*Human Molecular Genetics, 2003, Vol. 12, No. 11* **1313–1320**
*DOI: 10.1093/hmg/ddg137*

# Low conservation of alternative splicing patterns in the human and mouse genomes

**Ramil N. Nurtdinov[1], Irena I. Artamonova[2], Andrei A. Mironov[3,4] and Mikhail S. Gelfand[3,4,*]**

[1]Moscow State University, Department of Physics/Biophysics, GSP-2, Leninskie Gory, Moscow 119922, Russia, [2]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Miklukho-Maklaya 16/10, Moscow 117997, Russia, [3]IntegratedGenomics-Moscow, PO Box 348, Moscow 117333, Russia and [4]State Scientific Center GosNIIGenetika, 1st Dorozhny 1, Moscow 113545, Russia

Received January 28, 2003; Revised March 14, 2003; Accepted March 21, 2003

**Alternative splicing has recently emerged as a major mechanism of generating protein diversity in higher eukaryotes. We compared alternative splicing isoforms of 166 pairs of orthologous human and mouse genes. As the mRNA and EST libraries of human and mouse are not complete and thus cannot be compared directly, we instead analyzed whether known cassette exons or alternative splicing sites from one genome are conserved in the other genome. We demonstrate that about half of the analyzed genes have species-specific isoforms, and about a quarter of elementary alternatives are not conserved between the human and mouse genomes. The detailed results of this study are available at www.ig-msk.ru:8005/HMG_paper.**

## INTRODUCTION

Sequencing of the human (1) and mouse (2) genomes led to the identification of about 30 000 genes, in contrast to earlier estimates of up to 120 000 human genes (3). On the other hand, large-scale sequencing of human ESTs demonstrated that at least 30% of human genes are alternatively spliced (4,5), which again contrasted with the 5% accepted previously (6). Later this estimate was raised to 60% (1,7, reviewed in 8). The frequency of alternative splicing of mouse genes, estimated as 33% (9), was recently increased to 41% by analysis of full-length cDNAs (10). In fact, the observed frequency of alternative splicing depends on EST coverage (9,11) and on accepted thresholds for alternative splices to be considered functional (11).

Alternative splicing (AS) was shown to remove/insert complete protein domains rather than disrupt domains, and, in the latter case, to target functional sites within domains (12). Finally, regulated AS is involved in human genetic disease (13–15), and up to 15% of disease-causing point mutations are thought to be responsible for aberrant splicing (16). Thus alternative splicing emerges as a major mechanism of generating proteome diversity (17–19), although it has been suggested that some EST-based alternative variants might represent cellular noise (8,11,18).

Comparative analysis of exon–intron structure of eukaryotic genes demonstrated higher conservation of intron sequences near alternative splicing sites in different species of fruit flies *Drosophila* (20) and nematodes *Caenorhabditis* (21). Orthologous mammalian introns also show weak, but noticeable conservation at intron termini (F.A. and A.S. Kondrashov, personal communication). This is thought to reflect the presence of weak dispersed regulatory elements at intron termini. On the other hand, there is no avoidance of SNPs (single-nucleotide polymorphisms) in intron positions outside the splicing sites (F.A. and A.S. Kondrashov, personal communication) (22).

There exists anecdotal evidence of non-conserved alternative splicing patterns in different genomes, (for some examples see Table 1). Species-specific splicing variants are observed in transgenic constructions (23,24), and the same isoforms can be differently regulated in different genomes of mammals (23,25), between mammals and birds (26), or between different orders of insects, e.g. Diptera [fruit flies *D. melanogaster* and *D. virilis* (27) and fly *Megaselia scalaris* (28)] and Lepidoptera [silkworm *Bombyx mori* (29)]. However, conservation of alternative splicing patterns on the genome level has not yet been subject to large-scale computational analysis.

One reason for that is the incompleteness of EST and mRNA collections. This makes it difficult to directly compare the observed splicing isoforms and various normalizing techniques have to be used, producing wide prediction margins, e.g. 8–42% of human alternative splices were shown to be conserved in mouse (11).

*To whom correspondence should be addressed at: State Scientific Center GosNIIGenetika, 1st Dorozhny 1, Moscow 113545, Russia. Tel: +7 0951352041; Fax: +7 0951326080; Email: misha@imb.ac.ru

