DNA Research **16**, 371–383, (2009) doi:10.1093/dnares/dsp022

Identification and Functional Analyses of 11 769 Full-length Human cDNAs Focused on Alternative Splicing

Ai Wakamatsu[1], Kouichi Kimura[2], Jun-ichi Yamamoto[3], Tetsuo Nishikawa[3], Nobuo Nomura[4], Sumio Sugano[5], and Takao Isogai[1,3,*]

Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan[1]; Central Research Laboratory, Hitachi, Ltd, Kokubunji, Tokyo 185-8601, Japan[2]; Reverse Proteomics Research Institute, 1-9-11 Kaji, Chiyoda-ku, Tokyo 101-0044, Japan[3]; National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan[4] and Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1 Shiroganedai, Minato-ku, Tokyo 108-8639, Japan[5]

(Received 25 August 2009; accepted 1 October 2009; published online 30 October 2009)

Abstract

We analyzed diversity of mRNA produced as a result of alternative splicing in order to evaluate gene function. First, we predicted the number of human genes transcribed into protein-coding mRNAs by using the sequence information of full-length cDNAs and 5′-ESTs and obtained 23 241 of such human genes. Next, using these genes, we analyzed the mRNA diversity and consequently sequenced and identified 11 769 human full-length cDNAs whose predicted open reading frames were different from other known full-length cDNAs. Especially, 30% of the cDNAs we identified contained variation in the transcription start site (TSS). Our analysis, which particularly focused on multiple variable first exons (FEVs) formed due to the alternative utilization of TSSs, led to the identification of 261 FEVs expressed in the tissue-specific manner. Quantification of the expression profiles of 13 genes by real-time PCR analysis further confirmed the tissue-specific expression of FEVs, e.g. OXR1 had specific TSS in brain and tumor tissues, and so on. Finally, based on the results of our mRNA diversity analysis, we have created the FLJ Human cDNA Database. From our result, it has been understood mechanisms that one gene produces suitable protein-coding transcripts responding to the situation and the environment.

Key words: full-length cDNA; alternative splicing; alternative transcription start site; mRNA diversity; tissue-specific expression

1. Introduction

One of the most interesting findings revealed by the Human Genome Project is that the human genome contains only 20 000–25 000 protein-coding genes.[1] This number is unexpectedly too small. To explain this unexpected result and to understand functions of genes, it is necessary to analyze mRNA diversity.

Biologically, multiple transcripts can be generated from a single gene by alternative splicing (AS). According to several reports on genome research, AS occurs in 30–60% of human genes.[2–5] It has been reported that AS of a single gene could produce transcripts coding for multiple proteins, each exhibiting different biochemical properties including binding, intracellular localization and regulation of enzymatic activities.[6] AS is also of interest to the pharmaceutical research because unwanted AS of genes could lead to various genetic diseases and cancers.[7] We have particularly focused on the analysis of AS patterns that are produce by utilizing alternative transcription start sites (TSSs). Indeed, multiple transcripts were produced from a gene by utilizing variable TSSs.[8,9]

Edited by Osamu Ohara

[*] Corresponding author. E-mail: tisogai@mol.f.u-tokyo.ac.jp

For example, the *Pcdh* gene, which contained variable TSSs, was shown to produce different transcripts;[10] similarly, UGTs (UDP-glucuronosyltransferases), which contained more than 10 TSSs.[11] From these findings, it is clear that to elucidate gene function, we have to further our knowledge on and understanding of all transcripts made from each gene, particularly those of the protein-coding transcripts. However, identification of all protein-coding transcripts have so far been difficult due to the fact that a large number of EST data accumulated in the databases are 3'-EST data, which were obtained by sequencing cDNAs from the polyA-end. Thus, even though sequences of a large number of mRNAs are already known, our understanding of these mRNAs remained incomplete because of the fragmentary nature and 3′-end bias of their sequences. Because of the lack of sequence information, it has been difficult to predict TSSs and to identify all the open reading frame (ORF) regions. Although the use of next generation sequencer helped in making advances in analyzing TSSs, it still remains extremely difficult to evaluate diversities of mRNAs transcribed by each gene because of their accumulation of short-length sequences (less than 50 bases) of cDNA clones.[12,13]

We sequenced ∼55 000 human full-length cDNAs, including 11 769 newly identified cDNAs described in this paper, and also obtained ∼1.45 million 5'-end-one-pass sequences (5'-EST).[14–17] We believe that these cDNA sequences are very useful in analyzing the diversity of protein-coding transcripts and would definitely contribute to our understanding of mRNA. First, our cDNA clones were isolated from full-length human cDNA libraries constructed by an optimized oligo-capping method, and therefore by utilizing their sequence information, we were able to identify the TSS with 90% or better accuracy.[14,18–20] Thus, we could easily and accurately identify TSSs of even low-expressing genes, for which up until now it required comparison of a large amount of data.[17] Second, our 5′-EST data contained, on the average, sequence information of ∼500 bases/cDNA clone, which covered two or more exons. Since the average length of the 5'-untranslated region is believed to be 125 bases,[21] it was possible to predict ORF regions using our 5'-EST data. Finally, the most important point is that all of our resources were obtained from the full-length cDNAs, including the TSS and the polyA site. Moreover, we could obtain various findings on protein expression from our full-length cDNAs.[16] These findings could not be obtained from sequences of short mRNA fragments. Since AS of genes could potentially create a large number of protein-coding transcripts, analyzing full-length cDNAs might be immensely valuable in understanding gene function.

Here, we report on our analysis of 11 769 full-length cDNAs, which were identified from our full-length cDNA libraries, and contained ORFs as a result of AS. We also present our analysis on the splice patterns and expression profiles of the identified cDNAs to explore the correlation between the mRNA diversity and gene function. Furthermore, we describe 261 full-length cDNAs with unique TSSs known as multiple variable first exon (FEV) and report on their expression profiles. Finally, we report establishing the FLJ Human cDNA Database based on the results of our analysis of the variable protein-coding transcripts generated from each gene by AS.

2. Materials and methods

2.1. Construction of full-length cDNA libraries

Most total RNAs isolated from various tissues and cells were purchased from Clontech and Ambion. Cells were cultured following established protocol, and cytoplasmic total RNAs were extracted from these cultured cells following a standard RNA purification method. The list of total RNAs used in this study was shown in Supplementary Table S1. We constructed cDNA libraries from total RNAs by an optimized oligo-capping method (detailed method for the optimized oligo-capping is provided in the Supplementary Method 1).[18,19] Briefly, total RNAs were treated with bacteria alkaline phosphatase (TaKaRa) and tobacco acid pyrophosphatase. After that, total RNAs were ligated to the oligo-RNA using the RNA ligase (TaKaRa). Oligo-capped polyA(+) RNAs were then isolated oligo-dT columns. The first-strand cDNAs were synthesized using the Superscript II reverse transcriptase (Invitrogen), the synthesized cDNAs were amplified using the Gene Amp XL PCR kit (ABI) and the amplified product was digested with the restriction enzyme SfiI. Fragments longer than 2 kb were selected and purified by agarose gel electrophoresis and cloned into the DraIII-digested pME18SFL3 vector following the standard methods. The 5′-end-one-pass sequences of cloned cDNAs were analyzed using the ABI 377 and 3700 sequencers (ABI). The 5'-end fullness rate of the constructed oligo-capped cDNA libraries was evaluated as described previously,[22,23] and the detailed method for determining the 5'-end fullness rate is provided in the Supplementary Method 2.

2.2. Genome mapping and clustering

The 5′- and 3′-ends of cDNA sequences and the full-length cDNA sequences (Supplementary Table S2) were mapped onto the human genome (UCSC hg 18 NCBI Build 36.1). Possible local alignments between the cDNAs and genome sequences were identified by using the NCBI Mega BLAST program (ftp://ftp.ncbi.nih.gov/blast/). For each cDNA, best mapping of the sequence was determined from these local alignments using a dynamic programming technique that optimized the identity, coverage and topology of exons. The joining portions of consecutive local alignments were refined so as to restore the consensus sequence in the canonical splice sites. On the basis of the mapping results clustering of cDNA sequences were performed as follows: two cDNA sequences were grouped into the same cluster if their mapped positions shared at least one base on the genome. In general, each cluster corresponded to a single gene locus.

