mRNA Retrotransposition Coupled with 5' Inversion as a Possible Source of New Genes

Kenji K. Kojima and Norihiro Okada

Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Japan

Human long interspersed nuclear element-1 (L1) occupies one-sixth of our genome and has contributed to genome evolution in various ways. Approximately 10% of human L1 copies are composed of two L1 segments; the 5' segment and 3' segment are in head-to-head (i.e., 5'-inverted) orientation. Besides mediating their own retrotransposition, L1 has the ability to mobilize mRNA "in trans," and the number of retrotransposed mRNA sequences (retrocopies) is estimated to be >6,000. In this study, we identified 48 human-specific retrocopies and 95 chimpanzee-specific retrocopies by comparing the human and chimpanzee genomes. Among these retrocopies, 12 were 5'-inverted. The characteristics of these 5'-inverted retrocopies were similar to those of 5'-inverted L1 copies, indicating that the 5' inversion is generated by the same mechanism. With these findings, we examined the possibility that 5' inversion of the retrocopy generates a new gene that codes for a peptide with a different N terminus. We identified several potential 5'-inverted retrogenes, including those of thymopoietin beta (TMPO) and eukaryotic translation initiation factor 3 subunit 5 (EIF3F). The most interesting candidate was the 5'-inverted retrocopy of small nuclear ribonucleoprotein polypeptide N (SNRPN). This retrocopy was transcribed in the reverse orientation in several organs, had multiple transcript variants, and encoded a protein containing a peptide fragment derived from the N-terminal portion of SNRPN. Our results suggest that mRNA retrotransposition coupled with 5' inversion may be a mechanism to generate new genes distinct from parental genes.

Introduction

Human long interspersed nuclear element-1 (LINE-1; L1) is one of the best-characterized non-long terminal repeat (non-LTR) retrotransposon families. More than 500,000 L1 copies occupy one-sixth of our genome, but most L1 copies are severely 5'-truncated (Lander et al. 2001). The expansion of L1 copy number is an outstanding feature of mammalian genomes. Furthermore, L1 is responsible for the mobilization of Alu short interspersed nuclear elements (SINEs), which occupy an additional 10% of our genome (Lander et al. 2001; Dewannieux et al. 2003). SVA composite retrotransposon is also likely to be retrotransposed dependent on L1 (Ostertag et al. 2003). Present knowledge of L1 is mainly based on studies of L1s from human, mouse, and rat. Retrotransposons closely related to mammalian L1s are distributed in amphibians, fishes, sea squirts, sea urchins, and many other organisms (Kojima and Fujiwara 2004) (Repbase, http://www.girinst.org/repbase/index.html), but their characteristics remain unclear.

The retrotransposition of non-LTR retrotransposons, including L1, is initiated by single-strand DNA cleavage catalyzed by the endonuclease encoded by each non-LTR retrotransposon (Luan et al. 1993; Cost et al. 2002) (fig. 1A). The cleaved 3'-end is used as a primer to initiate reverse transcription catalyzed by the reverse transcriptase (RT) encoded by non-LTR retrotransposons, a process called target-primed reverse transcription (fig. 1D), followed by second-strand cDNA synthesis (fig. 1E–G). Integration is finished by ligating the nicks at both retrotransposon ends (fig. 1H). The L1 endonuclease generally cleaves the second DNA strand downstream from the cleavage site of the first DNA strand (fig. 1D). This distance (fig. 1D, yellow boxes) causes target site duplications (TSDs), which are short direct repeats on both sides of L1.

There are several features of the L1 retrotransposition mechanism that have not been observed in other non-LTR retrotransposons. One of the features specific to L1 is 5' inversion. Approximately 10% of human L1 copies are composed of two L1 segments; the 5' segment and 3' segment are in head-to-head (i.e., 5'-inverted) orientation (fig. 1O) (Kazazian and Moran 1998; Szak et al. 2002). The inversion point varies between different L1 copies. L1 5' inversion has been observed in only three mammalian species: human, mouse, and dog (Choi et al. 1999; Martin et al. 2005). It remains to be elucidated whether L1 in other organisms and non-LTR retrotransposons other than L1 can retrotranspose with 5' inversion. Ostertag and Kazazian (2001) proposed the "twin-priming" mechanism for the 5' inversion of L1 copies (fig. 1). In standard retrotransposition, one 3' end of double-stranded DNA is used as a primer to initiate reverse transcription (fig. 1C), whereas the other 3' end is used as a primer for second-strand cDNA synthesis following reverse transcription (fig. 1G). In twin priming, however, both 3' ends are used as primers for reverse transcription (fig. 1I). One end is annealed to the 3' terminus of L1 RNA as in standard retrotransposition, whereas the other end is annealed to the same L1 RNA internally. As a result of annealing, microhomologies—nucleotides homologous to both L1 cDNA and the target DNA—are often observed at the 5' and 3' ends of 5'-inverted L1 copies (fig. 1M and P, the left and right regions marked "MH"). Twin priming likely needs two RT molecules to generate the 5'-inverted structure. After reverse transcription (fig. 1J–L), L1 RNA is degraded and the two cDNA strands are annealed at a region of short complementarity (fig. 1M, indicated by the middle region marked MH). This annealing is presumed to occur because of the presence of a microhomology at the inversion junction (fig. 1P, the middle region marked MH). Finally, the second-strand cDNA is synthesized to fill in the single-stranded regions at both sides of the annealed sequences (fig. 1N). Microhomologies at three junctions are observed in many but not all inverted L1 copies, indicating that annealing is not essential for twin priming.

Key words: retrotransposition, retrocopy, retrogene, inversion, L1, twin priming.

E-mail: nokada@bio.titech.ac.jp.

doi:10.1093/molbev/msp050
Advance Access publication March 16, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org
There is apparently no strict sequence requirement in the 3' end for recognition by RT during L1 retrotransposition, except for the polyA tail (Moran et al. 1999). In contrast to L1, the RTs of other non-LTR retrotransposons, such as unaL2 and SART1, recognize specific sequences of their own RNAs in the 3' untranslated regions (UTRs) (Kajikawa and Okada 2002; Osanai et al. 2004). PolyA recognition by L1 RT probably causes two phenomena characteristic for L1: 3' transduction and mRNA retrotransposition. Human L1 retrotransposons can produce DNA transduction events in which unique DNA segments downstream from L1 are mobilized as part of retrotransposition (Moran et al. 1999; Goodier et al. 2000; Pickeral et al. 2000). 3' Transduction is considered to occur when transcription stops at a polyadenylation signal downstream of the L1 3' end and reverse transcription is initiated from the downstream polyA tail. L1 proteins also mobilize mRNAs probably by recognizing polyA tails "in trans," although the frequency is much lower than that of L1 RNA mobilization (Esnault et al. 2000; Wei et al. 2001). Retrotransposed mRNAs are called "retrocopies," "processed pseudogenes," or "intronless pseudogenes," because they usually have neither introns
nor promoters. The abundance of retrocopies is remarkable in mammalian genomes, although genomes of organisms other than mammals also include some retrocopies (Yu et al. 2007). Approximately 4,000 retrocopies have been characterized in the human genome, and the total number of retrocopies is estimated at over 6,000 (Harrison et al. 2002; Ohshima et al. 2003; Marques et al. 2005).

