Transcriptional interferences in cis natural antisense transcripts of human and mouse

Naoki Osato, Yoshiyuki Suzuki, Kazuho Ikeo, Takashi Gojobori

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima 411-8540, Japan
Short running head: Transcriptional interferences

Key words: *Cis* natural antisense transcripts; Functional RNA; Non-protein-coding RNA; Regulation of gene expression; SAGE

Corresponding author:

Takashi Gojobori

Center for Information Biology and DNA Data Bank of Japan,
National Institutes of Genetics,
Yata 1111, Mishima, Shizuoka 411-8540, Japan.
Tel: +81-55-981-6847
Fax: +81-55-981-6848
E-mail: tgojobor@genes.nig.ac.jp
ABSTRACT

For a significant fraction of mRNAs, their expression is regulated by other RNAs including cis natural antisense transcripts (cis-NATs) that are complementary mRNAs transcribed from opposite strands of DNA at the same genomic locus. The regulatory mechanism of mRNA expression by cis-NATs is unknown, though a few possible explanations have been proposed. To understand this regulatory mechanism, we conducted a large-scale analysis of the currently available data and examined how the overlapping arrangements of cis-NATs affect their expression level. Here, we show that for both human and mouse the expression level of cis-NATs decreases as the length of the overlapping region increases. In particular, the proportions of the highly expressed cis-NATs in all cis-NATs examined were about 36% and 47% for human and mouse, respectively, when the overlapping region was less than 200 bp. However, both proportions decreased virtually to zero when the overlapping regions are more than 2,000 bp in length. Moreover, the distribution of the expression level of cis-NATs changes according to different types of the overlapping pattern of cis-NATs in the genome. These results are consistent with the transcriptional collision model for the regulatory mechanism of gene expression by cis-NATs.

INTRODUCTION

Biological processes such as development, metabolism, and response to external stimuli are conducted by the cooperative activities of many genes. To understand a biological process it is essential to understand the regulatory network of genes composing the biological process. After genome sequences had been determined, attempts to reveal regulatory networks of genes have started (LEVINE and DAVIDSON 2005, WYRICK and YOUNG 2002, CARNINCI et al. 2005, ENCODE PROJECT CONSORTIUM 2004). Regulation of gene expression can be conducted mainly by proteins such as transcription factors. However, it has been found that about
20%–30% of mammalian transcripts are targets of microRNAs, which bind to complementary mRNAs and inhibit their activation (LEWIS et al. 2005, KREK et al. 2005, STARK et al. 2005, CARTHEW 2006). This suggests that the regulation of gene expression by RNAs is more ubiquitous and important than we thought (MATTICK 2001, MATTICK 2004).

*Cis* natural antisense transcripts (*cis*-NATs) are composed of a pair of mRNAs that are transcribed from the opposite strands of DNA at the same genomic locus. The antisense mRNA regulates the expression level of the sense mRNA in a pair. As a result, they affect the developmental processes such as neural, eye and tooth formation (POTTER and BRANFORD 1998, KORNEEV *et al.* 1999, KORNEEV and O'SHEA 2005, ALFANO *et al.* 2005, COUDERT *et al.* 2005), and various molecular functions such as X-inactivation, genomic imprinting, DNA methylation, RNA editing, and alternative splicing (LEE *et al.* 1999, MOORE *et al.* 1997, TUFARELLI *et al.* 2003, KUMAR and CARMICHAEL 1997, MUNROE and LAZAR 1991).

Although the number of experimentally verified *cis*-NATs was about 40, more than 2,000 *cis*-NATs were predicted in the analyses of the genomic and cDNA sequences in human and mouse (LEHNER *et al.* 2002, SHENDURE and CHURCH 2002, YELIN *et al.* 2003, KIYOSAWA *et al.* 2003, KATAYAMA *et al.* 2005). Recently, an analysis of human oligo microarray showed that as much as 60% of surveyed loci on human chromosome 10 were predicted to encode *cis*-NATs (CHENG *et al.* 2005). Other chromosomes were also expected to encode as many *cis*-NATs. In addition, the presence of *cis*-NATs has been predicted for other eukaryotes as well as prokaryotes (VANHEE-BROSSOLLET and VAQUERO 1998, WAGNER and SIMONS 1994, MAKALOWSKA *et al.* 2005). These observations implied that the regulation of gene expression by *cis*-NATs would occur more frequently than previously considered. Regulation of gene expression by RNAs may be evolutionarily advantageous, because it regulates gene expression quickly and saves energy and time
in synthesizing proteins. Chen et al. (2005b) showed that cis-NATs were encoded in genes with shorter intron sequences than other mRNAs.

