Published online 24 July 2008 *Nucleic Acids Research, 2008, Vol. 36, No. 15* ***4833–4844***
doi:10.1093/nar/gkn470

Trans-natural antisense transcripts including noncoding RNAs in 10 species: implications for expression regulation

Jiong-Tang Li[1], Yong Zhang[1], Lei Kong[1], Qing-Rong Liu[2] and Liping Wei[1,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing, 100871, P.R. China and [2]Department of Health and Human Services (DHHS), Molecular Neurobiology Branch, National Institute on Drug Abuse-Intramural Research Program (NIDA-IRP), NIH, Box 5180, Baltimore, MD 21224, USA

Received May 29, 2008; Revised and Accepted July 4, 2008

ABSTRACT

Natural antisense transcripts are at least partially complementary to their sense transcripts. *Cis*-Sense/Antisense pairs (*cis*-SAs) have been extensively characterized and known to play diverse regulatory roles, whereas *trans*-Sense/Antisense pairs (*trans*-SAs) in animals are poorly studied. We identified long *trans*-SAs in human and nine other animals, using ESTs to increase coverage significantly over previous studies. The percentage of transcriptional units (TUs) involved in *trans*-SAs among all TUs was as high as 4.13%. Particularly 2896 human TUs (or 2.89% of all human TUs) were involved in 3327 *trans*-SAs. Sequence complementarities over multiple segments with predicted RNA hybridization indicated that some *trans*-SAs might have sophisticated RNA–RNA pairing patterns. One-fourth of human *trans*-SAs involved noncoding TUs, suggesting that many noncoding RNAs may function by a *trans*-acting antisense mechanism. TUs in *trans*-SAs were statistically significantly enriched in nucleic acid binding, ion/protein binding and transport and signal transduction functions and pathways; a significant number of human *trans*-SAs showed concordant or reciprocal expression pattern; a significant number of human *trans*-SAs were conserved in mouse. This evidence suggests important regulatory functions of *trans*-SAs. In 30 cases, *trans*-SAs were related to *cis*-SAs through paralogues, suggesting a possible mechanism for the origin of *trans*-SAs. All *trans*-SAs are available at http://trans.cbi.pku.edu.cn/.

INTRODUCTION

Natural antisense transcripts are at least partially complementary to their endogenous sense RNAs. Transcripts in a *cis*-sense/antisense (*cis*-SA) pair are transcribed from the opposite strands at the same genomic locus and thus display perfect RNA–RNA sequence complementarity, whereas transcripts from a *trans*-sense/antisense (*trans*-SA) pair are transcribed from different genomic loci and may have imperfect sequence complementarity (1). *Cis*-antisense transcripts have been extensively studied both computationally (2–5) and experimentally, and found to play a variety of regulatory roles including involvement in imprinting (6), alternative splicing (7) and transcriptional interference (8). Studies of *trans*-antisense RNAs (also termed *trans*-encoded RNAs or *trans*-acting RNAs) have mainly focused on small RNAs such as small interfering RNAs (siRNAs) and microRNAs (miRNAs) which function in a *trans* base-pairing mechanism with their targets and play important regulatory roles such as in mRNA degradation and translational repression (9).

There is evidence suggesting that long *trans*-antisense RNAs may also perform versatile regulatory functions (10). In bacteria, there are several examples of functional long *trans*- antisense transcripts (11–13). In eukaryotes, to date in only three cases has the activity of long *trans*-antisense been experimentally characterized: *Lymnaea* anti-*NOS* prevents the translation of the nNOS protein from its encoding mRNA, resulting in substantial suppression of nNOS enzyme activity (14,15); variant δ of mouse *Msh4* forms double-stranded RNA (dsRNA) with *Hspa5*, possibly inducing *Hspa5* RNA degradation and resulting in cell death (16); *MBP*'s antisense RNA reduces the expression of the *MBP* gene, either by inhibition of *MBP* transport from the nucleus or by selective

*To whom correspondence should be addressed. Tel: +86 10 6275 5206; Fax: +86 10 6276 4970; Email: weilp@mail.cbi.pku.edu.cn

degradation of the *MBP*–antisense-*MBP* RNA duplex in myelin-deficient mutant mice (17).

Three groups have used computational methods to identify long *trans*-SAs from mRNAs and full-length cDNAs in human and *Arabidopsis*: Lehner *et al*. identified 92 putative *trans*-SA transcript pairs in human mRNA sequences (18); Li *et al*. identified 161 *trans*-SA transcript pairs in human RefSeq mRNAs (10) and Wang *et al*. identified 1320 *trans*-SA transcript pairs in *Arabidopsis* from full-length cDNA sequences (19). Studies on long *trans*-SAs are far from complete, especially given that these analyses have not utilized the more extensively available EST sequences. Therefore, the prevalence of *trans*-SAs together with other features has not been elucidated. Moreover, there have been no studies in animal species other than human.

A number of long noncoding RNAs were found to function as *cis*-antisense transcripts to regulate their overlapping sense genes (20,21). It was recently reported that around 98% of all transcripts in human were likely to be noncoding RNAs (22). The functions of most long noncoding RNAs remain unknown. It is important to explore whether they may function through *trans*-acting RNA–RNA interaction.

The goal of our present work was to: (i) estimate the prevalence of *trans*-SAs in different animals; (ii) classify the pairing and overlapping patterns of *trans*-SAs that may correspond to different regulatory mechanisms; (iii) investigate how many long noncoding RNAs might function by means of *trans*- RNA–RNA interaction in RNA-mediated gene regulation; (iv) determine in what biological events TUs in *trans*-SAs were most often involved; (v) study whether some, if not all, of *trans*-SAs may have correlated expression profiles; (vi) find *trans*-SAs that were conserved between different species and (vii) investigate how *trans*-SA originated by analyzing the relationship between *trans*-SAs and *cis*-SAs.

MATERIALS AND METHODS

Identification of *trans*-SAs

We identified *trans*-SAs in human and nine other animal species: *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Bos taurus* (cattle), *Canis lupus familiaris* (dog), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm) and *Ciona intestinalis* (sea squirt). These species were chosen because they allowed a range of cross-species comparisons with human and had relatively complete genomic sequences and abundant ESTs.

Lessons from previous studies of *cis*-SAs indicated that more SAs could be identified using ESTs in addition to mRNAs, rather than mRNAs alone (4,5). Using ESTs, however, is computationally more challenging because there are an-order-of-magnitude more ESTs than mRNAs and ESTs tend to have lower quality than mRNAs. Identifying *trans*-SAs is also more challenging than identifying *cis*-SAs because *trans*-SAs are transcribed from separate genomic loci. Previous work often clustered ESTs/mRNAs based on genomic overlap. This, however, could accidentally cluster neighboring genes into one. An additional confounding factor is that many genes in higher organisms are alternatively spliced. All these aspects indicated the need for a new and accurate identification pipeline of *trans*-SAs.

Our *trans*-SA identification pipeline was summarized in Supplementary Figure S1 and described below. We first downloaded from NATsDB (4) all EST/mRNA sequences which were (i) mapped unambiguously to the genome and (ii) had corrected orientations. For quality control, NATsDB applied stringent criteria to filter ESTs/mRNAs by aligning ESTs/mRNAs to genome, filtering low-quality mapping, selecting best unique mapping and inferring the correct orientation of ESTs/mRNAs (4,5). By comparing with human full-length cDNAs in the H-Invitational database (23), we found that our procedure inferred the correct orientation for 186 219 (or 99.5%) out of the 187 156 transcripts in the database, achieving an accuracy similar to other reports (2,24).