Although retrocopies were originally considered to be nonfunctional, recent analyses revealed many functional retrocopies in mammalian genomes. These functional retrocopies are called “retrogenes.” Retrogene sometimes serve as tissue-specific isotypes of their parental genes. For example, the mouse retrogene Utp14b is expressed at high levels in pachytene spermatocytes in which the parental intron-containing gene Utp14a is suppressed (Bradley et al. 2004; Rohozinski and Bishop 2004). Loss-of-function mutation of Utp14b causes juvenile spermatogonial depletion (jd). The human glutamate dehydrogenase 2 gene (GLUD2) is the retrogene of the glutamate dehydrogenase 1 gene (GLUD1) (Burki and Kaessmann 2004). GLUD2 is highly adapted to functioning in the brain, compared with the housekeeping gene GLUD1, which is expressed in many tissues. Retrogenes can also be generated by the formation of a chimeric gene. TRIMCyp was generated by retrotransposition of the cyclophilin A (CypA) gene into the eighth intron of the TRIM5 β gene (Sayah et al. 2004). The chimeric protein encoded by TRIMCyp blocks human immunodeficiency virus type 1 (HIV-1) infection in owl monkeys. mRNA retrotransposition is an important source of new gene formation.

Here, we propose a novel mechanism to generate new genes through the combination of 5’ inversion and mRNA retrotransposition. We provide evidence that mRNA retrotransposition can also be coupled with 5’ inversion, and we identified several potential 5’-inverted retrogenes by searching 5’-inverted retrocopies in mammalian annotated RNA sequences and human expressed sequence tag (EST) sequences. Our results suggest that mRNA retrotransposition coupled with 5’ inversion may be a mechanism to generate new genes distinct from parental genes.

Materials and Methods

Genome Sequences

Genome sequences of human Homo sapiens, common chimpanzee Pan troglodytes, Sumatran orangutan Pongo abelii, and rhesus macaque Macaca mulatta, and the pairwise alignment between human and chimpanzee were downloaded from the University of California–Santa Cruz Genome Browser web site (http://hgdownload.cse.ucsc.edu/downloads.html). Annotated RNA sequences of human, chimpanzee, rhesus macaque, mouse Mus musculus, rat Rattus norvegicus, dog Canis familiaris, cattle Bos taurus, horse Equus caballus, opossum Monodelphis domestica, and platypus Ornithorhynchus anatinus, and the EST database were downloaded from the National Center for Biotechnology Information (NCBI) ftp site (ftp://ftp.ncbi.nih.gov/genes/, ftp://ftp.ncbi.nih.gov/blast/db/). Genome sequences of northern white-cheeked gibbon Nomascus leucogenys, and common marmoset Callithrix jacchus were downloaded from the Washington University Genome Sequencing Center web site (http://genome.wustl.edu/genome_group.cgi?GROUP=1). Genome sequences of Gamet’s greater bushbaby Otomolor garnetti, grey mouse lemur Microcebus murinus, and northern tree shrew Tupaiia belangeri were downloaded from the Broad Institute mammalian genome project web site (http://www.broad.mit.edu/mammals/). Sequence traces of genomes of gorilla Gorilla gorilla, baboon Papio hamadryas, and Philippine tarier Tarsius syrichta were acquired by searching Trace Archive databases at the NCBI web site (http://www.ncbi.nlm.nih.gov/blast/blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&BLAST_SPEC= TraceArchive&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch).

Characterization of Species-Specific Retrocopy Insertions

We extracted all human-specific insertions and deletions (indels) and chimpanzee-specific indels on the basis of the UCSC pairwise alignment, and selected indels satisfying the following criteria: 1) both junctions could be determined precisely without unaligned nucleotides, 2) the length was between 50 and 10,000 bp, 3) there were no unsequenced nucleotides, 4) ≥5-bp direct repeats were present at both junctions, 5) both 30-bp flanking sequences included no unsequenced nucleotides, 6) the flanking sequences were not repetitive sequences, and 7) the sequences and flanking 30-bp sequences were not completely identical to those of other indels. To distinguish species-specific insertions from species-specific deletions, we used the rhesus macaque and orangutan genomes as references. We joined both 20-bp flanking sequences of each indel to represent its putative uninserted sequence. If either the rhesus macaque genome or the orangutan genome contained one sequence that showed >90% nucleotide identity with the putative uninserted sequence but did not contain additional sequences with >70% identity, we considered the indel to be a species-specific insertion. Finally, we removed short duplications by comparing the insertion sequences and flanking sequences. Using these criteria, we identified 6,197 human-specific insertions and 2,910 chimpanzee-specific insertions. The longest direct repeats at both junctions of insertions were interpreted as TSDs. In this analysis, we wrote several Ruby scripts and used them in combination with the NCBI Blast 2.2.11 package (ftp://ftp.ncbi.nih.gov/blast/) and Repeat-Masker (http://www.reapemasker.org/).

To characterize species-specific retrocopies, we performed BlastN to the RefSeq RNA database with all insertions as queries after excluding repetitive sequences with the aid of RepeatMasker. All identified insertions were collected and manually investigated as to whether they had parental genes at different loci. We removed insertions for which the parental gene was not annotated in human or chimpanzee. When transcript variants were annotated for the parental gene, the variant that contained the longest sequence similar to the retrocopy was selected as parental mRNA.

Sequence logos of target sequences of retrocopies were generated with the aid of WebLogo (http://weblogo.berkeley.edu/).
Search for 5′-Inverted Retrocopies in the Human Genome

We compared all RefSeq RNA sequences from a certain species with one another using BlastN after using RepeatMasker. The BlastN results were parsed with the aid of Ruby scripts to identify RNAs that had sequences corresponding to both the sense and antisense sequences of another RefSeq RNA. To exclude inversion derived by recombination, we selected only 5′-inverted retrocopies for which the parental genes contained introns. Retrocopies without either TSDs or A-rich tails were also manually removed.

