6386–6395 *Nucleic Acids Research, 2008, Vol. 36, No. 20* *Published online 5 October 2008*
doi:10.1093/nar/gkn677

Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs

Jun-ichi Takeda[1,2], Yutaka Suzuki[2], Ryuichi Sakate[1], Yoshiharu Sato[1], Masahide Seki[2], Takuma Irie[2], Nono Takeuchi[2], Takuya Ueda[2], Mitsuteru Nakao[3], Sumio Sugano[2], Takashi Gojobori[1,4] and Tadashi Imanishi[1,*]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo 135-0064, [2]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, [3]Laboratory for Plant Genome Informatics, Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818 and [4]Center for Information Biology and DDBJ, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Received July 8, 2008; Revised September 12, 2008; Accepted September 23, 2008

ABSTRACT

Using full-length cDNA sequences, we compared alternative splicing (AS) in humans and mice. The alignment of the human and mouse genomes showed that 86% of 199 426 total exons in human AS variants were conserved in the mouse genome. Of the 20 392 total human AS variants, however, 59% consisted of all conserved exons. Comparing AS patterns between human and mouse transcripts revealed that only 431 transcripts from 189 loci were perfectly conserved AS variants. To exclude the possibility that the full-length human cDNAs used in the present study, especially those with retained introns, were cloning artefacts or prematurely spliced transcripts, we experimentally validated 34 such cases. Our results indicate that even retained-intron type transcripts are typically expressed in a highly controlled manner and interact with translating ribosomes. We found non-conserved AS exons to be predominantly outside the coding sequences (CDSs). This suggests that non-conserved exons in the CDSs of transcripts cause functional constraint. These findings should enhance our understanding of the relationship between AS and species specificity of human genes.

INTRODUCTION

Alternative splicing (AS), the recombination of exons to produce novel transcripts, is thought to contribute substantially to the functional complexity of human proteins. AS frequently produces diverse transcripts from a single genetic locus. The consequent exon changes alter the corresponding amino acid sequences and generate functionally divergent proteins. The ENCODE and GENCODE projects have found that about 60% of all loci encode AS isoforms, with an average of more than 5.4 isoforms per locus (1–3). In the ENCODE project, a product consisting of two exons of the *DONSON* gene and three exons of the *ATP50* gene was cloned and sequenced by real-time reverse transcriptase polymerase chain reaction (RT–PCR) to generate several novel connected transcripts. Because many transcripts are derived from non-conserved sequences, however, the investigators in those projects emphasized the need for further studies exploring the neutrality of genome evolution. Several other reports have suggested that frequently occurring AS isoforms are not evolutionarily conserved and thus may be attributable to transcriptional noise or cloning artefacts.

Analyses of AS have relied on partially sequenced cDNA information (i.e. on expressed sequence tags, or ESTs). Although EST-based genome-wide approaches have identified thousands of instances of AS in human genetic transcripts (4,5), they do not provide information about the positions and combinations of AS exons in

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

full-length transcripts. This information is needed for detailed analysis of the impact of AS on protein function and evolutionary turnover. Furthermore, the standard methods for estimating selective pressure by calculating and comparing the rates of synonymous and non-synonymous substitutions are not easily applied to partially sequenced cDNA.

In the work reported here, we collected and analysed instances of AS in the context of full-length cDNAs, which are ideal resources with which to analyse AS because information regarding the complete form of a particular transcription unit (i.e. an isoform) allows us to determine the relevance of AS to conserved protein functions. The functional domains of proteins are often embedded over a wide region of the protein sequences, so it is possible that not every combination of AS exons is allowed. Full-length cDNA-based approaches also facilitate better coverage at the 5′-ends of AS sequences than do EST-based approaches. Our data are based on a large collection of physical cDNA clones whose complete sequences have been determined and therefore can thus be used to directly validate the functions of certain genes. In a previous study, we analyzed 56 419 full-length human cDNAs whose sequences were checked by expert scientists and by using computational methods specific to the full-length cDNA annotation conferences H-Invitational (H-Inv) (6) and H-Inv 2 (7). From that dataset, we created a catalogue of 18 297 AS variants at 6877 loci (8).

Our AS catalogue is not the perfect tool with which to understand the functional diversity of human genes. Information about evolutionary conservation must be added to our functional annotations if one is to use our catalogue to ascertain which of the AS sequences in specific parts of proteins should be prioritized in future functional analyses. Many studies have demonstrated that a significant proportion of AS sequences are not conserved, and those sequences can be assumed to have species-specific biological roles. Also, it is suggested that not a few genes seem to exist in a species-specific manner. We therefore compared human AS variants with AS variants in the mouse genome. We selected 431 conserved AS variants at 189 loci for which full-length human and mouse cDNAs were available. Interestingly, a significant number of the AS variants that were not directly supported by full-length mouse cDNAs, which contained non-conserved exons, were translated to proteins. Here, we compare the evolutionary conservation of AS transcripts in humans and mice by using full-length cDNAs.

MATERIALS AND METHODS

Full-length cDNA sequences from humans and mice

We obtained 64 034 full-length human cDNAs sequenced in four projects—Human Unidentified Gene-Encoded Large Proteins (HUGE), Full-Length cDNA Japan (FLJ), Munich Information Centre for Protein Sequences (MIPS) and Mammalian Gene Collection (MGC)—involving six institutions: Kazusa DNA Research Institute, Tokyo University, Helix Research Institute, German Cancer Research Centre, the United States National Institutes of Health and the Chinese National Human Genome Centre. These institutions provided full-length human cDNA clones that had been annotated at the H-Inv and H-Inv 2 conferences (Supplementary Table 1A). We also obtained 118 775 full-length mouse cDNAs that had been sequenced in the Functional Annotation of Mouse (FANTOM) 3 and MGC projects (Supplementary Table 1B). All sequences are registered in release 66 of the DNA Data Bank of Japan (DDBJ) and are freely available at http://www.ddbj.nig.ac.jp/.

