DNA Research 17, 169–183, (2010)
Advance Access Publication: 17 April 2010

doi:10.1093/dnares/dsq007

Characterization of Transcription Start Sites of Putative Non-coding RNAs by Multifaceted Use of Massively Paralleled Sequencer

Nuankanya Sathira[1,†], Riu Yamashita[2,†], Kousuke Tanimoto[1], Akinori Kanai[1], Takako Arauchi[1], Soutaro Kanematsu[1], Kenta Nakai[2], Yutaka Suzuki[1,*], and Sumio Sugano[1]

Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan[1] and Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan[2]

*To whom correspondence should be addressed. Tel. +81 4-7136-3607. Fax. +81 4-7136-3607.
E-mail: ysuzuki@hgc.jp

Edited by Katsumi Isono
(Received 24 December 2009; accepted 4 February 2010)

Abstract

On the basis of integrated transcriptome analysis, we show that not all transcriptional start site clusters (TSCs) in the intergenic regions (iTSCs) have the same properties; thus, it is possible to discriminate the iTSCs that are likely to have biological relevance from the other noise-level iTSCs. We used a total of 251 933 381 short-read sequence tags generated from various types of transcriptome analyses in order to characterize 6039 iTSCs, which have significant expression levels. We analyzed and found that 23% of these iTSCs were located in the proximal regions of the RefSeq genes. These RefSeq-linked iTSCs showed similar expression patterns with the neighboring RefSeq genes, had widely fluctuating transcription start sites and lacked ordered nucleosome positioning. These iTSCs seemed not to form independent transcriptional units, simply representing the by-products of the neighboring RefSeq genes, in spite of their significant expression levels. Similar features were also observed for the TSCs located in the antisense regions of the RefSeq genes. Furthermore, for the remaining iTSCs that were not associated with any RefSeq genes, we demonstrate that integrative interpretation of the transcriptome data provides essential information to specify their biological functions in the hypoxic responses of the cells.

Key words: non-coding RNA; integrated transcriptome analysis; transcriptional start site cluster (TSC); intergenic transcript; antisense transcript

1. Introduction

Since human genome and several other organisms' genomes are completely sequenced, gene regulation is realized to be more immense and more complex than expected. The functional properties are known not only on protein-coding sequences but also on non-coding or untranslated sequences. Recent cDNA studies[1,2] suggested that transcription start sites (TSSs) are widespread throughout the human genomes,[3,4] and ENCODE consortium reported that major part of such TSSs should produce long non-coding RNAs (lncRNAs),[3] especially from intergenic regions.[4–7] Furthermore, a large number of putatively lncRNAs were discovered from the antisense regions of the protein-coding genes.[8,9] Although tens of thousands of non-coding RNAs (ncRNAs) have been identified in mammalian cells,[10–13] their identification, regulation mechanisms and functional characterization are still incomplete.[14–17]

With the growing number of the uncharacterized lncRNAs, concerns have been raised that they may be mere transcriptional noise of mammalian cells.[18–22] A large number of lncRNAs are not evolutionarily conserved, and their expression levels seem

[†] These authors contributed equally to this work.

© The Author 2010. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

too low to imply any biological functions, frequently below the limit of the northern blotting analysis.[23–25] Nonetheless, a recent study demonstrated that the lncRNAs with potential biological significance could be selected by examining chromatin signatures,[5] proving that at least some of the lncRNAs do have biological relevance.

Although information about millions of TSSs has become available,[1,26,27] it is still insufficient to represent comprehensive overview of intergenic TSSs in a quantitative manner. In order to understand the nature of intergenic TSSs of unknown functions, further in-depth and wider variety of biological data, focusing on a particular cell type or a cellular response in a more quantitative manner, are essential.

We have been collecting and analyzing full-length cDNAs of human genes by utilizing our cap selection method, oligo-capping.[26–28] Recently, to increase the throughput of the cDNA sequencing, we combined oligo-capping with the massively paralleled sequencing technology, Illumina GA.[29] In this method, named TSS Seq, sequence adaptor, which is necessary for Illumina GA sequencing, is directly introduced to the cap site of the mRNA (TSS tag).[30] In addition to TSS Seq, various high-throughput analyses have been enabled by using the massively paralleled sequencer. It is now possible to perform genome-wide analyses to examine nucleosome structure (Nucleosome Seq),[31,32] binding of a transcription factor (ChIP Seq),[33] complete cDNA sequences (cDNA shotgun), identification and characterization of RNAs in a particular subcellular locations or an experimental condition (RNA Seq)[34,35] with feasible cost and efforts.

In our previous study, we identified 371 849 intergenic TSS clusters (iTSCs) that were identified by the TSS Seq analysis of 12 human cell lines and tissues. We found that, among these TSS clusters, only 6039 are having significant expression levels (R. Yamashita *et al.*, submitted). In this paper, we report our first integrative transcriptome analyses of the intergenic TSS using a total of 254 076 723 short-read sequences (Supplementary Fig. S1). We show that not all intergenic TSSs have the same properties; thus, it is possible to classify them and discriminate intergenic TSSs which are likely to have biological relevance from the other noise-level TSSs (Fig. 1).

2. Materials and methods

2.1. Cell culture and tissues

Human cell lines, DLD-1 cells (ATCC number: CCL-221), were maintained in Dulbecco's modified Eagle's medium (DMEM; Invitrogen) supplemented with 10% fetal calf serum, 4.5 g/l glucose and antibiotics. HEK293, MCF7, TIG3, BEAS2B and Ramos cells (ATCC number: CRL-1573, ATCC number: HTB-22 and Japan Cell Resource Bank number: JCRB0506) were cultured in standard conditions. For hypoxic induction, DLD-1, HEK 293, MCF7 and TIG3 cells were incubated in 1% O_2 and 5% CO_2 for 24 h. In order to get the RNA, six million cells were cultured and harvested for RNA extraction using RNeasy (Qiagen). The quality and quantity of the obtained total RNAs were assessed, using BioAnalyzer (Agilent). Human tissue RNAs (brain, heart, kidney, fetal brain, fetal heart, fetal kidney) were purchased from Clontech (catalog numbers: 636526, 636529, 636530, 636532, 636583 and 636584).

2.2. Construction of the TSS Seq libraries and analysis of the TSS tags

Two hundred micrograms of the obtained total RNA were subjected to oligo-capping; after the successive treatment of the RNA with 2.5 U BAP (TaKaRa) at 37°C for 1 h and 40 U TAP (Ambion) at 37°C for 1 h, the BAP-TAP-treated RNAs were ligated with 1.2 μg of RNA oligo(5′-AAUGAUACGGCGACCACCGA GAUCUACACUCUUUCCCUACACGACGCUCUUCCGAU CUGG-3′) using 250 U T4 RNA ligase (TaKaRa) at 20°C for 3 h. After the DNase I treatment (TaKaRa), polyA+ RNA was selected using oligo-dT powder (Collaborative). First-strand cDNA was synthesized from 10 pmol of random hexamer primer (5′-CAAG CAGAAGACGGCATACGANNNNNNC-3′) using Super Script II (Invitrogen) by incubating at 12°C for 1 h and at 42°C overnight. Template RNA was degraded by alkaline treatment. For polymerase chain reaction (PCR), one-fifth of the first-strand cDNAs were used as the PCR template. Gene Amp PCR kits (PerkinElmer) were used with the PCR primers 5′-AATGATACGGCGACCACCGAG-3′ and 5′-CAAGCAGAA GACGGCATACGA-3′ under the following reaction conditions: 15 cycles of 94°C for 1 min, 56°C for 1 min and 72°C for 2 min. The PCR fragments were size fractionated by 12% polyacrylamide gel electrophoresis and the fraction of 150–250 bp was recovered. The quality and quantity of the obtained single-stranded first-strand cDNAs were assessed, using BioAnalyzer (Agilent). One nanogram of the size-fractionated cDNA was used for the sequencing reactions with the Illumina GA. The sequencing reactions were performed according to the manufacturer's instructions.

