Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes

Kouichi Kimura,[1] Ai Wakamatsu,[2,3] Yutaka Suzuki,[4,14] Toshio Ota,[2,11] Tetsuo Nishikawa,[1,2,3] Riu Yamashita,[5] Jun-ichi Yamamoto,[2,3] Mitsuo Sekine,[6] Katsuki Tsuritani,[5] Hiroyuki Wakaguri,[4] Shizuko Ishii,[2,3] Tomoyasu Sugiyama,[2,12] Kaoru Saito,[2] Yuko Isono,[2,3] Ryotaro Irie,[2] Norihiro Kushida,[6] Takahiro Yoneyama,[6] Rie Otsuka,[6] Katsuhiro Kanda,[7] Takahide Yokoi,[7] Hiroshi Kondo,[7] Masako Wagatsuma,[7] Katsuji Murakawa,[8] Shinichi Ishida,[8] Tadashi Ishibashi,[8] Asako Takahashi-Fujii,[9,13] Tomoo Tanase,[9,13] Keiichi Nagai,[1,2,10] Hisashi Kikuchi,[6] Kenta Nakai,[5] Takao Isogai,[2,3] and Sumio Sugano[4]

[1]Life Science Research Laboratory, Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo, 185-8601, Japan; [2]Helix Research Institute, Kisarazu, Chiba, 292-0812, Japan; [3]Reverse Proteomics Research Institute, Kisarazu, Chiba 292-0818, Japan; [4]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Minato-ku, Tokyo, 108-8639, Japan; [5]Human Genome Center, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, 108-8639, Japan; [6]Genome Analysis Center, Department of Biotechnology, National Institute of Technology and Evaluation, Shibuya-ku, Tokyo, 151-0066, Japan; [7]Life Science Group, Hitachi, Ltd., Kawagoe, Saitama, 350-1165, Japan; [8]Hitachi Science Systems, Ltd., Kokubunji, Tokyo, 185-8601, Japan; [9]Takara Shuzo Co., Ltd., Noji-cho, Kusatsu, Shiga, 525-0055, Japan; [10]Advanced Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo, 185-8601, Japan

By analyzing 1,780,295 5′-end sequences of human full-length cDNAs derived from 164 kinds of oligo-cap cDNA libraries, we identified 269,774 independent positions of transcriptional start sites (TSSs) for 14,628 human RefSeq genes. These TSSs were clustered into 30,964 clusters that were separated from each other by more than 500 bp and thus are very likely to constitute mutually distinct alternative promoters. To our surprise, at least 7674 (52%) human RefSeq genes were subject to regulation by putative alternative promoters (PAPs). On average, there were 3.1 PAPs per gene, with the composition of one CpG-island-containing promoter per 2.6 CpG-less promoters. In 17% of the PAP-containing loci, tissue-specific use of the PAPs was observed. The richest tissue sources of the tissue-specific PAPs were testis and brain. It was also intriguing that the PAP-containing promoters were enriched in the genes encoding signal transduction-related proteins and were rarer in the genes encoding extracellular proteins, possibly reflecting the varied functional requirement for and the restricted expression of those categories of genes, respectively. The patterns of the first exons were highly diverse as well. On average, there were 7.7 different splicing types of first exons per locus partly produced by the PAPs, suggesting that a wide variety of transcripts can be achieved by this mechanism. Our findings suggest that use of alternate promoters and consequent alternative use of first exons should play a pivotal role in generating the complexity required for the highly elaborated molecular systems in humans.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to DDBJ under accession nos. DA000001–DA999999, DB000001–DB294747, DB294748–DB384947, BP192706–BP383670, AU279383–AU280837, and AU116788–U160826.]

One of the most striking findings revealed by the Human Genome Project is that the human genome contains only 20,000–25,000 kinds of protein-coding genes (International Human Genome Sequencing Consortium 2004). This number is unexpectedly small compared with the total gene numbers in yeast, fly, and worm genomes, which are estimated to be 6,000, 14,000, and 19,000, respectively (Goffeau et al. 1996; *C. elegans* Sequencing Consortium 1998; Adams et al. 2000). It is supposed that there must be other factors in addition to mere gene numbers to satisfy the prerequisites that enable the human genome to fabricate such highly elaborated systems as the brain and immune systems. To explain this, it has been hypothesized that multifaceted use of the genes should play a pivotal role in functional

Present Addresses: [11]Tokyo Research Laboratories, Kyowa Hakko Kogyo Co., Ltd., Machida, Tokyo, 194-8533, Japan; [12]School of Bionics, Tokyo University of Technology, Hachioji, Tokyo, 192-0982, Japan; [13]Takara Bio Inc., Otsu, Shiga, 520-2193, Japan.
[14]Corresponding author.
E-mail ysuzuki@hgc.jp; fax +81 4 7136 3607.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4039406.

diversification of human genes without affecting the total gene number (Ewing and Green 2000).

Multifaceted use of the genes would be enabled either by the production of slightly different transcripts, which are finely tuned for specific purposes from a single gene locus, or by employing essentially the same transcript in different circumstances, or by the combination of these mechanisms. As for the first possibility, recent reports showed that alternative splicing (AS) is employed in about half of all human genes, producing more than three different transcript variants per locus on average (Lander et al. 2001). Various transcripts produced by AS are consequently translated into proteins with slightly different structures and functions, and thus this mechanism is thought to provide a molecular mechanism for the fine tuning of the gene functions of a single locus (Lopez 1998; Black 2000). As for the second possibility, the use of alternative promoters (APs) has been presumed. By utilizing APs, which consist of different modules of transcriptional regulatory elements, diversified transcriptional regulation should be enabled at a single locus (Landry et al. 2003). Combinatory use of these two possibilities (AS and APs) would even further increase the potential complexity of the products expressed from a single gene; for example, multiple separated promoters might independently direct transcription from different genomic positions and the subsequent variation in the first exons might result in the production of N-terminally different proteins.

Actually, for some human genes of particular interest, in vitro and in vivo experiments have verified that such complex diversification takes place within a cell. For example, the *SHC1* gene has two APs and produces three different transcripts encoding protein isoforms of 46, 52, and 66 kD (Luzi et al. 2000). The transcript encoding p46/p52 is transcribed from the proximal promoter with a ubiquitous expression pattern. On the other hand, the transcript encoding p66, whose biological functions are completely different from those of p46/p52 because of the presence of one additional collagen homology domain at its N terminus, is driven by a distal promoter and is specifically expressed in limited types of cells. The promoters of these two isoforms are approximately 4 kb apart from each other and the repertories of the predicted potential *cis*-acting regulatory elements are completely different. In addition, recent studies also demonstrated that the histone acetylation and cytosine methylation statuses are significantly different between the two APs (Ventura et al. 2002).

