FEBS Letters 581 (2007) 4184–4188

Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse

Motomu Matsui[a,b], Nozomu Yachie[a,b], Yuki Okada[a,b], Rintaro Saito[a,c,*], Masaru Tomita[a,b,c]

[a] *Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan*
[b] *Bioinformatics Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-8520, Japan*
[c] *Department of Environment and Information Studies, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan*

Received 11 January 2007; revised 3 July 2007; accepted 24 July 2007

Available online 3 August 2007

Edited by Robert B. Russell

Abstract **RNA decay is thought to exert an important influence on gene expression by maintaining a steady-state level of transcripts and/or by eliminating aberrant transcripts. However, the sequence elements which control such processes have not been determined. Upstream open reading frames (uORFs) in the transcripts of several genes are reported to control translational initiation by stalling ribosomes and thereby promote RNA decay. We therefore performed bioinformatic analysis of the tissue-wide expression profiles and mRNA half-life of transcripts containing uORFs in humans and mice to assess the relationship between RNA decay and the presence of uORFs in transcripts. The expression levels of transcripts containing uORF were markedly lower than those not containing uORF. Moreover, the half-life of the uORF-containing transcripts was also shorter. These results suggest that uORFs are sequence elements that down-regulate RNA transcripts via RNA decay mechanisms.**

Keywords: Upstream open reading frames (uORF); RNA decay; Post-transcription control; 5′ Untranslated region; Kozak consensus

1. Introduction

RNA decay plays a major role in the post-transcriptional control of gene expression, maintaining the balance between the synthesis and degradation of RNA transcripts [1]. Although the complex processing pathways involved in RNA degradation have been described, the key determinants of instability remain unclear. Previous research has suggested that longer mRNA transcripts are less stable [2]. However, transcript stability was recently shown to not be correlated with overall transcript size, length of poly (A) tract, number of ribosomes, expression level, or codon usage [3].

The nonsense-mediated mRNA decay (NMD) pathway is thought to be an important surveillance mechanism that promotes the degradation of aberrant transcripts coding for non-functional or harmful proteins [4–6]. Nonsense or frameshift mutations introduce premature translation termination codons (PTCs) into the open reading frames (ORFs) of mRNAs and are a common cause of genetic disorders. PTCs usually lead to rapid mRNA degradation by NMD, which affects decapping, deadenylation, and $5' \rightarrow 3'$ exonucleolytic activities [7]. Upstream open reading frames (uORFs) are small open reading frames located in the 5′ untranslated regions (5′ UTR) of mRNA and also have important post-transcriptional effects. They are believed to function via *cis*-acting peptide products that reduce the initiation of translation of downstream ORFs by stalling the ribosome at the end of the uORF, thereby exposing the mRNA to degradation [8,9]. Genome-wide comparison of the human and mouse genomes suggest that the majority of uORFs are strongly conserved at the peptide level [10,11]. Furthermore, expression of the proteins encoded by human uORFs has been confirmed by mass spectrometry [9]. However it is not clear how pervasive a role the uORFs play in RNA decay. Thus, for example, a small peptide encoded within the 5′UTR of Yap2 mRNA modulates NMD in yeast [12] whereas the uORF encoded by the upstream region of the cytokine thrombopoietin transcript has been shown to not induce NMD in humans [13].

In this study we used a bioinformatics approach to assess whether uORFs in general affect RNA degradation. We first predicted the uORFs in the human and mouse transcriptomes and then compared the tissue-wide expression profile and decay rate of uORF-containing and non-uORF-containing transcripts. We found that the average level of expression of uORF-containing transcripts was markedly lower than that of the non-uORF-containing transcripts, and their decay rates were higher.

2. Materials and methods

2.1. Prediction of uORF-regulated transcripts

The human and mouse transcripts in the RefSeq database (release 23, accessed 23 May 2007) and human UniGene database (build 202, accessed 23 May 2007) were obtained via the National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nlm.nih.gov). Then all the transcripts were categorized into four levels through the following steps, as summarized in Fig. 1.

