Published online 3 December 2009 *Nucleic Acids Research, 2010, Vol. 38, No. 4* ***1163–1171***
doi:10.1093/nar/gkp1098

Discovery of short pseudogenes derived from messenger RNAs

Goro Terai[1,2,*], Aya Yoshizawa[1,2], Hiroaki Okida[1,2], Kiyoshi Asai[3,4] and Toutai Mituyama[3]

[1]INTEC Systems Institute Inc., Koto-ku 136-0075, [2]Japan Biological Informatics Consortium (JBIC), [3]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku 135-0064, Tokyo and [4]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8583, Japan

Received August 24, 2009; Revised and Accepted November 9, 2009

ABSTRACT

More than 40% of the human genome is generated by retrotransposition, a series of *in vivo* processes involving reverse transcription of RNA molecules and integration of the transcripts into the genomic sequence. The mechanism of retrotransposition, however, is not fully understood, and additional genomic elements generated by retrotransposition may remain to be discovered. Here, we report that the human genome contains many previously unidentified short pseudogenes generated by retrotransposition of mRNAs. Genomic elements generated by non-long terminal repeat retrotransposition have specific sequence signatures: a poly-A tract that is immediately downstream and a pair of duplicated sequences, called target site duplications (TSDs), at either end. Using a new computer program, TSDscan, that can accurately detect pseudogenes based on the presence of the poly-A tract and TSDs, we found 654 short ($\leq$300 bp), previously unknown pseudogenes derived from mRNAs. Comprehensive analyses of the pseudogenes that we identified and their parent mRNAs revealed that the pseudogene length depends on the parent mRNA length: long mRNAs generate more short pseudogenes than do short mRNAs. To explain this phenomenon, we hypothesize that most long mRNAs are truncated before they are reverse transcribed. Truncated mRNAs would be rapidly degraded during reverse transcription, resulting in the generation of short pseudogenes.

INTRODUCTION

Retrotransposition in eukaryotes can be divided into two types; the long terminal repeat (LTR) type and the non-LTR type. The latter accounts for the majority of retrotransposition events in human (1). Various types of RNA molecules, including *Alu* RNAs, LINE RNAs, mRNAs and small noncoding RNAs (2–6) are copied via non-LTR retrotransposition. An increasing number of versatile roles for retrotransposition have been recognized, such as the generation of novel functional genes and modulation of gene expression. Insertion of LINE-1 and *Alu* in a 3′ UTR may reduce gene expression (7). Retrotransposition may have expanded regulatory elements in the promoter region (8), and some endogenous siRNAs are derived from mRNA pseudogenes (9). Retrotransposition of LINE-1 may mediate exon shuffling (10). Retrotransposition of mRNA is one mechanism for generating functional genes (11).

Non-LTR retrotransposition is mediated by the protein encoded by the second open reading frame of LINE-1 (hereafter L1-ORF2p). This protein has both reverse transcriptase and endonuclease activity (12) and promotes retrotransposition of LINE-1 RNAs themselves (3). The endonuclease activity of L1-ORF2p creates a cleavage site in genomic DNA (12); the cleavage site is used as a primer, and reverse transcription of template RNAs and integration of the resultant cDNAs into the genome occur simultaneously. This integration process is called target-site-primed reverse transcription (TPRT) (13,14). In addition to LINE-1 RNAs, L1-ORF2p recognizes *Alu* RNAs (2) and the mRNAs of protein-coding genes and promotes their retrotransposition, although its recognition efficiency for protein-coding genes is much lower than for LINE-1 and *Alu* (2,4,15).

*To whom correspondence should be addressed. Tel: +81 3 5665 5011; Fax: +81 3 5665 5095; Email: terai_goro@intec-si.co.jp

In many LINE-1 and mRNA pseudogenes observed in the human genome, the 5′-end region of the template transcript has been truncated (1,16,17). This has long been explained by the inability of L1-ORF2p to copy the entire length of the template RNA during retrotransposition, or degradation of the template RNA before completion of reverse transcription (1). However, full-length (nontruncated) LINE-1 are also frequently observed (18–21). The mechanism for the preferential generation of full-length LINE-1 has not been explained.

In mammals, there are three types of sequence signatures around a retrotransposed element (Figure 1). The first is a poly-A tract found immediately downstream of the 3′-end of a retrotransposed element (1). The second is a pair of duplicated sequences surrounding the retrotransposed element, called target site duplications (TSDs) (1). The third is the TTAAAA consensus sequence, which overlaps with the 5′-end of the 5′-TSD. This consensus sequence is recognized by L1-ORF2p endonuclease to create the cleavage site in genomic DNA, but is not always present (14,22). The mechanisms generating the poly-A tract and TSD are not fully understood, but the presence of these sequence signatures is an established phenomenon and can be used to detect retrotransposed elements.

In this study, we developed a novel algorithm for detecting retrotransposed elements based on the presence of the poly-A tract and TSDs and implemented this algorithm as the TSDscan program. Because TSDscan uses general sequence signatures surrounding retrotransposed elements, it is able to detect any type of sequence element generated by non-LTR retrotransposition. TSDscan detected many previously unknown short pseudogenes generated by retrotransposition of mRNA. TSDscan also allows us to analyze detailed characteristics of pseudogenes, such as the length distribution of TSDs and poly-A tracts, which are useful for further study of the molecular mechanisms of pseudogenes generation. From this analysis, we found that short pseudogenes are more frequently generated from long mRNAs than from short mRNAs. In order to explain this phenomenon in the context of events previously reported to be associated with retrotransposition, we propose that two *in vivo* processes generate pseudogenes: short parent mRNAs use template-jumping to generate a full-length pseudogene, whereas long parent mRNAs are more likely to be truncated and degraded, after which microhomology generates short pseudogenes of the mRNA. The findings we have presented here provide new insights into the mechanism of retrotransposition.