Although our understanding of the comprehensive features of AS has been rapidly advancing with the compiling of EST data (Lee et al. 2003; http://www.bioinformatics.ucla.edu/ASAP/), very little is understood about the genome-wide features of the APs so far. Indeed, in spite of increasing general interest and need, almost no reports or databases have provided a genome-wide view of which population of human genes is regulated by APs and what biological consequences such diversification of the transcriptional modulation would bring about. In our opinion, this is because of a general lack of information about the transcriptional start sites (TSSs) and adjacent putative promoter regions (PPRs). For systematic identification of the APs, highly redundant sequence data would be essential. However, the coverage of the EST data at the 5′-ends has generally been low, since the conventional cDNAs are constructed utilizing the 3′-end poly(A) without any selection method for the opposite end, the 5′-end cap structure (Suzuki et al. 1997). Besides, even if available cDNA sequences have already covered the corresponding regions, it cannot be assured that their 5′-ends correspond to the real TSSs without in-depth analysis. For these reasons, the massively accumulated current EST data could not be directly used for the identification and analyses of the APs. Although a pioneering study was done by Zavolan et al. (2003) making use of full-length cDNAs, their data were limited to about 60,000 mouse cDNAs.

We have been collecting full-length cDNAs of human genes by constructing cDNA libraries using the oligo-capping method (Suzuki and Sugano 2003; Ota et al. 2004). Using the full-length cDNA sequence data, we have also demonstrated that the exact positions of the TSSs and the adjacent putative promoters could be retrieved from the human genomic sequences in a high-throughput manner (Suzuki et al. 2001b). Our human full-length cDNA data accumulated so far and the retrieved adjacent upstream PPR information are integrated in our database, DBTSS (Suzuki et al. 2004; http://dbtss.hgc.jp) and have been made publicly available now. In this study, we further expand our 5′-end sequence data up to 1.8 million cDNAs (all of the sequence data have been registered in DDBJ and the physical cDNA clones are available on request). With the increase of the entire dbEST data (6 million entries as of March, 2005) by about 30%, these data extensively complement the 5′-ends in the current EST collection. By using our new 5′-end oligo-cap cDNA data, we were able to glimpse for the first time an overview of how the TSSs are clustered and in what manner the adjacent putative promoters are used multifacetedly in humans. Here we report our first genome-wide analysis of the alternative use of putative promoters using our unprecedented collection of 5′-end cDNA data.

Results and Discussion

Massive identification of the transcriptional start sites using 1.8 million 5′-end sequences of putative full-length cDNAs

We computationally mapped 1,780,295 5′-end sequences of putative full-length cDNAs onto the human genome (hg_17 of UCSC Genome Browser; Hsu et al. 2005; http://genome.ucsc.edu/). The cDNAs were collected from full-length-enriched cDNA libraries constructed from 164 kinds of human tissues and cultured cells using our cap selection method, the "oligo-capping" (for further details of the protocol for the library construction, see Suzuki and Sugano 2003). The library information is summarized in Supplemental Table 1. The number of cDNA sequences that were successfully mapped from their first base was 1,536,604. Using the positional information of the exactly and uniquely mapped 5′-ends of the oligo-cap cDNAs, we defined the TSSs as the independent genomic positions to which the first bases of the oligo-cap cDNAs were mapped (Fig. 1).

Before generating the final TSS data set, we excluded those cDNA sequences that had been mapped inside the second or later exons of any other cDNAs or RefSeq sequences, which represent previously characterized transcripts (Pruitt et al. 2005; http://www.ncbi.nim.nih.gov/RefSeq/). This was intended to minimize dubious identification of TSSs due to erroneously cloned truncated cDNAs. Although it is relatively minor, our "oligo-cap" cDNA collection does include a certain population of those erroneously cloned species (see also Suzuki et al. 2001a). We presumed 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 presump-

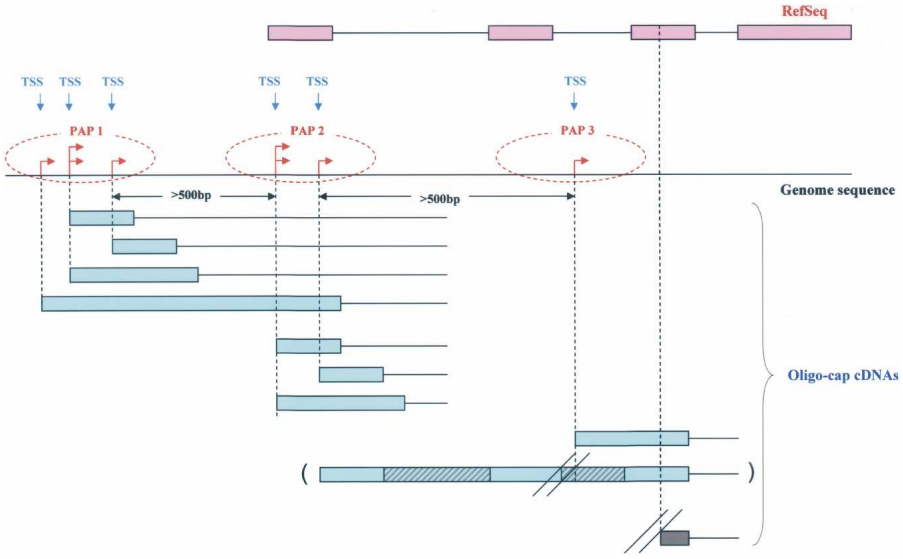


Figure 1. Identification of the putative alternative promoters in human genes. Schematic representation of the mapping of the 5′-ends of the oligo-cap cDNAs, the determination of the TSSs, and clustering of the TSSs to identify the PAPs. The boxes and lines represent exons and introns, respectively. The RefSeq sequences and the oligo-cap cDNAs are in red and blue, respectively. The lowest gray oligo-cap cDNA is excluded from the data set, since its 5′-end is located within an internal exon of the RefSeq. The third-lowest oligo-cap cDNA is accepted because the truncation of the erroneously sliced second-lowest transcript would otherwise need to be hypothesized to explain its presence, and the chance of the combination of such events should be low. The shaded boxes represent the retained introns. Altogether, this case consists of 8 "full-length" oligo-cap cDNAs that are mapped at 6 TSSs, clustered into 3 PAPs.

tion is based on the fact that the combination of multiple errors, for example, truncation followed by erroneous oligo-capping that occurred on an immaturely spliced form, would be required to erroneously generate such cDNAs (see example in Fig. 1). Because the expected rate of such tandem errors should be very small, we considered that most of the selected 5′-ends of the oligo-cap cDNAs should correspond to actual TSSs in vivo and that such selection should be helpful to reduce the false-positive identification of TSSs. We did not exclude the TSSs that mapped in the first exons because it has been shown that the locations of TSSs fluctuate to some extent in most genes (Suzuki et al. 2001a). The probability that a truncation event occurred within the typically relatively short first-exonic region and that spurious oligo-capping occurred on the same transcript can be considered to be low, and thus we considered that most TSSs that mapped in first exons should be the result of TSS fluctuations and not of truncated forms.