First, transcripts having no definite CDS annotations were eliminated. Then all of the longest ORFs satisfying the following three conditions were defined as uORFs; (1) the ORF (AUG) started in the 5′UTR, (2) the end of the ORF was not identical to the stop codon of the annotated downstream CDS, and (3) the end of the ORF was

*Corresponding author. Address: Department of Environment and Information Studies, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan. Fax: +81 466 47 5099.
E-mail address: rsaito@sfc.keio.ac.jp (R. Saito).

Abbreviations: ORF, open reading frame; uORF, upstream open reading frame; NMD, nonsense-mediated mRNA decay; PTC, premature termination codon

0014-5793/$32.00
doi:10.1016/j.febslet.2007.07.057

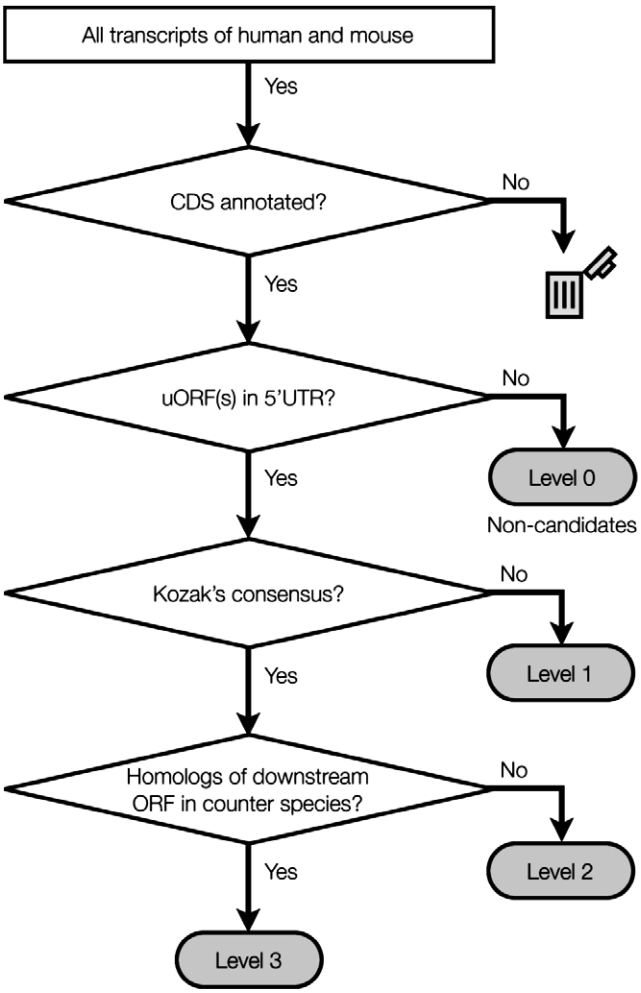

Fig. 1. Schematic representation of the categorization of the uORF-regulated candidates into four levels.

not in the 3′UTR [14]; the transcripts having no uORF were defined as "level 0" uORF candidates (non-candidates) and those having uORF(s) were defined as "level 1" candidates.

The efficiency of translation initiation from a given AUG codon is determined in part by the local sequence context around that codon [15,16]. Therefore, we selected those which had the Kozak consensus motif, 5′-[G/A]CC(AUG)G-3′ (the sequence in parentheses denotes the start codon), which is the most efficient context for the start codons of true ORFs, and re-defined them as "level 2" candidates.

We further selected those whose downstream ORFs were conserved in the human and mouse from the candidates in level 2, and re-defined these as "level 3" candidates. We used the BLAT program [17] with the E-value cutoff value 1e–50 to search for homologous gene pairs in the human and mouse. In other words, both transcripts in the human and mouse homologous pairs have uORFs in level 3.

In order to consider the effect of natural sense–antisense transcripts to gene expression, we checked the presence of overlapping transcripts in the antisense strand of the transcripts in each of the four levels. The dataset of 1233 natural antisense transcripts of humans and 4398 of the mouse were downloaded via the NATsDB website (http://natsdb.cbi.pku.edu.cn/, accessed 23 May 2007).