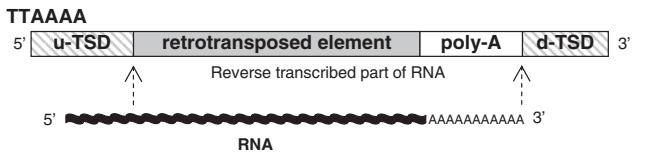

Figure 1. Signature sequences of a retrotransposed element. Several features can be used to identify a retrotransposed element. A poly-A tract (usually 5–30 bp) is located downstream of the retrotransposed element and a pair of duplicated sequences, the TSDs, are located at either side of the retrotransposed element—poly-A structure. The TSD is also usually 5–30 bp. The AAAA in the TTAAAA consensus sequence overlaps with the u-TSD (22). Reverse transcription of an RNA starts from its 3′-end and does not always copy the entire length (the reverse-transcribed region is indicated with arrows). Therefore, a retrotransposed element often lacks the 5′ region of its parent RNA.

MATERIALS AND METHODS

TSDscan: an algorithm to detect retrotransposed elements

TSDscan is an algorithm that not only detects retrotransposed elements but can also predict the length and boundaries of sequence signatures surrounding the retrotransposed elements. In TSDscan, upstream and downstream sequences of a retrotransposed element are aligned and scored with a specific scheme. To detect the sequence signatures shown in Figure 1, we need to consider that the lengths of poly-A tracts and TSDs are variable and that random mutations accumulate in these sequence signatures. Because the upstream TSD (u-TSD) and downstream TSD (d-TSD) are similar, TSDs can be detected by assigning positive scores to a base match and assigning negative scores to a base mismatch or gaps, analogous to the usual alignment technique for detecting similar regions in genes (23). The poly-A tract is detected by assigning positive scores to the insertion of a poly-A sequence immediately before the d-TSD. Figure 2A is an example of the alignment and the scores assigned to each alignment column. For alignment columns corresponding to TSDs (dotted rectangles), a pair of aligned nucleotides has a score defined in the HOXD matrix (24). The first gap and the gaps following it have scores of −400 and −30, respectively. In a poly-A tract located before a d-TSD, insertion of nucleotide A and of other nucleotides (G, T and C) have scores of +100 and −100, respectively. The total score is the sum of scores assigned to each alignment column. In the case of Figure 2A, the total score is 1129.

Figure 2B is an example of an alignment containing the TTAAAA consensus sequence. Alignment columns corresponding to the TTAAAA consensus are shown in

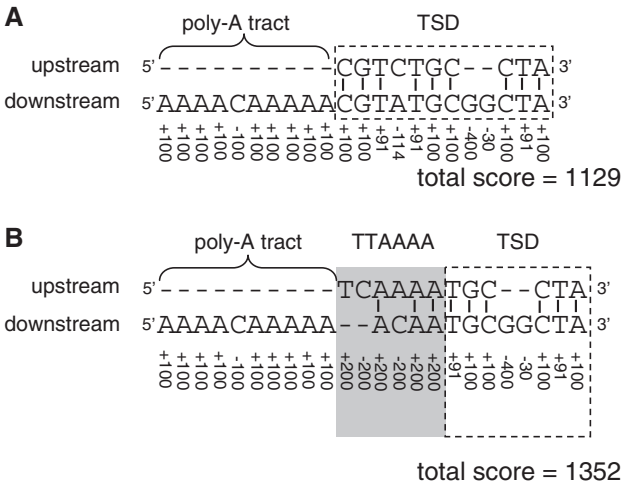

Figure 2. Examples of how scores are assigned to each potential retrotransposed element. Examples of alignments containing (**A**) a poly-A tract and TSD and (**B**) containing a TTAAAA consensus sequence, poly-A tract and TSD.

the gray area. In the first two columns of the gray area, insertion of nucleotide T and of other nucleotides (A, G and C) have scores of +200 and −200, respectively. In the last four columns, an A–A nucleotide match and the other aligned nucleotide pairs have scores of +200 and −200, respectively. Scores assigned to columns outside the gray area are the same as in Figure 2A. In the case of Figure 2B, the total score is 1352.

Next, we consider the positions of the poly-A tract and TSDs. The poly-A tract and TSDs are inserted immediately outside of a retrotransposed element (Figure 1). Therefore, poly-A tracts and TSDs that are distant from a retrotransposed element should be penalized. In TSDscan, each nucleotide insertion between u-TSD and the 5′-end of the retrotransposed element, and between the 3′-end of the retrotransposed element and a poly-A tract, has a score of −50.

In TSDscan, the alignment maximizing the total score is detected with a dynamic programming algorithm. Details of the algorithm are provided in Supplementary Text S1. In addition, the source code of TSDscan (perl and C++ versions) is available at http://www.intec-si.co.jp/technology/rna/, which may help with understanding the algorithmic details.

