Retrotransposition as a Source of New Promoters

Kohji Okamura and Kenta Nakai

Human Genome Centre, Institute of Medical Science, University of Tokyo, Tokyo, Japan; and Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Kawaguchi, Japan

The fact that promoters are essential for the function of all genes presents the basis of the general idea that retrotranspositions give rise to processed pseudogenes. However, recent studies have demonstrated that some retrotransposed genes are transcriptionally active. Because promoters are not thought to be retrotransposed along with exonic sequences, these transcriptionally active genes must have acquired a functional promoter by mechanisms that are yet to be determined. Hence, comparison between a retrotransposed gene and its source gene appears to provide a unique opportunity to investigate the promoter creation for a new gene. Here, we identified 29 gene pairs in the human genome, consisting of a functional retrotransposed gene and its parental gene, and compared their respective promoters. In more than half of these cases, we unexpectedly found that a large part of the core promoter had been transcribed, reverse transcribed, and then integrated to be operative at the transposed locus. This observation can be ascribed to the recent discovery that transcription start sites tend to be interspersed rather than situated at 1 specific site. This propensity could confer retrotransposability to promoters per se. Accordingly, the retrotransposability can explain the genesis of some alternative promoters.

Introduction

Retrotransposition is the molecular mechanism that leads to the formation of processed pseudogenes in genomes of a wide range of species. In the process of retrotransposition, a spliced mRNA is first reverse transcribed to cDNA by endogenous reverse transcriptase activity, which could be provided by LINE-1 in mammals, and randomly inserted into a certain chromosomal region. Although several possibilities have been proposed for the machinery involved in this phenomenon, a decisive model has yet to be demonstrated (Lewin 1983; Sharp 1983; Vanin 1985). For instance, a single-stranded cDNA containing a poly(T) head could be integrated via the hybridizing a poly(A) tract in genomes. Subsequently, the DNA repair system might synthesize the second strand to complete the double strand. In any case, because transcription initiates downstream from a core promoter region, the promoter is not thought to be transcribed; hence, retrotransposed genes are generally transcriptionally inactive. In addition, the decreased functional restraint means that these genes tend to accumulate mutations, thus becoming pseudogenes. A number of studies have focused on these pseudogenes (Zheng et al. 2007) and on retroelements such as *Alu* and LINE-1 (Sellis et al. 2007) in an attempt to determine their implication in molecular evolution (Nouvel 1994; Yu et al. 2007).

It has been shown that the retroduplication of genes can also generate new functional genes (Soares et al. 1985; Marques et al. 2005), as seen in segmental duplication (Bailey et al. 2002; Cheung et al. 2003; Ward and Thornton 2007). Whereas almost all retrotransposition events presumably result in processed pseudogenes or are negatively selected, some of them must somehow result in the acquisition of a new active promoter. One such example is the human *KLF14* gene, which encodes a member of the Krüppel-like family of transcription factors (Parker-Katiraee et al. 2007). It has been suggested that retrotransposition of the *KLF16* gene posterior to the divergence between eutherians and marsupials gave rise to this intronless gene. Its exclusive expression from the maternal allele has been demonstrated in both human and mouse, suggesting that the retrotransposed gene is undoubtedly transcribed. Incidentally, it has also been suggested that retrotransposition can cause distinctive gene expression such as genomic imprinting (Yoder et al. 1997; Suzuki et al. 2007).

To advance our knowledge of the evolutionary construction of promoters, it is therefore crucial to assess the degree of acquisition of regulatory sequences by retroduplicated genes. A chimeric structure between a pre-existing gene and part of the transposed coding sequence is often found in retrotransposed genes (Long et al. 2003). Although this seems to be a predominant mechanism for promoter acquisition, in the present study, we sought to probe for other possible processes that might explain the observations described above. To this end, we set out to stringently identify plausible gene pairs consisting of a retrotransposed gene and its source gene. LINE-1 was documented to be transcribed from its internal promoter by RNA polymerase III (Kurose et al. 1995). The present study was restricted to RNA polymerase II promoters by using a 5′ cap–dependent method. After comparing the promoter regions of the selected gene pairs with reference to our database of transcription start sites (DBTSS; Yamashita et al. 2006), we unexpectedly observed that core promoters are occasionally transcribed along with their downstream exonic sequences. This observation can be ascribed to the fact that the locations of transcription start sites (TSSs) fluctuate to some extent in most genes (Suzuki et al. 2001; Frith et al. 2008). If a TSS upstream of the promoter region is used, a large part of the core promoter may be transcribed. This idea could be extended to the speculation that promoters per se have retrotransposability, which might also explain the genesis of alternative promoters at unrelated or distinct loci.

Materials and Methods

Identification of the 29 Gene Pairs

From 24,837 human RefSeq entries (September 2007), intronless genes and intron-containing genes were selected to construct a BLASTN query set and the database,

Key words: promoter, transcription start site, retrotransposition, alternative promoter, human genome, molecular evolution.

E-mail: knakai@ims.u-tokyo.ac.jp

Mol. Biol. Evol. 25(6):1231–1238. 2008
doi:10.1093/molbev/msn071
Advance Access publication March 25, 2008

respectively. As described in Results and Discussion, histone cluster genes, keratin-associated protein genes, protocadherin genes, taste receptor genes, olfactory receptor genes, and other G protein–coupled receptor genes were precluded from analysis. As a result, 631 intronless genes and 23,599 other genes were selected as the candidates for retrotransposed genes and their source genes, respectively. The BLASTN search was performed with "-F F -e 0.01 -S 1" options. After selecting alignments whose scores were no less than 500, we obtained 137 hits, which contain redundancy. Manual curation finally organized these into the 29 gene pairs. Standard criteria (Gardiner-Garden and Frommer 1987) were used to determine whether each gene contains a CpG island.