Although a large number of cis-NATs have been predicted from various species, the regulatory mechanisms of gene expression by cis-NATs remain unclear. To understand the regulatory mechanisms, it will be crucial to know the features of cis-NATs that are important in the regulation of gene expression. Thus far, three models have been proposed for the regulation of gene expression by cis-NATs (LAVORGNA et al. 2004).

The first model asserts that cis-NATs form a double strand through their complementary sequences, which leads to the inhibition of the function of mRNAs including protein synthesis. In this model, it is expected that cis-NATs are overlapped in at least 6-8 bp regions to form stable double strands of RNA (LAI 2002, LEWIS et al. 2005).

The second model involves epigenetic regulations such as the methylation of promoters and the conversion of the chromosome structure (WUTZ et al. 1997, REIK and WALTER 2001, TUFARELLI et al. 2003). Through unknown mechanisms, the antisense mRNAs methylate promoters of sense mRNAs and inhibit the transcription of sense mRNAs. In addition, the antisense mRNAs convert the chromosomal structures in which cis-NATs are located and regulate the expression of the sense mRNAs. For the model of epigenetic regulations, the features of cis-NATs essential for the regulations are unclear.

The third model is transcriptional collisions. In the transcription of cis-NATs, RNA polymerases bind to the promoters of genes encoding sense and antisense mRNAs and synthesize mRNAs, moving toward the 3’ end of the genes. RNA polymerases clash in the overlapping region and inhibit their transcription (Figure 4). This model was implied from the analyses of the expression levels of cis-NATs in yeast (PETErSON and MYERS 1993, PUIG et al. 1999, PRESCOTT and PROUDFOOT
In the analyses, the expression level of cis-NATs decreased as the length of the overlapping region of the cis-NATs increased. Moreover, the expression level of adjacent transcripts in tandem on the same strand of the yeast genome decreased when the terminator of the upstream transcript was removed. This suggested that RNA polymerases did not stop at the terminator of the upstream transcript and affected the transcription of the downstream transcript. Recently, the collision of *Escherichia coli* RNA polymerases was observed by atomic force microscopy (CRAMPTON *et al.* 2006). This observation showed that RNA polymerases do not pass each other or displace one another, but instead stall against each other.

Here, we report the effects of length and pattern of the overlapping regions of cis-NATs for human and mouse on their expression level. Moreover, human and mouse adjacent transcripts, in a particular position, affect their expression level. These results are consistent with the transcriptional collision model, implying that the regulation of the expression of cis-NATs by transcriptional collisions is common among species.

**MATERIALS AND METHODS**

**Analysis of the expression level of human cis-NATs:** To predict cis-NATs from human cDNA sequences, we collected a total of 46,675 human cDNA sequences, which consisted of 39,530 human Ensembl cDNA sequences (February 2006) (HUBBARD *et al.* 2005), 6,501 human Ensembl non-protein-coding sequences (February 2006), and 644 non-protein-coding sequences in RNAdb (PANG *et al.* 2005) that were mapped to the human genome by BLAT software (KENT 2002). We predicted 8,964 cis-NATs, which had at least 1-bp-long overlapping region, based on their genomic location using in-house Perl scripts. Redundant cis-NATs were merged into the same group when cis-NATs overlapped in at least 1-bp-long region in the genome. The number of the groups of cis-NATs was 2,496. To examine the expression levels of human cis-NATs,
we employed 15.5 million NlaIII human SAGE tags (November 2005) of all tissues in
the NCBI SAGEmap database (LASH et al. 2000). Human SAGE tags were searched
against a total of 46,675 human cDNA sequences according to the protocol for SAGE
(VELCULESCU et al. 1995) using in-house Perl scripts. When a SAGE tag matched
to more than one transcript, these transcripts were removed from further analyses. As
a result, among the 46,675 human cDNA sequences, 28,009 had unique SAGE tag
assignments, and among the 2,496 groups of human cis-NATs, 728 groups had unique
SAGE tag assignments to all transcripts in each group. Some SAGE tags are supposed
to be assigned to overlapping regions of cis-NATs. Although cDNA microarrays
cannot distinguish the expression of mRNAs encoded in plus and minus strands of the
genome in the same locus, SAGE tags are strand-specific. Therefore, even though
SAGE tags are produced from the overlapping regions of cis-NATs, the expression
levels of sense and antisense mRNAs of cis-NATs can be measured separately.