Thirty-six-base pair-read TSS tags were mapped onto the human genome sequence (hg18 as of UCSC Genome Browser) using ELAND. Uniquely and completely (without any mismatch) mapped TSS tags were used for the analysis. TSSs located from the 3′-end boundaries to −50 kb of the 5′-end

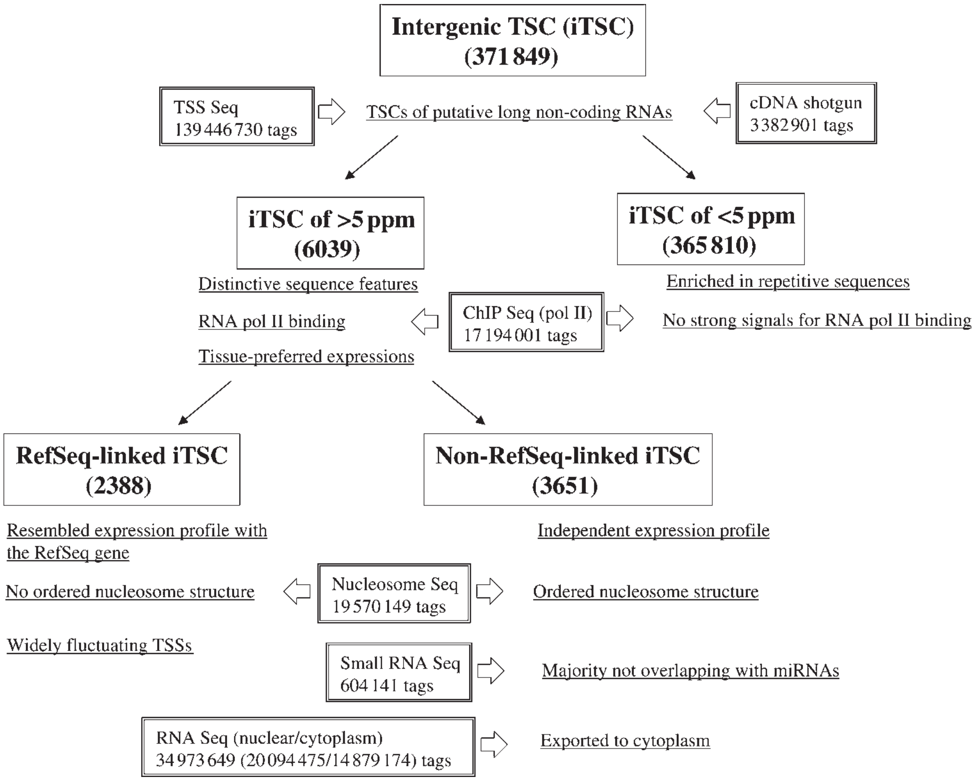

Intergenic TSC (iTSC)
(371 849)
TSS Seq
139 446 730 tags
TSCs of putative long non-coding RNAs
cDNA shotgun
3 382 901 tags
iTSC of >5 ppm
(6039)
iTSC of <5 ppm
(365 810)
Distinctive sequence features
RNA pol II binding
Tissue-preferred expressions
ChIP Seq (pol II)
17 194 001 tags
Enriched in repetitive sequences
No strong signals for RNA pol II binding
RefSeq-linked iTSC
(2388)
Non-RefSeq-linked iTSC
(3651)
Resembled expression profile with the RefSeq gene
Independent expression profile
No ordered nucleosome structure
Nucleosome Seq
19 570 149 tags
Ordered nucleosome structure
Widely fluctuating TSSs
Small RNA Seq
604 141 tags
Majority not overlapping with miRNAs
RNA Seq (nuclear/cytoplasm)
34 973 649 (20 094 475/14 879 174) tags
Exported to cytoplasm

Figure 1. Classification of the iTSCs. Schematic representation of the classification of the iTSCs is shown. Number of the iTSCs belonging to each category is shown in parenthesis. Characteristic features of each category of the iTSCs are shown in the margin. Analysis methods to identify them are also shown in the double-line box and the number of the sequence tags used is shown in the parenthesis.

boundaries of the adjoining RefSeq genes were selected. Then, the clustering of the intergenic TSS tags into 500-bp bin was performed. TSS clusters were removed when all the belonging TSS tags were located at the internal exonic region of the corresponding RefSeq genes. The RefSeq information as to the genomic coordinate, position of the protein-coding region and so on is as of hg18 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/). Gene ontology terms (as of 1 June 2009) were correlated with RefSeq using loc2go GO (as of 1 June 2009) using NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene).

2.3. Construction of the Nucleosome Seq library and analysis of the nucleosome tags

DLD-1 cells were cultured at 80% confluency and were treated with micrococcus nuclease to generate mononucleosomes using ChIP-IT Express Enzymatic (Active motif). The cells were formaldehyde-cross-linked before isolating nucleosomes. Cross-linking was achieved with fixation solution (DMEM, 0.01% formaldehyde) for 10 min at room temperature. The cells were then washed with PBS, and cross-linking was stopped with Glycine Stop-Fix Solution (PBS, Glycine buffer) for 5 min at room temperature. After washing with PBS, ice-cold cell scraping solution was added to the dish and cells were harvested. The cells were re-suspended in ice-cold lysis buffer, were incubate on ice for 30 min and homogenized with 15–20 strokes with a pestle. The nuclei were suspended in digestion buffer and pre-warm this solution for 5 min at 37°C. Enzymatic cocktail dissolved (200 U/ml) was added to the sample and the digestion reaction was incubated at 37°C for 15 min. The reaction was stopped with EDTA, the nuclei were pelleted by centrifugation and the supernatants were collected. Formaldehyde cross-linking was reversed by adding 5 M NaCl, RNase and incubation at 65°C for over 4 h. Proteinase K was added and incubated at 42°C for 1.5 h. DNA was purified by phenol/chloroform extractions and ethanol precipitation. Using the extracted DNA, sample preparation for Illumina GA was made according to the manufacturer's instructions.

2.4. Shotgun sequencing of the cDNA

Colony PCR was performed for the cDNA clones using the common PCR primers of 5′-TCAGTGGATGT TGCCTTTAC-3′ and 5′-TGTGGGAGGTTTTTTCTCTA-3′. Obtained PCR products were nebulized so that the average fragments size is 200–500 bp. Sample preparation for Illumina GA was made according to the manufacturer's instructions. Using the generated 36-bp sequence tags, cDNA sequences were assembled based on the mapping information to the human genome sequence. Details of the computational procedure for the assembling and their quality are described elsewhere. Since the quality of the base-call varied, only the boundary information of the exons was extracted and actual nucleotide sequences were replaced with those of the reference human genome.

2.5. *Subcellular fractionation*

A total of 1 × 10^8 DLD-1 cells were incubated with the medium containing 0.1 mg/ml cycloheximide for 5 min at 37°C and washed with PBS (containing 0.1 mg/ml cycloheximide). Cell pellets were re-suspended in lysis buffer [20 mM Tris–HCl (pH 7.5), 150 mM NaCl, 15 mM $MgCl_2$, 1% Triton X-100, 0.1 mg/ml cycloheximide, 0.1 mM dithiothreitol, RNase inhibitor, Complete Protease Inhibitor Cocktail (Roche)] and lysed on ice for 10 min. Lysate was separated into cytoplasmic fraction (supernatant) and nuclear fraction (pellet) by centrifugation. Nuclear pellet was re-suspended in lysis buffer, homogenized and lysed on ice for 10 min. A portion of cytoplasmic fraction was layered on the top of a 10-ml 15–50% (w/v) sucrose gradient and centrifuged at 100 000 g in Beckman SW41Ti rotor for 2 h 15 min at 4°C. Cytoplasmic and nuclear fractions were treated with 200 μg/ml proteinase K, and RNA was extracted by TRIzol LS (Invitrogen). Obtained RNA concentration was analyzed by 2100 Bioanalyzer (Agilent Technologies). For validation of the precise separation of the subcellular fractions, see Supplementary Fig. S8.

2.6. *Quantitative reverse transcription–PCR analysis*

Twenty kinds of normal human tissues were selected for expression analysis, both fetal and adult major organs, included adrenal gland, bone marrow, brain (whole), fetal brain, fetal liver, heart, kidney, liver, lung (whole), placenta, prostate, salivary gland, skeletal muscle, testis, thymus, thyroid gland, trachea, uterus, colon w/mucosa and spinal cord. Total RNAs were purchased from Clontech (Human total RNA Master Panel II). The qualitative assessment of all total RNA was done utilizing the Agilent 2100 Bioanalyzer. Total RNA (8 μg) from each sample was reverse transcribed using an oligo(dT) and the Superscript II reverse transcriptase kit (Invitrogen). Negative control samples (no first-strand synthesis) were prepared by performing reverse transcription reactions in the absence of reverse transcriptase. Expression profile was performed using the quantitative reverse transcription–PCR (qRT–PCR). Gene-specific primer pairs were designed using Primer3 software, with an optimal primer size of 20 bp, amplification size of 100–500 bp and annealing temperature of 55°C. The primers were purchased from Operon. Quantitative real-time PCR was carried out with 100 pg of total RNA per test well using the Power SYBR Green PCR Master Mix (Applied Biosystem). The PCR reactions were performed using an ABI 7900HT Fast Real-Time System (Applied Biosystems) using the following cycling protocols: 40 cycles of 15 s at 95°C and 60 s at 60°C. The threshold cycle (Ct) value was calculated from amplification plots, in which the fluorescence signal detected was plotted against the PCR cycle. We defined express sequence when the threshold cycle values (Ct) <37 in the presence of reverse transcriptase, but not when reverse transcriptase was absent. The product size was checked by 2% agarose gel electrophoresis, and melting curves were analyzed to monitor the specificity of the PCR.

2.7. *Computational procedures*

Overlap between the iTSCs and miRNA and snoRNA, from miRBase (http://microrna.sanger.ac.uk/sequences/index.shtml) and snoRNABase (http://www-snorna.biotoul.fr/index.php), respectively, were examined. Statistical significance of the difference in the sequence features around the iTSCs was evaluated by the indicated methods.