To identify 5′-inverted retrocopies in the human genome, we split human genomic sequences into 10,000-bp fragments and performed BlastN with each fragment as a query to the human RefSeq RNA database. The BlastN results were parsed with the aid of Ruby scripts to identify fragments that had sequences corresponding to both the sense and antisense sequence of a certain RefSeq RNA. All these fragments were manually checked to see whether they included inverted retrocopies. TSDs were manually characterized.

To characterize transcribed 5′-inverted retrocopies, we independently performed BlastN with the inverted segment and noninverted segment of retrocopies as queries to the human EST database. Retrocopies for which the inverted and noninverted segments hit the same EST in the same orientation were selected. We compared EST sequences and retrocopies at Blast 2 sequences in the NCBI web site (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi).

Results

Human-Specific and Chimpanzee-Specific Retrocopy Insertions

To obtain direct evidence for mRNA retrotransposition coupled with 5′ inversion, we first identified “bona fide” insertions of retrocopies by comparing the human H. sapiens and common chimpanzee P. troglodytes genomes. To distinguish species-specific insertions from species-specific deletions, we used the genomes of Sumatra orangutan P. abelii and rhesus macaque M. mulatta as outgroups indicating uninserted sequences. Consequently, we identified 48 human-specific retrocopy insertions and 95 chimpanzee-specific retrocopy insertions (table 1; supplementary material online).

Both human-specific retrocopies and chimpanzee-specific retrocopies showed characteristics of L1-dependent insertions that have been reported previously (Vanin 1985; Jurka 1997). The target sequence preference of retrocopy insertion was similar to those of L1 and Alu (fig. 2A and B); retrocopy, L1, and Alu were all preferentially inserted between TT and AAAA. The peak TSD length of 14–15 bp was also similar to those of L1 and Alu (data not shown). These data indicated that the retrocopies identified in this analysis were inserted at sites cleaved by the L1 endonuclease.

5′-Inverted Retrocopies

Interestingly, among 143 retrocopies, we found 6 human-specific 5′-inverted retrocopies and 6 chimpanzee-specific 5′-inverted retrocopies (table 1; fig. 3; supplementary fig. S1, Supplementary Material online). The structures of these retrocopies were similar to those of 5′-inverted copies of L1, as described below (Ostertag and Kazazian 2001). They did not have introns, and they were positioned on chromosomes different from those carrying their parental intron-containing genes, supporting their formation by retrotransposition.

We found two examples of 5′-inverted retrocopies of zinc finger protein 532. Using NCBI Blast, we also found 10 noninverted retrocopies of zinc finger protein 532 in the human genome (data not shown). This copy number is as high as those of ribosomal protein genes reported by Ohshima et al. (2003). Therefore, it is likely that the zinc finger protein 532 gene is highly transcribed in germ cells and the relative abundance of zinc finger protein 532 mRNA in the germline could cause multiple insertions of retrocopies. Another possible reason that two inverted retrocopies of this gene were found is its long mRNA length at ~6,500 nt. Most RNAs that have many retrocopies, including ribosomal protein genes, are shorter than 1,000 nt. Reverse transcription of short RNA is likely to finish before the top strand cleavage, and therefore, 5′ inversion seldom occurs in the retrotransposition of short mRNA. Longer mRNA is likely to be retrotransposed with inversion more frequently if the mRNA abundance is similar.

Our analysis based on species-specific insertions allowed us to determine the insertion junction without ambiguity. The 12 5′-inverted retrocopies were flanked by short direct repeats at both ends (fig. 3, red). If inversion occurred after the completion of retrotransposition, the TSDs would become inverted and positioned at the 3′ end and at the inversion junction. Hence, short direct repeats at both ends of the retrocopies were TSDs generated by retrotransposition coupled with inversion. All retrocopies were not inserted in the orangutan genome. Thus, the retrocopies shown in figure 3A–C are human-specific insertions and that shown in figure 3D is a chimpanzee-specific insertion. The TSD length was between 8 and 16 bp and the target sequences of these 5′-inverted retrocopies were similar to TTAAAA, suggesting their L1-dependent retrotransposition (figs. 2C and 3, and supplementary fig. S1, Supplementary Material online).

The junction structures also indicated their L1 dependence. The twin-priming model proposed by Ostertag and Kazazian (2001) predicts microhomologies at the 5′ junction between the target DNA and the inverted segment of the retrocopy as the result of annealing between the template RNA and the cleaved DNA end (fig. 1I–M, left region marked MH). To measure the contribution of annealing to twin priming, Ostertag and Kazazian (2001) calculated the average number of homologous nucleotides in the 3′ terminal four nucleotides in the 5′ TSD to be 3.1 in L1. This number in 5′-inverted retrocopies was 2.75, similar to that of L1. The frequency of complementarity decreased with the distance from the 3′ end of the 5′ TSD in inverted retrocopies, similar to inverted L1 copies. As in L1 inversion, we also found microhomologies (2.58 bp average) at the inversion junction of the retrocopies (fig. 3, highlighted in black). These microhomologies likely resulted from annealing between the two cDNA strands.
All inverted L1 copies described by Ostertag and Kazazian (2001) have nucleotides that are either deleted or duplicated at the point of inversion. They divided the inversions into three categories: small (1–50 bp) duplications, small (1–50 bp) deletions, and large (>50 bp) deletions. We can also divide inverted retrocopies into these three categories. Five inverted retrocopies (6_191778, 7_73776, 12_26524, 13_46085, and 35_3445; fig. 3 and supplementary fig. S1, Supplementary Material online) showed 1–15-bp duplications. In the twin-priming model shown in figure 1, duplications can be generated by the mechanism in which nucleotides that have been reverse-transcribed by the upstream RT molecule are reverse-transcribed again by the downstream RT molecule (fig. 1L). On the other hand, deletions are observed when the downstream RT molecule does not proceed with the reverse transcription to the start point of reverse transcription by the upstream RT molecule (fig. 1K). Four inverted retrocopies (3_84131, 35_3445, 10_48155, and 50_3415; fig. 3 and supplementary fig. S1, Supplementary Material online) showed 6–23-bp deletions. The other three retrocopies showed >50-bp deletions.

