Cell Research (2008) 18:695-700.

www.nature.com/cr

ORIGINAL ARTICLE

Finding noncoding RNA transcripts from low abundance expressed sequence tags

Chenghai Xue[1, 2, *], Fei Li[1, 3, *], Fei Li[1, §]

[1]Department of Entomology, Nanjing Agricultural University, Nanjing 210095, China; [2]MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China; [3]The First Hospital of Tsinghua University, Beijing 10084, China

It has been proved that noncoding RNA (ncRNA) genes are much more numerous than expected. However, it remains a difficult task to identify ncRNAs with either computational algorithms or biological experiments. Recent reports have suggested that ncRNAs may also appear in the expressed sequence tags (EST's) database. Nevertheless, intergenic ESTs have received little attention and are poorly annotated owing to their low abundance. Here, we have developed a computational strategy for discovering ncRNA genes from human ESTs. We first collected ESTs that are located in the intergenic regions and do not have detailed annotations. The intergenic regions were divided into non-overlapping 50-nt windows and PhastCons scores obtained from the UCSC database were assigned to these windows. We kept conserved windows that had PhastCons scores of over 0.8 and that had at least three supporting ESTs to act as seeds. Each cluster of ESTs corresponding to the seeds was assembled into a long contig. We used two criteria to screen for ncRNA transcripts from these contigs: the first was that the longest predicted open reading frame was less than 300 nt and the second was that the likely Pol-II promoters exist within 2 000 nt upstream or downstream of the contigs. As a result, 118 novel ncRNA genes were identified from human low abundance ESTs. Of seven randomly selected candidates, six were transcribed in human 2BS cells as shown by RT-PCR. Our work proves that the EST is a 'hidden treasure' for detecting novel ncRNA genes.

***Keywords*:** ncRNA, EST, computational identification, RT-PCR

Cell Research (2008) **18**:695-700. doi: 10.1038/cr.2008.59; published online 27 May 2008

Introduction

Noncoding RNA (ncRNA) has received increasing attention because it has a diverse range of functions and participates in many biological pathways [1, 2]. Recent transcriptome analysis of the human genome showed that the number of expressed transcripts is remarkably higher than expected from protein-coding sequences. A large number of transcripts are outside any known gene regions [3, 4], which implies that ncRNA genes are widely distributed in the genome. However, identification of ncRNA genes is still a difficult task, partially owing to the lack of common features [5]. For example, ncRNAs do not have apparent open reading frames (ORFs) and codon information. This makes the existing gene-finding algorithms fail to identify ncRNA genes. In addition, the current biological method for discovering novel ncRNA genes is still inefficient and costly.

There are lots of expressed sequence tags (ESTs) located in un-annotated intergenic regions, and these are mostly expressed at low levels [3, 4]. These low abundance ESTs have traditionally been considered unimportant 'noise' transcripts [6]. Most gene-finding algorithms did not analyze these ESTs because of their unreliability and their lack of ORFs [7, 8]. Recently, the analysis on a *Drosophila* cDNA collection revealed that some noncoding transcripts are polyadenylated and appear in the cDNA databases [9]. In *Arabidopsis*, ESTs have also been used in retrieving ncRNAs from the genome [10]. These pieces of evidence indicate that ncRNA transcripts might exist in the ESTs.

Here, we have developed a computational strategy for identifying novel ncRNA transcripts from intergenic ESTs in human. The reliability of the predicted ncRNA transcripts

*These two authors contributed equally to this work.
Correspondence: Fei Li§
Tel/Fax: +86-25-84399025
E-mail: lifei@njau.edu.cn
Received 11 February 2007; revised 2 July, 15 October 2007; accepted 21 December 2007; published online 27 May 2008

has been improved through comparative genomic analysis and promoter prediction. One hundred and eighteen contigs were predicted to be putative ncRNA transcripts. In addition, six of seven randomly selected candidates were verified using RT-PCR experiments in human 2BS cells.

