Nucleic Acids Research, 2003, Vol. 31, No. 15 ***4385–4390***
DOI: 10.1093/nar/gkg496

The human genome contains many types of chimeric retrogenes generated through *in vivo* RNA recombination

Anton Buzdin*, Elena Gogvadze, Elena Kovalskaya, Pavel Volchkov, Svetlana Ustyugova, Anna Illarionova, Alexey Fushan, Tatiana Vinogradova and Eugene Sverdlov

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow 117871, Russia

Received April 29, 2003; Revised and Accepted June 4, 2003

DDBJ/EMBL/GenBank accession nos CB333829, CB333830

ABSTRACT

L1 retrotransposons play an important role in mammalian genome shaping. In particular, they can transduce their 3′-flanking regions to new genomic loci or produce pseudogenes or retrotranscripts through reverse transcription of different kinds of cellular RNAs. Recently, we found in the human genome an unusual family of chimeric retrotranscripts composed of full-sized copies of U6 small nuclear RNAs fused at their 3′ termini with 5′-truncated, 3′-poly(A)-tailed L1s. The chimeras were flanked by 11–21 bp long direct repeats, and contained near their 5′ ends T_2A_4 hexanucleotide motifs, preferably recognized by L1 nicking endonuclease. These features suggest that the chimeras were formed using the L1 integration machinery. Here we report the identification of 81 chimeras consisting of fused DNA copies of different RNAs, including mRNAs of known human genes. Based on their structural features, the chimeras were subdivided into nine distinct families. 5′ Parts of the chimeras usually originated from different nuclear RNAs, whereas their 3′ parts represented cytoplasmic RNAs: mRNAs, including L1 mRNA and Alu RNA. Some of these chimeric retrotranscripts are expressed in a variety of human tissues. These findings suggest that RNA–RNA recombination during L1 reverse transcription followed by the integration of the recombinants into the host genome is a general event in genome evolution.

INTRODUCTION

L1 retrotransposons, which comprise ~17% of the human genomic DNA (1), are mostly inactive, transpositionally deficient 5′-truncated copies (2–5) with only a low number (30–60) of full-sized elements actively transposing in the human genome (2). L1 transposition is known to proceed in several steps including Pol II transcription of an active element, reverse transcription of the RNA formed with L1-encoded RNA-dependent DNA polymerase (reverse transcriptase), and integration of the cDNA into a new position within the genome (4). An L1 element contains two open reading frames: ORF1 that encodes a 40 kDa RNA binding protein, p40, co-localized with L1 RNA in cytoplasmic ribonucleoprotein particles (RNPs) which mediate retrotransposition (6,7), and ORF2 encoding the reverse transcriptase and endonuclease (2). Due to the so-called '*cis*-preference', the enzymatic machinery of a retrotransposition-competent L1 predominantly transposes its own copies (8). However, L1s are capable of transposing other sequences, mostly Alu retroposons, but also cDNAs of different types of cellular RNAs, thus forming pseudogenes (9). Recently, we found in the human genome a family of chimeric pseudogenes or retrotranscripts formed by full-sized copies of U6 small nuclear RNAs (snRNAs) fused at their 3′ ends with 5′-truncated L1 copies (10). In the human genome, this family was represented by 56 sequences poly(A) tailed at their 3′ termini and flanked by short direct repeats. All of them harbored either a T_2A_4 hexanucleotide preferably recognized by L1 nicking endonuclease, or its derivatives with single nucleotide substitutions at their 5′ ends. These features suggest that the chimeras' integration was mediated by the integration machinery of L1 retrotransposons (11). In addition, the U6-L1 chimeras structural peculiarities suggested them to be integrated in the genomic DNA as preformed units. We proposed a mechanism for the chimeras formation, which includes a template switch from L1 mRNA to U6 snRNA during L1 reverse transcription, followed by the integration of the chimeric cDNAs in the human DNA (10). Here we present evidence that the formation of similar chimeric retrogenes using the L1 retroposition machinery occurred also with the involvement of other classes of cellular transcripts and 'selfish' genomic elements. Locus-specific PCR analysis shows that various chimeras were being formed for at least 47 million years, and their formation still continues to the present day. We demonstrate also the transcriptional activity of some such retroelements.