As a result, we obtained a data set of 1,355,367 TSSs. Given this number, the overall frequency of the full-length cDNAs could be estimated as 88%, which is closely consistent with our previous evaluations of the full-length-ness of the libraries. Furthermore, we validated the accuracy of the TSSs identification by directly comparing the positions of the previously reported TSSs in Eukaryotic Promoter Database (EPD Release 82; http://www.epd.isb-sib.ch/; Schmid et al. 2004) with those identified in the present study. As shown in Figure 2, at least 86% (155/181) of the positional information of the TSSs previously detected using different methods [92% (1056/1144) if the TSSs determined by high-throughput methods were also summed] were consistent with our results. Although much more analysis would need to be done, including exhaustive RACE analyses, before we could actually conclude that all of the TSSs identified in this study correspond to real TSSs from which transcription was truly initiated in vivo, we consider that the error rate of the current data set should be sufficiently low to glimpse the general features of the APs via the analysis presented below.

Among the identified TSSs, 1,171,916 TSSs corresponded to RefSeq genes (LocusLink locus; now superseded by Entrez Genes) (Maglott et al. 2005; http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene). Those TSSs collectively covered 14,628 RefSeq genes, that is, 85% of all, RefSeq genes (17,217 genes). The base composition of the exact site of the TSSs was 51%, 25%, 17%, and 7% for A, G, C, and T, respectively, which is consistent with our previous observations using a smaller data set (Suzuki et al. 2001a). Interestingly, the base composition of the 183,415 TSSs that did not correspond to RefSeq genes was significantly different: 41% (A), 17% (G), 17% (C), and 24% (T). Especially notably, the frequency of T was more than three times higher than in RefSeq genes. Assuming that the major part of the protein-coding genes are covered by RefSeq genes and the nonprotein-coding transcripts are enriched in the noncovered population, the above observation might imply that the nature of the transcriptional initiation mechanism is different between protein-coding and nonprotein-coding RNAs. In this study, we focused on the TSSs of RefSeq genes for the analyses described below, although it would also be interesting to analyze the nature of this emerging class of non-protein coding RNAs further (Okazaki et al. 2002).

More than half of human genes are regulated by alternative promoters

To identify possible AP-containing genes, we clustered the TSSs of RefSeq genes by binning them by >500-bp intervals (Fig. 1). We tentatively employed this criterion because our previous studies showed that the range over which the TSSs are scattered is on average 62 bp, with a standard deviation of 20 bp (Suzuki et al. 2001a), and thus the chance that TSSs belonging to a single cluster would be scattered over a range of >500 bp would be very low. Also, the statistics shown in Supplemental Figures 1 and 2 support the notion that a 500-bp interval is a strict enough parameter to make a meaningful separation of the clusters of TSSs while at the same time avoiding the erroneous separation of single TSS clusters into multiples. Besides, >500-bp intervals would allow us to hypothesize that the identified putative alternative promoters (PAPs) should not share most of the surrounding sequence context of the TSSs and the major part of the transcriptional regulatory modules, which constitute the direct docking platform of the general transcription factors and the RNA polymerase complex, and the immediate upstream sequences, to which other transcription regulatory factors are most likely to be recruited.

The clustering analyses revealed that a total of 269,774 different positions of TSSs were clustered into 30,964 groups. Each

A

	#EPD TSS total	#EPD TSS in DBTSS genes	#EPD TSSs within 100 bp of DBTSS TSS	#EPD TSSs >100 bp downstream of DBTSS TSS	#EPD TSSs >100 bp upstream of DBTSS TSS
Determined by high-throughput methods	1608	1020	955	33	32
Determined by conventional methods	286	181	155	5	21
total	1799	1144	1056	38	50

B

	EPD AP total	EPD AP covered by DBTSS "gene"	EPD AP overlapping with DBTSS "PPRs"
#gene	62	59 (95 %)	47 (76 %)
#promoter	134	128 (96 %)	104 (78 %)