Results

Intergenic regions matched with low abundance ESTs

Although most ESTs were aligned to protein-coding genes, some low abundance ESTs could be perfectly matched to intergenic regions in the human genome, which we named as intergenic ESTs. To screen ncRNA transcripts in the intergenic ESTs, a sliding window of 50-nuclotides (nt) was used to scan the human genome without any overlap. The number of ESTs matched with each 50-nt window was counted. As expected, few ESTs match with intergenic windows, whereas much more are aligned to the windows in exon regions (Figure 1). In total, 1 101 439 of the windows perfectly matched at least one EST in the intergenic regions. To improve reliability, only those intergenic windows that

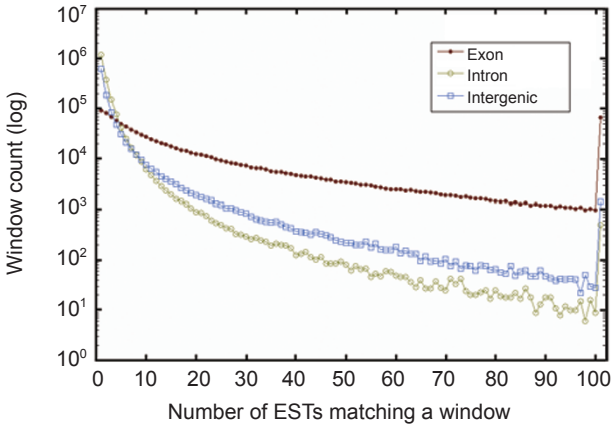

Figure 1 A sliding window of 50-nuclotides (nt) was used to scan the human genome without allowing any overlap between windows. The number of ESTs matching each given 50-nt window was counted separately. The *X*-axis represents the number of ESTs matching a 50 nt window and the *Y*-axis is the corresponding count to each *X*-axis value. The windows matching more than 100 ESTs are counted together. The results indicate that there are many ESTs that can be matched perfectly to intergenic regions.