2.2. Preparation of data for comparing the expression intensities and half-lives of mRNA transcripts

We obtained 4935 RNA expression profiles from 79 different human tissues and 16617 profiles from 61 different mouse tissues, annotated with RefSeq IDs, from the SymAtlas database (http://symatlas.gnf.org/, accessed 23 May 2007) [18]. The expression profiles in the SymAtlas database were examined using high-density oligonucleotide arrays, and the custom arrays were generated using a non-redundant set of documented and predicted genes compiled from RefSeq [19], Celera [20], Ensembl, and RIKEN [21]. We used the expression data normalized by the gcRMA algorithm [22,23].

The decay rates of human mRNAs were obtained from a previous study which investigated the decay rates of 5245 individual transcripts in human cells [24]. The data were downloaded from the Genome Research website (http://www.genome.org/) and we used decay data on 2948 transcript IDs matching those in the UniGene database.

3. Results and discussion

We divided mRNAs into four categories, i.e. level 0, level 1, level 2 and level 3, according to the uORF predictions, the presence or absence of Kozak consensus motifs around the start codons and the conservation of uORF between the human and mouse (conservation of sequence patterns of uORFs were not considered). Among the 38927 human and 46627 mouse mRNAs obtained from the RefSeq database, and the 6731038 human mRNAs from the UniGene database, the CDSs of 33670, 42934, and 58745 mRNAs, respectively, were unambiguously annotated. Using the RefSeq data, we extracted candidates for uORF-regulated transcripts; 13174, 12711, 242 and 365 of these were classified into levels 0–3, respectively, in the human case. In the mouse, 15198, 14424, 263 and 440 candidates were classified into each of these four levels. Similarly, we extracted candidates from the UniGene database and the number of transcripts which were classified into levels 0–3 were 53137 39429, 820 and 1146, respectively. In addition, we extracted mRNAs which had a natural antisense transcript (NAT) in the RefSeq and UniGene databases to investigate the relationship between the presence of NATs and gene expression. Approximately, one-fourth of mRNAs in the human and mouse exhibited a NAT regardless of the level they were classified. In other words, the presence of NATs and uORFs in the transcripts were independent. Approximately, half of all the mRNAs were predicted to be uORF-regulated, in agreement with previous work using human, mouse, and rat mRNA sequences in the RefSeq database [10,11]. We speculate that the majority of the predicted uORFs possess the potential to be scanned by ribosomes, and so to control the translation reinitiation of downstream ORFs [12] by stalling and occupying ribosomes at their stop codons [12,25], acting in *cis* at the peptide level [11], or promoting NMD [6]. The RefSeq database contains entries whose transcription initiation sites were not defined [21,26], and many of the 5′UTRs were not completely identified. It is possible therefore that some of the transcripts categorized as level 0 candidates might actually contain uORFs. However, as the aim of this study was to compare the overall trends of the expression intensities and half-lives of the different categories using massive and statistical approaches, we believe that some small number of mis-predictions of uORF-regulated genes should not have a major impact on the results. Similarly, some of the candidate transcripts (levels 1–3) may not be regulated by uORFs but this again should not influence our results.

We compared the amount of transcripts within the corresponding categories of human and mouse mRNA using the SymAtlas database. The numbers of transcripts in each category (that were) cross-linked between the RefSeq and the SymAtlas database are given in Table 1. There were no marked

Table 1
The number of human and mouse uORF-regulated transcript candidates in which the intensity of RNA expression was examined

	Human (RefSeq)				Mouse (RefSeq)			
	All		Sense–antisense pair		All		Sense–antisense pair	
Level 0	2330	(47.2%)	566	(11.5%)	8952	(53.9%)	2479	(14.9%)
Level 1	2493	(50.5%)	644	(13.0%)	7325	(44.1%)	1830	(11.0%)
Level 2	43	(0.9%)	6	(0.1%)	98	(0.6%)	20	(0.1%)
Level 3	69	(1.4%)	17	(0.3%)	242	(1.5%)	69	(0.4%)
Total	4935	(100.0%)	1233	(25.0%)	16617	(100.0%)	4398	(26.5%)