Search of All Human cDNA Sequences for Similarity to Unrelated Promoters

To find promoters that might be generated by a retrotransposition of an unrelated gene, we performed a BLASTN search using a database consisting of all human RefSeq genes. From the DBTSS site, 32,122 promoter sequences, including alternative promoters, were obtained for all human genes and used as the query set. The length of each sequence was 301 bp, including region 100 bp upstream and 200 bp downstream of the representative TSSs defined in DBTSS. We selected hits whose aligned length and sequence identity (found by BLASTN search, see below) were above 100 bp and 80%, respectively. Self-hits were also removed. Finally, those hits whose subjects contain exon junctions or a polyadenylation site were accepted for analysis.

Sequence Alignments

Final sequence alignments between 2 paralogous sequences were performed using ClustalW 1.83 (Thompson et al. 1994). The sequence similarity between the *LDHAL6B* and *DKFZp686H1233* promoters were identified by BLAT (Kent 2002), available on the UCSC Web site.

Results and Discussion

Retrotranspositions are particularly commonly observed in mammalian genomes (Zhang et al. 2004; Marques et al. 2005). To further investigate this phenomenon and to utilize the DBTSS based on the oligo-capping method (Maruyama and Sugano 1994), we decided to analyze the human genome to identify gene pairs consisting of a transcriptionally active retrotransposed gene and its source gene. As a nonredundant protein-coding gene set, we employed NCBI RefSeq (Pruitt et al. 2007) mRNAs whose accession numbers begin with "NM_". We undertook a strategy in which genes that appeared likely candidates for retrotransposed genes were searched against a database of the nonredundant gene set to identify their parental genetic entities.

Among the RefSeq mRNA data set (24,837 sequences), intronless genes (1,242 genes) are good candidates for having undergone retrotransposition. Although genes that have few introns could also be considered as retrotransposed genes (Soares et al. 1985), they might have arisen from the chimeric rearrangements mentioned above. Because we intended to exclusively study promoter creation, only intronless genes were taken into account. In addition, intronless gene could in turn be a source of additional retrotransposed genes. In this case, however, it becomes quite difficult to distinguish these entities from those resulting from segmental duplications, in which promoters are often copied along with the transcribed region. To ensure that the events studied there are indeed retrotranspositions, intronless genes were excluded as candidates for source genes. Furthermore, some classes of gene families, for example, olfactory receptor genes, tend to be highly clustered in certain chromosomal regions by segmental duplication. These genes were therefore also excluded from our study (see Materials and Methods for details) to simplify the analysis and improve its reliability.

A total of 137 gene pairs were obtained from a BLASTN search (Altschul et al. 1997) consisting of 631 queries against 23,599 subjects, using conservative criteria (see Materials and Methods). These pairs included a number of hits among mRNAs transcribed from the same locus due to alternative splicing or usage of alternative promoters. Moreover, the expression of certain intronless genes is uncertain because their transcription cannot be confirmed by the detection of splicing. We therefore defined transcriptionally active genes, in addition to accreditation by RefSeq, as genes that are supported by at least 1 oligo-capped clone (Yamashita et al. 2006). In the database, an expressed sequence tags (EST) that cannot be mapped to a single locus due to the existence of extensive genomic sequence similarity is discarded to maintain the reliability of the information. We finally compiled a list of 29 gene pairs that have undergone retrotransposition and that are likely to have generated transcriptionally active retrocopies in the human genome (table 1).

To investigate the core promoter architecture of retroduplicated genes, we then sought to align the genomic sequences between the gene counterparts. The promoters were first compared. If they were not aligned, the promoter of the retrotransposed gene was compared with the protein-coding sequence (CDS) of its source gene. If they still did not align, we then searched the human genome for sequences similar to the core promoters of retrotransposed genes, in an attempt to determine their origin. For the purpose of this study, we loosely define a core promoter as a relatively short region surrounding a cluster of TSSs, which is essential for recruitment of the transcription complex and the initiation of transcription. The 29 gene pairs were readily grouped into 1 of 4 categories based on the way in which the new copies acquired their promoters: I) the promoter of the source gene was transcribed, reverse transcribed, and integrated along with its downstream exonic region, thus becoming the new promoter; II) part of the CDS of the source gene became the new promoter; III) a promoter of an unrelated gene was copied and became the new promoter; and IV) acquisition of the new promoter could not be explained by sequence similarity (table 1).