To examine the expression levels of the human cis-NATs, we compared the
expression level of human cis-NATs with that of other human transcripts (i.e. human
transcripts excluding cis-NATs, pseudogenes and non-protein-coding mRNAs). We
calculated the ratio of the expression level of a human cis-NAT to that of the other
human transcripts. However, the expression levels of transcripts are known to be
affected by the overall length of the transcripts (CASTILLO- DAVIS et al. 2002).
Thus, we compensated the expression level of cis-NATs according to their overall
length. The expression level of a human cis-NAT was compared with that of the other
human transcripts with almost the same overall length of the cis-NAT. We removed
human cis-NATs, pseudogenes and non-protein-coding mRNAs from human cDNA
sequences, and selected 2,000 human transcripts of which overall length in the genome
was close to the overall length of a human cis-NAT in the genome. We calculated the
median of the expression levels of the selected 2,000 human transcripts (Supplemental
Figure 1A). The median of the expression levels of the selected human transcripts was
defined as a ratio of 1.0, and then the ratio of the expression level of a cis-NAT was calculated as follows:

\[ R_{\text{cis-NAT}} = \frac{T_{\text{cis-NAT}}}{T_{(\text{all} - \text{cis-NAT} - \text{pseudo - non-coding})}} \]

\( R_{\text{cis-NAT}} \): Ratio of the expression level of a human cis-NAT to that of other human protein-coding transcripts

\( T_{\text{cis-NAT}} \): Expression level of a human cis-NAT.

\( T_{(\text{all} - \text{cis-NAT} - \text{pseudo - non-coding})} \): Median of the expression level of other human transcripts (i.e. human transcripts excluding cis-NATs, pseudogenes and non-protein-coding mRNAs) with almost the same overall length of the cis-NAT.

Sense mRNAs of cis-NATs are not located only in the plus or minus strand of the genome, but in both strands of the genome. Some of the sense mRNAs encode proteins and others encode non-protein-coding mRNAs that may have some biological function such as the regulation of mRNA translation and stability (MATTICK and MAKUNIN 2006). The antisense mRNA in each group of cis-NATs is known to decrease the expression level of the sense mRNA in the group and inhibit the activation of the sense mRNA (WAGNER and SIMONS 1994, KUMAR and CARMICHAEL 1998, VANHEE-BROSSOLLET and VAQUERO 1998). Therefore, we recognized the cis-NAT with the lowest expression level in each group as the sense mRNA in the group. In this study, we used the expression level of the sense mRNA in each group to examine the expression levels of cis-NATs according to the overlapping arrangements in the genome. However, there is an assumption that under some regulatory mechanisms of cis-NATs such as RNA masking and double-stranded RNA-dependent mechanism, the expression level of a sense mRNA may not be the lowest in the group of cis-NATs and an inverse correlation of the expression levels may not be found between a sense and an antisense mRNA (LAPIDOT and PILPEL 2006, CHEN et al. 2005a, KATAYAMA et al. 2005). Therefore, we also examined the expression level of cis-NATs randomly selected (i.e. selected by
non-expression-level-dependent manner) (See SUPPLEMENTAL MATERIAL).

We examined the expression levels of \textit{cis-}NATs as the overlapping regions increased in length. When \textit{cis-}NATs included more than two transcripts (\textit{i.e.}, sense and antisense transcripts on both strands of the genome), the length of the overlapping region was defined as the distance between the furthest upstream and the furthest downstream genomic locations of the overlapping regions of \textit{cis-}NATs in a group.

\textbf{Analysis of the expression level of mouse \textit{cis-}NATs:} To predict the mouse \textit{cis-}NATs, we collected a total of 35,486 mouse cDNA sequences, which consisted of 33,252 mouse Ensembl cDNA sequences (April 2006), 1,752 mouse Ensembl non-protein-coding sequences (April 2006), and 482 non-protein-coding sequences in RNAdb that were mapped to the mouse genome using BLAT. We predicted 5,491 \textit{cis-}NATs from the mouse cDNA sequences and redundant \textit{cis-}NATs were merged into the same group. There were 1,868 groups of mouse \textit{cis-}NATs. We examined the expression levels of mouse \textit{cis-}NATs using 3.6 million \textit{NlaIII} mouse SAGE tags of all tissues in the NCBI SAGEmap database (November 2005) to compare them with the expression levels of human \textit{cis-}NATs. Among the 35,486 mouse cDNA sequences, 21,982 had unique SAGE tag assignments, and among the 1,868 groups of mouse \textit{cis-}NATs, 704 groups had unique SAGE tag assignments to all transcripts including alternative forms in each group. We examined the expression levels of mouse \textit{cis-}NATs in the same way as human.