Occupancy of the nucleosome was calculated according to the reference.[32] First, we defined the positions of nucleosome centers i by adding 75 bp from the 5′-end of each mapped nucleosome tag. The counts of nucleosome centers $c(i)$ were converted to the nucleosome signals $s(p_j)$ throughout the whole genome using the logarithm of a weighted average as follow:

$$s(p_j) = \log_2\left[\frac{\sum_{i=p_j-75}^{p_j+75} w(i)c(i)}{\sum_{i=p_j-75}^{p_j+75} w(i)} + 1\right],$$

where $p_j = 5 + 10j$ ($j = 0, 1, 2, \ldots$) and $w(i)$ is the Gaussian distribution with mean p_j and standard deviation 20.

Enrichment of the RNA Seq tags in the respective subcellular fractions was similarly evaluated using the Poisson distribution:

$$p(x, \lambda) = 1 - \sum_{t=0}^{x} \frac{e^{-\lambda}\lambda^t}{t!},$$

where $p(x, \lambda)$ is the probability of enrichment, λ the expected tag number in the cytoplasmic fraction for each gene based on the tag number in the total fraction, and *x* the observed tag number in the cytoplasmic fraction for each gene.

3. Results and discussion

3.1. *Identification and initial characterization of iTSCs*

Schematic representation of the TSS Seq and statistics of the 139 446 730 thirty-six-base pair TSS tags, which were collected from 12 cell lines and tissues of humans and used in this study, are shown in Supplementary Fig. S1 (GenBank accession number for each of the TSS tag data set is also shown there). All of the TSS information is publicly available from our database (http://dbtss.hgc.jp). The TSS tags that were mapped to the intergenic regions, from 3′-end boundaries to −50 kb of the 5′-end boundaries of the adjoining RefSeq genes (Fig. 2), were selected. The 5′-end of the RefSeq genes was mixed with 50-kb upstream regions to remove the possible alternative promoters of the RefSeq genes (Supplementary Fig. S1E). Then, the selected TSSs were clustered into 500 bp bins (iTSCs) in every cell type. The clustering analysis revealed that the numbers of iTSCs with >5 ppm TSS tags were only around 500–2000 depending on cell types. Taken all TSS data from each cell types together, there were 371 849 iTSCs; however, the number of iTSCs with >5 ppm TSS tags at least in one cell type was only 6039. We tentatively selected iTSCs of >5 ppm, since 5 ppm is supported to be roughly corresponding to 5 copies/cell, assuming that each cell has 1 million mRNA copies. We considered that such a transcript level should be necessary to robustly realize biological functions (see Supplementary Fig. S2, for further detailed discussion).

3.2. *Complete sequencing of the transcripts of the iTSCs*

For both iTSCs of >5 ppm and those of <5 ppm, we characterized what transcripts were transcribed from the identified iTSCs. We selected completely sequenced full-length DNA clones which overlapped the iTSCs from MGC and FLJ cDNA collections. Among the total 371 849 iTSCs, there were 395 and 1617 iTSCs which overlapped the 5′-ends of the MGC and FLJ cDNAs, respectively (Supplementary Fig. S3A). We examined the cDNA sequences and found that about 60% of the cDNAs contained open reading frames (ORFs) of no more than 100 amino acids. When including the cDNAs which were possible targets for the nonsense mRNA decay or had 5′-untranslated region of >750 bases, 85% of them had features which may hamper efficient translation and, thus, were likely to be ncRNAs.

In order to exclude the possible selection bias of the cDNA clones in the respective cDNA projects (note these cDNA projects aimed to select protein-coding transcripts), we selected and determined the complete sequences of the cDNA from our cDNA collection anew. In order to expedite the sequencing, we employed shotgun sequencing of the cDNA using Illumina GA. Details of the computational procedure for the assembling of the cDNA sequences are described in the ‘Materials and methods’ section. As a result, 361 and 103 cDNAs which correspond to the iTSCs of >5 ppm and those of <5 ppm were successfully assembled without any gap using a total of 3 382 901 short-read sequence tags. We analyzed the assembled sequences and found that the 80% of them contained no ORF longer than 100 amino acids and 91% of them had either of the above translational caveats (Supplementary Fig. S3B). These results strongly suggested that, either >5 ppm or <5 ppm, the iTSCs identified here should represent TSSs of ncRNAs.

3.3. *Expression profiles of the iTSCs of >5 ppm*

In our previous study, we found that the iTSCs of very low expression levels are generally: (i) located in AT rich regions, rarely associated with CpG island; (ii) evolutionarily poorly conserved, often associated

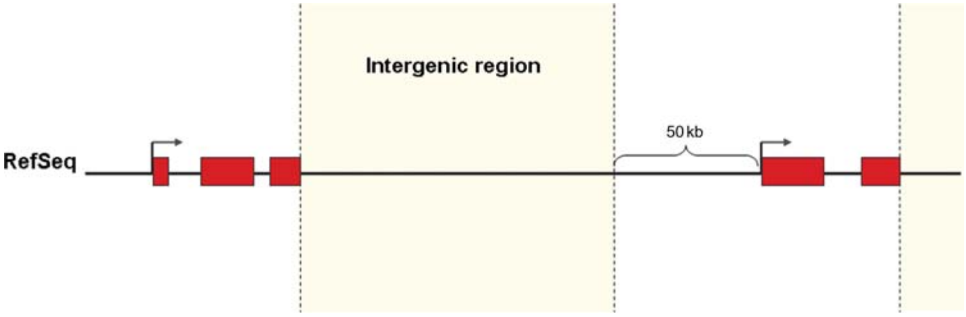


Figure 2. Intergenic region. Intergenic regions were defined as regions from 3′-end boundaries to the −50 kb of the 5′-end boundaries of the adjoining RefSeq genes; ‘50 kb margins’ was adopted at the 5′-ends of the RefSeq genes, considering that some of the iTSCs located within these regions might be alternative promoters of downstream RefSeq genes.

with repetitive sequences (also see Supplementary Fig. S4A). We also found that, unlike the cases in significantly expressed iTSCs, iTSCs of very low expression levels are rarely associated with RNA polymerase II binding sites based on the results of the ChIP-Seq analysis (also see Supplementary Figs S4C and S5). In summary, the iTSCs of significant expression levels have clearly distinctive features, which are rather similar to the TSSs of protein-coding transcripts. On the other hand, sequence features observed for iTSCs of low expression levels seemed consistent with the hypothesis that very minor iTSCs occasionally occur from universal genomic sequences as biological noise, thus evolutionarily unstable since no selection pressure is being extorted. Further details of this analysis are described in the reference (Yamashita *et al*., submitted).

In order to further characterize the iTSCs of >5 ppm, we examined their expression profiles. For this, we first used digital TSS Seq tag counts in 12 cell types. Among the 6039 iTSCs of >5 ppm, we counted number of cell types which gave TSS tag counts of >5 ppm. We found that 3739 (62%) of them were expressed at >5 ppm only in one cell type, suggesting that many of them are expressed in a tissue-preferred manner (Fig. 3A). Among these, fetal brain was the richest source of the iTSCs of >5 ppm. In all of 1812 iTSCs expressed at >5 ppm in fetal brain, 722 (40%) were expressed only in fetal brain and additionally 166 (9%) iTSCs were expressed in fetal brain and brain (Fig. 3B). On the other hand, the poorest source of the tissue-specific iTSCs was HEK293 [nevertheless, still 66 (25%) of the iTSCs showed specific expressions in HEK293; Fig. 3C]. Generally, tissue-specific expressions of the iTSCs were observed more frequently in normal tissues than cultured cells, possibly reflecting the reduced complexity of the gene functions in cultured cell lines.

In order to observe the expression patterns of the iTSCs in wider cell types, we performed qRT–PCR assays in additional 20 kinds of human normal tissues using the PCR primers designed based on the complete cDNA sequences. For 44 out of 67 iTSCs of >5 ppm, we observed clear signals at least in one tissue. Again, we observed that the expression of the iTSCs of >5 ppm generally had tissue preference (Supplementary Table S1). We also examined the expression of the 93 iTSCs of <5 ppm by qRT–PCR. In 60 out of 93 iTSCs of <5 ppm, weak but clear signals were obtained. Of these expressed iTSCs, 16 iTSCs (27%) showed clear tissue specificity. Although