In the next step, our *trans*-SA identification pipeline refined exon/intron boundaries. BLAT cannot always ensure precise genome mapping for low-quality ESTs, resulting in some ambiguous exon boundaries (25). Our pipeline refined exon boundaries by integrating multiple evidence including standard splicing sites (GT-AG), sequence type (mRNA or EST), boundary frequency and chromosomal mapping quality: (i) If one boundary encoded a standard splicing site whereas another did not, the former was selected as the correct boundary. (ii) If the first criterion did not apply, the boundary inferred from mRNAs was used in preference to that from ESTs. Our pipeline also allowed users to manually rank their input sequences and boundaries according to their own estimate of quality. (iii) If the second criterion also failed, for instance, if all input sequences were ESTs and the user did not provide any ranking, the correct boundary was selected to be the one which occurred most frequently among the input sequences. (iv) If two boundaries occurred with the same frequency, the correct boundary was defined as the one inferred from the sequence with higher mapping identity.

A gene may encode multiple transcripts due to alternative splicing and a transcript may correspond to multiple ESTs due to redundant sequencing. Thus, following Riken's definition, we grouped ESTs/mRNAs into transcripts and clustered alternatively spliced transcripts into one transcriptional unit (TU), if the ESTs/mRNAs shared exonic overlap of at least one nucleotide and had the same chromosomal orientation. Grouping ESTs/mRNAs into TUs was advantageous over previous pipelines that grouped them into genomic clusters (2,3,5) or 'transcriptional forests' (26) because a cluster or a transcriptional forest may accidentally include multiple TUs because of the existence of hundreds of triplet and quadruplet cases of *cis*-SAs that we had previously discovered (5).

We used the Splicing Variant Analysis Platform (SVAP) (http://svap.cbi.pku.edu.cn) to assemble ESTs/mRNAs into alternatively spliced transcripts using splice graphs. Similar to existing variant finders such as ASPIC (27) and ESTGene (28), SVAP represented a splice graph as a directed acyclic graph (DAG), using nodes to represent

exons and edges to represent relationship between exons. SVAP walked through all possible paths of the splicing DAG and created a variant for each path. To filter out possible false positives, SVAP discarded those variants with at least one exon that was not covered by any sequence evidence (24).

The result of SVAP was a huge collection of isoforms assembled into TUs, from which we then identified *trans*- and *cis*-SAs. We ran BLASTN using the collection of isoforms as queries against their reverse complement sequences (by setting BLASTN parameter −S to 2). Previous work has used two different criteria to define *trans*-SAs: (i) an *e*-value cutoff of 10^{-9} (18) and (ii) an *e*-value cutoff of 10^{-9} with an identity threshold of 98% (10). We chose the former criteria because the sequence identity between the partners of two validated *trans*-SAs in *Lymnaea*, antiNOS-1/Lym-nNOS and antiNOS-2/Lym-nNOS, was both 80%, lower than the identity threshold in the latter criteria (14). miRNAs are also known to form imperfect base-pairing with their *trans* targets. If a sequence and the reverse complement sequence of another sequence had a pairwise BLASTN *e*-value lower than 10^{-9} and were mapped to nonoverlapping genomic coordinates, they were considered a *trans*-SA pair. Otherwise, if they had overlapping genomic coordinates, they were considered a *cis*-SA pair. We then condensed *trans*-SA pairs into *trans*-SA TU pairs, which were reported and analyzed in this manuscript.

Finally, we applied additional filters to further ensure the high quality of the *trans*-SAs identified. First, we removed *trans*-SAs that had any pairing regions mapped to known repeats (defined by UCSC Genome Browser) (29). Second, we mapped the *trans*-SA TUs to known pseudogenes downloaded from www.pseudogene.org and discarded all TUs having overlapping chromosomal coordinates with a pseudogene. Third, to avoid false TUs resulting from incorrectly inferred transcript orientation, we followed the strategy published by Engstrom *et al*. (24) and discarded a TU if the number of transcripts mapped to it were less than a threshold t. t was set to be the smallest integer greater than 2 for which $P(\mathrm{Bin}(N, P)\ t) \leq 0.01$, where N was the total number of ESTs/mRNAs mapped to the TU, P was the estimated rate of misorientation of ESTs/mRNAs (0.5% in our case) and Bin stood for binomial distribution (24). This stringent filter offered an additional benefit of eliminating possible noisy transcription or genomic sequence contamination because these were less likely to have large number of transcripts.

Determination of coding potential

Our pipeline for determining coding potential of a TU was summarized in Supplementary Figure S2. We first identified protein-coding TUs based on NCBI gene annotations (30). For unannotated TUs, we used Coding Potential Calculator (CPC) to predict their coding potential (31). CPC extracts six features from the transcript's sequence and inputs these features into a support vector machine (SVM) classifier. Using several reference databases as testing datasets, the accuracy of CPC was shown to be over 91%, outperforming other existing prediction algorithms (31). If CPC predicted that a TU had both noncoding RNAs and protein-coding RNAs, as some noncoding RNAs are known to be transcribed from protein-coding genes (32), we considered this TU as protein-coding. Only those TUs in which all RNAs were predicted to lack coding potential were considered as noncoding.

Calculation of hybridizing potential of *trans*-SAs by DINAMelt

DINAMelt, a RNA hybridization prediction program (33,34), used an *in silico* RNA hybridization model to predict the hybridizing regions between two RNAs and was able to handle matched pairs, mismatches and symmetric and asymmetric interior loops. The thermodynamic parameters of oligonucleotides by DINAMelt had good agreement with experimental data (34) and it had been successfully used in many area such as the design of gene probes and microarray oligonucleotides and the prediction of miRNA targets (35). The parameters of DINAMelt were set as default: the concentrations of two molecules were equal and two molecules were hybridized at 37°C.

Finding functional categories and pathways enriched in *trans*-SAs

To study the functional categories involved, we mapped all TUs to Entrez Gene (30) and retrieved their functional categorization based on Gene Ontology (GO) (36) annotations available in Entrez Gene. We then used GO::TermFinder (37) to find statistically enriched GO terms. Assuming a hypergeometric distribution, GO::TermFinder compared the number of *trans*-SA TUs that fell into each functional category against the number of TUs in the whole genome that fell into the same functional category and calculated a *P*-value for each enriched GO term. GO::TermFinder also implemented a 1000-simulation-based correction for multiple hypotheses testing for every *P*-value. The functional categories with corrected $P\text{-value} \leq 0.01$ were considered statistically enriched in *trans*-SAs.

To study the pathways involved, we used the KOBAS software (38) to assign TUs to metabolic pathways based on sequence similarity to sequences in known pathways in the KEGG database (39). KOBAS then compared the number of *trans*-SA TUs in each pathway against the number of all TUs in the genome in the same pathway, assuming a hypergeometric distribution (38). To reduce Type-1 errors, KOBAS performed an FDR correction. Pathways with $q\text{-value} \leq 0.01$ were considered to be statistically enriched.

Analysis of expression profiles of *trans*-SAs

Significant positive or negative expression correlation could give hints on antisense RNAs' concordant or reciprocal regulation (40). We used the enlarged human SAGE library consisting of 309 samples on GPL4 platform, downloaded from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL4). We reconstructed 22 'tissue-type' SAGE libraries (adrenal cortex, blood, bone, brain, breast, colon, eye, kidney, liver,

lung, muscle, ovary, placenta, prostate, skin, stem cell, stomach, thyroid, pancreas, peritoneum, sciatic nerve and sperm). After being normalized to tags per million (*tpm*), only tags that could be mapped unambiguously to TUs by SAGEMap (http://www.ncbi.nlm.nih.gov/SAGE/) were used. The expression level of a TU in a tissue library was set to the sum of *tpm* of all SAGE tags containing the TU in that tissue library (41). Only TUs with at least 3 *tpm* across all the tissues were kept in order to eliminate potential sequencing errors in low-abundance SAGE tags (42). We adopted the correlation coefficient r following the same strategy proposed by RIKEN group (40). We made one important improvement: we tested whether a correlation coefficient r was statistically significantly different from what one would expect by chance in which the population correlation coefficient ρ was equal to zero. We calculated the P-value using t-test. To reduce Type-1 errors, we then performed an FDR correction (43) with a stringent cutoff of q-value ≤ 0.01.