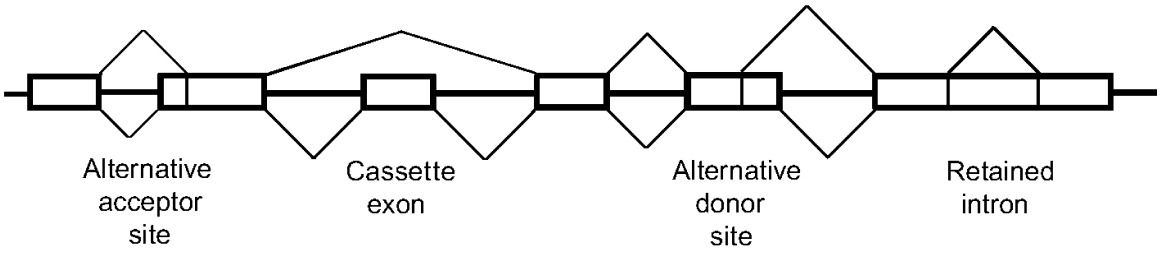


**Figure 1.** Types of elementary alternatives.

**Table 1.** Examples of species-specific alternative splicing

| Gene and product | Genomes | References |
|---|---|---|
| ALAS (erythroid 5-aminolevulinate synthase) | Human, mouse | (66) |
| GAP (Ras GTPase) | Human, monkey, mouse, rat, pig, sheep | (67) |
| p53 (tumor supressor) | Human, mouse, rat | (68,46,48) |
| PTP-S (non-membrane protein tyrosine phosphatase) | Human, mouse | (69) |
| Aggrecan | Human, cow, dog | (47) |
| HSL (hormone-sensitive lipase) | Human, mouse | (25) |
| H4P (inter-alpha-inhibitor) | Human, rat | (70) |
| ER-alpha (estrogen receptor) | Human, mouse | (71) |
| CS (calpastatin) | Human, mouse, rat | (72) |
| Ad4BP/SF-1 (NR5A1) | Mouse, rat | (73) |
| SERCA (sarco/endoplasmic reticulum Ca(2+)ATPase) | Human, rat | (74) |
| MCOLN1/Mcoln1 (mucolipidosis type IV) | Human, mouse | (75) |
| PECAM-1 (cell adhesion protein) | Human, mouse, rat | (76) |

We apply a conservative approach: we analyze whether a cassette exon, retained intron, or an alternative splicing site (Fig. 1) observed in one species is present in the genome of the other species. This variant may be non-functional due to changes in regulation, as demonstrated by some examples listed above and in Table 1. Thus our results may underestimate the true extent of non-conserved alternative splicing. On the other hand, unless we accept that alternatively spliced regions may evolve at a much faster rate than constitutive regions, so that alternative coding regions are no longer recognizable by the similarity search, we will not lose any splicing variants, and thus will not underestimate the conservatism of alternative splicing.

## RESULTS

Analysis of GenBank (30), draft human genome (1), and two databases of alternative splicing, AsMamDB (31) and HASDB (32) resulted in identification of 166 pairs of orthologous alternatively spliced human and mouse genes. Of these, mRNA and EST evidence of AS in both genomes was available for 84 gene pairs, in the human genome, for 42 genes, and in the mouse genome, in 40 genes.

We considered four types of elementary alternatives: cassette (on/off) exons, alternative donor and acceptor splicing sites, and retained introns (Fig. 1). We did not distinguish cassette and alternative (mutually exclusive) exons, as the comparative genomic approach does not allow one to take into account dependencies between elementary alternatives. One hundred and twenty six alternatively spliced human genes contained 177 cassette exons, 51 alternative acceptor sites, 52 alternative donor sites, and 12 retained introns. The total number of known human elementary alternatives was 285. One hundred and twenty four alternatively spliced mouse genes contained 123 cassette exons, 46 alternative acceptor sites, 53 alternative donor sites and 29 retained introns. The total number of known mouse elementary alternatives was 252. Only 51 elementary alternatives were described in both genomes.

All alternative isoforms were translated and aligned with genomic sequences of both species (see *Data and Methods*). As the EST collections might contain aberrant non-functional splicing variants, we considered separately elementary alternatives supported by mRNA and EST evidence. The results are shown in Table 2. Graphical maps for all gene pairs in this study genes are available at www.ig-msk.ru:8005/HMG_paper.

Only 69–83% of elementary alternatives are conserved. The degree of conservation is higher in mRNA-derived alternatives (76–83%) compared with EST-derived ones (69–75%), and is higher for mouse alternatives (73–83%) than for human ones (69–76%). However, it is encouraging that the differences between the four columns in Table 2 (human/mouse, mRNA/EST) are not high (lower values for EST-derived alternatives were expected, see below). No significant differences in behavior of different types of elementary alternatives could be observed.

Only 51 elementary alternatives were observed in both genomes in the initial sample. This is ~20% of all considered elementary alternatives and lies between the estimates of 8% conserved alternative splices and 42% conserved reliable alternative splices reported in (11). It is noteworthy that