MATERIALS AND METHODS

DNA sequence analysis

Pseudogene consensus sequences were taken from the RepBase Update database (http://www.girinst.org/server/

*To whom correspondence should be addressed. Tel: +7 095 3306329; Fax: +7 095 3306538; Email: anton@humgen.siobc.ras.ru

Table 1. The most frequent human pseudogenes and chimeric retrotranscripts detected by public human genome databases screening

Pseudogene[a]	Number[b]	Chimeras[c]
5S rRNA	40	3 (7.5%)
4,5S rRNA	7	–
tRNA Asn	24	–
L7 r.p.[d]	45	–
L7A r.p.	21	–
L23A r.p.	33	–
L10 r.p.	34	–
L31 r.p.	40	2 (5%)
L28 r.p.	7	–
7SK	33	–
7SL (SRP)	43	1 (2.3%)
E1 snlRNA	5	–
E2 snlRNA	5	–
E3 snlRNA	4	–
hY1	40	–
CYCLO	33	–
XBR (a-fet.)[e]	2	–
XTR (a-fet.)	2	–
U1 snRNA	40	–
U2 snRNA	40	–
U3 snRNA	40	8 (20%)
U4 snRNA	20	–
U5 snRNA	21	1 (4.8%)
U6 snRNA	161	66 (41.0%)
All types	740	81 (10.9%)

[a]Name of the consensus sequence used to search the databases.
[b]The number of full-sized pseudogenes found in the human genome databases, moderately (<10%) diverged from the corresponding consensus sequence.
[c]The number and proportion of chimeric retrotranscripts identified among the representatives of a given class of pseudogenes.
[d]r.p., ribosomal protein.
[e]a-fet., a-fetoprotein.

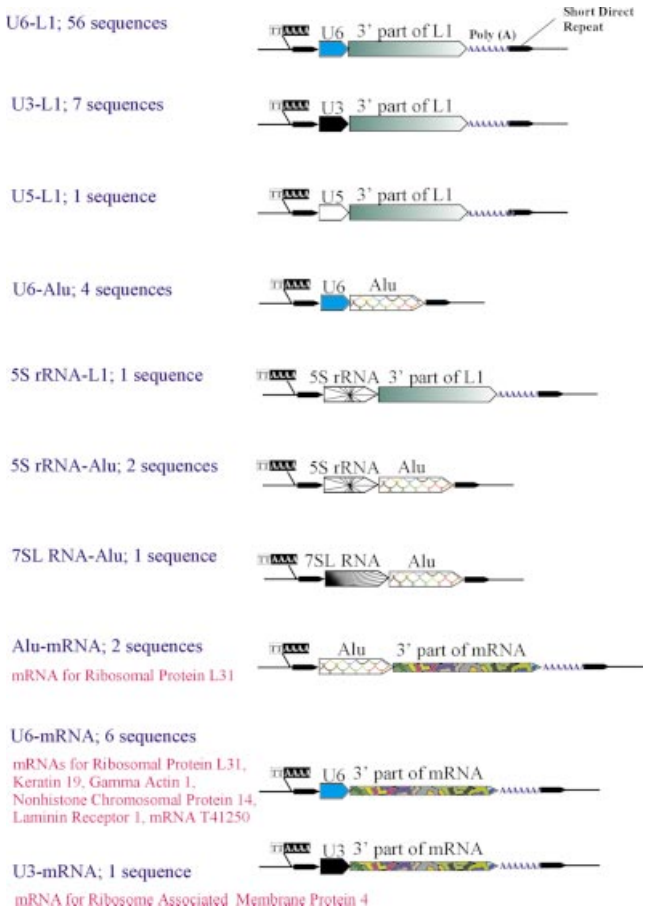

Figure 1. Schematic representation of the chimeric retrogenes identified in public databases. The chimeras' genomic locations and GenBank accession numbers are listed in Supplementary Material, Table S1.