Identification of human alternatively spliced variants

Human AS variants were identified as previously described (8). To remove potential 5′- or 3′-truncated cDNAs, we excluded sequences with 5′- or 3′-ends in the second or downstream exons of other sequences. We included cDNAs with 5′- or 3′-ends in the first or last exon and which were therefore considered to vary in their transcriptional initiation or termination sites. We assumed that cDNAs with 5′-ends outside the exonic regions of other sequences were not truncated forms of known transcripts. For a detailed discussion of this topic, see Kimura *et al.* (9). The resulting set of putative full-length cDNAs was used to compare the genomic positions of each exon–intron boundary with that of other transcripts from the same locus, with an allowance of 10 bp to remove potential sequencing or mapping errors (Supplementary Table 2). If part of the cDNA exonic sequence was in the first or last exon of another cDNA intronic region, that sequence was considered a '5′/3′-end' AS variant. If part of the internal cDNA exonic sequence was in the confirmed intronic region of another cDNA, then that sequence was considered an 'internal' AS variant. We then removed annotated genomic 'rearrangement' genes, such as those encoding immunoglobulin (Ig) and the T-cell receptor (TCR) and anomalously highly polymorphic genes, such as those of the major histocompatibility complex (MHC). If a group of AS variants contained two or more variants with the same genomic structure, the one of median length was selected as a representative of the group. If the number of variants in the group is even, then the longer of the two near-median-length variants was selected as a representative of the group.

Functional annotation of human AS isoforms

We identified AS variants containing full-length open reading frames (ORFs), which means the start codon (ATG) encoded methionine and the final codon was TAA, TAG or TGA. Full-length ORFs were defined as coding sequences (CDSs). Of the 20 392 representative human AS variants, 16 103 transcripts (i.e. 14 597 AS variants) were determined to be isoforms with CDSs. Four types of protein functions in these transcripts were annotated: protein motifs, gene ontology (GO) terms, sub-cellular localization signals and transmembrane domains. Protein motifs and GO terms were identified with InterProScan (10); sub-cellular localization signals were predicted with WoLF PSORT (11) and TargetP (12) and transmembrane domains were predicted with TMHMM

(13) and SOSUI (14) software. These annotation analyses of protein function were automatically executed by TACT (15), an integrated annotation tool for genome and transcriptome analyses. Further details regarding the functional annotation pipeline are available from H-InvDB (http://www.h-invitational.jp/) (16). The results of the computational identification and annotation analyses were visually inspected by members of the AS annotation team of H-Inv. Controversial annotations were flagged to identify possible errors.

Detection of retrotransposons and exonic splicing enhancers in human AS variants

RepeatMasker [A. F. A. Smit, R. Hubley and P. Green, RepeatMasker Open-3.0. 1996 to 2004 http://www.repeatmasker.org)] was used to detect retrotransposons [long terminal repeats (LTRs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs)] in the DNA sequences of human AS variants. Exonic splicing enhancers (ESEs) were detected in the DNA sequences of human AS variants by using 238 candidate hexamer sequences obtained from the RESCUE-ESE Web Server (17).

Alignment of the human and mouse genomes

Human and mouse genome sequences (versions hg18 and mm8) were downloaded from UCSC (http://genome.ucsc.edu/) and a modified pairwise genome alignment was constructed using BLASTZ (18) version 7, with parameter $C = 2$. Regions of redundant alignment were removed using reciprocal best hits. This method was originally used to construct two satellite databases of H-InvDB: the comparative genomics database G-compass (19) and the orthologue database Evola (20).

Genomic comparison of human AS variants and the mouse genome

The human–mouse genome alignment was used to compare the position of each human AS variant exon with that of the corresponding exon in mice. If the mapped coverage and identity of the overlapping regions surrounding the full-length or CDS in the exon exceeded threshold values (70% for coverage and 60% for identity), the region was considered 'genome-conserved'. If the full-length or CDS in the exon corresponded to a mouse counterpart transcript with coverage and identity equal to or greater than the threshold values, this region was considered 'transcript-conserved'. If the exon did not map to the genome alignment, or if the mapped coverage or identity was less than the threshold value, the exon was considered 'non-conserved' (Figure 1A).

The exon information was used to determine the conservation category for each transcript. If at least one non-conserved exon was identified in the transcript, the transcript was considered 'non-conserved'. If the transcript had no non-conserved exons but contained genome-conserved exons, it was considered 'genome-conserved'. If the transcript had only transcript-conserved exons, it was considered 'transcript-conserved' (Figure 1B). If all the exons in the human and mouse