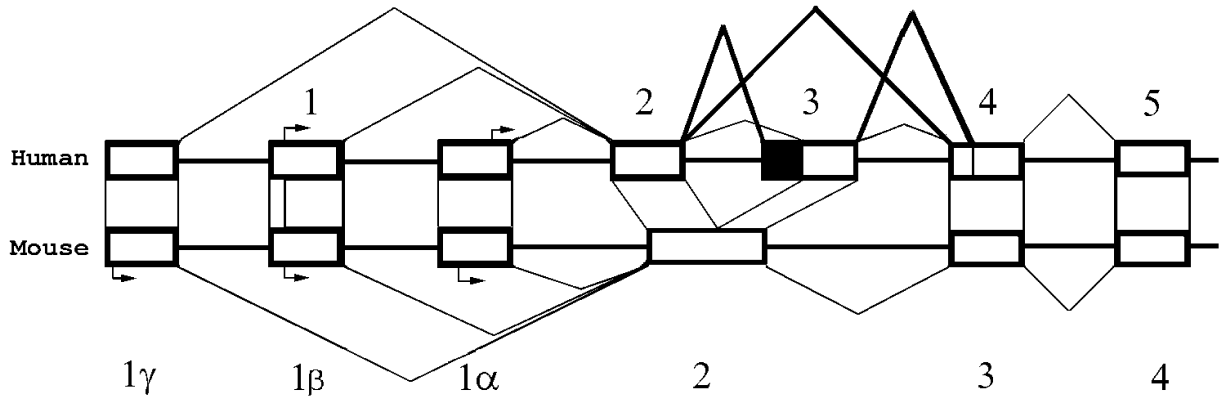


**Figure 2.** Exon–intron structure of Fxyd2/FXYD2 gene encoding Na/K-ATPase. Arrows: start codons. Bold lines: alternatives generating species-specific isoforms. The scheme is not to scale.

**Table 2.** Conservation of elementary alternatives from mRNAs and ESTs

| | Human | | | | Mouse | | | |
|---|---|---|---|---|---|---|---|---|
| | mRNA | | EST | | mRNA | | EST | |
| | C | NC | C | NC | C | NC | C | NC |
| Cassette exons | 56 | 25 | 74 | 26 | 70 | 5 | 39 | 9 |
| Alternative donor sites | 18 | 7 | 16 | 10 | 24 | 6 | 17 | 6 |
| Alternative acceptor sites | 13 | 5 | 19 | 15 | 15 | 6 | 16 | 9 |
| Retained introns | 4 | 3 | 5 | 0 | 8 | 7 | 10 | 4 |
| Total alternatives | 96 | 30 | 114 | 51 | 117 | 24 | 82 | 28 |
| Degree of conservation | 76% | | 69% | | 83% | | 75% | |
| Genes | 45 | 28 | 41 | 44 | 68 | 22 | 30 | 26 |
| Total genes | C: 57 (45%), NC: 69(55%) | | | | C: 79 (64%), NC: 45 (36%) | | | |
| Total gene pairs | C: 74 (45%), NC: 82 (55%) | | | | | | | |

C, conserved elementary alternatives (for genes, all isoforms conserved); NC, non-conserved elementary alternatives (for genes, at least one non-conserved isoform).

retained introns, the main type of aberrant or artefactual data in EST databases are mainly conserved both in mRNA and EST comparisons. In fact, the degree of conservation of retained introns could even be overestimated, as the data were specifically filtered in order to remove ESTs arising from unspliced transcripts. Because of that conservatism of retained introns from mRNAs appears to be higher than conservatism of retained introns from ESTs.

Twenty-three gene pairs of specific interest were selected for exhaustive manual analysis using as complete literature search as possible. Among these pairs there were 11 pairs with seven or more elementary alternatives, five pairs with six elementary alternatives, of which at least half were non-conserved, three pairs with complicated combinations of elementary alternatives, three pairs with alternatives involving an additional human intron not present in the mouse gene and, finally, gene p53 selected because of its functional importance.

First of all, analysis of the literature revealed four elementary alternatives not represented in the initial sample and two cases of multiple alternatives caused by the use of alternative promoters; they were taken into account in Table 2. As EST do not retain information about long-range correlation between elementary alternatives, we did not attempt to calculate the number of new isoforms. On the other hand, there were several well-studied human genes, for which only one isoform has been published, but analysis of ESTs suggests alternative splicing. Some examples are cell cycle checkpoint control protein (33) (GenBank U53174, UniGene Hs.240457), creatine transporter (34) (GenBank AAC41688, UniGene Hs.187958), and methyl-CpG binding protein 4 (35) (GenBank AF072250, Hs.35947). It remains to be seen in experiment whether the new candidate isoforms are functional. Finally, there are cases when isoforms are known under different names, e.g. MutS homolog 5 (36) (MSH5, GenBank AAC62533) and G7 protein (37) (GenBank CAB52406).

We have also observed numerous alternatives in non-coding regions; some prominent examples with multiple non-coding alternative exons are neuronatin (Peg5) (38) and menin (MEN1) (39). As discussed above, they could not be analyzed by the present approach, and we did not consider them in detail. However, we have found several cases of seemingly recent inactivation of exons. In these cases an alternative is distinguishable in DNA comparisons, but is no longer translated due to frameshifts or the absence of start codons. It is unlikely that such situation would represent the emergence of a protein-coding region from conserved non-coding DNA.