C

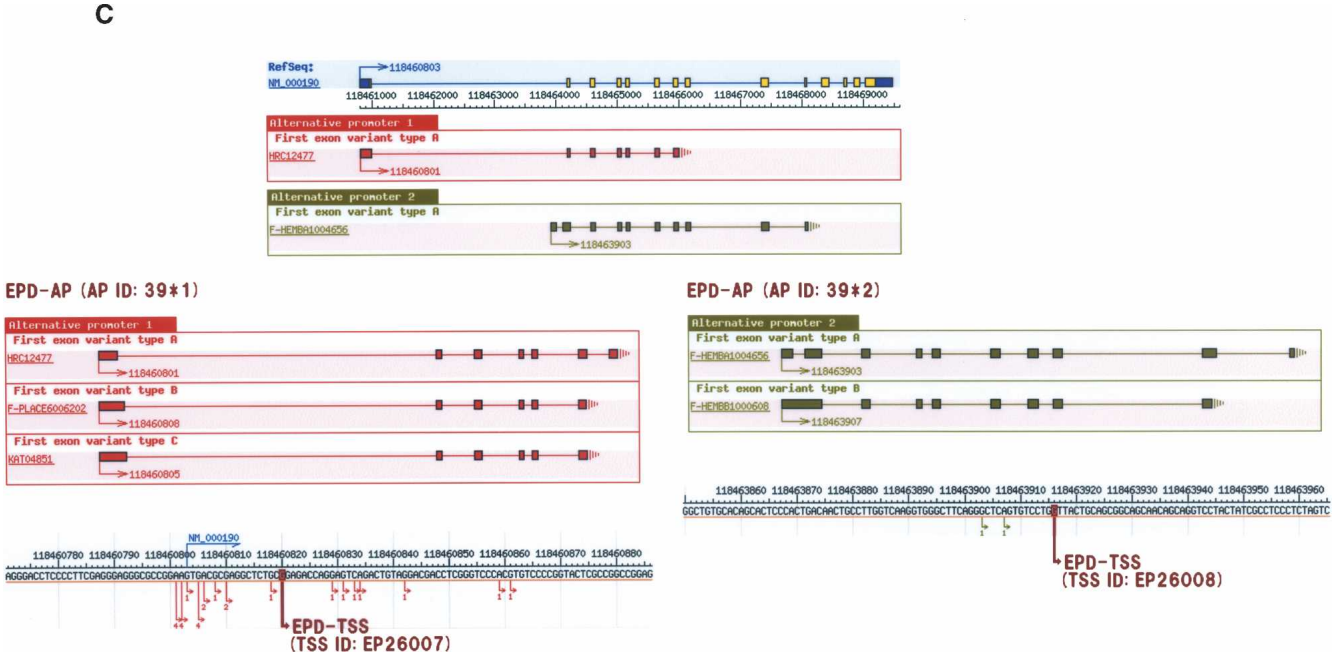


Figure 2. Comparison of the DBTSS data with the previously characterized TSSs and APs. TSSs (*A*) and APs (*B*) identified by the DBTSS data were compared with those characterized in previous studies. When a TSS/AP registered in EPD was located within 100 bp of one in DBTSS, they were counted as "overlapping." The margin of 100 bp was allowed considering fluctuations of the TSSs (Suzuki et al. 2001a). (*A*) The "overlapping" was counted separately for the TSS data obtained from high-throughput cDNA cloning methods (like ours) and that from conventional methods, such as RACE and nuclease protection assays. Note that as some of the TSSs were identified by multiple methods, the total numbers in the third line are not always the sum of the above two. (*B*) First column is the total number of EPD genes registered as "alternative promoter-containing genes" and the number of the corresponding promoters; second column is the coverage of the DBTSS against EPD at the gene level; third column is coverage of the DBTSS against EPD at the promoter level (all APs were covered by DBTSS PPRs). (*C*) The case in which EPD data and DBTSS data were overlapping with each other is exemplified by the case of the human hydroxymethylbilane synthase gene (NM_000190). RefSeq exons are shown in blue (non-coding regions) and yellow (coding regions) boxes and the DBTSS exons are shown in red (PAP group 1) and green (PAP group 2) boxes. The *lower* panels are magnifications of the *upper* panel(s). The TSSs are represented by arrows of the corresponding colors. The IDs of corresponding EPD data are shown. Note that there are variations in the first exon patterns even within the same PAP group (alternative donor in PAP1 and retaining intron in PAP2) and the TSSs are fluctuating. For additional examples, see Supplemental Table 3.

cluster identified was considered to represent a PPR, and for gene loci with multiple PPRs, each individual PPR was designated a PAP. (Note that they are "putative" promoters, because our TSS determination is mostly based on a single method, oligo-capping, and thus confirmation using other methods as well as functional analyses of the promoters remains essential).

To our surprise, 7674 (52%) of the RefSeq genes contained more than a single TSS cluster (24,010 clusters in total), which may be regulated by distinct APs. If any, there were 3.1 distinctive PAPs per gene, on average. By way of comparison, genes having alternative splicing forms are thought to constitute 60% of human genes and, if any, there are about 3 AS variants per locus (Lander et al. 2001; Modrek and Lee 2002). (Note that the EST data set from which the AS variants were identified did not include the majority of the new data used in the present study. The number of available cDNAs at the time the AS variant studies

were performed was about 2 million, which is almost the same as the number in the currently used data set.) Our analysis indicates that alternative promoters occur at least with the same order of frequency as the AS variants (see also the section on first exon patterns below). The regulation of transcription thus appears to be more dynamic than previously thought. Although it has been reported that splicing events are highly variable in human genes, this is the first report showing that alternative transcriptional regulation is about equally prevalent in human genes. All of the information produced in this study has been integrated in our database, DBTSS, and the raw data have been made freely downloadable for anonymous public users.

We also examined how many of the previously characterized APs were represented in our newly generated PAP data set. As shown in Figure 2B, there were 134 APs in 62 genes registered in the latest version of EPD. Among them, 104 promoters (78%) in 47 genes (76%) were correctly represented in our data set. Figure 2C shows a case in which the identified APs were supported by both EPD and the present DBTSS data. Additional examples are presented in Supplemental Table 3, in which some of the PAPs were consistent with the EPD data and others were unique in our data set. Based on the high coverage [about 95% (59/62 and 128/134) at the gene and promoter levels] and the consistency [about 80% (47/59 and 104/128) at the gene and promoter levels] with the previous data, we consider that our PAP data should be a useful resource for exploring the genome-wide features of the APs, providing accurate representations of the APs with high reliability of each entry.

Although the frequency of the PAPs was unexpectedly large, the estimated frequency should be near the lower boundary of the actual complexity. First, although we used an unprecedented amount of TSS data (80 TSSs per gene on average, which led to the identification of 3.1 PPRs; see Table 1), additional PAPs could remain undiscovered. Actually, the PAPs used with very low efficiency or PAPs used in very limited cell types might not be included in our data set. Secondly, the procedure of clustering the TSSs with the 500-bp bin might have resulted in a number of falsely clustered distinct PAPs. Lastly and most importantly, to exclude possible truncated cDNAs, we had to neglect the possibility that some of the TSSs are really located within the internal exons of other variant transcripts in this study. Indeed, among the 24 EPD promoters that corresponded to DBTSS PAP-containing genes but did not overlap with the DBTSS PAPs (Fig. 2B), 4 were located in internal positions of the second or later exons. Although they should correspond to genuine APs, they were removed from our data set in exchange for increasing the reliability of the data. We were concerned that the error rate (frequency of truncated cDNAs) should be much higher among those cDNAs whose 5′-ends were mapped internal exonic regions than those cDNAs that were used in the present study. Unlike in the latter group, only single-step errors (erroneous oligo-capping on the truncated mRNAs) would be accounted for dubious generation of the truncated cDNAs in the former group (see Fig. 1, compare the bottom two lines), which we estimated to occur by the frequency of 10%–15%. Therefore, for this purpose, we considered extensive filtration should be indispensable (see Supplemental Fig. 5). In spite of the technical difficulty, further detailed analyses focusing on this subject should be especially fruitful, since in such "internal" PPRs, the corresponding genomic regions should serve as both exons and promoters, and thus should have several features different from other canonical promoters.

Table 1. Distribution of the putative alternative promoters

No. PAPs	No. locus	No. included TSS positions	No. cDNA clones (avg.)
1 (PAP-less)	6954 (48%)	70,175	43
2	3724 (26%)	67,846	83
3	1821 (12%)	44,455	115
4	1003 (7%)	32,582	160
5	490 (3%)	19,962	166
6	294 (2%)	13,937	159
7	147 (1%)	7948	184
8	85 (0.6%)	4912	194
9	42 (0.3%)	2167	163
10	25 (0.2%)	1650	164
>10	43 (0.3%)	4140	341
total	14,628	269,774	80