RepBase/). We used BLAT search (http://genome.ucsc.edu/cgi-bin/hgBLAT) to find the full-sized 5S rRNA, 4.5S rRNA, tRNA, L7, L7A, L23A, L10, L31, L28, 7SK, 7SL, E1, E2, E3, hY1, CYCLO, XBR, XTR, U1, U2, U3, U4, U5 and U6 genes and pseudogenes in the human genome databases, and to determine their genomic locations. Flanking regions of pseudogenes were investigated with the RepeatMasker program (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker; A. F. A. Smit and P. Green, unpublished data). Homology searches against GenBank were done using the BLAST Webserver at NCBI (http://www.ncbi.nlm.nih.gov/BLAST) (12). L1 and Alu sequences adjacent to pseudogenes were assigned to subfamilies according to the RepBase nomenclature. For multiple alignments the Clustal W program (13) was used.

Oligonucleotide primers

Oligonucleotide primers for PCR amplification were synthesized using an ASM-102U DNA synthesizer (Biosan, Novosibirsk, Russia). Their structures can be found in Supplementary Material, Table S2.

Chimeric retrotranscripts insertion analysis

The insertion polymorphism of 12 selected chimeric retrotranscripts in the primate genomes was studied by PCR analysis. Forty nanograms each of two chimpanzee (*Pan paniscus* and *Pan troglodytes*), two gorilla (*Gorilla gorilla*), two orangutan (*Pongo abelii*), two gibbon (*Hylobates lar* and *Hylobates syndectilus*), four Old World monkey (*Macaca nemestrina*, *Macaca arctoides*, *Mandrillus sphinx*, *Colobus g. kikuy*) and six New World monkey (*Callitrix pigmaea* and *Saimiri sciureus*) blood DNA samples as well as 10 human placenta DNA samples, taken as templates, were PCR amplified with unique G1 and G2 primers. G1 and G2 are genomic primers flanking the chimeric insertions (for structures see Supplementary Material, Table S2). The PCR was conducted at 95°C for 15 min, 56°C for 10 min, 72°C for 1 h 30 min, for 30 cycles. The PCR products were separated in 1.2% agarose gels, transferred to Hybond N filters (Gibco BRL), and hybridized with radiolabeled probes specific to the sequences of the 5′ and 3′ parts of the chimeras, and to the corresponding genomic pre-integration site sequences. U6-, U3-, 5S-, Alu-, L1- and mRNA-specific probes were obtained by genomic PCRs using 0.2 µM of primers shown in Supplementary Material, Table S2. Human placenta genomic DNA was used as a template and PCR conditions were as described above. The probes for pre-integration sequences were the products of PCRs with G1 and G2 primers and primate genomic DNA templates corresponding to the absence of chimera insertions from genomic loci under study. These probes were randomly labeled with ^{32}P using a Prime-a-Gene labeling system (Promega), and hybridized at 65°C according to the membrane manufacturer's recommendations.