One such example is the gene for Na/K-ATPase gamma subunit (Fig. 2). The mouse gene *Fxyd2* has three isoforms

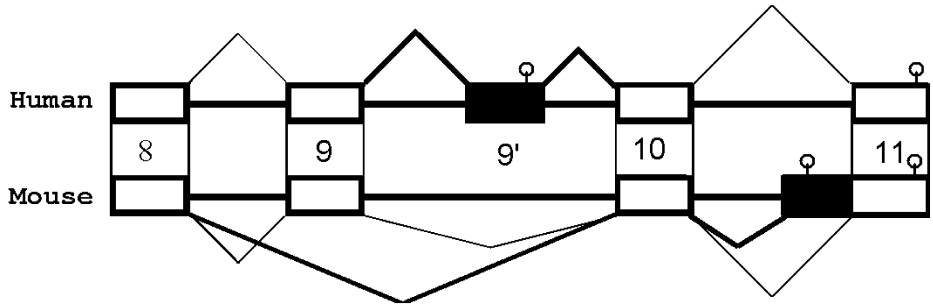


**Figure 3.** Fragment of the exon–intron structure of the p53 gene. Circles: (candidate) stop codons. Other notation as in Figure 2.

with different initial exons 1γ, 1β, 1α having tissue- and stage-specific pattern of expression (40) (GenBank AY035583, UniGene Mm.22742). The human gene *FXYD2* retains exon 1β; exon 1α is conserved on the nucleotide level, but it has a different start codon ATG and the respective protein N-termini cannot be aligned due to frameshifts; and finally, exon 1γ also is conserved on the nucleotide level, but has no ATG at all (41) (GenBank AF316896, UniGene Hs.19520). However, all three exons were observed in human mRNAs and thus are probably functional. Additionally, the human gene has an intron inserted in exon 2, which has now become two exons 2α and 2β (or exon 3 in the human numbering), the latter becoming a cassette exon. Finally, there are ESTs produced by alternative acceptor sites at exons 2β/3 and 3/4.

A complicated, rather unclear, situation is that of the mouse gene *CW17* (a.k.a. *zfp162* and *mzfm*) (42) and the orthologous human gene *SF1* (a.k.a. *ZFM1*) (43). Indeed, as noted in the original publication, one human isoform (ZFM1-A) has non-canonical splicing sites at both termini of the intron between exons 13 and 14 (in some isoforms this intron is retained; exon 13 itself is a cassette exon). Accordingly, the acceptor splicing site of this intron is not conserved. However, an isoform where this intron is spliced out has been observed in mouse as well (44). Further, some clones that apparently contain retained introns (45) are probably results of inefficient splicing, as these introns are not conserved in the mouse genome.

Alternative splicing of gene *p53* (Fig. 3) has been the subject of intensive experimental study. Our analysis included an unpublished mouse isoform that is known as GenBank entry AJ297973. This isoform lacks exon 9, that is a cassette exon. Our analysis cannot distinguish cassette exons if the longer (exon-containing) isoform is conserved; however, it should be noted that human isoforms without exon 9 were not observed, and thus the mouse alternative seems to be species-specific. This gene illustrates one more limitation of our analysis and the importance of taking into account all available data when drawing conclusions about specific genes. The mouse exon 11 (the terminal one) has an alternative acceptor site. This site is conserved in human and rat genes; however, isoforms utilizing this site have not been observed in human despite directed effort, and the frequency of such isoforms in rat is much lower than in mouse (46). This shows that conservation of a site is not sufficient to predict the existence of an isoform (a similar situation exists in the cartilage aggrecan gene, where the exon encoding the EGF1 domain is conserved with splicing sites intact, but is incorporated in human and bovine mRNAs, but not those of dog and rat) (47). Thus the two published species-specific isoforms generate proteins with different C-termini. The human isoform utilizing exon 9′ has a stop codon in this exon and generates a truncated protein. This isoform is probably functional, as the mRNA is observed in significant amounts in normal quiescent cells whereas the derived protein lacks the tetramerization domain and thus has specific transcriptional properties (48). Similarly, the mouse isoform utilizing the alternative acceptor site in exon 11 has a stop codon in the alternative region. The derived protein has increased non-specific and specific DNA-binding activity (49). Finally, we could not find counterparts for possible alternative splicing at the junction of exons 2 and 3 reported for *Danio rerio* (50), and it is likely that this issue will be resolved only when the zebrafish genome becomes available.

At present there is insufficient data to make a general statement about functionality of non-conserved alternative isoforms. The agreement of mRNA-derived and EST-derived estimates of the amount of non-conserved elementary alternatives shows that most non-conserved isoforms are expressed and are not due to experimental artifacts. If we accept a tissue-specific mRNA expression pattern or changed protein properties as evidence of functionality, to the discussed examples of the *Fxyd2*/*FXYD2* and *p53* we can add four more genes with functional non-conserved isoforms.