Characterization of the identified PAPs

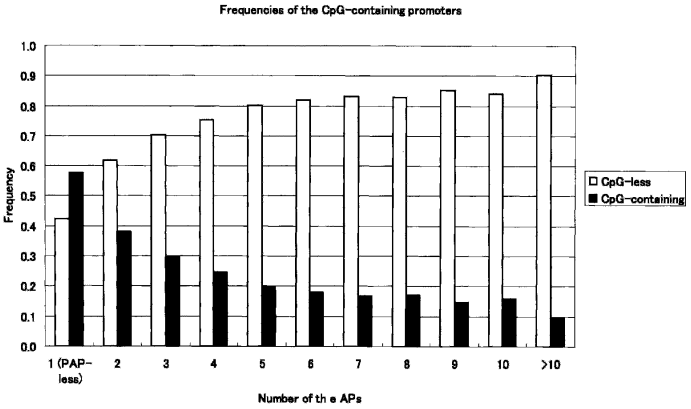
To study the biological implications of the presence of so many PAPs embedded in the human genome, we first examined the differences between the PAP-containing (i.e., with ≥2 PPRs) promoters and the PAP-less (i.e., with only 1 PPR) promoters with respect to the frequencies of basic promoter elements, namely, TATA boxes and CpG islands. As shown in Figure 3, the frequencies of the TATA box-containing (Fig. 2B) and the CpG island-containing (Fig. 3A) promoters were 17% and 58% for the PAP-less promoters, respectively. While the frequency of the TATA box-containing promoters was about the same in the PAP-containing population (11%), the frequency of the CpG island-containing promoters was less than half of that in the PAP-less population (28%). We further examined the ratio of the CpG-containing to CpG-less promoters as related to the number of PAPs. As shown in Figure 3A, the number of CpG-containing promoters per locus did not increase at all as the number of PAPs increased, indicating that the probability of the occurrence of the CpG islands was not proportional to the total number of PAPs. In contrast, the population of CpG-less promoters did increase as the number of PAPs increased. We previously showed that CpG-containing promoters predominantly consisted of promoters whose expression was ubiquitous (Yamashita et al. 2005). It may be generally true of the PAPs that there is one (or a few) ubiquitously expressed promoter(s), accompanied by, on average, 2.6 other promoters used for tissue-specific or signal-dependent expression.

Identification of tissue-specific PAPs

Having observed the above results, we examined whether some of the PAPs appear to contribute to different usage of the genes depending on the tissue types in which they are expressed. Since none of the cDNA libraries used in the present study were produced by a normalization method, it was possible to count the number of the represented cDNA clones within the library for the first approximation of their expression levels. To examine the tissue-specific use of a PAP, the sources of the cDNA libraries of all of the TSSs belonging to a given locus were noted and the information about which tissue they were derived from was collected. The probability of the enrichment of any particular tissue type as a source for cDNAs for a given PAP was judged by considering whether the tissue-source composition of the TSSs belonging to that PAP was skewed compared with the overall tissue-source composition of the TSSs belonging to the same locus. For this, we calculated the hypergeometric distribution of the component TSSs for each of the PAPs. After the Bonferroni

A

	CpG+	CpG-	Total
PAP-less	4011 (58%)	2943 (42%)	6954
PAP-containing	6686 (28%)	17324 (72%)	24010
total	10697 (34%)	20267 (65%)	30964



B

	TATA+	TATA-	Total
PAP-less	1181 (17%)	5773 (83%)	6954
PAP-containing	2720 (11%)	21290 (89%)	24010
total	3901 (13%)	27063 (87%)	30964

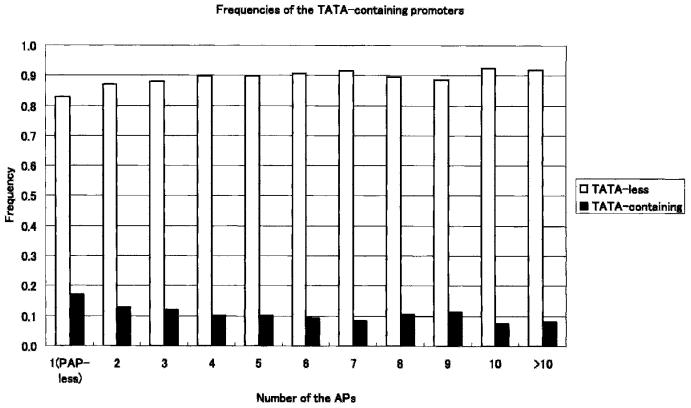


Figure 3. Relationship between the PAPs and the CpG islands and TATA boxes. Frequencies of the CpG island (*A*) and TATA box (*B*) containing PPRs. In the *right* panels, the relationship between the number of PAPs (*x*-axis) and the frequency of the corresponding promoter motif (*y*-axis) is shown.

correction, the hits with $P < 0.01$ were selected. The statistical procedure and the subsequent correction were designed so that the statistical bias depending on the coverage of the cDNA libraries (number of the cDNAs sequenced from the library) should be minimized.

Among the PAPs in 7674 PAP-containing loci, 1803 PAPs in 1333 loci (17%) met the above criteria and were designated as "tissue-specific PAPs" (Fig. 4). This indicates that one of the important potential roles of the PAPs should be independent regulation of a gene in a particular tissue-type, while the same gene is subject to different types of regulation in different cellular contexts. The tissue in which such specific usage of the PAPs was most frequently observed was testis tissue. It has been reported that the transcriptional regulation in testis is significantly different from that in other normal tissues. The chromosome condensation and methylation status of the genome are reported to be very special in testis. Actually, many of the antigens specifically expressed during the course of tumorigenesis are expressed in testis under normal conditions, and thus are so-called cancer-testis antigens (Maio et al. 2003; Kalejs and Erenpreisa 2005). The fact that testis was identified as the richest source of the PAPs may indicate that the development of the exclusive use of some types of transcriptional modulation has been indispensable to realizing the functional molecular systems from the heavily skewed genomic structures in germ-line cells. It was also intriguing that the second-richest origin of the "tissue-specific" PAPs was brain. The specific diversification of a large population of genes should also be indispensable for the fabrication of such an elaborate system as the brain. Considering that the largest populations of the tissue-specific PAPs were related to testis and brain, which separate human species from other organisms most distinctively, it seems likely that further detailed analyses of the PAPs from the molecular evolutional viewpoint would eventually lead to the identification of genes whose diversification has collectively provided a molecular basis for unique features in humans.

GO classification of the PAP-containing genes

We further investigated in what kinds of genes and with what frequency the alternative use of the first exons was likely to result in alteration of the encoded protein sequences, and what biological consequences it would bring about. We first categorized the PAP-containing genes according to the GeneOntology (GO) classification (Harris et al. 2004; http://www.geneontology.org/). The distribution of the GO categories was then subjected to hypergeometric distribution analysis. As shown in Table 2, GO categories of "ATP binding," "phosphorylation," and "kinase" were significantly enriched. The fact that PAP-containing genes were especially enriched in genes for these so-called signal transduction molecules seemed reasonable, since complex transcriptional regulation should be indispensable for appropriate transmission of signals depending on the dynamically changing conditions of the cells.