Conservation of *trans*-SAs between human and mouse

We sought to investigate whether any *trans*-SAs were conserved between human and mouse. We first mapped TUs across species by NCBI HomoloGene (44) and BLASTN search (5): (i) If both partners of a human *trans*-SA could be mapped to partners of a mouse *trans*-SA by HomoloGene, this *trans*-SA was considered conserved between human and mouse. (ii) If one partner of a human *trans*-SA could be mapped to a partner of a mouse *trans*-SA by HomoloGene but the other partner could not, we calculated the sequence similarity between the unmapped partners using pair-wise BLASTN. If BLASTN e-value $< 10^{-10}$, identity $> 80\%$ and alignment length >100 nt (5), we considered this *trans*-SA pair to be conserved between human and mouse. (iii) For the remaining human *trans*-SAs with neither partner mapped by HomoloGene, we performed all-against-all BLASTN search against mouse *trans*-SAs. A *trans*-SA was considered conserved between human and mouse if both partners passed the above BLASTN similarity cutoff.

We next considered a second evidence of conservation, complementary to the first one, by human-mouse BLASTZ genome alignment that identified homologous regions of the human and mouse genomes by alignment of large neutrally evolving regions, allowing nucleotide substitutions (45). We extracted the borders for all human TUs involved in *trans*-SAs and identified their homologous genomic regions in mouse, using the human-mouse BLASTZ CHAIN pair-wise alignment from the UCSC Genome Browser Database (assembly hg18, mm8) (29). We chose the CHAIN alignment because it permitted a region of the human genome to be aligned to more than one region in the mouse genome (46). If the genomic regions of a human *trans*-SA and a mouse *trans*-SA were homologous, the *trans*-SA was considered conserved between human and mouse.

RESULTS

A new pipeline identified *trans*-SAs in 10 species

Using the new pipeline, we identified *trans*-SAs in human and nine other animal species. Table 1 showed the number and percentage of *trans*-SAs identified. The percentage of TUs involved in *trans*-SAs among all TUs was as high as 4.13%. In particular, in human, we found that 2896 TUs (or 2.89% of all human TUs) were involved in 3327 *trans*-SAs (sometimes one TU may be involved in two or more *trans*-SAs). Our results showed that the number of human *trans*-SAs had previously been severely underestimated. Lehner *et al*. (18) and Li *et al*. (10) identified 92 and 161 human *trans*-SAs on the transcript level, respectively, whereas we identified thousands on the TU level. As another example, in mouse, we found that 1519 TUs (or 3.13%) were involved in *trans*-SAs. The chromosomal distribution of TUs involved in human *trans*-SAs could be viewed at http://trans.cbi.pku.edu.cn/html/trans_overview.php?spe=hs. A segment on human chromosome 10 was illustrated in Figure 1.

Table 1. Input and output statistics of *trans*-SAs in 10 species

Species	GoldenPath genome version	Number of orientation-reliable sequences mapped to exact genomic location from NATsDB	Total number of TUs	Number of *trans*-SA TU pairs	Number of TUs involved in *trans*-SAs[b]	Proportion of TUs involved in *trans*-SAs among all TUs (%)
Human	hg18	4 494 665	99 919	3327	2896	2.89
Mouse	mm8	2 100 305	57 416	1519	1799	3.13
Fly	dm2	310 319	15 375	76	122	0.79
Worm	ce2	291 395	19 815	41	69	0.35
Sea squirt	ci2	414 454	14 273	1216	589	4.13
Rat[a]	rn4	463 787	40 428	110	188	0.47
Cattle[a]	bosTau2	536 939	27 102	251	302	1.11
Dog[a]	canFam2	203 772	16 956	31	54	0.32
Chicken[a]	galGal2	299 931	21 426	37	66	0.31
Zebrafish[a]	danRer4	522 259	23 809	185	282	1.18

[a]Due to the relatively small amount of available ESTs compared to their genome size for these species, the absolute numbers reported for these five species may be low and the percentages may not be accurate. However, the candidate *trans*-SAs identified should still be reliable.
[b]Previous papers reported the number and percentages of genes forming *cis*-SAs in terms of clusters or transcriptional forests. Here, we reported the number and percentage in terms of TUs.

We stored *trans*-SAs from all 10 species in a MySQL 5.0 (http://www.mysql.com/) relational database and developed a web interface of the database using PHP (http://www.php.net/) and GD (http://www.boutell.com/gd/) graphical libraries. The database, named *Trans*-SAMap, is freely available at http://trans.cbi.pku.edu.cn/. Users could interactively browse the database or search the database by TUs, Entrez Gene synonyms (44), mRNA/EST accession numbers, chromosomal locations and sequences (by BLAST similarities). Each *trans*-SA was displayed graphically, showing the overlapping patterns and other annotations.

One-fourth of human *trans*-SAs involved noncoding RNAs

We next examined whether and how many human *trans*-SAs involved noncoding TUs. We classified *trans*-SAs into three types: protein-coding–protein-coding pairs (p–p pairs), noncoding–protein-coding pairs (n–p pairs) and noncoding–noncoding pairs (n–n pairs). In human, 332 noncoding TUs were involved in 830 *trans*-SAs (24.9% of all human *trans*-SAs), including 753 n–p pairs and 77 n–n pairs. On average, each noncoding TU was involved in 2.5 *trans*-SAs, indicating that some noncoding RNAs might not only have *trans*-antisense activities but might also potentially regulate multiple targets.

In 247 human n–p pairs, the protein-coding TU had coding regions (CDS) annotation. Among them, in 172 (69.6%) pairs, the partner noncoding TU overlapped with CDS of the protein-coding TU, 46 (18.6%) with 3′ UTR and 29 (11.8%) with 5′ UTR. Different locations of the overlapping in *trans*-SAs may be related to different regulatory mechanisms (1,10). Overlapping with CDS of a sense transcript may destroy the sense transcript through an RNA interference (RNAi) mechanism (18) or interfere with the sense transcript's interaction with their *trans*-acting proteins; an overlap in the 3′UTR might affect the sense-mRNA's stability in cytoplasm (47) or its transport out of the nucleus (1) and an overlap in the 5′UTR might regulate mRNA translation initiation in a manner similar to, for example, what occurs with RNA III (48).

Trans-SAs may form sophisticated RNA–RNA pairing patterns

We observed that *trans*-SAs had diverse sequence complementary patterns, which may relate to their being involved in different regulatory mechanisms. Particularly, for many *trans*-SAs, there was more than one region that was complementary between two transcripts. Based on different patterns of BLASTN High-Scoring Pairs (HSPs) in a *trans*-SA, we constructed a new classification schema consisting of four classes of *trans*-SAs, shown in Figure 2A and described below:

(i) 'Single': both transcripts contain only one HSP.
(ii) 'Optional': one transcript contains a region that can form HSPs with more than one distinct region in the paired transcript
(iii) 'Parallel': both transcripts contain more than one HSP, forming pairing regions in parallel.
(iv) 'Mixed': a pair of *trans*-SA transcripts with other, often more complicated, pairing patterns that cannot be classified into any of the above three classes. This class included transcripts that had simultaneously two or more of the above pairing patterns of HSPs (e.g. Figure 2A 'Mixed' which is a mixture of 'Optional' and 'Parallel') as well as transcripts in which the HSPs overlap in their sequence coordinates.

We observed that although the 'Single' class was predominant, a significant number of *trans*-SAs fell into the 'Optional', 'Parallel' and 'Mixed' classes (Figure 2B).

The number of pairs with a perfect sequence match and the median length of the overlapping region were listed in Supplementary Table S1. Most *trans*-SA overlapping regions were longer than 50 bps, which was clearly different from those of small *trans*-acting RNAs such as miRNAs. The majority of the pairings were imperfect, which was reasonable because, unlike *cis*-SAs, two partners of a *trans*-SA were transcribed from different chromosomal loci and small *trans*-acting RNAs such as miRNAs were known to have imperfect matches with their targets.