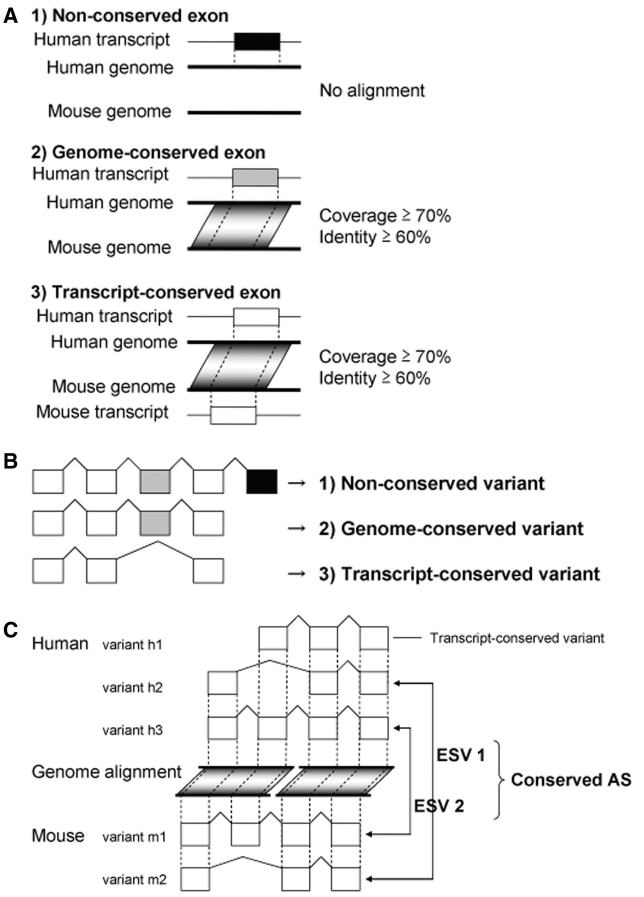


Figure 1. Comparative analysis of human and mouse AS sequences. (**A**) Categories of conserved exons. Full-length exons and coding regions of exons were included in the analysis, and the highest conservation level was selected. Boxes indicate exons; thin straight lines indicate introns. (**B**) Categories of conserved AS variants within the categorized exons in (A). Conservation levels were determined from the lowest conservation level of each transcript's exons. (**C**) Equally spliced variants (ESVs) and conserved AS sequences represent higher categories of transcript-conserved variants. ESVs are transcript-conserved variants with similar combinations of exons in different species. Conserved AS sequences are two or more different ESV pairs at a particular locus in multiple species. Additional details are available in the Results section and the Materials and methods section.

transcripts were transcript-conserved and the exon combination was also conserved, the corresponding transcripts were defined as equally spliced variants (ESVs). Finally, if different ESVs were identified at a particular human and mouse locus, the transcripts were considered 'conserved AS' (Figure 1C).

Experimental validation of 'retained-intron' AS sequences

Polysomal fractions were prepared, using $\sim 3 \times 10^7$ cells, as described elsewhere (21). Cell pellets were suspended in 1 ml of lysis buffer [20 mM Tris–HCl (pH 7.5), 10 mM NaCl, 3 mM $MgCl_2$, 0.04 M sucrose, 0.5% NP40 and 1 mM dithiothreitol] containing 100 units of RNase inhibitor. The pellets were then lysed by incubation on ice for 10 min. The nuclei and cell debris were removed by centrifugation at 1000*g* for 10 min at 4°C. The lysate was layered on top of 11 ml of a 15%/50% (w/v) sucrose gradient and centrifuged at 36 000*g* in a Beckman SW41Ti

rotor for 135 min at 4°C. A density-gradient fractionator (Towa Labo, Japan, Model 152–001) was used to separate the gradient into 11 equal fractions. Absorbance was monitored at 260 nm. Each fraction was treated with proteinase K. RNA was extracted using phenol and $CHCl_3$, precipitated with ethanol and analysed for each mRNA species.

First-strand cDNA synthesis was performed using a 17-bp dT primer and SuperScript II as directed by the manufacturer. The resulting cDNAs were quantified and used in quantitative RT–PCR analysis. The results were normalized to the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) PCR product. PCR-cycling parameters were 50°C for 2 min, 95°C for 10 min, followed by 35 cycles of 95°C for 30 s, 57°C for 1 min and 72°C for 1 min. The ABI PRISM HT7000 Sequence Detection System (Applied Biosystems) was used to detect the products. The primers used in the RT–PCR are described in Supplementary Table 3A. The sizes and integrity (the degree of smeared amplicons and primer dimers) of the final PCR products were confirmed by agarose gel electrophoresis.

RESULTS

Generation and characterization of evolutionarily conserved AS sequences

Evolutionarily conserved AS sequences were selected from the H-Inv and H-Inv 2 datasets. We examined the evolutionary conservation of exons in 20 392 AS variants of 64 034 human full-length cDNAs. The exons were defined as 'conserved', when the human genomic sequence to which the full-length or CDS regions of the exons were mapped could be aligned with the mouse genome with a coverage of >70% and an identity of >60%. Human–mouse alignments were performed using information available in the UCSC Genome Browser (versions hg18 and mm8). Of the 199 426 total exons, 171 547 (86%) were categorized as conserved (Table 1, panel A). Moreover, the exons in 12 096 transcripts of the 20 392 AS sequence variants were categorized as genome- or transcript-conserved (Table 1, panel B). On the other hand, in 944 human transcripts, all of the exons were not conserved. Therefore, the locus itself seemed not conserved. Although we did not analyze such putatively human-specific loci separately, they should be studied in greater detail (also see Supplementary Table 4).

The human transcripts were further examined to determine whether corresponding transcripts existed in mice. For this purpose, 118 775 full-length mouse cDNAs were collected and annotated during the FANTOM 3 and MGC projects to determine if any exons within the full-length mouse cDNAs mapped to corresponding genomic regions in humans. The conserved exons described above were separated into transcript-conserved and genome-conserved categories, if the coverage and identity of the mouse counterpart exon was more than the threshold value (Figure 1A). A transcript consisting of all transcript-conserved exons was deemed to be transcript-conserved (Figure 1B). The ESVs, in which all of the exons and their combinations were conserved, and the conserved AS sequences, in which both of the AS variants were ESVs, were more highly conserved categories of transcript-conserved variants (Figure 1C). These categories are explained in more detail in the Materials and methods section.

We then identified 4624 ESVs from 3570 loci and determined that only 431 AS variants from 189 loci were conserved (Table 1, panel C). Because these AS variants will undoubtedly prove interesting in future investigations, we confirmed their CDSs and looked for alternative splicing-mediated changes in the corresponding protein sequences and for the presence of protein motifs, GO terms, subcellular localization signals and transmembrane domains within the affected regions.