In contrast, the GO category 'extracellular' was reduced. For extracellular secretory proteins, there are protein-sorting motifs, called "signal peptides," at the N terminus. Because of this re-

A

Rank	Tissue	#Tissue-specific APs	#Locus	Ratio
1	Testis	416	408	0.24
2	Brain	268	261	0.15
3	Immune cells	84	81	0.05
	Total	1,731	1,287	

B

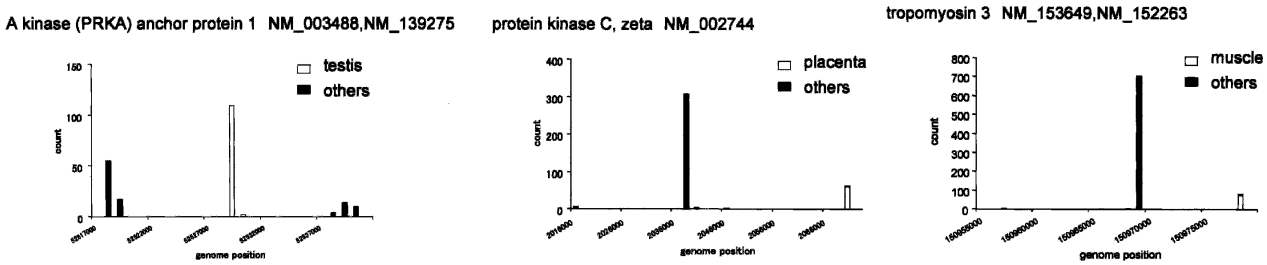


Figure 4. Tissue-specific usage of PAPs. (*A*) The number of PAPs that are used in a tissue-specific manner. For the detailed definition of the tissue specificity, see the Methods section. (*B*) Examples of tissue-specific PAPs. The *x*-axes represent the genomic positions and the bars represent the number of 5′-ends of the oligo-cap cDNAs mapped at the corresponding genomic positions (TSSs). White bars show the tissue-specific usage of the corresponding PAPs observed in the indicated tissues.

striction, the transcriptional regulation could not have freely diverged. It was also intriguing that the G-protein coupled receptors (GPCRs) were infrequent in the PAP-containing genes. Many of the GPCR families are supposed to have been generated via multiple rounds of duplication of the coding exons (Lander et al. 2001; Fredriksson et al. 2003). At least immediately after such gene duplications, it can be assumed that the degree of freedom in the genes themselves should have been relatively high. In such circumstances, transcriptional diversification could have been realized by alteration of the independent locus as a whole for different purposes, and this might account for the lower frequency of the PAPs.

Variation in the first exon patterns contributed by the PAPs and its influence on the downstream protein-coding regions

We also analyzed the first exon sequences to investigate the direct downstream consequences of the PAPs. The patterns of the first exons were classified according to the criteria shown in Figure 5. Indeed, the patterns of the first exons were highly variable as well, partly reflecting the presence of the complex PAPs. In addition, several patterns of first exon variations were observed for a given PPR. In total, even if 199,599 "TSS variables," which were derived from differing TSSs, were excluded, 54,302 different patterns of the first exon were observed. This means that, on average, there were 7.7 first exon variations and among them 3.1 variants were responsible for separating the PAP groups. Considering that the overall frequency of the hitherto identified AS variants, whose coverage is mostly concentrated in the later exons, is about 3 variants per locus in total, our findings suggest that the first exons are even more variable than the later exons combined.

As shown in Figure 5A, "mutually exclusive" was the most abundant type of first exon. This pattern was responsible for the largest population of the PAPs according to the criteria employed in this study. Although it is supposed that some of the variations belonging to the other categories such as "TSS variable" and "staggered" are actually derived from PAPs that are separated by ≤500 bp, this observation lessened the concern that technical errors had led to misidentification of the PAPs, since it is relatively simple to confirm this pattern both empirically and computationally. It should also be noted that "mutually exclusive" variations were more abundantly observed in the first exons than in the later exons. Among hitherto registered AS variants, the population of "mutually exclusive" variations accounted for only several percent (Modrek and Lee 2002; Imanishi et al. 2004). The first exons should have been relatively free from functional requirements to maintain the consistency of the open reading frames or the proper folding of the main parts of the proteins. Because of these characteristics of the

Table 2. Relationship between putative alternative promoters and the GO categories

GO id	GO term	No. locus containing APs	*P*-value
Enriched			
GO:0005524	ATP binding	678	6e-14
GO:0006468	protein amino acid phosphorylation	325	3e-09
GO:0005856	cytoskeleton	164	2e-08
GO:0004674	protein serine/threonine kinase activity	228	1e-07
GO:0007242	intracellular signaling cascade	205	2e-07
Reduced			
GO:0003735	structural constituent of ribosome	37	4.7e-16
GO:0005576	extracellular region	113	3e-10
GO:0007186	G-protein coupled receptor protein signaling pathway	104	9e-06
GO:0005739	mitochondrion	215	1e-04

A

Exon pattern	Definition	Total		>500bp separated TSSs (PAPs)	
		#Pattern	Average frequency (pattern/cDNA)	#Pattern	Average frequency (pattern/cDNA)
	TSS variable	199,599	26.1	0	0
	Mutually exclusive	32,251	4.2	20,895	2.7
	Alternative donor	1,040	0.1	0	0
	Contained	2,275	0.3	0	0
	Staggered	6,449	0.8	2	0
	Single exon*	12,287	1.6	3,342	0.4

B

	#gene	#exon pattern	relative difference in amino acid length
inter AP	3282	15794	234
intra AP	373	996	63
total	5880	16790	228

Figure 5. Patterns of the first exons in the PAP-containing genes. (*A*) Distributions of the patterns of the first exons are shown. The number of identified exon patterns was counted in total (third column) or between the populations which are separated by >500 bp, thus accounting for the separation of the PAPs. *Either of the first exon variations was a "single exon" transcript. Different criteria were employed for them because these transcripts cannot be regarded as "splicing" variants. (*B*) Alterations of the amino acids resulting from the exon variations occurring in the population of "inter-APs" (TSS distance >500) or "intra-APs" (TSS distance ≤500) were counted.

terminal exons, the occurrence of the "mutually exclusive variations," in which the degree of the freedom is higher than that of the other exon patterns, could have been permitted very frequently during evolution.

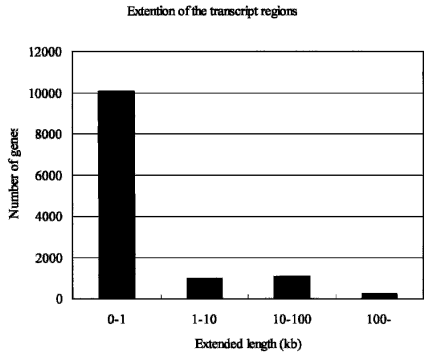
As shown in Figure 5B, the alternative use of the first exons following the PAPs caused alterations of the protein coding sequences (CDSs: the positions of the CDSs were defined according to the RefSeq information) in 3282 genes (43% of the 7674 PAP-containing genes), including 15,794 patterns in total. The average length of the alteration in amino acid length was 234 amino acids (a.a). Interestingly, even when no first exon variations changing the CDS were observed between different PAPs (inter PAPs), variations occurring within the same PAPs (intra PAPs) were observed to change the CDS in 373 genes (5%), including 996 patterns in total. Although their influence on the CDSs was relatively small (63 a.a. on average), it is possible that such slight changes also have biological consequences. A short stretch of amino acids could sufficiently mask an N-terminally located motif such as a secretion signal peptide.