We further investigated whether two transcripts in a *trans*-SA could in fact bind to each other based on the melting profiles calculated by DINAMelt (33,34). The HSPs in 89.42% of *trans*-SAs in the 'Single' class were covered by the hybridizing regions predicted by DINAMelt, indicating that, in addition to being complementary in sequence, they were likely to hybridize in solution. Investigating the other classes, we applied two criteria: a nonstringent one that specified that as long as any HSP was covered by the hybridizing regions predicted

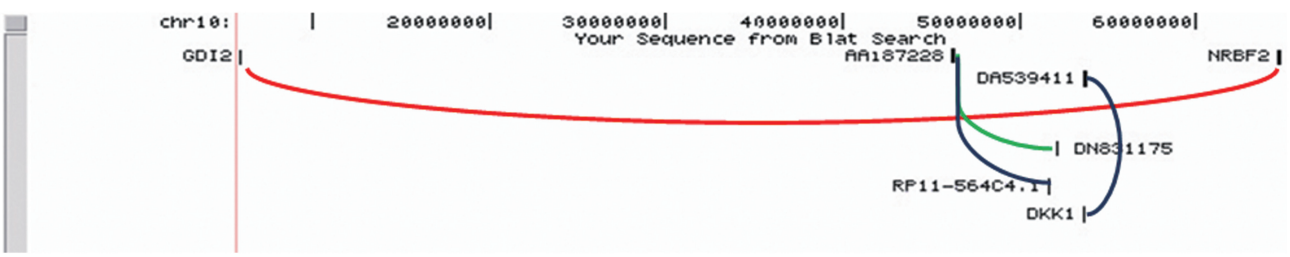

Figure 1. *Trans*-SAs within a segment on human chromosome 10. Arcs linking two partners of a *trans*-SA pair were color-coded: red arc linked protein-coding–protein-coding pairs; blue arc linked noncoding–protein-coding pairs and green linked noncoding–noncoding pairs. In particular, a noncoding TU, AA187228, could pair with a coding TU (RP11-564C4.1) as well as a noncoding TU (DN831175).

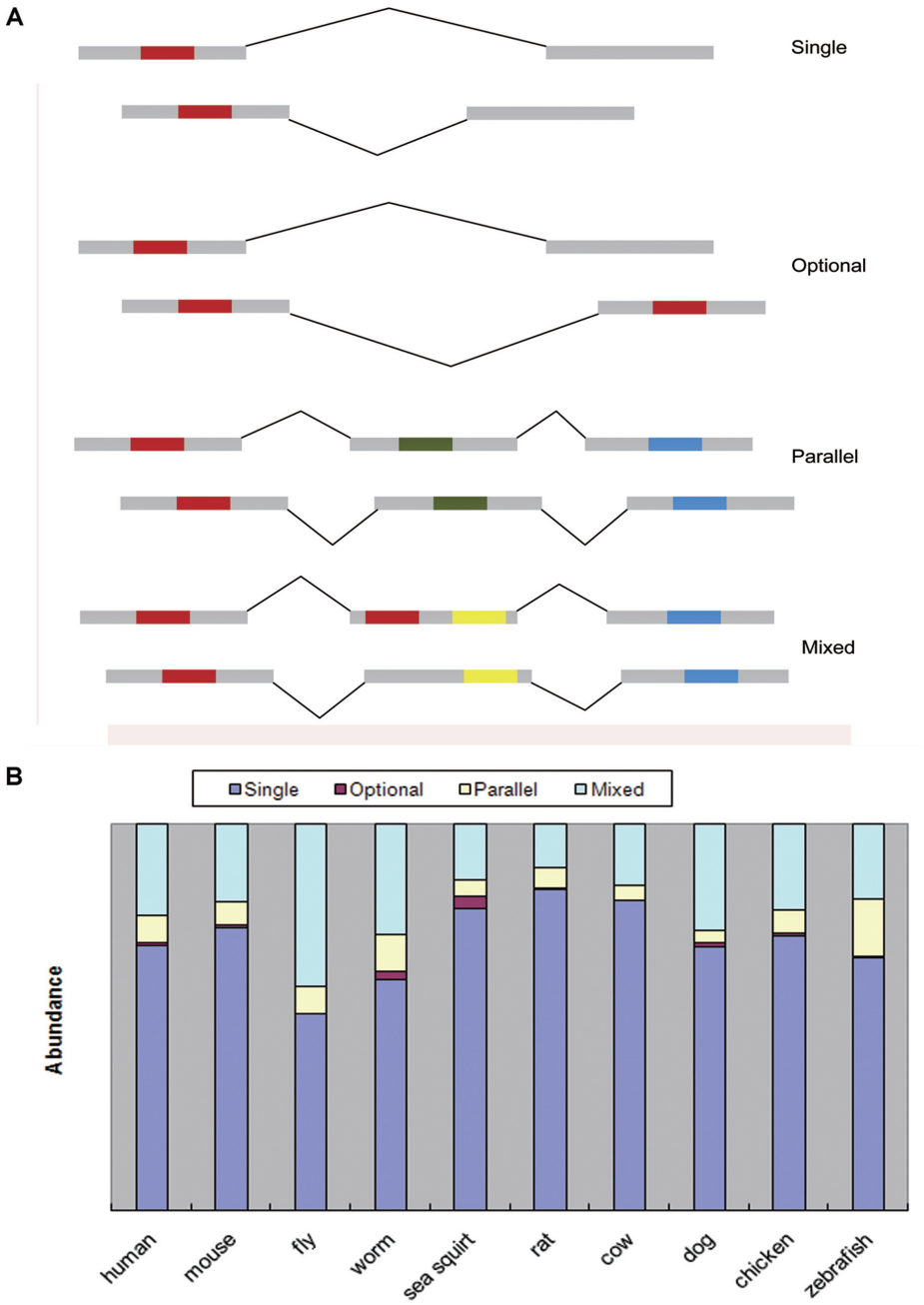

Figure 2. *Trans*-SAs may have complex pairing patterns. (**A**) *Trans*-SAs are classified into four classes based on their HSP patterns. Grey blocks indicate exons. Folding lines between blocks indicate splicing junctions. Red, blue, yellow and green blocks indicate complementary regions. (**B**) The abundance of four classes of *trans*-SAs in 10 species.

by DINAMelt, the *trans*-SA pair was considered to form dsRNA in solution and a stringent one that required that all HSP regions in both transcripts must be covered by the DINAMelt-predicted pairing regions. If a *trans*-SA pair satisfied the stringent criteria, then it was more likely that a complex pairing pattern indeed occurred in solution. Applying the nonstringent criteria, we found that 93.81% of the pairs in the 'Optional' class, 98.10% in 'Parallel' and 96.71% in 'Mixed' had at least one HSP covered by the hybridization regions predicted by DINAMelt, indicating that they may form dsRNA. Applying the more stringent criteria, we found that 44.62% of the pairs in the 'Optional' class, 19.43% in 'Parallel' and 74.45% in 'Mixed' had all HSPs in both transcripts covered by the hybridizing regions predicted by DINAMelt, indicating that complex pairing patterns might indeed occur.

TUs in *trans*-SAs were significantly enriched in nucleic acid binding, ion/protein binding and transporter/signal transducer functions and pathways

As shown in Supplementary Table S2 and S3, TUs in *trans*-SAs were enriched in functional categories and pathways

Table 2. Enriched Gene Ontology (GO) categories in *trans*-SAs that were consistent between human and mouse (corrected *P*-value ≤ 0.01, up to GO level 4)

Biological Process	Molecular function	Cellular component
Regulation of cellular process	Metal ion binding	Intracellular
Transcription	Cation binding	Intracellular organelle
Transport	DNA binding	Intracellular part
Organelle organization and biogenesis	Purine nucleotide binding	Intracellular organelle part
DNA packaging	RNA binding	Cytoplasm
Cellular localization	Hydrolase activity, acting on acid anhydride	Membrane
Intracellular transport	Cytoskeletal protein binding	Cytoplasmic part
Negative regulation of biological process	Transferase activity, transferring phosphorus-containing group	Membrane part
Positive regulation of biological process	Ligase activity, forming carbon–nitrogen bond	Cytoskeletal part
Negative regulation of cellular process	Unfolded protein binding	Plasma membrane
Protein transport		Ribonucleoprotein complex
Ion transport		Nuclear part
Positive regulation of cellular process		Organelle lumen
Regulation of cell cycle		Nuclear lumen
Vesicle-mediated transport		Chromosomal part
Organ development		Membrane fraction
Response to DNA damage stimulus		Nucleosome
Cytoplasm organization and biogenesis		Cell fraction
		Organelle membrane
		Mitochondrial membrane
		Voltage-gated sodium channel complex
		Endomembrane system
		Organelle inner membrane
		Transcription factor complex

The detailed *P*-values were listed in Supplementary Table S2.

involved in nucleic acid binding, similar to the functional bias of *cis*-SAs (5). In addition, TUs in *trans*-SAs were also more common in ion/protein binding and transporter activities as represented by the GO categories 'ion binding', 'ion transport', 'protein binding', 'protein transport', 'membrane-bound organelle' and 'voltage-gated sodium channel complex' and by the KEGG pathways 'Protein export', 'Adhesion junction' and 'Tight junction'. These were categories in which *cis*-SAs were not commonly found. Enriched functions that were consistent between human and mouse were shown in Table 2.