Examples of evolutionarily conserved AS sequences

An example of a conserved AS sequence is shown in Figure 2A. At the locus encoding the phosphoinositide-3-kinase (PI3-kinase) regulatory subunit, the 5′-exons of BC094795 are AS relative to another variant, BC030815. These human-transcript variants correspond to the mouse-transcript variants BC026146 and BC051106, respectively. All of the exons as well as their combinations and AS patterns are conserved in each of the human–mouse transcript pairs. Three AS variants (p85-α, p55-α and p50-α) were originally identified at the human

Table 1. Genomic conservation of AS in humans and mice

	Total	Non-conserved	Genome-conserved	Transcript-conserved
Panel A: exon-level conservation of human AS variants, compared with the mouse genome				
All exons	199 426	27 879 (14%)	23 412 (12%)	148 135 (74%)
AS exons	49 842	12 196 (25%)	9064 (18%)	28 582 (57%)
	Total	Non-conserved	Genome-conserved	Transcript-conserved
Panel B: transcript-level conservation of human AS variants, compared with the mouse genome and transcripts				
AS variants	20 392	8296 (41%)	4410 (21%)	7686 (38%)
AS loci	7601	1716 (23%)	1241 (16%)	4644 (61%)
	Transcript-conserved total	Non-equally spliced variant	ESV	Conserved AS
Panel C: higher conservation categories for transcript-conserved AS sequences				
AS variants	7686	2631 (34%)	4624 (60%)	431 (6%)
AS loci	4644	885 (19%)	3570 (77%)	189 (4%)

PI3-kinase locus in humans (22), with BC094795 corresponding to p85-α and BC030815 corresponding to p55-α. The functional domains specific to Rho GTPase-activating protein (RhoGAP) (IPR008936) and Src homology-3 (SH3) (IPR001452) are embedded in the N-terminal region of p85-α, which extends from the N-terminal region of p55-α. The p85-α and p55-α variants appear to have diverged functionally to the extent that they relay signals from the insulin receptor substrate (IRS) proteins to PI3-kinase with different efficiencies (23). Because it seems natural that such a fundamental diversification would be evolutionarily conserved in both mice and humans, we were not surprised to identify these AS variants in mice. Indeed, they were categorized as 'conserved' in our study.

Figure 2B shows an example of a locus at which we identified previously unknown functional protein changes caused by AS. The protein products of this locus have been annotated as cysteinyl-tRNA synthetase (CARS). A glutathione S-transferase (GST) C-terminal-like (IPR010987) protein motif was identified in the BX647906 AS variant, but was absent from the BC002880 variant because of the presence of a cassette-type AS pattern in the second exon of BX647906. The C-terminal region of GST plays an essential role in substrate activation, and this motif occurs not only in GST but also in many types of aminoacyl-tRNA synthetases (aaRSs) (24). Some types of mammalian aaRSs interact with translational elongation factor (EF)-1 via this GST C-terminal domain, and this interaction is thought to facilitate the vectorial transfer of aminoacylated tRNAs. Thus, the functional roles of the newly identified AS variants BX647906 and BC002880 may involve the delivery of the cognate tRNA from CARS to the EF complex, thereby controlling translation efficiency in the cell under specific circumstances (25).

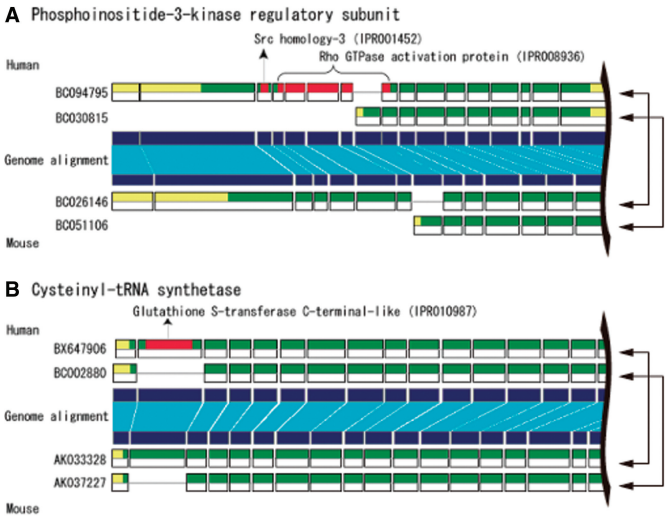


Figure 2. Conserved AS sequences, exemplified by (**A**) PI3-kinase regulatory subunit and (**B**) CARS. Constitutive introns are shortened here and additional details are available in the Results section. Boxes indicate exons. Filled regions within boxes indicate CDSs (green), protein motifs (red) and untranslated regions (yellow). The ESVs shared by humans and mice are indicated by arrows.

Statistical analysis of evolutionarily conserved AS sequences

Information regarding individual AS sequences and their annotations is freely available from our database H-DBAS (http://h-invitational.jp/h-dbas/) (26). We investigated differences between the conserved AS sequences and the total AS sequences and determined that the cassette-type pattern is more common in conserved AS sequences and that the retained intron-type pattern is less common in conserved AS sequences (Table 2, panel A). Our statistical analysis showed that the average length of the CDSs of conserved AS sequences was 52 amino acids, while that of the total AS variants was 87 amino acids. The frequency of AS-mediated changes in protein function also differed between the conserved AS variants (66%) and the total AS variants (45%) (Table 2, panel B). Because of the increased occurrence of the cassette-type pattern, conserved AS variants caused more radical changes than did the general AS population.

As shown in Table 3, we investigated the difference between the GO terms of the conserved AS loci and those of the total AS loci. For this analysis, we examined protein motifs embedded in the CDSs and considered the GO terms associated with those motifs. We found that the GO terms 'DNA binding (GO:0003677)', 'Peroxidase activity (GO:0004601)' and 'Response to oxidative stress (GO:0006979)' were significantly ($P < 0.01$) over-represented in conserved AS loci (Table 3, panel A). The genes having these functions seem to need evolutionarily invariant AS sequences to play basic roles in keeping cellular homeostasis. The 'nucleotide-binding', 'protein kinase' and 'WD40 protein' motifs are the most commonly affected motifs in conserved AS sequences, although these motifs are also relatively common in the general AS population. In contrast, the transforming growth factor (TGF)-β-stimulated clone-22 (TSC-22) Dip Bun (IPR000580) motif is significantly ($P < 0.01$) more common in the conserved AS loci than in the general AS loci (Table 3, panel B). This motif, which has homologues in mice and *Drosophila*, acts as a transcription factor and plays a fundamental role in cell differentiation (27). Another motif, the basic leucine zipper (bZIP)

