Published online 14 May 2008 Nucleic Acids Research, 2008, Vol. 36, No. 11 3707–3715
doi:10.1093/nar/gkn248

Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics

Riu Yamashita[1], Yutaka Suzuki[2], Nono Takeuchi[2], Hiroyuki Wakaguri[2], Takuya Ueda[2], Sumio Sugano[2] and Kenta Nakai[1,3,*]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-8562 and [3]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), 4-5-3 Chiyoda-ku, Tokyo, Japan

Received December 29, 2007; Revised March 25, 2008; Accepted April 17, 2008

ABSTRACT

Although the knowledge accumulated on the transcriptional regulations of eukaryotes is significant, the knowledge on their translational regulations remains limited. Thus, we performed a comprehensive detection of terminal oligo-pyrimidine (TOP), which is one of the well-characterized *cis*-regulatory motifs for translational controls located immediately downstream of the transcriptional start sites of mRNAs. Utilizing our precise 5′-end information of the full-length cDNAs, we could screen 1645 candidate TOP genes by position specific matrix search. Among them, not only 75 out of 78 ribosomal protein genes but also eight previously identified non-ribosomal-protein TOP genes were included. We further experimentally validated the translational activities of 83 TOP candidate genes. Clear translational regulations exerted on the stimulation of 12-*O*-tetradecanoyl-1-phorbol-13-acetate for at least 41 of them was observed, indicating that there should be a few hundreds of human genes which are subjected to regulation at translation levels via TOPs. Our result suggests that TOP genes code not only formerly characterized ribosomal proteins and translation-related proteins but also a wider variety of proteins, such as lysosome-related proteins and metabolism-related proteins, playing pivotal roles in gene expression controls in the majority of cellular mRNAs.

INTRODUCTION

Eukaryotic gene expressions are controlled at several levels. Compared with the knowledge on their transcriptional regulations, still limited knowledge is accumulated on their translational regulations. In this respect, there is an interesting set of genes: several vertebrate mRNAs which code ribosomal proteins or translation elongation factors (EF) have a 4–15 oligo-pyrimidine long tract on their 5′-end, and they are called TOP (terminal oligo-pyrimidine) genes (1–3). This sequence is thought to serve as a *cis*-regulatory element which inhibits the binding of translational regulatory proteins or the translational machinery itself. As a result, the translations of these genes are inhibited at the growth arrest of cells. More specifically, when a cell is faced with starvation or treated by some chemicals such as 12-*O*-tetradecanoyl-1-phorbol-13-acetate (TPA), mRNAs of TOP genes, which are normally associated with polysomes, change their state into the translationally inactive ‘sub-polysome’ while most non-TOP mRNAs stay in the ‘polysome’ state (1–4).

The TOP motif may also function as a part of a *cis*-regulatory element for transcription. For example, in a typical TOP gene, EF1-A, at least three T’s in the tract must exist for its high transcription activity (5). In fact, EF1-A is known as one of the most highly expressed genes in a cell, and its promoter region has remarkable activity for transcription (6). In addition, the conservation around the transcriptional start sites (TSSs) of ribosomal genes extends to upstream untranscribed regions, such as $(Y)_2$ | $CTY(T)_2(Y)_3$,where ‘|’ denotes TSS. The possibility that this motif is bound by some transcription factors has been implicated (7). Therefore, the TOP motif is unique in cooperatively controlling the gene expressions at both the transcription and translation levels.

In spite of potential importance and interests in the balance between the transcriptional and translational regulations there are only few estimations about the total number of translationally controlled genes or TOP genes, in particular, which include all of ribosomal protein

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: knakai@ims.u-tokyo.ac.jp genes and may be one of the largest groups of translationally regulated genes. Although many studies have been reported on the pathway of TOP-dependent translational regulation (8–11), they did not aim at comprehensive identification of those genes in the human genome. Indeed, there have been a few pre-genome era studies, which aimed at this subject. By finding a simple pattern C_mT_nC (where $m = 0,1,2$ and $n > 0$) against 1496 human full-length cDNAs, Kato *et al.* (12) identified 21 TOP genes besides ribosomal proteins. Based on this, Amaldi and Pierandrei-Amaldi (1) estimated that the total number of TOP genes should be at least 100. However, no update has been made since then and, to date, genes coding most ribosomal proteins (2,13), translation EFs (EF1-A, EF1-B and EF-2), hnRNP A1, laminin receptor 1, nucleophosmin 1, polyA binding protein 1 and tumor protein translationally controlled 1 have been considered as the only TOP genes (3).

For the genome-wide identification and characterization of TOP genes, it is essential to know the accurate position of TSSs because pseudo-TOP motif could occur frequently by chance on 5′-UTRs/upstream sequences. Thus, the 5′-end sequence information of full-length cDNAs is quite valuable. We have been collecting and analyzing 5′-end clones obtained by the oligo-capping method (14) as well as the cap trapper method (15). Our database, DataBase of Transcription Start Sites (DBTSS), contains a large number of 5′-end clones which are used to identify accurate TSSs in the genomes of various species (16–18). Using this information, we performed the first post-genome era comprehensive detection of TOP gene candidates in the human genome as well as their verification with a sedimentation experiment.

MATERIALS AND METHODS

Datasets for human and mouse TSSs

In DBTSS version 5.1, there are 425 117 human TSSs corresponding to 19 573 NCBI reference sequence (RefSeq) genes. Among them, we first used the set of 921 genes (921 TSSs) that have more than 10 clones, of which more than a half start from the same TSS. This dataset contained 48 known TOP genes and 873 presumed non-TOP genes. For the screening of TOP candidates, we used only TSSs that were indicated by multiple clones to increase reliability. This reduced the TSS number to 87 397 (13 717 RefSeq genes) for the human genome.

We also used mouse TSS information from DBTSS. 149 876 TSSs, which correspond to 14 745 mouse genes, are registered. Because the number of 5′-end clones is relatively small for mouse, we used all TSSs to detect TOP gene candidates.

Algorithms

A position specific weighted matrix (PSWM) of the TOP motif was constructed based on 48 known TOP genes as shown in Figure 1C. The score was calculated by the following formula:

$$\text{score} = \sum_{i=1}^{l} \log\left(\frac{(n_{if}+1)}{(N_i+4)^*(1/4)}\right)$$

where l: length of the matrix, which is taken to be 11; n_{if}: the observed number of base 'f' (A, C, G or T) at the i-th position in the training; N_i: sum of the observed number of all bases at i-th position, which is equal to 48; the base of the logarithm was taken to be 10. We defined a gene as a TOP gene candidate if (i) its score >0.1, (ii) its +1 position is 'c' (iii) the positions −1 to +4 are pyrimidines.

