

**Table 3.** Changes in the DNA methylation status between PAPs and between tissues, the numbers of TB-PAPs and non-TB-PAPs whose DNA methylation status was Changed (upper line) or Not-Changed (lower line) between tissues

Refseq# or Promoter ID	CpGisland	Biased expression	Testis	Brain	Spleen	Stomach	Liver
(A) The exceptional cases in which the DNA methylation patterns of a PAP perfectly matched with those of the accompanying PAPs in all five tissues							
NM_002065							
1	1	Testis	NL	NL	NL	NL	NL
2	0	Lung	NL	NL	NL	NL	NL
NM_002394							
1	1	Other	NL	NL	NL	NL	NL
2	0	Cancer	NL	NL	NL	NL	NL
NM_004046							
1	1	Testis	NL	NL	NL	NL	NL
2	1	Brain	NL	NL	NL	NL	NL
NM_005104							
2	1	Other	NL	NL	NL	NL	NL
3	0	Other	NL	NL	NL	NL	NL
NM_006759							
1	1	Other	NL	NL	NL	NL	NL
2	1	Other	NL	NL	NL	NL	NL
(B) The cases in which the DNA methylation patterns of a PAP matched with those of the accompanying PAPs in all five tissues, when one mismatch was allowed							
NM_005896							
1	1	Cancer	NL	NL	NL	NL	NL
2	1	Brain	NL	NL	NL	NL	NL
3	0	Other	NL	NL	NL	NL	CP
NM_018093							
1	0	Testis	NL	NL	NL	NL	CP
2	1	Non-specific	NL	NL	NL	NL	NL
			TB-PAPs			Non-TB-PAPs	
(C)							
Changed			77 (62%)			16 (53%)	
Not-Changed			47 (38%)			14 (47%)	
Total			124 (100%)			30 (100%)	

state was not seen in other tissues for testis-preferring PAPs or for non-TB-PAPs in the testis. These observations suggest that unique regulation may be exerted for TB-PAPs in the testis. We observed no such bias for the other PAPs of the genes that had testis-preferring PAPs with 'null' methylation in the testis (Table 4).

When the same analysis was performed for brain-preferring PAPs, we did not see any bias toward 'null'. However, interestingly, in brain, the DNA methylation patterns in accompanying PAPs of the same gene were skewed. As shown in Table 4, even when the brain-preferring PAPs were 'compositely' methylated, their counterpart PAPs rarely had 'null' status. Although further extensive data collection will be necessary, these observations indicate that there may be several distinct methylation mechanisms for tissue-biased gene expression depending on the tissue type. Nevertheless, it

seemed that actively used PAPs are generally less extensively methylated than their accompanying PAPs.

### 3.5. Relationship between DNA methylation status and mRNA expression level

In order to examine the relationship between the DNA methylation status and the level of mRNA expression, we measured the relative amount of mRNAs transcribed from each of the alternative promoters in the five tissues. For this purpose, we used real-time RT-PCR assays. Based on our full-length cDNA information, RT-PCR primers were designed at unique regions inside of the first exons of the transcript variants corresponding to each of the PAPs which showed different DNA methylation statuses between tissues. Ninety-three such primer pairs were considered, and in 49 cases, positive signals from either of the tissues were observed. (For an example, see Fig. 3A.) Fig. 3B shows the distribution

**Table 4.** Relation between the Tissue Preference and the DNA Methylation Status of the PAPs (A) Numbers of observed DNA methylation patterns in testis-preferring PAPs in testis (1st column) and in other tissues (2nd column), those in brain-preferring PAPs in brain (3rd column) and in other tissues (4th column) and those in other-tissue-preferring PAPs in testis (5th column) and in brain (6th column) (B) Number of observed DNA methylation patterns in the PAPs whose accompanying PAPs were testis-preferring and “null” methylated in testis (population marked with an asterisk in Table 4). (C) Number of observed DNA methylation patterns in the PAPs whose accompanying PAPs were brain-preferring and “compositely” methylated in brain (population marked with double asterisks in Table 4).

(A)	Testis-preferring PAPs in testis (population marked with an asterisk)	Testis-preferring PAPs in other tissues	Brain-preferring PAPs in brain (population marked with double asterisks)	Brain-preferring PAPs in other tissues	Other tissue-preferring PAPs in testis	Other tissue-preferring PAPs in brain
Complete	1 (5%)	10 (12%)	13 (31%)	38 (23%)	38 (31%)	26 (24%)
Null	12 (63%)*	43 (53%)	18 (43%)	79 (49%)	46 (37%)	43 (39%)
Composite	6 (32%)	26 (32%)	11 (26%)**	41 (25%)	34 (28%)	38 (35%)
Incomplete	0 (0%)	2 (3%)	0 (0%)	5 (3%)	5 (4%)	2 (2%)
Total	19 (100%)	81 (100%)	42 (100%)	163 (100%)	123 (100%)	109 (100%)

(B) Accompanying PAPs in testis when testis-preferring PAPs = Null

Complete	2 (13%)
Null	8 (50%)
Composite	2 (13%)
Incomplete	0 (0%)
n.d.	4 (25%)
Total	16 (100%)

(C) Accompanying PAPs in brain when brain-preferring PAPs = Composite

Complete	10 (36%)
Null	4 (14%)
Composite	9 (32%)
Incomplete	0 (0%)
n.d.	5 (18%)
Total	28 (100%)

of the relative expression level calculated separately for each of the DNA methylation patterns. Although the difference in the distributions of the relative expression levels between ‘null’ and ‘composite’ populations was not highly statistically significant as a whole, we still observed that ‘null’ or relatively light methylation patterns of PAPs were enriched ( $P < 0.02$ ; hypergeometric distributions calculated similarly as in Section 3.4.) in the tissues in which the mRNA expression was the strongest (relative expression level  $\geq 8$ ; also see Fig. 3A, Supplementary Table S2). We also observed that ‘complete’ DNA methylation was the only pattern in which the mRNA expression was the weakest (relative expression level  $\leq 1/8$ ). It is highly possible that the changes in the DNA methylation status are responsible for the modulations of the gene expression in these cases.

### 3.6. Relationship between DNA methylation status and the presence of CpG islands

We further attempted to clarify the relationship between the DNA methylation status and the presence of CpG islands. We compared the DNA methylation

status between the populations of CGI-containing PAPs and CGI-less PAPs and found a striking difference. In contrast to the so-far observed variable methylation patterns, we found that the CGI-containing PAPs were predominantly ‘null’ methylated throughout the five tissues ( $P < 1.0E-55$ ; evaluated by standard Fisher’s test; Table 5). Consequently, the DNA methylation status was unchanged throughout the tissues, remaining constantly ‘null’. These tendencies held irrespective of whether the PAPs were TB-PAPs or non-TB-PAPs ( $P < 30E-9$ ; evaluated by standard Fisher’s test; Table 5). In contrast, the DNA methylation in CGI-less PAPs was highly variable among tissues. When the CGI-containing PAPs were ‘null’ methylated, the pattern of the DNA methylation of the other PAPs of the same gene, which were usually CGI-less, was similar to the overall distribution patterns.

## 4. Discussion

In this study, we characterized the DNA methylation status at 181 PAPs in 61 genes which were identified

**A**

	Testis	Brain	Spleen	Stomach	Liver	(Ave. Ct)
<b>NM_005639</b>						
<b>Promoter ID 5</b>						
(Tissue-bias= Non-specific)						
<b>Ct value</b>	30	25	30	30	30	(29)
<b>Relative Expression Level</b>						
	1/2	16	1/2	1/2	1/2	
<b>DNA Methylation Pattern</b>						
	CL	CP	CL	CL	CL	

	Testis	Brain	Spleen	Stomach	Liver	(Ave. Ct)
<b>NM_004655</b>						
<b>Promoter ID 2</b>						
(Tissue-bias= Non-specific)						
<b>Ct value</b>	23	25	26	26	28	(26)
<b>Relative Expression Level</b>						
	8	2	1	1	1/4	
<b>DNA Methylation Pattern</b>						
	IC	CP	CP	IC	CL	

**B**

Relative Expression Level	≤ 1/8	1/8-1/4	1/4-1/2	1/2-2	2-4	4-8	≥ 8
<b>Complete</b>	2 (100%)	1 (17%)	10 (44%)	52 (34%)	6 (28%)	4 (45%)	1 (12%)
<b>Null</b>	0 (0%)	4 (66%)	7 (30%)	38 (25%)	5 (24%)	1 (11%)	4 (50%)
<b>Composite</b>	0 (0%)	1 (17%)	6 (26%)	52 (34%)	10 (48%)	3 (33%)	3 (38%)
<b>Incomplete</b>	0 (0%)	0 (0%)	0 (0%)	9 (6%)	0 (0%)	1 (11%)	0 (0%)
<b>Total</b>	2 (100%)	6 (100%)	23 (100%)	151 (100%)	21 (100%)	9 (100%)	8 (100%)

**Figure 3.** Methylation Pattern and mRNA Expression in the Five Tissues. (A) Examples of the semi-quantitative (real-time) RT-PCR results. The Ct values (constant of threshold; as of the default setting of ABI7900HT) and the relative expression levels are shown in the margin. The results of a similar analysis for the GAPDH mRNA are shown in Supplementary Figure S3, although we did not use GAPDH as a normalization standard. Ct: Constant of threshold as for default setting. Also note that relatively high expression levels are correlated with relatively light methylation patterns in both cases, even if their expression patterns are non-specific. (B) Composition of the DNA methylation patterns in the population of the PAPs with the indicated relative expression level. Based on the results of the semi-quantitative (real-time) RT-PCR assays, the relative expression levels were calculated as the deviation from the averaged expression level in the five tissues [also see (A) and Materials and Methods]. The relative expression levels observed for each of the PAPs in each of the tissues were correlated with their DNA-methylation status and summed for the five tissues.

by our previous full-length cDNA analyses. This is the first report describing the profiles of the DNA methylation of PAPs in a variety of tissues and may be the largest collection of data on the methylation of human promoters in general for normal tissues. In total, we analyzed the methylation status of 8612 CpG dinucleotides in 181 PAPs located in the proximal regions of transcriptional start sites (also see Supplementary Table S2). Although this number is small relative to the total number of CpG dinucleotides existing within the whole genome, this novel analysis of these CpG islands should be significant, because they may have direct biological consequences via modulating the transcription levels of genes.

Many of the previous studies using DNA-chip-based technologies were focused on comparing the DNA methylation status between cancerous cells and normal cells.<sup>21,22</sup> From the data we collected here, we found: (i) PAPs of the same gene tend to have different methylation pattern variation depending on the tissue; (ii) DNA methylation status differs among tissues for the majority of the individual PAPs; (iii) CGI-containing PAPs are an exception to observation, (ii) in that they tend to have a uniform 'null' pattern throughout the tissues examined.

The last result is somewhat of a confirmation of the findings of our previous study, although it was not previously tested on this scale. We made a comprehensive analysis of CpG islands on human chromosome 21 and observed that 84% of CpG islands located within promoter regions were 'null' methylated in leukocytes.<sup>19</sup> The general lack of methylation of promoter CpG islands, regardless of whether the corresponding genes are expressed or not, was also suggested by other previous studies, although the numbers of the assayed promoters and cell types were limited.<sup>23</sup> In this study, we assayed 181 PAPs in five normal tissues and observed that most of the CGI-containing PAPs were constantly 'null' methylated irrespective of the tissue (Table 5). Based on this observation, we concluded that methylation appears not to be the reason for the biased expression of most of the CGI-containing TB-PAPs.