Table 2. Comparison of AS features in total and conserved AS loci

	Total AS loci	Conserved AS loci
Panel A: AS patterns		
Total	7601	189
Cassette (Skipped exon)	3584 (35%)	66 (42%)
Internal acceptor (Alternative 3′ splice)	1988 (19%)	30 (19%)
Internal donor (Alternative 5′ splice)	1990 (20%)	33 (21%)
Mutually exclusive	237 (2%)	4 (2%)
Retained intron	2477 (24%)	26 (16%)
Panel B: Effects of AS on protein function		
Total	7601	189
Average difference in CDS length	87 aa	52 aa
Total effects of AS on function	3395 (45%)	125 (66%)
Protein-motif alterations	2423	86
GO alterations	1078	30
Sub-cellular localization changes	2305	75
Transmembrane domain changes	444	16

Table 3. GO terms and protein motifs frequently observed at conserved AS loci

GO ID	Definition	Number of conserved AS loci	Total number of AS loci	*P*
Panel A: GO terms				
GO:0003677	DNA binding	17	333	0.0085*
GO:0004601	Peroxidase activity	3	14	0.0077*
GO:0006979	Response to oxidative stress	3	14	0.0077*
InterPro ID	Definition	Number of conserved AS loci	Total number of AS loci	*P*
Panel B: protein motifs				
IPR004827	Basic-leucine zipper transcription factor	3	4	0.0005*
IPR000580	TSC-22/Dip/Bun	2	3	0.0057*

P-values were calculated using Fisher's exact test. They indicate the significance of the difference between the ratios to the conserved AS loci (189) and the total AS loci (7601). $^{*}P < 0.01$.

(IPR004827) motif, is also significantly more common in the conserved AS loci (Table 3, panel B). In eukaryotes, the bZIP transcription factor is responsible for initiating cellular responses to ultraviolet (UV) damage and osmotic stress (28). Because the switch of these functional motifs is essential to basic cellular functions, the corresponding AS sequences should also be conserved between species.

These results suggest that the functions of AS sequences differ between conserved AS variants, which constitute the core dataset of the AS sequences and have conserved diversification of gene functions, and other AS variants. In other words, a distinct subset of protein motifs appears to be responsible for the functional diversity of conserved and species-specific gene functions. Classification of the AS variants is necessary to further address this issue.

Experimental validation of retained introns

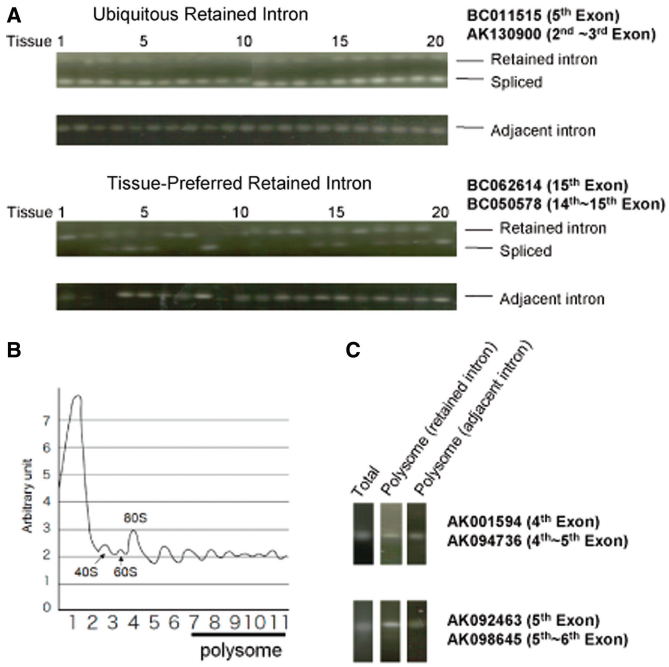
After selecting and manually annotating the AS variants, we sought to understand why there was such a large population of non-conserved variants. One possibility is that these non-conserved variants are artefacts derived from aberrant cDNA cloning (e.g. cDNAs produced from incompletely spliced mRNAs or genomic DNA contaminants). Sequence information is insufficient to distinguish truly functional retained-intron type AS sequences (which are less common among conserved AS sequences) from the rest of the population.

We selected 14 retained-intron AS variant pairs from the transcript-conserved population, 15 from the genome-conserved population, and 5 from the non-conserved population, which were manually annotated as 'highly likely protein coding' sequences. In the transcript-conserved population all of the exons were conserved at the transcriptional level (i.e. there were corresponding mouse cDNAs for both variants). In the genome-conserved population, there were no corresponding mouse transcripts but the corresponding genomic sequences for all the related exons were conserved. In the non-conserved population, at least one of the related exons was not conserved. We determined the expression patterns of the selected variants in 20 types of human tissues by using semi-quantitative real-time RT–PCR. As a control, we performed the same series of experiments on adjacent exons separated by moderately long introns. We observed no evidence of immature splicing in any of the controls. PCR products corresponding to the retained introns were observed in at least one tissue for 14 of the 14 transcript-conserved variants, for three of the 15 genome-conserved variants and for three of the five non-conserved variants. A wide variety of expression patterns was observed. For example, we observed ubiquitous expression of both type of some AS variants (Figure 3A, upper panel), whereas we observed the mutually exclusive expression of other AS variants in a tissue-preferred manner (Figure 3A, lower panel). These results suggest that a significant population of cDNAs, even those resulting from retained-intron AS variants, were derived from real transcripts. The overall expression patterns and mutual dependence of the AS variants appear to be controlled.