Gene expression preference using expression breadth

In order to estimate tissue specificity of the genes, we used 'expression breadth' (19,20). First, we obtained human gene expression data, based on Affymetrix microarray data, from the mammalian gene expression atlas (http://symatlas.gnf.org/SymAtlas/) (21). The data contains expression data for 79 human tissues. Affymetrix probe IDs were connected with RefSeq ID using the annotation table in the database. We defined that a gene which showed on expression level ≥200 in a given tissue is expressed, and counted the number of tissues in which it is expressed.

Gene Ontology (GO) annotation

We obtained the information of GO terms (22) from the 'gene2go' and 'gene2refseq' tables in NCBI (http://www.ncbi.nlm.nih.gov/). The 'GOslim' information in European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/GOA/downloads.html) was used to simplify the GO annotation by obtaining their top-level GO terms. Since some GO terms did not correspond to GO slim terms, we added such GO terms to our extended GO slim set. The *P*-value for each GO term was calculated assuming the hypergeometric distribution and was corrected according to the Bonferroni Correction (http://mathworld.wolfram.com/BonferroniCorrection.html)

Preparation of polysomal and sub-polysomal fractions

We used exactly the same methods and samples as those previously reported (23). Namely, HL-60 rapid growth cell (RG) derived from HL-60 cell lines was cultured. TPA treated cell lines (TPA+) and untreated cell lines (TPA−) were prepared. Fractionation of HL-60RG polysomes and isolation of RNA contained in the fractions were carried out with a modified published protocol (24). Approximately 3×10^7 cells were used for each gradient (growing cells, 4×10^5 cells/ml; TPA+ cells, 50 nM TPA, 48 h). Before harvesting, cells were incubated with the medium containing 100 µg/ml cycloheximide for 5 min and washed twice with PBS containing 100 µg/ml cycloheximide. Cell pellets were resuspended in 1 ml of lysis buffer (20 mM Tris-HCl (pH 7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.04 M sucrose, 0.5% Nonidet P40, 1 mM dithiothreitol) containing 100 units of RNase inhibitor and lysed by incubation on ice for 10 min with occasional shaking. Nuclei and cell debris were removed by centrifugation at 1000*g* for 10 min at 4°C. The lysate was layered on top of a 11 ml 15–50% (w/v) sucrose gradient and centrifuged at 36 000 rpm in a Beckman SW41Ti rotor for 2 h 15 min at 4°C.

Using a density gradient fractionator (Model 152–001 Towa Labo, Misaki-cho, Chiyoda-ku, Tokyo, Japan), gradients were separated into 11 equal fractions with monitoring absorbance at 260 nm. Each fraction was treated with proteinase K, and RNA was extracted by phenol/$CHCl_3$, precipitated with ethanol, and analyzed for each mRNA species. We defined the fractions 1–5 as the sub-polysomal fraction and 7–11 as the polysomal fraction, as shown in Figure 3.

Quantitative RT–PCR for detecting enrichment of the TOPs in the sub/polysome fractions

For negative controls, we sorted genes in DBTSS according to the number of 5′-ESTs expressed in HL-60RG cells. The top nine genes which were not TOP candidates were chosen. We also added the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene as a negative control. PCR primers were constructed by Primer3 (25).

Twenty microlitres PCR mixture contained 1× Power SYBR Green PCR mixture (Applied Biosystems, Lincoln Centre Drive Foster City, CA, USA), 1 ng/μl cDNA sample, and 0.125 pM primers. PCR amplification consisted of pre-heating (50°C for 2 min, 95°C for 10 min), and 35 cycles of 95°C for 30 s, 57°C for 1 min, 72°C for 1 min. We used the HT7000 Sequence Detection System (ABI PRISM) to measure the expression level of mRNAs. The HT7000 reports the number of the cycle when the first fluorescence is observed. We regarded the ratio of these values between TPA+ samples and TPA− samples as the expression difference between the two states. All data including experimental results and predicted TOP gene candidates can be downloaded at ftp://ftp.hgc.jp/pub/hgc/db/dbtss/Yamashita_et_al.

To estimate the rough number of ribosome on ORF or mRNAs, first we used the refGene.txt table from University of California Santa Cruz genome browser (http://genome.ucsc.edu/) to obtain ORF and mRNA length. Then we estimated 1 ribosome/150 bp, as was previously reported in eukaryotic cells (26).

RESULTS

Initial set of TOP genes

It is now established that most genes have multiple TSSs (27). For example, tumor protein translationally controlled 1, which is a known TOP gene, has a major TSS at 44 813 318 on minus-strand of chromosome 13, and more than 30 TSSs are observed around it (Supplementary Material Figure 1). Thus, to construct a firm criterion for finding TOP gene candidates, we initially selected 921 genes that have a dominant and stable TSS; more specifically, (i) genes whose TSS(s) was determined by no less than 10 clones were chosen and (ii) those where at least a half of these clones start from the same TSS were further selected (the list of these genes is shown in the Supplementary information 1). This gene set included 48 known TOP genes, which code 45 ribosomal proteins, tumor protein translationally controlled 1, eukaryotic translation EFs 1 and 2. Since they correspond to 56% of known or suggested TOP genes, TOP genes may tend to have dominant TSSs.

With the Sequence logo representation, these 48 genes are shown to have a clear TOP motif while the other genes seem to be characterized with the initiator sequence (28) (Figure 1B). The consensus motif sequence is almost the same with the previously-reported $(Y)_2$ | $CTY(T)_2(Y)_3$, which was obtained from the ribosomal protein genes only (7). To detect all potential TOP genes in our dataset, we constructed a PSWM based on the conserved −4 to +7 region of these 48 genes (Figure 1C and D). With this PSWM, the minimum score of known TOP genes was 0.12 (NM_021029: ribosomal protein L36, Supplementary Material Figure 2); thus, (i) we set our threshold value to 0.1. In addition, (ii) all TOP genes have C at +1 position and (iii) the bases from −1 to +4 were all pyrimidines. Therefore, we combined these three criteria to detect candidates. In the remaining 873 genes in our initial gene set, we further detected with these criteria 18 novel TOP gene candidates, which include eukaryotic translation initiation factors 2, 3 and 4A.