2.3. *Identification of alternatively spliced variants of mRNAs*

On the basis of the results of genome mapping and clustering analysis, ESTs that had different regions compared with known full-length cDNAs by AS were selected by Intris, a viewer for cDNA-genome alignments used for analysis of splicing variants and expression profiles.[24] To exclude the cDNA fragments derived from the immature mRNA and genomic DNA, reliability of mRNA was evaluated by using not only the human EST data but also the data conserved from other animals (Phastcons; obtained from UCSC Genome Browser). We predicted the ORF regions from the 5’-end sequences of full-length cDNAs on selected ESTs by using ATGpr (http://flj.lifesciencedb.jp/top/).[25] Next, we excluded those ESTs from the selected analytical targets when the predicted ORF regions of the selected ESTs were the same as the ORF regions of known full-length cDNAs. In addition, even if the predicted ORF regions were different from the ORF regions of known full-length cDNAs, we excluded cDNA clones containing extremely short ORF regions (mostly 60 amino acids or less) compared with the other full-length cDNAs that mapped in the same locus of the human genome. The selected cDNAs were further sequenced by primer walking method using an ABI3700 sequencer (ABI) to obtain information on 500 additional bases, and the ORF regions were predicted again by using the ATGpr.[25] We also evaluated the predicted ORF regions by using TRis,[26] translated region inspector, and examined their novelty of amino acid sequences by using ALVISION,[27] aligns two cDNA sequences that are splicing variants allowing large gaps. When the reliability of the predicted ORF region was insufficient, we excluded it from our list of analytical targets. When the predicted ORF regions of the selected cDNAs were judged reliable and different from those of the known full-length cDNAs, we then sequenced the full-length cDNA clone all the way up to the stop codon. Consequently, we completely sequenced 11 769 of full-length FLJ cDNAs and analyzed their tissue-specific expression. A detailed method for the analysis of the tissue-specific expression of the cDNAs is provided in the Supplementary Method 3. We have also constructed the FLJ Human cDNA Database (http://flj.lifesciencedb.jp) that contained these sequence information. A detailed method for the analysis of AS by using the information available in the FLJ Human cDNA Database is provided in the Supplementary Method 4. Sequences of 11 769 of our full-length cDNAs were also deposited in the DDBJ/GenBank/EMBL databases (AK293122–AK304890).

2.4. *Functional analysis of full-length cDNAs* in silico

Sequences of cDNAs were analyzed for the signal sequences, trans-membrane domains and motifs in the encoded proteins by using Signal P ver. 3.0 (http://www.cbs.dtu.dk/services/SignalP/), SOSUI ver. 1.5 (Mitsui Knowledge Industry) and Pfam 19.0 (November 2005; http://pfam.sanger.ac.uk/), respectively. We obtained information on motifs showing E-values of e-30 or more from the Pfam analysis, and based on these results, we then categorized each cDNA and the corresponding gene according to its gene ontology (GO) (http://www.geneontology.org/) classification by using InterPro (http://www.ebi.ac.uk/interpro/).

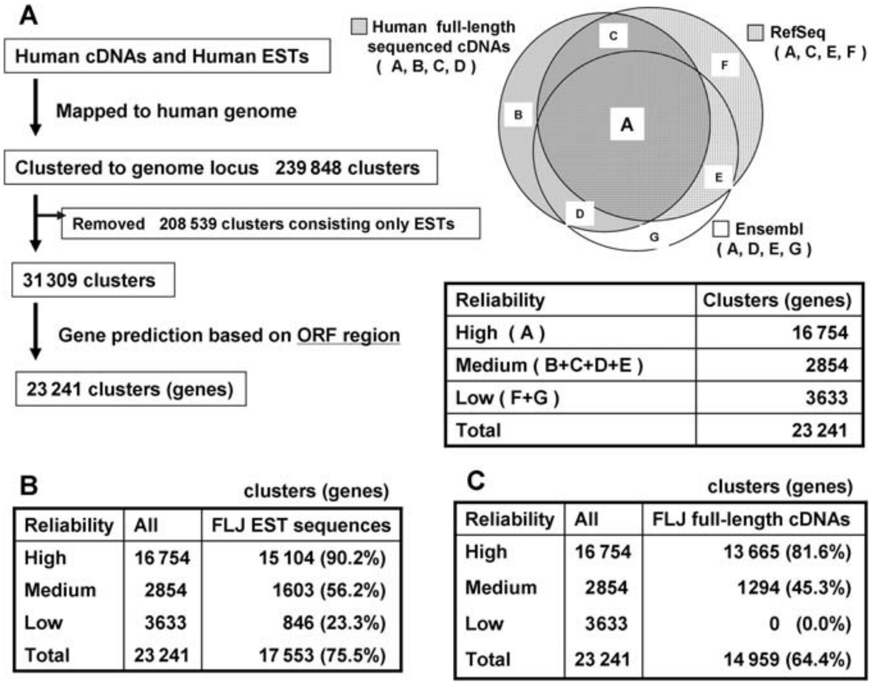
2.5. *Quantitative real-time PCR analysis*

Total RNAs derived from various tissues were purchased from Clontech, Ambion and STRATAGENE (listed in Supplementary Table S4). From 10 μg of each total RNA, first-strand cDNAs were synthesized using random primers and the Superscript III reverse transcriptase (Invitrogen) following the manufacturer’s instructions. Real-time PCR was performed using TaqMan Universal Master Mix (ABI) or SYBR Master Mix (ABI) on an ABI Fast7500 System (ABI) according to the manufacturer’s instructions. Approximately 300 ng of template cDNAs was used in each PCR reaction. Probes and primers were designed using the Primer Express3.0 (ABI) (refer to Supplementary Table S5 for the list of primers). The expression levels of genes were normalized with respect to that of the human GAPDH, and expression values of individual genes were calculated by comparing their Ct values to that of the control using the RQ software (ABI). The expression levels of genes were represented in $\log_{10}$ base. Samples were run in duplicates and the data shown are the average of two experiments.

3. Results and discussion

3.1. *Identification of human genes*

It is known that AS could produce mRNA diversity.[2–6] However, to analyze the mRNA diversity, it is necessary to identify human genes (i.e. the genome loci from where the protein-coding mRNAs are transcribed). We obtained 1.45 million human full-length cDNAs and sequenced their 5'-ends. We previously selected ~30 000 cDNAs from these full-length cDNAs based on the novelty analysis, and completely sequenced them.[14–16] Later, we also selected ~25 000 cDNAs based on the mRNA diversity and also sequenced them completely. In our quest to identify human genes, we used, for our analysis, the sequence information on these 55 000 full-length human cDNAs including 11 769 cDNAs reported in this paper (Supplementary Table S2). Furthermore, for the analysis, we not only used our own data but also data from 52 000 full-length human cDNA sequences available from the public databases, 30 000 human RefSeq (NCBI Reference Sequences; http://www.ncbi.nlm.nih.gov/RefSeq/) and 48 000 Ensembl, human gene transcripts (http://www.ensembl.org/index.html). In addition, we used EST sequences obtained by us and from other public databases (Supplementary Table S2). All the sequence data we collected were mapped onto the human genome and clustered. We then examined reliability of each full-length cDNAs by Intris[24] using sequences of all full-length cDNAs and ESTs mapped on the same locus of the genome, and based on this analysis, we selected only the reliable cDNAs for the gene identification analysis. We determined the genome locus of each one of the selected reliable cDNA and manually checked them one by one to identify the corresponding gene. As a result, we identified 23 241 human genes from this analysis (Fig. 1A). Each gene cluster was classified into three categories based on the reliability scores. The number of genes in the high reliability category (high category) were 16 754. Sequences of cDNAs belonging to the high-category group were found to be already analyzed because the genome locus was covered by sequence information available from the three types of databases, the human full-length cDNAs, RefSeq and Ensembl. It accounted for 72% of the total number of genes. The number of genes with intermediate reliability (medium category) was 2854. As for the medium-category group, the genome locus was covered by sequence information available from only the human full-length cDNAs or from two out of three of the above-mentioned databases. The number of genes with low reliability (low category) were 3633. As for the low-category group, the gene locus was covered by sequence information available only from the RefSeq or the Ensembl.