In contrast to the case of the CGI-containing TB-PAPs, we found surprisingly wide variation in the methylation pattern among the tissues examined. We observed that the majority of TB-PAPs showed alteration of their DNA methylation status among tissues (62%; Table 5). If we counted only CGI-less TB-PAPs, the proportion went up to 80% (Table 5). We also found

**Table 5.** Changes in the DNA Methylation Status between CGI-containing and CGI-less Promoters (A) Numbers of observed patterns of DNA methylation status in CGI-containing PAPs and CGI-less PAPs are shown. The DNA methylation patterns observed in five tissues were summed. (B) Numbers of PAPs with “Changed” or “Not-Changed” DNA methylation status in five tissues are shown. (C) Numbers of DNA methylation patterns of CGI-less PAPs when their accompanying CGI-containing PAPs were “null” methylated.

	CGI-containing PAPs				CGI-less PAPs	
(A)						
Complete	5 (2%)				170 (34%)	
Null	202 (87%)				126 (25%)	
Composite	23 (10%)				189 (37%)	
Incomplete	2 (1%)				19 (4%)	
Total	232 (100%)				504 (100%)	
	CGI-containing PAPs		CGI-less PAPs		PAP Total	
	TB-PAP	Non-TB-PAP	TB-PAP	Non-TB-PAP	TB-PAP	Non-TB-PAP
(B)						
Changed	10 (25%)	1 (12%)	67 (80%)	15 (68%)	77 (62%)	16 (53%)
Not-Changed	30 (75%)	7 (88%)	17 (20%)	7 (32%)	47 (38%)	14 (47%)
Total	40 (100%)	8 (100%)	84 (100%)	22 (100%)	124 (100%)	30 (100%)
Accompanying CGI-less PAPs when CGI-containing PAPs = Null						
(C)						
Complete					72 (34%)	
Null					49 (23%)	
Composite					83 (40%)	
Incomplete					6 (3%)	
Total					210 (100%)	

The DNA methylation patterns observed in five tissues were summed.

that TB-PAPs seemed to have more variation of their methylation patterns than non-TB-PAPs (Table 3). We examined only five tissues in this study. We expect that the proportion of PAPs that show a variable methylation pattern among the tissues may increase if the number of tissues assayed is increased. Since we assayed 181 PAPs that were located on various chromosomes, these results indicate that various regions in the human genome are differentially methylated among various tissues. This suggests that the methylation of promoters is more dynamically regulated from one tissue to another than previously anticipated.

Among 48 genes for which we compared PAPs within the same gene, only about 10% had the same methylation pattern in all five tissues. All of these methylation patterns were ‘null’, as described in Table 3. Actually, CGI-containing PAPs accounted for these uniform ‘null’ patterns. Seven PAPs contained CpG-islands out of 10 PAPs of 5 genes that showed perfect matches of the methylation patterns (Table 3). We could not find any simple correlation of methylation pattern variation between PAPs within the same gene. The methylation of PAPs thus seems to be individually controlled even within the same gene.

We also searched for specific patterns of methylation according to the expression pattern or expression level of the mRNA transcribed from a PAP. We observed

that high levels expression correlated with the ‘null’ or relatively light methylation pattern of PAPs, although the statistical significance was not as clear as in the above cases, possibly due to the limited size of the current dataset. There seems to be a tendency for CGI-containing PAPs to be ‘null’ methylated and to show high expression levels. Our previous findings suggested that most genes have only one CGI-containing PAP among all their PAPs.<sup>1</sup> It is possible that this unique CGI-containing PAP is the major promoter of the gene, playing a key role among the group of alternative promoters.

Together, in this study, we showed that information about DNA methylation is as indispensable as the mere DNA sequences of the promoters for a thorough understanding of the regulation of the expression of the genes encoded by the human genome. Modifications of the histones, methylation of the DNA, and the chromatin structure are suggested to be related with each other.<sup>24</sup> Thus, information about DNA methylation may in general be useful as an indicator of the chromatin structure. Advances in genome and transcriptome research have opened a way for decoding the gene expression regulation of the genomic DNA sequences. Actually, in yeast, a recent report demonstrated that it was possible to predict the level of gene expression directly from the genomic sequence with some accuracy,

using clustered microarray data as an education dataset.<sup>25</sup> The sequence information of the DNA together with its higher-level chromatin structure will enable us to better understand how the code of the DNA directs the complex regulation of gene expression in human cells. Based on that knowledge, we should eventually be able to understand how the alternative utilization of the promoters plays roles in the diversification of gene functions.

**Acknowledgements:** We are grateful to K. Abe, K. Toya and K. Imamura for technical support. We are also thankful to E. Nakajima for critical reading of the manuscript. This work was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO) project of the Ministry of Economy, Trade and Industry (METI) of Japan, the Japan Key Technology Center project of METI of JAPAN and a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan.

**Supplementary Data:** Supplementary data are available online at <http://www.dnaresearch.oxfordjournals.org>

## References

- 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.
- Kim, T. H., Barrera, L. O., Zheng, M., et al. 2005, A high-resolution map of active promoters in the human genome, *Nature*, **436**, 876–880.
- Kapranov, P., Drenkow, J., Cheng, J., et al. 2005, Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays, *Genome Res.*, **15**, 987–997.
- Carninci, P., Sandelin, A., Lenhard, B., et al. 2006, Genome-wide analysis of mammalian promoter architecture and evolution, *Nat. Genet.*, **38**, 608–609.
- Landry, J. R., Mager, D. L., and Wilhelm, B. T. 2003, Complex controls: the role of alternative promoters in mammalian genomes, *Trends Genet.*, **19**, 640–648.
- Khaitovich, P., Hellmann, I., Enard, W., et al. 2005, Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees, *Science*, **309**, 1850–1854.
- King, M. C. and Wilson, A. C. 1975, Evolution at two levels in humans and chimpanzees, *Science*, **188**, 107–116.
- Suzuki, Y., Tsunoda, T., Sese, J., et al. 2001, Identification and characterization of the potential promoter regions of 1031 kinds of human genes, *Genome Res.*, **11**, 677–684.
- 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.
- 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.
- Gardiner-Garden, M. and Frommer, M. 1987, CpG islands in vertebrate genomes, *J. Mol. Biol.*, **196**, 261–282.
- Jones, P. A. and Baylin, S. B. 2002, The fundamental role of epigenetic events in cancer, *Nat. Rev. Genet.*, **3**, 415–428.
- Laird, P. W. 2005, Cancer epigenetics, *Hum. Mol. Genet.*, **14**, R65–R76.
- Clark, S. J. and Melki, J. 2002, DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene*, **21**, 5380–5387.
- Murrell, A., Rakyian, V. K., and Beck, S. 2005, From genome to epigenome, *Hum. Mol. Genet.*, **14**, R3–R10.
- Fazzari, M. J. and Greally, J. M. 2004, Epigenomics: beyond CpG islands, *Nat. Rev. Genet.*, **5**, 446–455.
- Clark, S. J., Harrison, J., Paul, C. L., and Frommer, M. 1994, High sensitivity mapping of methylated cytosines, *Nucleic Acids Res.*, **22**, 2990–2997.
- Frommer, M., McDonald, L. E., Millar, D. S., et al. 1992, A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands, *Proc. Natl Acad. Sci. USA*, **89**, 1827–1831.
- Yamada, Y., Watanabe, H., Miura, F., et al. 2004, A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q, *Genome Res.*, **14**, 247–266.
- Kuhlenbaumer, G., Hannibal, M. C., Nelis, E., et al. 2005, Mutations in SEPT9 cause hereditary neuralgic amyotrophy, *Nat. Genet.*, **37**, 1044–1046.
- Hatada, I., Fukasawa, M., Kimura, M., et al. 2006, Genome-wide profiling of promoter methylation in human, *Oncogene*, **25**, 3059–3064.
- Fukasawa, M., Kimura, M., Morita, S., et al. 2006, Microarray analysis of promoter methylation in lung cancers, *J. Hum. Genet.*, **51**, 368–374.
- Antequera, F., Boyes, J., and Bird, A. 1990, High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines, *Cell*, **62**, 503–514.
- Fuks, F. 2005, DNA methylation and histone modifications: teaming up to silence genes, *Curr. Opin. Genet. Dev.*, **15**, 490–495.
- Beer, M. A. and Tavazoie, S. 2004, Predicting gene expression from sequence, *Cell*, **117**, 185–198.

# Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56 419 completely sequenced and manually annotated full-length cDNAs

Jun-ichi Takeda<sup>1,2</sup>, Yutaka Suzuki<sup>3</sup>, Mitsuteru Nakao<sup>4,5</sup>, Roberto A. Barrero<sup>6</sup>, Kanako O. Koyanagi<sup>7</sup>, Lihua Jin<sup>6</sup>, Chie Motono<sup>4</sup>, Hiroko Hata<sup>3</sup>, Takao Isogai<sup>8,9</sup>, Keiichi Nagai<sup>9,10</sup>, Tetsuji Otsuki<sup>9</sup>, Vladimir Kuryshev<sup>11</sup>, Masafumi Shionyu<sup>12</sup>, Kei Yura<sup>13,14</sup>, Mitiko Go<sup>11,15</sup>, Jean Thierry-Mieg<sup>16,17</sup>, Danielle Thierry-Mieg<sup>16,17</sup>, Stefan Wiemann<sup>11</sup>, Nobuo Nomura<sup>2</sup>, Sumio Sugano<sup>3</sup>, Takashi Gojobori<sup>2,6</sup> and Tadashi Imanishi<sup>2,7,\*</sup>

<sup>1</sup>Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, <sup>2</sup>Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, <sup>3</sup>Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan, <sup>4</sup>Computational Biology Research Center, National Institute of Advanced Science and Technology, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, <sup>5</sup>Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan, <sup>6</sup>Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan, <sup>7</sup>Graduate School of Information Science and Technology, Hokkaido University, North 14, West 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan, <sup>8</sup>Reverse Proteomics Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan, <sup>9</sup>Helix Research Institute, Inc. 1532-3, Yana, Kisarazu, Chiba 292-0812, Japan, <sup>10</sup>Central Research Laboratory, Hitachi, Ltd, 1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan, <sup>11</sup>Division of Molecular Genome Analysis, German Cancer Research Center, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany, <sup>12</sup>Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, 1266 Tamura-cho, Nagahama, Shiga 526-0829, Japan, <sup>13</sup>Quantum Bioinformatics Team, Center for Computational Science and Engineering, Japan Atomic Energy Agency, 8-1 Umemidai, Kizu, Souraku, Kyoto 619-0215, Japan, <sup>14</sup>Core Research for Evolution Science and Technology, Japan Science and Technology Agency, Japan, <sup>15</sup>Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan, <sup>16</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA and <sup>17</sup>Centre National de la Recherche Scientifique, Laboratoire de Physique Mathematique, Montpellier, France

Received April 22, 2006; Revised and Accepted July 3, 2006

## ABSTRACT

We report the first genome-wide identification and characterization of alternative splicing in human gene transcripts based on analysis of the full-length cDNAs. Applying both manual and computational analyses for 56 419 completely sequenced and precisely annotated full-length cDNAs selected for the H-Invitational human transcriptome annotation meetings, we identified 6877 alternative splicing genes with 18 297 different alternative splicing

variants. A total of 37 670 exons were involved in these alternative splicing events. The encoded protein sequences were affected in 6005 of the 6877 genes. Notably, alternative splicing affected protein motifs in 3015 genes, subcellular localizations in 2982 genes and transmembrane domains in 1348 genes. We also identified interesting patterns of alternative splicing, in which two distinct genes seemed to be bridged, nested or having overlapping protein coding sequences (CDSs) of different reading

\*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: 81 3 3599 8801; Email: imanishi@jbirc.aist.go.jp

© 2006 The Author(s).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

**frames (multiple CDS). In these cases, completely unrelated proteins are encoded by a single locus. Genome-wide annotations of alternative splicing, relying on full-length cDNAs, should lay firm groundwork for exploring in detail the diversification of protein function, which is mediated by the fast expanding universe of alternative splicing variants.**

## INTRODUCTION

Alternative splicing is a phenomenon in which different combinations of exons are spliced to produce distinct transcripts (1). Especially in higher eukaryotes, alternative splicing is frequently used as versatile means of producing diverse transcripts from a single gene locus. The alterations of the exons may cause changes in the encoded amino acid sequences and, at least in some cases, produce functionally divergent proteins by modifying, for example, the binding site of a growth factor receptor or an activation site of transcription factor (2,3). The most striking example of this is the *Drosophila DSCAM* gene, which is an axon guidance receptor gene. This gene consists of 17 exons, with 12, 48, 44 and 2 mutually exclusive alternative splicing exons for exons 4, 6, 9 and 17, respectively. Thus, even if not all combinations of these exons are allowed, this gene can encode thousands of protein products, which should enable functional diversification of the protein that could in turn assure precise axonal trajectory (4,5). Since the initial draft sequence of the human genome revealed that there seem to be an unexpectedly small number of genes embedded in the human genome (6), it has been hypothesized that alternative splicing is one of the most significant processes giving rise to the functional complexity of the human genome and that it might be indispensable for generating highly complex organisms such as humans (7).