Mouse autoimmune regulator gene *Aire* (Fig. 4) has a non-conserved alternative donor site in exon 6 (extending this exon by four codons) and a conserved acceptor site in exon 8 (extending this exon by one codon). Further, there are isoforms lacking exon 10 or exons 10 and 11 (51). None of these elementary alternatives have been observed in human ESTs. However, all 12 ($2 \times 2 \times 3$) combinatorial variants have been detected *in vivo*, and the conserved isoform constitutes only 23% of mRNA in thymus (51). Further, there exists a testis-specific isoform utilizing a non-conserved exon 7′ (52).

The MIA/CD-RAP gene whose product exhibits malnoma-inhibiting activity has a human-specific isoform lacking exon 2. It generates a protein with different C-terminus (53). The ratio of the conserved and non-conserved isoforms differs in melanocytes and other cells; further, the conserved isoform localizes mainly in the nucleus, whereas the non-conserved protein isoform is expressed mainly in the cytoplasm (53).

Finally, multiple non-conserved isoforms of the human *TRAD*/*R51H3*/*RAD51D* gene have different tissue specificity

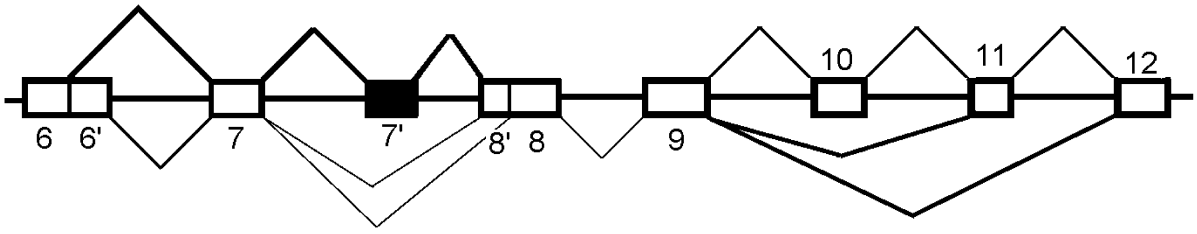


**Figure 4.** Fragment of the exon–intron structure of mouse gene AIRE. Notation as in Figure 2.

(54), whereas the transcription factor E2F6 gene has a mouse-specific cassette exon 2 containing a stop-codon; the corresponding isoform is the predominant one (55).

## DISCUSSION

The main question arising in all genomic analyses that involve EST data is that of reliability. Indeed, EST libraries are incomplete and known to contain a large number of artifacts arising both from experimental problems such as genome contamination and errors of cellular machinery such as aberrant splicing. However, the results observed on EST and mRNA data are very close, and overall about a quarter of all elementary alternatives seem to be species-specific, in contrast to the constitutive exon–intron structure that is generally conserved (2,56). About half of alternatively spliced genes have non-conserved isoforms. This suggests that the birth of new spliced isoforms is one of the major evolutionary mechanisms for generation of species-specific proteome diversity. Indeed, analysis of the literature demonstrated that functionality of at least some non-conserved isoforms is functional.

The higher than expected frequency of young (species-specific) isoforms agrees well with a number of other recent observations. It has been demonstrated that all human internal exons containing *Alu* sequences are alternatively spliced using sites within *Alu* sequences (57). As this family of repeats is specific to primates, all these cassette exons are necessarily specific to this taxonomic group. Cassette exons demonstrate lower rate of synonymous substitutions and marginally higher rate of non-synonymous substitutions compared with constitutive exons in pairs of human and non-primate mammalian genes (58). As the increased ratio of the frequency of non-synonymous to synonymous substitutions is a standard sign of positive selection (59), the alternatively spliced regions seem to be under positive selection. Similar results are observed in analysis of human SNPs, as there are less synonymous and more non-synonymous SNPs in alternative regions (V.E. Ramensky, A.A. Mironov, M.S. Gelfand, manuscript in preparation). Thus alternative splicing emerges as an evolutionary workshop for tinkering with the protein structure and function.

The recent origin of many splicing isoforms may partially explain the observed discrepancy between the human and mouse data and the mRNA and EST data. Indeed, species-specific isoforms can be expected to have narrower tissue- or stage-specificity than older isoforms. Consequently, they are less well known and thus under-represented in mRNA samples, as compared with EST data, and also under-represented in the mouse data. The observed difference in conservatism of mRNA- and EST-derived elementary alternatives is 8%, and 6% fewer human elementary alternatives are conserved compared with mouse. We expect that this gap will close as more reliable data become available.

Of course, all estimates made here should be treated with appropriate caution. Recent publication of the complete mouse genome makes it possible to repeat this study on a larger dataset, and this will be done. The conservative technique of genomic comparison (instead of direct EST comparison) developed here can be applied only to protein-coding regions of exons. It also cannot identify cases when splicing sites are conserved, but still do not function, and we know that this may happen, e.g. in the *p53* gene, nor can it identify exons that are cassette in one species and constitutive in the other. On the other hand, alternative splicing may be confused with allele-specific exon deletion, as in the human growth hormone receptor gene (60). At present these cases look as rare oddities, but the history of studying alternative splicing is ripe with oddities becoming a rule.