A
Human cDNAs and Human ESTs
Mapped to human genome
Clustered to genome locus 239 848 clusters
Removed 208 539 clusters consisting only ESTs
31 309 clusters
Gene prediction based on ORF region
23 241 clusters (genes)
Human full-length sequenced cDNAs (A, B, C, D)
RefSeq (A, C, E, F)
Ensembl (A, D, E, G)
A
B
C
D
E
F
G
B
C
clusters (genes)
clusters (genes)

Reliability	Clusters (genes)
High (A)	16 754
Medium (B+C+D+E)	2854
Low (F+G)	3633
Total	23 241

Reliability	All	FLJ EST sequences
High	16 754	15 104 (90.2%)
Medium	2854	1603 (56.2%)
Low	3633	846 (23.3%)
Total	23 241	17 553 (75.5%)

Reliability	All	FLJ full-length cDNAs
High	16 754	13 665 (81.6%)
Medium	2854	1294 (45.3%)
Low	3633	0 (0.0%)
Total	23 241	14 959 (64.4%)

Figure 1. Clustering of human cDNA sequences. (A) Estimation of the number of human genes from full-length cDNAs and ESTs. Outline of our gene prediction method from the human full-length cDNAs and ESTs mapped to human genome is schematically shown. For each one of the predicted genes, classification reliability was evaluated manually. (B) Cover rate of FLJ EST sequences and (C) cover rate of FLJ full-length sequenced cDNAs. Results of reliability analysis according to the category based on the cover rates of 1.45 million of ESTs (B) and 55 000 full-length cDNAs (C).

To further assess these reliabilities, we next calculated the cover rate of genes using our cDNAs. First, the cover rate was calculated using our 1.45 million FLJ ESTs, and we found a positive correlation between these reliabilities and the cover rate of FLJ ESTs (Fig. 1B). Next, we calculated the cover rate of genes using our 55 000 FLJ human full-length cDNA sequences. In this case, we also found a positive correlation between the reliability and the cover rate similar to that was observed for the ESTs (Fig. 1C). Thus, we were able to verify reliability irrespective of whether we used the sequences of our ESTs or full-length cDNAs in the analysis.

3.2. *Analysis of AS and functional classification of sequenced full-length cDNAs by GO*

We selected 25 000 full-length cDNAs from among the identified genes by focusing our attention on AS and subsequently sequenced them. In addition, from these cDNAs, we selected 11 769 of human full-length cDNAs in which the ORF regions were predicted to be different from the known full-length cDNAs, and then classified them by GO according to their predicted functions. First, ESTs exhibiting a different splicing pattern than the known full-length cDNAs were selected and were completely sequenced. From the sequence analysis, we were able to predict the ORF regions in only 30% of them (results not shown). Interestingly, a number of cDNA, for which we were unable to predict the ORF region, were thought to produced by AS. But, because our target was to be able to predict the function of the gene from the sequence of its transcript, it was necessary to select protein-coding transcripts efficiently. It is difficult to predict the ORF region correctly from the EST sequences lacking the TSS. However, our 5'-EST sequences not only contained the TSS but also contained sequence information on an average of 500 bases from the TSS. Therefore, we were able to correctly predict the ORF regions of our 5'-EST by using ATGpr.[25] As a result, the number of clones containing unpredictable ORF regions decreased to ~10%. Moreover, by using the tools such as TRins[26] for inspecting the translated region and ALVISION[27] for evaluating the novelty of amino acid sequences, we succeeded in identifying the ORF regions with high accuracy. Consequently, we obtained 11 769 of human full-length cDNAs in which the ORF regions were predicted to be different from the known full-length cDNAs (Supplementary Table S3). Ninety-six percent of these cDNAs-encoded proteins which differed in at least 10 amino acids from those encoded by their respective known full-length cDNAs, mainly because we selected them based on their altered ORF regions as a result of AS. These full-length cDNAs covered 7025 of 23 241 genes that we had originally identified.

Once it was established that human genes could produce multiple protein-coding transcripts, it was important to analyze their putative functions. The GO classification analysis was performed for all 11 769 our full-length cDNAs using Pfam, and their predicted functions, obtained from this analysis, are summarized in Table 1. The classification results revealed that a large number of our cDNA clones were listed under the GO molecular function categories 'nucleotide binding', 'nucleic acid binding', 'protein binding', 'hydrolase activity', 'transferase activity' and 'oxidoreductase activity'. Because 11 769 of our full-length cDNAs had ORF regions different from those of the known full-length cDNAs, we also analyzed their functions by predicting domains and motifs using Pfam, SOSUI and SignalP (Supplementary Table S3). Consequently, we discovered full-length cDNAs that encoded proteins with altered functional domains and signal sequences as a result of AS.

3.3. *Classification of splicing patterns of full-length cDNAs*

Up until now, majority of the ESTs entered in the public databases were 3'-EST. We succeeded in constructing full-length cDNA libraries efficiently by using the optimized oligo-capping method and obtained ~1.4 million 5'-ESTs of full-length cDNAs constructed by this method.[18,19] Our 5'-EST sequences were especially useful for the analysis of TSSs because 90% or more of our cDNAs contained the TSSs. We analyzed the splicing patterns of the 11 769 cDNAs by using the 5'-EST sequence data (Fig. 2). Results of this analysis revealed that 3403 cDNAs, which correspond to ~30% of all cDNAs, were transcribed using alternative TSSs (Type A), and thus, the predicted proteins contained new amino acid sequences at their N-terminal ends. In addition, 1962 cDNAs in Type A (designated as Type A1) contained FEV, due to transcripts originating from a TSS that was previously ignored because it was mapped in an intron region of the genome or transcripts originating from a TSS that was mapped upstream from the one that was analyzed before. Taken together, these results led to the discovery of new exons. We analyzed expression profiles of the genes containing multiple TSSs and discovered that the same gene could code for proteins with diverse function in different tissues by the proper utilization of alternative TSS. There were 8277 cDNAs (i.e. ~70% of all the full-length cDNAs) that were transcribed from the previously identified TSSs, but contained different ORF region because of AS; they were designated as Type

Table 1. Functional classification of the 11 769 full-length cDNAs based on the molecular function hierarchy of GO

Functional categorization (GO: molecular function)	Number of matched cDNAs
Binding	
Nucleotide binding	681
Nucleic acid binding	341
Protein binding	202
Ion binding	149
Lipid binding	28
Tetrapyrrole binding	27
Neurotransmitter binding	24
Carbohydrate binding	22
Other bindings	57
Catalytic activity	
Hydrolase activity	506
Transferase activity	479
Oxidoreductase activity	207
Ligase activity	85
Lyase activity	47
Helicase activity	38
Isomerase activity	26
Other catalytic activities	106
Enzyme regulator activity	
GTPase regulator activity	45
Enzyme inhibitor activity	44
Other enzyme regulator activities	21
Motor activity	
Microtubule motor activity	24
Other motor activities	20
Signal transducer activity	
Receptor activity	124
Receptor binding	25
Other signal transducer activities	40
Structural molecule activity	
Structural constituent of ribosome	25
Other structural molecule activities	56
Transcription regulator activity	
Transcription factor activity	138
Other transcription regulator activities	39
Translation regulator activity	
Translation factor activity, nucleic acid binding	25
Transporter activity	
Ion transporter activity	169
Carrier activity	90
Channel or pore class transporter activity	79
ATPase activity, coupled to movement of substances	39
Other transporter activities	131
Others	2
Molecular function unknown	45

If a protein was predicted to belong to two or more categories, all categories were included for counting.