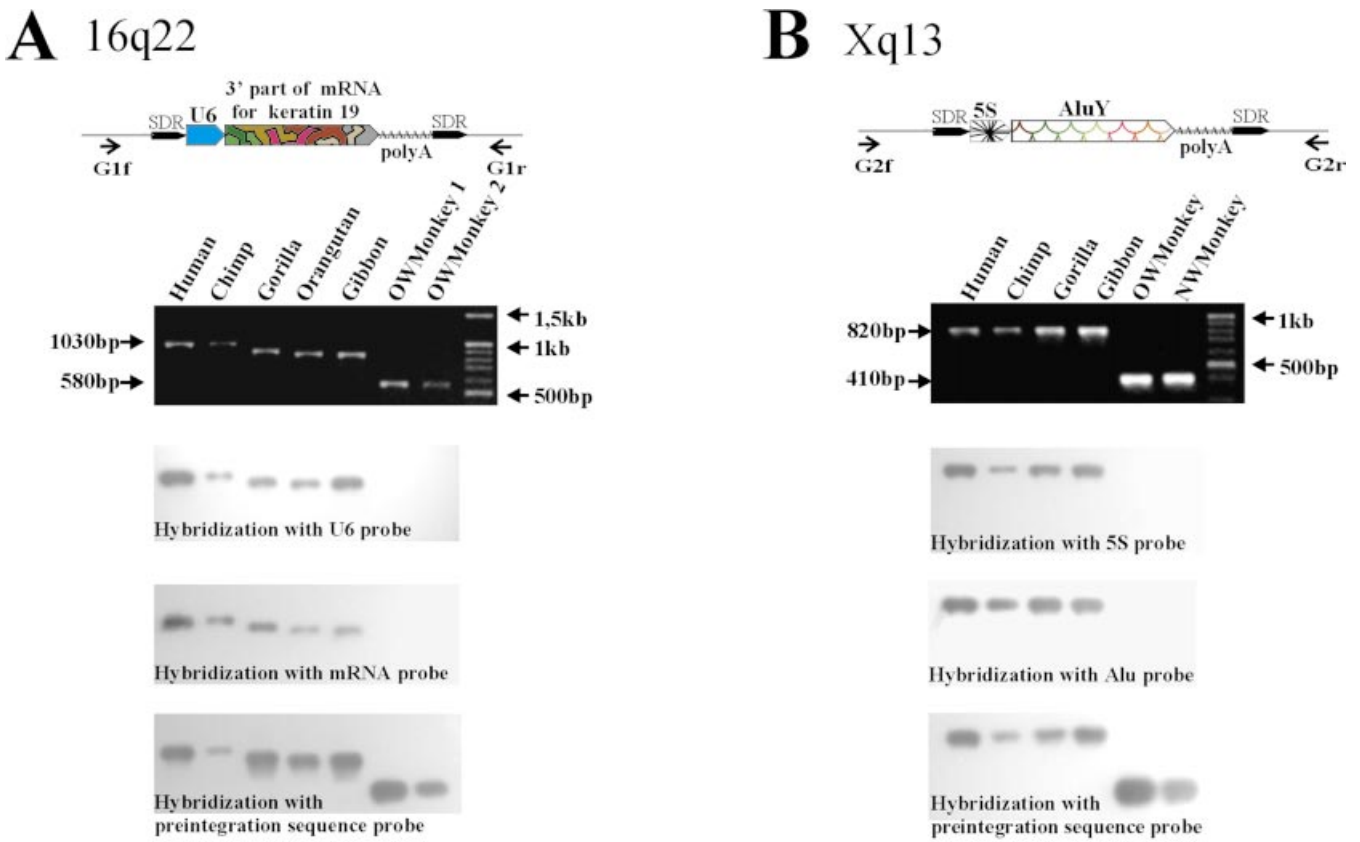

Figure 2. Two examples of chimeric insertion locus-specific PCR analysis. The PCR products obtained with primate genomic DNA templates and unique primers flanking the chimeras' insertions were transferred to membranes and separately hybridized with probes specific to the 5′ and 3′ parts of the chimeras and with probes specific to pre-integration sequences. (**A**) Analysis of U6-mRNA for Keratin 19 chimera insertion from the human 16q22 genomic locus (GenBank accession no. AC009131). (**B**) Analysis of 5S-AluY chimera insertion from the Xq13 locus (GenBank accession no. AL158069).

Tissue sampling

Human embryo brain samples were obtained from the Cancer Research Center (Moscow). Seminoma and testicular parenchyma were sampled from orhidectomy specimens with testicular germ cell tumors under non-neoplastic conditions. Representative samples were divided into two parts: one was immediately frozen in liquid nitrogen and the other was formalin-fixed and paraffin-embedded for histological analysis.

RNA isolation

Total RNA was isolated from frozen tissues pulverized in liquid nitrogen and from cell lines using an RNeasy Mini RNA purification kit (Qiagen). All RNA samples were further treated with DNase I to remove residual DNA. cDNA synthesis was performed according to a standard protocol using mixed oligo(dT) and random hexamer primers with or without the addition of AMV reverse transcriptase. The efficiency of cDNA synthesis was equal in all preparations, as verified using RT–PCR with primers specific for the beta-actin gene (Gene Checker Kit, Invitrogen).