We further investigated whether transcripts containing retained-intron AS sequences were translated into proteins. We did this by recovering RNAs from actively translating ribosomal fractions (i.e. polysomal fractions) in a human promyelocytic leukaemia cell line, HL60, with sucrose density-gradient purification (Figure 3B). RNAs purified from fractions 7–10 were analysed with real-time RT–PCR. RT–PCR of total RNA revealed that 14 of the transcript-conserved retained introns that were expressed in at least one tissue were expressed in HL60 cells. For nine of these, clear bands of the appropriate size were produced by RT–PCR of the polysome fractions (Figure 3C and D and Supplementary Table 3B). Interestingly, even among non-conserved AS sequences, we observed evidence of the translation of retained-intron AS sequences. Although further analysis is necessary, we have shown that even retained-intron AS variants, which were previously considered dubious cDNAs, are often expressed as proteins. Future investigators should therefore annotate these full-length cDNAs rather than automatically discard them.

Characterization of non-conserved exons

To further characterize the non-conserved exons, we first compared the characteristic features of the non-conserved



D

	transcript-conserved	genome-conserved	non-conserved
primer set	14	15	5
20 tissues	14	3	3
HL60 total RNA	14	3	3
HL60 polysome	9	0	1

Figure 3. Experimental validation of AS human transcripts by using the retained-intron pattern. (**A**) RNA expression of retained-intron AS sequences in 20 human tissues (1, adrenal gland; 2, bone marrow; 3, brain, cerebellum; 4, brain, whole; 5, fetal brain; 6, fetal liver; 7, heart; 8, kidney; 9, liver; 10, lung, whole; 11, placenta; 12, prostate; 13, salivary gland; 14, skeletal muscle; 15, testis; 16, thymus; 17, thyroid gland; 18, trachea; 19, uterus and 20, spinal cord). The upper panel exemplifies 'ubiquitous' retained-intron AS sequences and the lower panel exemplifies 'tissue-preferred' retained-intron AS sequences. (**B**) RT–PCR analysis using polysomal fractions isolated from the human promyelocytic leukaemia cell line HL60. (**C**) RNA expression of retained-intron AS sequences mixed with translating ribosome fractions (i.e. polysome fractions) from the HL60 cell line. (**D**) Number of expressed transcripts in each conserved category.

exons with those of the conserved exons. Our conserved exon data set consisted of the genome level and above, and our results showed that the majority of non-conserved exons are located at the 5′-end of the transcript and appear to function as AS exons (Supplementary Table 5). Consistent with our findings, recent studies have found that alternative 5′-exons derived from alternative promoters are abundant in humans and are often not evolutionarily conserved (29). In contrast, the coding regions of non-conserved exons overlap less frequently with AS exons than with constitutively spliced (CS) exons (Table 4). It is likely that previous studies reported overall higher frequencies for non-conserved exons among AS exons because the number of non-conserved exons is greatest at the 5′-ends of non-coding exons.

We also observed that non-conserved exons were excluded from protein-motif regions (Table 4). Even in cases where non-conserved AS exons overlap with protein motifs, the categories of overlapping motifs differ strikingly from those motifs overlapping conserved AS sequences. The most frequently observed protein motif in non-conserved AS exons was the KRAB box (IPR001909) (Supplementary Table 6) (30). This observation is consistent with the results of a previous study reporting that this motif, which is a well-known protein interaction domain, was significantly enriched in the sites of alternative splicing (31). The KRAB-type zinc finger is thought to control the transcription of the sex-determining region Y (*SRY*/*Sry*) gene and is therefore essential for the determination and differentiation of the testis (32). The fact that KRAB motifs tend to overlap with non-conserved AS exons may underscore the evolutionary uniqueness of the human transcriptional regulatory network in the reproductive system. A few protein motifs, including the GAGE motif (IPR008625), nuclear-pore-complex-interacting motif (IPR009443), phospholipase A2 active site motif (IPR013090) and the T-complex 11 motif (IPR008862), were also identified only in non-conserved AS exons (Supplementary Table 6). The GAGE motif is specific to humans and the nuclear-pore-complex-interacting motif is specific to primates (33,34). We further examined the expression patterns of these genes having GAGE, nuclear-pore-complex-interacting, phospholipase A2, active site and T-complex 11 motifs by using our expression profiling database H-ANGEL (Human Anatomical Gene Expression Library) (35). We found that the genes containing the GAGE and T-complex motifs were preferentially expressed in testis, that the expression pattern of genes containing the nuclear-pore-complex-interacting motif was ubiquitous and that there was no data on the expression pattern of genes containing the phospholipase-A2-active-site motif (Figure 4). Although the functions of GAGE-motif-containing genes are potentially very interesting, all that is known about that them is that they are primate specific. We were unable to investigate the AS of the GAGE motifs in this study because their parent genes are not found in mice (also see Supplementary Table 6). The parent genes for the T-complex-11 motifs, in contrast, are found in both humans and mice. Furthermore, human and mouse *TCP11*, which encodes the receptor for fertilization-promoting peptide (FPP), is reported to play an important role in sperm function (36). In any case, it is interesting that the GAGE and T-complex-11 motifs, which are enriched in the non-conserved AS exons, were associated with a species-specific factor involved in the reproductive system.

Our results also show that non-conserved exons are more frequently related to retrotransposons (Table 4). This observation is not surprising because human retrotransposons are primate specific (*Alu*) and mouse retrotransposons are rodent specific (37). Exonic-splicing enhancers (ESEs) are ubiquitously scattered in human AS variants (Table 4), so we could not observe the relation between ESE conservation and splicing (38). We found no clear differences between the non-conserved and conserved

Table 4. Relationship between conservation and splicing in human alternatively spliced variant exons

	Total	CDS-related	Protein-motif-related	Retrotransposon-related	Exonic-splicing-enhancer-related
C/CS exons	133 901	95 583 (71%)	27 805 (21%)	2523 (2%)	130 104 (97%)
C/AS exons	37 646	20 805 (55%)	6192 (16%)	2308 (6%)	35 702 (95%)
NC/CS exons	15 683	7030 (45%)	1898 (12%)	2549 (16%)	14 961 (95%)
NC/AS exons	12 196	3516 (29%)	812 (7%)	4544 (37%)	11 701 (96%)
All exons	99 426	126 934 (64%)	36 707 (18%)	11 924 (6%)	192 468 (97%)

C, conserved; NC, non-conserved.