\textbf{Comparison of \textit{cis-}NATs of human and mouse at the nucleotide level:} To find \textit{cis-}NATs conserved between human and mouse, we compared 46,675 human cDNA sequences with 35,486 mouse cDNA sequences and vice versa using all-against-all FASTA (PEARSON and LIPMAN 1988) procedure. An expected (\textit{E}) value cut-off of $1.0 \times 10^{-20}$ was used. We selected the human and mouse cDNA sequence matched reciprocally with an \textit{E}-value less than the square root of the lowest \textit{E}-value as candidates of orthologs. When human \textit{cis-}NATs in both strands of the genome are
orthologous to mouse \textit{cis}-NATs in both strands of the genome, we recognized the \textit{cis}-NATs as conserved between human and mouse.

**RESULTS**

**Length of the overlapping region of human \textit{cis}-NATs affects their expression level:** To predict \textit{cis}-NATs from human cDNA sequences, we searched 46,675 human cDNA sequences. A total of 8,964 \textit{cis}-NATs were predicted and were clustered into 2,496 groups, each of which consisted of sense and antisense transcripts as well as their alternative forms. To examine the expression levels of human cDNA sequences, about 15.5 million \textit{Nla}III serial analysis of gene expression (SAGE) tags were collected from all the human tissues available in the NCBI SAGEmap database (November, 2005) (LASH \textit{et al.} 2000) and were compared to 46,675 human cDNA sequences. Among the 8,964 (2,496 groups) \textit{cis}-NATs, 2,038 (728 groups) had unique SAGE tag assignments to all transcripts in each group.

To examine whether the length of overlapping regions in \textit{cis}-NATs affects their expression level, we investigated the relationship between the expression levels of \textit{cis}-NATs and the length of the overlapping region in the genome. It should be noted that Castillo-Davis \textit{et al} (2002) found that highly expressed transcripts tended to have short introns, implying that the short \textit{cis}-NATs may be expressed at a higher level than the long \textit{cis}-NATs. To eliminate the effect of the overall length of a transcript on its expression level, we selected non-overlapping transcripts whose overall lengths in the genome were almost the same as that of a \textit{cis}-NAT in the genome, and then compared the expression level of the \textit{cis}-NATs with that of the selected transcripts (See MATERIALS AND METHODS).

In Figure 1, the Y-axis represents the ratio of the expression level of the human \textit{cis}-NATs to that of the selected human transcripts. The proportion of highly expressed \textit{cis}-NATs (ratio greater than 1.0) in all \textit{cis}-NATs examined was 36% when the
overlapping region was between 1 bp and 200 bp. However, the proportion decreased virtually to zero when the overlapping regions were more than 2,000 bp long (Chi-square $P < 10^{-15}$). This result suggests that the expression level of cis-NATs decreases as the length of the overlapping region increases. In addition, the expression levels of cis-NATs may be influenced by other factors such as alternative transcripts of cis-NATs, the difference of the expression levels among SAGE tag libraries, GC content bias of SAGE tags (MARGULIES et al. 2001), the experimental methods for the analysis of gene expression, the ways of selecting sense mRNAs, the criteria for the length of the overlapping regions and a set of human cDNA sequences used in this analysis. However, these factors did not change the overall distribution of the expression levels significantly (See SUPPLEMENTAL MATERIAL).