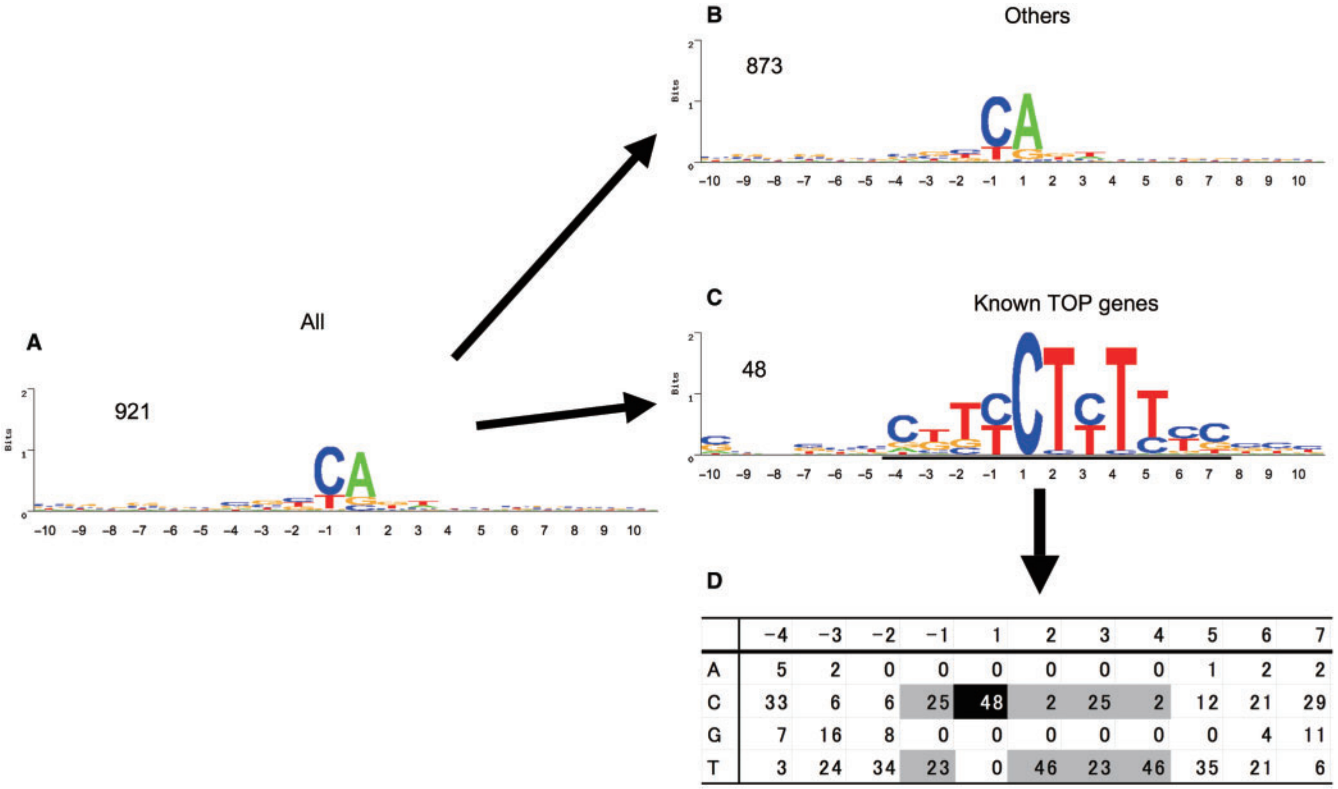
Detection of TOP gene candidates from all TSSs

The distribution of the scores of randomly sampled genome sequences (for a million times) showed the average −0.87 and the SD 0.53, which means that genes whose score exceed 0.1 are TOP gene candidates ($P < 0.05$). We then calculated the score for all 87 397 TSSs corresponding to 13 717 human genes (Note that they include all of the above 921 TSSs). Of these, 1645 genes (2772 TSSs) fulfilled the criteria (Supplementary Material table 2). They contained not only all known TOP genes but also 75 genes coding ribosomal proteins and 22 translation-related genes (such as the translation initiation factors or EFs) (Table 1, panel A, B).

We then detected 816 TOP gene candidates (967 TSSs) from 149 876 mouse TSSs (14 745 genes) under the same conditions. In the 1645 human TOP gene candidates, 1314 of them had homologous counterparts in the 14 745 mouse genes. But only 239 of them were also identified as TOP candidates in mouse. This subset contained 54 of the 58 known-TOP genes with mouse homologs. This consensus gene set also included 49 genes coding ribosomal proteins and 9 translation-related genes.

GO annotation and expression profiles of TOP gene candidates

We obtained two kinds of datasets for TOP gene candidates: the 1645 genes relying on only human information and the 239 genes relying on both human and mouse information. To characterize these TOP genes candidates, we used GO annotations. As shown in Table 2, panel A, 10 GO terms such as ‘protein biosynthesis’, ‘ribosome’ and ‘regulation of translation initiation’ were significantly frequent. Though the frequency of ‘translation elongation’ was not regarded as significant with our threshold, it was in the 11th position. The same analysis was performed using the GO slim terms for the detection of a more global and unbiased tendency (Table 2, panel B). Among those, terms such as ‘receptor activity’, ‘channel or pore class transporter activity’ and ‘cell communication’ were less frequent while ‘lysosome’, ‘cytoplasm’, ‘intracellular



	−4	−3	−2	−1	1	2	3	4	5	6	7
A	5	2	0	0	0	0	0	0	1	2	2
C	33	6	6	25	48	2	25	2	12	21	29
G	7	16	8	0	0	0	0	0	0	4	11
T	3	24	34	23	0	46	23	46	35	21	6

Figure 1. Sequence logo of the most fixed TSSs and known TOP genes. (**A**) We constructed the Sequence logo of the most fixed 931 TSSs corresponding to 931 genes. 873 other genes (**B**) and 48 TOP genes group (**C**) and (**D**) PSWM of TOP genes. The black box at +1 position in the table shows all of the TOP genes showing same nucleotide namely 'C'. The gray boxes in the table from −1:+4 shows the pyrimidine region. These two conditions were considered to detect TOP gene candidates.

activity', 'structure molecular activity' or 'translation regulator activity' were more frequent.