Table 1
Gene Pairs Consisting of a Retrotransposed Gene and Its Source Gene

		Source Gene				Retrotransposed Gene			
No.	Acquisition Type	Name	Accession No.	Chromosomal Location	Number of Clones[a]	Name	Accession No.	Chromosomal Location	Number of Clones[a]
1	I	*CTAGE5*	NM_005930	14q13	21	*CTAGE6*	NM_001008747	7q35	1
2	I	*GK*	NM_000167	Xp21	26	*GK2*[b]	NM_033214	4q13	43
3	I	*GLUDI*	NM_005271	10q23	32	*GLUD2*	NM_012084	Xq24	13
4	I	*GSPTI*	NM_002094	16p13	32	*GSPT2*	NM_018094	Xp11	102
5	I	*GUSB*	NM_000181	7q21	194	*LOC441046*	NM_001011539	4q31	2
6	I	*H3F3B*	NM_005324	17q25	122	*LOC440093*	NM_001013699	12p11	1
7	I	*HMGB1*	NM_002128	13q12	90	*HMG1L1*[b]	NM_001008735	20q13	20
8	I	*MORF4L1*	NM_006791	14q24	154	*MORF4*[b]	NM_006792	4q33	13
9	I	*NACA*	NM_005594	12q23	178	*NACA2*	NM_199290	17q23	1
10	I	*PABPC1*	NM_002568	8q22	266	*PABPC3*	NM_030979	13q12	49
11	I	*PAPOLA*	NM_032632	14q32	126	*PAPOLB*	NM_020144	7p22	18
12	I	*PDHA1*	NM_000284	Xp22	135	*PDHA2*	NM_005390	4q22	1
13	I	*RAB6A*	NM_002869	11q13	304	*RAB6C*	NM_032144	2q21	4
14	I	*RPL10*	NM_006013	Xq28	207	*RPL10L*	NM_080746	14q13	2
15	I	*TAF1*	NM_004606	Xq13	14	*TAF1L*	NM_153809	9p21	1
16	I	*TRAM1*	NM_014294	8q13	46	*TRAM1L1*	NM_152402	4q26	20
17	II	*CTBP2*	NM_001329	10q26	2	*MGC70870*	NM_203481	17	2
18	II	*WDR21A*	NM_015604	14q24	80	*WDR21B*	NM_001029955	4p13	8
19	II	*WDR21A*	NM_015604	14q24	80	*WDR21C*	NM_152418	8q21	17
20	III	*LDHAL6A*	NM_144972	11p15	1	*LDHAL6B*	NM_033195	15q22	14
21	IV	*ACTB*	NM_001101	7p15	11,128	*DKFZp686D0972*[b]	NM_001017992	5q11	1
22	IV	*BIRC4*	NM_001167	Xq25	7	*BIRC8*	NM_033341	19q13	11
23	IV	*KLHL13*[b]	NM_033495	Xq23	40	*KLHL9*	NM_018847	9p22	240
24	IV	*PGK1*	NM_000291	Xq13	482	*PGK2*[b]	NM_138733	6p12	65
25	IV	*RANBP5*[b]	NM_002271	13q32	14	*RANBP6*	NM_012416	9p24	77
26	IV	*RRAGB*	NM_006064	Xp11	58	*RRAGA*	NM_006570	9p22	105
27	IV	*TKTL1*	NM_012253	XQ28	154	*TKTL2*	NM_032136	4q32	19
28	IV	*WDR42A*	NM_015726	1q22	54	*WDR42B*[b]	NM_001017930	Xp21	1
29	IV	*WDR5*	NM_017588	9q34	8	*WDR5B*	NM_019069	3q21	67

[a] Number of oligo-capped clones deposited in DBTSS.

[b] Genes that lack CpG islands around the promoter region.

A genomic sequence alignment of *PABPC1* and *PABPC3* (No. 10 in table 1)—2 genes that code for a poly(A)-binding protein—is shown as an example of acquisition type I (fig. 1). The intronless *PABPC3* gene is thought to have retrotransposed from the *PABPC1* locus, which contains more than 10 introns. Amino acid sequence identity and similarity between the 2 are 92% and 97%, respectively. TSSs identified on the basis of oligo-capped clones are shaded in the figure 1. For both genes, TSSs are interspersed around promoter regions and even into protein-coding regions, as indicated by boxes. A similar scattered distribution of TSSs is observed in all genes classified into this category. Sequence alignments of other pairs can be seen in supplementary figure S1 (Supplementary Material online). These data, as well as the frequency of each TSS, can be retrieved from the DBTSS Web site (http://dbtss.hgc.jp).

Conventionally, transcription of a gene has been believed to start at a specific site located at the most upstream position of the transcript (fig. 2). To determine the total length and TSS of a given cDNA, researchers typically employ techniques such as RACE, among others. Because truncated cDNA fragments are common, shorter clones are not generally considered as full-length cDNAs; however, the accumulation of full-length cDNA clones by methods that recognize 5′ cap structure demonstrates that TSSs for a single gene exhibit a scattered distribution around its promoter region (Suzuki et al. 2001; Carninci et al. 2006; Frith et al. 2008) rather than being situated at a single fixed site (Hampsey 1998). As seen in the case of *PABPC1* and *PABPC3*, such broad TSS regions were notably observed in the source and retrotransposed gene sets; averages of standard deviations were determined to be 66 and 108 bp, respectively.

For the *PABPC3* gene, 49 oligo-capped clones are deposited in DBTSS and mapped to show each TSS. The RefSeq data entry of *PABPC3* assumes the most upstream TSS reported in DBTSS (9 sites are shown in fig. 1) as the working TSS of the gene, implying that the region from this site to the first methionine codon is its 5′ untranslated region (UTR). Beyond this 5′ UTR, the upstream genomic sequences are highly conserved between *PABPC1* and *PABPC3*, and there are a number of TSSs for *PABPC1* (fig. 1). However, the most frequently used TSS, endorsed by RefSeq, is located in a region further upstream where the 2 genomic sequences are no more aligned (85 bp upstream from the region shown in fig. 1). Although the integration boundary is unclear in some cases, it is likely that the retrotransposed fragment was transcribed from a TSS that resided in more upstream promoter region. A large part of the core promoter of the source gene, that is, *PABPC3*, appears to have been transcribed, reverse transcribed, and then integrated. Despite the high degree of observed sequence similarity, all the TSSs of *PABPC1* and *PABPC3* except