**Overlapping pattern of human cis-NATs affects their expression level:**
Cis-NATs are classified into three types based on the overlapping patterns in the genome: head-to-head, tail-to-tail, and full overlap (Figure 2). “Full overlap” describes cis-NATs where the sense mRNA entirely overlaps within the antisense mRNA. The numbers of head-to-head, tail-to-tail, and full-overlap types of human cis-NATs were 254, 476, and 1,766 groups of which 126, 230, and 356 groups had unique SAGE tag assignments to all transcripts in each group, respectively. To examine whether the overlapping pattern of cis-NATs affects their expression level, we analyzed the expression level of cis-NATs of human according to the overlapping patterns in the genome. Figures 3A, 3B, and 3C show the expression levels of human cis-NATs in the head-to-head, tail-to-tail and full-overlap manners, respectively. The highly expressed cis-NATs decreased in quantity as the overlapping region increased in length for all types of cis-NATs. However, highly expressed cis-NATs in head-to-head and full-overlap manners decreased in quantity more than those in a tail-to-tail manner. When the length of the overlapping region was less than 600 bp, 26.7% of cis-NATs in a head-to-head manner showed high expression (ratio greater than 1.0) and 43.4% of
cis-NATs in a tail-to-tail manner showed high expression. The proportion of highly expressed cis-NATs in a head-to-head manner (26.7\%) was 1.6 times smaller than that in a tail-to-tail manner (43.4\%) (Mann-Whitney U-test \( P < 10^{-2} \)). Similarly, when the length of the overlapping region was less than 600 bp, 24.5\% of cis-NATs in a full-overlap manner showed high expression. The proportion of highly expressed cis-NATs in a full-overlap manner (24.5\%) was 1.7 times smaller than that in a tail-to-tail manner (43.4\%) (Mann-Whitney U-test \( P < 10^{-3} \)). Among the 356 cis-NATs in a full-overlap manner, 314 were cis-NATs where a sense transcript overlapped only in the intron regions of the antisense transcript in the genome. The mRNAs that overlapped in the intron regions showed the same feature of expression.

As many as 1,450 human transcripts were found to be located within a distance of less than 1 kbp in the genome (ADACHI and LIEBER 2002, KOYANAGI et al. 2005). To examine whether the expression levels of nearby transcripts decreased, we investigated the expression levels of human nearby transcripts. Figure 3D shows the expression levels of nearby transcripts where the 5′ end of a transcript is near the 5′ end of another transcript in the genome. Here, we call them “nearby transcripts in a head-to-head manner.” (Figure 2) When the distance between the 5′ ends of transcripts was less than approximately 50 bp, highly expressed transcripts (a ratio greater than 1.0) were not observed (Chi-square \( P < 10^{-7} \)).

Figure 3E shows the expression levels of nearby transcripts where the 3′ end of a transcript is near the 3′ end of another transcript in the genome. Here, we call them “nearby transcripts in a tail-to-tail manner.” Contrary to nearby transcripts in a head-to-head manner, the expression levels of nearby transcripts in a tail-to-tail manner did not change, regardless of the distance of the nearby transcripts in the genome (Chi-square \( P = 0.7 \)).

There is a possibility that the length of some human transcripts registered in a database such as the Ensembl database may be shorter than natural transcripts.
Nearby transcripts found in the Ensembl database may, in fact, overlap in the genome, such that the expression levels of such artificial nearby transcripts seemed to decrease. However, almost all nearby transcripts in a head-to-head manner were expressed at a low level when the distance of the transcripts was less than approximately 50 bp. Therefore, the decrease in the expression level of nearby transcripts will be a natural phenomenon.

Overlapping arrangements of mouse cis-NATs affect their expression levels:

Cis-NATs have been predicted for various species (Wagner and Simons 1994, Vanhee-Brossollet and Vaquero 1998, Wagner and Flardh 2002, Makalowska et al. 2005). If the regulatory mechanisms of cis-NATs in gene expression are conserved among species, cis-NATs of another species are expected to show similar effects on their expression levels. To evaluate whether the relationship between the overlapping arrangements and the expression levels of cis-NATs is conserved among species, we examined the expression levels of mouse cis-NATs. Almost the same number of mouse cis-NATs (1,771 groups) as that of human was found in mouse cDNA sequences (Yelin et al. 2003, Kiyosawa et al. 2003). Mouse cis-NATs were compared to 3.6 million NlaIII mouse SAGE tags in NCBI SAGEmap database. Although the number of mouse SAGE tags and the number of mouse cis-NATs (705 groups) assigned to unique SAGE tags were smaller than human, the distribution of the expression levels of mouse cis-NATs showed the same features as human (Supplemental Figure 2A): highly expressed (ratio greater than 1.0) cis-NATs decreased in quantity when the overlapping regions in the genome increased in length. The proportion of highly expressed cis-NATs in all cis-NATs examined was 47% when the overlapping region was between 1 bp and 200 bp, and the proportion decreased virtually to zero when the overlapping regions were more than 2,000 bp long (Chi-square $P < 10^{-15}$).