To characterize the TOP candidates, we used 'expression breadth' to compare the expression preference between TOP and not-TOP genes. As shown in Figure 2, TOP candidates tend to be expressed in more tissues, in other words, in a more house-keeping manner while the other genes tend to be more tissue-specific. These tendencies were observed in both the 1645 group and the 239 group (data not shown) and were highly significant by Wilcoxon's rank-sum test (both of them showed $P < 1e{-}200$). We also counted the number of cDNA libraries where the 5′-end clones of each gene in DBTSS were obtained, and obtained the same tendency (Supplementary material Figure 3).

Experimental validation for translational regulation

We experimentally validated whether translational controls were observed upon TPA treatment for some of our predicted TOP gene candidates. The translational controls were evaluated by sedimentation experiments. For this, mRNAs were extracted for every fraction of either subpolysomes or polysomes (Figure 3A) before and after the TPA treatment in human HL-60RG cells (23) and the relative abundance of them was detected by semiquantitative real-time PCR. Before the sedimentation experiment, the RNA levels (transcription level) of all 239 genes with and without TPA treatment (TPA+/TPA-) were measured by real-time PCR. In the following experiments, we used 86 genes whose transcription level did not vary significantly on the TPA treatment.

Figure 3B depicts the results of the sedimentation experiments for two new TOP gene candidates (EEF1G and SDBCAG84), one known TOP gene (RPL19) and one negative control (GAPDH) (For further details on fraction distributions, see Supplementary Material Figure 4). Distributions of the mRNAs of RPL19, EEF1G and SDBCAG84 shifted to the sub-polysomal fractions after TPA treatment. In contrast, significant population of the mRNAs of GAPDH remained at the polysomal fractions. These results are consistent with the general notion of TOP genes. In the cases of the mRNAs of EEF1G and SDBCAG84, which are newly identified TOP candidates, the degree of translational controls estimated by the ratio between the sub-polysomal and polysomal fractions (Sub/Pol) were more than 1000-fold higher than the positive controls.

In total, we examined translational controls for 91 genes, which consist of 47 newly predicted genes, 34 known TOP genes and 10 negative controls. It contains RPL19, RPL13a and RPS27 genes which are well experimentally established as TOP genes (4). The mRNAs of most of the genes examined were enriched in the polysomal fraction. After TPA treatment, mRNAs of 63 genes (78%) showed some shift towards the sub-polysomal

Table 1. Relationship between TOP gene candidates and translation-related genes

RefSeq ID	Gene name	Definition
Panel A		
NM_001402	EEF1A1	Eukaryotic translation elongation factor 1-α
NM_001958	EEF1A2	Eukaryotic translation elongation factor 1-α
NM_001959	EEF1B2	Eukaryotic translation elongation factor 1-β
NM_001960	EEF1D	Eukaryotic translation elongation factor 1-δ
NM_001404	EEF1G	Eukaryotic translation elongation factor 1
NM_001961	EEF2	Eukaryotic translation elongation factor 2
NM_003907	EIF2B5	Eukaryotic translation initiation factor 2B,
NM_003908	EIF2S2	Eukaryotic translation initiation factor 2-β
NM_001415	EIF2S3	Eukaryotic translation initiation factor 2,
NM_013234	eIF3k	Eukaryotic translation initiation factor 3
NM_003758	EIF3S1	Eukaryotic translation initiation factor 3,
NM_003750	EIF3S10	Eukaryotic translation initiation factor 3,
NM_003757	EIF3S2	Eukaryotic translation initiation factor 3,
NM_003756	EIF3S3	Eukaryotic translation initiation factor 3,
NM_003755	EIF3S4	Eukaryotic translation initiation factor 3,
NM_003754	EIF3S5	Eukaryotic translation initiation factor 3,
NM_016091	EIF3S6IP	Eukaryotic translation initiation factor 3
NM_003753	EIF3S7	Eukaryotic translation initiation factor 3
NM_001967	EIF4A2	Eukaryotic translation initiation factor 4A,
NM_001417	EIF4B	Eukaryotic translation initiation factor 4B
NM_001418	EIF4G2	Eukaryotic translation initiation factor 4
NM_022170	WBSCR1	Eukaryotic translation initiation factor 4H
Panel B		
NM_004280	EEF1E1	Eukaryotic translation elongation factor 1
NM_001412	EIF1AX	X-linked eukaryotic translation initiation
NM_004681	EIF1AY	Eukaryotic translation initiation factor 1A, Y
NM_004836	EIF2AK3	Eukaryotic translation initiation factor 2-α
NM_001414	EIF2B1	Eukaryotic translation initiation factor 2B,
NM_014239	EIF2B2	Eukaryotic translation initiation factor 2B,
NM_020365	EIF2B3	Eukaryotic translation initiation factor 2B,
NM_015636	EIF2B4	Eukaryotic translation initiation factor 2B,
NM_012199	EIF2C1	Eukaryotic translation initiation factor 2C, 1
NM_012154	EIF2C2	Eukaryotic translation initiation factor 2C, 2
NM_024852	EIF2C3	Eukaryotic translation initiation factor 2C, 3
NM_017629	EIF2C4	Eukaryotic translation initiation factor 2C, 4
NM_004094	EIF2S1	Eukaryotic translation initiation factor 2,
NM_003752	EIF3S8	Eukaryotic translation initiation factor 3,
NM_003751	EIF3S9	Eukaryotic translation initiation factor 3,
NM_001416	EIF4A1	Eukaryotic translation initiation factor 4A,
NM_001968	EIF4E	Eukaryotic translation initiation factor 4E
NM_004095	EIF4EBP1	Eukaryotic translation initiation factor 4E
NM_004096	EIF4EBP2	Eukaryotic translation initiation factor 4E
NM_003732	EIF4EBP3	Eukaryotic translation initiation factor 4E
NM_004846	EIF4EL3	Eukaryotic translation initiation factor 4E-like
NM_019843	EIF4ENIF1	Eukaryotic translation initiation factor 4E
NM_004953	EIF4G1	Eukaryotic translation initiation factor 4
NM_003760	EIF4G3	Eukaryotic translation initiation factor 4
NM_001969	EIF5	Eukaryotic translation initiation factor 5
NM_001970	EIF5A	Eukaryotic translation initiation factor 5A
NM_015904	EIF5B	Translation initiation factor IF2
NM_005801	SUI1	Putative translation initiation factor
NM_005726	TSFM	Ts translation elongation factor, mitochondrial
NM_003321	TUFM	Tu translation elongation factor, mitochondrial

