Discovery of short pseudogenes derived from messenger RNAs

Goro Terai1,2,*, Aya Yoshizawa1,2, Hiroaki Okida1,2, Kiyoshi Asai3,4 and Toutai Mituyama3

1INTEC Systems Institute Inc., Koto-ku 136-0075, 2Japan Biological Informatics Consortium (JBIC), 3Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku 135-0064, Tokyo and 4Department 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 (≤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).
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.

**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...
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 
between 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 >30bp 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 (AluY, AluS and AluJ) 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:

Sensitivity = TP/(TP + FN)
Specificity = TN/(FP + 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 ≥5bp 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, AluY, AluS and 
AluJ).

Data for random genomic regions were generated by 
sampling 10000 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
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; (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 RTanalyzer 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...
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, AluY, AluS and AluJ). 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 and synthesis of the remaining strand, the 5' cap structure is no longer present, giving L1-ORF2p a good chance to complete reverse transcription. Accordingly, 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.

<table>
<thead>
<tr>
<th>TSDscan</th>
<th>TSDfinder</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC&lt;sub&gt;max&lt;/sub&gt;</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>L1P</td>
<td>0.926</td>
</tr>
<tr>
<td>L1M</td>
<td>0.834</td>
</tr>
<tr>
<td>AluY</td>
<td>0.991</td>
</tr>
<tr>
<td>AluS</td>
<td>0.975</td>
</tr>
<tr>
<td>AluJ</td>
<td>0.949</td>
</tr>
</tbody>
</table>

The ACC score is the average of sensitivity and specificity, and ACC<sub>max</sub> is the maximum ACC score.

**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.
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 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**