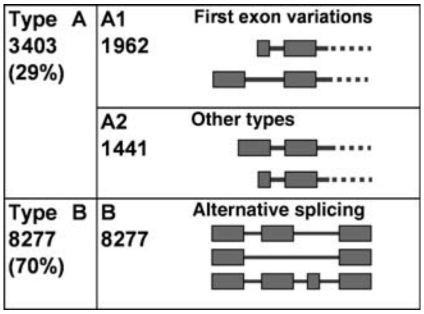


Figure 2. Classifications of the 11 769 full-length cDNAs based on splicing patterns. The 11 769 human full-length cDNAs were classified according to their TSS utilization. Type A: these cDNAs were derived from transcripts which were generated utilizing a TSS different than the previously analyzed TSS of the gene. Type A1: cDNAs contained a sequence variation known as FEV. Type A2: this class of cDNAs did not have the FEV feature. Type B: these cDNAs were derived from transcripts that were generated utilizing the same TSS as the previously analyzed TSS, but were found to be alternatively spliced. We could not classify 89 cDNAs because they coded for newly identified proteins.

B. Because we used our 5'-EST data for the selection, a lot of Type B cDNAs were predicted to contain N-terminal sequences different from those of the known cDNAs, except for a portion of cDNAs which were either selected by PCR or found during sequencing analysis. To assess whether AS or use of alternative TSS could alter the function of the predicted protein, we compared the GO functional categories of the Type A and Type B (Table 2). Our results showed that majority of the Type A belonged mainly to the GO molecular function categories of ‘neurotransmitter binding’, ‘enzyme activator activity’, ‘cyclase activity’, ‘ATPase activity, coupled to movement of substances’ and ‘GTPase regulator activity’. Thus, by using our 5'-EST data, a lot of valuable information were obtained regarding the diversity of TSS and amino acid sequences at the N-terminal ends of proteins. However, since only a portion of the full-length cDNAs was selected for this analysis, information on sequence diversity in regions beyond 500 bases from the TSSs were not obtained. We believe that there are additional alternately spliced transcripts which remained to be analyzed in the future studies.

Table 2. Functional classification of two types of splicing patterns of 11 769 full-length cDNAs based on GO category analysis

Functional categorization (GO: molecular function)	Number of matched cDNAs		
	Type A (%)	Type B (%)	Type A + B
Binding			
Lipid binding	4 (14.3)	24 (85.7)	28
Tetrapyrrole binding	5 (18.5)	22 (81.5)	27
Neurotransmitter binding	12 (50.0)*	12 (50.0)	24
Carbohydrate binding	4 (18.2)	18 (81.8)	22
Cofactor binding	3 (16.7)	15 (83.3)	18
Steroid binding	1 (10.0)	9 (90.0)	10
Catalytic activity			
Helicase activity	4 (10.5)	34 (89.5)	38
Small protein activating enzyme activity	2 (18.2)	9 (81.8)	11
Cyclase activity	6 (54.5)*	5 (45.5)	11
Enzyme regulator activity			
GTPase regulator activity	31 (68.9)*	14 (31.1)	45
Enzyme activator activity	6 (50.0)*	6 (50.0)	12
Structural molecule activity			
Structural constituent of ribosome	1 (4.0)	24 (96.0)	25
Transporter activity			
ATPase activity, coupled to movement of substances	23 (59.0)*	16 (41.0)	39
Electron transporter activity	2 (13.3)	13 (86.7)	15
Total	1344 (32.0)	2862 (68.0)	4206

The ratio of Type A and Type B is 3:7 as shown by total. Total is all the results of classification in the category of molecular function. If a protein was predicted to belong to two or more categories, all categories were included for counting.
*Functional categories biased to Type A.

3.4. *Analysis of genes showing tissue-specific expression*

We analyzed expression of genes producing multiple protein-coding transcripts by AS and found that many of these transcripts were expressed in specific tissues or cells, suggesting that the genes likely use this diversity according to the need and situation. We next analyzed expression profiles of 10 069 cDNAs, which corresponded to 5542 genes, out of 11 769 full-length cDNAs we identified in this study. As our cDNA libraries were constructed using RNAs derived from more than 100 different types of tissues and cells, we therefore used the 5′-EST data for analyzing gene expression. We next analyzed gene expression profiles of Type A1 cDNAs containing the FEV diversity and found that the FEVs of 261 cDNAs, which correspond to 155 genes, showed specific expression patterns that were different from those already obtained for the genes with alternative TSSs (Table 3). Thus, like the genes with alternative TSSs, the expression patterns of the genes with FEVs likely depended on the tissue and condition. Consequently, we found genes producing multiple protein-coding transcripts by AS.

3.5. *Analysis of expression patterns of tissue-specific expressed genes*

We quantified tissue-specific expressions of 13 out of 261 selected cDNAs by real-time PCR (Fig. 3). Results of our analysis especially suggested that there was a strong relationship between the tissue-specific expression and diversity of gene function or disease. We compared the expression profile of a specific gene by utilizing the TSS identified in this study with that of the same gene in which a previously identified TSS was utilized for expression. These results are summarized in Supplementary Table S6 and are discussed below in more detail.

First example, FGF13 is a gene that belongs to the FGF family and is believed to play roles in cell proliferation and differentiation, and also in neuronal differentiation.[28,29] FLJ57884 and FLJ57068 cDNAs exhibited different ORF regions as a result of FEV and were splicing variants of the known FGF13 cDNA. The TSSs we found in each one of them were located upstream from the TSS of FGF13. Whereas the known TSS of FGF13 was expressed highly in both fetal and adult brains, the TSSs of both FLJ57884 and FLJ57068 cDNAs were highly expressed only in the fetal brain. Moreover, the TSS of our FLJ57068 cDNA was also expressed highly in the kidney cancer (Fig. 3A). Second example, OXR1 is one of the oxidation stress receptivity genes localized in mitochondria.[30] The TSS of known OXR1 was expressed at equal levels in various tissues. But the TSS we identified in the FLJ56044 cDNA was located upstream from the known TSS of OXR1 and was highly expressed in brain, kidney cancer and lung cancer (Fig. 3B). Thus, these results suggested that these two genes were using different TSSs to regulate their expression levels in the brain. Moreover, our results also suggest that, for both genes, only one of the TSSs was preferentially recognized by the transcription machinery in the cancerous tissue.

Third example, C6orf142 (chromosome 6 ORF 142) is a gene of an unknown function. The known TSS of C6orf142 was highly expressed in the heart. However, the TSS we identified in the FLJ58494 cDNA, which was located downstream from the previously identified TSS of C6orf142, was highly expressed in both fetal and adult brains (Fig. 3C).