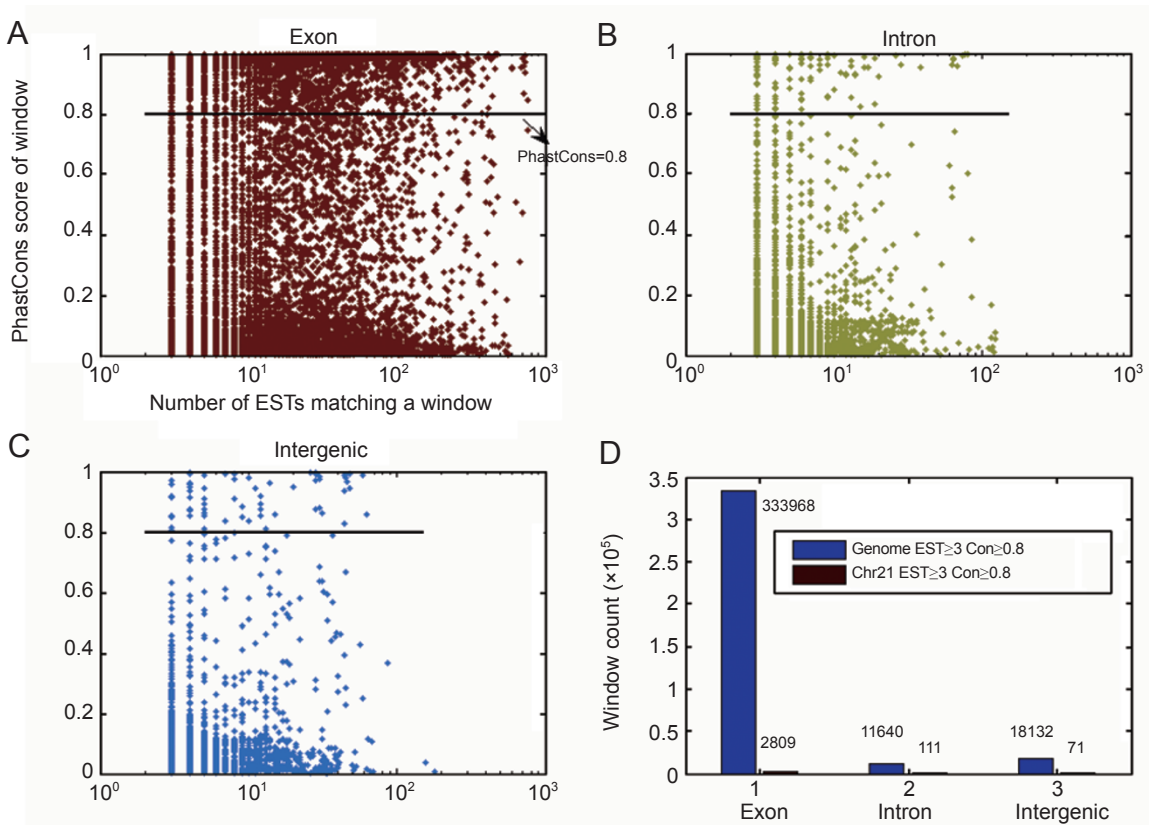

Figure 2 Supporting ESTs and PhastCons scores for the 50-nt windows. The data from chromosome 21 are presented as an example in **(A**-**C)**. The PhastCons score is the average score of 50 positions in the window. The conserved windows (PhastCons score ≥ 0.8, indicated by a black horizontal line) are much more abundant in exons **(A)** than in introns **(B)** and intergenic regions **(C)**. Intergenic conserved windows with ≥ 3 supporting ESTs were kept as candidates. The numbers of 50-nt windows meeting the above criteria from the whole genome are summarized in **(D)**.

are supported by at least three ESTs were kept, which gave us 288 277 candidate windows.

Conservation of intergenic ESTs across different species

The PhastCons score from the UCSC annotation database was used to estimate the sequence conservation of all 288 277 intergenic windows. We defined the PhastCons score as the average score of 50 positions in a given window [11]. If the PhastCons score was greater than 0.8, the corresponding intergenic window was used as a 'seed window' for further analysis. Figure 2 shows data from human chromosome 21 as an example. In total, there were 18 132 'seed windows' in the human genome that satisfied the above criteria (Figure 2D).

Electronic elongation of putative ncRNA transcripts

To get putative ncRNA transcripts that were as long as possible, all 18 132 seed windows were used for electronic elongation. Each cluster of ESTs corresponding to a given seed window was assembled as a contig, which was the longest putative ncRNA transcript. This produced 3 457 potential ncRNA transcripts, which did not overlap with the RefGene and the KnownGene annotations.

Screening novel ncRNA transcripts using ECgene annotation

Although we did not use intronic ESTs in this work, it was possible that these predicted intergenic transcripts were from novel 5′-end or 3′-end exons of protein-coding genes.