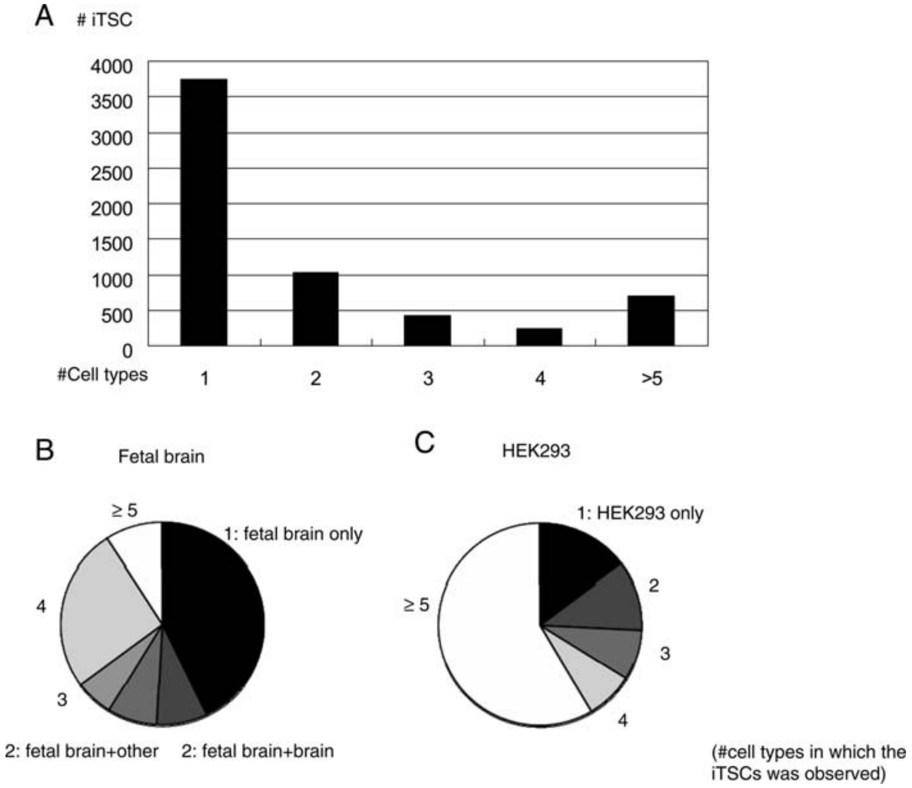


Figure 3. Expression profile of the iTSCs of >5 ppm. (**A**) Tissue specificity of the iTSCs of >5 ppm. The number of the cell types in which iTSCs of >5 ppm were observed is indicated in *x*-axis. (**B** and **C**) As for iTSCs which were observed in fetal brain (B) and HEK293 cells (C), whether they were also observed in different cell types was examined. Populations of the iTSCs whose expressions were observed in the indicated number of the cell types are shown.

the presence of the transcripts for those iTSCs was confirmed, their expression levels determined by qRT–PCR were lower than iTSCs of >5 ppm across the tissues examined.

3.4. *Correlation between a group of iTSCs of >5 ppm and RefSeq genes*

We further scrutinized the iTSCs of >5 ppm and found that iTSCs of >5 ppm are not a uniform group. To precisely examine the genomic position of the iTSCs, 10 kb windows of the genome position were defined relatively to the RefSeq gene positions. We found that a number of iTSCs of >5 ppm were located in the proximal regions of RefSeq genes (Fig. 4). Although 23% (1382 cases) of the iTSCs of >5 ppm were located within 10 kb downstream of the RefSeq gene in the same strand (RefSeq-linked iTSCs), the frequency of such iTSCs in the iTSC group of <5 ppm was only 7% (24 486 cases). We compared the expression patterns of the iTSCs of >5 ppm with their neighboring RefSeq genes by calculating Pearson's correlation of the respective TSS tag counts. We found that the expressions of the RefSeq-linked iTSCs were generally resembled to the expressions of their nearest RefSeq genes (Fig. 5A). Such correlation was not detected in the iTSCs located further than 10 kb from the nearest RefSeq genes (non-RefSeq-linked iTSCs; Fig. 5B). In order to observe the expression correlation in wider cell types, we performed qRT–PCR assays in 20 normal tissues using 12 representative cases. Again, we observed general correlation of the expression patterns between the RefSeq-linked iTSCs of >5 ppm and the nearby RefSeq genes (exemplified in Fig. 5C), and no correlation was observed for the non-RefSeq-linked iTSCs of >5 ppm (exemplified in Fig. 5D).

It was unlikely that these RefSeq-linked iTSCs of >5 ppm were derived from the 5′-end truncated products of the RefSeq gene transcripts due to the errors in the cap selection procedure.[27] At least in 186 out

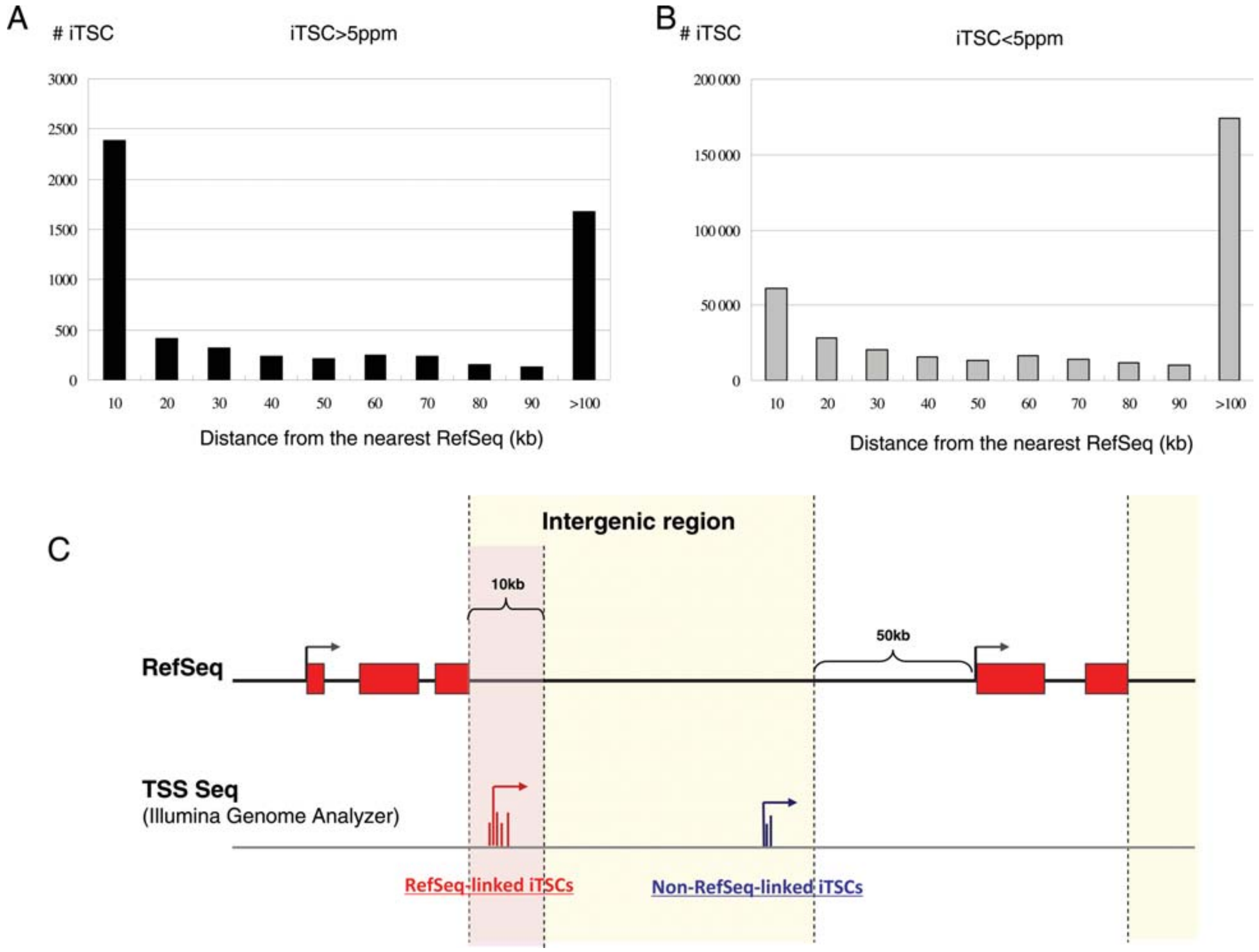


Figure 4. Distance from the iTSCs to the nearest RefSeq genes. Distribution of the distance between the iTSCs and the nearest RefSeq genes is shown in the case of the iTSCs of >5 ppm (**A**) and the those of <5 ppm (**B**). According to this positioning observation, the iTSCs were separated into two groups: (**C**) RefSeq-linked iTSCs (iTSCs which lie within 10 kb of RefSeq genes) and non-RefSeq-linked iTSCs (iTSCs which locate further than 10 kb from the nearest RefSeq genes).