Here we do not account for varying EST coverage, and thus some of the isoforms may be due to aberrant splicing and non-functional, especially as many clone libraries are derived from cancer cell lines where increased error rate of many cellular processes may be expected (61). However, reasonable agreement of mRNA- and EST-derived estimates and conservation of retained introns, the latter being the most frequent case of aberrant alternatives (11), makes it unlikely that our data were heavily contaminated.

Similarly, although identifying orthologs in the absence of a complete mouse genome might lead to creation of spurious pairs formed by paralogous genes, the applied criteria of orthology were sufficiently stringent not to let it happen unless we encountered a very recent duplication. In the latter case conservation of all alternatives on the genome level could be expected anyway, although some of isoforms may be specific for only one copy, as in the following example. The mouse genome contains two paralogous genes encoding intercellular adhesion molecule, *Ceacam1* and *Ceacam2*, whereas both human and rat genomes contain only one such gene (62). Of these, only the second paralog was shown to be alternatively spliced. In fact, systematic comparison of alternative splicing of recent paralogs could be the subject of an independent study.

This study opens a door for many questions. Is there anything specific about the function of young isoforms? Are

they tissue- or stage-specific, and if yes, can this specificity be linked to differences in human and mouse anatomy or physiology? How do young splices influence the protein structure, as compared with conserved alternative splicing events? How conserved is alternative splicing at larger evolutionary distances, e.g. between human and fugu? Completion of additional mammalian genomes, projects aimed at cataloging of all mRNAs of human and mouse, and increasing EST coverage of human, mouse and other genomes make it likely that at least some answers will be found in near future.

## MATERIALS AND METHODS

Initially 431 groups of mouse alternative splicing isoforms from AsMamDB 1.0 (31) (http://166.111.30.65/ASMAMDB.html) were considered. Of these, we retained 201 groups containing genomic sequences (the complete mouse genome was unavailable at the time of this study). The exon–intron structure of these 201 genes was reconstituted by spliced alignment mRNA sequences and genomic DNA sequences using Pro-EST (4). We then removed sequences with non-alignable overhangs and, mRNAs that differed only at start or end points, and sequences with alternatives in non-coding regions. This resulted in 76 groups of protein-coding isoforms with well-established exon–intron structure and, in particular, canonical GT-AG dinucleotides at all intron termini. The longest isoform was translated into protein and used to find the human ortholog in the draft version of the human genome (1) using TBLASTN (63) (www.ncbi.nlm.nih.gov/BLAST/). Human homologs were accepted if they produced strong alignment throughout the entire protein length and had the same exon–intron structure (including conservation of exon lengths up to several amino acids) in constitutively spliced regions. In addition, absence of close paralogs was required. This resulted in 62 pairs of orthologous mouse–human gene pairs with known alternative splicing in mouse.

All 2121 complete mouse genes from GenBank (30) (www.ncbi.nlm.nih.gov/) with known genomic sequence and exon–intron structure, excluding immunoglobulin and T-cell receptor genes, were compared with a database of alternatively spliced human genes HASDB (32) (www.bioinformatics.ucla.edu/~splice/HASDB/). To do that, one mRNA was selected from each UniGene (64) (www.ncbi.nlm.nih.gov/UniGene/) cluster corresponding to one of 6201 HASDB entries and compared with translated complete mouse genes using BLASTX. After establishing pairs consisting of a mouse gene and a UniGene/HASDB cluster with confirmed alternative splicing, the corresponding human genes were selected from the draft version of the human genome (1). This resulted in 116 human–mouse genes pairs with known alternative splicing in human. Of these, 33 pairs were already generated by analysis of mouse alternatively spliced genes, and thus had known AS in both species.

Additionally, we considered 21 mouse gene with alternative splicing annotated in GenBank entries. Their human orthologs were identified in UniGene and the human genome using TBLASTN. Evidence of alternative splicing was found for eight human genes. Finally, for all mouse genes, ESTs from UniGene with elementary alternatives not covered by previously identified isoforms were selected; the standard precautions against contamination by unspliced or partially spliced transcripts were taken.

The final sample consisted of 166 human–mouse gene pairs with mRNA and ESTs showing alternative splicing in at least one genome. Of these, 84 pairs were known to be alternatively spliced in both genomes, 42 genes in the human genome and 40 genes in the mouse genome.

Protein sequences corresponding to the observed isoforms were aligned to genomic sequences using Pro-Frame (65). Alternative splicing sites were assumed to be conserved if (a) the alternatively spliced regions could be aligned and (b) the canonical dinucleotides, GT for donor sites and AG for acceptor sites, were observed in the matching positions in the alignment with only short insertions of an integer number of codons allowed. An alternative exon was assumed to be conserved if (a) it could be aligned at the same level of similarity as the rest of the protein and (b) it was bounded by canonical dinucleotides. Finally, a retained intron was considered conserved if (a) it could be well aligned and (b) its length was a multiple of 3, unless this intron contained an in-frame stop-codon. Missing cassette exons or alternatively spliced parts of exons were searched for in corresponding introns using Pro-Frame and TBLASTN.