The number of transcripts which were used for the expressional analyses are listed in the ''All'' column, separated by their levels. Their ratios among the total transcripts are shown in the brackets. In the ''Sense–antisense pair'' column, the number of transcripts having overlapping transcripts in the antisense strand is shown.

differences in the proportions of the respective categories in the RefSeq sequences and those linked in the SymAtlas database (data not shown). The average expression intensities of levels 1, 2 and 3 candidates in all human and mouse tissues were significantly lower than those of the level 0 candidates (see supplementary materials for details) (P-value two-sided t-test < 0.01). The maximum and minimum reductions in average expression levels in the level 3 candidates compared to the level 0 candidates were 55.9% and 11.7% in humans, and 67.2% and 15.2% in the mouse, respectively. The average intensities throughout all the tissues for candidates in each level are shown in Fig. 2. The average intensities of levels 1, 2 and 3 candidates were significantly lower than those of the level 0 candidates. These results clearly show that the amount of expressed mRNAs containing uORF(s) is lower compared to those which do not contain any uORFs. The difference between levels 1 and 2 was slight and thus we did not find the Kozak consensus sequence pattern as an apparent factor promoting down-regulation of transcripts. However, the lower intensity of the level 3 candidates compared with those of level 2 candidates (P-value < 0.01; two-sided t-test) might suggest that the uORFs which are involved in the down-regulation of mRNA expressions are evolutionary conserved.

The average expression levels of transcripts having NATs were significantly lower than those of all transcripts (Fig. 2), presumably reflecting expressional regulation by the sense–antisense mechanism of RNA [27,28]. However, we found that even if the transcripts did have NATs, the expression levels of the levels 1, 2 and 3 transcripts were significantly lower than those of level 0 (Fig. 2). Thus, the down-regulation of mRNAs by uORFs may be independent of antisense transcripts.

Although the absolute levels of transcripts in cells are not necessarily closely reflected by the intensities on oligonucleotide arrays due to problems such as lack of probe specificity, we believe that the relative abundance of sets of mRNAs are reflected reasonably well due to the massive comparison of such large amounts of intensity data.

The levels of transcripts in cells are determined by their rates of transcription and degradation. Thus far, there is no evidence that uORFs themselves regulate transcription. Therefore, in the absence of a relationship between the presence of uORFs and certain other factors (e.g. aberrant transcripts, probe specificities, etc.) influencing the measurements of the transcript levels, the significant differences in expression levels between the uORF-containing and non-uORF-containing mRNAs suggest that uORFs regulate mRNA degradation in all tissues.