Detection of mRNA pseudogenes

Deleting repeats from a genomic sequence. If *Alu* or LINE-1 were inserted within a pseudogene, the pseudogene would be disrupted. Determining both ends of such a pseudogene becomes a bit complicated, and we needed a way to cope with this complexity, because, in our method, both ends of pseudogenes need to be determined fairly precisely. To circumvent the complexity, we created a genomic sequence from which *Alu* and LINE-1 are deleted. In addition, tandem repeat sequences detected by tandem repeats finder (TRF) (25) were masked by 'N'. Hereafter, the genomic sequence thus created is called the 'processed genome'.

Searching homologous regions of mRNAs. We obtained the nucleotide sequences of all human mRNAs from the 'Human mRNAs' track of the human genome (version hg17) in the UCSC genome browser (26). mRNAs without 3′ UTR annotation were excluded from further study. Then we deleted *Alu* and LINE-1 from the mRNA sequences and masked tandem repeat sequences detected by TRF. Using these mRNAs as queries, we searched the processed genome using blastz with the default parameters. Among blastz hits, we excluded those that did not contain the 3′ UTR of query mRNAs, because a pseudogene of an mRNA should contain the 3′ UTR.

Excluding overlap with known genes. We converted positions of blastz hits in the processed genome into positions in the original genome. Then we excluded blastz hits that overlapped with exons of human mRNAs.

Excluding redundancy of blastz hits. Blastz hits often overlap with each other. In such cases, we excluded the blastz hit with the lower score.

Applying TSDscan. We applied TSDscan to the regions 100 bp upstream and downstream of blastz hits. Then we extracted blastz hits with a TSDscan score of 1100 or more. Among 10 000 genomic regions that we randomly selected, only ∼3% had a score of 1100 or more (Supplementary Figure S1). We excluded blastz hits having TSDs or poly-A tracts of >30 bp even if the score was at least 1100 because such long TSDs and poly-A tracts were rarely seen for LINE-1 and *Alu* (Supplementary Figures S2 and S3), and thus they may be false positives.

Evaluating the accuracy of TSDscan and TSDfinder

For our evaluation study, we designated two types of LINE-1 subfamilies (L1P: primate specific LINE-1 and L1M: mammalian-wide LINE-1) and three types of *Alu* subfamilies (*Alu*Y, *Alu*S and *Alu*J) as positive samples, and randomly selected genomic regions as negative samples. The detection accuracy is measured by the ACC score, which is the average of sensitivity and specificity. Sensitivity and specificity are defined as:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN}),$$

where TP, FP, TN and FN are the number of true positives, false positives, true negatives, and false negatives, respectively.

TSDfinder (27) is a program that defines the boundaries of a retrotransposed element based on the presence of TSDs. The TSDfinder program consists of several steps, including merging and determining the boundaries of a retrotransposed element, obtaining sequences surrounding the retrotransposed element, detecting potential TSDs, and scoring TSDs. To evaluate TSDfinder using exactly the same test data as we used for TSDscan, we modified the TSDfinder program such that we could input sequence data directly.

Sequence data of *Alu*, LINE-1 and random genomic regions

We used RepeatMasker (http://www.repeatmasker.org/) to detect sequence data for *Alu* and LINE-1. RepeatMasker often detects poly-A tracts in the 3′-end of *Alu* and LINE-1 as a part of repetitive sequences. To avoid this, we excluded poly-A tracts from the 3′-end of the consensus sequences of LINE-1 and *Alu*, and ran RepeatMasker using the truncated consensus sequences as queries. Among the *Alu* and LINE-1 sequences detected by RepeatMasker, we excluded those which lacked ≥5 bp of the 3′-end, because *Alu* and LINE-1 should contain the 3′-end of their original transcripts at the time they are retrotransposed. We also excluded *Alu* and LINE-1 if their upstream and downstream 100 bp contained other repetitive sequences detected by RepeatMasker. Among the remaining *Alu* and LINE-1 sequences, we randomly selected 1000 samples for each *Alu* and LINE-1 subfamily (L1P, L1M, *Alu*Y, *Alu*S and *Alu*J).

Data for random genomic regions were generated by sampling 10 000 genomic regions that did not overlap

with repetitive sequences identified by RepeatMasker and TRF.

RESULTS AND DISCUSSION

Detection of pseudogenes derived from mRNA

We obtained 84 332 mRNA sequences with a 3′ UTR annotation from the 'Human mRNAs' track in the UCSC genome browser, human genome version hg17 (26). Using these mRNA sequences as queries, we performed homology searches against the human genome by using blastz (24). We excluded blastz hits that did not have homology to the 3′ UTR, because, as shown in Figure 1, reverse transcription starts from the 3′-end of the mRNA, and an mRNA pseudogene should contain the 3′ UTR. After also excluding overlapping blastz hits, we obtained 27 465 hits, which we considered candidate pseudogenes. We applied TSDscan to the 100-bp upstream and downstream regions of these pseudogene candidates and extracted those with a score of 1100 or higher. Among 27 465 candidate pseudogenes, 6982 passed the score threshold. Then, we excluded candidate pseudogenes having TSDs or poly-A tracts of >30 bp even if the score was at least 1100. Among the 6982 high scoring candidates, 4464 passed the length threshold of Poly-A tract and TSDs, and we considered these to be mRNA pseudogenes. Genomic coordinates of the mRNA pseudogenes are provided in Supplementary Table S1.