The columns show, respectively, RefSeq ID: NCBI reference sequence ID, gene name: gene name in short form, definition: definition of the gene. Panel A detected TOP gene candidates. Panel B genes not detected as TOP gene candidates.

fraction. Especially in 40 genes (49%) the sub-polysomal fraction became the major fraction (Figure 3C). Among them, 31 of 34 known TOP genes were included. On the other hand, mRNAs of all of the negative control genes still remained in the polysomal fraction (Figure 3D). Interestingly, some of the new TOP candidates showed no significant increase in the ratio of Sub/Pol in the fractions corresponding to the ORF regions. For example, in the case of the TMEM18 mRNA, which has relatively long 3′-UTR (1645 bases out of 2122), the ratio was 1.2 (the most bottom line of Figure 3B). However, in this case, when the ratio of Sub/Pol corresponding to the entire mRNA was assessed, it increased to 1374. This observation may suggest that in some mRNAs, ribosomes are also associated outside of the ORF and play roles in controlling the translational efficiency.

Table 2. GO analysis of TOP genes

GO ID	Category	Definition	All	1645 genes		239 genes	
				No. of genes	*P*-value	No. of genes	*P*-value
Panel A							
GO:0016499	Process	Protein biosynthesis	249	127	7.1E−52	60	4.0E−52**
GO:0016505	Function	Structural constituent of ribosome	144	86	1.4E−42	52	2.2E−55**
GO:0016511	Component	Ribosome	116	73	1.6E−38	42	1.1E−44**
GO:0016519	Component	Cytosolic large ribosomal subunit (sensu Eukaryota)	25	22	9.0E−18	16	8.7E−23**
GO:0016520	Component	Cytosolic small ribosomal subunit (sensu Eukaryota)	12	12	9.3E−12	9	2.9E−14**
GO:0016559	Function	RNA binding	382	91	7.8E−11	33	2.8E−14**
GO:0016584	Process	Regulation of translational initiation	23	14	3.8E−08	2	6.1E−02*
GO:0016600	Function	Translation initiation factor activity	54	21	4.6E−07	4	1.5E−02*
GO:0016601	Function	rRNA binding	11	8	5.3E−06	6	1.2E−08**
GO:0016624	Component	Lysosome	97	28	6.8E−06	3	2.4E−01*
GO:0016742	Process	Translational elongation	15	9	1.4E−05	4	1.10E−04
Panel B							
GO:0004872	*Function*	*Receptor activity*	*1065*	*88*	*2.0E−05*	*12*	*6.0E−02**
GO:0015267	*Function*	*Channel or pore class transporter activity*	*312*	*16*	*2.0E−05*	*2*	*8.5E−02**
GO:0007154	*Process*	*Cell communication*	*2412*	*242*	*3.0E−04*	*37*	*2.1E−01**
GO:0005737	Component	Cytoplasm	2923	520	6.8E−25	128	1.2E−27**
GO:0009058	Process	Biosynthesis	972	226	3.6E−24	78	1.2E−31**
GO:0005198	Function	Structural molecule activity	542	135	3.3E−17	61	1.9E−32**
GO:0005622	Component	Intracellular	6139	879	2.7E−13	160	1.1E−11**
GO:0043170	Process	Macromolecule metabolism	3342	507	3.9E−10	107	7.8E−12**
GO:0045182	Function	Translation regulator activity	98	29	2.7E−06	8	3.4E−04**
GO:0008152	Process	Metabolism	5838	782	2.7E−05	144	6.6E−08**

The columns show, respectively, GO ID: GO ID, category: one of the category of the GO ID, definition: definition of GO ID, all: observed number of terms in whole of the gene set, 1645 genes *P*-value: the *P*-value of 1645 gene set according to hypergeometrical test, 239 genes p-value: the *P*-value of the 239 gene set according to hypergeometrical test. All of the GO terms which are overrepresented (normal) or underrepresented (italic) under the threshold ($P < 0.05$ with Bonferroni correction) are shown. * and ** indicate statistical significance in either of the two sets and in both sets, respectively. Panel A all of the 4753 GO observed in the dataset. After Bonferroni correction, we set the threshold to 1.1e−5. Panel B The results of 33 GO slims. After Bonferroni correction, we set the threshold to 1.5e−3.

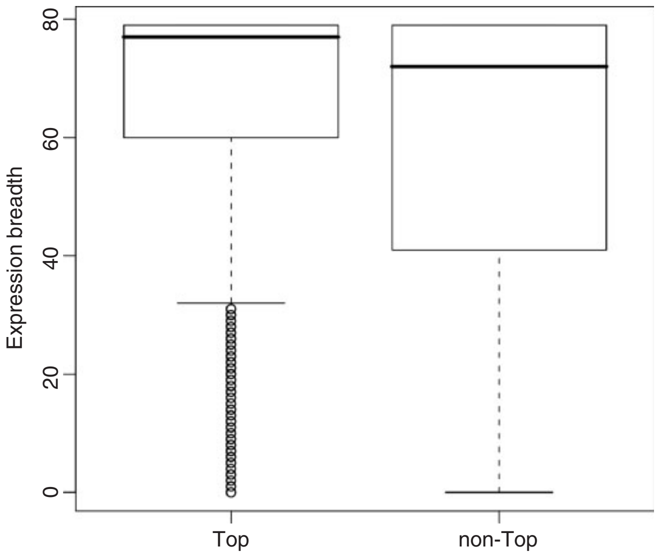


Figure 2. Tissue specificity of TOP genes candidates with a box-and-whisker plot. The horizontal axis shows 1645 TOP genes candidates (TOP) and 6174 not TOP candidates (not-TOP). The vertical axis shows 'expression breadth'. This figure is drawn with 'boxplot' in R package.

Effects of mRNA lengths on the change of translation status

We examined the correlation between the TOP activity and the lengths of 3′-UTR, 5′-UTR, ORF and RNA. Ledda *et al.* (29) previously reported that the 3′-UTR length of mRNAs coding TOP genes affects their translational efficiency. We also observed a significant correlation between the length of 3′-UTR and the ratio between the two fractions in our data of 84 TOP gene candidates ($r = -0.56$) (Figure 4A). Moreover, we observed similar correlation between the translational efficiency and total mRNA length ($r = -0.61$), the ORF length ($r = -0.53$) and the 5′-UTR length ($r = -0.42$) (Figure 4B–D). The fact that the correlation coefficient with mRNA length was the greatest may indicate that entire mRNA parts are involved in translational control as well as the UTR and ORF parts in many cases.