However, despite the growing interest on the impact of alternative splicing in various aspects of the biological processes, our understanding of alternative splicing is still very primitive and its mechanisms of control are mostly unknown (8). In order to advance our understanding of the biological significance of alternative splicing in humans, it is essential to identify and characterize the genes that are subject to alternative splicing and which splicing patterns are used in what context in a genome-wide manner.

For this reason, large-scale attempts to identify alternative splicing have been initiated by several groups [i.e. (9)], mainly using bioinformatics analysis of partially sequenced cDNAs (ESTs). The EST sequences are clustered and compared to evaluate if the differences in their sequences might have been resulted from alternative splicing. So far, millions of human ESTs have been analyzed, and the newly identified alternative splicing variants are presented in several databases, such as AceView (130 576 alternative splicing variants from 19 557 genes; <http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/>; Y. Kohara *et al.*, in preparation), ASAP [30 793 alternative splicing variants from 7991 genes; <http://www.bioinformatics.ucla.edu/ASAP/>; (10)] and ASD [73 340 alternative splicing variants from 16 236

genes; <http://www.ebi.ac.uk/asd/>; (11)]. However, these recent high-throughput approaches have limitations. The first is the skewed coverage of the ESTs over the mRNAs. Since the ESTs are generally scarce around the 5' ends of an mRNA, previous data could have been biased towards the 3' ends. Secondly, some combinations of the alternative splicing exons may not be allowed. In those cases, analysis of only partially sequenced cDNAs would have no chance to identify mutual dependence of the combinations in alternative splicing.

Full-length cDNAs provide an ideal solution to all of these limitations. Moreover, not only full-length cDNAs can be utilized to extract complete cDNA sequence information, but are also useful as physical reagents indispensable for experimental analysis to examine the functional consequences of the identified alternative splicing. In this study, we used 56 419 cDNA sequences of human genes selected for the H-Invitational human transcriptome annotation meetings (12). These cDNAs were enriched for full-length cDNAs using various methods (13–16). In addition, they were fully sequenced with a sequence reliability higher than 99% [Phred values greater than 30; (17)]. Also, potentially problematic sequences, such as vectors and poly(A) tails were precisely trimmed. Thus, this cDNA collection constitutes an outstanding resource for comprehensive studies of alternative splicing. Out of the 56 419 cDNAs, 55 036 were successfully mapped onto the human genomic sequence (UCSC hg16; <http://hgdownload.cse.ucsc.edu/downloads.html#human>) and clustered into 24 425 loci. Of these, 10 127 loci contained two or more cDNAs and both manual and computational inspection allowed us to identify 18 297 alternative splicing variants encoded in 6877 loci [Table 1; (12)]. General statistics and a part of the related information have been published as a part of H-Invitational paper (12). In this paper, we describe further detailed features of alternative splicing. Here, we report the large-scale genome-wide identification and analysis of human alternative splicing based on fully sequenced and precisely annotated full-length cDNAs.

## MATERIALS AND METHODS

### Dataset description

In the present study, the set of the 56 419 cDNAs, selected for the H-Invitational human transcriptome annotation meetings was used. Contributors and attributes of each of the cDNA data subsets are described in Supplementary Table 1, the reference (12) and further references therein.

### Computational procedures to identify and characterize the alternative splicing variants

We mapped the full-length cDNAs to the human genome (UCSC hg16; <http://hgdownload.cse.ucsc.edu/downloads.html#human>). Alignments were generated using EST2GENOME (<http://emboss.sourceforge.net/apps/est2genome.html>); alignments having at least 95% identity and 90% coverage were selected. The cDNAs mapped on the same genomic region (at least one base overlap) were clustered and regarded

**Table 1.** Statistics of the data processing and of the alternative splicing variants and exons identified

	#Locus	#cDNA	#Total exon	#Alternative exon	#Constitutive exon
<b>H-Invitational total</b>	25 585	56 419	389 895 <sup>a</sup>	44 727	345 168
<b>Successfully mapped</b>	24 425	55 036	389 895	44 727	345 168
<b>≥2 cDNAs per locus</b>	10 127	35 030	331 924	44 727	287 197
<b>Identified alternative splicing</b>	6877	18 297	176 505	37 670	138 835
<b>5' end alternative splicing</b>	4568	7494	18 297	7494	10 803
<b>Internal alternative splicing</b>	5565	11 156	139 911	25 236	114 675
<b>3' end alternative splicing</b>	2933	4940	18 297	4940	13 357
<b>5'-UTR alternative splicing</b>	3216	4750	18 262	6398	11 864
<b>CDS alternative splicing</b>	6005	13 409	148 242	28 728	119 514
<b>3'-UTR alternative splicing</b>	797	1034	5877	1401	4476

<sup>a</sup>Unmapped cDNAs' exons could not be counted.

as a putative 'locus'. For further details of the mapping and clustering, see Ref. (12).

Since this study focuses on complete transcript variants, we conducted a sequence inspection of the mapped cDNAs to identify and discard 5' end-truncated cDNAs from the dataset. To this end, we excluded cDNAs whose 5' ends had been located inside the second or later exons of any other cDNAs with compatible exon structure in the same locus. We accepted the cDNAs whose 5' ends were located inside of the first exons and considered as variations in the exact transcriptional start sites. We also assumed that those cDNAs whose 5' ends were located outside of the exonic regions of any other clones could not be truncated forms of any known types of transcripts, at least. This assumption is based on the fact that the combination of multiple errors, for example, truncation followed by erroneous oligo-capping that occurred on an immature form, would be required to erroneously generate such cDNAs [for further detailed discussion of this subject, see Ref. (18)]. Using similar concepts, we examined the completeness of the 3' end using the same procedure and similarly removed all possible 3' end-truncated cDNAs. The 3' ends located inside of the last exons were allowed and considered as alternative polyadenylation sites.

Computational identification of the alternative splicing variants was then performed using the resulting filtered set of full-length cDNAs as follows: (i) The genomic position of each exon-intron boundary was compared with those of the other transcripts belonging to the same locus. For the comparison, a 10 bp allowance was made; (ii) If a cDNA had a part of the exonic sequence in the first/last exon inside confirmed intronic regions of the other cDNAs, it was regarded as being a '5'/3' end' alternative splicing variants. (iii) If a cDNA had a part of an internal exonic sequence inside a confirmed intronic region of other cDNAs, it was recognized as being an 'internal' alternative splicing variants (for a schematic representation, see Figure 1).

In order to evaluate and characterize the outcomes of the identified alternative splicing events on the encoded protein sequences, we used the information of the ORFs (i.e. positions and reading frames) annotated during the H-Invitational meetings. Differences in the length of the ORFs were evaluated in a pair-wise manner between all alternative splicing variants within the locus and the average ORF length difference was calculated for each locus. Possible targets for nonsense-mediated decay were selected as variants in which

the stop codon mapped more than 50 bp upstream of the last exon junctions. For the detection of Alu-like elements, RepeatMasker was run with default settings and for the detection of exonic splice enhancers (ESEs) (19), the RESCUE-ESE program was run as described previously (20).

Based on the deduced amino acid sequences for alternative splicing variants, protein motifs and Gene Ontology (GO) terms were predicted using InterProScan (<http://www.ebi.ac.uk/interpro/>) with default parameters. GO terms were automatically added to each of the variants when a protein motif(s) was recognized and could be associated with functional annotation. Enrichments of the motifs and GO terms were statistically evaluated using a hypergeometric distribution by using the following equation:

$$\sum_{x=k}^m \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

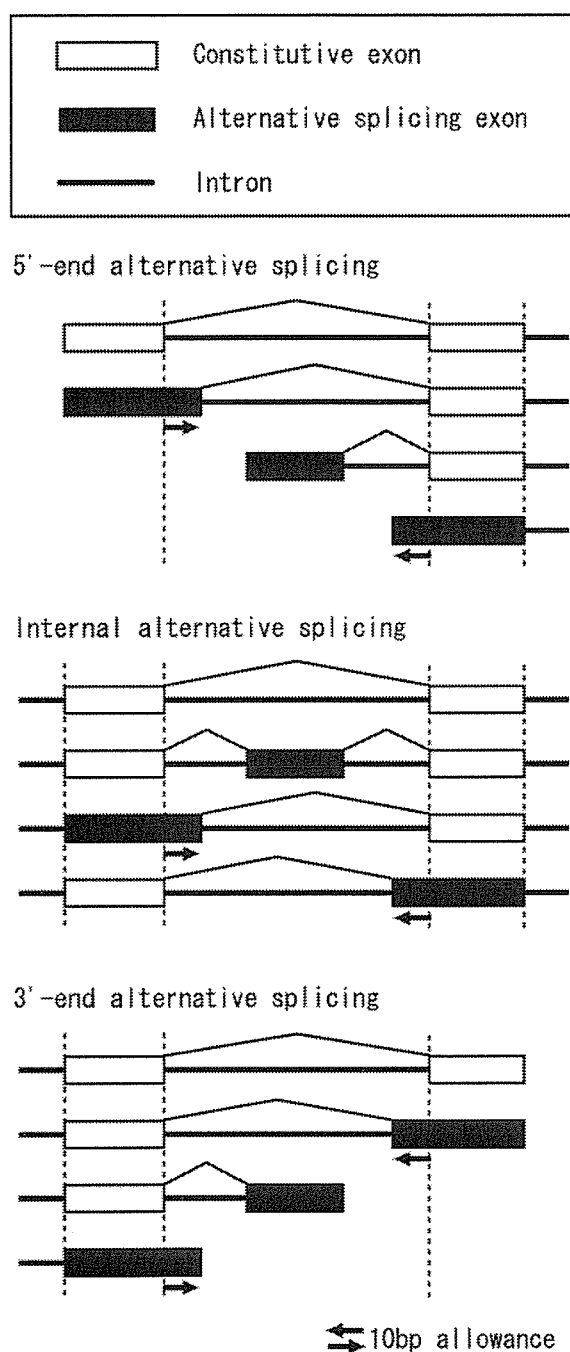
Here,  $N = 24\,425$  (number of loci containing successfully mapped cDNAs),  $n = 6877$  (number of loci containing alternative splicing variants),  $M = 12\,764$  (number of motif-containing loci),  $k = 5523$  (number of motif-containing loci with alternative splicing variants) in the case of Table 2. Similar calculations were done to evaluate the statistical enrichments for particular motifs and GO terms in Tables 3 and 4.