Discovery of short pseudogenes derived from mRNAs

Most of the novel pseudogenes, that is, pseudogenes that did not overlap with the existing pseudogene annotations (16,17), were short (Figure 3A). Median length of the novel pseudogenes is 356 bp, which is much shorter than that of all pseudogenes identified in this study (691 bp). By discovering the novel pseudogenes, the length distribution of pseudogenes in the human genome changed significantly (Figure 3B). Our results contained 645 new pseudogenes with lengths of ≤300 bp. To verify that the new short pseudogenes were not caused by a limitation of the TSDscan algorithm, we investigated the length distribution of TSDs for short pseudogenes. The TSD length distribution of the 645 short pseudogenes, as well as that of all pseudogenes detected in this study, had a peak ~15 bp (Figure 3C). The TSD length distribution

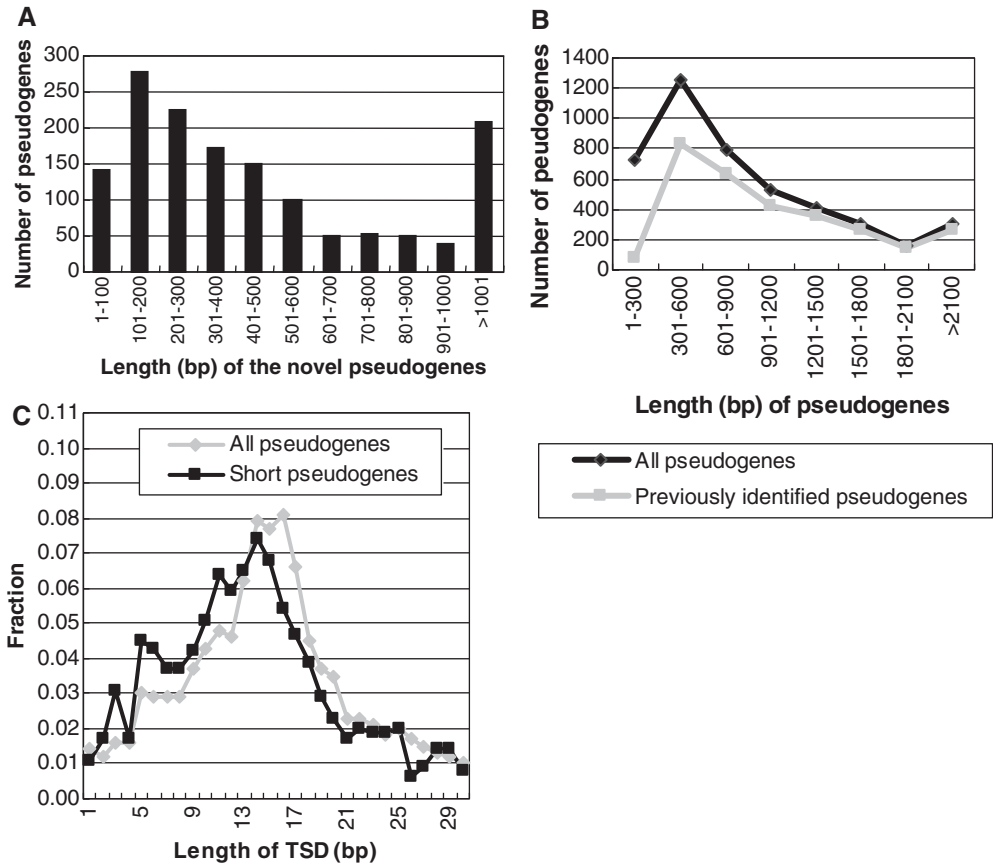

Figure 3. Length distribution of pseudogenes and their TSDs. (**A**) Length distribution of pseudogenes newly found in this study. The median length is 356 bp (mean 533.2 with SD 562.6 bp). (**B**) Comparison of length distributions of previously identified pseudogenes and all pseudogenes identified with TSDscan. Previously identified pseudogenes are the pseudogenes identified by TSDscan that overlap with the existing pseudogene annotation (16,17). The median length of previously identified pseudogenes is 877 bp (mean 1096.7 with SD 756.9 bp) and that of all pseudogenes identified with TSDscan is 691 bp (mean 911.0 with SD 747.3 bp). (**C**) Length distributions of TSDs of the 654 short pseudogenes and all pseudogenes identified with TSDscan.