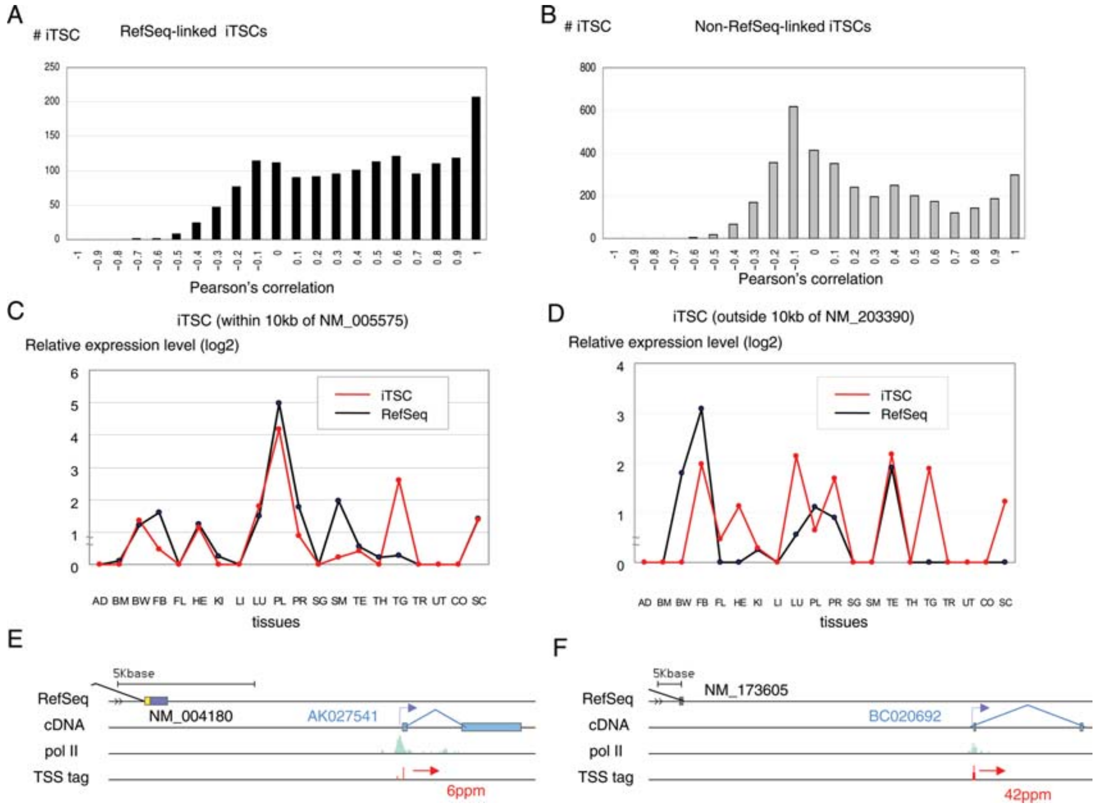


Figure 5. Correlation between the iTSCs and the nearest RefSeq genes. Distribution of Pearson's correlation of the TSS tag counts between the iTSCs located within 10 kb of the RefSeq genes (**A**) and those located outside 10 kb of the nearest RefSeq genes (**B**). (**C**) Correlation of the expression patterns monitored by quantitative RT-PCR in 20 normal human tissues (AD, adrenal gland; BM, bone marrow; BW, brain (whole); FB, fetal brain; FL, fetal liver; HE, heart; KI, kidney; LI, liver; LU, lung; PL, placenta; PR, prostate; SG, salivary gland; SM, skeletal muscle; TE, testis; TH, thymus; TG, thyroid gland; TR, trachea; UT, uterus; CO, colon w/mucosa; SC, spinal cord) for a RefSeq-linked iTSC and the linked RefSeq (NM_005575; 244 bp apart from the iTSC). (**D**) Result from the similar analysis as (C) using a non-RefSeq-linked iTSC and the nearest RefSeq gene (NM_203390; 11 581 bp apart from the iTSC). Typical examples for the RefSeq-linked iTSC (**E**) and non-RefSeq-linked iTSC (**F**). Completely sequenced cDNA and distributions of the TSS Seq tags and ChIP-Seq tags (pol II) in the surrounding region are shown.

of 356 (52%) RefSeq-linked iTSCs which were expressed at >5 ppm in DLD-1 cells, their presence were also clearly supported by bindings of RNA polymerase II by ChIP-Seq analysis (Supplementary Fig. S4C). Of these, 50 RefSeq-linked iTSCs were further supported by cDNA clones, which did not overlap the RefSeq transcripts (Fig. 5E and F). It was rather likely that these RefSeq-linked iTSCs might be activated simply because of the generally loose chromatin structure which was formed to facilitate the transcription of the upstream RefSeq gene (mentioned as 'transcriptional rippling noise' in a recent paper[36]).

In order to examine this hypothesis, we analyzed the nucleosome structure in their surrounding regions by Nucleosome Seq using DLD-1 cells. A total of 19 570 149 sequence tags were collected and occupancy of the nucleosome was calculated (see the 'Materials and methods' section, Supplementary Fig. S6 and the references[37–39]). Orderly nucleosome structure was not observed for the RefSeq-linked iTSCs of >5 ppm (Fig. 6B) and those of <5 ppm (Fig. 6C). On the contrary, for the non-RefSeq-linked TSCs of >5 ppm, we observed the formation of the nucleosome-free regions just upstream of the TSSs and periodic positioning of the nucleosome immediately downstream of the TSS (Fig. 6A). These results may reflect that transcriptions of RefSeq-linked iTSCs of >5 ppm seem not to form independent transcriptional 'regulatory' units, in spite of their relatively high expression levels, whereas non-RefSeq-linked iTSCs of >5 ppm are subjected to strict independent regulations. Consistently, we further examined and found that the TSSs were more widely fluctuating (also see Suzuki *et al.*[40]) in RefSeq-linked iTSCs of >5 ppm than non-RefSeq-linked iTSCs of >5 ppm (7e−16; Wilcoxon test). The average sizes of the regions which covered 80% of the TSSs within a particular TSC were 264 and 122 bp, respectively.

3.5. *Characterization of the non-RefSeq-linked iTSCs*

For the remaining 3651 non-RefSeq-linked iTSCs of >5 ppm, we considered that some of them might

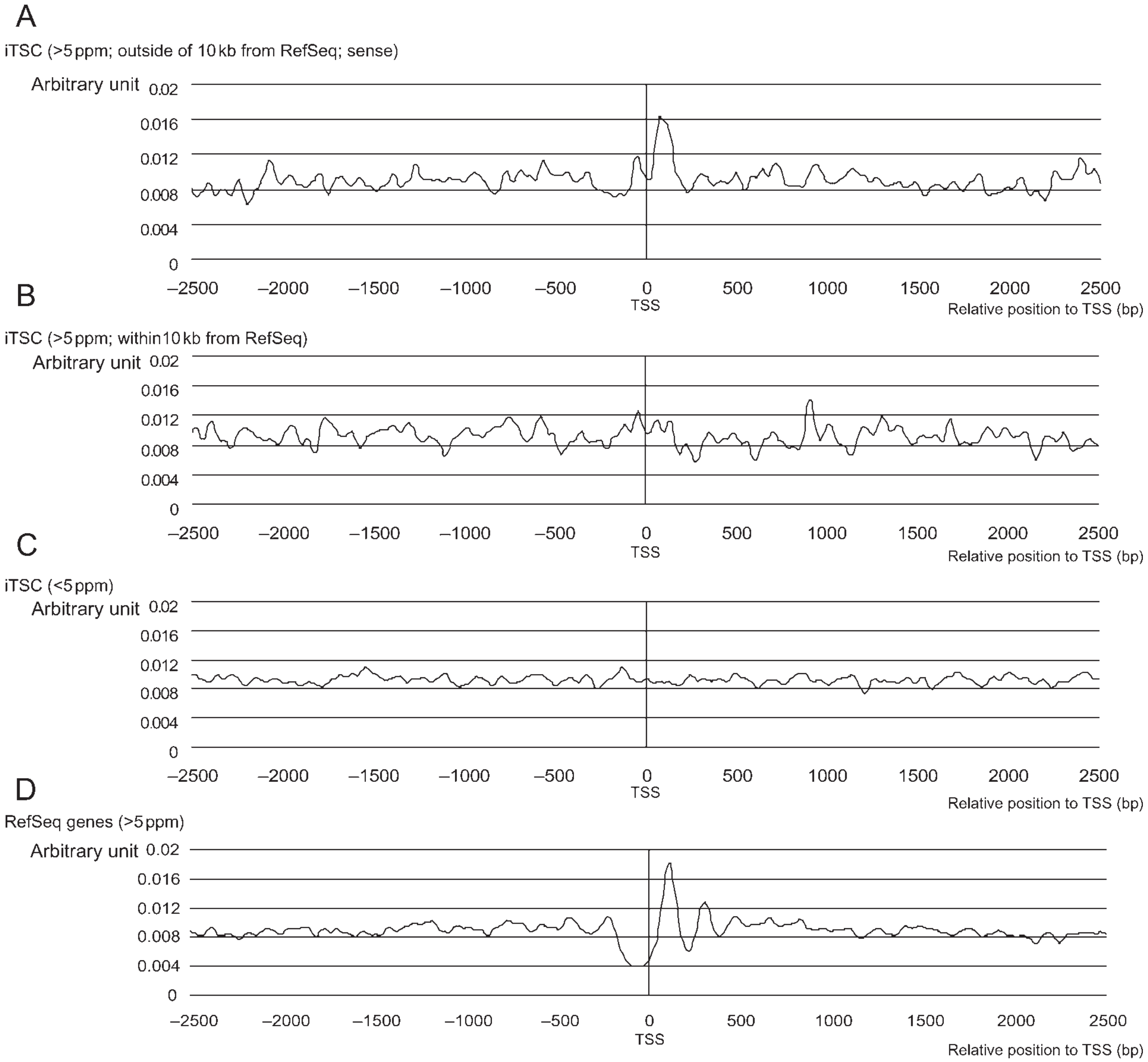


Figure 6. Nucleosome structure of the surrounding regions of the iTSCs. Nucleosome structure in the surrounding regions of the iTSCs of >5 ppm, located outside of 10 kb from the nearest RefSeq genes on the same strand (**A**), those of >5 ppm, located within 10 kb from the nearest RefSeq genes on the same strand (**B**), and those of <5 ppm (**C**). (**D**) Nucleosome structure in the surrounding regions of RefSeq genes. Six hundred and fifteen (total number of the RefSeq-linked iTSCs plus non-RefSeq-linked RNAs) RefSeq genes were randomly selected so that the distribution of their expression levels should be similar to that of iTSCs. For each group, nucleosome occupancy scores (*y*-axis) were calculated for the indicated genomic position (*x*-axis) relative to the TSS, according to the standard method shown in the reference.