Our study revealed, for the first time, the statistical functional bias of *trans*-SAs in animals. Results were in agreement with those from the small number of previous experimental studies of individual *trans*-SAs in prokaryotes and plants. For instance, in *Escherichia coli*, *RyhB*, a 90 nt noncoding RNA, is found to down-regulate a group of iron-storage and iron-using proteins by *trans*-acting when iron is limited (49); *MicF* gene encodes a 93 nt noncoding antisense RNA and regulates target *ompF* expression via a *trans*-pairing mechanism that inhibits translation and induces mRNA degradation, thereby affecting stress response cellular processes (50). In plants, transcripts involved in *trans*-SAs were found to be over-represented in function categories such as signal transducer activity and transporter activity (19).

Some *trans*-SAs showed significant concordant (positively correlated) or reciprocal (negatively correlated) expression patterns

Significant positive or negative expression correlation between *trans*-SAs could suggest antisense RNAs' concordant or reciprocal regulation. Among 392 *trans*-SAs that co-occurred in at least 3 tissues and were used in subsequent expression analysis, 164 (or 42%) showed significant expression correlation. Three pairs, including two p–p pairs (*ZNF227* versus *CR593740*, *HIST1H2BG* versus *HIST1H2AK*) and one n–p pair (*CN480497* versus *CYP4A11*), showed a significant negative expression correlation (q-value ≤ 0.01), whereas remarkably, 161 pairs, including 146 p–p pairs and 15 n–p pairs, showed a significant positive expressional correlation (q-value ≤ 0.01).

Some *trans*-SAs were conserved between human and mouse

Our multi-species dataset enabled us to identify how many *trans*-SAs were conserved between human and mouse. Among *trans*-SAs where both partners had one-to-one mapping by HomoloGene (496 in human and 321 in mouse by orthologous mapping), five p–p *trans*-SAs in human (1.0%) were conserved in mouse (Table 3, panel A). To test whether the percentage of conserved *trans*-SAs differed from what one would expect by chance, we extracted all human and mouse genes in HomoloGene and constructed 9600 pseudo human SAs and 10 254 pseudo mouse SAs by pairing one gene to another randomly without repetition. No pseudo human SA pairs were found to be conserved in mouse. Fisher test confirmed that the fraction of conserved *trans*-SAs significantly differed from what would be expected by chance (*P*-value ~2.94e-7). Among *trans*-SAs where one partner was mapped to HomoloGene and the other was mapped by BLASTN similarity (1801 pairs in human and 832 in mouse), an additional four p–p and two n–p *trans*-SAs in human were conserved in mouse (Table 3, panel B).

Table 3. List of human *trans*-SAs that were conserved between human and mouse

Trans-SA	Gene symbol	Gene ID	TU ID
Panel A			
1	FER	2241	hs_27889_p.14
	THOC2	57187	hs_37092_m.1
2	TUBB3	10381	hs_13578_p.3
	LOC401565	401565	hs_36040_m.0
3	KCNA6	3742	hs_7160_p.0
	KCNA10	3744	hs_1560_p.3
4	LOC401565	401565	hs_36040_m.0
	TUBB4	10382	hs_16192_m.0
5	PPAN	55337	hs_16298_p.0
	ANGPTL2	23452	hs_35783_m.3
Panel B			
1 (n–p)[a]	GPD1L	23171	hs_23040_p.1
	CA450491[b]		hs_7693_m.2*
2 (n–p)[a]	MAFK	7975	hs_30950_p.0
	DB018139[b]		hs_15104_p.0*
3	TUBB2C	10383	hs_36040_p.0
	MTAP	4507	hs_34800_p.2
4	H2AFJ	55766	hs_7373_p.0
	MGC12935	84780	hs_29114_p.0
5	TUBA1C	84790	hs_7676_p.2
	MGC16703	113691	hs_22137_m.1
6	TUBA1C	84790	hs_7676_p.2
	LOC730222	730222	hs_2584_m.0
Panel C			
1	BX326232[b]		hs_11717_m.0
	C14orf37	145407	hs_10399_m.0
2 (n–p)[a]	LOC729226	729226	hs_13114_p.0*
	LOC390616	390616	hs_4793_m.0
3	BU662546[b]		hs_13878_p.1
	LOC390616	390616	hs_4793_m.0
4 (n–p)[b]	BG398362[b]		hs_18569_p.0
	LOC400794	400794	hs_2189_m.2*
5	CR602957[b]		hs_19485_m.8
	C14orf37	145407	hs_10399_m.0
6 (n–p)[a]	LOC400794	400794	hs_2189_m.2*
	CDC27	996	hs_14521_m.10
7 (n–p)[a]	BQ935655[b]		hs_8564_p.2
	LOC400794	400794	hs_2189_m.2*
8	ADRA1B	147	hs_28453_p.0
	BE535870[b]		hs_36406_m.0
9	BE535870[b]		hs_36406_m.0
	DB239600[b]		hs_20661_m.4
10 (n–p)[a]	BM477121[b]		hs_36425_m.0
	LOC400794	400794	hs_2189_m.2*
11	LOC338739	338739	hs_5542_p.3
	BG612114[b]		hs_30109_p.0
12	BC036362[b]		hs_8886_m.6
	BE535870[b]		hs_36406_m.0

Panel A, List of human *trans*-SAs that were conserved in mouse identified by cross-referencing HomoloGene. Panel B, List of human *trans*-SAs that were conserved in mouse identified by HomoloGene cross-reference of one TU and pair-wise BLAST similarity of the other TU. Panel C, Human *trans*-SAs that were conserved in mouse identified by pair-wise BLASTN similarity of both TUs.

[a](n–p) indicates a *trans*-SA pair in which a noncoding TU pairs with a protein-coding TU. The TUs without coding potential were marked with asterisk. When not specified, the *trans*-SA pair is a protein-coding–protein-coding (p–p) pair. We did not observe any n–n pairs conserved between human and mouse.

[b]These TUs have no NCBI GeneID or Gene name. Thus we used an isoform to represent the TU.

Among the remaining *trans*-SAs both partners of which were not covered in HomoloGene but mapped by BLASTN similarity (1030 pairs in human and 366 in mouse), another twelve human *trans*-SAs were conserved in mouse (7 conserved p–p pairs and 5 n–p pairs, shown in Table 3, panel C).

Two thousand five hundred and fifty six human *trans*-SAs could be mapped to mouse using BLASTZ alignment as genome-based evidence of conservation. Among them, 1017 (39.7%) formed *trans*-SAs in mouse and were considered conserved between human and mouse (Supplementary Table S4), including 826 p–p pairs, 188 n–p pairs and 3 n–n pairs. We shuffled the human and mouse *trans*-SAs and constructed equal number of human and mouse pseudo *trans*-SAs. Only 125 out of 2443 pseudo human *trans*-SAs were found to be conserved in mouse following the same protocol. Fisher test confirmed that the fraction of the conserved *trans*-SAs we detected was significantly higher than what would be expected by chance in the pseudo dataset (P-value $< 2.2e\text{-}16$). These results demonstrated that the conserved *trans*-SAs we observed were unlikely a chance event and suggested that they might have important functions.