Even alterations that did not change the CDS should not be overlooked. In these regions, various sequence motifs involved in modulation of the translational efficiency have been shown to be embedded (e.g., see UTRdb; Mignone et al. 2005). Actually not a few examples have been reported in which very small changes in the first exons play pivotal roles in controlling gene functions. In chicken proinsulin mRNA, a 32-base extension in the 5′-end sequence, possibly derived from fluctuation of the TSSs, was shown to play important biological roles due to harboring two so-called upstream ATGs that throttle the translation from the downstream authentic ATG. Further detailed manual and computational functional annotations of the PAP data obtained in the present study to assess gains and losses of protein motifs as well as the sequence motifs in the 5′UTRs would further enrich our knowledge about the functional diversification of the protein functions resulting from the alternative use of the first exons.

PAPs as an additional complex trait of the human gene transcriptome

In addition to the alterations in the transcriptional regulation and downstream coding sequences resulting from the use of APs in a single gene, the combinatory use of multiple APs between adjacent genes would enable further complex regulation of gene expression. Recent progress in genome research suggests that the transcriptome of human genes has greater than previously thought complexity (Kapranov et al. 2005), including transcripts generated from many so-called overlapping genes and anti-sense genes (Kiyosawa et al. 2003; Imanishi et al. 2004; Dahary et al. 2005). Newly identified additional PAPs (and thus first exons) have expanded the transcript regions (from TSSs to transcript termination sites), which had been delimited by RefSeqs, in many cases. Figure 6A shows the length distribution of the transcript regions extended towards the 5′-end by the present study. Occasionally, the extensions led to the identification of novel cases in which adjacent gene pairs are overlapping with each other either in the same or opposite direction (210 and 360 gene pairs, respectively; Fig. 6B; typical examples are shown in Supplemental Fig. 4). In those cases, it is possible that the alternative first exons are also involved in the regulation of the adjacent genes by interfering with the transcription from the genomic regions that overlap with each other. Even among PAPs belonging to the same gene, it is possible that some mutual influences are exerted. In the yeast SER3 gene, transcription from the upstream region of the canonical promoter is indispensable for the proper transcriptional control of the gene, masking the regulatory DNA elements of the downstream promoters, and thereby repressing the transcription (Martens et al. 2004). Similarly, for some PAPs in human genes, transcription from the distal (upstream) promo-

A



B

	same strand	opposite strand	total
RefSeq	362	458	820
RefSeq+DBTSS	210	360	570
total	572	818	1390

(gene pair)

Figure 6. Identification of putative overlapping and anti-sense gene pairs. (*A*) Length distribution of the RefSeq (LocusLink) regions extended by the additional DBTSS data. (*B*) Number of putative gene pairs identified using the indicated data set.

ters might be occurring and might consequently inhibit transcription initiation from the proximal (downstream) promoters. Although it still largely remains unclear whether there is mutual dependence of the transcriptional regulatory effects between multiple promoter units, the data set of the PAPs (and overlapping/antisense genes) produced in the present study will serve as an important basis for further exploration of the complex world of the transcriptome.

Conclusions

Here we described the genome-wide identification and characterization of the PAPs in human genes. In this study, based on analyses of the unprecedented amount of our newly generated oligo-cap cDNA data, we demonstrated that an unexpectedly large population of human genes is subject to regulation by more than a single promoter. The diversity of the transcriptional modulation and the consequent alterations of the N termini of the encoded proteins revealed that the PAPs have the potential to play major roles in generating the complexity of human gene expression. Indeed, PAPs were shown to have at least as much potential as AS in this regard.

This is the first report that describes what is actually occurring in the immediately proximal regions around the sites of transcription initiation within the cell and shows that these events increase the functional complexity of the genes. Although there was a pioneering study of potential PAPs using EST data by Landry et al. (2003), the 5′-end completeness of each PAP, and thus whether it represented an AP or AS of an internal exon, was not clearly discriminated in their survey. Besides, our analysis of the first exon patterns (Fig. 4) indicated that our newly and massively generated 5′-end data are qualitatively different from the previously available dbEST data, and therefore precise analysis of the PAPs was previously essentially impossible.

For a number of other mammals, such as mouse, rat, dog, and various kinds of primates, full-length cDNA data together with the genomic sequences have now become available or will become available in the near future. Recently developed high-throughput TSS detection technologies (Shiraki et al. 2003; Hashimoto et al. 2004) will further accelerate the accumulation of positional information of the TSSs. Detailed comparative studies of each of the PAPs, for example, studies of which PAPs and which of their parts are conserved during evolution and which have been diversified, would provide us with key information about the putative ancestral forms of transcriptional modulation as well as information about how such modulation has evolved in each organismal lineage (Xie et al. 2005). Indeed, it has long been supposed that the genetic basis for various anatomical and physiological similarities/differences between humans and other organisms lies in alterations in the expression of genes rather than changes in the functions of their encoded protein products (King and Wilson 1975; Tautz 2000; Boguski 2002). To address these issues, various types of data regarding biological activities and evolutional conservation of the promoters are actually being accumulated. In particular, it should be extremely fruitful to interpret the upcoming data from the ongoing ENCODE project (ENCODE Consortium 2004; http://www.genome.gov/10005107) in the context of the PAPs identified in the present study. This study gives a firm base for further exploring the molecular basis underlying the complexity of the transcriptional regulation of the human genome.

Methods

Mapping of the 5′-end sequences of the full-length cDNAs, filtration of the possibly dubious cDNAs, and classification of the putative TSSs into putative alternative promoter groups

We isolated 1,780,295 human cDNA sequences from full-length cDNA libraries constructed from 164 kinds of human tissues and cultured cells. The detailed library information is provided as Supplemental information. Each of the cDNA libraries was constructed using the oligo-capping method. Part of the sequence data was used for further selecting putative novel transcripts in the FLJ cDNA project (Ota et al. 2004). The computational mapping of the cDNA sequences onto the genomic sequence (hg_17; obtained from UCSC Genome Browser) was done by sequential use of BLAT and SIM4, following the procedure described in our previous reports (Suzuki et al. 2004). Positional information of the RefSeq genes is as of hgc_17. Those cDNA sequences that could not be mapped from their first bases were excluded from the data set for the precise identification of the TSSs. Furthermore, to minimize the frequency of erroneously oligo-capped truncated cDNAs, the 5′-ends of the cDNAs that mapped inside of the second or later exons of any other cDNAs or RefSeq were excluded. As described in the text, this filtration was intended for removing the dubiously identified truncated cDNAs from the data set. We presumed that those cDNAs whose 5′-ends were located outside of the exonic regions of any other cDNAs could not be truncated forms of any known types of transcripts at least (for schematic representation, see Fig. 1, bottom line). We would like to add the caveat that the PPRs that were genuinely located inside the exonic regions of other transcript variants had been missed in exchange by this procedure. To rescue them, we also proposed an extensive procedure as shown in Supplemental Figure 5. Ten-base margins were also allowed for each of the exon-intron boundaries for the judgment. This was done to diminish the influence of inaccurate mapping. The cDNA sequences overlapping with any of the RefSeq sequences belonging to a given Locus ID were considered to correspond to that Locus ID (as of February 10, 2005; obtained from the NCBI download site). The locus-grouped sequences were further separated into several TSS clusters by binning the TSSs by 500-bp intervals. All of the raw data together with the obtained results are presented in our DBTSS data set (the Web site will be open to the public on acceptance of this manuscript).