DISCUSSION

This is the first article which describes the identification of the translational regulatory motifs by utilizing TSS information. Based on exact positional information of the TSSs, we could identify and characterize TOP genes in a highly accurate and comprehensive manner. Previously, use of the TSS information has been mainly focused on the analysis of transcriptional regulatory elements in promoters. We showed that TSS information is also advantageous to identify *cis*-regulatory elements in mRNAs, which also play roles in determining the expression levels of final protein products.

So far, there were only eight genes reported to be human TOP genes besides ribosomal protein genes (3).

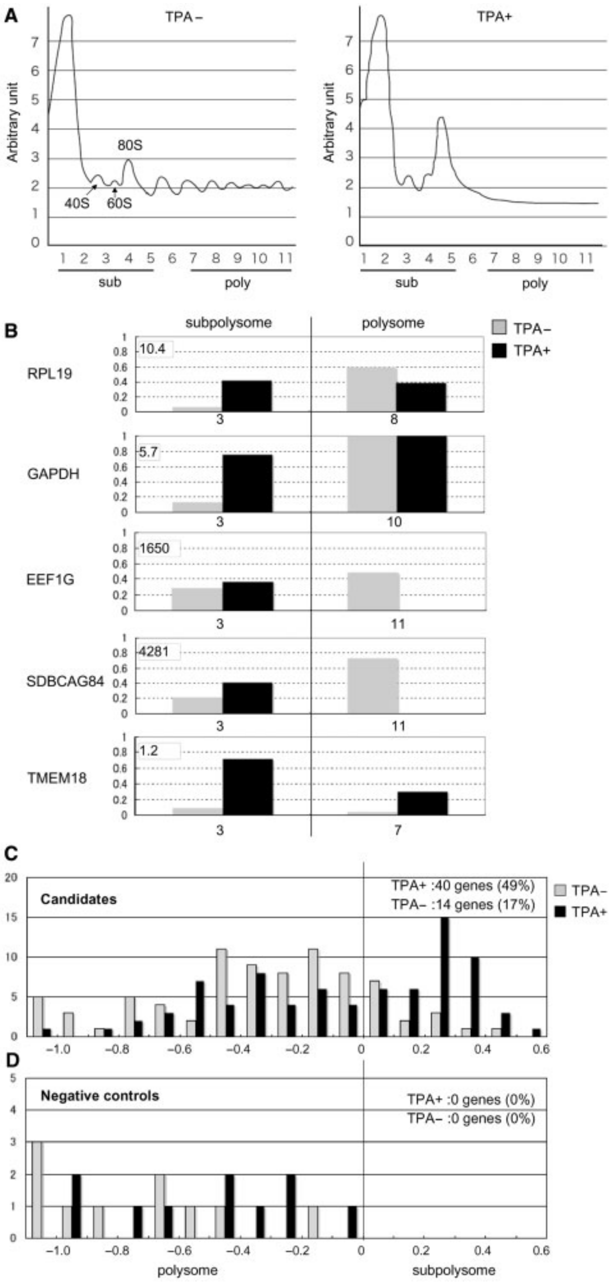


Figure 3. Expression profiles for TOP genes candidates. (**A**) HPLC fraction of cell elutions. The fractions were divided into Sub: sub-polysomal groups (1–5) and Poly: polysomal groups (7–11). Left: fraction of cells not treated with TPA, Right: fraction of cells treated with TPA. (**B**) Several examples of expression of detected TOP genes candidates. Relative expression levels, the highest one corresponding to 1, are shown. RPL19: ribosomal protein L19 for positive control, GAPDH for negative control, SDBCAG84: serologically defined breast cancer antigen 84, EEF1G: eukaryotic translation EF1-G, TMEM30A: transmembrane protein 30A. We showed the fraction corresponding to the potential ribosome number on ORF assuming one ribosome/150 bp. The number in each figure indicates the ratio of (TPA + :Sub/TPA + :Pol)/(TPA−:Sub/TPA−:Pol). For further details on fraction distributions, see Supplementary Research Data Figure 3. (**C**) Expression ratios of 81 candidates. (**D**) Expression ratios of 10 negative controls. Black bars correspond to the expression ratio with TPA treatment, and gray bars correspond to the expression ratio without TPA treatment. The *y*-axis shows the ratio of mRNA expression in polysome and sub-polysome fractions. These ratios were converted as log ratio.

How many TOP genes exist in the entire genome remained to be elucidated. We performed a systematic survey using the accurate TSS information. We detected as much as 1645 candidates, which included 83 (97%) of 86 known TOP genes. The three missing genes were overlooked because their clones were mapped to multiple loci and were not stored in DBTSS. Therefore, we believe that there are not many false negatives in our candidate set.

In the 1645 gene set, several plausible genes were included. For example, there were many translation initiation factors, such as initiation factors 2, 2β, 2B, 3, 4, 4A, 4B and 4H (Table 1, panel A). On the other hand, initiation factors 2C, 4E and 5 were not detected as TOP genes (Table 1, panel B). There are 11 paralogous genes coding initiation factor 3 in RefSeq and 9 of them were detected as TOP gene candidates. Although initiation factors have not been reported as TOP genes, it is very likely because all known TOP genes are related to the translational activity.

In a previous study, Levy *et al.* (30) reported that a purine, mostly a G, was frequently found at the end of pyrimidine stretch. In fact, they reported 8 out of 12 TOP genes had ‘G’ at the end of pyrimidine stretch. To check this idea, we picked up the first base after the pyrimidine stretch of each TOP candidates. From 1645 genes, we obtained 2772 TSSs which showed TOP positive. We observed 891 ‘A’ terminal and 1881 ‘G’ terminal among them, so that similarly to the previous research, we also found frequent ‘G’: the ratio was A:G = 1:2.11.