A proposed mechanism of origination of some *trans*-SAs via *cis*-SAs and gene duplication

Comparing *trans*-SAs and *cis*-SAs in human, we found that 616 TUs involved in *trans*-SAs also had *cis*-SA partners. Based on information in GeneCards (51), we found that, for 30 such TUs, their *cis*-SA partner and *trans*-SA partner were members of the same gene family (Supplementary Table S5). Given the high abundance of *cis*-SAs and the relative simplicity with which *cis*-SAs can originate (52,53), it was more likely that in these dual cases, *cis*-SAs originated prior to *trans*-SAs. The following example illustrated one possible scenario of *trans*-SA origination. Human *MYH11* (myosin, heavy polypeptide 11) and *MYH9* (myosin, heavy polypeptide 9) are paralogs from the same family (based on GeneCards (51)) encoded at separate chromosomal loci. A third gene, *NDE1* (nudE nuclear distribution gene E homolog 1), formed a *cis*-SA pair with *MYH11* through a polyA-sharing mechanism. This *cis*-SA pair appeared to have originated recently because in both mouse and chicken, *NDE1* and *MYH11* were simply neighboring genes on the opposite strands without any overlap. Due to the high sequence similarity between *MYH11* and *MYH9*, *NDE1* forms a *trans*-SA pair with *MYH9* (shown in Figure 3A).

A second example involves genes duplicated by segmental duplications. The pairing regions in 500 human *trans*-SAs (15% of all human *trans*-SAs) could be mapped to known segmental duplication regions based on UCSC genome browser database. For example, the *RFPL* (Ret finger protein-like) gene family is known to have undergone a series of duplications from an ancestor containing a RING-B30 domain (54). *RFPL1* and *RFPL3* are two members of this family, and *RFPL1* is believed to be the older of the two evolutionarily (54). *RFPL1S* formed a *cis*-SA pair with *RFPL1*, and on a separate genomic locus, *RFPL3S* formed a *cis*-SA pair with *RFPL3*.

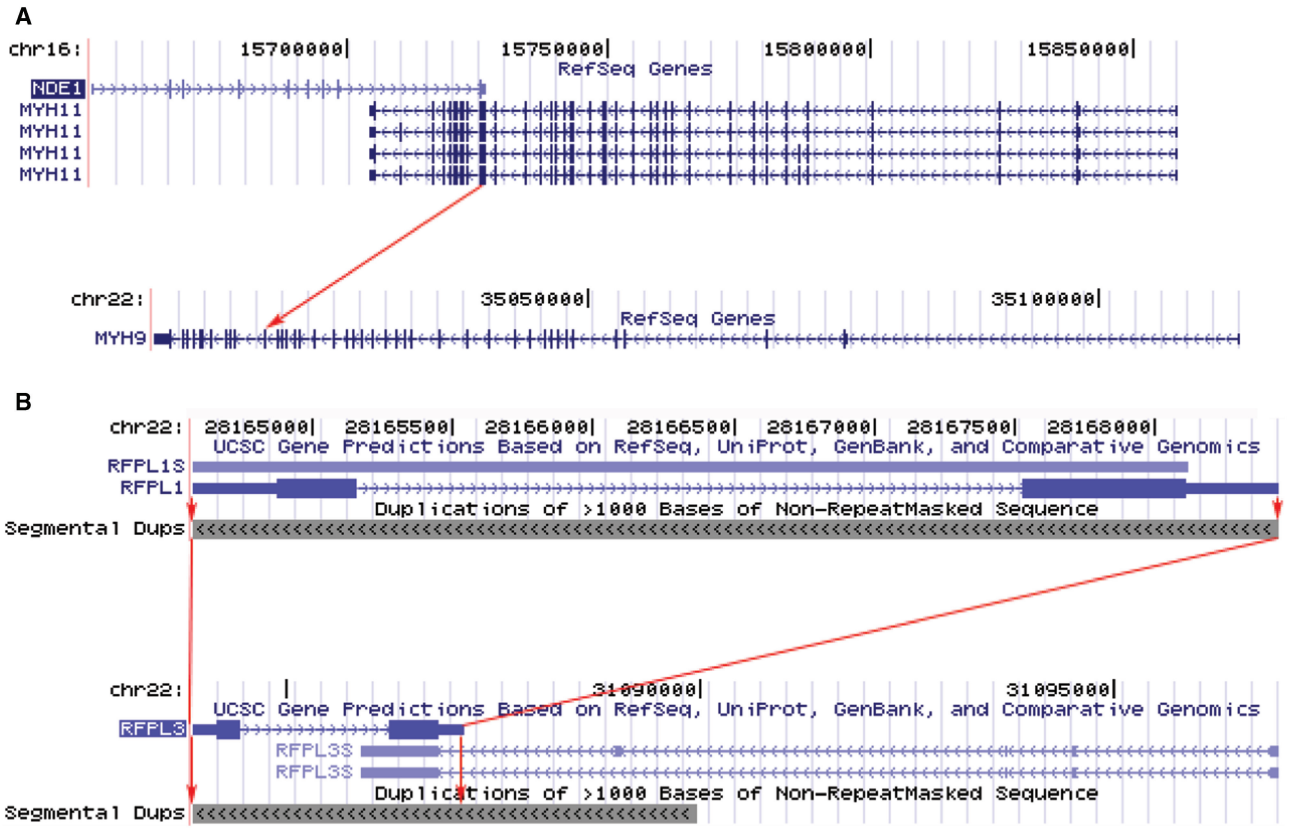

Figure 3. Some *trans*-SAs were related to *cis*-SAs through paralogues. (**A**) *NDE1* and *MYH11* formed a *cis*-SA pair sharing the last exon of *NDE1* (99 bp overlap with perfect complementarity), from UCSC genome browser. The arrow indicates *MYH9*'s paralogous exon to the *cis*-pairing region of *MYH11*. *NDE1* shares this paralogous exon and thus *trans*-pairs with *MYH9*. (**B**) *RFPL1* and *RFPL1S* formed a *cis*-SA (293 bp overlap with perfect complementarity). *RFPL3* and *RFPL3S* formed another *cis*-SA (588 bp overlap with perfect complementarity). *RFPL3S* and *RFPL1* formed a *trans*-SA (591 bp overlap with an *e*-value of 0.0 and an identity of 95%). *RFPL3* and *RFPL1S* formed another *trans*-SA (519 bp overlap with an *e*-value of 0.0 and an identity of 96%). Red arrows indicate the *cis*-pairing region of *RFPL1* and *RFPL1S* was duplicated to the locus of *RFPL3* and *RFPL3S*.

Because *RFPL1* and *RFPL3* shared 95% sequence identity across their full length (54), *RFPL1* and *RFPL3S* constituted a *trans*-SA pair, while *RFPL3* and *RFPL1S* constituted another *trans*-SA pair (shown in Figure 3B).

DISCUSSION

Our pipeline had the highest coverage compared to previous works (10,18,19) because of the use of ESTs in addition to mRNAs. At the same time, our pipeline was also rigorous, applying stringent quality control filters. We identified an order-of-magnitude more *trans*-SAs in human (at least 20 times higher than previous studies (10,18)), as well as *trans*-SAs in nine other animal species for the first time, allowing new features of *trans*-SAs to be revealed. For instance, we observed that *trans*-SAs could potentially form more complex RNA–RNA pairing patterns than previously thought (10,18).