participate in transcription of the precursors of miRNAs or other small RNAs, since the structures of the primary transcripts of them were still mostly unknown. We examined overlap of the iTSCs with previously reported miRNAs and snoRNAs registered in miRBase[41] and snoBase,[42] respectively. We found that 23 miRNAs and 123 snoRNAs were located within 50 kb downstream of the identified iTSCs (note 50 kb downstream sequences were extended from identified iTSCs, representing transcription regions of iTSC). Particularly, for 22 miRNAs and 11 snoRNAs, complete cDNA sequences directly showed that the mature forms of the small RNAs were actually located within the transcribed regions following the iTSCs (exemplified in Supplementary Fig. S7).

In order to further identify the iTSCs which transcribe the primary transcripts of miRNAs, we performed small RNA Seq analysis using DLD-1 cells. In total, 604 141 sequences of size-selected RNAs of 21-25 bp were determined. Obtained short-read sequences were clustered, and 21-bp sequences which had expression levels of >50 ppm were selected (in this case, the ppm value was calculated against the total number of the sequenced small RNAs, whose gross expression levels was estimated 1/10–1/100 compared with mRNAs/lncRNAs.

Thus, 50 ppm is expected to be at the similar level with 5 ppm in the case of mRNAs/lncRNAs). Among the identified 1266 putative miRNAs (49 overlapped the previously reported miRNAs), 24 miRNAs were located within 50 kb of the non-RefSeq-linked iTSCs of >5 ppm in DLD-1 cells (4% of the total 575 non-RefSeq-linked iTSCs). Of these putative miRNAs, only two were directly supported by cDNA clones. Having observed that most of the iTSCs were not correlated with the miRNAs, the majority of non-RefSeq-linked iTSCs are not used for transcribing primary transcripts of miRNAs.

We examined subcellular localization of their transcripts by RNA Seq of RNAs extracted from nuclear and cytoplasmic fractions using DLD-1 cells. As shown in Supplementary Fig. S8, precise separation of the cytoplasmic fraction from the nuclear fraction was confirmed by qRT–PCR, using nuclear scaRNAs and other snoRNAs, and western blotting analysis using nuclear lamin A/C proteins and cytoplasmic GAPDH protein. A total of 20 094 475 and 14 879 174 tags were collected from the nuclear and cytoplasmic RNAs, respectively. RNA Seq tags were associated with the iTSCs when their mapped genomic positions overlapped. Among the 575 non-RefSeq-linked iTSCs with expression levels of >5 ppm in DLD-1 cells, 337 iTSCs (59%) overlapped at least three RNA Seq tags from the cytoplasmic fraction. Of these, 99 iTSCs (17%) were particularly enriched in cytoplasm ($P < 0.01$). These results showed that a large number of iTSC transcripts are actually exported into cytoplasm without losing the surrounding sequences of the TSSs, possibly in their native forms as lncRNAs.

3.6. *Integration of the transcriptome data for inferring biological functions of the non-RefSeq-linked iTSCs*

We wished to demonstrate that specific functions of the iTSCs underlying particular biological events could be inferred from the integrative interpretation of hereby described various types of transcriptome data. We focused the analysis on the iTSCs whose functions may be associated with hypoxic response of the cell. For this, we generated additional 36 761 810 thirty-six-base pair TSS tags in DLD-1, MCF7, HEK293 and TIG3 cells which were cultured in 1% O_2 (Supplementary Fig. S9). As shown in Table 1, 508 and 365 iTSCs of >5 ppm showed expression induction and repression by more than 5-fold in response to the hypoxic shock in at least one cell type, respectively. Among the 508 hypoxia-induced iTSCs, 232 were RefSeq-linked iTSCs and 276 were non-RefSeq-linked iTSCs. Among the 276 non-RefSeq-linked hypoxia-induced iTSCs, iTSC of an lncRNA, H19, whose induction in response to hypoxia has been recently reported to be essential in the tumorigenecity of the cell, was included.[43]

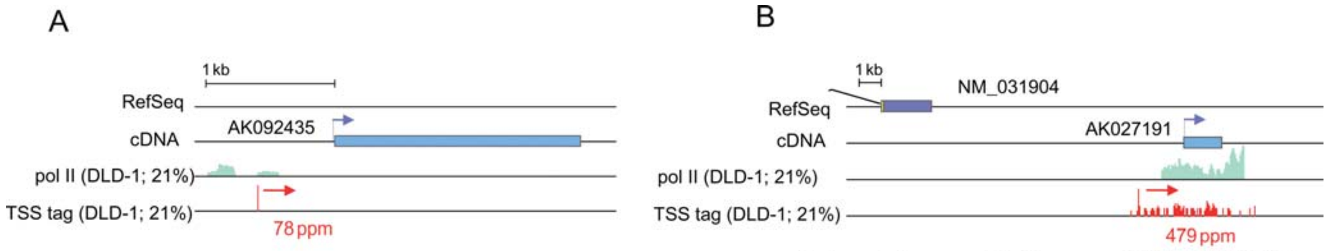
On the basis of the collected TSS Seq data, we examined whether expression patterns of the iTSCs were correlated with any groups of the RefSeq genes. We associated the RefSeq genes with the iTSC when Pearson's correlation coefficient was >0.8 (Pearson's correlation coefficients between the TSS tag counts of an lncRNA, H19, and those of p57kip2 and angiogenin, which are reported direct effectors of H19 in the hypoxic responses,[30,43,44] were 0.98 and 0.78, respectively). Then, we examined whether particular GO terms[45] were enriched ($P < 1e-6$; hypergeometric distribution) in the group of RefSeq genes which were associated with the iTSCs.

As exemplified in Fig. 7A, a non-RefSeq-linked iTSC, which was 9.6-fold induced in MCF-7 cells by hypoxia, was correlated with the 34 RefSeq genes having GO terms, 'hemophilic cell adhesion', in common ($P = 7e-6$). Subcellular localization of the RNA products of this iTSC identified by the RNA Seq analysis was 'cytoplasm' ($P = 0.0$). The biological function of this lncRNA might be to interact directly with cell adhesion protein molecules in modulating cellular mobility. Full list of the iTSCs for which biological functions were inferred by the GO terms is shown in Supplementary Table S2. Interestingly, although various GO terms were associated, the GO term 'hemophilic cell adhesion' appeared far more frequently than the other GO terms ($P < 4e-13$). In total, 53 iTSCs were correlated with this GO term. One of the major biological roles of iTSCs in response to hypoxia might be to control cell–cell interactions.

On the other hand, expressions of a significant number of RefSeq-linked iTSCs were also changed by hypoxia, as exemplified in Fig. 7B. However, in this group of iTSCs, generally fewer RefSeq genes other

Table 1. Statistics of the TSS Seq analysis monitoring hypoxic responses of the cultured cells

	Cell type	DLD-1 1% O_2	HEK293_1% O_2	MCF-7 1% O_2	TIG-3 1% O_2	Total
Fold (1%/21% O_2) <0.2	Total	127	99	99	69	365
	Non-RefSeq-linked iTSC	79	63	56	27	213
	RefSeq-linked iTSC	48	36	43	42	152
Fold (1%/21% O_2) >5	Total	264	111	105	59	508
	Non-RefSeq-linked iTSC	169	45	56	19	276
	RefSeq-linked iTSC	95	66	49	40	232



	A	B
Distance to the nearest RefSeq	>10 kb	9.3 kb (NM_031904)
cDNA (longest ORF)	AK092435 (73 a.a.)	AK027191 (39 a.a.)
Associated GO term	Cell adhesion	None
Subcellular localization (RNA Seq)	Cytoplasm ($P = 0.0$)	Nucleus ($P = 0.004$)
Maximum fold change by hypoxia (1%/21% O_2)	9.6 (in MCF-7)	0.1 (in MCF-7)

Figure 7. Integration of the transcriptome data obtained from various viewpoints. Transcriptome information collected by the indicated analyses is shown for a non-RefSeq-linked iTSC (**A**) and a RefSeq-linked iTSC (**B**).

than their neighboring RefSeq genes had correlated expression patterns with those iTSCs and fewer GO terms were associated. These results may reflect the fact that inductions of the RefSeq-linked iTSCs are not well coordinated in the transcriptional regulatory network of human genes.