Table 3. Expressions of a selected list of 261 FEV-containing cDNAs (155 genes)

FLJ ID	Specific expression	Gene symbol	FLJ ID	Specific expression	Gene symbol	FLJ ID	Specific expression	Gene symbol	FLJ ID	Specific expression	Gene symbol
FLJ50079	Brain	NRK	FLJ52319	Trachea	GNE	FLJ55043	FB, NT	PDZRN3	FLJ57051	Brain	Pld5
FLJ50162	Brain	LARGE1	FLJ52354	Brain, NT	CHRNB1_pre	FLJ55050	Brain	EPS15	FLJ57068	FB	FGF13
FLJ50199	Brain	ARHGEF6	FLJ52356	Testis	ARMC4	FLJ55194	Brain	Unknown	FLJ57107	Brain, NT	CHRNB1_pre
FLJ50365	Trachea	CRISPLD1	FLJ52358	Testis	TP73	FLJ55226	FB	CHST10	FLJ57108	Brain	SNAP91
FLJ50390	Brain	GRIA1_pre	FLJ52367	Testis	IQGAP2	FLJ55256	Synovial	TFEC	FLJ57207	Im	Unknown
FLJ50398	Testis	IQGAP2	FLJ52368	Testis, Trachea	ARMC4	FLJ55265	Im	Unknown	FLJ57232	Testis	PRCP_pre
FLJ50459	Brain	ETV1	FLJ52384	Im	PTPN3	FLJ55281	Heart, Fetal heart	SLC5A1	FLJ57269	Brain	BTBD10
FLJ50460	Brain	DLG4	FLJ52407	Testis	CRB1_pre	FLJ55284	FB, NT	MAGI2	FLJ57290	Trachea	CRISPLD1
FLJ50484	Brain	SLC26A4	FLJ52427	Brain	AMPD3	FLJ55338	FB	CLASP1	FLJ57298	Brain	RAPGEF4
FLJ50494	Brain	ETV1	FLJ52435	Testis	MARCH7	FLJ55344	Brain	DYSF	FLJ57302	Brain	RAPGEF4
FLJ50523	Brain	PEX5L	FLJ52438	Brain	RIMS1	FLJ55381	FB	SLC44A5	FLJ57330	Brain	APBB1
FLJ50526	Brain	PEX5L	FLJ52453	Testis	AMPD3	FLJ55423	Placenta	NRK	FLJ57521	Tu	PPFIBP2
FLJ50533	Brain	SLC6A9	FLJ52496	Brain	TSPAN5	FLJ55434	Testis	POMGNT1	FLJ57884	FB	FGF13
FLJ50539	Brain, NT	DCAMKL1	FLJ52520	FB	EOMES	FLJ55460	Brain	SEMA5B_pre	FLJ57888	Brain	SGCB
FLJ50557	Brain	MAP7	FLJ52731	Brain	SPRED2	FLJ55461	NT	KLHL13	FLJ57953	Brain	STAU
FLJ50577	FB	DLG4	FLJ52750	Brain	ARHGEF7	FLJ55481	NT	RGMA_pre	FLJ58008	Brain	PPP2R2B
FLJ50619	NT	ELAVL4	FLJ52810	Testis	GABRB3_pre	FLJ55495	Testis	PCYT2	FLJ58099	Brain	CLTCL1
FLJ50623	Brain, NT	DCAMKL1	FLJ53109	Testis	PPP2R5E	FLJ55504	Testis	KLHL13	FLJ58366	Brain	RIMS1
FLJ50641	Brain	ETV1	FLJ53114	Testis	NCAM2_pre	FLJ55514	Brain, Tu	EGFR_pre	FLJ58368	Brain	RAPGEF4
FLJ50646	FB	DLG4	FLJ53167	NT	CUL4B	FLJ55516	Tu	LIMS1	FLJ58494	Brain	Unknown
FLJ50725	Testis	ATPAF1	FLJ53184	Brain	PPFIA2	FLJ55607	Brain, Trachea	HDAC9	FLJ58753	Brain	ARHGEF3
FLJ50745	Testis	CCNA1	FLJ53222	FB	MLLT3	FLJ55622	Testis	MMRN1_pre	FLJ58755	Brain	CHN2
FLJ50761	Brain	LRIG1_pre	FLJ53242	Testis	CLASP1	FLJ55627	Testis	MOV10L1	FLJ58966	Im	RAB37
FLJ50773	Brain	CALB1	FLJ53247	Testis	IDE	FLJ55628	Testis	LOXHD1	FLJ59303	Brain	DOCK4
FLJ50776	Brain	ARHGEF6	FLJ53252	Testis	CDH2_pre	FLJ55641	Brain, NT	JARID2	FLJ59333	Tu	RARG
FLJ50810	FB, NT	MAGI2	FLJ53320	Brain	DLGAP1	FLJ55662	Im	FGR	FLJ59338	Tu	RARG
FLJ50844	Brain	WARS2_pre	FLJ53324	Brain	TJP2	FLJ55664	Testis	NTRK3_pre	FLJ59345	Brain	PPFIA2
FLJ50917	Testis	PCCB_pre	FLJ53330	Brain, NT	EXOC4	FLJ55778	Brain	CLASP1	FLJ59425	Placenta	SH3KBP1
FLJ50956	Brain	RAPGEF4	FLJ53518	Testis	POMGNT1	FLJ55834	Brain, NT	FGF11	FLJ59496	Brain	CHN2
FLJ50959	Brain	RAPGEF4	FLJ53578	Brain	Rims1	FLJ55856	Testis	ARHGEF3	FLJ59502	Brain	PPFIA2
FLJ50961	Brain	TMEM16C	FLJ53606	NT	AKT1	FLJ55859	Testis	ST7L	FLJ59511	Brain	GRIA1_pre
FLJ50989	FB	EOMES	FLJ53680	Testis	KIF2C	FLJ55865	Im	SLC43A2	FLJ59545	Brain	EML2
FLJ51025	Kidney	NOX4	FLJ53829	Brain	APBB1	FLJ55903	FB	GPR161	FLJ59625	Brain	ARHGEF7
FLJ51027	Kidney	NOX4	FLJ53875	Brain	APBB1	FLJ55905	Im	FGD4	FLJ59641	Testis	PPFIA2

FLJ51073	FB	EOMES	FLJ53929	Im	PTPN4	FLJ55906	Testis	KIFC3	FLJ59648	Im	DYSF
FLJ51155	Testis	Unknown	FLJ53980	Brain	PPM1F	FLJ55918	Brain	EML2	FLJ59678	Brain	PEX5L
FLJ51157	Testis	HDAC4	FLJ53990	Brain	GABRB3_pre	FLJ55961	Brain	GRM4_pre	FLJ59684	Brain	PLEKHG5
FLJ51174	Im	HDAC4	FLJ53997	Brain	CTNNA2	FLJ55997	Brain	CPNE6	FLJ59710	Brain	MCF2
FLJ51177	Im	HDAC4	FLJ53999	Brain	GAB1	FLJ56033	Testis	Unknown	FLJ59717	FB	TBR1
FLJ51210	Brain	KIFC3	FLJ54008	Brain	TPCN1	FLJ56036	Tu	KIFC3	FLJ59769	Im	PLEKHG5
FLJ51383	Testis	PPP2R5A	FLJ54011	Brain	PPFIA2	FLJ56037	Testis, Prostate	CUL2	FLJ59799	Testis	CTNNA2
FLJ51528	Im	BTNL8_pre	FLJ54016	Testis	DIP13B	FLJ56038	Small intestine	Unknown	FLJ59802	Testis	ADCY5
FLJ51566	Brain	PDK1	FLJ54093	Brain	GPHN	FLJ56044	Brain	OXR1	FLJ59806	Im	HDAC4
FLJ51606	Trachea	HABP2_pre	FLJ54100	Brain	CHN2	FLJ56093	Brain	PTPRR_pre	FLJ60503	Brain	LARGE1
FLJ51663	Testis	CPS1_pre	FLJ54331	Brain, Osteoclast	Unknown	FLJ56095	Brain	KLHL13	FLJ60665	Tu	SLC44A5
FLJ51675	Brain	ETV1	FLJ54394	Testis	CRB1_pre	FLJ56110	FB	GOLSYN	FLJ60667	Tu	SLC44A5
FLJ51685	Testis	MCF2	FLJ54513	Testis	WDR59	FLJ56116	FB	APLP1	FLJ60693	FB	PHF21B
FLJ51695	Im	TP74	FLJ54541	FB	EXOC4	FLJ56136	NT	SLC2A14	FLJ60998	Testis	INPP4B
FLJ51706	Testis	RAPGEF4	FLJ54577	NT	HDAC9	FLJ56137	Im	Unknown	FLJ61124	Brain	RAB37
FLJ51734	Uterus	TMEM16C	FLJ54580	NT	HDAC9	FLJ56142	NT	AMOTL2	FLJ61133	FB	EXOC4
FLJ51737	Brain	ARHGEF6	FLJ54612	Brain	SH3KBP1	FLJ56148	Brain	PLEKHG5	FLJ61370	FB	SNCAIP
FLJ51769	Testis	IQGAP2	FLJ54642	Brain	APBB1	FLJ56167	Testis	KLHL12	FLJ61443	Testis	LARGE1
FLJ51805	Brain	RIMS2	FLJ54658	Brain	LSAMP_pre	FLJ56226	NT	SNCAIP	FLJ61560	Trachea	TJP2
FLJ51859	Brain	APBB1	FLJ54672	Brain	DOCK4	FLJ56370	Testis, Prostate	FKBP8	FLJ61674	Brain	PEX5L
FLJ51873	Brain, NT	AGPS_pre	FLJ54673	Brain	Unknown	FLJ56376	Brain	MTMR1	FLJ61679	Brain	APBB1
FLJ51910	FB	GTPBP3	FLJ54674	Brain	TPCN1	FLJ56411	Brain	GRIA2_pre	FLJ53199	Brain ↓	NEDD4L
FLJ51934	Im	AOAH_pre	FLJ54690	Brain	BACE1_pre	FLJ56420	Testis	DNPEP	FLJ59993	Brain ↓	RIMS1
FLJ51957	NT	ELAVL4	FLJ54693	Brain	BACE1_pre	FLJ56452	Brain	EML2	FLJ55591	Brain ↓	ARHGEF3
FLJ51977	Brain	Unknown	FLJ54702	Brain	DLGAP1	FLJ56634	Brain	GRM4_pre	FLJ56152	Brain ↓	ARHGEF7
FLJ52027	Testis	ATPAF1	FLJ54724	FB	DLG2	FLJ56895	Testis	EML2	FLJ58411	FB ↓	CACNB3
FLJ52034	Im	Unknown	FLJ54738	Brain	PDZRN3	FLJ56912	Uterus	FBLN2_pre	FLJ58949	FB ↓	CACNB3
FLJ52037	Im	GRAP2	FLJ54742	Testis	Slmap	FLJ56913	Placenta, Uterus	FBLN2	FLJ57810	Tu ↓	A2ML1
FLJ52039	Im	GRAP2	FLJ54746	NT	PDZRN3	FLJ56957	Brain	TMEM16C	FLJ53545	Tu ↓	RARG
FLJ52041	Im	Unknown	FLJ54751	NT	SUV420H1	FLJ56961	Brain	CLTCL1			
FLJ52042	Im	GRAP2	FLJ54906	Trachea	TMC5	FLJ56973	Brain	TMEM16C			
FLJ52288	Testis	ARMC4	FLJ54987	FB	PHF21B	FLJ56979	Brain	MYRIP			