We followed the common practice of filtering out repeats and pseudogenes in order to identify *trans*-SAs that were most likely to be functional (10,18). However, although repeats were previously thought to be genomic parasites (55) and their transcripts were major targets of RNAi pathways (56), recent evidence has suggested repeats' roles in the regulation of transcription or post-transcriptional events (57–59). Some repeats were found to be under strong selective constraint (55). Some were found to function as antisense (60,61). Similarly, pseudogenes were often considered evolutionary 'dead-ends' and filtered out. However, recently a large number of novel endogenous siRNAs were reported to be derived from a protein-coding gene and a *cis*-antisense transcript of its pseudogene, indicating the functional role of pseudogene (or at least its *cis*-antisense) in regulating the parental mRNA's level (62,63). Interestingly, we found that, if we had not filtered out repeats and pseudogenes in our pipeline, the final set of *trans*-SAs would be 20 times larger. The potential functions of *trans*-SAs involving repeats and pseudogenes are currently under further investigation.

Our results indicated that long noncoding RNAs, aside from functioning as *cis*-antisense transcripts such as *Air* (21) and *Xist* (20), might also function via a *trans*-acting mechanism in animals and could regulate more than one target. It was previously known that *DsrA*, an 87 nt noncoding RNA in *Escherichia coli*, regulates at least five different genes via *trans*-acting RNA–RNA interaction by different pairing regions (64). Therefore, a *trans*-acting mechanism for noncoding RNAs appeared to be capable

to regulate more targets than a *cis*-antisense mechanism in which a noncoding RNA could regulate only one target at the same genomic locus on the opposite strand. This is reminiscent of the difference between short *trans*-SAs, i.e. miRNAs, which can regulate multiple targets at different genomic loci, and siRNAs, which most often regulate one target at the same genomic locus.

The expression profiles of *trans*-SAs pointed toward a complexity of regulatory interactions. Well-known short *trans*-acting RNAs and their targets have been shown to have different patterns of expression correlation. miRNAs could act through at least two mechanisms: imperfect base-pairing within 3′UTR that blocks translation, which does not affect target mRNA's abundance, or perfect base-pairing with miRNA-mediate cleavage and inactivation of target mRNA, which results in a decrease of target mRNA's abundance (65). Data have shown that the long *trans*-acting RNA, *DsrA*, has opposite expression effects on two different targets both mediated by RNA–RNA interaction: it decreases *H-NS* mRNA level but increases *RpoS* mRNA level (64). Among 392 *trans*-SAs having expression data in human tissues, we identified 164 *trans*-SAs that had significant expression correlations that might be involved in inhibition or stabilization effect of antisense transcripts. As for the remaining *trans*-SA transcripts, they might be implicated in other aspects of RNA function such as translation, RNA editing, trafficking and subcellular localization.

We used the large human SAGE library consisting of 309 samples on GPL4 platform for our expression studies. Admittedly this SAGE data set was still far from complete because it provided enough data for only 12% of all human *trans*-SAs. Thus, it would be important to continue to use high-throughout technologies to further investigate *trans*-SAs' expression profiles. Nevertheless, we were able to find a remarkable 161 *trans*-SA pairs with positive expression correlation. Positive expression correlation of *trans*-SAs could be caused by three possible mechanisms. The first was a dsRNA-mediated stabilization function (66,67). Scheele *et al*. observed that a mammalian noncoding antisense molecule (ncNAT) could increase the abundance of its *cis*- target mRNA (svPINK1) in a 'tail-tail' overlapping pattern (67). In the overlapping region we detected neither AU-rich elements (AREs) nor other rapid decay-associated motifs using UTRscan (68) and theorized a possible mechanism: the 'tail-tail' overlap might affect mRNA stability by reducing mRNA decay where transcripts undergo 3′→5′ exonucleolytic decay subsequent to poly(A) shortening (47,69). This mechanism could potentially account for some of the positively-correlated *trans*-SAs. Among *trans*-SAs where at least one partner had CDS annotation, 15 p–p pairs overlapped at 3′UTR and two n–p pairs overlapped at 3′UTR. A second possible mechanism was that *trans*-antisense might interrupt target intramolecular base-pairing (13) and thereby stabilize target mRNA. For example, in *Escherichia coli*, *DsrA* pairing stabilizes *RpoS* mRNA and stimulates *RpoS* translation, freeing the translation initiation region from intramolecular base-pairing and allowing increased translation and mRNA stability (70). The third mechanism was that *trans*-antisense might function as a transcription activator (48). In *Staphylococcus aureus*, aside from being an activator of translation, RNAIII acts on the initiation of transcription of virulence genes by means of intermediary protein factors (71). Similarly, some *trans*-antisense RNAs in animals might also initiate transcription of their target genes.

We performed several analyses on the evolutionary conservation of *trans*-SAs between human and mouse and provided evidence for *trans*-SAs' conservation. We observed that, as a group, *trans*-SAs were significantly more conserved between human and mouse than what would be expected by chance. However, the lack of detected conservation of many *trans*-SAs did not necessarily mean that they were not functional. First, the ENCODE project recently reported that as many as 50% of the experimentally identified functional elements do not show evidence of evolutionary constraint across mammals, especially for many types of noncoding functional elements (72). Second, TUs in some *trans*-SAs might be under rapid sequence evolution. As widely known, a subset of human protein-coding genes are fast evolving, driven by positive Darwinian selection, such as genes involved in immune response and reproduction (73). Some functional noncoding RNAs were also found to be subject to positive selection when associated with phenotypic radiation (74). Third, some pairs might be species-specific and arose recently after the divergence of human and mouse.

The mechanism of *trans*-SA origination was not studied before. Gene duplication including segmental duplication may contribute to *trans*-SA origination. Segmental duplication is known to play fundamental roles in both gene evolution and genomic disease (75). The pairing regions in 500 human *trans*-SAs (15% of all human *trans*-SAs) could be mapped to known segmental duplication regions (accounting for 2% of all human segmental duplications). The majority (85%) of human *trans*-SAs were not related to segmental duplications

In summary, our comprehensive analysis offers not only candidates for further experiments but also important clues to the function and evolution of *trans*-SAs in multiple animal species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Drs Manyuan Long, Jane Wu, Louis Tao, Zicai Liang and Ge Gao for insightful suggestions. We thank Dr Laurie Goodman for proofreading the manuscript. This work was supported by China Ministry of Science and Technology 863 Hi-Tech Research and Development Programs (No. 2006AA02Z334, 2006AA02Z314, 2006AA02A312, 2007AA02Z165) and 973 Basic Research Programs (No. 2006CB910404, 2007CB946904), and China Ministry of Education 111 project (No. B06001). This study was also supported in part by the National Institutes of Health (NIH)-Intramural Research

Program, National Institute on Drug Abuse. Funding to pay the Open Access publication charges for this article was provided by China Ministry of Education 111 Project (No. B06001).

Conflict of interest statement. None declared.