For overlapping patterns in the genome, the distribution of the expression levels in
mouse changed in the same way as in human (Mann-Whitney U-test $P = 0.52$ between human and mouse cis-NATs in a head-to-head manner, $P = 0.92$ between those in a tail-to-tail manner, $P = 0.17$ between those in a full-overlap manner) (Supplemental Figures 2B, 2C, and 2D). When the length of the overlapping region was less than 600 bp, 27.6% of cis-NATs in a head-to-head manner showed high expression (ratio greater than 1.0) and 44.1% of cis-NATs in a tail-to-tail manner showed high expression. The proportion of highly expressed cis-NATs in a head-to-head manner (27.6%) was 1.6 times smaller than that in a tail-to-tail manner (44.1%) (Mann-Whitney U-test $P < 10^{-2}$).

Similarly, 25.9% of cis-NATs in a full-overlap manner showed high expression (ratio greater than 1.0) and the proportion of highly expressed cis-NATs in a full-overlap manner (25.9%) was 1.7 times smaller than that in a tail-to-tail manner (44.1%) (Mann-Whitney U-test $P < 0.05$). With nearby transcripts, when the distance between nearby transcripts in a head-to-head manner was less than approximately 50 bp, highly expressed transcripts were not observed as found in human (Chi-square $P < 10^{-5}$) (Supplemental Figure 2E). The expression levels of nearby transcripts in a tail-to-tail manner did not change, regardless of the distance of the nearby transcripts in the genome (Chi-square $P = 0.9$) (Supplemental Figure 2F). These results suggest that the expression levels of mouse cis-NATs are affected by the overlapping arrangements in the genome in the same way as human. This implies that the regulatory mechanisms of cis-NATs in gene expression are conserved between human and mouse.

However, there was a possibility that cis-NATs showed a similar distribution of the expression level between human and mouse because most of human and mouse cis-NATs were orthologous (LIAO and ZHANG 2006). To address this possibility, we compared the distribution of the expression levels of cis-NATs that are not conserved between human and mouse. First, we compared human and mouse cDNA sequences by using FASTA (PEARSON and LIPMAN 1988), and found that only 329 groups (11.9%) of human cis-NATs were conserved in mouse cis-NATs, although 34,670
(76.0%) of human cDNA sequences were conserved in mouse. Human and mouse cis-NATs were supposed to be highly divergent in terms of cDNA sequences (VEERAMACHANENI et al. 2004, MAKALOWSKA et al. 2005). We removed the 329 groups of cis-NATs from 2,765 groups of human and 1,704 groups of mouse cis-NATs, which left 710 groups of human and 605 groups of mouse cis-NATs with SAGE tags assigned to all transcripts in each group. We examined the expression level of the human and mouse cis-NATs according to the length and pattern of the overlapping region in the genome. They showed almost the same distribution of the expression level as those including cis-NATs conserved between human and mouse (Mann-Whitney U-test $P = 0.22$ and $P = 0.88$ for human and mouse respectively). These results suggested that the similarity of the distribution of the expression level of human and mouse cis-NATs was not due to the conservation of the cDNA sequences of the cis-NATs.

DISCUSSION

We found that the expression level of cis-NATs changed according to the overlapping arrangements of cis-NATs in the human and mouse genomes. The expression level of cis-NATs decreased when the overlapping regions increased in length. Moreover, the overlapping pattern of cis-NATs affects their expression level. Nearby transcripts in a particular position decreased their expression levels.

Here, we examined the expression level of cis-NATs using SAGE tags and oligonucleotide arrays in public databases. We obtained the same distribution of the expression level of cis-NATs at least in the same tissue or cell such as fetal brain, embryonic stem cell, and liver (SUPPLEMENTAL MATERIAL and Supplemental figures 3 and 4). However, all SAGE libraries and the expression data of oligonucleotide arrays were produced from cells and tissues of human and mouse, not from a single cell. In addition, some cis-NATs may be expressed only in some
developmental stages or at a specific time. Therefore, some cis-NATs may not be expressed concurrently in the same single cell.

Thus far, three models have been proposed for the regulation of gene expression by cis-NATs. Forming double strands of cis-NATs requires a minimum 6-8 bp overlapping region of cis-NATs. We found that the expression level of human and mouse cis-NATs decreased consecutively as the length of the overlapping region increased. However, currently there is no report that this result is brought about by this model. In addition, this model does not intend to explain the change of the expression levels of nearby transcripts. Epigenetic regulations are also considered to be involved in the regulation of cis-NATs in gene expression. However, currently there is no report that by this model the expression level of human and mouse cis-NATs decrease consecutively as the length of the overlapping region increase.