We analyzed expression profiles of the first exons of ~1.5 million 5’-ESTs constructed by the oligo-capping method. From this analysis, we selected 261 full-length cDNAs based on the expression levels of their FEVs in specific tissues. Expression levels of cDNAs indicated without any label and with a ‘↓’ label were high and low, respectively, in the respective tissues.

*NT: NT2 cell induced by retinoic acid; FB, fetal brain; Im, immune tissues; Tu, tumor tissues; pre, precursor; unknown, function unknown.

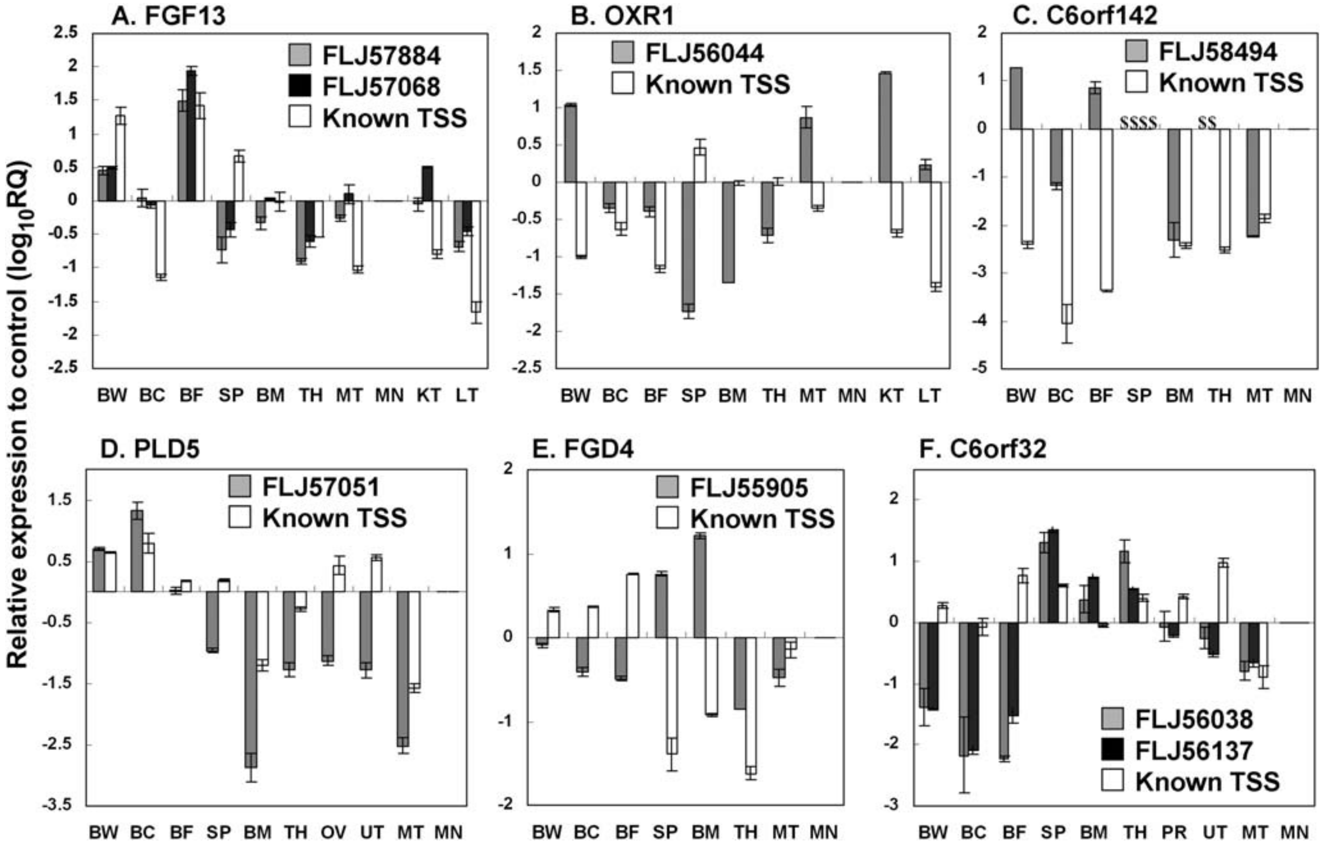


Figure 3. Quantitative evaluation of selected genes by real-time PCR. Expression levels of the first exon regions of the selected genes were analyzed by real-time PCR. The data were normalized with respect to that of the human GAPDH as described in the Materials and methods section. The expression levels of genes were represented in $\log_{10}$ base. Expression levels of cDNAs labeled '$$' represent the very low expression level or undetected. (A) FGF13, (B) OXR1, (C) C6orf142, (D) PLD5, (E) FGD4, (F) C6orf32. BW, brain, whole; BC, brain, cerebellum; BF, fetal brain; SP, spleen; BM, bone marrow; TH, thymus; OV, ovary; PR, prostate; UT, uterus; MT, mixture of tumor human tissues; MN, control, mixture of normal human tissues; KT, kidney tumor; LT, lung tumor.