Characterization of the promoter motifs

The presence of CpG islands was determined using NEWCPGREPORT contained in the EMBOSS package 2.8.0 (Rice et al. 2000). The CpG islands covering the TSSs were counted as "CpG island-containing" promoters. For the search of TATA boxes, MATCH was run for TRANSFAC database version 8.2 (Wingender 2004; http://www.gene-regulation.com/) using the matrices V$TATA_01 and V$TATA_C with the search conditions of: plus strand; −43 to −27; cut-off values of minFN74.prf. The obtained hits were counted as "TATA-containing" promoters.

Categorization of the first exons

For every combination of the first exons belonging to the same PAP group, the exon patterns were classified into the categories illustrated in Figure 4. Concerning the 3′-end boundaries of the first exons and the 5′-end boundaries of the second exons, only the patterns with a mutual difference >10 bp were counted. For counting the total number of "TSS variables," which is the total number of the 5′-end boundaries of the first exons (TSSs) in this population, such an allowance was not made since the precise

mappings were assured at the initial mapping process. Then, to identify the first exon patterns that were responsible for separating PAP groups, the patterns whose TSSs were separated by >500 bp were counted.

Calculation of the statistical significance

Statistical significance of the tissue-specific usage of the PAPs and enrichment of the GO terms was evaluated by calculating the hyper-geometric distribution using the following equation.

$$\sum_{x=i}^{Np} \frac{\binom{Np}{x} \bullet \binom{N(1-P)}{n-x}}{\binom{N}{n}}$$

where N = total number of cDNAs belonging to the locus, P = expected frequency of the cDNAs associated with a given tissue or a given GO term within the locus (number of associated cDNAs/N), n = total number of cDNAs belonging to the PAP of concern, and i = number of the cDNAs associated with a given tissue or a given GO term within the AP.

For the tissue classification of the cDNA libraries, see Supplemental Table 1. GO terms were related to the LocusLink genes according to "loc2go" obtained from LocusLink.

Acknowledgments

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.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Black, D.L. 2000. Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* **103:** 367–370.

Boguski, M.S. 2002. Comparative genomics: The mouse that roared. *Nature* **420:** 515–516.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Dahary, D., Elroy-Stein, O., and Sorek, R. 2005. Naturally occurring antisense: Transcriptional leakage or real overlap? *Genome Res.* **15:** 364–368.

ENCODE Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

Ewing, B. and Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.* **25:** 232–234.

Fredriksson, R., Lagerstrom, M.C., Lundin, L.G., and Schioth, H.B. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63:** 1256–1272.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32:** D258–D261.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., and Matsushima, K. 2004. 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22:** 1146–1149.

Hsu, F., Pringle, T.H., Kuhn, R.M., Karolchik, D., Diekhans, M., Haussler, D., and Kent, W.J. 2005. The UCSC Proteome Browser. *Nucleic Acids Res.* **33:** D454–D458.

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.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Kalejs, M. and Erenpreisa, J. 2005. Cancer/testis antigens and gametogenesis: A review and "brain-storming" session. *Cancer Cell Int.* **5:** 4.

Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.***15:** 987–997.

King, M.C. and Wilson, A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., and Hayashizaki, Y. 2003. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13:** 1324–1334.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, 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.

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.

Lee, C., Atanelov, L., Modrek, B., and Xing, Y. 2003. ASAP: The Alternative Splicing Annotation Project. *Nucleic Acids Res.* **31:** 101–105.

Lopez, A.J. 1998. Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32:** 279–305.

Luzi, L., Confalonieri, S., Di Fiore, P.P., and Pelicci, P.G. 2000. Evolution of Shc functions from nematode to human. *Curr. Opin. Genet. Dev.* **10:** 668–674.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. 2005. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **33:** D54–D58.

Maio, M., Coral, S., Fratta, E., Altomonte, M., and Sigalotti, L. 2003. Epigenetic targets for immune intervention in human malignancies. *Oncogene* **22:** 6484–6488.

Martens, J.A., Laprade, L., and Winston, F. 2004. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429:** 571–574.

Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C., and Pesole, G. 2005. UTRdb and UTRsite: A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **33:** D141–D146.

Modrek, B. and Lee, C. 2002. A genomic view of alternative splicing. *Nat. Genet.* **30:** 13–19.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420:** 563–573.

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. *Nat. Genet.* **36:** 40–45.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33:** D501–D504.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16:** 276–277.

Schmid, C.D., Praz, V., Delorenzi, M., Périer, R., and Bucher, P. 2004. The Eukaryotic Promoter Database EPD: The impact of in silico primer extension. *Nucleic Acids Res.* **32:** D82–D85.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100:** 15776–15781.

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.

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: 149–156.
Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., et al. 2001a. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep.* **2:** 388–393.
Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T. Tanaka, T., Nakamura, Y., et al. 2001b. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.* **11:** 677–684.
Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K. 2004. DBTSS, DataBase of Transcriptional Start Sites: Progress report 2004. *Nucleic Acids Res.* **32:** D78–D81.
Tautz, D. 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10:** 575–579.
Ventura, A., Luzi, L., Pacini, S., Baldari, C.T., and Pelicci, P.G. 2002. The p66Shc longevity gene is silenced through epigenetic modifications of an alternative promoter. *J. Biol. Chem.* **277:** 22370–22376.
Wingender, E. 2004. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.* **4:** 55–61.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. 2005. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350:** 129–136.
Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D.A., Hayashizaki, Y., and Gaasterland, T. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* **13:** 1290–1300.

Web site references

http://www.bioinformatics.ucla.edu/ASAP/; ASAP.
http://dbtss.hgc.jp/; DBTSS.
http://www.genome.gov/10005107/; ENCODE.
http://www.epd.isb-sib.ch/; EPD.
http://www.geneontology.org/; GO.
http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene; LocusLink (Entrez Gene).
http://www.ncbi.nlm.nih.gov/RefSeq/; RefSeq.
http://www.gene-regulation.com/; TRANSFAC.
http://genome.ucsc.edu/; UCSC Genome Browser.
http://bighost.area.ba.cnr.it/srs6/; UTRdb.

Received April 15, 2005; accepted in revised form September 19, 2005.