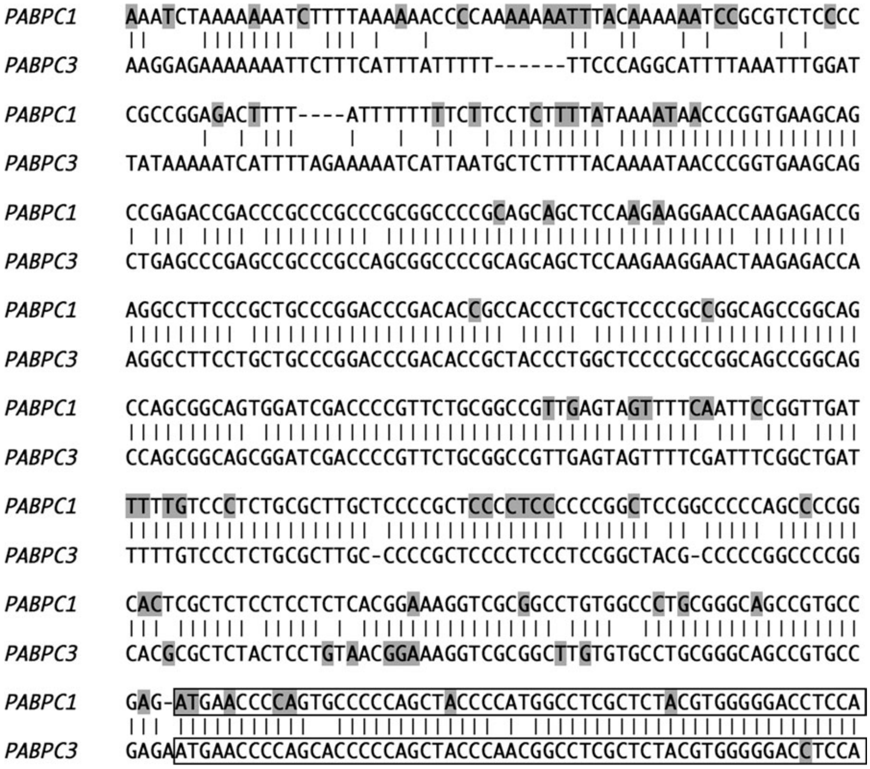

PABPC1 AAATCTAAAAAAATCTTTTAAAAAACCCCAAAAAAATTTACAAAAAATCCGCGTCTCCCC
PABPC3 AAGGAGAAAAAAATTCTTTCATTTATTTTT------TTCCCAGGCATTTTAAATTTGGAT
PABPC1 CGCCGGAGACTTTT----ATTTTTTTTCTTCCTCTTTTATAAAATAACCCGGTGAAGCAG
PABPC3 TATAAAAATCATTTTAGAAAAATCATTAATGCTCTTTTACAAAATAACCCGGTGAAGCAG
PABPC1 CCGAGACCGACCCGCCCGCCCGCGGCCCCGCAGCAGCTCCAAGAAGGAACCAAGAGACCG
PABPC3 CTGAGCCCGAGCCGCCCGCCAGCGGCCCCGCAGCAGCTCCAAGAAGGAACTAAGAGACCA
PABPC1 AGGCCTTCCCGCTGCCCGGACCCGACACCGCCACCCTCGCTCCCCGCCGGCAGCCGGCAG
PABPC3 AGGCCTTCCTGCTGCCCGGACCCGACACCGCTACCCTGGCTCCCCGCCGGCAGCCGGCAG
PABPC1 CCAGCGGCAGTGGATCGACCCCGTTCTGCGGCCGTTGAGTAGTTTTCAATTCCGGTTGAT
PABPC3 CCAGCGGCAGCGGATCGACCCCGTTCTGCGGCCGTTGAGTAGTTTTCGATTTCGGCTGAT
PABPC1 TTTTGTCCCTCTGCGCTTGCTCCCCGCTCCCCTCCCCCCGGCTCCGGCCCCCAGCCCCGG
PABPC3 TTTTGTCCCTCTGCGCTTGC-CCCCGCTCCCCTCCCTCCGGCTACG-CCCCCGGCCCCGG
PABPC1 CACTCGCTCTCCTCCTCTCACGGAAAGGTCGCGGCCTGTGGCCCTGCGGGCAGCCGTGCC
PABPC3 CACGCGCTCTACTCCTGTAACGGAAAGGTCGCGGCTTGTGTGCCTGCGGGCAGCCGTGCC
PABPC1 GAG-ATGAACCCCAGTGCCCCCAGCTACCCCATGGCCTCGCTCTACGTGGGGGACCTCCA
PABPC3 GAGAATGAACCCCAGCACCCCCAGCTACCCAACGGCCTCGCTCTACGTGGGGGACCTCCA

FIG. 1.—Sequence alignment of promoter regions of *PABPC1* and *PABPC3*. Recent studies have demonstrated that TSSs are often interspersed around promoter regions, as seen here for *PABPC1* and *PABPC3*. Each TSS is shaded, and protein-coding sequences are indicated by boxes. The length of the 5′ UTR, which spans from 1 TSS to the first methionine codon, varies in each case. TSSs were observed even in coding sequences. Frequency of each TSS is available in graphical format at the DBTSS Web site.

for 1 do not coincide with each other. A short region (approximately 50 bp long) in the retrotransposed fragment functions as a core promoter for *PABPC3*, without a drastic change in the nucleotide sequence. A similar structural pattern was also seen in all the pairs grouped into category I; most of the promoter of the source gene was retrotransposed, and it functions as a new promoter of the retroduplicated gene.