All 12 5’-inverted retrocopies had recognizable A-rich tracts at their 3’ termini (fig. 3, the 3’ ends of sequences in blue). Of these retrocopies, 5 (6_191778, 7_73776, 13_46085, 11_137428, and 7196_6; fig. 3 and supplementary fig. S1, Supplementary Material online) ended with A-rich tracts just at or near the polyadenylation sites of the parental mRNAs. The parental gene for 7196_6 was annotated to have six transcript variants and three of them shared the polyadenylation site with the retrocopy. The other seven retrocopies ended with A-rich tracts that were not included in the parental mRNA sequences, suggesting internal polyadenylation (see below). Among these seven retrocopies, the parental genes for 3_84131, 4_82164, and 10_48155 were annotated to have multiple transcript variants, but none of the annotated transcripts used the same polyadenylation sites as the retrocopies of these genes. We were able to characterize canonical polyadenylation signals (AA-TAAA or ATTAAA) 16–29 bp upstream from the A-rich tracts in all inverted retrocopies (fig. 3, underlined). Their parental mRNA sequences also contained polyadenylation signals. Hence, the retrocopies that appear to be internally polyadenylated may have resulted from insertion of unidentified transcript variants containing alternative polyadenylation sites. Otherwise, the mRNA terminus may have been incorrectly predicted.

Ostertag and Kazazian (2001) reported that the points of 5’ inversion of L1 are highly clustered in a small 269-bp region. They described the possibility that the secondary structure of the L1 RNA is important for determining the sites of internal annealing. In the present study, we
characterized two 5′-inverted retrocopies of the gene for zinc finger protein 532 (fig. 3B and C). The inverted segment of 6_191778 corresponded to nucleotides 4,999–5,443 of the predicted zinc finger protein 532 mRNA, and that of 12_26524 corresponded to nucleotides 5,083–5,461. In each case, the junction of 5′ inversion was in a similar region of the zinc finger protein 532 mRNA, but the junctions did not share homologous sequences. Our data are consistent with the notion that the concentration of inversion points in a certain small region of the template RNA is the universal feature for inversion coupled with retrotransposition. However, secondary structures predicted by mfold (http://mfold.bioinfo.rpi.edu/) suggested that the 5′ junction, where the top strand end is likely to anneal the template RNA, is not always single stranded.
The TSD length, target sequence preference, microhomologies, duplications and deletions of mRNA sequences, A-rich tracts, and polyA signals, all indicate that these 12 5′-inverted retrocopies were generated by the same machinery as 5′-inverted L1 copies, namely, twin priming. Thus, these features suggest that the 5′-inverted retrocopies were not generated by recombination after retrotransposition.

5′-Inverted Retrocopy as a Possible Source of New Genes

We revealed that 5′ inversion occurs during mRNA retrotransposition. The noninverted segment of the 5′-inverted retrocopy encodes the C-terminal portion of its parental protein, whereas the inverted segment is not likely to encode a protein because of the antisense orientation. However, there is a possibility that the inverted segment contains an open reading frame (ORF) extending into the noninverted segment. If this is the case, 5′ inversion could potentially generate a new protein in which the C-terminal portion shares the function with the parental protein and the N-terminal portion is unique. This phenomenon might be an evolutionary force to generate new genes. We hereafter examine this possibility.

Chimpanzee-Specific 5′-Inverted Retrocopy of TMPO

Eleven of 12 human-specific retrogenes and chimpanzee-specific retrocopies with 5′ inversions did not encode long peptides. The remaining retrocopy (10_48155) was annotated as a part of the putative gene LOC465562 (fig. 4A), and contained the last 441 bp (fig. 4A, yellow box) of the protein-coding region of its parental gene, thymopoietin beta (TMPO). Thymopoietin beta is the same protein as lamina-associated polypeptide 2 (LAP2) beta. This retrocopy included only 2 nt differences in the inverted segment and 2 nt differences in the noninverted segment compared with the chimpanzee parental gene (supplementary table S3, Supplementary Material online). The few substitutions made it impossible to show positive or negative selection in this retrocopy using the nonsynonymous-to-synonymous substitution rate ratio with PAML (Yang 2007).

The entire sequence except the last 6 bp of the inverted segment of the retrocopy was annotated as an intron in LOC465562. Therefore, the inverted segment can encode only two residues. The first exon of LOC465562 was located far upstream from the inverted retrocopy. The noninverted segment encoded 147 residues similar to TMPO. A large part of the lamin B-binding domain and the transmembrane region at the C terminus were included in the retrocopy. Whether this retrocopy is transcribed and translated remains to be determined. If this retrocopy is functional, the inverted segment serves as a splice acceptor, but it contributes little toward generating a new peptide sequence.

Searching for 5′-Inverted Retrogenes

To identify transcribed 5′-inverted retrogenes in which the inverted segment contributes toward generating a new protein sequence, we screened retrocopies using two strategies that can detect non–species-specific retrocopies in addition to the species-specific retrocopies described above. First, we searched transcripts in which the 5′ region showed sequence similarity with the antisense of the 5′ region of other transcripts. We downloaded annotated RNA sequences of human, rhesus macaque, mouse M. musculus, rat R. norvegicus, dog C. familiaris, cattle B. taurus, horse E. caballus, opossum M. domestica, and platypus O. anatinus. These RNA sequences comprise sets of all annotated RNA transcripts, and most of them have not been validated by experiments. We made comparisons among all RNA sequences of the same species.

We identified five putative 5′-inverted retrogenes containing A-rich tails and TSDs from four mammalian species (table 2). Like human-specific and chimpanzee-specific 5′-inverted retrocopies, these putative retrogenes were sandwiched by ~14-bp TSDs rich in adenine. There is no evidence as yet for transcription and translation of these putative 5′-inverted retrogenes, but these putative retrogenes could potentially encode proteins if they are transcribed. Among the 5 putative 5′-inverted retrogenes, only the putative retrogene of the eukaryotic translation initiation factor 3 subunit 5 (EIF3F) gene encoded a long peptide in its inverted segment (described in the next section). The other four putative retrogenes encoded only 8–12 residues in their inverted segment either because the first AUG was positioned just ~30 bp upstream of the inversion junction or the splice acceptor was positioned ~20 bp upstream of the inversion junction. Thus, the inverted segments of these four putative retrogenes do not substantially change the protein sequences, and therefore, we will not discuss them further.

In parallel, we collected 5′-inverted retrocopies from the human genome without the restriction that they were species specific and transcribed. From among these retrocopies, we selected transcribed retrocopies using the EST database. We found 139 5′-inverted retrocopies with A-rich tails (supplementary table S4, Supplementary Material online). Among them, 90 retrocopies contained TSDs and showed characteristics of L1-dependent 5′ inversion including microhomologies at the 5′ and inversion junctions. This group includes all six human-specific inverted retrocopies (table 1) and the putative retrogene of EIF3F, demonstrating the validity of this analysis.