Thus, non-conserved alternatives were defined as those that had no counterpart in the initial genome, or could not function because of loss of the splicing sites at their boundaries, or disrupted the reading frame downstream. The quality of obtained alignments was assessed by eye, however, in no case any ambiguity was encountered. To account for lower conservation of alternatively spliced regions (cassette exons, retained introns, or exon extensions generated by alternative splicing sites), the threshold for accepting alignments of such regions was 60% identity, whereas for constitutive exons it was 70%.

If the reading frame of an alternative region could not be determined, all three possible reading frames were considered. Alternatives involving the start or stop codons were considered only within the coding region, and non-coding leader (respectively, trailer) sequences were ignored.

## ACKNOWLEDGEMENTS

We are grateful to V. Yu. Makeev for useful discussions. This study was partially supported by grants from the Howard Hughes Medical Institute (55000309) and the Ludwig Institute for Cancer Research (CRDF RB0-1268).

## REFERENCES

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Mouse Genome Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L. and Quackenbush, J. (2000) Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.*, **25**, 239–240.
4. Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
5. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. and Bork, P. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, 83–86.
6. Sharp, P.A. (1994) Split genes and RNA splicing. *Cell*, **77**, 805–815.

7. Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.
8. Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
9. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.
10. The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I and II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
11. Kan, Z., States, D. and Gish, W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.
12. Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S. and Sunyaev S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet.*, **9**, 124–128.
13. Cooper, T.A. and Mattox, W. (1997) The regulation of splice-site selection, and its role in human disease. *Am. J. Hum. Genet.*, **61**, 259–266.
14. Caceres, J.F. and Kornblihtt, A.R. (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.*, **18**, 186–193.
15. Savkur, R.S., Philips, A.V. and Cooper, T.A. (2001) Aberrant regulation of insulin receptor alternative splicing is associated with insulin resistance in myotonic dystrophy. *Nat. Genet.*, **29**, 40–47.
16. Krawczak, M., Reiss, J. and Cooper, D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
17. Black, D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367–370.
18. Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic word. *Trends Genet.*, **17**, 100–107.
19. Roberts, G.C. and Smith, C.W.J. (2002) Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol.*, **6**, 375–383.
20. Thackeray, J.R. and Ganetzky, B. (1995) Conserved alternative splicing patterns and splicing signals in the *Drosophila* sodium channel gene *para*. *Genetics*, **141**, 203–214.
21. Kent, W.J. and Zahler, A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.
22. Majewski, J. and Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.*, **12**, 1827–1836.
23. Pret, A.M. and Fiszman, M.Y. (1996) Sequence divergence associated with species-specific splicing of the nonmuscle beta-tropomyosin alternative exon. *J. Biol. Chem.*, **271**, 11511–11517.
24. Aigner, B., Pambalk, K., Reichart, U., Besenfelder, U., Bosze, Z., Renner, M., Gunzburg, W.H., Wolf, E., Muller, M. and Brem, G. (1999) Species-specific alternative splicing of transgenic RNA in the mammary glands of pigs, rabbits, and mice. *Biochem. Biophys. Res. Commun.*, **257**, 843–850.
25. Laurell, H., Grober, J., Vindis, C., Lacombe, T., Dauzats, M., Holm, C. and Langin, D. (1997) Species-specific alternative splicing generates a catalytically inactive form of human hormone-sensitive lipase. *Biochem. J.*, **328**, 137–143.
26. Jin, J.P. (1996) Alternative RNA splicing-generated cardiac troponin T isoform switching: a non-heart-restricted genetic programming synchronized in developing cardiac and skeletal muscles. *Biochem. Biophys. Res. Commun.*, **225**, 883–889.
27. Hertel, K.J., Lynch, K.W., Hsiao, E.C., Liu, E.H. and Maniatis, T. (1996) Structural and functional conservation of the *Drosophila doublesex* splicing enhancer repeat elements. *RNA*, **2**, 969–981.
28. Kuhn, S., Sievert, V. and Traut, W. (2000) The sex-determining gene *doublesex* in the fly *Megaselia scalaris*: conserved structure and sex-specific splicing. *Genome*, **43**, 1011–1020.