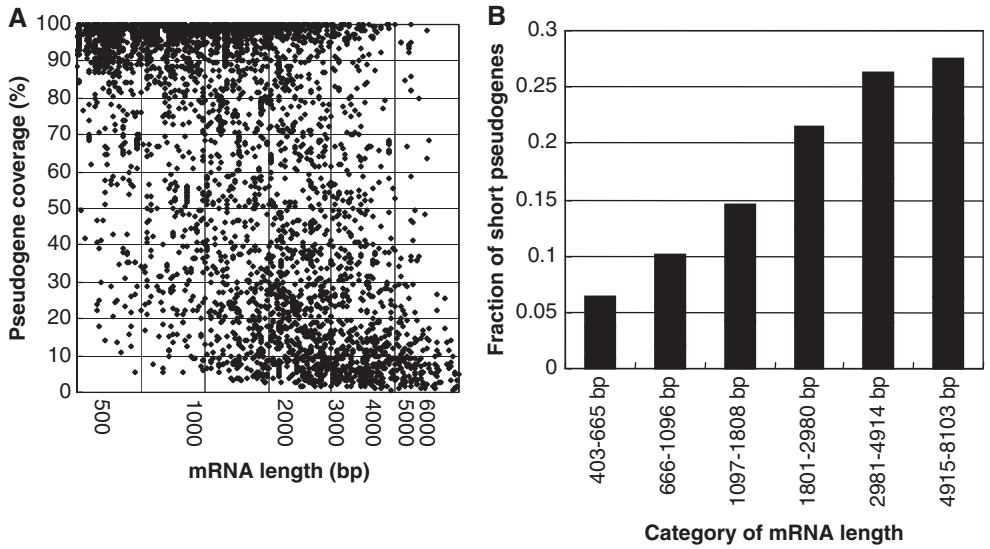

Figure 4. (**A**) Relationship between parent mRNA length and pseudogene coverage. Parent mRNA length (*X*-axis) is shown in a logarithmic scale. Pseudogene coverage (*Y*-axis) is the length of a pseudogene divided by that of its parent mRNA. (**B**) Fraction of short pseudogenes (≤300 bp) calculated for each mRNA length category.

peak is consistent with that of LINE-1 and *Alu* (22,27,28; Supplementary Figure S2), supporting the validity of the short pseudogenes we detected.

Relationship between parent mRNA and pseudogene length

Because our method could accurately predict the boundaries of a pseudogene, we could investigate the relationship between parent mRNA and pseudogene length. Figure 4A is a two-dimensional plot of parent mRNA length (*X*-axis) versus pseudogene coverage, i.e. the length of the pseudogene relative to that of the parent mRNA (*Y*-axis). A cluster can be seen along the line of pseudogene coverage = 100%, indicating that full-length pseudogenes are frequent. Another cluster of plots can be seen in the lower right area, suggesting that short and truncated pseudogenes are frequently generated from long mRNAs. The fraction of full-length pseudogene gradually decreased as the length of the parent mRNA became longer, and conversely, the fractions of short and truncated pseudogenes gradually increased as the length of the parent mRNAs became longer (Supplementary Figure S4). In addition, it can be seen that (i) most pseudogenes derived from short mRNAs are full length; (ii) most pseudogenes derived from long mRNAs are short and truncated; and (iii) for medium-to-long mRNAs, both full-length and short truncated pseudogenes are frequent. Therefore, the length distribution of pseudogenes of medium-to-long mRNAs has two peaks, similar to the bimodal length distribution already reported for LINE-1 elements (18–21,29).

To explicitly show the relationship between the fractions of short (≤300 bp) pseudogenes and parent mRNA length, we divided parent mRNA length into categories and calculated the fraction of the short pseudogenes for each mRNA length category (Figure 4B). The boundaries of the categories are shown by the vertical lines in Figure 4A. As can be seen in Figure 4B, short pseudogenes were generated more frequently from long mRNAs than from short mRNAs, and the longer a parent mRNA was, the more frequently short pseudogenes were generated. Reverse transcription of template RNAs by L1-ORF2p is an essential step of retrotransposition, but poor processivity of L1-ORF2p alone cannot explain why long mRNAs more frequently generate short pseudogenes. Here, we hypothesize that most long mRNAs are truncated before they are reverse transcribed. Details are described later in this section.

Comparison of TSDscan with existing methods of identifying retrotransposed elements

Two other software packages, RTanalyzer (30) and TSDfinder (27), are available to detect retrotransposed elements. RTanalyzer detects retrotransposed elements based on the presence of a poly-A tract and TSDs. Potential TSDs are first identified by local alignment, and the final score is calculated based on the presence of the poly-A tract and the TTAAAA consensus sequence. However, because RTanalyzer is available only through a web interface, we could not evaluate its accuracy using our large test data set and therefore could not include RTanlyzer in our comparison.

TSDfinder is a program that defines the boundaries of a retrotransposed element based on the presence of TSDs. In TSDfinder, potential TSDs are identified by aligning the upstream and downstream regions of a retrotransposed element and detecting those that have perfect nucleotide matches in at least 9 consecutive base pairs (27). The final score is calculated by considering both the TSD position and alignment score. To compare TSDscan with TSDfinder, we measured the detection accuracy by using the ACC score, an average of sensitivity

Table 1. Comparison of the detection accuracy of TSDscan and TSDfinder

	TSDscan			TSDfinder		
	ACC_{max}[a]	Sensitivity	Specificity	ACC_{max}[a]	Sensitivity	Specificity
L1P	0.926	0.914	0.937	0.628	0.263	0.994
L1M	0.834	0.821	0.848	0.525	0.056	0.995
*Alu*Y	0.991	0.991	0.991	0.788	0.582	0.994
*Alu*S	0.975	0.968	0.981	0.679	0.365	0.994
*Alu*J	0.949	0.941	0.957	0.596	0.198	0.994

[a]The ACC score is the average of sensitivity and specificity, and ACC_{max} is the maximum ACC score.

and specificity scores (see 'Materials and Methods' section). Table 1 shows the detection accuracy of TSDscan and TSDfinder for five types of retroposons (L1P, L1M, *Alu*Y, *Alu*S and *Alu*J). In all five types, the ACC score of TSDscan was higher than that of TSDfinder; therefore, for the purpose of detecting retrotransposed elements, TSDscan is superior to TSDfinder. The sensitivity of TSDfinder was relatively low (Table 1), which may be due to the stringent criterion of at least nine perfect nucleotide matches for detecting TSDs. In contrast, our method has greater sensitivity because of its flexible requirement for detecting TSDs.