A BlastN search of the human EST database revealed that six 5′-inverted retrocopies are transcribed (table 3). Among them, only one retrocopy contained an ORF extending from the inverted segment to the noninverted segment and encoded a peptide corresponding to part of the parental protein. This was the retrocopy of the small nuclear ribonucleoprotein polypeptide N (SNRPN) (NT_004350.18 1761725–1762268). We did not analyze the remaining five retrocopies further because they did not encode proteins that had novel peptides besides fragments similar to their parental proteins.

Putative 5′-Inverted Retrogene of EIF3F

The 5′-inverted retrocopy of EIF3F was annotated as a part of mRNA XM_372447 and the gene LOC390282
At the 5′ junction of this retrocopy, 2-bp microhomologies (GG) were observed (fig. 5B, shaded). The inversion junction at the middle of the retrocopy contained two nontemplated nucleotides, AG. Nontemplated segments 1–7 bp in length were reported at the inversion junctions of two 5′-inverted L1 copies (Ostertag and Kazazian 2001). The canonical polyadenylation signal AATAAA was positioned 28 bp upstream of the polyA tract in the retrocopy (fig. 5B, underlined). Nucleotides 149–702 of EIF3F mRNA were not copied into the retrocopy. The identity between the human retrocopy and the human parental gene was ~90% in both the inverted and noninverted segments (supplementary table S3, Supplementary Material online).

The start codon of LOC390282 was positioned 2,785 bp upstream of the 5′ TSD (fig. 5A and C). A 2,707-bp intron was annotated upstream of the retrocopy. The first exon encoded 23 residues and the second exon encoded 182 residues. The noninverted segment encoded 130 residues corresponding to the C-terminal 134 residues of the parental EIF3F protein (fig. 5A and C; yellow). The inverted segment was 132 bp in length and encoded 44 residues.

The parental gene EIF3F encodes a subunit of the mammalian eIF3 multiprotein complex that binds to the 40 S ribosome.

Table 2

<table>
<thead>
<tr>
<th>Accession No.</th>
<th>Target Site Duplications</th>
<th>Parental (Original) Gene (Accession No.)</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>XM_372447.4</td>
<td>AAGAAGGGAAGAGTGG</td>
<td>EIF3F eukaryotic translation initiation factor</td>
<td>Human</td>
</tr>
<tr>
<td>XM_484859.4</td>
<td>AGATAATACCAATGCG</td>
<td>Myo6 myosin, light polypeptide 6, alkali, smooth muscle and nonmuscle</td>
<td>Mouse</td>
</tr>
<tr>
<td>XM_001068804.1</td>
<td>AAATTTATTACAT</td>
<td>Rpsa ribosomal protein SA (NM_017138)</td>
<td>Rat</td>
</tr>
<tr>
<td>XM_001249030.1</td>
<td>AAAAATTTATTCCA</td>
<td>BRMS1L breast cancer metastasis-suppressor 1-like (XM_592281)</td>
<td>Cattle</td>
</tr>
<tr>
<td>XM_001252998.1</td>
<td>AAAATTTCAGATA</td>
<td>RAN RAN member RAS oncogene family (NM_001034705)</td>
<td>Cattle</td>
</tr>
</tbody>
</table>
ribose and promotes the binding of methionyl-tRNA and mRNA (Dong and Zhang 2006). In the C-terminal region shared by the retrocopy and the parental EIF3F gene, there were 11 residue differences and a four-residue deletion. Because the protein sequence of this region of EIF3F is completely identical between human and mouse, all changes have occurred in the retrocopy since its retrotransposition. The function of the C-terminal region is unknown, but the high conservation of the protein sequence indicates its importance; it may bind other subunits of the eIF3 complex. The N-terminal region of EIF3F lost in the retrocopy corresponds to the domain Mov34/MPN/PAD-1 family (pfam01398). This domain binds the C-terminal kinase domain of cyclin-dependent kinase 11 (CDK11) during apoptosis (Shi et al. 2003). Thus, it is obvious that the 5’-inverted putative retrogene of EIF3F does not retain this function.

To elucidate the age of this 5’-inverted retrocopy, we performed BlastN to primate assembled genomes and sequence traces. We found insertions of the 5’-inverted retrocopy with TSDs in all examined apes, that is, chimpanzee, gorilla G. gorilla, orangutan, and white-cheeked gibbon N. leucogenys (fig. 5D). In contrast, we found uninserted loci in white-tufted-ear marmoset Callithrix jacchus and northern treeshrew T. belangeri. Because TSDs are generated by duplication of a unique sequence during retrocopy insertion, an uninserted locus contains a single sequence corresponding to TSDs (in fig. 5D, sequences corresponding to TSDs are aligned with 5’ TSDs, but they can be also aligned with 3’ TSDs). Marmosets belong to New World monkeys and treeshrews belong to Scandentia, a close relative of primates. The genome sequences of two Old World monkeys, rhesus macaque and baboon P. hamadryas, do not have sequences corresponding to TSDs or the retrocopy itself. There are two possibilities to explain the absence of both the TSDs and the retrocopy in Old World monkeys. One is that the retrocopy was not inserted in Old World monkeys and the deletion of the sequence corresponding to TSDs occurred in the common ancestor of Old World monkeys after branching from apes. The other possibility is that the retrocopy was inserted in the common ancestor of Old World monkeys and apes, and Old World monkeys underwent a large deletion including the TSDs and the retrocopy. Whichever possibility is true, our data showed that the 5’-inverted retrocopy of EIF3F was inserted in the common ancestor of apes 18–40 million years ago (Ma), after the split from New World monkeys and before the branching of gibbons, according to the species-splitting times adopted from Goodman et al. (1998).

In addition to the human copy, we were able to characterize the complete sequences of the 5’-inverted retrocopy of EIF3F of chimpanzee and orangutan (data not shown). The EIF3F retrocopy of orangutan contained two frameshift mutations, and therefore, the retrocopy is unlikely to be functional. The chimpanzee retrocopy was nearly identical to the human retrocopy. The protein sequence coded by the chimpanzee retrocopy contained only three residue differences from that of human. We detected neither positive nor negative selection of this retrocopy by calculating the nonsynonymous-to-synonymous substitution rate ratio (data not shown). At present, there are no data supporting the functionality of this putative 5’-inverted retrogene of EIF3F.

Transcribed Putative 5’-Inverted Retrogene of SNRPN

We found the transcribed 5’-inverted retrocopy of SNRPN (table 3 and fig. 6A). The parental gene SNRPN encodes SmN, a component of the spliceosomal complex. SmN is the brain-specific isotype of SmB/B’ encoded by the small nuclear ribonucleoprotein polypeptide B/B’ (SNRPB). Although SNRPA and SNRPN are highly conserved at the protein level, at the nucleotide level it is clearly discernable that this retrocopy was derived from SNRPN. SmN, SmB, and other Sm proteins have the Like Sm (LSM) domain (pfam01423) at their N termini in common. The LSM domain is likely to associate with RNA.