Our results are consistent with the transcriptional collision model that has been proposed following the analyses of adjacent transcripts and cis-NATs in yeast (PUIG et al. 1999, PRESCOTT and PROUDFOOT 2002) and the observation of the collision of RNA polymerases by atomic force microscopy (CRAMPTON et al. 2006). RNA polymerases bind to the upstream regions of genes and synthesize mRNAs, moving toward the 3′ ends of the genes. When opposite genomic strands in the same locus encode complementary mRNAs like cis-NATs, an RNA polymerase bound on a strand of the genome collides with the RNA polymerase bound on the opposite strand during the transcription of both strands of mRNAs (Figure 4). This leads to the inhibition of transcription. From this model, the frequency of the collisions of RNA polymerases is expected to increase when the overlapping regions increase in length. Moreover, overlapping patterns in the genome would affect the frequency of the collisions of RNA polymerases. In the case of cis-NATs in a head-to-head manner, the 5′ ends of mRNAs are the start position for the transcription of mRNAs. Overlapping at the 5′ end would inhibit the initiation of transcription and decrease the expression level of the cis-NATs.
Contrary to cis-NATs in a head-to-head manner, overlapping at the 3′ end would not decrease the expression level significantly when the overlapping region is short (Figure 3B). In the case of cis-NATs in a full-overlap manner, both the 5′ and 3′ ends of a transcript are overlapped. This would decrease the expression level, even when the length of overlapping regions is short (Figure 3C). For nearby transcripts, highly expressed nearby transcripts in a head-to-head manner decreased in quantity when the distance between the nearby transcripts was less than 50 bp (Figure 3D). However, the level of highly expressed nearby transcripts in a tail-to-tail manner did not decrease (Figure 3E). These results would be because of whether or not the start or end positions of transcripts were close to each other. In the initiation of transcription, RNA polymerases bind to the start position of transcription of mRNAs and cover the region between 55 bp upstream and 20 bp downstream (-55 to 20) of the start position (MURAKAMI et al. 2002, KORZHEVA et al. 2000, LEE and YOUNG 2000). Therefore, it suggests that nearby transcripts in a head-to-head manner inhibited the binding of RNA polymerases to the upstream regions of the transcripts, when the distance between the nearby transcripts in a head-to-head manner was less than 50 bp. In addition, the model of transcriptional collisions for cis-NATs may explain an observation that experiments of northern hybridization showed smear bands of mRNAs at the genomic regions where cis-NATs were located (KIYOSAWA et al. 2005). This implies that various lengths of single stranded mRNAs may be produced by the inhibition of the transcription and unusual movements of RNA polymerases.

Among the 2,462 groups of human cis-NATs, 874 groups did not include alternative forms of sense and antisense mRNAs, and among the 874 groups, 542 (62%) groups consisted of cis-NATs where a sense mRNA overlapped only the intron regions of the gene encoding the antisense mRNA in the genome. As shown in Figures 1 and 3, the expression level of cis-NATs overlapping the intron regions also decreased as the length of the overlapping region increased. As for the regulatory mechanisms of gene
expression by cis-NATs overlapping the intron regions, double-stranded RNA-dependent mechanism would be difficult to explain the decrease of the expression level of the cis-NATs, because they cannot form double strands of mRNAs after transcription and pre-mRNA splicing of mRNAs. Although double strands of mRNAs may be formed before pre-mRNA splicing after transcription, it is unclear whether it occurs. In the meantime, transcriptional collisions reasonably explain that cis-NATs overlapping the intron regions affected the expression of the cis-NATs.

Our observations are consistent with the transcriptional collision model. However, this does not mean that they will exclude other regulatory mechanisms from the regulation of cis-NATs in gene expression. In addition to the regulation by general transcription factors in gene expression, cis-NATs will employ several regulatory mechanisms including transcriptional collisions (LAVORGNA et al. 2004). Our findings would be useful for the examination and understanding of the regulatory mechanisms of cis-NATs in gene expression and furthermore help elucidate the regulatory network of genes and their evolution.