Gene pairs consisting of *WDR21A* and *WDR21B*/*C* (Nos 18 and 19 in table 1) are an example of promoter acquisition category II. All 3 genes encode a WD (tryptophan-aspartate) repeat–containing protein. *WDR21A* is assumed to be the source gene of retrotransposed genes *WDR21B* and *WDR21C*. In fact, it is difficult to conclude that both *WDR21B* and *WDR21C* were generated independently via retrotransposition of *WDR21A*. It is also possible that *WDR21C* was caused by segmental duplication of *WDR21B*, or vice versa. Because the neighboring genes (genomic context) surrounding *WDR21B* and *WDR21C* are different, we deduced that there were probably 2 independent retrotransposition events. An alignment of *WDR21A* and *WDR21C* is shown in figure 3. In contrast to the previous example, part of the protein-coding region of the source gene functions as the promoter of *WDR21C*. Because the 2 sequences do not align around the promoter of *WDR21A*, it is unlikely that the core promoter was retrotransposed. *WDR21C* employs a downstream AUG codon as its translation start sites. Its upstream coding sequence is degenerated, but it might be transformed into a new promoter. This is also the case with *WDR21B*, although the sequences and TSS positions are different between *WDR21B* and *WDR21C* (supplementary fig. S2, Supplementary Material online). Due to a preferential affinity of the transcription initiation complex to exons rather than to introns, weak alternative exonic promoters have been proposed (Carninci et al. 2006; Sandelin et al. 2007). It might be important to consider these exonic promoters in order to understand the evolutionary construction of promoters.

In other gene pairs, promoter sequences could not be aligned, unlike the observations for promoter acquisition categories I and II. It seems that the acquisition of promoters cannot be explained by sequence similarity for these retrotransposed genes. In the case of the *LDHAL6B* gene (No. 20 in table 1), however, the promoter is similar to that of *DKFZp686H1233* (accession number AL833331). These 2 loci are mapped to human chromosomes 15 and 12, respectively. Hence, 1 of these promoters was generated from the other via either retrotransposition or interchromosomal segmental duplication. If splicing was seen in 1 of the 2, we could conclude that the event was a retrotransposition.

In any case, the scattered distribution of TSSs seems to enable the promoter to have retrotransposability per se, by initiating transcription from an upstream TSS. This notion could also explain the appearance of alternative promoters,

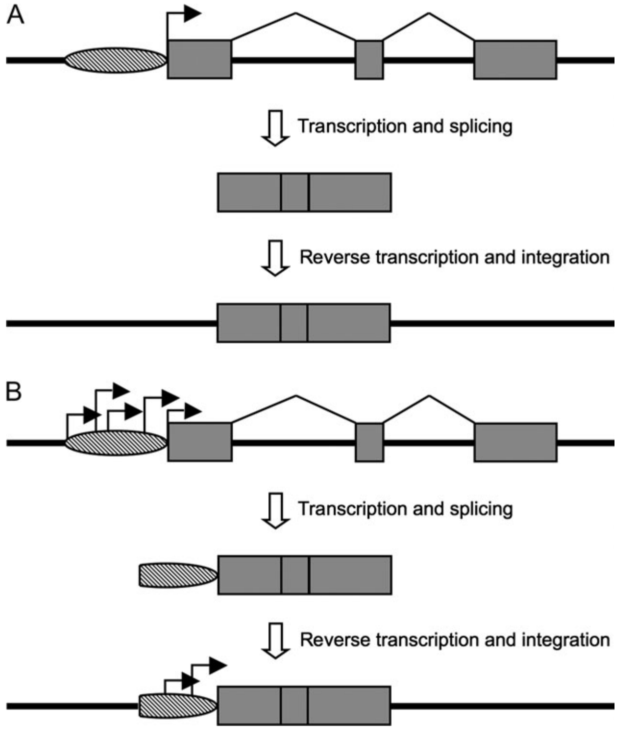

A
Transcription and splicing
Reverse transcription and integration
B
Transcription and splicing
Reverse transcription and integration

FIG. 2.—Schematic representation of possible models of retrotransposition, showing conventional (*A*) and proposed (*B*) views of retrotransposition. Ovals, arrows, and arrow heights indicate promoters, TSSs, and their relative expression levels, respectively.

which are found in more than half of all human genes (Kimura et al. 2006). When a reverse-transcribed fragment containing a core promoter is integrated around the 5′ region of a gene, it is possible for this retrotransposed fragment to function as a new promoter for the preexisting gene. As mentioned above, this assumption appears to be supported by the fact that a chimeric structure between them is often observed. A newly combined gene can be expressed either with or without splicing. Because splice donor sites are shorter than acceptor sites, it is possible that a transposed sequence behaves as a donor to a downstream splice acceptor site that resides at the 5′ end of an internal exon of the preexisting gene.

Finally, to test the above assumption, we searched all human promoters for evidence of retrotransposition. If a promoter is produced by a retrotransposition, it may bear, downstream, a processed sequence of its source locus. We constructed a database consisting of 24,837 RefSeq cDNA sequences for BLASTN analysis. For its query set, 32,122 promoters were collected from DBTSS, in which a 500-bp interval was adopted as a reliable parameter to separate 2 discrete clusters of TSSs (Kimura et al. 2006). Because it is difficult to discriminate alternative promoters from canonical promoters, all promoters were considered together in this study. Using our strict conditions, 7 hits that contain splice junctions only in subject were obtained (fig. 4 and supplementary fig. S3, Supplementary Material online). The 7 retrotranspositions occurred into different chromosomes, and the inserted cDNA fragment gave rise to a new promoter in each case. The results appear to support our hypothesis.