For the subcellular localization signal predictions, PSORT II (<http://psort.ims.u-tokyo.ac.jp/>) was run as indicated previously (21) and for the predictions of the transmembrane domains, TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>) was run with the cut-off value of 0.8. In cases where a protein motif(s) and/or a predicted localization signal(s), including transmembrane domains, were altered between two alternative splicing variants belonging to the same locus, the corresponding locus was defined as a 'Motif/GO/Subcellular localization/Transmembrane domain-changed' locus.

#### Manual procedures to inspect the identified alternative splicing variants

The results of the computational identification and annotation of the alternative splicing were visually inspected by the members of the alternative splicing annotation team of





**Figure 1.** Schematic representation of the identification of the alternative splicing. Essentially, the illustrated patterns of the exon pairs were searched for and selected as alternative splicing exons in both computational and manual annotations.

H-Invitational meetings by using the G-integra human genome browser [(12); Supplementary Figure 1] to verify the accuracy of the detected alternative splicing. Several items were manually evaluated, including correct discrimination of truncated cDNAs, proper identification of the alternative splicing types, and the corresponding functional annotations related to them. Further details of the annotation items and the records of the manual inspection are available at our web site (see below). The results of the manual and computational annotations were compared with each other, and in

**Table 2.** Relation between alternative splicing genes and motifs

	#Motif-related locus	#NOT motif-related locus	Total
Alternative splicing locus (3015; motif-changed)	5523 <sup>a</sup>	1354	6877
NOT alternative splicing locus	7241	10 307	17 548
<b>Total</b>	<b>12 764</b>	<b>11 661</b>	<b>24 425</b>

<sup>a</sup>*P*-value < 10<sup>-16</sup>.

cases where the results were consistent between the two approaches, the alternative splicing and the related annotations were defined as 'validated'. The results obtained for the annotations for each of the loci were made public and freely available at our web site ([http://jbirc.jbic.or.jp/h-inv2\\_as/](http://jbirc.jbic.or.jp/h-inv2_as/)).

### Identification of uncommon patterns of alternative splicing

Three 'uncommon' patterns of alternative variations were defined as follows: (i) 'bridged': a locus in which two alternative splicing variants are arrayed tandemly without sharing any exons and another transcript 'bridged' these two variants, sharing at least some of its exons with both of them; (ii) 'nested': a locus in which protein coding sequence (CDS) of one alternative splicing variant is not shared with another variant and (iii) 'multiple CDS': a locus in which different ORFs >200 amino acids in length are annotated independently for different alternative splicing variants having overlapping CDSs of different reading frames.

### RT-PCR of the 'bridged' transcript

RT-PCR was performed using the primers: primer A 5'-CG-TGAGCTCGCCCGCCAGAAG-3'; primer B 5'-TCCAACCTCCAGCTCCACATC-3'; primer C 5'-CGAGATGACGGGC-TTTCTGC-3'; primer D 5'-GGAATGCCATCGGTGCTGG-3'; primer E 5'-CCGACTATGCAGAGGAGAAG-3'; primer F 5'-GCGTTCGTGCTGCTCGAG-3'; primer (GAPDH fw) 5'-TCGGAGTCAACGGATTTGGT-3'; primer (GAPDH rv) 5'-TGACGGGTGCCATGGAATTTG-3', using ABI Prism 7900 Real Time PCR (ABI) with standard reaction conditions. The template RNAs (50 ng for each PCR) used were an RNA panel (BD Biosciences). For a negative control, 50 ng of human genomic DNA (Promega) was used as a template.

## RESULTS

### Identification of alternative splicing variants by manual and computational methods

A total of 56 419 human full-length cDNAs (Supplementary Table 1) were mapped onto the human genome and 55 036 cDNAs were unambiguously mapped. The mapped cDNAs were clustered into 24 425 loci, resulting in 2.3 cDNAs per locus on average. Single exon transcripts and the sole transcript in the locus were then removed. As a result, 10 127 loci contained at least two cDNAs (for further details on the mapping and clustering procedures, see Materials

**Table 3.** Most frequently observed motifs which were affected by alternative splicing variants

InterPro ID	Motifs in alternative splicing locus	Motifs in all locus	Ratio	Significance of enrichment ( <i>P</i> -value)	Definition
003598	417	495	0.84	$<10^{-16}$	Immunoglobulin C-2 type
000005	237	245	0.97	$<10^{-16}$	Helix–turn–helix, AraC type
000867	73	79	0.92	$<10^{-16}$	Insulin-like growth factor-binding protein (IGFBP)
000345	114	211	0.54	$10^{-15}$	Cytochrome <i>c</i> heme-binding site
003962	55	78	0.71	$10^{-15}$	Fibronectin, type III subdomain
002017	56	88	0.64	$10^{-12}$	Spectrin repeat
000379	62	103	0.6	$10^{-11}$	Esterase/lipase/thioesterase
002035	42	60	0.7	$10^{-11}$	von Willebrand factor, type A
000595	22	25	0.89	$10^{-10}$	Cyclic nucleotide-binding domain
003034	31	42	0.74	$10^{-9}$	DNA-binding SAP

**Table 4.** Most frequently observed GO terms which were affected by alternative splicing variants

GO ID	GOs in alternative splicing locus	GOs in all locus	Ratio	Significance of enrichment ( <i>P</i> -value)	GO term
0003676	451	1112	0.41	$<10^{-16}$	Nucleic acid binding
0003700	327	518	0.63	$<10^{-16}$	Transcription factor activity
0003677	276	603	0.46	$<10^{-16}$	DNA-binding
0004713	164	318	0.52	$<10^{-16}$	Protein tyrosine kinase activity
0005215	164	299	0.55	$<10^{-16}$	Transporter activity
0008270	148	276	0.54	$<10^{-16}$	Zinc ion binding
0005520	73	79	0.92	$<10^{-16}$	Insulin-like growth factor-binding
0005524	379	967	0.39	$10^{-14}$	ATP binding
0003824	190	429	0.44	$10^{-13}$	Catalytic activity
0016491	116	237	0.49	$10^{-11}$	Oxidoreductase activity

and Methods). These 10 127 loci, containing 35 030 cDNAs in total, were subjected to computational and manual inspection schemes to find if alternative splicing variants (defined as ‘complete forms of the transcripts’, with fully sequenced cDNAs) were included. We used this strategy since we considered that both the manual and computational methods should have advantages and disadvantages. Concerning the manual annotation, human errors are inevitable and for computational methods, detection of spurious alternative splicing due to various errors/ambiguities inherent to automated analyses is problematic. In order to maximize the accuracy of the analysis, the results of the manual and computational analyses should be compared with each other. When the results from computational analyses were ‘approved’ by manual inspections, the results were regarded as ‘validated’ (for the criteria of the annotation, see Figure 1).

Out of 35 030 cDNAs, the cDNAs annotated to be derived from ‘5′/3′ end-truncated’ or ‘immature’ transcripts either by manual or computational annotation were also excluded. In total, 5308 (15%), 706 (2%) and 787 (2%) cDNAs were defined as ‘5′ end’, ‘3′ end’ and ‘5′/3′-both side’ truncated, respectively (for further confirmation of the 5′ end completeness of each of the transcripts, see Supplementary Figure 2). A total of 1297 (4%) cDNAs were defined as ‘immature’. Also note that as for 3913 (11%) cDNA sequences, some of the annotators reported concerns that they might contain sequence problems due to cloning errors. The largest population of them was suspected spontaneous deletion of a part of cDNA insert in bacteria. The cDNAs supported by no ESTs were also identified. However, we did not remove these indecisive cDNAs, because it was not always straightforward to distinguish them from non-canonical rarely-occurring, but, biologically interesting alternative splicing

(also see below). We provide independent statistics using the dataset which did not include them in Supplementary Table 2. Also, caveats for possible errors were precisely annotated for each of them in our web site ([http://jbcirc.jbic.or.jp/h-inv2\\_as/](http://jbcirc.jbic.or.jp/h-inv2_as/)).

As a result, a representative ‘validated’ dataset of 6877 loci (68%), in which 18 297 ‘unique’ alternative splicing variants (2.7 variants per locus) were made of 37 670 alternative splicing exons (2.1 exons per variant; also see Supplementary Figure 3), was obtained and used for the subsequent analyses (Table 1). Also, with respect to each of the analyses described below, we chose the same strategy; to perform computational calculations first and then manually check the results. Therefore, each of the subsequent statistical analyses was based on the data which had passed both computational and manual inspections. A schematic representation of the computational calculations and the web interfaces used for the manual checks are shown in Supplementary Figure 1. The final results of each of the annotations were made public and freely available from our web site.

Interestingly, a surprisingly large population of the alternative splicing variants identified in the present study (as full-length forms) did not match in Ensembl (<http://www.ensembl.org>). If H-Invitational transcripts were defined as identical to Ensembl transcripts, when all of the exon–intron boundaries were corresponded to those of Ensembl transcripts with 10 bp allowance, we found 11 704 out of 18 297 (64%) of the H-Invitational transcripts were represented in Ensembl. When this was counted at the locus level, 6284 out of 6877 (91%) of the alternative splicing locus presented here contained novel alternative splicing relationship. This low level of overlap might be reflecting the fact that Ensembl is mainly based on the conservative analyses of the EST

sequences and do not put much stress on the full-length cDNAs.

### Patterns of the identified alternative splicing variants

Using the obtained dataset of the 18 297 alternative splicing variants from 6877 loci, we first examined the genome-wide features of alternative splicing in terms of their complete variants. Although similar analyses have been carried out by several groups (22–24), most of them were based either on partial EST sequences or analyses using smaller datasets or smaller numbers of cDNAs subjected to limited analyses (25).

The alternative splicing variants were first classified with regard to the splicing types, as often employed in previous studies (8). As shown in Figure 2, ‘cassette’ type alternative splicing variants were most frequently observed in our dataset, which is consistent with a previous result for the alternative splicing genes located on chromosome 22 (26). This type of alternative splicing variants may be preferred, because diversification of the transcripts can be achieved more flexibly than with the other categories of splice variations. ‘Retained intron’ type alternative splicing variants were observed in 1970 loci. However, it became a concern to us that cDNAs derived from unspliced, immature forms of transcripts might be classified as ‘retained intron’ types. To address this concern, we further checked how many of the alternative splicing variants of this category might be subject to nonsense-mediated decay [NMD; see Materials and Methods for the procedure to identify possible NMD-target variants (27)]. In 402 cases, one of the alternative splicing variants were annotated as a possible NMD-target, thus, might not have biological relevance. However, in the remaining cases, the alternative splicing variants with a ‘retained intron’ seemed to encode proteins with no explicit defects. So, it may indicate that most variants with ‘retained






Alternative splicing pattern	Definition	#Locus	#cDNA
	Cassette	3020	8166
	Internal acceptor	1758	4896
	Internal donor	1686	4537
	Mutually exclusive	210	636
	Retained intron	1970	2803

Figure 2. Patterns of the identified alternative splicing.

Table 5. Characteristics of the identified alternative splicing exons

	Exon-intron junction type		Containing Alu-like element	Containing ESE	Total
	Canonical	Non-canonical			
Alternative splicing exon	26 888 (96.6%)	954 (3.4%)	12%	8%	27 842
Constitutive exon	129 293 (99.2%)	1073 (0.8%)	2%	10%	130 366
Total	156 181 (98.7%)	2027 (1.3%)	4%	9%	158 208

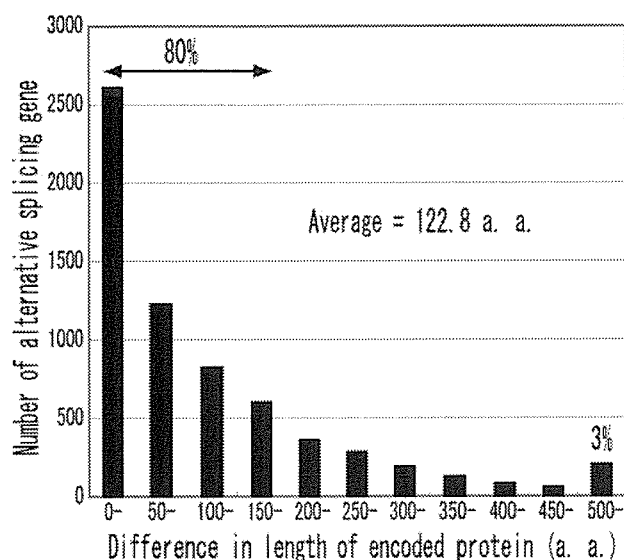
introns’ were beneficial for the diversification of human proteome.