3.7. *Identification and characterization of TSCs in the antisense of the RefSeq genes*

About 1–5% of the TSS tags were mapped to the antisense of the RefSeq genes, depending on cell types (Table 2). Similar to the case in the iTSCs, we clustered and analyzed those TSS tags (antisense TSCs; aTSCs). Again, we observed that the aTSCs having significant expression levels are only minor part. Taken all cell types together, although there were 127 036 aTSCs, aTCSs of >5 ppm were only 3301 (Table 3; also see Supplementary Fig. S2). We examined the characteristic features of the aTSCs and found that they showed somewhat similar features with iTSCs of >5 ppm; located in evolutionarily conserved GC-rich regions and rarely associated with repetitive elements (Supplementary Fig. S4B). We also confirmed that the presence of the aTSCs of >5 ppm is clearly supported by pol II binding in some cases (Supplementary Fig. S4C). Interestingly, those aTSCs of >5 ppm were concentrated around the 5′- and 3′-ends of their harboring genes (Fig. 8A). Nevertheless, when we further examined their expression patterns with the overlapping RefSeq genes by calculating Pearson's correlation based on the digital TSS tag counts, we found generally correlation between the expression patterns of aTSCs and those of the antisense RefSeq genes (Fig. 8B). Also, we observed that general lack of the ordered nucleosome positioning in the surrounding regions (Fig. 8C and D), perhaps reflecting the lack of strict transcriptional regulations. These results may imply that, similar to the case in the RefSeq-linked iTSCs, major part of the aTCSs, if not all, are also 'transcriptional rippling' products of the transcriptions of the RefSeq genes and have no biological consequences on their own. Distal regions of the genes might be particularly susceptible to such transcriptions, although the possibility that at least a part of those aTSCs are involved in transcriptional regulations of RefSeq genes has not been completely excluded.

3.8. *Conclusions*

In this paper, we described our first attempt to characterize the iTSCs by multifaceted use of massively paralleled sequencer. Integrated analysis of various types of transcriptome data revealed that different types of iTSCs have different properties. Particularly, we showed that non-RefSeq-linked iTSCs of >5 ppm have several features which should be essential to realize biologically relevant transcriptions. And they could be clearly discriminated from other type of iTSCs, many of which might represent noise products of the transcriptions inherent to the long and complex human genome. We also showed that major parts of the aTSCs have the similar properties with the iTSCs which seemed not to form independent transcriptional units. It is true that we used the conservative computational filters placed on the sequencing results. For example, we considered only TSCs of >5 ppm in the limited cell types. We also removed TSCs located within 50-kb upstream regions of the RefSeq genes and defined the iTSCs located within 10-kb downstream regions of the RefSeq genes as 'RefSeq-linked' iTSCs. Therefore, classification using further optimized parameters is necessary to re-defining the characteristics of the respective iTSCs group. However, it should be significant that we

Table 2. Statistics of the sequence tags categorized to the indicated criteria

Cell type	DLD-1	Ramos	BEAS2B	HEK293	MCF7	TIG3	Brain	Kidney	Heart	Fetal brain	Fetal kidney	Fetal heart	Total
#TSS tag in RefSeq region	7 679 112 (87%)	13 875 727 (83%)	20 106 118 (89%)	8 880 335 (91%)	7 476 053 (89%)	8 883 872 (90%)	7 761 770 (66%)	7 285 131 (64%)	5 379 867 (57%)	9 607 689 (82%)	6 783 071 (77%)	7 982 779 (77%)	111 70 1 524 (80%)
#TSS tag in intergenic region	957 301 (11%)	2 355 840 (14%)	1 744 854 (8%)	736 616 (8%)	796 086 (10%)	783 112 (8%)	3 472 374 (29%)	3 458 501 (31%)	3 653 984 (38%)	1 542 402 (13%)	1 592 092 (18%)	1 817 819 (18%)	22 910 981 (16%)
#TSS tag in antisense of RefSeq region	213 119 (2%)	408 592 (3%)	623 077 (3%)	117 363 (1%)	82 131 (1%)	217 908 (2%)	602 851 (5%)	579 412 (5%)	441 998 (5%)	559 169 (5%)	487 960 (5%)	500 645 (5%)	4 834 225 (4%)
Total	8 849 532	16 640 159	22 474 049	9 734 314	8 354 270	9 884 892	11 836 995	11 323 044	9 475 849	11 709 260	8 863 123	10 301 243	139 446 730
#iTSC (total)	47 347	58 755	59 529	27 959	23 914	18 450	91 915	41 251	53 204	67 897	50 950	64 144	371 849
#iTSC (>5 ppm)	931	1066	1004	833	1088	861	1207	1013	688	1812	1434	1323	6039

Table 3. Statistics of the aTSCs of >5 ppm and those of <5 ppm

Cell type	DLD-1	Ramos	BEAS2B	HEK293	MCF7	TIG3	Brain	Kidney	Heart	Fetal brain	Fetal kidney	Fetal heart	Total
# aTSC (total)	17 516	19 757	23 836	8914	6491	7294	28 972	20 174	25 522	20 284	23 870	23 815	127 036
# aTSC (>5 ppm)	575	453	519	203	197	457	765	897	608	945	1331	856	3301

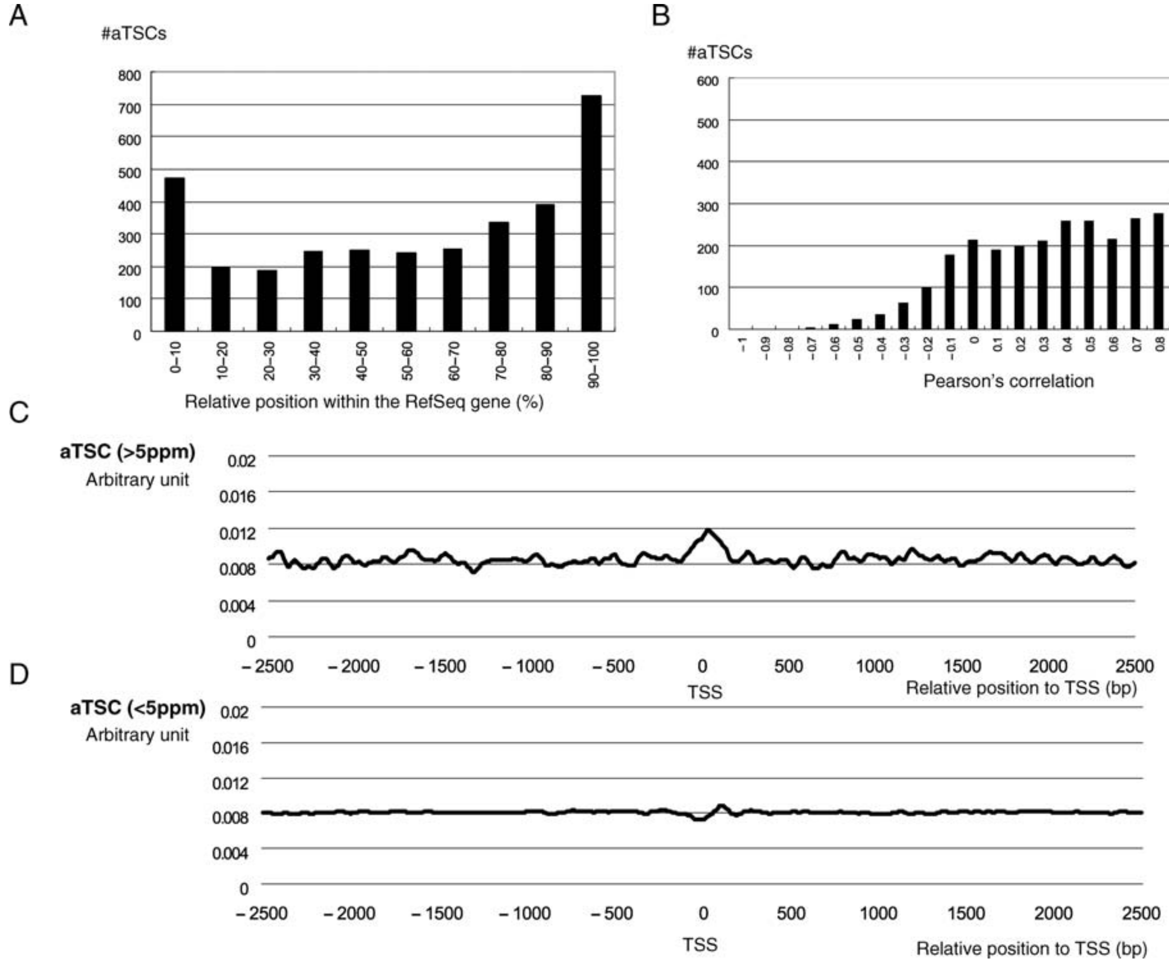


Figure 8. aTSCs of the RefSeq genes. In addition, we performed integrated analysis of aTSCs of the RefSeq genes. Interestingly, aTSCs of >5 ppm were concentrated around the 5′- and 3′-ends of their harboring genes (**A**). Distribution of Pearson's correlation of the TSS tag counts between aTSCs and their harboring genes revealed that there are generally correlation between the expression patterns of aTSCs and those of the antisense RefSeq genes (**B**). By analyzing the nucleosome structure, we observed a general lack of the ordered nucleosome positioning in the surrounding regions of both aTSCs of >5 ppm (**C**) and those of <5 ppm (**D**).

could observe not all the iTSCs have the same features even as the first approximation.