Conclusions

To rule out a possible contribution of segmental duplication to our observation, we first compiled intronless genes and then searched for their parental genes bearing introns. The gene pairs were likely generated by processed retrotransposition, therefore offering a unique opportunity to investigate how a new gene acquires its promoter. Because the accumulation of single-nucleotide substitutions seems too slow to construct a functional promoter, a new promoter is more likely to be a copy of another functional promoter unit from somewhere else in the genome. As expected, we found that the *LDHAL6B* gene gained a copy of another gene's promoter, following the retrotransposition from its parental gene, *LDHAL6A*. Nevertheless, this kind of acquisition was just 1 of 29 events identified in this study.

Although it is believed that retrotransposition involves only exonic regions, the results of the present study highlight the fact that promoter sequences can also be transcribed and integrated (fig. 2). This phenomenon can be ascribed to a recent finding that TSSs tend to be interspersed around the core promoter rather than positioned at 1 or a few specific sites. If an upstream TSS is employed, a large part of the promoter region is indeed transcribed. More surprisingly, as is seen in figures 1 and 3, the positions of TSSs and expression levels inferred from numbers of clones (table 1) are generally different between a retrotransposed gene and its source gene despite of a high degree of similarity between the 2 sequences.

All retrotranspositions that lead to a transcriptionally active gene or an alternative promoter result from integration into either an intergenic region or an intron. This is plausible because the total length of these regions greatly exceeds that of exons and UTRs in the human genome. Moreover, if such an integration event occurs into an exon or a UTR, the affected gene might be disrupted by the insertion and negatively selected. Other functional elements, such as promoters, splice donors, and acceptors, would also be intolerant to retrotransposition insertions. It is possible that the appearance of an alternative promoter causes a dominant-negative mutation. If an insertion occurs on the opposite strand, it may generate an antisense transcript that could be harmful. It is therefore likely that the integration sites per se have dramatically decided the fate of the lineage; the individual or its offspring is negatively, neutrally, or positively selected depending on the position of the insertion. Evolutionarily, these retrotranspositions had to have occurred in the germ line to be observed by researchers.

Scattered TSSs are particularly commonly present in CpG islands (Yamashita et al. 2005); indeed, 27 of the 29 source genes studied here have a CpG island in their 5′ region (table 1). This frequency is much higher than the ratio of genes that possess CpG islands in relation to all human genes (Bird 1986; Yamashita et al. 2005). All the 17 source genes classified into category I have a CpG island. It seems