We then examined the sequences of the alternative splicing exons and found that remnants of Alu-like elements were detected in a significant population (12%), which is considerably higher than the frequency in constitutive exons (2%; Table 5). This is consistent with previous results demonstrating that integration and subsequent alteration of the Alu-like elements play a significant role in the birth of alternative splicing exons (28). We also examined exon–intron junctions and detected canonical GT-AG (22) splice junctions in 156 181 (98.7%) out of 158 208 sites in total. Non-canonical AT-AC sites, the processing of which is using a splicing machinery with different and specific components (29), were also found in 89 cases. Another non-canonical splice site, GC-AG, were found in 57 cases. Taken together, we found that non-canonical junctions were enriched in alternative splicing exons (3.4%) compared to constitutive exons (0.8%). It is also intriguing those ESEs (20), which are suggested to play a role in efficient splicing, were also less frequent in alternative splicing exons. These findings suggest that the alternatively spliced junctions may be less deterministic than constitutive junctions, thus allowing versatile patterns of splicing events.

### Positions of the identified alternative splicing exons

In our dataset of the 6877 alternative splicing genes, alternative splicing events located at the 5′ end, internal and 3′ end exons were observed in 4568, 5565 and 2933 genes, and consisted of 7494, 11 156 and 4940 ‘5′ end’, ‘internal’ and ‘3′ end’ alternative splicing variants, respectively (Table 1). Considering that the total numbers of the exons examined were 18 297, 139 911 and 18 297, the average frequency of the alternative splicing exons were 0.41, 0.08 and 0.27 for each of the positions, respectively. It was intriguing that the ‘5′ end’ alternative splicing variants formed the most frequent category. We further examined these 5′ end alternative splicing exons and found that, among the 7494 ‘5′ end alternative splicing genes’, 3495 genes (47%) contained 5′ end alternative splicing variants, which were separated by more than 500 bp from each other. It is likely that these alternative splicing exons were produced as a consequence of the use of alternative promoters (30). Therefore, the biological significance of those 5′ end alternative splicing variants should be comprehensively analyzed together with the results of recent studies demonstrating that alternative use of promoters is prevalent in human genes (18,31,32). It is possible that the 5′ end alternative splicing exons were most abundant because of the requirements for the diversification of the transcriptional modulation and vice versa.

In order to characterize the biological consequences of the identified alternative splicing, we examined the relative



**Figure 3.** Distribution of the length difference between the alternative splicing variants. The percentages show the populations belonging to the corresponding groups.

position of the alternative splicing variants compared to their CDSs in 6555 genes in which two or more alternative splicing variants are annotated as protein coding. We found that alternative splicing variants were located in the 5'-untranslated regions (5'-UTRs), CDSs, and 3'-UTRs in 4750, 13 409 and 1034 alternative splicing variants in 3216, 6005 and 797 genes, respectively. In the majority of the alternative splicing genes, alternations in their CDSs were observed. Notice that 3' end alternative splicing was counter-selected in our analyses, by performing *in silico* filtration of NMD. Although in 80% of the alternative splicing variants, the alteration of the polypeptide length was less than 200 amino acid, 3% of the alternative splicing variants resulted in alteration of the polypeptide length of more than 500 amino acid (Figure 3). For this population, alternative splicing should have the most significant impact on the functions of the encoded proteins. It is even possible that these loci encode two functionally different proteins simultaneously, in which the fraction of the transcripts that is shared and their biological functional similarities may widely vary. Extreme examples of these cases will be discussed later.

#### Possible biological relevance of the identified alternative splicing variants to the diversification of protein functions

The influence of the identified alternative splicing variants on the encoded protein sequences and their possible biological functions was further evaluated from the following viewpoints. The complete records of each of the analyses are available for each of the entries from our web site ([http://jbirc.jbic.or.jp/h-inv2\\_as/](http://jbirc.jbic.or.jp/h-inv2_as/)).

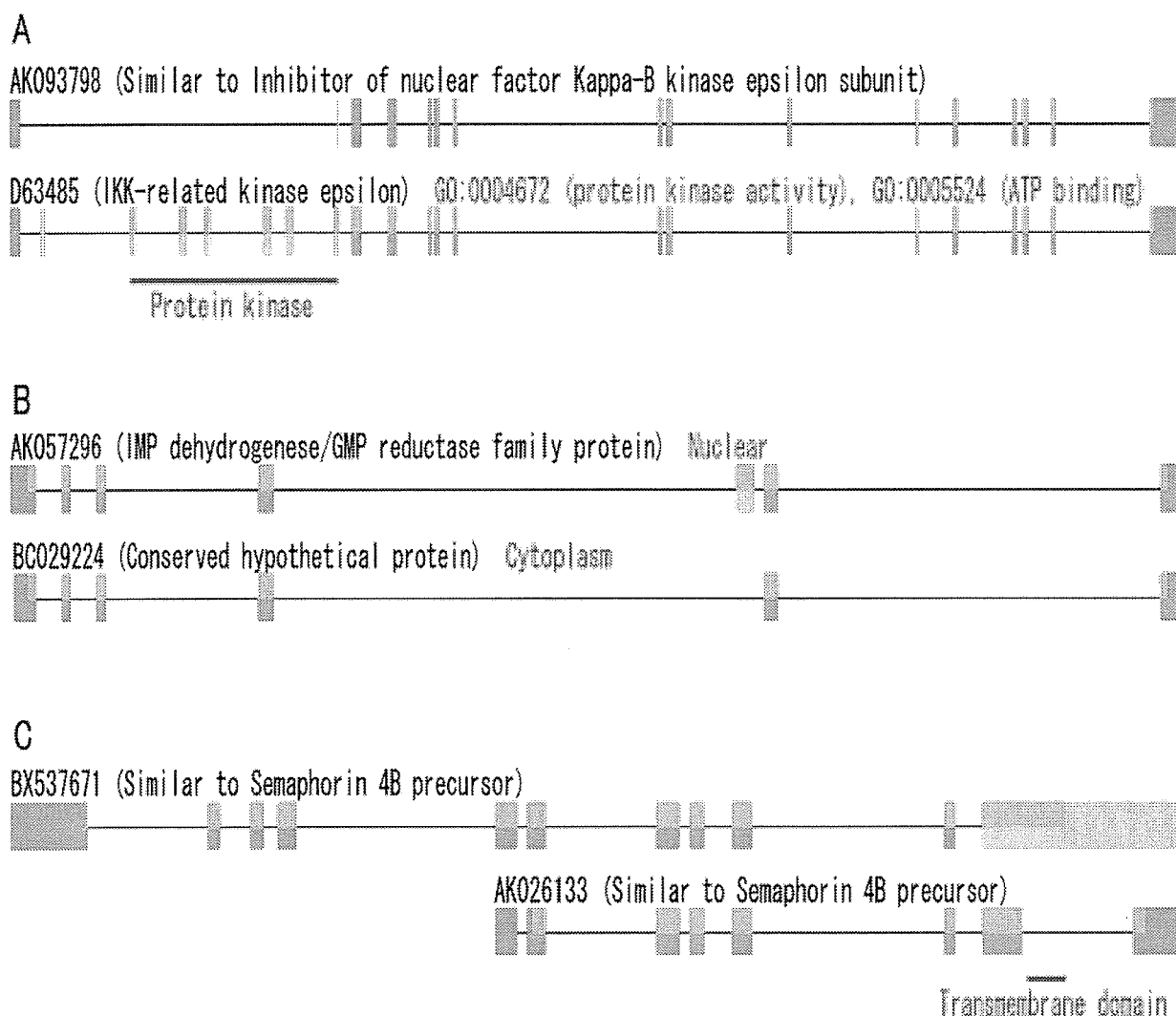
**Motif.** Protein motifs were frequently affected by the alternative splicing located within the CDSs. Out of 6005 alternative splicing variants residing in the CDSs, 3015 were located in protein motifs (Table 6). As shown in Figure 4, the alternative splicing variants influenced a wide variety of protein motifs. For example, we identified a novel alternative splicing

variant, AK093798 in the I $\kappa$ B kinase epsilon gene (IKK $\epsilon$ ; NM\_014002). This novel variant specifically lacks the kinase domain, while the rest of the CDS remained intact (Figure 4A). The IKK complex plays a pivotal role in immune and inflammatory responses by transmitting a variety of signals to a transcription factor, NF- $\kappa$ B (33). It has been reported that a kinase-deficient dominant-negative mutant of IKK $\epsilon$  blocks the induction of NF- $\kappa$ B invoked by T cell receptor but has no effect on its activation invoked by TNF $\alpha$  (34). The natural kinase-deficient variant of IKK $\epsilon$  represented by AK093798 could serve as a modulator between these two signaling pathways, and thus provide cells with the opportunity to regulate the relative amounts of the signals that the two pathways receive at a single point.

In total, among 6877 genes with alternative splicing, alterations of the motifs due to the alternative splicing were observed in 3015 genes (44%; Table 2). In 2508 genes (36%), alternative splicing did not change the annotated motifs. The remaining 1354 genes (20%) contained no annotated motif. The alternative splicing genes had a higher frequency of motifs overall (80%), as among 17 548 non-alternative splicing genes, only 7241 genes (41%) contained annotated motifs ( $P < 10^{-16}$ ; Table 2). We also found that alternative splicing exons were enriched in motifs. On average, one motif was contained per 1.6 alternative splicing exons, while the average frequency of motifs was only one per 3.0 constitutive exons. A similar tendency was also suggested from a recent EST-based study (35). The motifs should be actively associated with alternative splicing, and in many cases direct switching of the motifs is enabled by alternative splicing.

**Subcellular localization.** We also examined whether the predicted protein motifs determining the subcellular localization signals of the proteins, such as secretion signal peptides, mitochondria targeting signals and transmembrane domains, were affected by the alternative splicing (Table 6). For the subcellular localization signal predictions, PSORT II was used. The subcellular localization signal was predicted for each of the alternative splicing variants, and in cases where the alternative splicing variants were predicted to localize in different subcellular compartments, the alternative splicing were categorized as 'subcellular localization-changed' (shown in Figure 4B). Similarly, 'transmembrane domain-changed' alternative splicing were identified using TMHMM. In total, 2982 subcellular localizations and 1348 transmembrane domains were affected (Table 6). Figure 4C shows a case in which the transmembrane domain was altered. The most frequently observed switching of the predicted subcellular localization signal was between 'nuclear' and 'cytoplasm' (2455 cases). Switching between 'secretory' and 'plasma membrane' was detected in 1145 cases. These findings suggested that the proteins produced from the same loci as a result of the alternative splicing are actively utilized in a multi-faceted manner at different locations or compartments in the cells. A similar tendency of the use of several localization signals in the protein isoforms was also indicated by a recent bioinformatics study (36).

