 3b



References

[1] J. r. Kazazian HH, C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, S. E. Antonarakis, Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man., Nature 332 (6160) (1988) 164–166.

[2] B. Morse, P. G. Rotherg, V. J. South, J. M. Spandorfer, S. M. Astrin, Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma., Nature 333 (6168) (1988) 87–90.

[3] Y. Miki, I. Nishisho, A. Horii, Y. Miyoshi, J. Utsunomiya, K. W. Kinzler, B. Vogelstein, Y. Nakamura, Disruption of the apc gene by a retrotransposal insertion of L1 sequence in a colon cancer., Cancer Res 52 (3) (1992) 643–645.

[4] S. E. Holmes, B. A. Dombroski, C. M. Krebs, C. D. Boehm, J. r. Kazazian HH, A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion., Nat Genet 7 (2) (1994) 143–148.

[5] J. V. Moran, S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke, J. r. Kazazian HH, High frequency retrotransposition in cultured mammalian cells., Cell 87 (5) (1996) 917–27.

[6] T. P. Naas, R. J. DeBerardinis, J. V. Moran, E. M. Ostertag, S. F. Kingsmore, M. F. Seldin, Y. Hayashizaki, S. L. Martin, H. H. Kazazian, An actively retrotransposing, novel subfamily of mouse L1 elements., EMBO J 17 (2) (1998) 590–597.

[7] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S.

Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, J. Szustakowki, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, Initial sequencing and analysis of the human genome., Nature 409 (6822) (2001) 860–921.

[8] I. Ovchinnikov, A. Troxel, G. Swergold, Genomic characterization of recent human LINE-1 insertions: Evidence supporting random insertion, Genome Res 11 (2001) 2050–2058.

[9] I. Ovchinnikov, A. Rubin, G. D. Swergold, Tracing the LINEs of human evolution., Proc Natl Acad Sci U S A 99 (16) (2002) 10522–10527.

[10] J. S. Myers, B. J. Vincent, H. Udall, W. S. Watkins, T. A. Morrish, G. E. Kilroy, G. D. Swergold, J. Henke, L. Henke, J. V. Moran, L. B. Jorde, M. A. Batzer, A comprehensive analysis of recently integrated human Ta L1 elements., Am J Hum Genet 71 (2) (2002) 312–26.

[11] T. A. Morrish, N. Gilbert, J. S. Myers, B. J. Vincent, T. D. Stamato, G. E. Taccioli, M. A. Batzer, J. V. Moran, DNA repair mediated by endonuclease-independent LINE-1 retrotransposition., Nat Genet 31 (2) (2002) 159–65.

[12] N. Okada, M. Hamada, I. Ogiwara, K. Ohshima, SINEs and LINEs share common 3' sequences: a review., Gene 205 (1-2) (1997) 229–43.

[13] Z. Yang, D. Boffelli, N. Boonmark, K. Schwartz, R. Lawn, Apolipoprotein(a) gene enhancer resides within a line element., J Biol Chem 273 (2) (1998) 891–897.

[14] K. Kobayashi, Y. Nakahori, M. Miyake, K. Matsumura, E. Kondo-Iida, Y. Nomura, M. Segawa, M. Yoshioka, K. Saito, M. Osawa, K. Hamano, Y. Sakakihara, I. Nonaka, Y. Nakagome, I. Kanazawa, Y. Nakamura, K. Tokunaga, T. Toda, An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy., Nature 394 (6691) (1998) 388–92.

[15] G. L. Bratthauer, T. G. Fanning, Active LINE-1 retrotransposons in human testicular cancer., Oncogene 7 (3) (1992) 507–10.

[16] J. Skowronski, M. F. Singer, Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line., Proc Natl Acad Sci U S A 82 (18) (1985) 6050–6054.

[17] G. Alves, M. T. Kawamura, P. Nascimento, C. Maciel, J. A. Oliveira, A. Teixeira, G. C. Mda, DNA release by line-1 (L1) retrotransposon. could it be possible?, Ann N Y Acad Sci 906 (2000) 129–33.

[18] A. R. Florl, R. Lower, B. J. Schmitz-Drager, W. A. Schulz, DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas., Br J Cancer 80 (9) (1999) 1312–21.

[19] M. Ehrlich, DNA methylation in cancer: too much, but also too little., Oncogene 21 (35) (2002) 5400–13.

[20] E. M. Ostertag, J. r. Kazazian HH, Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition., Genome Res 11 (12) (2001) 2059–65.

[21] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. K. H. Lehvaslaiho,

P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, M. Clamp, Ensembl, Nucleic Acids Res 30 (1) (2002) 38–41.
URL http://www.ensembl.org/

[22] W. Wei, N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, J. D. Boeke, J. V. Moran, Human L1 retrotransposition: cis preference versus trans complementation., Mol Cell Biol 21 (4) (2001) 1429–39.

[23] D. M. Sassaman, B. A. Dombroski, J. V. Moran, M. L. Kimberland, T. P. Naas, R. J. DeBerardinis, A. Gabriel, G. D. Swergold, J. r. Kazazian HH, Many human L1 elements are capable of retrotransposition., Nat Genet 16 (1) (1997) 37–43.

[24] U. Schwahn, S. Lenzner, J. Dong, S. Feil, B. Hinzmann, G. van Duijnhoven, R. Kirschner, M. Hemberger, A. A. Bergen, T. Rosenberg, A. J. Pinckers, R. Fundele, A. Rosenthal, F. P. Cremers, H. H. Ropers, W. Berger, Positional cloning of the gene for X-linked retinitis pigmentosa 2., Nat Genet 19 (4) (1998) 327–32.

[25] M. L. Kimberland, V. Divoky, J. Prchal, U. Schwahn, W. Berger, J. r. Kazazian HH, Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells., Hum Mol Genet 8 (8) (1999) 1557–60.

[26] B. A. Dombroski, S. L. Mathias, E. Nanthakumar, A. F. Scott, J. r. Kazazian HH, Isolation of an active human transposable element., Science 254 (5039) (1991) 1805–1808.

[27] J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice., Nucleic Acids Res 22 (22) (1994) 4673–80.
URL http://www.ebi.ac.uk/clustalw/

[28] G. D. Swergold, Identification, characterization, and cell specificity of a human LINE-1 promoter., Mol Cell Biol 10 (12) (1990) 6718–29.

[29] S. Boissinot, P. Chevret, A. V. Furano, L1 (LINE-1) retrotransposon evolution and amplification in recent human history., Mol Biol Evol 17 (6) (2000) 915–28.

[30] P. Medstrand, L. N. van de Lagemaat, D. L. Mager, Retroelement distributions in the human genome: variations associated with age and proximity to genes., Genome Res 12 (10) (2002) 1483–95.

[31] S. Boissinot, A. Entezam, A. Furano, Selection against deleterious LINE-1-containing loci in the human lineage., Mol Biol Evol 18 (6)

(2001) 926–35.

[32] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, D. L. Wheeler, Genbank, Nucleic Acids Res 30 (1) (2002) 17–20.
URL http://www.ncbi.nlm.nih.gov/

[33] R. Staden, K. F. Beal, J. K. Bonfield, The Staden package, 1998., Methods Mol Biol 132 (2000) 115–30.
URL http://www.mrc-lmb.cam.ac.uk/pubseq/