RT–PCR

For RT–PCR we designed pairs of primers specific to the 3′ and 5′ parts of the chimeras (for sequences see Supplementary Material, Table S2). Prior to the RT–PCR analysis, the priming efficiency of the primers was examined by genomic PCRs (performed at 95°C for 20 min, 58°C for 20 min and at 72°C for 40 min, for 25 cycles) with 40 ng each of the human genomic DNA templates isolated from all of the tissues used for RT–PCR. Then the RT–PCR was performed with cDNA samples of human embryo thalamus and hippocampus, mature seminoma and normal testicular parenchyma, an equivalent of 20 ng of total RNA being used as a template in each PCR. PCR was performed in a final volume of 40 µl using the primers against the 3′ and 5′ parts of the chimeras (see above). Aliquots (6 µl) of the reaction mixture after 24, 27, 30, 33, 36 and 39 cycles of the amplification were analyzed by electrophoresis in 1.5% agarose gels. All RT–PCR experiments were reproduced at least twice using independent cDNA preparations.

RESULTS

Using BLAT and UCSC Human Genome Browser software, we have managed to identify in the human genome databases 740 full-sized pseudogenes 0–10% diverged from their consensus sequences and belonging to 24 most abundant types (1) (Table 1). In addition to U6-L1 chimeras described previously, a detailed structural analysis of these pseudogenes revealed 25 more chimeric retrotranscripts. Being organized like the former, the latter chimeras were assembled from other

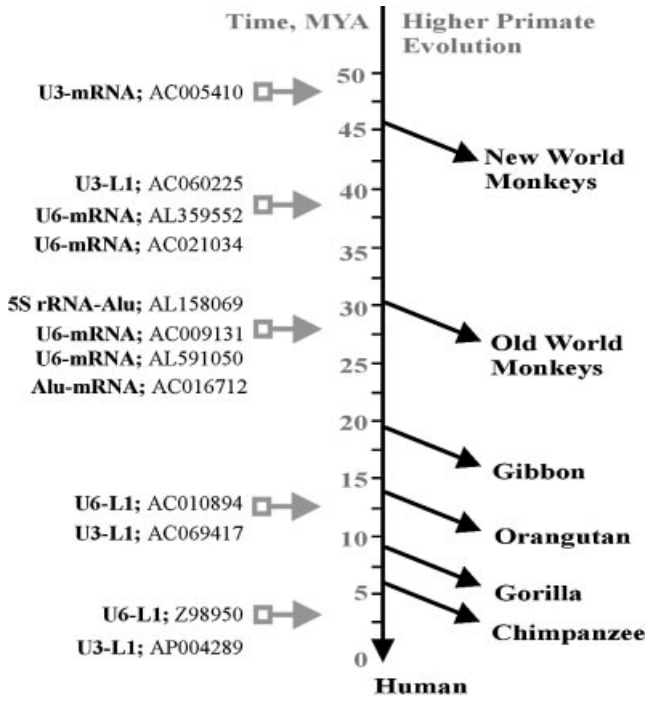

Figure 3. Results of the 12 chimeric retrogenes insertional polymorphism study. The chimeras' integration times were estimated according to the presence/absence of the inserts in genomic DNAs of different primate species.