The retrocopy of SNRPN contained a 1-bp microhomology at the inversion junction, a canonical polyadenylation signal, a polyA tail, and 12-bp TSDs (fig. 6B). The nucleotide identity between the human retrocopy and the human parental gene in the inverted segment was 89%, whereas that in the noninverted segment was 94% (supplementary table S3, Supplementary Material online). The inverted segment corresponded to the last 51 bp of the 5’ UTR and the first 204 bp of the protein-coding sequence of the parental SNRPN mRNA. The noninverted segment corresponded to the last 145 bp of the protein-coding sequence and the full 3’ UTR. If the transcription is in sense orientation, the noninverted segment encodes 48 residues corresponding to the C terminus of SNRPN. However, the National Institute of Health Mammalian Gene Collection Program (Strausberg et al. 2002) revealed that this retrocopy was transcribed in the reverse orientation, such that the inverted segment became sense and the noninverted segment became antisense. The full-length cDNA sequence BC065369 started

### Table 3

Transcribed 5’-Inverted Retrocopies in the Human Genome

<table>
<thead>
<tr>
<th>Accession No., Position</th>
<th>Parental (Original) Gene (Accession No.)</th>
<th>EST Organ (Accession No.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_004350.18 1761725–1762268</td>
<td>SNRPN small nuclear ribonucleoprotein polypeptide N (NM_003097)</td>
<td>Spleen (BQ710845), mixed (EL736139)</td>
</tr>
<tr>
<td>NT_005403.16 48104443–48102757</td>
<td>RPL4 ribosomal protein L4 (NM_000968)</td>
<td>Spleen (BQ706658)</td>
</tr>
<tr>
<td>NT_016354.18 69024265–60024888</td>
<td>HNRNPA1 heterogeneous nuclear ribonucleoprotein A1 (NM_002136)</td>
<td>Lung tumor (AA610842)</td>
</tr>
<tr>
<td>NT_009733.14 22115263–22116827</td>
<td>C1orf98 chromosome 12 ORF 48 (NM_017915)</td>
<td>Pooled skin (CB999723)</td>
</tr>
<tr>
<td>NT_025028.13 13402052–13400875</td>
<td>KRT8 keratin 8 (NM_002273)</td>
<td>Neuroblastoma (AU253661), testis (DB340669)</td>
</tr>
<tr>
<td>NT_006576.15 61669904–61670588</td>
<td>HNRNPCL1 heterogeneous nuclear ribonucleoprotein C (NM_003157)</td>
<td>Sciatic nerve (BQ948362)</td>
</tr>
<tr>
<td>NT_006576.15 61669904–61670588</td>
<td>HNRNPA1 heterogeneous nuclear ribonucleoprotein A1 (NM_002136)</td>
<td>Sciatic nerve (BQ948362)</td>
</tr>
</tbody>
</table>

SNRPN data showed that the 5’-inverted retrocopy of EIF3F was inserted in the common ancestor of apes 18–40 million years ago (Ma), after the split from New World monkeys and before the branching of gibbons, according to the species-splitting times adopted from Goodman et al. (1998).
25 bp away from the 3′ TSD, and was 1,780 bp in length ending with a polyA tail (fig. 6A and C). Transcripts have been detected from various organs: spleen (EST accession number BQ710845), prostate (CF123022), lung (BF507812, CA313794, CN478849, and D31016), and kidney (AW295379). Three EST sequences (CA313794, CN478849, and AW295379) had polyA tails; 450 nt upstream of the polyadenylation site of BC065369. There were polyadenylation signals AATAAA; 20 nt upstream of both polyadenylation sites (fig. 6C, highlighted in red). Both polyadenylation signals were downstream of the protein-coding region, indicating the existence of two transcript variants coding for the same protein.

The full-length cDNA encoded a 208-residue protein in which the N-terminal 23 residues were encoded on the noninverted segment and the middle 85 residues were on the inverted segment of the retrocopy. Due to the direction of transcription, peptides similar to the parental protein were encoded on the inverted segment. Because 51 bp of the inverted segment corresponded to the 5′ UTR of the SNRPN gene, only the middle 68 residues (41–108) had counterparts in the parental protein (fig. 6A and C, yellow). This region corresponded to the LSM domain. The LSM domain is always positioned near the N termini of Sm proteins. However, the retrocopy encoded a protein containing the LSM domain in the middle of its sequence.
FIG. 6.—Transcribed putative 5'-inverted retrogene of SNRPN. (A) Schematic structure of the 5'-inverted retrocopy of SNRPN and the mRNA BC065369. It is marked as in figure 4A. Arrowheads indicate indel positions observed only in human. (B) Junction sequences. All markers and coloring are the same as in figure 3. (C) Coding sequence. The sequence shown is in reverse orientation. All markers and coloring are the same as in figure 4B. The sequence corresponding to BC065369 is boxed. Polyadenylation signals are highlighted in red. (D) Distribution of the retrocopy of SNRPN. Coloring is the same as in figure 3.
Furthermore, although the LSM domain is composed of Sm1 and Sm2 motifs, Sm2 was lost in the retrocopy. Therefore, it is unlikely that this region has the original function of the LSM. Thus, if it is translated, the protein coded by the retrocopy may have a function unrelated to Sm proteins.

We characterized orthologous loci of this retrocopy in other primates (fig. 6D). The retrocopy insertion was observed in chimpanzee, gorilla, orangutan, gibbon, and rhesus macaque. The marmoset locus did not contain the insertion. Hence, this insertion occurred in the common ancestor of Old World monkeys and apes, after the split from New World monkeys. The insertion time was estimated between 25 and 40 Ma (Goodman et al. 1998). The comparison of insertion sequences revealed 1 nt deletion (from TTTT to TTT at positions 327–329 in fig. 6C) and 1 nucleotide insertion (from CCC to CCCC at positions 567–570 in fig. 6C) in the lineage leading to human; these indels made it possible for this human retrocopy to encode a protein. This raises the possibility that this retrocopy became a functional retrogene recently only in the human lineage. Consistent with our assumption that this retrocopy evolved neutrally until its recent functionalization, we detected neither positive nor negative selection of this retrocopy by calculating the nonsynonymous-to-synonymous substitution rate ratio (data not shown). However, given that the protein has not been detected, it is possible that this retrocopy is still under neutral evolution.