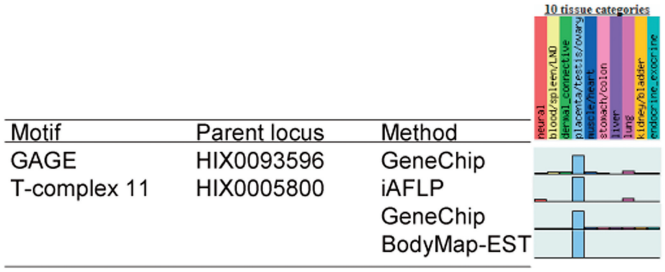


Figure 4. Expression pattern of parent genes in H-ANGEL that have GAGE or T-complex 11 motifs. The expressed pattern was examined by iAFLP, GeneChip and BodyMap-EST. The height of bars indicates the percentage in 10 tissue categories. The name of tissues are as follows (from the left): neural, blood/spleen/lymph node dissection (LND), dermal/connective, placenta/testis/ovary, muscle/heart, stomach/colon, liver, lung, kidney/bladder and endocrine/exocrine. The expressed specific tissue here was testis according to more detailed categories in H-ANGEL.

exons in terms of their length distributions, base compositions, absolute base sizes and relative positions in introns, or frequencies of particular AS relationships (data not shown).

DISCUSSION

We performed a large-scale comparative study of the AS variants in human transcripts, a study based on completely sequenced and intensively annotated full-length human and mouse cDNAs. Our dataset of AS full-length cDNAs is unique in that all combinations of AS exons have been defined as single entities in the complete form of the transcript. To the best of our knowledge, this is the largest available dataset of its kind. Although some of the AS variants identified here, especially those having an easily detectable internal AS pattern like a cassette, have been previously identified with EST-based or microarray-based approaches and are in the RefSeq and Ensembl databases, the previous identification of those variants was based on the interpretation of fragmented partial cDNA sequences and computational gene predictions. Indeed, 12 710 of the 13 705 cassette AS exons identified in this study (93%) had been identified in more than one EST-based study (Supplementary Table 7A). On the other hand, only 11% of the cassettes AS exons we identified in this study are in the RefSeq database and only 38% of them are in the Ensembl database (Supplementary Table 7B). We think that one of the major barriers to these AS exons being included in the representative gene models has been a lack of precise annotations, which are now available in our dataset. Our manually and computationally inspected data allowed us to examine the genome-wide features of AS sequences in a far more reliable manner than was previously possible, and our findings will enhance our understanding of the complex biological nature of AS sequences. A very interesting paper based on our AS database revealing the relation between transcriptional start sites (TSSs) and AS has already been published (39).

This study provides the first glimpse of the evolutionary turnover of AS sequences in full-length transcripts. Throughout evolution, the genomes of higher eukaryotes have expanded and mutations have been introduced by environmental factors (e.g. ultraviolet radiation) or internal factors (e.g. transposons). Mammals and other higher eukaryotes are thought to have accumulated numerous novel AS variants, and many AS sequences identified in this study appear to have occurred in a human-specific manner. By integrating information about evolutionary conservation with functional annotations, we have determined that non-conserved exons appear to have unexpectedly been excluded from the CDSs (Table 4). We had hypothesized that the CDSs would be more likely to contain non-conserved AS exons than conserved ones because the selective pressure against their biological functions is weak. We however found evidence of selective pressure against non-conserved AS exons in the CDS. Thus, it is likely that exons are easily generated by alternative splicing after speciation but are more likely to produce malfunctioning proteins when they occur in the CDS. It is also interesting that most AS variants are produced in the brain and testis (40). Because these tissues contain large numbers of neurons and sperm cells that could be functionally served as mutual substitutions or because variability itself may be important in these tissues, the occasional malfunction of individual proteins might be relatively well tolerated there. Thus, cells in the brain and testis might take advantage of the opportunity to experiment with AS variants, and this experimentation might act as a driving force in evolution and speed the process of speciation.

We shall soon overcome the current problems associated with the use of full-length cDNA data, the greatest of which—the limited number of sequences available—prevents us from analysing all AS variants. Researchers are already initiating large-scale projects by, using the next generation of DNA sequencers (e.g. Illumina 1G and SOLiD) to sequence previously isolated and partially

sequenced cDNA clones. The same type of full-length cDNA data now available for mice and men will soon become available for other mammals. Further enrichment of the expression information will also allow us to precisely analyse in what tissue the species-specific AS sequences are utilized. The current presumption that the products of AS are found most frequently in brain and testis should be tested by non-EST-based sampling-bias-free analysis. Cataloguing full-length cDNAs and creating related databases are the first steps toward understanding how AS variants contribute to the functional and evolutionary diversification of gene functions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Y. Kawahara, A. Matsuya, H. Nakaoka, T. Habara, F. Todokoro and C. Yamasaki for their assistance with genomic mapping and ORF predictions. They also thank E. Sekimori and H. Wakaguri for their technical support during the computational analysis and are grateful to all who annotated the full-length human cDNAs at the H-Inv and H-Inv 2 conferences.

FUNDING

Genome Information Integration Project of the Ministry of Economy; Trade and Industry of Japan; the Ministry of Education, Culture, Sports, Science and Technology of Japan; Japan Biological Informatics Consortium (JBIC). Funding for open access charge: JBIC.

Conflict of interest statement. None declared.