components (see Fig. 1; GenBank accession numbers for all the chimeras are given in Supplementary Material, Table S1). Similar to the U6-L1 fusions, all these chimeras were flanked by direct repeats of 8–25 nt, and for the most part contained poly(A) tails at their 3′ ends. All of them harbored at their 5′ ends either a T_2A_4 hexanucleotide preferably recognized by L1 nicking endonuclease, or its derivatives with one or two single nucleotide substitutions. These features suggest the involvement of the L1 retroposition machinery in the formation of the chimeras. The types of chimeric retrotranscripts revealed are shown in Figure 1. Most often the 3′ ends of the chimeras were 5′-truncated L1 retroelements, but in 20% of cases the 3′ ends were either pseudogenes of various mRNAs (11%) or Alu retroposons (9%). The 5′ ends of the chimeras were pseudogenes of snRNAs—U6 (82%), U3 (10%) and U5 (1%), or sequences of Alu, 7SL RNA and 5S rRNA (3, 1 and 4%, respectively). All the 5′ parts were full-length pseudogenes, except U3 and U5 snRNAs, which always presented as 3′-truncated copies corresponding to positions 1–72 and 1–75 of the U3 and U5 consensus sequences, lacking 145 and 41 3′ end nucleotides, respectively. Interestingly enough, the 5′ components of the chimeras are mostly localized in nuclei, whereas the 3′ components are the copies of cytoplasmic RNAs.

The structural peculiarities of the chimeras suggest that they did not have any common ancestor, but were rather formed due to multiple independent events. Similar to the U6-L1 fusions, the divergences of the 3′ and 5′ components of the chimeras from their consensus sequences were linearly correlated (see Supplementary Material, Table S1). Together with the features of the chimeras' integration sites, this correlation indicates that the 3′ and 5′ parts of the retrotranscripts were simultaneously integrated into the genomic DNA. To confirm this, we carried out a PCR analysis of 12 integration sites of various retrotranscript types. To this end, unique primers specific to the sequences flanking the integration sites were made (Fig. 2; for primer sequences see Supplementary Material, Table S2). The primers were used for PCR on panels of human and non-human primate genome DNA samples including those of chimpanzee, gorilla, orangutan, gibbon as well as Old and New World monkeys. In this way, the PCR products formed on the sites containing integrated chimeras were longer than those formed on orthologous loci lacking insertions. The PCR products were transferred to membranes and separately hybridized with labeled probes for 3′ and 5′ parts of chimeras, or with a probe for the pre-integration sequence (Fig. 2). The latter probe hybridized to all the PCR products, whereas the former probes could selectively hybridize only to 'long' PCR products. The primers against the 3′ and 5′ parts of one and the same chimera always hybridized pairwise supporting simultaneous

Table 2. Chimeric retrogenes expressed in human tissues

Name	GenBank accession no.[a]	EST/mRNA[b]	Tissue[c]	Expression level (molecules/cell)[d]
U6-L1PA7	AL121883	BQ720168	Sympathetic trunk	Unknown
		AI095257	Senescent fibroblasts	Unknown
U6-L1PA3	AC010894	BQ447264	Osteoartritic cartilage	Unknown
U6-AluY	AC004128	AA581502	Ovary bulk tumor	Unknown
U6-mRNA for γ-actin	AL591050	AK056682	Peripheral blood mononuclear cells	Unknown
U6-mRNA T41250	AL354668	CB333830	Embryo brain: thalamus	~10
			Embryo brain: hippocampus	~10
			Normal testicular parenchyma	~10
			Seminoma	~10
U6-mRNA for non-histone chromosomal protein 14	AC021037	CB333829	Embryo brain: thalamus	~10
			Embryo brain: hippocampus	~200
			Normal testicular parenchyma	~100
			Seminoma	0

[a]GenBank chimeras' accession numbers.
[b]GenBank accession numbers of the corresponding EST/mRNA.
[c]Human tissues where the chimeras are expressed.
[d]The number of chimeric transcripts per cell estimated by RT–PCR.