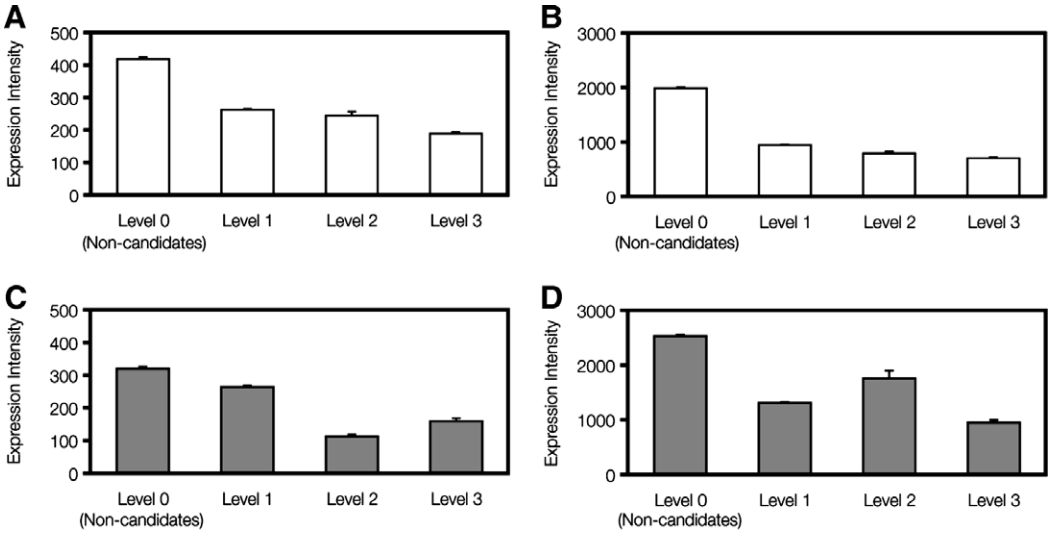

Fig. 2. Average expression intensities of (A, B) human and (C, D) mouse transcripts which contain putative uORFs in each level. Results for all transcripts are displayed in (A) and (C), and results for those having antisense transcripts are displayed in (B) and (D), respectively. The error bars are 95% posterior probability intervals.

Table 2
The number of putative uORF-regulated transcripts in which the RNA half-life in humans was examined

	Human (UniGene)	
Level 0	2003	(67.9%)
Level 1	907	(30.8%)
Level 2	10	(0.3%)
Level 3	28	(0.9%)
Total	2498	(100.0%)

Transcripts were separated into four levels.

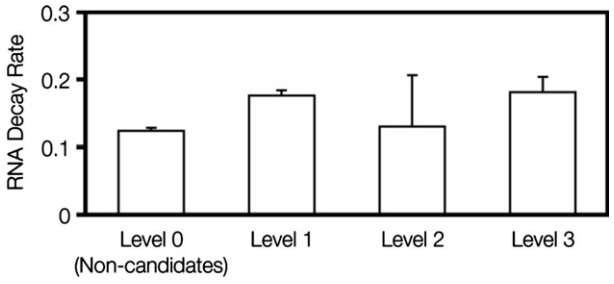

Fig. 3. Average RNA decay rates of human transcripts which contain putative uORFs in each level. The error bars indicate the 95% confidence limits.

The human transcript rates of decay were obtained from a previous report [24], and each transcript was linked to those in the UniGene database (for detail, see Table 2) in order to investigate differences in RNA half-life between the different categories. The average decay rates of candidates for each level are shown in Fig. 3. Although no marked difference was observed between the level 2 candidates and level 0 candidates, presumably due to the small amount of data (only 10 transcripts were identified as level 2 candidates), the decay rates of the level 1 and level 3 candidates were significantly higher than those of level 0 candidates (P-value < 0.01; two-sided t-test). This finding is consistent with the results for the expression levels. We therefore suggest that uORF-containing transcripts are generally degraded by a specific RNA degradation mechanism.

In conclusion, we have presented evidence that uORFs control the post-transcriptional levels of their downstream genes and promote the degradation of these transcripts. We further speculate that the mechanism of RNA decay promoted by uORFs is similar to the NMD pathway proposed previously [12].

Acknowledgements: This research was supported in part by the Japan Society for the Promotion of Science (JSPS). The authors are grateful to Katsunori Komatsu for technical support, and the members of the MGSP project at the Institute for Advanced Biosciences, Keio University, for critical discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2007.07.057.