We are grateful to members of the Laboratory for DNA Data Analysis at the National Institute of Genetics for discussion and comments on the manuscript. This work was supported by (1) a research grant from the Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST); (2) a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

LITERATURE CITED


ALFANO, G., C. VITIELLO, C. CACCIOPPOLI, T. CARAMICO, A. CAROLA et al.,


HUBBARD, T., D. ANDREWS, M. CACCAMO, G. CAMERON, Y. CHEN et al., 2005

KATAYAMA, S., Y. TOMARU, T. KASUKAWA, K. WAKI, M. NAKANISHI et al.,
2005 Antisense transcription in the mammalian transcriptome. Science 309:
1564-1566.


KIYOSAWA, H., N. MISE, S. IWASE, Y. HAYASHIZAKI and K. ABE, 2005
Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to
be poly(A) negative and nuclear localized. Genome Res 15: 463-474.

KIYOSAWA, H., I. YAMANAKA, N. OSATO, S. KONDO and Y. HAYASHIZAKI,
2003 Antisense transcripts with FANTOM2 clone set and their implications for

KORNEEV, S., and M. O'SHEA, 2005 Natural antisense RNAs in the nervous system.
Rev Neurosci 16: 213-222.

KORNEEV, S. A., J. H. PARK and M. O'SHEA, 1999 Neuronal expression of neural
nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA

KORZHEVA, N., A. MUSTAEV, M. KOZLOV, A. MALHOTRA, V. NIKIFOROV et

KOYANAGI, K. O., M. HAGIWARA, T. ITOH, T. GOJOBORI and T. IMANISHI,
2005 Comparative genomics of bidirectional gene pairs and its implications for

KREK, A., D. GRUN, M. N. POY, R. WOLF, L. ROSENBERG et al., 2005

KUMAR, M., and G. G. CARMICHAEL, 1997 Nuclear antisense RNA induces
extensive adenosine modifications and nuclear retention of target transcripts.


MAKALOWSKA, I., C. F. LIN and W. MAKALOWSKI, 2005 Overlapping genes in


POTTER, S. S., and W. W. BRANFORD, 1998 Evolutionary conservation and


FIGURE LEGENDS

FIGURE 1. Distribution of the expression level of human cis-NATs as the overlapping region in the genome increased in length

X-axis shows the length of the overlapping exon and intron regions of human cis-NATs in the genome. Y-axis shows the ratio of the expression level of a human cis-NAT to the median of the expression levels of human transcripts of which overall lengths in the genome are close to the overall length of the human cis-NAT in the genome. The median of the expression levels of human transcripts is defined as a ratio of 1.0. Because the expression levels of human transcripts are affected by their overall length, we compensated for the expression level of a cis-NAT according to its overall length (See MATERIALS AND METHODS).

FIGURE 2. Classification of cis-NATs and nearby transcripts based on their relative positions in the genome

Cis-NATs are classified based on their relative positions in the genome: (a)
cis-NATs in a head-to-head manner (5′ end to 5′ end), (b) those in a tail-to-tail manner (3′ end to 3′ end), and (c) those in a full-overlap manner. Full overlap describes the cis-NATs where a transcript on a strand of the genome is overlapped by the entire length of the other transcript on the opposite strand of the genome. (d) nearby transcripts in a head-to-head manner where the 5′ end of a transcript is near the 5′ end of another transcript in the genome. (e) nearby transcripts in a tail-to-tail manner where the 3′ end of a transcript is near the 3′ end of another transcript in the genome.

FIGURE 3. Distribution of the expression level of human cis-NATs according to overlapping patterns in the genome

X-axis shows the length of the overlapping exon and intron regions of human cis-NATs in the genome. Y-axis shows the ratio of the expression level of a cis-NAT to the median of the expression levels of human transcripts whose overall lengths in the genome are close to those of the human cis-NAT in the genome. The median of the expression levels of human transcripts is defined as a ratio of 1.0. Because the expression levels of human transcripts are affected by their overall length, we compensated the expression level of a cis-NAT according to its overall length (See MATERIALS AND METHODS). (A) cis-NATs in a head-to-head manner, (B) those in a tail-to-tail manner, (C) those in a full-overlap manner, (D) nearby transcripts in a head-to-head manner, and (E) nearby transcripts in a tail-to-tail manner.

FIGURE 4. Transcriptional collision model

When cis-NATs are transcribed by RNA polymerases, RNA polymerases bind to the upstream region of a gene encoding a sense mRNA and synthesize the complementary mRNA, moving to the 3′ end of the gene. Similarly, RNA polymerases that are bound to the upstream region of a gene encoding an antisense mRNA move to the 3′ end of the gene. Then, RNA polymerases collide with each
other in the overlapping region of the genes, thereby inhibiting the transcription.