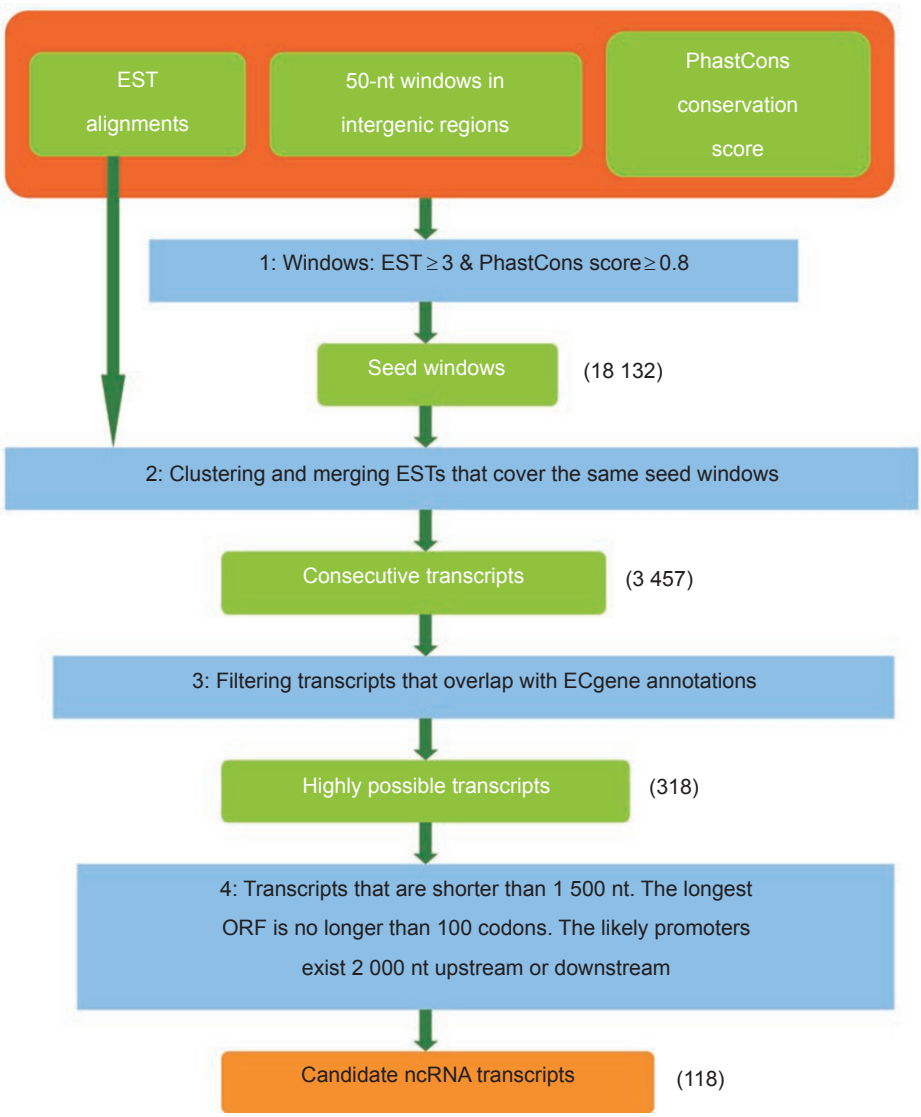

Figure 3 Computational pipeline for identifying ncRNA genes from low abundance ESTs.

In order to exclude this possibility, ECgene annotations at a medium confidence level were used to filter any transcripts that overlapped with alternative spliced regions or alternative transcription start/alternative poly A sites. This gave us 318 potential ncRNA transcripts.

Putative ncRNA transcripts with probable promoters

We reasoned that most of the ncRNA transcripts that were predicted from ESTs were transcribed from Pol-II promoters. Therefore, we predicted the presence of transcription starting sites and core promoter regions within 2 000 nt upstream or downstream of potential ncRNA transcripts using Promoter 2.0. Only those with probable promoters remained.

Two additional criteria were also used: the first was that the length of the predicted ncRNA transcripts was less than 1 500 nt, and the second was that the length of any predicted ORF was not more than 300 nt. As a result, 118 novel ncRNA transcripts satisfying these stringent criteria were obtained (Supplementary information, Table S1). The detailed computational procedure is shown in Figure 3. We also calculated the distance between predicted ncRNAs and their neighboring annotated exons. The average distances were 66 109 bp according to the RefGene annotation and 49 585 bp according to the KnownGene annotation.

Validation of putative noncoding transcripts

Two strategies were used to validate the predicted ncRNA transcripts. First, we compared our results with recent transcriptome data from 10 human chromosomes. All available information on transcribed fragments (transfrags) obtained from tiling array experiments was downloaded from the UCSC database. Of the 118 putative ncRNAs, 36 transcripts were located in the 10 chromosomes selected for tiling array analysis. Also, 23 of the 36 predicted ncRNA transcripts (63.8%) were also detected by the tiling array (Supplementary information, Table S2) [12].

RT-PCR experiments were used to verify our results. Primers designed from neighboring exons of a human housekeeping gene were used as a control to ensure that there was no genomic DNA contamination. Seven putative transcripts, including two candidates filtered by ECgene annotation, were selected for verification, of which six were successfully detected by PCR (Table 1 and Figure 4). The PCR products of candidates E1 (E2), ncR118 and ncR95 were sequenced.

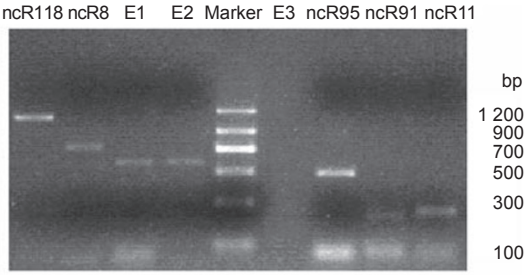

Figure 4 PCR results from the experimental validation of eight predicted ncRNA genes in the human 2BS cell. The ncRNA transcripts were amplified to their maximum length, as the PCR primers were designed at both ends of clustered transcripts. Three candidates (E1, E2 and E3) were filtered using the ECgene annotation. E1 and E2 are repeat sequences.