**GO.** In a large population of the genes (1779 genes; 27%; Table 6), the GO terms attached to the transcripts were



**Figure 4.** Examples of the alternative splicing variants detected as 'motif-changed' (A), 'subcellular localization-changed' (B) and 'transmembrane domain-changed' (C). Exons and introns are represented by green boxes and lines, respectively. The violet boxes are protein coding regions and yellow boxes are alternative splicing exons. The positions of the detected motifs and transmembrane domains are shown beneath the transcripts. In the uppermost panel, GO terms attached to the transcript indicated by the lower line are shown.

altered between the alternative splicing variants even within the same locus because the protein motifs and subcellular localizations were affected by the alternative splicing variants, as mentioned above. For example, in the case shown in Figure 4A, the GO term, 'kinase activity' (and related terms), was assigned to the kinase domain-containing variants, but not to the kinase negative-variants. Some of the motifs and GO terms were significantly enriched in alternative splicing genes (Table 4). It is especially noteworthy that motifs and GO terms associated with signal transduction and transcriptional regulation were most frequently influenced by alternative splicing (Tables 3 and 4). Fine adjustments of the resultant protein functions might be one of the most important functions of alternative splicing in higher organisms, such as humans.

#### Uncommon patterns of alternative splicing

During manual annotation, we noticed that a number of transcripts undergo novel patterns of alternative

splicing. Definitions of them are described in the legend for Figure 5.

*Bridged.* As shown in the upper panel of Figure 5A, a cDNA, AK000438, was identified as an alternative splicing variant of AK000479 and AF161485. However, it was more likely that AK000438 was overlaid on two adjacent loci rather than 'a variant of either of the loci', representing a 'bridging' transcript from two genes, the SERF2 gene (NM\_005770) and the HYPK gene (NM\_016400). It is noteworthy that both genes are associated with neuromuscular diseases. These genes were originally identified respectively as a candidate modifying gene for spinal muscular atrophy (37) and a *Huntingtin* interacting protein (38). We excluded the possibility that AK000438 was derived from artifacts, such as chimeric transcripts produced during the cDNA cloning process because of the facts that: (i) genes immediately adjacent to each other are bridged; (ii) the transcript was connected exactly at exon-intron junctions satisfying the GT-AG rule; (iii) there are supporting dbEST sequences which correspond to the

**Table 6.** Numbers of the genes in which alternative splicing variants should influence the possible protein functions

	#Locus	#cDNA
<b>Alternative splicing affecting functional annotation (total)</b>	<b>4481</b>	<b>12 542</b>
Motif-changed	3015	8727
Subcellular localization-changed	2982	8624
GO-changed	1779	5179
Transmembrane domain-changed	1348	3933
<b>Uncommon alternative splicing pattern (total)</b>	<b>316</b>	<b>1033</b>
Bridged	129	604
Nested	172	390
Multiple CDS	27	56

'bridging' transcript and (iv) RT-PCR using primers sets at the SERF2 and HYPK transcripts produced direct evidence that the 'bridging' transcript exists *in vivo* (Figure 5A lower panel). A previous study demonstrated that this kind of 'bridging' transcript is produced between the GALT gene and the IL-11Ra gene, and results in an mRNA encoding a protein retaining the functions of both proteins (39). This kind of 'bridging' transcript has been reported to be especially frequent in *Caenorhabditis elegans*, (179 out of 3829 alternative splicing genes) as are complex genes generating two completely distinct proteins in addition to a fusion of both. Recent reports suggested that such transcripts seem abundant in humans as well [(40); also see <http://www.ncbi.nlm.nih.gov/IEB/Research/Assembly/index.html?human>]. Although, further experimental characterization is required to clarify their biological consequences, it is possible that such bridged transcripts play some roles by combining the respective gene functions in general.

*Nested.* Figure 5B shows another uncommon type of alternative splicing ('nested'). In this case, all exons of AB508739 except for a shared 3' end exon, are embedded in the last intron of AK130874 (Figure 5B, upper panel). Neither a mapping error of the cDNAs nor an assembly error of the genome sequence is likely to account for this observation, since the sequence identity/coverage between the genome and cDNA sequences are almost 100% for each of the exons and genomic sequences are 'finished' in this region. In this case, these transcripts may share some modulatory elements, such as binding sites of regulatory proteins or non-coding RNAs, but their biological functions should be completely different. A similar specific case of alternative splicing was also described with regard to the 5' end exons, on which promoters seemed to be shared. (Figure 5B, lower panel). In a very recent paper, two otherwise independent genes were shown to be co-expressed from one promoter, giving yield to tissue-specific expression of an alternative form of an otherwise B-cell specific gene (41).

*Multiple CDS.* Alternative splicing even seemed to allow a particular genomic sequence to encode two distinct amino acids in different reading frames. The last exon of AK097244 overlapped with the second exon of AK000272 (Figure 5C, upper panel). Both of the exons were coding exons, however, their reading frames were different. Similarly,

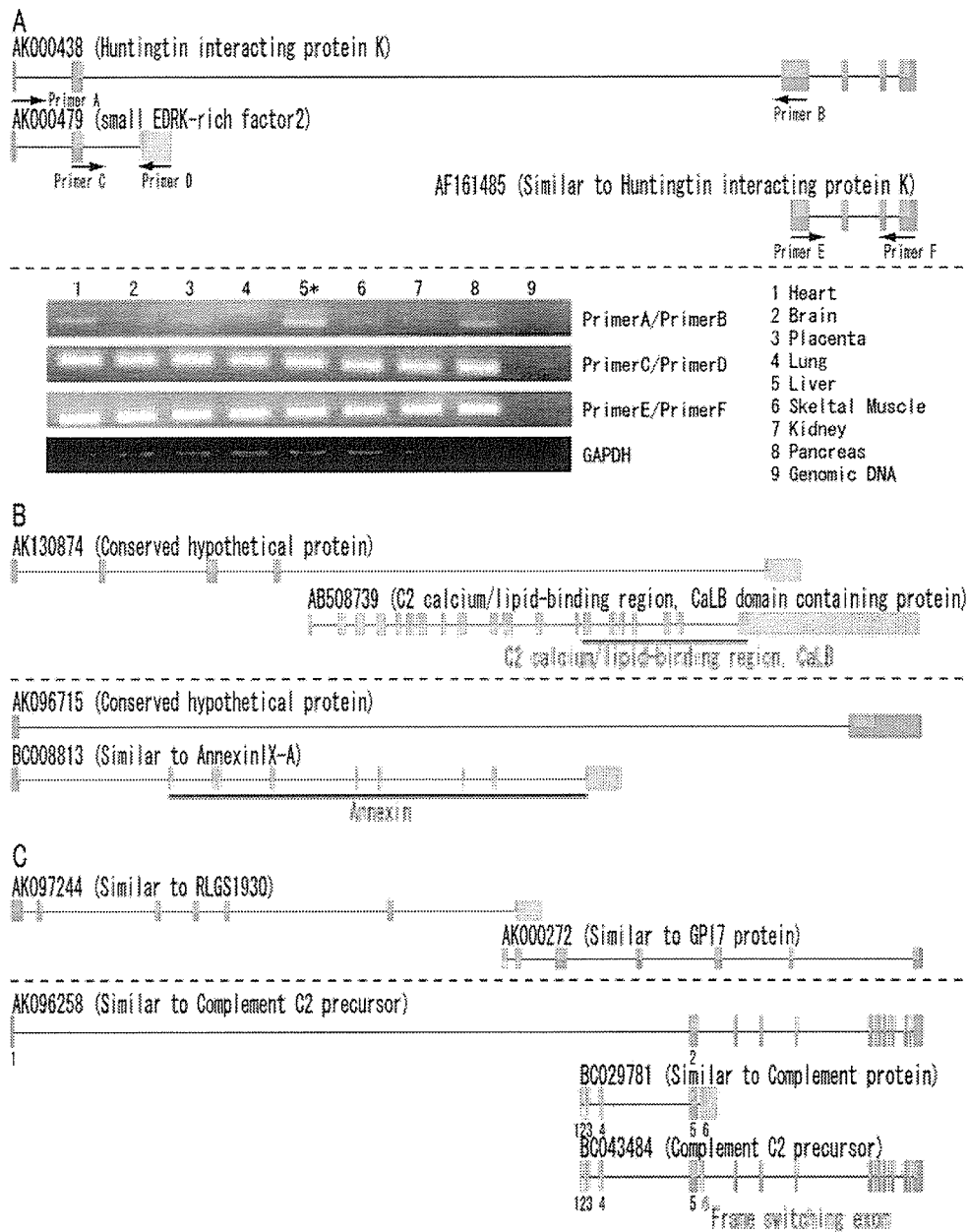
the reading frames of the second exon of AK096258 and fifth exons of BC029781 were different (Figure 5C, lower panel). Interestingly, in this gene, another alternative splicing variant, BC043484 used the latter-type reading frame for the N-terminal half of the protein and the former-type reading frame for the C-terminal half. In this variant, sixth exon seemed to serve as a 'frame switching' exon. Although the gene function of this locus still remains unknown, it is possible that the novel proteins discovered in the present study should also play roles which are distinct from but, at the same time, correlate with each other. Consistent with this possibility, emerging evidence also suggested that the cases in which mRNAs transcribed from a single locus are used as templates for two independent proteins seem not extremely rare (42). Further experimental validation should clarify which gets translated to produce functional proteins and which have regulatory roles.

The numbers of cases identified for each of the above-mentioned 'uncommon' patterns of the alternative splicing are shown in Table 6. In any case, it would be more appropriate to regard these cases as two 'genes' merged into a single 'locus' rather than as mutually 'alternative' variants occurring from the same locus, since, obviously, it is not probable that these variants share the same protein function. If we consider the evolutionary origins of these loci, they could originally have been two neighboring genes that then evolved to share some of their exons as a result of mutational changes. In each case, the extents to which multiple loci are merged vary. In this study, we applied strict criteria to select them (see Materials and Methods). Therefore, the numbers shown in Table 6 should be the minimum values, suggesting that it is not extremely rare to observe such 'uncommon' alternative splicing events in cells. Such mechanisms may allow for further diversification of the transcriptome of human genes. Although careful evaluations should be necessary (i.e. whether the cDNAs are originated from cancerous cells) before concluding whether the particular examples described here should have biological relevance, if any, such mechanism would enable the multi-faceted use of the biological information encoded in a given genomic region.

It should also be noted that without a combination of manual and computational analyses, this kind of alternative splicing variants would have been overlooked. By performing detailed manual and computational analyses, we could precisely identify these novel patterns of alternative splicing. Also, if no more than partial information of the alternative splicing exons had been available, these findings would have been utterly impossible. Judging from these results, alternative splicing is likely to be used in a more versatile manner to enable the diversification of the gene functions of human genes than was previously thought.

## DISCUSSION

In this paper we described the genome-wide identification and characterization of alternative splicing variants of human gene transcripts based on completely sequenced human full-length cDNAs in details. Starting with 56 419 cDNAs, we identified 18 297 complete alternative splicing



**Figure 5.** Examples of the ‘uncommon’ patterns of alternative splicing; ‘bridged’ (A), ‘nested’ (B) and ‘multiple CDS’ (C). These ‘uncommon’ patterns of alternative variations were defined as following: i) ‘bridged’: a locus in which two alternative splicing variants were arrayed tandemly without sharing any exons and another transcript ‘bridged’ these two variants, sharing at least some of its exons with both of them; ii) ‘nested’: a locus in which CDS region of one alternative splicing variant was not shared with another variant and iii) ‘multiple CDS’: a locus in which different ORFs >200 amino acid in length were annotated independently for different alternative splicing variants having overlapping CDSs of different reading frames. In the lower panel of (A), the results of RT-PCR are shown. Each photograph shows the amplicons of the indicated RT-PCR using the indicated primers. Tissue origins of the template RNAs are shown in the margin. The asterisk indicates a non-specific band. The coloring of the figures is the same as in Figure 4.

variants at 6877 loci of human genes and precisely annotated each of them.