A proposed model for generating short pseudogenes

This is the first large-scale analysis of short pseudogenes derived from mRNAs in the human genome. In previous studies, pseudogenes were detected mostly based on their lack of introns and the accumulation of random mutations in their protein-coding sequences (16,17,31). However, these methods cannot be applied to short pseudogenes, because most short pseudogenes are derived from last exons that lack protein-coding sequences, where homology searches do not work effectively. Therefore, short pseudogenes have escaped detection in previous studies. In addition to discovering novel short pseudogenes, our method accurately predicts the boundaries of pseudogenes. This enabled us to closely investigate the essential characteristics of pseudogenes.

Using TSDscan, we made the novel discovery that long mRNAs tend to produce a higher percentage of short pseudogenes than do short mRNAs. Although TPRT is the currently accepted mechanism of retrotransposition (1), it does not explain this length-dependent phenomenon. Here, we propose that two *in vivo* processes generate pseudogenes in a length-dependent manner. We hypothesize that most long mRNAs are truncated before they are reverse transcribed (Figure 5A). Because RNAs without a 5′ cap are rapidly digested (32), the template RNA may be removed during reverse transcription. After removal of the template RNA, a single-stranded cDNA is exposed, which is integrated into the genome by the microhomology-mediated mechanism proposed by Zingler *et al.* (28). If reverse transcription is completed before digestion of the template RNA, L1-ORF2p moves to a genomic 3′ overhang via template jumping (33). After the genomic 3′ overhang region is reverse transcribed, removal of the template RNA and synthesis of the remaining strand occur. In contrast, when mRNAs are short, they are rarely truncated (Figure 5B). The 5′ cap of an mRNA protects it from digestion, giving L1-ORF2p a good chance to complete reverse transcription. Subsequently, L1-ORF2p moves to a genomic 3′ overhang via template jumping (33). After the genomic 3′ overhang region is reverse transcribed, removal of the template RNA and synthesis of the remaining strand occur to generate a full-length pseudogene. Although the role of the 5′ cap structure in retrotransposition has not been studied, it has been strongly suggested that LINE-1 RNAs also have the 5′ cap structure because of the frequent guanines at the 5′-end of full-length LINE-1 elements (34). Our hypothesis (Figure 5) can explain the bimodal length distribution of LINE-1 elements, which has been reported by many researchers (18–21,29), and which cannot be explained by the TPRT mechanism alone.

If our hypothesis is true, how are RNAs truncated in a length-dependent manner? We infer that each nucleotide in all RNAs is cleaved with roughly equal probability, and thereby long mRNAs are more likely to be truncated. Assuming that the cleavage of each nucleotide is a rare event and occurs with the same probability, λ, the number of cleaved nucleotides in each RNA molecule should follow a Poisson distribution. The probability that there is no cleaved nucleotide in a given RNA is expressed as follows:

$$P^{\text{Full-length}}(L) = e^{-\lambda L},$$

where L is the length of the RNA. By taking logarithms on both sides, we obtain the following simple equation:

$$Y = -\lambda L,$$

where Y is $\log_e[P^{\text{full-length}}(L)]$. The fractions of full-length pseudogenes we found are well fitted by the above expression ($P = 1.95 \times 10^{-5}$ by F-test; Figure 6), supporting our inference of equiprobable nucleotide cleavage in the RNAs being retrotransposed.

CONCLUSION

In this study, we developed a novel method for detecting retrotransposed elements and found that the human genome contains many previously unidentified short pseudogenes generated by retrotransposition of mRNAs,