Amaldi and Pierandrei-Amaldi (1) estimated that the number of TOP genes are no less than 100. Davuluri *et al.* (31) predicted 152 (6.6%) TOP genes from a 2312 full-length gene set. Our results suggest that there are much more TOP genes beyond the category of so-called ‘translation-related’ genes. For example, we identified transmembrane protein 30 as a TOP candidate. This gene is a homolog of yeast CDC50 genes, which is necessary for subcellular localization of yeast Bni1p and plays a pivotal role in asymmetrical cell division (32). Although their gene functions in humans remain elusive, it is possible that the wider variety of genes involved in various cellular functions, such as transmembrane proteins and signal transducing proteins, are also subjected to translational regulations unlike previously thought. Of course, there still remains some possibilities that our set contains a significant number of false positives. Indeed, there were only 239 genes that are predicted to be TOP genes in both human and mouse. However, it is as well possible that translational regulations might be even more evolutionary diverged than transcriptional regulations, for which significant species-specific traits have been identified (33). If our estimation of the frequency of the TOP genes is totally reliable, the translational control of the mRNAs should take place very abundantly within human cells. We showed that the mRNA expression levels of the TOP genes are generally higher and ubiquitously expressed (Figure 2). It has been estimated that ∼15% of total cellular mRNAs is occupied with ribosomal protein mRNAs (34); therefore, >20% of total mRNA could be translationally regulated (35). Based on our prediction, the percentage is even higher: 41.9% of total 1 034 085 5′-end

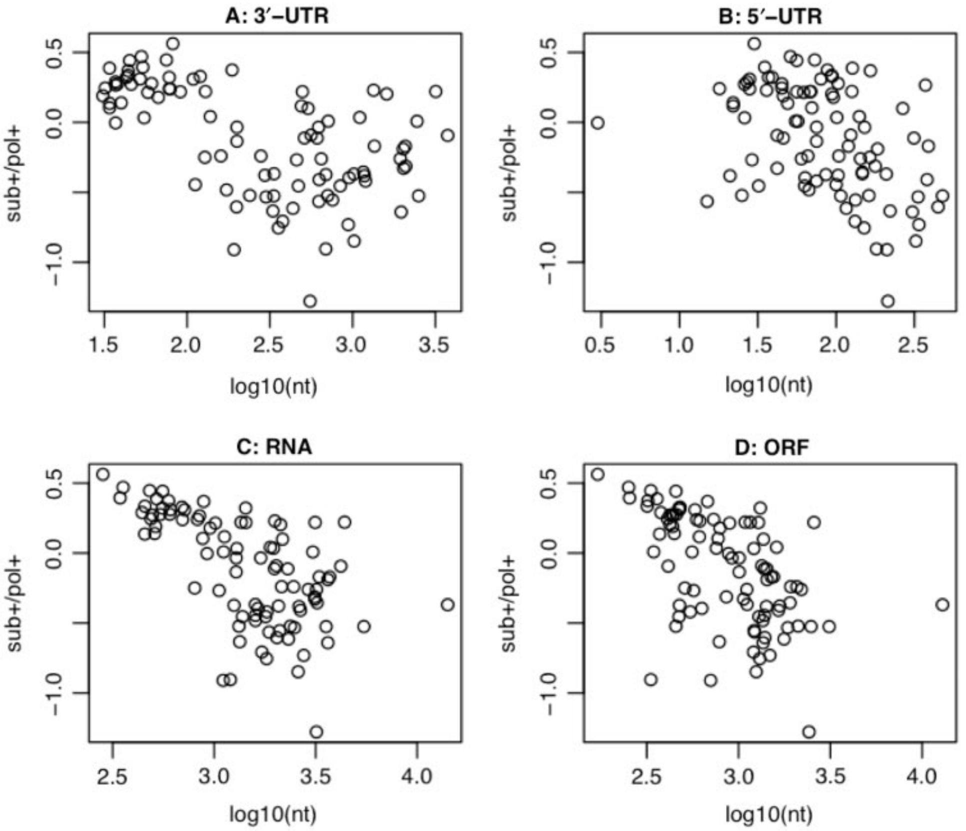


Figure 4. Correlation between translational regulation and mRNA features. Figures (**A–D**) show the correlation between translational regulation and the 3′-UTR, 5′-UTR, RNA and ORF length, respectively. The horizontal axis shows each mRNA length, and the vertical axis shows the ratio of mRNA level in sub-polysome and polysome fractions. The correlation coefficients were 3′-UTR: −0.56 ($P < 1.1e{-}8$), 5′-UTR: −0.42 ($P < 2.9e{-}5$), RNA: −0.61 ($P < 1.1e{-}12$), ORF: −0.53 ($P < 5.1e{-}8$).

clones in our cDNA collection could be translationally regulated.

To our surprise, the extent of the translational induction differed between genes. It seems that some of the 239 candidates showed relatively small change of translational status upon TPA treatment (Figure 3B). It seems that the TPA effect on TOP mRNAs is dependent on several factors. Ledda *et al.* (29) reported that the 3′-UTR length affects the translational regulation of TOP genes. We could confirm this with our own data but the correlation was even stronger with the mRNA length (Figure 4). Considering the fact that the correlation between the mRNA length and the degree of the ‘TOP-ness’ is the most strict (left upper corner in Figure 4C), the TOP effects seem to be the most straightforward for the mRNAs that are less than 1000 bases long with relatively short UTRs. Actually, this population contains 64 (56%) known TOP genes. Furthermore, we also observed that in some mRNAs the TOP-mediated depletion of the polysomes are not clear in the ORF regions but evident when the entire mRNA regions, including UTRs, were assessed. When a *trans*-factor binds to the 5′-end of TOP mRNAs under particular cellular conditions, it invokes the dissociation of bound ribosomes; in TOP genes of longer mRNAs or 3′-UTR, however, ribosomes might remain associated outside of the ORF regions, which are hard to be dissociated, thus, showing some buffering effects on the TOP activity.

In this study, we performed a genome-wide detection of human TOP genes. Our results strongly support the idea that the TOP genes are not restricted to genes of limited functional category, but control the expression of a wider variety of genes and a major population of cellular mRNAs. Further analyses on the additional factors to modulate the TOP activity of the newly identified candidates should also reveal underlying molecular mechanisms which realizes the translational control at the versatile levels. Our study should lay foundation for an in-depth understanding of the genome-wide figure of translational controls of gene expressions, for which relatively little knowledge has been accumulated.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the members of Dynacom Co. Ltd. for technical supports in the database construction. We also thank Nicolas Sierro for critical reading of the manuscript. Computation time was provided by the Super Computer