Lander *et al.* (6) proposed that there should be five alternative splicing variants per locus in human genes. Although, our dataset used here is the largest one from a full-length cDNA collection, it is much smaller than that from dbEST, which includes more than 7 million ESTs. Therefore, some of the alternative splicing variants that actually exist in human genes might not be represented in our dataset (we

identified alternative splicing variants from 68% of the loci examined, with 2.7 variants per locus). However, our dataset should have two major advantages that largely compensate for the shortage of coverage. First, our dataset had been validated by heuristic annotations, and therefore various computational errors which could be easily discriminated by the human eye were excluded. Secondly, each of the identified alternative splicing variants was supported by full-length cDNAs whose sequences had been completely determined.



This enabled us to assure that each of the alternative splicing variants identified in this study corresponds to a complete form of a particular transcription unit. This feature was especially important when the relevance of the alternative splicing variants to the protein motifs or various subcellular localizations was evaluated. If comprehensive information of the cDNA sequences had not been available, those analyses would have been less reliable. Sometimes protein motifs are embedded over a wide region of the protein sequences, and all of the combinations of the alternative splicing exons may not be allowed. Besides, for certain types of subcellular targeting signals, such as signal peptides, the position within the protein sequence is critical.

Recently, it was reported that many of the alternative splicing variants seemed not to be evolutionarily conserved, and thus the biological significance of their existence might be questionable (43,44). The same argument may be used in the other direction as well. Alternative splicing could be the easiest road to diversification of the species. Future analyses of our full-length alternative splicing data should be useful to clarify which alternative splicing should be evolutionarily wobbly and which should provide a molecular basis for various species-specific biological features. Indeed, it will be important in future studies to discriminate which of the identified transcript variants have a general *raison d'être* and which play species-specific roles, because lack of such knowledge would severely restrict the potential power of the comparative genomic approaches utilizing the genomic sequences of many organisms which will be determined in the next few years. The information described here together with the availability of the accompanying physical full-length cDNA clone resources should lay firm groundwork for exploring how alternative splicing generates the functional diversification of the human transcriptome and proteome.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## ACKNOWLEDGEMENTS

The authors thank Y. Fujii, Y. Sato, H. Sakai, T. Habara, C. Yamasaki and M. Tanino for genome mapping and ORF prediction of H-Invitational full-length cDNA dataset. The authors thank F. Todokoro, H. Kawashima, E. Sekimori and H. Wakaguri for technical support of computational analysis. The authors thank K. Abe for experimental validation of the alternative splicing. The authors are also grateful to E. Nakajima for critical reading of the manuscript. This research was financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the Bund sministerium für Bildung und Forschung (BMBF, Grant NGFN-01GR0420), and the Japan Biological Informatics Consortium (JBIC). Funding to pay the Open Access publication charges for this article was provided by JBIC.

*Conflict of interest statement.* None declared.

## REFERENCES

- Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
- Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, **32**, 279–305.
- Smith, C.W. and Valcarcel, J. (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.*, **25**, 381–388.
- Schmucker, D. and Flanagan, J.G. (2004) Generation of recognition diversity in the nervous system. *Neuron*, **44**, 219–222.
- Wojtowicz, W.M., Flanagan, J.J., Millard, S.S., Zipursky, S.L. and Clemens, J.C. (2004) Alternative splicing of *Drosophila* Dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, **118**, 619–633.
- Lander, E.S., Linton, L.M., Birren, B., Nussbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Ewing, B. and Green, P. (2000) Analysis of expressed sequence tags indicates 35 000 human genes. *Nature Genet.*, **25**, 232–234.
- Ladd, A.N. and Cooper, T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, **3**, reviews0008.1–0008.16.
- Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
- Lee, C., Atanelov, L., Modrek, B. and Xing, Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
- Stamm, S., Riethoven, J.J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L. and Thanaraj, T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
- Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. *et al.* (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
- Zhang, Q.H., Ye, M., Wu, X.Y., Ren, S.X., Zhao, M., Zhao, C.J., Fu, G., Shen, Y., Fan, H.Y., Lu, G. *et al.* (2000) Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells. *Genome Res.*, **10**, 1546–1560.
- Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
- Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F. *et al.* (2002) Generation and initial analysis of more than 15 000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
- Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K. *et al.* (2004) Complete sequencing and characterization of 21 243 full-length human cDNAs. *Nature Genet.*, **36**, 40–45.
- Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
- Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
- Fairbrother, W.G., Yeh, R.F., Sharp, P.A. and Burge, C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
- Fairbrother, W.G., Yeo, G.W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P.A. and Burge, C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.
- Nakao, M. and Nakai, K. (2002) Improvement of PSORT II protein sorting prediction for mammalian proteins. *Genome Informatics*, **13**, 441–442.



22. Burset, M., Seledtsov, I.A. and Solovyev, V.V. (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.*, **28**, 4364–4375.
23. Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P. and Mattick, J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
24. Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
25. Kochiwa, H., Suzuki, R., Washio, T., Saito, R., Bono, H., Carninci, P., Okazaki, Y., Miki, R., Hayashizaki, Y. and Tomita, M. (2002) Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data. *Genome Res.*, **12**, 1286–1293.
26. Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C. and Kelso, J.F. (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.*, **11**, 1848–1853.
27. Lejeune, F. and Maquat, L.E. (2005) Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell Biol.*, **17**, 309–315.
28. Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, **300**, 1288–1291.
29. Will, C.L. and Luhrmann, R. (2005) Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol. Chem.*, **386**, 713–724.
30. Landry, J.R., Mager, D.L. and Wilhelm, B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.*, **19**, 640–648.
31. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
32. Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, **436**, 876–880.
33. Karin, M. (1999) How NF-kappaB is activated: the role of the IkappaB kinase (IKK) complex. *Oncogene*, **18**, 6867–6874.
34. Peters, R.T., Liao, S.M. and Maniatis, T. (2000) IKKepsilon is part of a novel PMA-inducible IkappaB kinase complex. *Mol. Cell*, **5**, 513–522.
35. Xing, Y., Resch, A. and Lee, C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.*, **14**, 426–441.
36. Nakao, M., Barrero, R.A., Mukai, Y., Motono, C., Suwa, M. and Nakai, K. (2005) Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res.*, **33**, 2355–2363.
37. Scharf, J.M., Endrizzi, M.G., Wetter, A., Huang, S., Thompson, T.G., Zerres, K., Dietrich, W.F., Wirth, B. and Kunkel, L.M. (1998) Identification of a candidate modifying gene for spinal muscular atrophy by comparative genomics. *Nature Genet.*, **20**, 83–86.
38. Faber, P.W., Barnes, G.T., Srinidhi, J., Chen, J., Gusella, J.F. and MacDonald, M.E. (1998) Huntingtin interacts with a family of WW domain proteins. *Hum. Mol. Genet.*, **7**, 1463–1474.
39. Magrangeas, F., Pitiot, G., Dubois, S., Bragado-Nilsson, E., Chereil, M., Jobert, S., Lebeau, B., Boisteau, O., Lethe, B., Mallet, J. *et al.* (1998) Cotranscription and intergenic splicing of human galactose-1-phosphate uridylyltransferase and interleukin-11 receptor alpha-chain genes generate a fusion mRNA in normal cells. Implication for the production of multidomain proteins during evolution. *J. Biol. Chem.*, **273**, 16005–16010.
40. Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. and Sorek, R. (2006) Transcription-mediated gene fusion in the human genome. *Genome Res.*, **16**, 30–36.
41. Wiemann, S., Kokocinski, A.K. and Poustka, A. (2005) Alternative pre-mRNA processing regulates cell-type specific expression of the IL411 and NUP62 genes. *BMC Biol.*, **3**, 16.
42. Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T. and Sugano, S. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res.*, **14**, 2048–2052.
43. Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature Genet.*, **34**, 177–180.
44. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl Acad. Sci. USA*, **102**, 2850–2855.

# DBTSS: DataBase of Human Transcription Start Sites, progress report 2006

Riu Yamashita, Yutaka Suzuki<sup>1,\*</sup>, Hiroyuki Wakaguri<sup>1</sup>, Katsuki Tsuritani, Kenta Nakai and Sumio Sugano<sup>1</sup>

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan and <sup>1</sup>Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Received September 15, 2005; Revised and Accepted October 21, 2005

## ABSTRACT

DBTSS was first constructed in 2002 based on precise, experimentally determined 5' end clones. Several major updates and additions have been made since the last report. First, the number of human clones has drastically increased, going from 190 964 to 1 359 000. Second, information about potential alternative promoters is presented because the number of 5' end clones is now sufficient to determine several promoters for one gene. Namely, we defined putative promoter groups by clustering transcription start sites (TSSs) separated by <500 bases. A total of 8308 human genes and 4276 mouse genes were found to have putative multiple promoters. Third, DBTSS provides detailed sequence comparisons of user-specified TSSs. Finally, we have added TSS information for zebrafish, malaria and schyzo (a red algae model organism). DBTSS is accessible at <http://dbtss.hgc.jp>.

## INTRODUCTION

Recently, a huge amount of comprehensive expression profile data obtained by various experiments, such as microarrays, has been made available. It is a challenging problem to uncover the regulatory networks among the expressed genes from these data. Information about promoters, which contain most of the binding sites of transcription factors, is indispensable for solving this question. To define promoter regions, precise information about transcription start sites (TSSs) is also required. Such data, however, are not easily obtained because the cDNA sequence data in repository sequence databases provide no guarantees regarding the 5' end of the sequences and because the computational prediction of promoters and TSSs still remains problematic (1). To overcome these difficulties several databases (2), including DBTSS (DataBase of

Transcription Start Sites) have been constructed. DBTSS contains TSS information of genes based on specific experiments (3,4). Clones constructed by full-length cDNA methods such as oligo-capping (5,6) or CAP-trapper (7,8) are mapped on to genome sequences to determine TSSs. Each TSS is determined based on the 5' end of the corresponding clone. DBTSS was first constructed in 2002, and has been improved by several major and minor updates. The original version (version 1) contained only human data (3). Two years later, we reported the addition of mouse TSS information (9) in version 3 (4). Here we introduce the new updates and additions since version 3, the most important one being the addition of putative alternative promoter information.

## NEW FEATURES

The current version of DBTSS, version 5, includes some notable improvements since the previous report, in addition to minor updates such as modifications of the interface and the result views.

One major improvement is that the amount of data for human TSSs has been significantly increased: in our report in 2002, we described 190 964 human clones which corresponded to 11 234 NCBI reference sequence cDNAs (RefSeq) (4). Because we added data from a new full-length cDNA project (10), DBTSS now contains 1 359 000 clones corresponding to 19 753 RefSeq cDNAs (Table 1). Since RefSeq cDNAs contain splicing variants as separate entries, we performed clustering of clones' information depending on their coordinate in the genome sequence; if their sequences overlapped, we regard them as the same locus. After clustering, our data correspond to 15 262 genes (Table 1). This is one of the largest collections of human 5' end cDNA sequences.

To check the quality of our TSS data, we compared DBTSS with the Eukaryote Promoter Database (EPD) (2). In EPD Release 82, there are 1871 promoters collected from the literature. Among them, we could map 1767 promoter

\*To whom correspondence should be addressed. Tel: +81 4 7136 3607; Fax: +81 4 7136 3607; Email: [ysuzuki@hgc.jp](mailto:ysuzuki@hgc.jp)

sequences to the human genome; 1639 of them mapping within 100 bases of the DBTSS TSSs, indicating that the data in DBTSS are consistent with the data obtained from ordinary methods.

In the next two sections, we will discuss two other major updates: alternative promoters (APs) and promoter comparison.