References

[1] Linz, B., Koloteva, N., Vasilescu, S. and McCarthy, J.E. (1997) Disruption of ribosomal scanning on the 5′-untranslated region, and not restriction of translational initiation *per se*, modulates the stability of nonaberrant mRNAs in the yeast *Saccharomyces cerevisiae*. J. Biol. Chem. 272 (14), 9131–9140.
[2] Santiago, T.C., Purvis, I.J., Bettany, A.J. and Brown, A.J. (1986) The relationship between mRNA stability and length in *Saccharomyces cerevisiae*. Nucleic Acids Res. 14 (21), 8347–8360.
[3] Shapiro, R.A., Herrick, D., Manrow, R.E., Blinder, D. and Jacobson, A. (1988) Determinants of mRNA stability in *Dictyostelium discoideum* amoebae: differences in poly(A) tail length, ribosome loading and mRNA size cannot account for the heterogeneity of mRNA decay rates. Mol. Cell Biol. 8 (5), 1957–1969.
[4] Maquat, L.E. and Carmichael, G.G. (2001) Quality control of mRNA function. Cell 104 (2), 173–176.
[5] Mendell, J.T. and Dietz, H.C. (2001) When the message goes awry: disease-producing mutations that influence mRNA content and performance. Cell 107 (4), 411–414.
[6] Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F. and Dietz, H.C. (2004) Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat. Genet. 36 (10), 1073–1078.
[7] Lejeune, F., Li, X. and Maquat, L.E. (2003) Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. Mol. Cell 12 (3), 675–687.
[8] Vilela, C. and McCarthy, J.E. (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Mol. Microbiol. 49 (4), 859–867.
[9] Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T. and Sugano, S. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. 14 (10B), 2048–2052.
[10] Iacono, M., Mignone, F. and Pesole, G. (2005) uAUG and uORFs in human and rodent 5′ untranslated mRNAs. Gene 349, 97–105.
[11] Crowe, M.L., Wang, X.Q. and Rothnagel, J.A. (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics 7.
[12] Vilela, C., Linz, B., Rodrigues-Pousada, C. and McCarthy, J.E. (1998) The yeast transcription factor genes YAP1 and YAP2 are subject to differential control at the levels of both translation and mRNA stability. Nucleic Acids Res. 26, 1150–1159.
[13] Stockklausner, C., Breit, S., Neu-Yilik, G., Echner, N., Hentze, M.W., Kulozik, A.E. and Gehring, N.H. (2006) The uORF-containing thrombopoietin mRNA escapes nonsense-mediated decay (NMD). Nucleic Acids Res. 34 (8), 2355–2363.
[14] Kozak, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. Gene 299, 1–34.
[15] Kozak, M. (1984) Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin *in vivo*. Nature 308 (5956), 241–246.
[16] Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361, 13–37.
[17] Kent, W.J. (2002) BLAT– the BLAST-like alignment tool. Genome Res. 12 (4), 656–664.
[18] Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R. and Hogenesch, J.B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA 101 (16), 6062–6067.
[19] Pruitt, K.D. and Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29 (1), 137–140.
[20] Kerlavage, A., Bonazzi, V., Tommaso, M. di., Lawrence, C., Li, P., Mayberry, F., Mural, R., Nodell, M., Yandell, M., Zhang, J. and Thomas, P. (2002) The Celera discovery system. Nucleic Acids Res. 30 (1), 129–136.
[21] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002) FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team, Analysis of the mouse

transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420 (6915), 563–573.
[22] Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19 (2), 185–193.
[23] Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31 (4), e15.
[24] Yang, E., van Nimwegen, E., Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M. and Darnell Jr., J.E. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res. 13 (8), 1863–1872.
[25] Gaba, A., Jacobson, A. and Sachs, M.S. (2005) Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. Mol. Cell 20 (3), 449–460.
[26] Imanishi, T. et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol. 2 (7), e62.
[27] Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K.C., Hallinan, J., Mattick, J., Hume, D.A., Lipovich, L., Batalov, S., Engstrom, P.G., Mizuno, Y., Faghihi, M.A., Sandelin, A., Chalk, A.M., Mottagui-Tabar, S., Liang, Z., Lenhard, B. and Wahlestedt, C.RIKEN Genome Exploration Research Group, Genome Science Group, FANTOM Consortium (2005) Antisense transcription in the mammalian transcriptome. Science 309 (5740), 1564–1566.
[28] Kiyosawa, H., Mise, N., Iwase, S., Hayashizaki, Y. and Abe, K. (2005) Disclosing hidden transcripts: mouse natural sense–antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res. 15 (4), 463–474.