Table 1 Eight candidate ncRNA transcripts used for RT-PCR validation

Candidate ncRNAs	Chr no	Matched genome start	Matched genome end	Length of prediction	Length of PCR products	Result of PCR	Forward primer (5′→3′)	Reverse primer (5′→3′)
ncR118	X	134768266	134769481	1216	1088	Amplified	gcaaggcaaaatctcagaagc-tgtaaagac	ggaaatcgttatgattaaagc-atcaaggaa
ncR8	1	151484199	151485258	1060	730	Amplified	cattgctaatctccttgaaagcc-aattctc	gtgtggagaaaactgggcac-tagactgaac
E1	1	141416885	141417509	625	575	Amplified	gtgcagagcctttgcttcagtaa	cacagtcgaatgtgtctccattt
E2	1	120502675	120503298	624	575	Amplified	gtgcagagcctttgcttcagtaa	cacagtcgaatgtgtctccattt
E3	7	123264778	123265466	689	–	No	aagcctggcccaaggactctgg	ttagattaaaaagaaagtaaaa
ncR95	15	94689348	94690137	790	492	Amplified	ggtctccgagtgtgacagtact-cagcatag	caagccaaagcagagatacc-tgttatcaaa
ncR91	15	19185890	19186343	454	232	Amplified	tatcacaacaagaaaaaaaag-ccaatgcgt	atgctgtggtagacgaggttg-gagttagaa
ncR11	2	29325508	29325991	484	311	Amplified	agttgataccaataaggacga	tatttaactcacagaccatctg-aaggggctc

Three candidates (E1, E2, E3) that overlapped with the ECgene annotations.

Discussion

There are huge numbers of ESTs available for mammals, insects, nematodes and plants. Although most ESTs are derived from protein-coding genes, there are still lots of ESTs that are mapped to intergenic regions without detailed annotation. Here, we performed a large-scale analysis of intergenic ESTs to screen for novel ncRNA transcripts. Because most intergenic ESTs are expressed at low levels, sequence conservation and promoter prediction were adopted to increase the reliability of our results. Some predicted ncRNA genes were confirmed by either tiling array transcriptome data or RT-PCR experiments. The transcripts were successfully detected from cDNA synthesized with oligo (dT) as the anchor primers, suggesting that most, if not all, ncRNA genes are transcribed by RNA polymerase II.

It is very likely that some real ncRNA transcripts were filtered out in our work, as only 118 ncRNA transcripts were discovered from more than five million ESTs. This is probably due to the strict criteria used, which improve the reliability but also reduce the sensitivity. For example, we required that the average PhastCons score should be larger than 0.8. This removed a large number of the intergenic ESTs and only kept 6.3% (18 132/288 277) seed windows. However, it has been reported that some ncRNAs are only structurally conserved and their primary sequences share low similarities [13-16]. Therefore, if we adjusted the criteria or performed a structural conservation analysis, more ncRNA transcripts would be discovered. Actually, there were 26 predicted ncRNAs overlapping with previous results based on structural analysis [16] (Supplementary information, Table S1). We suggest that the actual selection criteria should depend on the purpose of the work. In this paper, we have focused on introducing a computational strategy.

ESTs have been well studied for their role in discovering protein-coding genes, but they are seldom used for ncRNA transcript analysis. Our work proves that ESTs could be a 'hidden treasure' for studying ncRNA transcripts. Although tiling array analysis and other genome-scale transcriptome analysis has become useful in identifying noncoding genes, it still remains time-consuming and is costly because most ncRNAs genes are spatiotemporally expressed. Together with comparative genomics analysis, the method we propose is a helpful strategy to exploit large amounts of EST data for discovering ncRNA genes.

Materials and Methods

EST alignment

We used EST alignment resources between human ESTs and the human genome from the UCSC database, which contained 5 977 963 EST alignment entries (hg17, May 2004). To improve the reliability of our analysis, the following ESTs were removed: ESTs that were aligned to multiple regions in the human genome and ESTs that shared less than 90% sequence similarity with the human genome. In total, 3 946 573 EST entries remained.

Intergenic ESTs

The intergenic regions were defined according to the RefGene and the KnownGene tables in the UCSC annotation database [13]. Overlapping exons and alternatively spliced variants were merged into a single consecutive region. Then, non-overlapping introns and intergenic boundaries were obtained accordingly. All intergenic regions were divided into 50-nt windows that were not overlapping. The number of ESTs that matched each 50-nt window was counted.

PhastCons score

To estimate the conservation of a given 50-nt window, we used the UCSC PhastCons conservation score in an 8-way vertebrate alignment of human, chimp, mouse, rat, dog, chicken, fugu and zebrafish [13]. The average score of 50 positions in the window was defined as the PhastCons score. Human alignment data (hg17) were downloaded from the UCSC website, http://hgdownload.cse.ucsc.edu/goldenPath/hg17/ [13].