REFERENCES

1. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
2. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
3. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7 (**Suppl. 1), S41–S49.
4. 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. *Nat. Genet.*, **34**, 177–180.
5. Pritsker,M., Doniger,T.T., Kramer,L.C., Westcot,S.E. and Lemischka,I.R. (2005) Diversification of stem cell molecular repertoire by alternative splicing. *Proc. Natl Acad. Sci. USA*, **102**, 14290–14295.
6. 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.
7. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
8. Takeda,J., Suzuki,Y., Nakao,M., Barrero,R.A., Koyanagi,K.O., Jin,L., Motono,C., Hata,H., Isogai,T., Nagai,K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.*, **34**, 3917–3928.
9. 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.
10. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
11. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
12. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
13. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
14. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
15. Yamasaki,C., Kawashima,H., Todokoro,F., Imamizu,Y., Ogawa,M., Tanino,M., Itoh,T., Gojobori,T. and Imanishi,T. (2006) TACT: transcriptome auto-annotation conducting tool of H-InvDB. *Nucleic Acids Res.*, **34**, W345–W349.
16. Yamasaki,C., Koyanagi,K.O., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
17. 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.
18. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
19. Fujii,Y., Itoh,T., Sakate,R., Koyanagi,K.O., Matsuya,A., Habara,T., Yamaguchi,K., Kaneko,Y., Gojobori,T. and Imanishi,T. (2005) A web tool for comparative genomics: G-compass. *Gene*, **364**, 45–52.
20. Matsuya,A., Sakate,R., Kawahara,Y., Koyanagi,K.O., Sato,Y., Fujii,Y., Yamasaki,C., Habara,T., Nakaoka,H., Todokoro,F. *et al.* (2008) Evola: ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.*, **36**, D787–D792.
21. Takeuchi,N. and Ueda,T. (2003) Down-regulation of the mitochondrial translation system during terminal differentiation of HL-60 cells by 12-O-tetradecanoyl-1-phorbol-13-acetate: comparison with the cytoplasmic translation system. *J. Biol. Chem.*, **278**, 45318–45324.
22. Inukai,K., Funaki,M., Ogihara,T., Katagiri,H., Kanda,A., Anai,M., Fukushima,Y., Hosaka,T., Suzuki,M., Shin,B.C. *et al.* (1997) p85alpha gene generates three isoforms of regulatory subunit for phosphatidylinositol 3-kinase (PI 3-Kinase), p50alpha, p55alpha, and p85alpha, with different PI 3-kinase activity elevating responses to insulin. *J. Biol. Chem.*, **272**, 7873–7882.
23. Ueki,K., Algenstaedt,P., Mauvais-Jarvis,F. and Kahn,C.R. (2000) Positive and negative regulation of phosphoinositide 3-kinase-dependent signaling pathways by three different gene products of the p85alpha regulatory subunit. *Mol. Cell. Biol.*, **20**, 8035–8046.
24. Simader,H., Hothorn,M., Kohler,C., Basquin,J., Simos,G. and Suck,D. (2006) Structural basis of yeast aminoacyl-tRNA

synthetase complex formation revealed by crystal structures of two binary sub-complexes. *Nucleic Acids Res.*, **34**, 3968–3979.
25. Kim,J.E., Kim,K.H., Lee,S.W., Seol,W., Shiba,K. and Kim,S. (2000) An elongation factor-associating domain is inserted into human cysteinyl-tRNA synthetase by alternative splicing. *Nucleic Acids Res.*, **28**, 2866–2872.
26. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
27. Kawamata,H., Fujimori,T. and Imai,Y. (2004) TSC-22 (TGF-beta stimulated clone-22): a novel molecular target for differentiation-inducing therapy in salivary gland cancer. *Curr. Cancer Drug Targets*, **4**, 521–529.
28. Deppmann,C.D., Alvania,R.S. and Taparowsky,E.J. (2006) Cross-species annotation of basic leucine zipper factor interactions: insight into the evolution of closed interaction networks. *Mol. Biol. Evol.*, **23**, 1480–1492.
29. Tsuritani,K., Irie,T., Yamashita,R., Sakakibara,Y., Wakaguri,H., Kanai,A., Mizushima-Sugano,J., Sugano,S., Nakai,K. and Suzuki,Y. (2007) Distinct class of putative 'non-conserved' promoters in humans: comparative studies of alternative promoters of human and mouse genes. *Genome Res.*, **17**, 1005–1014.
30. Urrutia,R. (2003) KRAB-containing zinc-finger repressor proteins. *Genome Biol.*, **4**, 231.
31. Resch,A., Xing,Y., Modrek,B., Gorlick,M., Riley,R. and Lee,C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J. Proteome Res.*, **3**, 76–83.
32. Oh,H.J. and Lau,Y.F. (2006) KRAB: a partner for SRY action on chromatin. *Mol. Cell Endocrinol.*, **247**, 47–52.
33. Johnson,M.E., Viggiano,L., Bailey,J.A., Abdul-Rauf,M., Goodwin,G., Rocchi,M. and Eichler,E.E. (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature*, **413**, 514–519.
34. Zendman,A.J., Van Kraats,A.A., Weidle,U.H., Ruiter,D.J. and Van Muijen,G.N. (2002) The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma. *Int. J. Cancer*, **99**, 361–369.
35. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. *et al.* (2005) The human anatomic gene expression library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**, D567–D572.
36. Ma,Y., Zhang,S., Xia,Q., Zhang,G., Huang,X., Huang,M., Xiao,C., Pan,A., Sun,Y., Lebo,R. *et al.* (2002) Molecular characterization of the TCP11 gene which is the human homologue of the mouse gene encoding the receptor of fertilization promoting peptide. *Mol. Hum. Reprod.*, **8**, 24–31.
37. Sorek,R. (2007) The birth of new exons: mechanisms and evolutionary consequences. *RNA*, **13**, 1603–1608.
38. Parmley,J.L., Chamary,J.V. and Hurst,L.D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.*, **23**, 301–309.
39. Chern,T.M., Paul,N., van Nimwegen,E. and Zavolan,M. (2008) Computational analysis of full-length cDNAs reveals frequent coupling between transcriptional and splicing programs. *DNA Res.*, **15**, 63–72.
40. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.*, **5**, R74.