REFERENCES

1. Vanhee-Brossollet,C. and Vaquero,C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.
2. Chen,J., Sun,M., Kent,W.J., Huang,X., Xie,H., Wang,W., Zhou,G., Shi,R.Z. and Rowley,J.D. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.*, **32**, 4812–4820.
3. Yelin,R., Dahary,D., Sorek,R., Levanon,E.Y., Goldstein,O., Shoshan,A., Diber,A., Biton,S., Tamir,Y., Khosravi,R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, **21**, 379–386.
4. Zhang,Y., Li,J., Kong,L., Gao,G., Liu,Q.R. and Wei,L. (2007) NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res.*, **35**, D156–D161.
5. Zhang,Y., Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
6. Wutz,A., Smrzka,O.W., Schweifer,N., Schellander,K., Wagner,E.F. and Barlow,D.P. (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature*, **389**, 745–749.
7. Munroe,S.H. and Lazar,M.A. (1991) Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J. Biol. Chem.*, **266**, 22083–22086.
8. Prescott,E.M. and Proudfoot,N.J. (2002) Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA*, **99**, 8796–8801.
9. Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
10. Li,Y.Y., Qin,L., Guo,Z.M., Liu,L., Xu,H., Hao,P., Su,J., Shi,Y., He,W.Z. and Li,Y.X. (2006) In silico discovery of human natural antisense transcripts. *BMC Bioinformatics*, **7**, 18.
11. Hirsch,M. and Elliott,T. (2002) Role of ppGpp in rpoS stationary-phase regulation in Escherichia coli. *J. Bacteriol.*, **184**, 5077–5087.
12. Moller,T., Franch,T., Udesen,C., Gerdes,K. and Valentin-Hansen,P. (2002) Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev.*, **16**, 1696–1706.
13. Good,L. (2003) Translation repression by antisense sequences. *Cell Mol. Life Sci.*, **60**, 854–861.
14. Korneev,S. and O'Shea,M. (2002) Evolution of nitric oxide synthase regulatory genes by DNA inversion. *Mol. Biol. Evol.*, **19**, 1228–1233.
15. Korneev,S.A., Park,J.H. and O'Shea,M. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.*, **19**, 7711–7720.
16. Hirano,M. and Noda,T. (2004) Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. *Gene*, **342**, 165–177.
17. Okano,H., Aruga,J., Nakagawa,T., Shiota,C. and Mikoshiba,K. (1991) Myelin basic protein gene and the function of antisense RNA in its repression in myelin-deficient mutant mouse. *J. Neurochem.*, **56**, 560–567.
18. Lehner,B., Williams,G., Campbell,R.D. and Sanderson,C.M. (2002) Antisense transcripts in the human genome. *Trends Genet.*, **18**, 63–65.
19. Wang,H., Chua,N.H. and Wang,X.J. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol.*, **7**, R92.
20. Pauler,F.M., Koerner,M.V. and Barlow,D.P. (2007) Silencing by imprinted noncoding RNAs: is transcription the answer? *Trends Genet.*, **23**, 284–292.
21. Sleutels,F., Zwart,R. and Barlow,D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature*, **415**, 810–813.
22. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
23. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–799.
24. Engstrom,P.G., Suzuki,H., Ninomiya,N., Akalin,A., Sessa,L., Lavorgna,G., Brozzi,A., Luzi,L., Tan,S.L., Yang,L. *et al.* (2006) Complex Loci in human and mouse genomes. *PLoS Genet.*, **2**, e47.
25. Gupta,S., Zink,D., Korn,B., Vingron,M. and Haas,S.A. (2004) Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics*, **5**, 72.
26. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
27. Bonizzoni,P., Rizzi,R. and Pesole,G. (2005) ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. *BMC Bioinformatics*, **6**, 244.
28. Eyras,E., Caccamo,M., Curwen,V. and Clamp,M. (2004) ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res.*, **14**, 976–987.
29. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
30. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–25.
31. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
32. Gerstein,M.B., Bruce,C., Rozowsky,J.S., Zheng,D., Du,J., Korbel,J.O., Emanuelsson,O., Zhang,Z.D., Weissman,S. and Snyder,M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, **17**, 669–681.
33. Dimitrov,R.A. and Zuker,M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, **87**, 215–226.
34. Markham,N.R. and Zuker,M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, **33**, W577–581.
35. Burgler,C. and Macdonald,P.M. (2005) Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics*, **6**, 88.
36. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
37. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
38. Mao,X., Cai,T., Olyarchuk,J.G. and Wei,L. (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, **21**, 3787–3793.
39. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
40. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al.* (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
41. Chen,J., Sun,M., Hurst,L.D., Carmichael,G.G. and Rowley,J.D. (2005) Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet.*, **21**, 326–329.
42. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
43. Dalmasso,C., Broet,P. and Moreau,T. (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.

44. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.
45. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*., **13**, 103–107.
46. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA*, **100**, 11484–11489.
47. Lavorgna,G., Dahary,D., Lehner,B., Sorek,R., Sanderson,C.M. and Casari,G. (2004) In search of antisense. *Trends Biochem. Sci*., **29**, 88–94.
48. Morfeldt,E., Taylor,D., von Gabain,A. and Arvidson,S. (1995) Activation of alpha-toxin translation in Staphylococcus aureus by the trans-encoded antisense RNA, RNAIII. *EMBO J*., **14**, 4569–4577.
49. Masse,E. and Gottesman,S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl Acad. Sci. USA*, **99**, 4620–4625.
50. Delihas,N. and Forst,S. (2001) MicF: an antisense RNA gene involved in response of Escherichia coli to global stress factors. *J. Mol. Biol*., **313**, 1–12.
51. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
52. Keese,P.K. and Gibbs,A. (1992) Origins of genes: "big bang" or continuous creation? *Proc. Natl Acad. Sci. USA*, **89**, 9489–9493.
53. Shintani,S., O'HUigin,C., Toyosawa,S., Michalova,V. and Klein,J. (1999) Origin of gene overlap: the case of TCP1 and ACAT2. *Genetics*, **152**, 743–754.
54. Seroussi,E., Kedra,D., Pan,H.Q., Peyrard,M., Schwartz,C., Scambler,P., Donnai,D., Roe,B.A. and Dumanski,J.P. (1999) Duplications on human chromosome 22 reveal a novel Ret Finger Protein-like gene family with sense and endogenous antisense transcripts. *Genome Res*., **9**, 803–814.
55. Nishihara,H., Smit,A.F. and Okada,N. (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res*., **16**, 864–874.
56. Bernstein,E. and Allis,C.D. (2005) RNA meets chromatin. *Genes Dev*., **19**, 1635–1655.
57. Britten,R.J. (1997) Mobile elements inserted in the distant past have taken on important functions. *Gene*, **205**, 177–182.
58. Brosius,J. (1991) Retroposons–seeds of evolution. *Science*, **251**, 753.
59. Peaston,A.E., Evsikov,A.V., Graber,J.H., de Vries,W.N., Holbrook,A.E., Solter,D. and Knowles,B.B. (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell*, **7**, 597–606.
60. Tchurikov,N.A. and Kretova,O.V. (2007) Suffix-specific RNAi leads to silencing of F element in Drosophila melanogaster. *PLoS ONE*, **2**, e476.
61. van de Lagemaat,L.N., Medstrand,P. and Mager,D.L. (2006) Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol*., **7**, R86.
62. Watanabe,T., Totoki,Y., Toyoda,A., Kaneda,M., Kuramochi-Miyagawa,S., Obata,Y., Chiba,H., Kohara,Y., Kono,T., Nakano,T. *et al*. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539–543.
63. Tam,O.H., Aravin,A.A., Stein,P., Girard,A., Murchison,E.P., Cheloufi,S., Hodges,E., Anger,M., Sachidanandam,R., Schultz,R.M. *et al*. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
64. Lease,R.A. and Belfort,M. (2000) A trans-acting RNA as a control switch in Escherichia coli: DsrA modulates function by forming alternative structures. *Proc. Natl Acad. Sci. USA*, **97**, 9919–9924.
65. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al*. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
66. Moorwood,K., Charles,A.K., Salpekar,A., Wallace,J.I., Brown,K.W. and Malik,K. (1998) Antisense WT1 transcription parallels sense mRNA and protein expression in fetal kidney and can elevate protein levels in vitro. *J. Pathol*., **185**, 352–359.
67. Scheele,C., Petrovic,N., Faghihi,M.A., Lassmann,T., Fredriksson,K., Rooyackers,O., Wahlestedt,C., Good,L. and Timmons,J.A. (2007) The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics*, **8**, 74.
68. Pesole,G. and Liuni,S. (1999) Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. *Trends Genet*., **15**, 378.
69. Day,D.A. and Tuite,M.F. (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J. Endocrinol*., **157**, 361–371.
70. Majdalani,N., Cunning,C., Sledjeski,D., Elliott,T. and Gottesman,S. (1998) DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc. Natl Acad. Sci. USA*, **95**, 12462–12467.
71. Novick,R.P., Ross,H.F., Projan,S.J., Kornblum,J., Kreiswirth,B. and Moghazeh,S. (1993) Synthesis of staphylococcal virulence factors is controlled by a regulatory RNA molecule. *EMBO J*., **12**, 3967–3975.
72. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
73. Furlong,R.F. and Yang,Z. (2003) Comparative genomics coming of age. *Heredity*, **91**, 533–534.
74. Pang,K.C., Frith,M.C. and Mattick,J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*., **22**, 1–5.
75. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*., **11**, 1005–1017.