# Scattered Distribution of Transcription Start Sites Confers Retrotransposability to the Promoters

**Kohji Okamura** kohji@hgc.jp
**Kenta Nakai** knakai@ims.u-tokyo.ac.jp

Human Genome Centre, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato Ward, Tokyo 108-8639, Japan

**Keywords**: retrotransposition, transcription start site (TSS), core promoter, alternative promoter

## 1 Introduction

Retrotransposition is the molecular mechanism of forming processed pseudogenes in a wide range of genomes. In the process of a retrotransposition, a processed mRNA is first reversed transcribed to cDNA by endogenous reverse transcriptase and randomly integrated into certain chromosomal regions. Since transcription initiates downstream from a promoter sequence, the promoter is generally thought not to be transcribed. However, it has recently been shown that retroduplication of mRNA has also generated new functional genes [2] as often seen in segmental duplication. While most of these events have presumably resulted in pseudogenes or been negatively selected, others must have somehow resulted in the acquisition of a new active promoter. Hence, it is likely that promoter analysis of transcriptionally active retrotransposed genes offers a unique opportunity to investigate the evolutionary construction of promoter sequences. In this study, we identified 29 pairs of such genes in the human genome and found a retrotransposability of promoters *per se*.

## 2 Method and Results

Using our database of transcription start sites (DBTSS) [5] based on oligo-capping method [3], we assessed the human genome to identify gene pairs. Basically we employed a strategy in which queries of candidates of retrotransposed genes were searched against a database of the non-redundant gene set to identify their parental genes. Intronless genes are good candidates for such retrotransposed genes. Although genes that have few introns can also be considered, they may have a chimeric structure between the integrated cDNA sequence and a previously existing gene at the locus. Since we intended to study promoter construction, only intronless genes were taken into account. In addition, it is critical to distinguish retrotransposition events from segmental duplication, in which promoters can also be copied along with exonic sequences. We precluded gene families that tend to be highly clustered in certain chromosomal region, *e.g.* olfactory receptor genes.

As a result of a BLASTN search consisting of 631 queries against 23,599 subjects with our conservative criteria, 137 gene pairs were obtained. In addition to being a RefSeq gene, we only selected genes whose expression is supported by at least one oligo-capped clone as transcriptionally active genes. We finally made a list of 29 gene pairs or retrotransposition events that had generated transcriptionally active copies (Table 1). The genomic sequences were aligned and compared. Subsequently, the pairs were readily grouped into one of four categories based on how new copies gained promoters. These included CORE: the core promoter of the source gene was also transcribed, reverse transcribed with its downstream exonic region, and became the new promoter; EXONIC: part of either 5' untranslated region or protein-coding sequence became the new promoter; OTHER: a promoter of an unrelated gene was copied and became the new promoter; and NEW: acquisition of the new promoter could not be explained by sequence similarity.

Since more than half of the retrotransposed genes retain their original promoter sequences, it is likely that a gene can be transposed with its upstream region. This idea was also supported by another comprehensive analysis of alternative promoters, which are often found in the human genome [1]. Existing part of an unrelated exonic sequence downstream of alternative promoters suggests that the origin of some alternative promoters could be explained by retrotransposition of promoters.

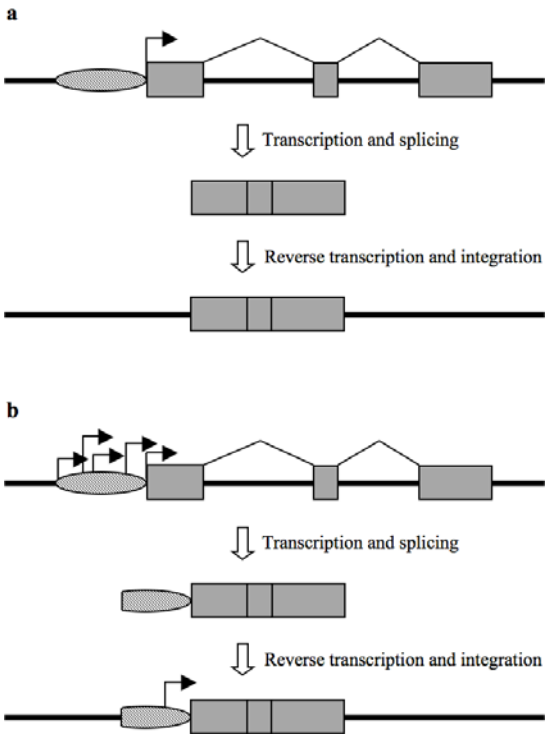


Figure 1: Conventional (a) and proposed (b) views of retrotransposition. Ovals, arrows, and their heights indicate promoters, transcription star sites, and their expression levels, respectively.

Table 1: Pairs of a retrotransposed gene and its source gene.

| No. | Source gene | Retrotransposed gene | Acquisition type |
|---|---|---|---|
| 1 | *HMGB1* | *HMG1L1* | EXONIC |
| 2 | *CTAGE5* | *LOC441294* | CORE |
| 3 | *GUSB* | *LOC441046* | CORE |
| 4 | *H3F3B* | *LOC440093* | CORE |
| 5 | *WDR42A* | *WDR42B* | NEW |
| 6 | *ACTB* | *DKFZp686D0972* | NEW |
| 7 | *WDR21A* | *WDR21B* | EXONIC |
| 8 | *WDR21A* | *WDR21C* | EXONIC |
| 9 | *PDHA1* | *PDHA2* | CORE |
| 10 | *RRAGB* | *RRAGA* | NEW |
| 11 | *MORF4L1* | *MORF4* | CORE |
| 12 | *GLUD1* | *GLUD2* | CORE |
| 13 | *RANBP5* | *RANBP6* | NEW |
| 14 | *GSPT1* | *GSPT2* | CORE |
| 15 | *KLHL13* | *KLHL9* | NEW |
| 16 | *WDR5* | *WDR5B* | NEW |
| 17 | *PAPOLA* | *PAPOLB* | CORE |
| 18 | *PABPC1* | *PABPC3* | CORE |
| 19 | *TKTL1* | *TKTL2* | NEW |
| 20 | *RAB6A* | *RAB6C* | CORE |
| 21 | *LDHAL6A* | *DHHAL6B* | OTHER |
| 22 | *GK* | *GK2* | CORE |
| 23 | *BIRC4* | *BIRC8* | NEW |
| 24 | *RPL10* | *RPL10L* | CORE |
| 25 | *PGK1* | *PGK2* | NEW |
| 26 | *TRAM1* | *TRAM1L1* | CORE |
| 27 | *TAF1* | *TAF1l* | CORE |
| 28 | *NACA* | *NACA2* | CORE |
| 29 | *CTBP2* | *MGC70870* | EXONIC |

## 3 Discussions

Although it has been believed that retrotransposition involves only exonic regions, this study shows promoter sequences are also transcribed and integrated (Figure 1). This can be ascribed to a recent finding that TSSs tend to be interspersed around a core promoter rather than reside at one or a few specific sites [4]. If an upstream TSS is employed, a large part of the promoter region is indeed transcribed. Scattered TSSs are particularly seen in CpG islands. Accordingly, most of the 29 source genes have CpG islands at 5' regions. As seen in some alternative promoters, this retrotransposability of promoters seems to have contributed to a large variety of transcripts in mammalian genomes.

## References

[1] Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., *et al.*, Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, *Genome Res.*, 16:55-65, 2006.

[2] Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H., Emergence of young human genes after a burst of retroposition in primates, *PLoS Biol.*, e357, 2005.

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

[4] Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., *et al.*, Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites, *EMBO Rep.* 2:388-93, 2001.

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