**ALTERNATIVE PROMOTERS**

Several genes are known to have multiple promoters which could be regulated in a different manner. These promoters,

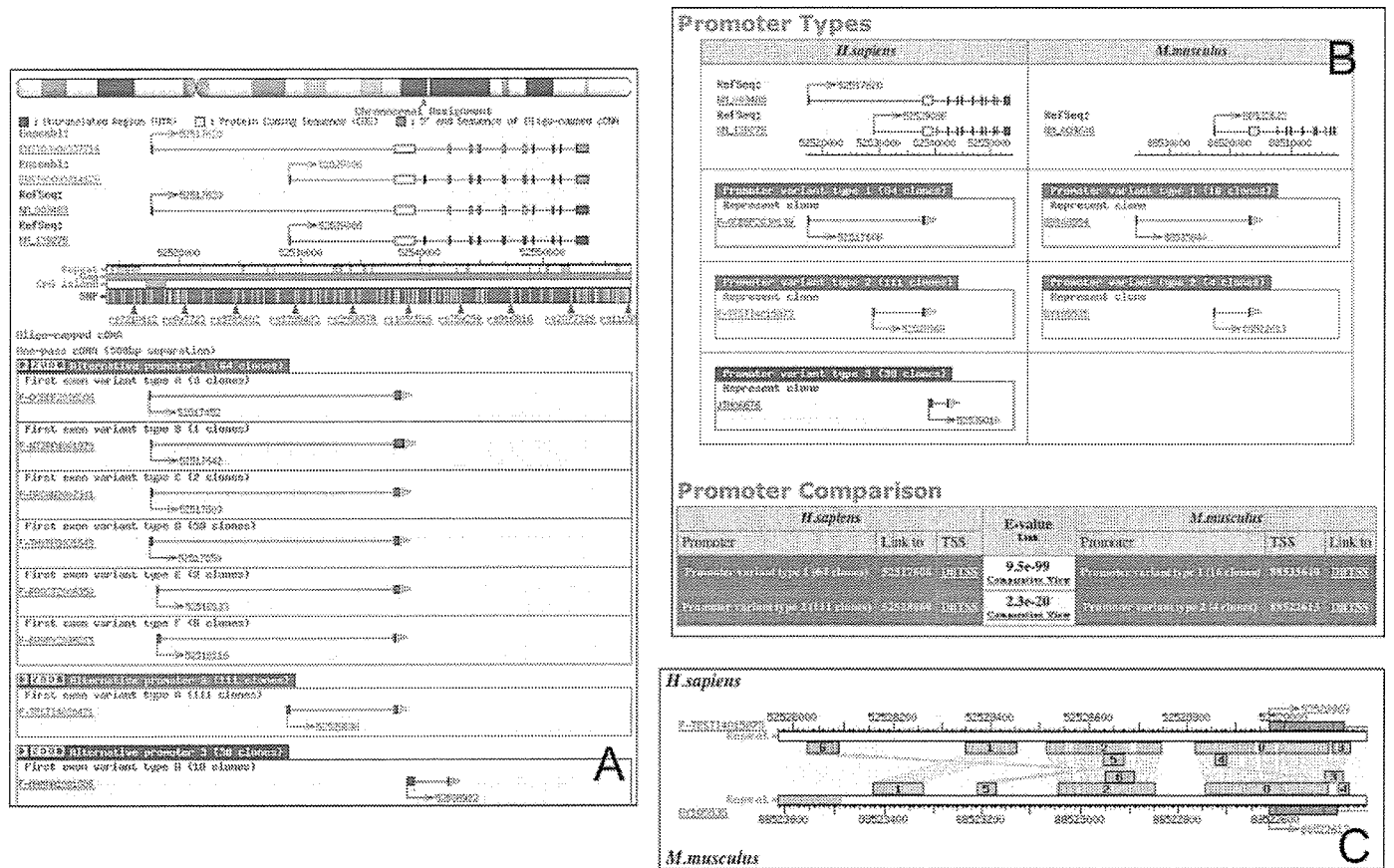
**Table 1.** Statistics of DBTSS

	No. of genes/ no. of RefSeq	No. of promoters	No. of TSSs	No. of clones
Human	15 262/19 753	30 964	452 117	1 359 000
Mouse	14 162/14 746	19 023	149 876	364 487
Zebrafish	3061/3075	3382	15 198	32 263
Malaria	1527/NA	NA	6908	10 236
Schyzon	3635/NA	NA	14 029	22 923

labeled as APs, could be useful to maximally exploit the relatively limited number of genes in the genome (11). However, no estimation of how many genes might have alternative promoters is available to date. Since DBTSS now has enough 5' end clones from human and mouse, we performed this estimate. This is the most important addition in version 5. Although the details of our analysis will be reported elsewhere (12), the procedure is summarized below.

To determine APs, we first collected all the TSSs from the same locus. TSSs located inside a RefSeq gene exon, with the exception of the first one, were removed in order to avoid artifacts caused by truncated 5' ends. We used several intervals to define AP clusters. The distribution of the number of putative alternative promoter containing genes shows a plateau before the interval size reaches 500 bp (12). We, therefore, clustered the clones using a 500 base interval, and defined each cluster as an promoter. We obtained 30 964 promoters, and 26 784 (86.5%) of them are within 500 bp. According to this procedure, 6954 human loci and 9886 mouse loci have only one promoter while 8308 human loci and 4276 mouse loci have two promoters or more. Figure 1A shows the three alternative promoters found in the gene encoding human A kinase

**A kinase anchor protein 1**



anchor protein 1 (AKAP1). It is notable that DBTSS also provides comparative information between human and mouse promoters. Figure 1B shows an example of comparative promoter analysis between orthologous genes. Two promoters were identified for the mouse gene for AKAP1. From this view, the representative APs are also available for alignment. By clicking ‘Comparative View’ in ‘Promoter Comparison’ in Figure 1B, the LALIGN-based alignment view, shown in Figure 1C is obtained.

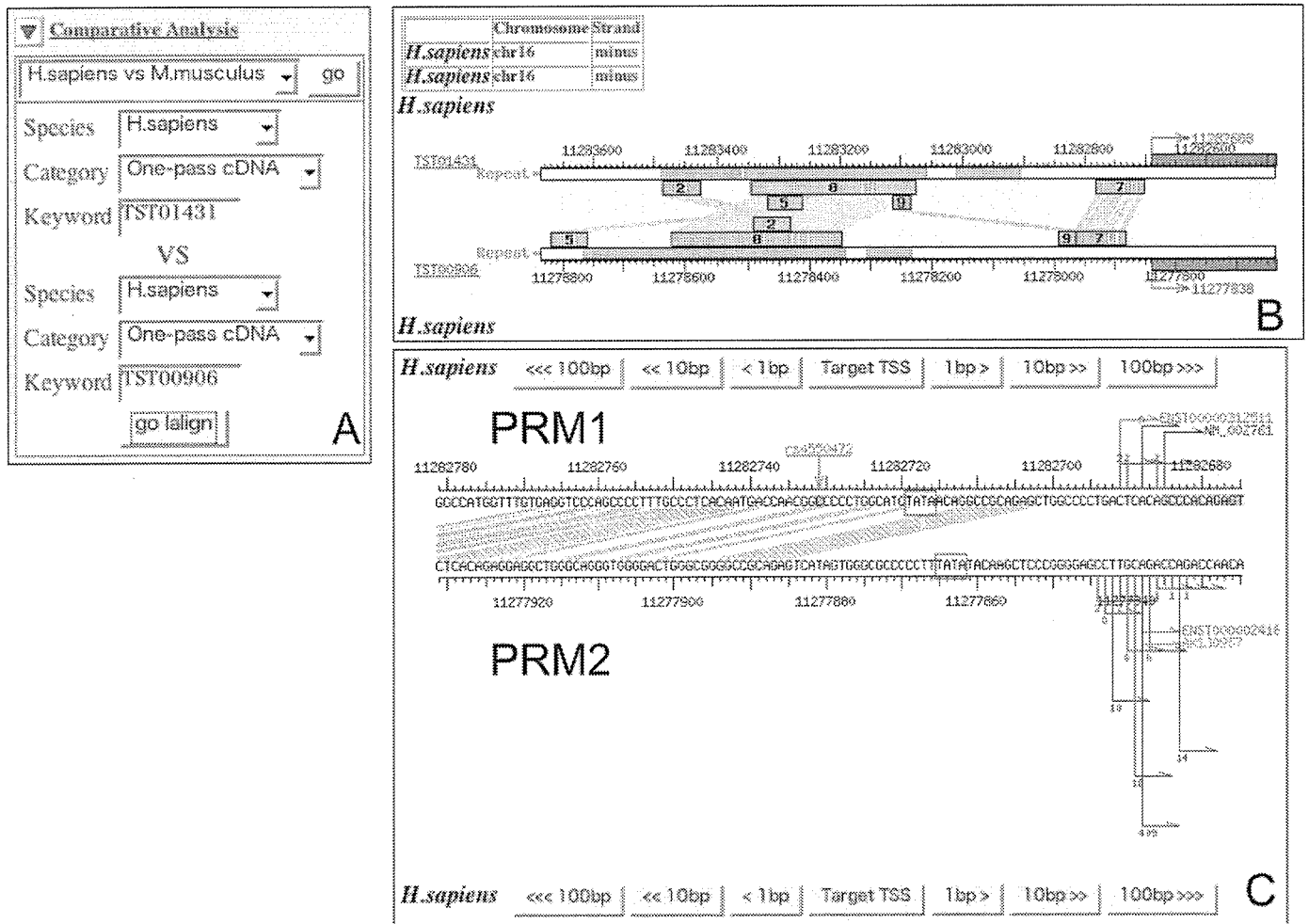
**COMPARATIVE PROMOTER ANALYSIS**

In the previous section, we showed an example of alternative promoter comparison between human and mouse. Before version 5 of DBTSS, these were precomputed, and the user could only obtain alignments between orthologous human and mouse genes. Despite being a useful idea, this sometimes failed to answer the user’s need for alignments of arbitrary promoter pairs, for instance, promoters of paralogous genes. We therefore implement a dynamic viewer allowing the alignment of any two TSSs present in DBTSS. Such analyses are

necessary to understand how transcriptional regulatory elements were conserved or diverged during gene and exon duplication. For example, in Figure 2A, the clones TST01431 of protamine 1 (PRM1: NM\_002761) and TST00906 of protamine 2 (PRM2: NM\_002762) are selected for alignment. Both genes are expressed in testis and are paralogous to each other. PRM1 is found in nearly all mammals while PRM2 is observed in relatively few mammals including human and mouse (13). In human, both genes are on chromosome 16, separated by ~5 kb (14). The obtained alignment and the determined conserved regions are shown in Figure 2B. In this case, the blocks ‘0’ and ‘7’ are highly conserved. The details of the alignment of both TSS regions are also available, as shown in Figure 2C. Especially, it is noteworthy that the putative TATA-box is inside block ‘7’ for the PRM1 promoter and outside of it for the PRM2 promoters (15).

**FUTURE PERSPECTIVE**

As shown in Table 1, we have added data from 32 263 zebrafish (*Danio rerio*) (16), 10 236 malaria (*Plasmodium*



**Figure 2.** An example of comparative analysis with any pair of TSSs. We show paralogous genes, protamine 1 (PRM1: NM\_002761) and protamine 2 (PRM2: NM\_002762), as an example. (A) By inputting the IDs of clones of PRM1 (TST01431) and PRM2 (TST00906) representative TSSs, users can obtain the results (B and C). (B) LALIGN analysis between two sequences. Note: smaller numbers indicate more highly conserved blocks. In this figure, the most conserved region between a pair is block 0; however, it includes *Alu* repeats. (C) The detail of the alignment of block 7. The putative TATA-boxes are marked with boxes.