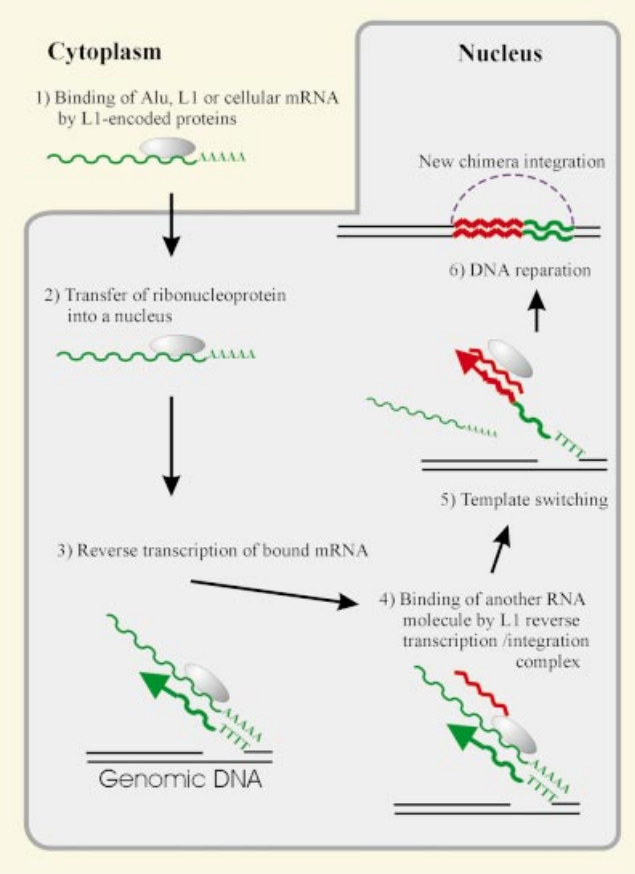

Figure 4. A probable mechanism for the chimeras' formation. (Step 1) An L1 pre-integration complex binds L1, Alu or the host mRNA in the cytoplasm. (Step 2) The ribonucleoprotein formed is transferred to the nucleus. (Step 3) Reverse transcription of the bound mRNA primed by a genomic DNA single-stranded break within the TTTTAA sequence. (Step 4) Another (nuclear) RNA binds to the L1 reverse transcription/integration complex. (Step 5) Switch of templates for the reverse transcription. (Step 6) The DNA reparation mediated formation of a new chimeric retrogene insertion flanked by short direct repeats and carrying a poly(A) sequence at the 3′ terminus.

integration of both parts into primate DNA for all the chimeras studied.

Moreover, the PCR analysis showed that the formation of chimeric retrotranscripts in the primate genomes had occurred for at least 47 million years, starting from the divergence of the New World monkeys branch (14) and up to evolutionarily recent time. In particular, some of the chimeric insertions are specific for the human genome (Fig. 3), and at least one of them is polymorphic in the human population (10).

To find out if the chimeras are expressed in human tissues, all the 81 chimeric retrotranscripts identified so far were searched for against EST databases. For four of them the corresponding cellular mRNAs were detected in GenBank (Table 2). We have also performed an RT–PCR analysis for six chimeras containing 3′ terminal fragments of human mRNAs. As a result, two of these chimeras were shown to be expressed in human tissues. One of them, U6-cDNA T41250, was expressed at ~10 mRNA molecules per cell in all four tissues studied, whereas the other, U6-mRNA for the non-histone chromosomal protein 14, was expressed tissue specifically at 0 to ~200 mRNA molecules per cell (Table 2).

DISCUSSION

The data obtained here and by Buzdin *et al.* (10) suggest that the human genome contains a lot of chimeric products formed due to RNA–RNA hybridization of various cellular transcripts and then integrated into DNA, and that these integrations occurred for at least 47 million years. It can be supposed that, similar to U6-L1 retrotranscripts, all the chimeras revealed were formed due to a template switch during the reverse transcription of L1 RNA (10) (Fig. 4), as also observed for an R2 non-LTR retrotransposon (15) and in retroviral genome recombination (16,17). The frequency of such events is rather low. As judged from 60 'abundant' U6-containing chimeras having been formed for 50 million years, it is approximately as low as one recombination event per million years. However, this figure seems to be seriously underestimated, inasmuch as we used only a limited number of pseudogenes to screen

Table 3. Assignment of 5′ and 3′ terminal parts of chimeras to known RNA sequences