Fourth example, PLD5 is one of the phospholipid-splitting enzymes presumably involved in the intracellular signaling.[31] Although the known TSS of PLD5 was expressed equally in various tissues, the TSS we identified in the FLJ57051 cDNA, which was located downstream of the previously identified TSS of PLD5, was highly expressed in the brain (Fig. 3D). Fifth example, SPRED2 is a Ras inhibitory factor belonging to the Sprouty/Spred family.[32] The TSS we identified in the FLJ52731 cDNA, which was located downstream from the known TSS of SPRED2, was expressed highly in the brain (Supplementary Table S6). Sixth example, SEMA5B is a nerve guidance factor which is involved in organogenesis, angiogenesis and oncogenesis.[33] The TSS we identified in the FLJ55460 cDNA, which was located downstream from the known TSS of SEMA5B, also was expressed highly in the brain (Supplementary Table S6). Seventh example, CACNB3 is a calcium channel beta-3 subunit, which is involved in modifying sympathetic nervous system, olfaction and control of blood pressure.[34] Although the known TSS of CACNB3 was expressed highly in both fetal and adult brains, the newly identified TSSs of FLJ58949 and FLJ58411 cDNAs, both of which were located downstream from the known TSS of CACNB3, were expressed at a low level in the brain (Supplementary Table S6). These cDNAs exhibited different ORF regions as a result of AS. Eighth example, BACE1 is a peptide hydrolase that cleaves the amyloid precursor protein and is one of the factors involved in Alzheimer's disease.[35] The known TSS of BACE1 was expressed equally in various tissues. However, the TSS we identified in the FLJ54690 cDNA, which was located downstream from the known TSS of BACE1, was expressed highly in the brain (Supplementary Table S6). Thus, these six genes regulated their expression levels in the brain using a specific TSS in each gene.

Ninth example, FGD4 is a gene that seemed to be involved in the regulation of the actin in the cytoskeleton and cell shape and also have various roles in proliferation, differentiation, transcriptional regulation and development.[36] The known TSS of FGD4 was highly expressed in the nervous system tissues such as brain, spinal cord and testis. However, the TSS we identified in the FLJ55905 cDNA, which was located downstream from the known TSS of FGD4, was highly expressed in the immune system tissues such as bone marrow and spleen (Fig. 3E). Tenth example, C6orf32 is a gene of unknown function whose expression level increased during the myoblast differentiation of the embryo.[37] FLJ56038 and FLJ56137 cDNAs exhibited different ORF regions as a result of FEV and were splicing variants of the known C6orf32 cDNA. The known TSS of C6orf32 was expressed at equal levels in various tissues. However, the TSSs we found in FLJ56038 and FLJ56137 cDNAs were located upstream of the known TSS of C6orf32, and both of these newly identified TSSs were highly expressed in the immune system tissues such as bone marrow, spleen and thymus (Fig. 3F). Eleventh example, PTPN4 is a gene belonging to the PTP (tyrosine escape phosphoric acid enzyme) family that works as a transmitter and controls various cellular processes like cell proliferation, differentiation, mitotic cycle and oncogenesis.[38] The known TSS of PTPN4 was highly expressed in the brain, but the TSS we identified in the FLJ53929 cDNA, which was located downstream from the known TSS of TPN4, was highly expressed in the immune system tissues such as bone marrow and spleen (Supplementary Table S6). Twelfth example, BTNL8 is one of the butyrophilin-like proteins and seemed to be involved in conferring immunity.[39] The known TSS of BTNL8 was found to be expressed at equal levels in various tissues. However, the TSS we identified in the FLJ51528 cDNA, which was located downstream from the known TSS of BTNL8, was highly expressed in the lung and thymus (Supplementary Table S6). Thus, it seems that these four genes regulated their expression levels in the immune system tissues by using specific TSSs.

Thirteenth example, AKT1 is a gene involved in apoptosis and neuronal differentiation and also may have a role in schizophrenia, especially in the neurotransmission system.[40] The TSS we identified in the FLJ53606 cDNA, which was located downstream from the known TSS of AKT, was highly expressed in the retinoic acid-induced NT2 cells (Supplementary Table S6). Thus, this gene uses a specific TSS during the neuronal differentiation.

Thus, among the newly identified genes we have analyzed in this study, the TSSs of a number of these genes revealed specific expression patterns. These results suggest that a single gene could use alternative TSS for tissue-specific transcription. We also found a close relationship between the predicted function of a gene and its tissue-specific expression. Thus, our results suggest a strong correlation between the mRNA diversity and function of a gene.

3.6. *Construction and use of the FLJ Human cDNA Database*

We constructed the FLJ Human cDNA Database ver. 3.0 (http://flj.lifesciencedb.jp) based on the results of our analysis of variable protein-coding transcripts produced from a gene by AS. A detailed description of our DB is available at the DB website. The DB graphically displays mapping of all the full-length cDNAs in the human genome and their ORF regions and thus provides a lot of useful information on the mRNA diversity. Moreover, the DB not only contain sequence information on full-length human cDNAs but also contain sequence information on a huge number of human ESTs generated using the oligo-capping method, allowing us to obtain useful information on ESTs mapped on the same genome locus. Because the average length of our EST sequences was ∼500 bases, the diversity of mRNAs produced as a result of AS could be efficiently analyzed by using this information. Because we were able to accurately identify TSSs using our 5′-EST data, we believe that they could be used to understand the relationship between the variable utilization of TSSs and biological functions of genes. Moreover, one could analyze the expression profiles of the transcriptional region of genes using the data from our high accuracy 5′-EST sequences, although in some cases the results might be different from those obtained using the 3′-EST data.

Despite these useful features, our database specializes on 5′-end sequences, and therefore these data are not suitable for predicting AS in the C-terminal end. Then, a lot of AS-related information still remain to be extracted from our 1.4 million cDNA resources as all of them were not sequenced to completion. Because our cDNA resources are mostly full-length cDNAs including the TSS and the polyA site, complete sequencing of these cDNA clones will add to our understanding of the mRNA diversity. In addition, every full-length sequenced FLJ cDNAs is available from the National Institute of Technology and Evaluation (http://www.nite.go.jp/). We will continue to add new information on our resources to our database, and these resources will be very useful in the analysis of gene functions.

Because our interest was on the mRNAs with ORF regions different from those of already known mRNAs, we stopped sequencing the cDNA once we found that the predicted ORF region of the transcript was not different from the known mRNA (for instance, where the alternative TSS only existed in the 5′-untranslated region). We, however, found that there is a tissue specificity in the expression patterns of these genes where the variation in TSS existed in the 5′-untranslated region (results not shown). Collectively, these results suggest that depending on the situation and environment, the transcription machinery utilizes alternative TSS to regulate the expression of a transcript, even when the translated protein is same. These results are also included in our DB. We also did not complete sequencing the clones for which we were unable to predict the ORF regions of their mRNAs. However, we have also included these clones in the DB with the belief that one could obtain some new and useful information by analyzing these clones.

We discovered a lot of genes had mRNA diversity due to, for example, FEVs. We also found a lot of tissue-specific splicing patterns. Especially, in the case of FEVs that we analyzed, genes used different regions of the genome loci as the first exon, which seemed to be dependent on the tissue and its condition. We also discovered genes, the TSSs of which were located further away on the same genome locus of the gene. In these cases, there exists a high possibility that their transcription is controlled by individual transcription factors. As the mechanisms for controlling the transcription are closely related to the function, by understanding these mechanisms one could be able to artificially control the expression of an appropriate transcript in the future.

In this study, we have identified multiple transcripts producing genes, and we believe that each one of these genes is transcribed into an appropriate transcript according to the need and circumstance. Now, it will be important to know whether there is any correlation between the expression of one of the transcripts produced by a gene and a disease. For example, in the case of transcripts containing FEVs, which we analyzed in detail, only the first exon regions were different from the other previously characterized transcripts. Since the first exon regions of these transcripts are unique, it is possible to distinguish them easily from the other transcripts. It may be possible to control the expression of a specific mRNA from a group of mRNAs transcribed from a gene by targeting the first exon. As we accumulate more information on mRNA diversity of genes using approaches similar to what we have described in this study, we might be able to identify candidate genes as novel targets for the development of drugs with lower side effects.