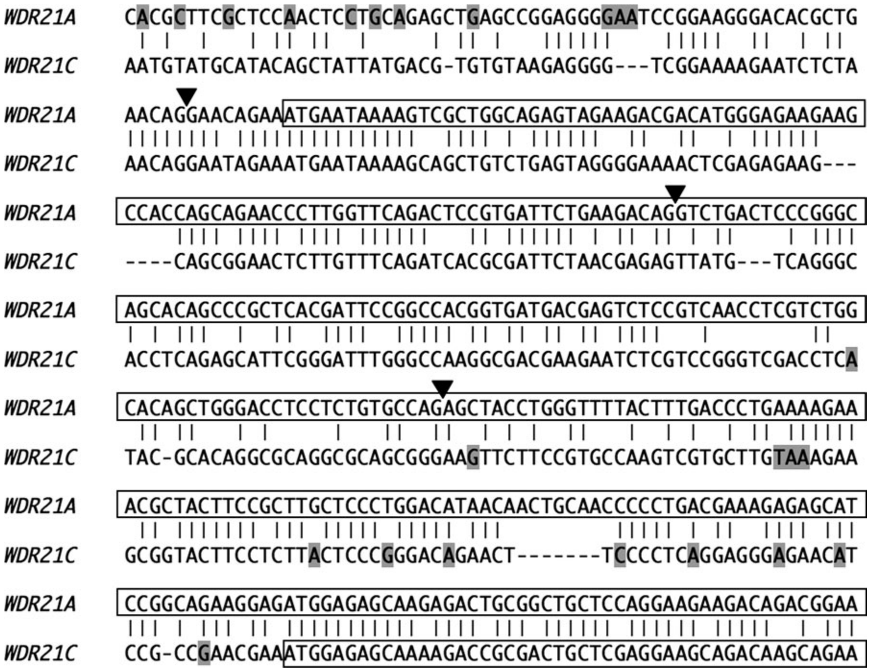

WDR21A CACGCTTCGCTCCAACTCCTGCAGAGCTGAGCCGGAGGGGAATCCGGAAGGGACACGCTG
WDR21C AATGTATGCATACAGCTATTATGACG-TGTGTAAGAGGGG---TCGGAAAAGAATCTCTA
WDR21A AACAGGAACAGAAATGAATAAAAGTCGCTGGCAGAGTAGAAGACGACATGGGAGAAGAAG
WDR21C AACAGGAATAGAAATGAATAAAAGCAGCTGTCTGAGTAGGGGAAAACTCGAGAGAAG---
WDR21A CCACCAGCAGAACCCTTGGTTCAGACTCCGTGATTCTGAAGACAGGTCTGACTCCCGGGC
WDR21C ----CAGCGGAACTCTTGTTTCAGATCACGCGATTCTAACGAGAGTTATG---TCAGGGC
WDR21A AGCACAGCCCGCTCACGATTCCGGCCACGGTGATGACGAGTCTCCGTCAACCTCGTCTGG
WDR21C ACCTCAGAGCATTCGGGATTTGGGCCAAGGCGACGAAGAATCTCGTCCGGGTCGACCTCA
WDR21A CACAGCTGGGACCTCCTCTGTGCCAGAGCTACCTGGGTTTTACTTTGACCCTGAAAAGAA
WDR21C TAC-GCACAGGCGCAGGCGCAGCGGGAAGTTCTTCCGTGCCAAGTCGTGCTTGTAAAGAA
WDR21A ACGCTACTTCCGCTTGCTCCCTGGACATAACAACTGCAACCCCCTGACGAAAGAGAGCAT
WDR21C GCGGTACTTCCTCTTACTCCCGGGACAGAACT-------TCCCCTCAGGAGGGAGAACAT
WDR21A CCGGCAGAAGGAGATGGAGAGCAAGAGACTGCGGCTGCTCCAGGAAGAAGACAGACGGAA
WDR21C CCG-CCGAACGAAATGGAGAGCAAAAGACCGCGACTGCTCGAGGAAGCAGACAAGCAGAA

FIG. 3.—Sequence alignment of promoter regions of *WDR21A* and *WDR21C*. The cDNA sequence is shown for the *WDR21A* gene. The region consists of 4 exons, and the splice junctions are indicated by inverted triangles. It is reasonable to make this kind of alignment because introns are generally precluded in retrotransposed genes. Each TSS is shaded, and protein-coding sequences are indicated by boxes. The TSS cluster of *WDR21A* extends further upstream, where the 2 sequences are no longer aligned.

that transcripts regulated by a CpG island tend to have an increased ability to copy their promoters by retrotransposition. Nonetheless, not every retrotransposition generates transcriptionally active genes. Whereas only transcribed retrotransposed genes are selected and studied here, a large number of processed pseudogenes have been cataloged (Zhang et al. 2003). Most of them are transcriptionally inactive, probably due to the absence of promoters. Even if an operative promoter is present, these genes may be rendered silent by the lack of appropriate enhancers, the presence of insulators or silencers, DNA methylation, or chromatin modifications.

We note that the *KLF14* gene mentioned in the Introduction was not detected in our study because the alignment score, 198, between *KLF14* (NM_138693) and *KLF16* (NM_031918) is below our set threshold of 500. Relaxation of this threshold would have allowed us to detect this pair. However, conservative criteria can significantly reduce the rate of false-positive discoveries. Indeed, these 2 genes are too degenerated to align in their promoter regions, as is shown that the retrotransposition is an ancient event (Parker-Katiraee et al. 2007). In contrast to intronic and intergenic sequences, protein-coding sequences have a high CpG content (Okamura et al. 2006). Hence, splicing of premRNA molecule can condense the frequency of the CpG dinucleotide and possibly create a novel CpG island that generally escapes DNA methylation. This holds true for *KLF14*, suggesting the existence of additional explanation for promoter creation. Retrotranspositions may have been occurring throughout eukaryotic evolution. In general, recent events preserve the sequences and ancient events do not, thereby rendering a quantitative analysis of this phenomenon difficult and unreliable (Soares et al. 1985). However, our stringent analysis seems to suggest a high frequency of promoter retrotransposition.

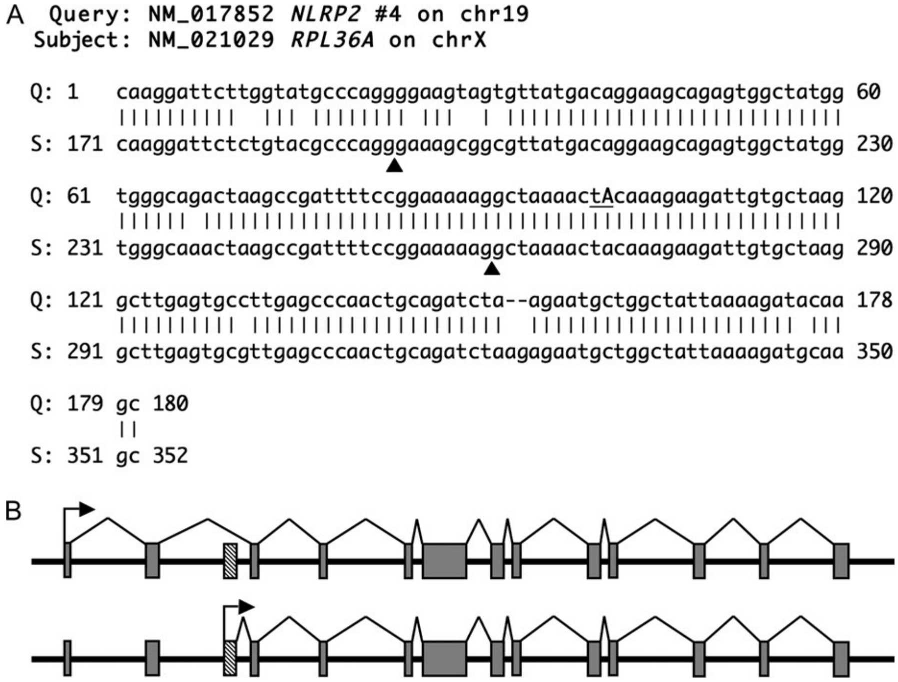
In the present study, we described the retrotransposability of promoters. As seen in some alternative promoters, this phenomenon seems to have contributed to a wide variety of transcripts in mammalian genomes. However, 9 pairs grouped into type IV could not be explained by our hypothesis. Further investigation of such retrotransposed genes and their parental genetic entities would be necessary to determine how genes construct their promoters.