Name[a]	Number and proportion[b]	Function[c]	RNA localization[d]
5′ terminal parts			
U6	66 (81.5%)	snRNA, splicing	Nucleus
U3	8 (9.9%)	snRNA, splicing	Nucleus
U5	1 (1.2%)	snRNA, splicing	Nucleus
5S rRNA	3 (3.7%)	rRNA, ribosome	Cytoplasm, nucleus
7SL	1 (1.2%)	SRP, protein sorting	Cytoplasm, nucleus
Alu	2 (2.5%)	Selfish RNA	Cytoplasm, nucleus
3′ terminal parts			
L1	65 (80.3%)	mRNA, selfish	Cytoplasm, nucleus
Alu	7 (8.6%)	Selfish RNA	Cytoplasm, nucleus
mRNA	9 (11.1%)	mRNAs of seven known cellular genes	Cytoplasm

[a]Names of the cellular transcripts corresponding to a given part of the chimeras.
[b]Number and proportion of the sequences among the chimeric retrotranscripts identified so far.
[c]Function of the corresponding RNA in the cell.
[d]Cellular localization of the corresponding RNAs.

for chimeras. Moreover, certain chimeras might be strongly diverged and therefore undetectable by the screening procedure used.

Interestingly, 93% of the 5′ parts of the chimeras identified are DNA copies of snRNAs involved in spliceosomes: U6, U3 and U5 snRNAs (82, 10 and 1% of the cases, respectively; Table 3). Such a high frequency of template switches to the spliceosomal RNAs might imply a relationship between L1 retrotranspositions and the splicing machinery, as well as a close spatial location of L1 reverse transcription/integration complexes and spliceosomes.

Some of the chimeric retrotranscripts are expressed in human tissues. Therefore, the phenomenon of the chimerization revealed can be considered a previously unknown mechanism of the formation of new genes by combining parts of pre-existing expressing sequences.

Formation of certain L1 families might also involve RNA–RNA recombination due to a template switch after L1 mRNA major part is reverse transcribed, resulting in the fusion of the L1 3′ part with an entirely new nucleotide sequence. In particular, 5′-untranslated regions and the first third of ORF1 of human L1, murine L1 families and of the L1s of rat and rabbit are known to be not homologous to each other (4). However, these speculations need more detailed analysis.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Professor Victor Potapov and Dr Nadezhda Skaptsova (Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia) for synthesis of oligonucleotides, and Dr Boris Glotov (Institute of Molecular Genetics, Moscow, Russia) for invaluable comments on the manuscript. The work was supported by INTAS-01-0759 and Russian Foundation for Basic Research 02-04-48614a and 00-15-97945 grants.

REFERENCES

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Kazazian,H.H,Jr (2000) Genetics. L1 retrotransposons shape the mammalian genome. *Science*, **289**, 1152–1153.
3. Ovchinnikov,I., Troxel,A.B. and Swergold,G.D. (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res.*, **11**, 2050–2058.
4. Furano,A.V. (2000) The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 255–294.
5. Buzdin,A., Ustyugova,S., Gogvadze,E., Lebedev,Y., Hunsmann,G. and Sverdlov,E. (2003) Genome-wide targeted search for human specific and polymorphic L1 integrations. *Hum. Genet.*, **112**, 527–533.
6. Kolosha,V.O. and Martin,S.L. (2003) High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J. Biol. Chem.*, **278**, 8112–8117.
7. Hohjoh,H. and Singer,M.F. (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.*, **15**, 630–639.
8. 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.
9. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.*, **24**, 363–367.
10. Buzdin,A., Ustyugova,S., Gogvadze,E., Vinogradova,T., Lebedev,Y. and Sverdlov,E. (2002) A new family of chimeric retrotranscripts formed by a full copy of u6 small nuclear RNA fused to the 3′ terminus of l1. *Genomics*, **80**, 402–406.
11. Jurka,J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA*, **94**, 1872–1877.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) 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**, 4673–4680.
14. Sverdlov,E.D. (2000) Retroviruses and primate evolution. *Bioessays*, **22**, 161–171.
15. Bibillo,A. and Eickbush,T.H. (2002) The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.*, **316**, 459–473.
16. Negroni,M. and Buc,H. (2001) Mechanisms of retroviral recombination. *Annu. Rev. Genet.*, **35**, 275–302.
17. 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.