Twin Priming Combined with a Template Jump

In the process of searching for 5'-inverted retrocopies in the human genome, we encountered a complex retrocopy of chloride intracellular channel 4 (CLC4), a nuclear gene encoding a mitochondrial protein (fig. 7A and B). This retrocopy has accumulated many mutations and could not encode a protein. This retrocopy consisted of three segments of its parental gene. The 5'-segment was inverted and corresponded to the sequence 2,269–1,180. The 3'-segment was not inverted and corresponded to 3,135–4,295, but it had a 3'-extension containing a polyadenylation signal and a short A-rich tail (AAATATATA). The middle segment was also not inverted and corresponded to 429–1,155. The complex retrocopy of CLC4 was inserted in human, chimpanzee, gorilla, orangutan, gibbon, macaque, and marmoset, but was not inserted in Philippine tarsier T. syrichta, Garnett’s greater bushbaby O. garnetti, grey mouse lemur M. murinus, or treeshrew (fig. 7C). The distribution of this retrocopy showed that it was inserted in the common ancestor of New World monkeys, Old World monkeys, and apes after the split of tarsiers 40–58 Ma. TSDs were found at both sides of the retrocopy, indicating that the insertion of three segments of the retrocopy was the result of a single retrotransposition event.

The complex structure of this retrocopy can be explained by a template jump during twin priming (figs. 1 and 7D). The overall mechanism is the same as twin priming, but one step is inserted between steps K and M in figure 1. After step K, the downstream RT molecule jumps to the upstream mRNA (fig. 7D). This scenario can explain why the middle segment is in the forward orientation, why the sequence originally the 5'-terminal is copied into the middle, and why the three segments do not overlap.

This structure shows how dramatically mRNA retrotransposition can recompose an original gene sequence.

Discussion

Previous Identification of 5'-Inverted Retrocopies

We characterized six human-specific and six chimpanzee-specific 5'-inverted retrocopies. Furthermore, we characterized 139 5'-inverted retrocopies (including the six human-specific retrocopies above) from the human genome. Several unusual retrocopies similar to 5'-inverted retrocopies have been reported, although the mechanism of generation has not been determined (Rozmahel et al. 1997; Courseaux and Nahon 2001; Ejima and Yang 2003). Ejima and Yang (2003) reported a retrotransposed copy of antisense RNA of ATM (AT, mutated) gene with inversion, and proposed two mechanisms to generate 5'-inverted retrocopies, namely, 3' transduction and transl-mobilization, both of which are theoretically possible. The retrotransposed sequence reported by Rozmahel et al. (1997) contains an inverted L1 segment, a noninverted L1 segment, and an antisense sequence of cystic fibrosis transmembrane conductance regulator (CFTR) including introns and exon 9. This structure was likely generated by L1 3' transduction coupled with 5' inversion. However, all 145 (139 in human and 6 in chimpanzee) 5'-inverted retrocopies identified in this study did not contain L1 sequences. Therefore, even if some 5'-inverted retrocopies were generated by L1 3' transduction coupled with 5'-inversion, most 5'-inverted retrocopies were likely generated by mRNA retrotransposition in trans.

Courseaux and Nahon (2001) reported two unusual retrocopies of melanin-concentrating hormone (MCH) gene. These two retrocopies were duplicated by large segmental duplication 5–10 Ma and the original retrocopy was inserted ~25 Ma. The MCH gene is transcribed bidirectionally and the antisense transcript was named AROM for antisense-RNA-overlapping-MCH gene. It is very likely that the reported retrocopies are 5'-inverted retrocopies of AROM transcript generated by twin priming, although they proposed a different model for generating these retrocopies. It is obvious that they are retrotransposed products of AROM because they include sequences corresponding to MCH introns and a polyA tail. They consist of two segments; the 5' segment is inverted, whereas the 3' segment is not inverted. Their structures are similar to those of 5'-inverted retrocopies. These two inverted retrocopies are transcribed. They potentially encode a fragment of the MCH protein in their inverted segments because the inverted (antisense) segments of AROM are the sense segments of MCH. However, the full length of the protein fragment is encoded within the inverted segment. Furthermore, western blotting failed to detect the protein (Schmieder et al. 2008). Therefore, they are not examples of the type of 5'-inverted retrogene that we describe here.

5'-Inverted Retrocopy: A Possible Source of New Genes

Identification of 5'-inverted retrocopies led us to propose that 5' inversion of the retrocopy replaces the N-terminal portion of the parental protein with a novel peptide
while conserving the function of the C-terminal domain. We identified examples of 5'-inverted retrocopies that can encode proteins. Although transcription of the retrocopies of TMPO and EIF3F has not been reported, their structures indicate that the inverted segments can encode novel peptides fused to fragments of their parental proteins. The retrocopy of SNRPN is expressed as an EST in the reverse orientation but can encode a protein. We originally hypothesized that the C-terminal portion of the protein encoded by the retrogene resembles the C-terminal portion of the parental protein. Because the stop codon is conserved at the same position between the parental gene and the retrogene, the parental C-terminal region can be placed at neither the middle nor the N terminus. Therefore, 5'-inverted retrogenes necessarily encode proteins in which the N-terminal region has been replaced. The retrocopies of TMPO and EIF3F showed such presumed structures. However, when transcription occurs in the reverse orientation, more complex alterations can occur. For example, the N-terminal 68 residues of the parental SNRPN protein corresponded to the middle 68 residues sandwiched by unique sequences in the protein coded by its retrogene. Transcription in the reverse orientation enhances the potential for generating new protein structures originating from 5'-inverted retrocopies.

Low Frequency of 5' Inversion in mRNA Retrotransposition

We found 12 5'-inverted retrocopies and 131 noninverted retrocopies from human-specific and chimpanzee-specific insertions. It was reported that 8% of L1 insertions are inverted (Szak et al. 2002), but the frequency increases to ~25% when only recent L1 insertions are counted (Ostertag and Kazazian 2001). Because we analyzed human-specific and chimpanzee-specific retrocopy insertions, which have been retrotransposed during the last ~6 My, we can compare the frequency of 5' inversion of the retrocopy with that of recent L1 insertions. We confirmed that ~30% of human-specific and chimpanzee-specific L1 insertions were coupled with 5' inversion (Kojima KK, Okada N, unpublished data). Hence, the frequency of 5' inversion of retrocopies at 8.4% (12 of 143) is much lower than that of recent L1 insertions.