We also demonstrated that integrative transcriptome data gives useful starting point for further investigation of the biological roles of the iTSCs, specifically in the case of the hypoxic response of the cells. Indeed, it is important to consider various types of transcriptome data. Importantly, as shown in this study, the number of iTSCs/aTSCs whose biological relevance seems somewhat questionable is far larger than the iTSCs/aTSCs which have presumed biological functions. Without any biological information in addition to cataloged TSSs/cDNAs, it is practically impossible to select and conduct meaningful functional assays for the increasing number of iTSCs/aTSCs. In our opinion, the biggest advantage of the massively paralleled sequencer is that various approaches of genome-wide analysis are simultaneously enabled on a common platform. Further intensive use of the massively paralleled sequencer will shed brighter light on dynamically interacting worlds of ncRNAs and proteins, which collectively realize complex human gene network. With such knowledge, we will finally understand the biological meaning of every transcription initiation event in the human genome.

Acknowledgements: We are grateful to K. Abe, E. Sekimori and K. Imamura for technical supports. We are also thankful to Dr Sierro, Dr M. Frith and Mr F. Sathira for critical reading of the manuscript.

Supplementary data: Supplementary data are available at *DNA Research* online.

Funding

This work was supported by Ministry of Economy, Trade and Industry (METI) of Japan [the Japan Key Technology Center project]; Ministry of Health, Labour and Welfare [Third-Term Comprehensive 10-Year Strategy for Cancer Control, Grants-in-Aid for Cancer Research]; Ministry of Education, Culture, Sports, Science and Technology-Japan [Grant-in-Aid for Scientific Research on Priority Areas].

References

1. Hashimoto, S., Suzuki, Y., Kasai, Y., et al. 2004, 5′-end SAGE for the analysis of transcriptional start sites, *Nat. Biotechnol.*, **22**, 1146–9.
2. Carninci, P., Kasukawa, T., Katayama, S., et al. 2005, The transcriptional landscape of the mammalian genome, *Science*, **309**, 1559–63.
3. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., et al. 2007, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, **447**, 799–816.
4. Bertone, P., Stolc, V., Royce, T.E., et al. 2004, Global identification of human transcribed sequences with genome tiling arrays, *Science*, **306**, 2242–6.
5. Guttman, M., Amit, I., Garber, M., et al. 2009, Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, *Nature*, **458**, 223–7.
6. Babak, T., Blencowe, B.J. and Hughes, T.R. 2005, A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription, *BMC Genomics*, **6**, 104.
7. Rinn, J.L., Euskirchen, G., Bertone, P., et al. 2003, The transcriptional activity of human chromosome 22, *Genes Dev.*, **17**, 529–40.
8. Yelin, R., Dahary, D., Sorek, R., et al. 2003, Widespread occurrence of antisense transcription in the human genome, *Nat. Biotechnol.*, **21**, 379–86.
9. Rosok, O. and Sioud, M. 2004, Systematic identification of sense-antisense transcripts in mammalian cells, *Nat. Biotechnol.*, **22**, 104–8.
10. Khaitovich, P., Kelso, J., Franz, H., et al. 2006, Functionality of intergenic transcription: an evolutionary comparison, *PLoS Genet.*, **2**, e171.
11. Nakaya, H.I., Amaral, P.P., Louro, R., et al. 2007, 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.*, **8**, R43.
12. Numata, K., Kanai, A., Saito, R., et al. 2003, Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection, *Genome Res.*, **13**, 1301–6.
13. Ravasi, T., Suzuki, H., Pang, K.C., et al. 2006, Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome, *Genome Res.*, **16**, 11–9.
14. Szell, M., Bata-Csorgo, Z. and Kemeny, L. 2008, The enigmatic world of mRNA-like ncRNAs: their role in human evolution and in human diseases, *Semin. Cancer Biol.*, **18**, 141–8.
15. Perez, D.S., Hoage, T.R., Pritchett, J.R., et al. 2008, Long, abundantly expressed non-coding transcripts are altered in cancer, *Hum. Mol. Genet.*, **17**, 642–55.
16. Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F. and Mattick, J.S. 2008, Specific expression of long noncoding RNAs in the mouse brain, *Proc. Natl Acad. Sci. USA*, **105**, 716–21.
17. Ponting, C.P., Oliver, P.L. and Reik, W. 2009, Evolution and functions of long noncoding RNAs, *Cell*, **136**, 629–41.
18. Ponjavic, J., Ponting, C.P. and Lunter, G. 2007, Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs, *Genome Res.*, **17**, 556–65.
19. Arias, A.M. and Hayward, P. 2006, Filtering transcriptional noise during development: concepts and mechanisms, *Nat. Rev. Genet.*, **7**, 34–44.
20. Struhl, K. 2007, Transcriptional noise and the fidelity of initiation by RNA polymerase II, *Nat. Struct. Mol. Biol.*, **14**, 103–5.
21. Brosius, J. 2005, Waste not, want not–transcript excess in multicellular eukaryotes, *Trends Genet.*, **21**, 287–8.
22. Huttenhofer, A., Schattner, P. and Polacek, N. 2005, Non-coding RNAs: hope or hype?, *Trends Genet.*, **21**, 289–97.
23. Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S. and Sidow, A. 2004, Characterization of evolutionary rates and constraints in three mammalian genomes, *Genome Res.*, **14**, 539–48.
24. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. 2005, Distribution and intensity of constraint in mammalian genomic sequence, *Genome Res.*, **15**, 901–13.
25. Wang, J., Zhang, J., Zheng, H., et al. 2004, Mouse transcriptome: neutral evolution of ‘non-coding’ complementary DNAs, *Nature*, **431**, 1 p following 757; discussion following 757.
26. Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. 2008, DBTSS: database of transcription start sites progress report, *Nucleic Acids Res.*, **36**, D97–101.
27. Suzuki, Y. and Sugano, S. 2003, Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method, *Methods Mol. Biol.*, **221**, 73–91.
28. Ota, T., Suzuki, Y., Nishikawa, T., et al. 2004, Complete sequencing and characterization of 21,243 full-length human cDNAs, *Nat. Genet.*, **36**, 40–5.
29. Bentley, D.R. 2006, Whole-genome re-sequencing, *Curr. Opin. Genet. Dev.*, **16**, 545–52.
30. Tsuchihara, K., Suzuki, Y., Wakaguri, H., et al. 2009, Massive transcriptional start site analysis of human genes in hypoxia cells, *Nucleic Acids Res.*, **37**, 2249–63.
31. Schones, D.E., Cui, K., Cuddapah, S., et al. 2008, Dynamic regulation of nucleosome positioning in the human genome, *Cell*, **132**, 887–98.
32. Albert, I., Mavrich, T.N., Tomsho, L.P., et al. 2007, Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome, *Nature*, **446**, 572–6.
33. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. 2007, Genome-wide mapping of in vivo protein-DNA interactions, *Science*, **316**, 1497–502.
34. Seila, A.C., Calabrese, J.M., Levine, S.S., et al. 2008, Divergent transcription from active promoters, *Science*, **322**, 1849–51.

35. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat. Methods*, **5**, 621–8.
36. Ebisuya, M., Yamamoto, T., Nakajima, M. and Nishida, E. 2008, Ripples from neighbouring transcription, *Nat. Cell Biol.*, **10**, 1106–13.
37. Gupta, S., Dennis, J., Thurman, R.E., Kingston, R., Stamatoyannopoulos, J.A. and Noble, W.S. 2008, Predicting human nucleosome occupancy from primary sequence, *PLoS Comput. Biol.*, **4**, e1000134.
38. Jiang, C. and Pugh, B.F. 2009, Nucleosome positioning and gene regulation: advances through genomics, *Nat. Rev. Genet.*, **10**, 161–72.
39. Ozsolak, F., Song, J.S., Liu, X.S. and Fisher, D.E. 2007, High-throughput mapping of the chromatin structure of human promoters, *Nat. Biotechnol.*, **25**, 244–8.
40. Suzuki, Y., Taira, H., Tsunoda, T., et al. 2001, Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites, *EMBO Rep.*, **2**, 388–93.
41. Griffiths-Jones, S., Saini, H.K., van Dongen, S. and Enright, A.J. 2008, miRBase: tools for microRNA genomics, *Nucleic Acids Res.*, **36**, D154–8.
42. Xie, J., Zhang, M., Zhou, T., Hua, X., Tang, L. and Wu, W. 2007, Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs, *Nucleic Acids Res.*, **35**, D183–7.
43. Matouk, I.J., DeGroot, N., Mezan, S., et al. 2007, The H19 non-coding RNA is essential for human tumor growth, *PLoS One*, **2**, e845.
44. Cai, X. and Cullen, B.R. 2007, The imprinted H19 non-coding RNA is a primary microRNA precursor, *RNA*, **13**, 313–6.
45. GeneOntologyConsortium 2006, The gene ontology (GO) project in 2006, *Nucleic Acids Res.*, **34**, D322–6.