System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was supported by Grants-in-Aid for Young Scientists (B) from the Ministry of Education, Science, Sports and Culture of Japan and BIRD of Japan Science and Technology Agency (JST). Funding to pay the Open Access publication charges for this article was provided by Grants-in-Aid for Young Scientists (B) from the Ministry of Education, Science, Sports and Culture of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Amaldi,F. and Pierandrei-Amaldi,P. (1997) TOP genes: a translationally controlled class of genes including those coding for ribosomal proteins. *Prog. Mol. Subcell Biol.*, **18**, 1–17.
2. Meyuhas,O., Avni,D. and Shama,S. (1996) *Translational Control of Ribosomal Protein mRNAs in Eukariotes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Meyuhas,O. and Hornstein,E. (2000) *Translational Control of TOP mRNAs*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
4. Krichevsky,A.M., Metzer,E. and Rosen,H. (1999) Translational control of specific genes during differentiation of HL-60 cells. *J. Biol. Chem.*, **274**, 14295–14305.
5. Shibui-Nihei,A., Ohmori,Y., Yoshida,K., Imai,J., Oosuga,I., Iidaka,M., Suzuki,Y., Mizushima-Sugano,J., Yoshitomo-Nakagawa,K. and Sugano,S. (2003) The 5′ terminal oligopyrimidine tract of human elongation factor 1A-1 gene functions as a transcriptional initiator and produces a variable number of Us at the transcriptional level. *Gene*, **311**, 137–145.
6. Kim,D.W., Uetsuki,T., Kaziro,Y., Yamaguchi,N. and Sugano,S. (1990) Use of the human elongation factor 1 alpha promoter as a versatile and efficient expression system. *Gene*, **91**, 217–223.
7. Perry,R.P. (2005) The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.*, **5**, 15.
8. Ruvinsky,I. and Meyuhas,O. (2006) Ribosomal protein S6 phosphorylation: from protein synthesis to cell size. *Trends Biochem. Sci.*, **31**, 342–348.
9. Thomas,G. (2000) An encore for ribosome biogenesis in the control of cell proliferation. *Nat. Cell Biol.*, **2**, E71–E72.
10. Edgar,B.A. (1999) From small flies come big discoveries about size control. *Nat. Cell Biol.*, **1**, E191–193.
11. Hamilton,T.L., Stoneley,M., Spriggs,K.A. and Bushell,M. (2006) TOPs and their regulation. *Biochem. Soc. Trans.*, **34**, 12–16.
12. Kato,S., Sekine,S., Oh,S.W., Kim,N.S., Umezawa,Y., Abe,N., Yokoyama-Kobayashi,M. and Aoki,T. (1994) Construction of a human full-length cDNA bank. *Gene*, **150**, 243–250.
13. Yoshihama,M., Uechi,T., Asakawa,S., Kawasaki,K., Kato,S., Higa,S., Maeda,N., Minoshima,S., Tanaka,T., Shimizu,N. *et al.* (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.*, **12**, 379–390.
14. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
15. Carninci,P., Westover,A., Nishiyama,Y., Ohsumi,T., Itoh,M., Nagaoka,S., Sasaki,N., Okazaki,Y., Muramatsu,M., Schneider,C. *et al.* (1997) High efficiency selection of full-length cDNA by improved biotinylated cap trapper. *DNA Res.*, **4**, 61–66.
16. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of transcriptional start sites: progress report 2004. *Nucleic Acids Res.*, **32**, D78–D81.
17. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
18. Yamashita,R., Suzuki,Y., Wakaguri,H., Tsuritani,K., Nakai,K. and Sugano,S. (2006) DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res.*, **34**, D86–D89.
19. Jordan,I.K., Marino-Ramirez,L. and Koonin,E.V. (2005) Evolutionary significance of gene expression divergence. *Gene*, **345**, 119–126.
20. Jordan,I.K., Marino-Ramirez,L., Wolf,Y.I. and Koonin,E.V. (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.*, **21**, 2058–2070.
21. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
22. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
23. Takeuchi,N. and Ueda,T. (2003) Down-regulation of the mitochondrial translation system during terminal differentiation of HL-60 cells by 12-O-tetradecanoyl-1-phorbol-13-acetate: comparison with the cytoplasmic translation system. *J. Biol. Chem.*, **278**, 45318–45324.
24. Ruan,H., Brown,C.Y. and Shama,S. (1997) *Analysis of Ribosome Loading onto mRNA Species: Implications for Translational Control*. Academic Press Inc., NY.
25. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
26. Arava,Y., Wang,Y., Storey,J.D., Liu,C.L., Brown,P.O. and Herschlag,D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc. Natl Acad. Sci. USA*, **100**, 3889–3894.
27. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al.* (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.*, **2**, 388–393.
28. Kraus,R.J., Murray,E.E., Wiley,S.R., Zink,N.M., Loritz,K., Gelembiuk,G..W. and Mertz,J.E. (1996) Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. *Nucleic Acids Res.*, **24**, 1531–1539.
29. Ledda,M., Di Croce,M., Bedini,B., Wannenes,F., Corvaro,M., Boyl,P.P., Caldarola,S., Loreni,F. and Amaldi,F. (2005) Effect of 3′UTR length on the translational regulation of 5′-terminal oligopyrimidine mRNAs. *Gene*, **344**, 213–220.
30. Levy,S., Avni,D., Hariharan,N., Perry,R.P. and Meyuhas,O. (1991) Oligopyrimidine tract at the 5′ end of mammalian ribosomal protein mRNAs is required for their translational control. *Proc. Natl Acad. Sci. USA*, **88**, 3319–3323.
31. Davuluri,R.V., Suzuki,Y., Sugano,S. and Zhang,M.Q. (2000) CART classification of human 5′ UTR sequences. *Genome Res.*, **10**, 1807–1816.
32. Katoh,Y. and Katoh,M. (2004) Identification and characterization of CDC50A, CDC50B and CDC50C genes in silico. *Oncol. Rep.*, **12**, 939–943.
33. Odom,D.T., Dowell,R.D., Jacobsen,E.S., Gordon,W., Danford,T.W., MacIsaac,K.D., Rolfe,P.A., Conboy,C.M., Gifford,D.K. and Fraenkel,E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, **39**, 730–732.
34. Geyer,P.K., Meyuhas,O., Perry,R.P. and Johnson,L.F. (1982) Regulation of ribosomal protein mRNA content and translation in growth-stimulated mouse fibroblasts. *Mol. Cell Biol.*, **2**, 685–693.
35. Caldarola,S., Amaldi,F., Proud,C.G. and Loreni,F. (2004) Translational regulation of terminal oligopyrimidine mRNAs induced by serum and amino acids involves distinct signaling events. *J. Biol. Chem.*, **279**, 13522–13531.