Screening candidates

Since we used low abundance ESTs, four strict criteria were chosen to improve the reliability: (1) that at least one 50-nt window could perfectly match with three ESTs or more; (2) that the 50-nt window had a PhastCons score of over 0.8 (for brevity, we defined the 50-nt window that satisfied the above two criteria as the seed window); (3) that a highly likely promoter could be predicted by the Promoter 2.0 software to lie within 2 000 nt upstream or downstream of the predicted ncRNA transcripts; and (4) that the length of any predicted ORF was not more than 300 nt. The ESTs that satisfied the above criteria were kept as candidates.

Filtering known ncRNA genes

Because the ECgene data set contained more gene annotation information than the RefGene and the KnownGene databases, we used the ECgene annotation database at a medium confidence level (version 1.2, hg17) to further filter any known transcripts, such as predicted ncRNAs and alternative spliced regions [14, 15]. The candidate ESTs were removed if they had been already annotated in the ECgene data set.

RT-PCR experiments

Total RNA was isolated from human 2BS cells with TRIzol reagent (Life Technologies, Gaithersburg, MD) following the manufacturer's instructions. RNA was treated with Dnase I enzyme following the standard procedure. Primers that were designed from neighboring exons of a human housekeeping gene were used to detect genomic DNA contamination. If there was genomic DNA in the template, two bands were amplified. One was from cDNA that does not have introns, and the other was from DNA that has introns. First-strand cDNA was synthesized from total RNA using SuperScriptTMII RNase H-Reverse Transcriptase (Invitrogen) with oligo (dT) as anchor primers. PCR analysis was performed in accordance with standard procedures with 2 μM of each primer and 2 U ex-Taq DNA polymerase (Takara Corporation). The products were resolved

by electrophoresis on 1% w/v agarose gel in TAE buffer (40 mmol/L Tris-acetate, 2 mmol/L $Na_2EDTA•2H_2O$) and stained with GlodenView. DNA bands were viewed using a UVP-GDS-8000 system UV transilluminator and the Lab-Works program (UVP, Inc). Some PCR results were sequenced for validation. The primer sequences are listed in Table 1.

Acknowledgments

The authors wish to thank Professor Xuegong Zhang at the Tsinghua University for intriguing discussions. Ms Hongyan Han at the Military Medical Academy kindly provided human 2BS cells. We also thank Meisch Francoise for suggestions on writing. This work is in part supported by the National Science Foundation of China (60405001, 60702002, 30771417), and the Natural Science Foundation of Jiangsu Province (BK2007524), the China Postdoctoral Science Foundation (20060400060) and the program of New Century Excellent Talents (NCET) to Fei Li.

References

1 Storz G. An expanding universe of noncoding RNAs. *Science* 2002; **296**:1260-1263.
2 Moulton V. Tracking down noncoding RNAs. *Proc Natl Acad Sci USA* 2005; **102**:2269-2270.
3 Kampa D, Cheng J, Kapranov P, *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004; **14**:331-342.
4 Kapranov P, Cawley SE, Drenkow J, *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002; **296**:916-919.
5 Eddy SR. Computational genomics of noncoding RNA genes. *Cell* 2002; **109**:137-140.
6 Lee S, Bao J, Zhou G, *et al.* Detecting novel low-abundant transcripts in Drosophila. *Rna* 2005; **11**:939-946.
7 Imanishi T, Itoh T, Suzuki Y, *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2004; **2**:e162.
8 Okazaki Y, Furuno M, Kasukawa T, *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 2002; **420**:563-573.
9 Tupy JL, Bailey AM, Dailey G, *et al.* Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 2005; **102**:5495-5500.
10 MacIntosh GC, Wilkerson C, Green PJ. Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol* 2001; **127**:765-776.
11 Karolchik D, Baertsch R, Diekhans M, *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* 2003; **31**:51-54.
12 Cheng J, Kapranov P, Drenkow J, *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 2005; **308**:1149-1154.
13 Nakaya HI, Amaral PP, Louro R, *et al.* Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol* 2007; **8**:R43.
14 Pedersen JS, Bejerano G, Siepel A, *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2006; **2**:e33.
15 Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* 2006; **16**:885-889.
16 Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 2005; **23**:1383-1390.

(**Supplementary information** is linked to the online version of the paper on the Cell Research website.)