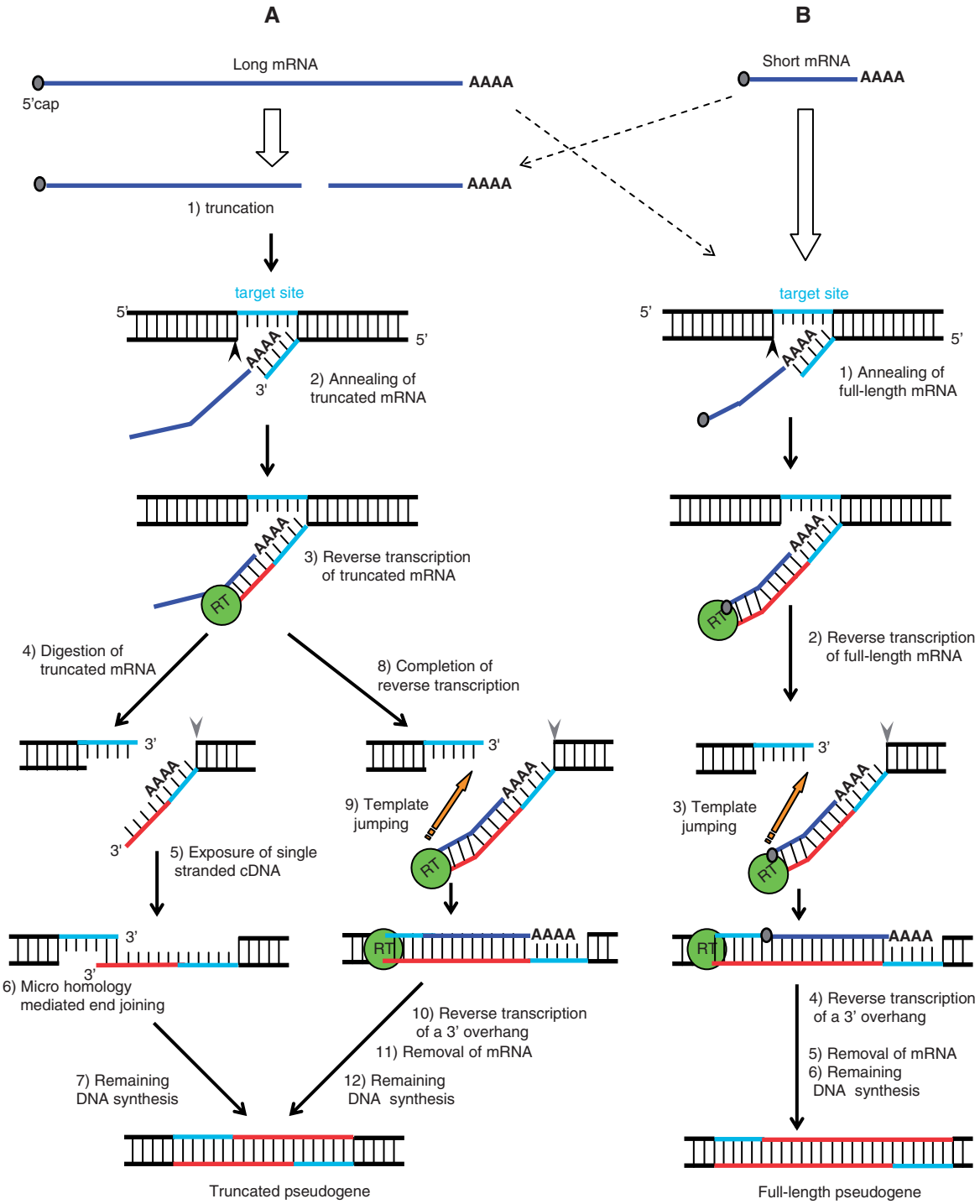

Figure 5. Hypothetical mechanism for the generation of short pseudogenes. (**A**) Generation of short truncated pseudogenes from long mRNAs. 1) Most long mRNAs are truncated before they are reverse transcribed. 2) After first strand cleavage (indicated by a black arrowhead), truncated mRNAs are annealed at the nick. 3) Reverse transcription of truncated mRNAs by L1-ORF2p proceeds. 4) Truncated mRNAs are digested before the completion of reverse transcription. 5) After second-strand cleavage (indicated by a gray arrowhead), a single-stranded cDNA is exposed, and 6) it base-pairs with a genomic 3′ overhang (microhomology-mediated end joining) (28). Finally, 7) the remaining DNA synthesis is completed. 8) If reverse transcription is completed before digestion of the template RNA, 9) the L1-ORF2p jumps from the template mRNA onto a genomic 3′ overhang. 10) After the genomic 3′ overhang region is reverse transcribed, 11) removal of the template RNA and 12) synthesis of the remaining strand occur. (**B**) Generation of full-length pseudogenes from short mRNAs. 1) Full-length mRNAs are annealed at the nick. 2) Reverse transcription of full-length mRNAs by L1-ORF2p proceeds. 3) After reverse transcription is completed, the L1-ORF2p jumps from the template mRNA onto a genomic 3′ overhang. 4) After reverse transcription of the genomic 3′ overhang is completed, 5) the template mRNA is removed and 6) the remaining DNA synthesis is completed.

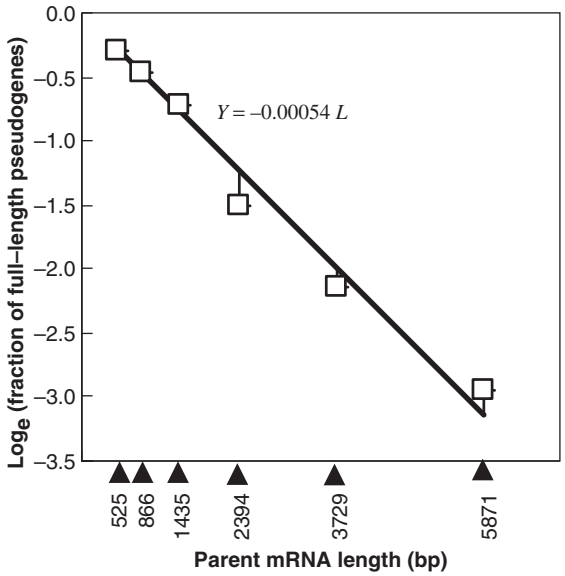

Figure 6. Relationship between the fraction of full-length pseudogenes and parent mRNA length. White boxes represent natural logarithms of the fractions of full-length pseudogenes. Pseudogenes were considered to be full-length if they were longer than 90% of the parent mRNA length. Black arrowheads at the bottom of the figure indicate mean lengths of mRNAs in each mRNA length category. The regression line was obtained by the least squares method.

which gives more complete view of pseudogenes in the human genome. By utilizing our findings, we performed comprehensive analyses of pseudogenes and their parent mRNAs, which presented interesting propensities: short pseudogenes are more likely sourced from long mRNAs than short mRNAs. Importantly, this length-dependent phenomenon cannot be explained by the currently accepted mechanism of retrotransposition alone. Therefore, in order to explain this phenomenon, we propose a novel mechanism in which two different *in vivo* processes, previously reported to be associated with retrotransposition, are involved in the generation of pseudogenes. The findings we have presented here provide important insights into the mechanism of retrotransposition.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank members of the bioinformatics group at the National Institute of Advanced Industrial Science and Technology (AIST) and the members of the Japan Biological Information Consortium (JBIC) for useful discussions. They also thank Yasuo Tabei for discussing the dynamic programming used in our method, Dr Martin Frith for a useful suggestion about alignment parameters and Dr Yasunori Aizawa for discussing our model for generating pseudogenes.