Supplementary Materials

Supplementary figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Dr Riu Yamashita and other members of Dr Nakai Laboratory for helpful discussions. Computation time was provided by the supercomputer system at Human Genome Centre, Institute of Medical Science, University of Tokyo. This work was supported by Institute for Bioinformatics Research and Development, Japan Science and Technology Agency.



```
Query: NM_017852 NLRP2 #4 on chr19
Subject: NM_021029 RPL36A on chrX

Q: 1   caaggattcttggtatgcccaggggaagtagtgttatgacaggaagcagagtggctatgg 60
       ||||||||||  ||| |||||||| |||  | ||||||||||||||||||||||||||||
S: 171 caaggattctctgtacgcccagggaaagcggcgttatgacaggaagcagagtggctatgg 230

Q: 61  tgggcagactaagccgattttccggaaaaaggctaaaactAcaaagaagattgtgctaag 120
       |||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||
S: 231 tgggcaaactaagccgattttccggaaaaaggctaaaactacaaagaagattgtgctaag 290

Q: 121 gcttgagtgccttgagcccaactgcagatcta--agaatgctggctattaaaagatacaa 178
       |||||||||| |||||||||||||||||||||  |||||||||||||||||||||| |||
S: 291 gcttgagtgcgttgagcccaactgcagatctaagagaatgctggctattaaaagatgcaa 350

Q: 179 gc 180
       ||
S: 351 gc 352
```

FIG. 4.—Genesis of an alternative promoter by retrotransposition in the *NLRP2* locus. (*A*) A BLASTN search result shows that a promoter sequence (query) hits an unrelated cDNA sequence (subject). In the subject, exon–intron junctions (indicated by triangles) suggest that the promoter was generated by a retrotransposition of the subject cDNA (*RPL36A*). In 1 other case, its retrotransposition event was validated by polyadenylation. Sequence alignments of all 7 cases are shown in supplementary figure S3 (Supplementary Material online). TSSs, which are written in uppercase, reside at position 101 in the queries. They are either A or G, and 6 of them follow pyrimidine–purine (YR) consensus, as indicated by underlines. The promoter number in a gene designated in DBTSS follows a sharp sign, for example, #4. (*B*) Two splicing patterns of the *NLRP2* gene are drawn roughly to scale. A retrotransposition from the *RPL36A* locus into the second intron of *NLRPN2* gave rise to an alternative promoter and alternative first exon, as highlighted by diagonal lines. In DBTSS, 4 promoters are reported for this gene. Its major transcript is shown in the uppermost schematic representation.

Literature Cited

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science. 297: 1003–1007.

Bird AP. 1986. CpG-rich islands and the function of DNA methylation. Nature. 321:209–213.

Carninci P, Sandelin A, Lenhard B, et al. (41 co-authors). 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 38:626–635.

Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol. 4:R25.

Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. 2008. A code for transcription initiation in mammalian genomes. Genome Res. 18:1–12.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. J Mol Biol. 196:261–282.

Hampsey M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev. 62:465–503.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.

Kimura K, Wakamatsu A, Suzuki Y, et al. (32 co-authors). 2006. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res. 16:55–65.

Kurose K, Hata K, Hattori M, Sakaki Y. 1995. RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. Nucleic Acids Res. 23:3704–3709.

Lewin R. 1983. How mammalian RNA returns to its genome. Science. 219:1052–1054.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:e357.

Maruyama K, Sugano S. 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene. 138:171–174.

Nouvel P. 1994. The mammalian genome shaping activity of reverse transcriptase. Genetica. 93:191–201.

Okamura K, Feuk L, Marquès-Bonet T, Navarro A, Scherer SW. 2006. Frequent appearance of novel protein-coding sequences by frameshift translation. Genomics. 88:690–697.

Parker-Katiraee L, Carson AR, Yamada T, et al. (16 co-authors). 2007. Identification of the imprinted KLF14 transcription factor undergoing human-specific accelerated evolution. PLoS Genet. 3:e65.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet. 8:424–436.

Sellis D, Provata A, Almirantis Y. 2007. Alu and LINE1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws. Mol Biol Evol. 24:2385–2399.

Sharp PA. 1983. Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. Nature. 301:471–472.

Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol. 5:2090–2103.

Suzuki S, Ono R, Narita T, et al. (13 co-authors). 2007. Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. PLoS Genet. 3:e55.

Suzuki Y, Taira H, Tsunoda T, et al. (15 co-authors). 2001. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2:388–393.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Vanin EF. 1985. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 19:253–272.

Ward JJ, Thornton JM. 2007. Evolutionary models for formation of network motifs and modularity in the *Saccharomyces* transcription factor network. PLoS Comput Biol. 3:1993–2002.

Yamashita R, Suzuki Y, Sugano S, Nakai K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. Gene. 350:129–136.

Yamashita R, Suzuki Y, Wakaguri H, Tsuritani K, Nakai K, Sugano S. 2006. DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. Nucleic Acids Res. 34:D86–D89.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 13:335–340.

Yu Z, Morais D, Ivanga M, Harrison PM. 2007. Analysis of the role of retrotransposition in gene evolution in vertebrates. BMC Bioinformatics. 8:308.

Zhang Z, Carriero N, Gerstein M. 2004. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 20:62–67.

Zhang Z, Harrison PM, Liu Y, 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.

Zheng D, Frankish A, Baertsch R, et al. (16 co-authors). 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res. 17:839–851.

Takashi Gojobori, Associate Editor

Accepted March 18, 2008