What causes this lower frequency of 5' inversion of retrocopies? Shorter length of mRNA than L1 RNA could be the primary factor. Most mRNAs that have many retrocopies, including ribosomal protein genes, are shorter than 1,000 nt. Abundance of short mRNA retrotransposition reduces the frequency of 5' inversion of retrocopy. The secondary structure of RNA is another possible reason. Ostertag and Kazazian (2001) raised the possibility that the secondary structure of the L1 RNA is important for determining sites that are amenable to invasion by the internal primer. L1 may have an appropriate site for the second primer to invade, whereas most mRNAs may not.

The difference between retrotransposition “in cis” and in trans could also contribute to the lower frequency of 5' inversion. It is likely that two RT molecules reverse transcribe one RNA molecule simultaneously in twin priming (fig. 1). In
contrast, standard retrotransposition is not considered to need two RT molecules because the second-strand cDNA can be synthesized by cellular DNA polymerases instead of RT, as proposed by Zingler et al. (2005). L1 proteins tend to retrotranspose the RNA encoding themselves; L1 mobilizes cellular mRNAs at a much lower frequency (0.01–0.05%) than L1 RNAs (Esnault et al. 2000; Wei et al. 2001). Even in the L1 ribonucleoprotein in vitro which included L1 proteins, L1 RNA, and other mRNAs, L1 RT more preferentially recognized L1 RNA than other mRNAs (Kulpa and Moran 2006). It is possible that two RT molecules recognize the same RNA in trans less often than one RT molecule.

5’ Inversion as a Marker for mRNA Mobilization by L1 Proteins

Given that L1 elements in reptiles and amphibians have not been characterized in detail, mammalian L1s are most closely related to fish Swimmer/L1 elements and sea urchin L1 elements (Kojima and Fujiwara 2004). Fish genomes contain fewer retrocopies than those of mammals (Yu et al. 2007). One explanation for this difference is that fish L1-like elements have a weak ability to retrotranspose mRNAs. Ichiyanagi et al. (2007) reported that the target sequence preference of L1 for TTAAAA was acquired by a certain lineage of L1 during the evolution of vertebrates, suggesting that the characteristics of human L1 are not always conserved among L1 elements in all organisms. Like the target sequence preference of L1, the trans-mobilization of mRNA by L1 proteins could have developed during vertebrate evolution.

To elucidate the contribution of L1 to mammalian genome evolution, it is important to characterize the mobilization of mRNA by L1 proteins in each organism, especially with regard to new gene formation. The mobilization of mRNA by L1 proteins has been shown experimentally only in human (Esnault et al. 2000; Wei et al. 2001). In other mammals, the involvement of L1 proteins in retrocopy formation is supported by several characteristics shared between L1 copies and retrocopies. The target sequence preference of the L1 endonuclease for TTAAAA is used as a marker for mobilization by L1 proteins (Jurka 1997). However, this characteristic is not definitive because the L1 endonuclease cleaves target DNA without strict sequence specificity and therefore sequence variations of TSDs are observed. Furthermore, ancient retrocopy insertions have often lost recognizable TSDs due to point mutations, deletions, and recombination. In contrast, 5’ inversion is clearly distinguishable even if mutations have been accumulated. If 5’-inverted L1 copies and 5’-inverted retrocopies of mRNA are found in the same organism, it strongly supports the mobilization of mRNA by L1 proteins in that organism. For example, the mouse genome contains both 5’-inverted L1 copies (Martin et al. 2005) and 5’-inverted retrocopies (table 2), supporting the idea that mouse L1 proteins can mobilize mRNA in trans. Thus, 5’ inversion of retrocopies can be used as a marker for the mobilization of mRNA by L1 proteins.

Of course, when there are no 5’-inverted retrocopies in a certain species, it does not necessarily mean that L1 proteins in this species do not have the ability to mobilize mRNAs. This is because retrocopies showing 5’ inversion are a minority of all retrocopies. Furthermore, twin priming may have evolved with the evolution of L1.

The History of Twin Priming

The 5’ inversion of L1 copies has been observed only in three mammalian species: human, mouse, and dog (Kazazian and Moran 1998; Choi et al. 1999; Szak et al. 2002; Martin et al. 2005). This raises the possibility that twin priming has evolved with the evolution of mammalian L1 elements.

We determined the ages of several 5’-inverted retrocopies. We characterized human-specific and chimpanzee-specific 5’-inverted retrocopies (table 1). These retrocopies have been inserted during the last ~6 My. The retrocopies of EIF3F, SNRPN, and CLIC4 were inserted in the common ancestor of apes 18–40, 25–40, and 40–58 Ma, respectively (figs. 5–7). In the evolution of primates, mRNA retrotransposition exploded 40–50 Ma, preceding the diversification of anthropoids (monkeys and apes) (Ohshima et al. 2003; Marques et al. 2005). This retrotransposition explosion could have generated many 5’-inverted retrocopies.

The beginning of twin priming can be traced back further. We also found 5’-inverted retrocopies from mouse, rat, and cattle (table 2). Primates and rodents (mouse and rat) belong to Euarchontoglires (Supraprimates) (Murphy et al. 2001). Dog and cattle are members of Laurasiatheria. Euarchontoglires and Laurasiatheria together form Boreotheria. The parsimonious explanation is that L1 in the common ancestor of Boreotheria had an ability to perform twin priming. It further indicates that 5’-inverted retrocopies have been generated during the last 80–90 My of the evolution of Boreotheria. It remains to be elucidated whether L1s from other mammals and nonmammals can generate 5’-inverted retrocopies. Our preliminary research has found 5’-inverted LINE insertions other than L1 (Kojima KK, Okada N, unpublished data). Therefore, it is quite likely that there are 5’-inverted retrocopies in other mammals and nonmammals.

Contributions of 5’-Inverted Retrocopies to Genome Evolution

In this study, we found only one candidate for a 5’-inverted retrogene that is transcribed and can encode a >100-residue protein. However, considering the progress of EST and full-length cDNA cloning, we believe that more transcribed 5’-inverted retrocopies will be discovered. It was reported that more than one-third of retrocopies are transcribed (Vincenbosch et al. 2006). In a simple estimate, 46 (139/3) 5’-inverted retrocopies can be transcribed in human. In addition, because the putative retrogene of SNRPN has evolved to encode a protein only in the human lineage, it is possible that 5’-inverted retrogene formation can occur long after retrocopy insertion. This suggests that mRNA retrotransposition coupled with inversion is one of the mechanisms to generate species-specific genes.
Supplementary Material

Supplementary tables S1–S4 and supplementary figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science, Technology. K.K.K. is the recipient of a Grant-in-Aid from the Japan Society for the Promotion of Science for Young scientists.

Literature Cited


Naoko Takezaki, Associate Editor

Accepted March 10, 2009