FUNDING

The Functional RNA Project funded by the New Energy and Industrial Technology Development Organization (NEDO). Funding for open access charge: Computational Biology Research Center (CBRC) and National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Ostertag,E.M. and Kazazian,H.H. Jr (2001) Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.*, **35**, 501–538.
2. Dewannieux,M., Esnault,C. and Heidmann,T. (2003) LINE-mediated retrotransposition of marked *Alu* sequences. *Nat. Genet.*, **35**, 41–48.
3. Moran,J.V., Holmes,S.E., Naas,T.P., DeBerardinis,R.J., Boeke,J.D. and Kazazian,H.H. Jr (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917–927.
4. Esnault,C., Maestre,J. and Heidmann,T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.*, **24**, 363–367.
5. Buzdin,A., Gogvadze,E., Kovalskaya,E., Volchkov,P., Ustyugova,S., Illarionova,A., Fushan,A., Vinogradova,T. and Sverdlov,E. (2003) The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res.*, **31**, 4385–4390.
6. Perreault,J., Noël,J.F., Brière,F., Cousineau,B., Lucier,J.F., Perreault,J.P. and Boire,G. (2005) Retropseudogenes derived from the human Ro/SS-A autoantigen-associated hY RNAs. *Nucleic Acids Res.*, **33**, 2032–2041.
7. Faulkner,G.J., Kimura,Y., Daub,C.O., Wani,S., Plessy,C., Irvine,K.M., Schroder,K., Cloonan,N., Steptoe,A.L., Lassmann,T. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.
8. Bourque,G., Leong,B., Vega,V.B., Chen,X., Lee,Y.L., Srinivasan,K.G., Chew,J.L., Ruan,Y., Wei,C.L., Ng,H.H. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–1762.
9. Watanabe,T., Totoki,Y., Toyoda,A., Kaneda,M., Kuramochi-Miyagawa,S., Obata,Y., Chiba,H., Kohara,Y., Kono,T., Nakano,T. *et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
10. Moran,J.V., DeBerardinis,R.J. and Kazazian,H.H. Jr (1999) Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530–1534.
11. Babushok,D.V., Ostertag,E.M. and Kazazian,H.H. Jr (2007) Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol. Life Sci.*, **64**, 542–554.
12. Feng,Q., Moran,J.V., Kazazian,H.H. Jr and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
13. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
14. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J.*, **21**, 5899–5910.
15. 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.
16. Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
17. Zhang,Z., Harrison,P.M., Liu,Y. and Gerstein,M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the

processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.
18. Boissinot,S., Chevret,P. and Furano,A.V. (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.*, **17**, 915–928.
19. Pavlícek,A., Paces,J., Zíka,R. and Hejnar,J. (2002) Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. *Gene*, **300**, 189–194.
20. Myers,J.S., Vincent,B.J., Udall,H., Watkins,W.S., Morrish,T.A., Kilroy,G.E., Swergold,G.D., Henke,J., Henke,L., Moran,J.V. *et al.* (2002) A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.*, **71**, 312–326.
21. Salem,A.H., Myers,J.S., Otieno,A.C., Watkins,W.S., Jorde,L.B. and Batzer,M.A. (2003) LINE-1 preTa elements in the human genome. *J. Mol. Biol.*, **326**, 1127–1146.
22. Jurka,J. (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA*, **94**, 1872–1877.
23. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) Pairwise alignment. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, pp. 12–45.
24. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R., Haussler,D. and Miller,W. (2003) Human–Mouse Alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
25. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
26. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
27. Szak,S.T., Pickeral,O.K., Makalowski,W., Boguski,M.S., Landsman,D. and Boeke,J.D. (2002) Molecular archeology of L1 insertions in the human genome. *Genome Biol.*, **3**, research0052.
28. Zingler,N., Willhoeft,U., Brose,H.P., Schoder,V., Jahns,T., Hanschmann,K.M., Morrish,T.A., Löwer,J. and Schumann,G.G. (2005) Analysis of 5′ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5′-end attachment requiring microhomology-mediated end-joining. *Genome Res.*, **15**, 780–789.
29. Babushok,D.V., Ostertag,E.M., Courtney,C.E., Choi,J.M. and Kazazian,H.H. Jr (2006) L1 integration in a transgenic mouse model. *Genome Res.*, **16**, 240–250.
30. Lucier,J.F., Perreault,J., Noël,J.F., Boire,G. and Perreault,J.P. (2007) RTAnalyzer: a web application for finding new retrotransposons and detecting L1 retrotransposition signatures. *Nucleic Acids Res.*, **35**, W269–W274.
31. Ohshima,K., Hattori,M., Yada,T., Gojobori,T., Sakaki,Y. and Okada,N. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and *Alu* repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.*, **4**, R74.
32. Newbury,S.F. (2006) Control of mRNA stability in eukaryotes. *Biochem. Soc. Trans.*, **34**, 30–34.
33. Bibillo,A. and Eickbush,T.H. (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J. Biol. Chem.*, **279**, 14945–14953.
34. Lavie,L., Maldener,E., Brouha,B., Meese,E.U. and Mayer,J. (2004) The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.*, **14**, 2253–2260.