Acknowledgements: We thank the members of the NEDO functional analysis of protein and research application project for cDNA sequencing and clone stock, especially thanks to M. Yamazaki, K. Watanabe, A. Sugiyama, Y. Ono, T. Takayama (Japan Biological Informatics Consortium) and K. Fujita (National Institute of Technology and Evaluation, Japan). We also thank the members of the NEDO human full-length cDNA sequencing project for EST sequencing.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was partly supported by a grant from New Energy and Industrial Technology Developmental Organization (NEDO) project of the Ministry of Economy, Trade and Industry of Japan.

References

1. International Human Genome Sequencing Consortium 2004, Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931–45.
2. Lander, E.S., Linton, L.M., Birren, B., et al. 2001, Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921.
3. Lopez, A.J. 1998, Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation, *Annu. Rev. Genet.*, **32**, 279–305.
4. Black, D.L. 2000, Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology, *Cell*, **103**, 367–70.
5. Modrek, B. and Lee, C. 2002, A genomic view of alternative splicing, *Nat. Genet.*, **30**, 13–9.
6. Stamm, S. 2002, Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome, *Hum. Mol. Genet.*, **11**, 2409–16.
7. Bracco, L. and Kearsey, J. 2003, The relevance of alternative RNA splicing to pharmacogenomics, *Trends Biotechnol.*, **21**, 346–53.
8. Landry, J.R., Mager, D.L. and Wilhelm, B.T. 2003, Complex controls: the role of alternative promoters in mammalian genomes, *Trends Genet.*, **19**, 640–8.
9. Zhang, T., Haws, P. and Wu, Q. 2004, Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation, *Genome Res.*, **14**, 79–89.
10. Wu, Q. and Maniatis, T. 2000, Large exons encoding multiple ectodomains are a characteristic feature of protocadherin genes, *Proc. Natl Acad. Sci. USA*, **97**, 3124–9.
11. Strassburg, C.P., Oldhafer, K., Manns, M.P. and Tukey, R.H. 1997, Differential expression of the UGT1A locus in human liver, biliary, and gastric tissue: identification of UGT1A7 and UGT1A10 transcripts in extrahepatic tissue, *Mol. Pharmacol.*, **52**, 212–20.
12. Wang, E.T., Sandberg, R., Luo, S., et al. 2008, Alternative isoform regulation in human tissue transcriptomes, *Nature*, **456**, 470–6.
13. Licatalosi, D.D., Mele, A., Fak, J.J., et al. 2008, HITS-CLIP yields genome-wide insights into brain alternative RNA processing, *Nature*, **456**, 464–9.
14. 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.
15. Otsuki, T., Ota, T., Nishikawa, T., et al. 2005, Signal sequence and keyword trap in silico for selection of full-length human cDNAs encoding secretion or

membrane proteins from oligo-capped cDNA libraries, *DNA Res.*, **12**, 117–26.

16. Goshima, N., Kawamura, Y., Fukumoto, A., et al. 2008, Human protein factory for converting the transcriptome into an in vitro-expressed proteome, *Nat. Methods.*, **5**, 1011–7.
17. Kimura, K., Wakamatsu, A., Suzuki, Y., et al. 2006, Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes, *Genome Res.*, **16**, 55–65.
18. Maruyama, K. and Sugano, S. 1994, Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, **138**, 171–4.
19. 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.
20. Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. and Sugano, S. 1997, Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library, *Gene*, **200**, 149–56.
21. Suzuki, Y., Ishihara, D., Sasaki, M., et al. 2000, Statistical analysis of the 5' untranslated region of human mRNA using 'Oligo-Capped' cDNA libraries, *Genomics*, **64**, 286–97.
22. Nishikawa, T., Ota, T., Kawai, Y., et al. 2002, Database and analysis system for cDNA clones obtained from full-length enriched cDNA libraries, *In. Silico Biol.*, **2**, 5–18.
23. Nishikawa, T., Ota, T., Kawai, Y., et al. 2001, Comparison of sequences of cDNA clones obtained from oligo-capping cDNA libraries with those from unigene, *DNA Res.*, **8**, 255–62.
24. Kimura, K., Nishikawa, T., Nagai, K., Sugano, S. and Isogai, T. 2002, Intris: A viewer for cDNA-genome alignments enabling efficient detection of splicing variants and expression profiles, *Genome Inform.*, **13**, 548–50.
25. Salamov, A.A., Nishikawa, T. and Swindells, M.B. 1998, Assessing protein coding region integrity in cDNA sequencing projects, *Bioinformatics*, **14**, 384–90.
26. Kimura, K., Nishikawa, T., Nagai, K., Sugano, S., Nomura, N. and Isogai, T. 2003, The translated region inspector for cDNA sequences, *Genome Inform.*, **14**, 456–7.
27. Yamamoto, J., Hatano, N., Araki, H., et al. 2003, A cDNA evaluation system for highly efficient sequencing of splicing variant cDNAs, *Genome Inform.*, **14**, 430–1.
28. Facchiano, A., Russo, K., Facchiano, A.M., et al. 2003, Identification of a novel domain of fibroblast growth factor 2 controlling its angiogenic properties, *J. Biol. Chem.*, **278**, 8751–60.
29. Greene, J.M., Li, Y.L., Yourey, P.A., et al. 1998, Identification and characterization of a novel member of the fibroblast growth factor family, *Eur. J. Neurosci.*, **10**, 1911–25.
30. Durand, M., Kolpak, A., Farrell, T., et al. 2007, The OXR domain defines a conserved family of eukaryotic oxidation resistance proteins, *BMC Cell Biol.*, **8**, 13.
31. Foster, D.A. and Xu, L. 2003, Phospholipase D in cell proliferation and cancer, *Mol. Cancer. Res.*, **1**, 789–800.
32. Nonami, A., Kato, R., Taniguchi, K., et al. 2004, Spred-1 negatively regulates interleukin-3-mediated ERK/mitogen-activated protein (MAP) kinase activation in hematopoietic cells, *J. Biol. Chem.*, **279**, 52543–51.
33. Adams, R.H., Betz, H. and Puschel, A.W. 1996, A novel class of murine semaphorins with homology to thrombospondin is differentially expressed during early embryogenesis, *Mech. Dev.*, **57**, 33–45.
34. Yamada, Y., Masuda, K., Li, Q., et al. 1995, The structures of the human calcium channel alpha 1 subunit (CACNL1A2) and beta subunit (CACNLB3) genes, *Genomics*, **27**, 312–9.
35. De Pietri Tonelli, D., Mihailovich, M., Di Cesare, A., Codazzi, F., Grohovaz, F. and Zacchetti, D. 2004, Translational regulation of BACE-1 expression in neuronal and non-neuronal cells, *Nucleic Acids Res.*, **32**, 1808–17.
36. Chen, X.M., Splinter, P.L., Tietz, P.S., Huang, B.Q., Billadeau, D.D. and LaRusso, N.F. 2004, Phosphatidylinositol 3-kinase and frabin mediate Cryptosporidium parvum cellular invasion via activation of Cdc42, *J. Biol. Chem.*, **279**, 31671–8.
37. Yoon, S., Molloy, M.J., Wu, M.P., Cowan, D.B. and Gussoni, E. 2007, C6ORF32 is upregulated during muscle cell differentiation and induces the formation of cellular filopodia, *Dev. Biol.*, **301**, 70–81.
38. Young, J.A., Becker, A.M., Medeiros, J.J., et al. 2008, The protein tyrosine phosphatase PTPN4/PTP-MEG1, an enzyme capable of dephosphorylating the TCR ITAMs and regulating NF-kappaB, is dispensable for T cell development and/or T cell effector functions, *Mol. Immunol.*, **45**, 3756–66.
39. Rhodes, D.A., Stammers, M., Malcherek, G., Beck, S. and Trowsdale, J. 2001, The cluster of BTN genes in the extended major histocompatibility complex, *Genomics*, **71**, 351–62.
40. Brugge, J., Hung, M.C. and Mills, G.B. 2007, A new mutational AKTivation in the PI3K pathway, *Cancer. Cell*, **12**, 104–7.