29. Suzuki, M.G., Ohbayashi, F., Mita, K. and Shimada, T. (2001) The mechanism of sex-specific splicing at the doublesex gene is different between *Drosophila melanogaster* and *Bombyx mori*. *Insect Biochem. Mol. Biol.*, **31**, 1201–1211.
30. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) GenBank. *Nucl. Acids Res.*, **30**, 17–20.
31. Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X. and Li, Y. (2001) AsMamDB: an alternative splice database of mammals. *Nucl. Acids Res.*, **29**, 260–263.
32. Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucl. Acids Res.*, **29**, 2850–2859.
33. Lieberman, H.B., Hopkins, K.M., Nass, M., Demetrick, D. and Davey, S. (1996) A human homolog of the *Schizosaccharomyces pombe rad9+* checkpoint control gene. *Proc. Natl Acad. Sci. USA.* **93**, 13890–13895.
34. Snow, R.J. and Murphy, R.M. (2001) Creatine and the creatine transporter: a review. *Mol. Cell. Biochem.*, **224**, 169–181.
35. Hendrich, B. and Bird, A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.*, **18**, 6538–6547.
36. Her, C. and Doggett, N.A. (1998) Cloning, structural characterization, and chromosomal localization of the human orthologue of *Saccharomyces cerevisiae MSH5* gene. *Genomics*, **52**, 50–61.
37. Snoek, M., Albertella, M.R., van Kooij, M., Wixon, J., van Vugt, H., de Groot, K. and Campbell, R.D. (2000) *G7c*, a novel gene in the mouse and human major histocompatibility complex class III region, possibly controlling lung tumor susceptibility. *Immunogenetics*, **51**, 383–386.
38. Kagitani, F., Kuroiwa, Y., Wakana, S., Shiroishi, T., Miyoshi, N., Kobayashi, S., Nishida, M., Kohda, T., Kaneko-Ishino, T. and Ishino, F. (1997) *Peg5*/Neuronatin is an imprinted gene located on sub-distal chromosome 2 in the mouse. *Nucl. Acids Res.*, **25**, 3428–3432.
39. Khodaei-O'Brien, S., Zablewska, B., Fromaget, M., Bylund, L., Weber, G. and Gaudray P. (2000) Heterogeneity at the 5′-end of *MEN1* transcripts. *Biochem. Biophys. Res. Commun.*, **276**, 508–514.
40. Jones, D.H., Golding, M.C., Barr, K.J., Fong, G.H. and Kidder, G.M. (2001) The mouse Na+-K+-ATPase gamma-subunit gene (*Fxyd2*) encodes three developmentally regulated transcripts. *Physiol. Genomics*, **6**, 129–135.
41. Sweadner, K.J., Wetzel, R.K. and Arystarkhova, E. (2000) Genomic organization of the human FXYD2 gene encoding the gamma subunit of the Na,K-ATPase. *Biochem. Biophys. Res. Commun.*, **279**, 196–201.
42. Wrehlke, C., Schmitt-Wrede, H.P., Qiao, Z. and Wunderlich, F. (1997) Enhanced expression in spleen macrophages of the mouse homolog to the human putative tumor suppressor gene *ZFM1*. *DNA Cell Biol.*, **16**, 761–767.
43. Kramer, A., Quentin, M. and Mulhauser, F. (1998) Diverse modes of alternative splicing of human splicing factor SF1 deduced from the exon–intron structure of the gene. *Gene*, **211**, 29–37.
44. Wrehlke, C., Wiedemeyer, W.R., Schmitt-Wrede, H.P., Mincheva, A., Lichter, P. and Wunderlich, F. (1999) Genomic organization of mouse gene zfp162. *DNA Cell. Biol.*, **18**, 419–428.
45. Toda, T., Iida, A., Miwa, T., Nakamura, Y. and Imai, T. (1994) Isolation and characterization of a novel gene encoding nuclear protein at a locus (D11S636) tightly linked to multiple endocrine neoplasia type 1 (*MEN1*). *Hum. Mol. Genet.*, **3**, 465–470.
46. Laverdiere, M., Beaudoin, J. and Lavigueur. A. (2000) Species-specific regulation of alternative splicing in the C-terminal region of the *p53* tumor suppressor gene. *Nucl. Acids Res.*, **28**, 1489–1497.
47. Fulop, C., Cs-Szabo, G. and Glant, T.T. (1996) Species-specific alternative splicing of the epidermal growth factor-like domain 1 of cartilage aggrecan. *Biochem J.*, **319**, 935–940.
48. Flaman, J.-M., Waridel, F., Estreicher, A., Vannier, A., Limacher, J.-M., Gilbert, D., Iggo, R. and Frebourg, T. (1996) The human tumour suppressor gene p53 is alternatively spliced in normal cells. *Oncogene*, **12**, 813–818.
49. Bayle, J.H., Elenbaas, B. and Levine, A.J. (1995) The carboxyl-terminal domain of the p53 protein regulates sequence-specific DNA binding through its nonspecific nucleic acid-binding activity. *Proc. Natl Acad. Sci. USA*, **92**, 5729–5733.
50. Cheng, R., Ford, B.L., O'Neal, P.E., Mathews, C.Z., Bradford, C.S., Thongtan, T., Barnes, D.W., Hendricks, J.D. and Bailey, G.S. (1997) Zebrafish (*Danio rerio*) *p53* tumor suppressor gene: cDNA sequence and expression during embryogenesis. *Mol. Mar. Biol. Biotechnol.*, **6**, 88–97.
51. Ruan, Q.-G., Wang, C.-Y., Shi, J.-D. and She, J.-X. (1999) Expression and alternative splicing of the mouse autoimmune regulator gene (*Aire*). *J. Autoimmunity*, **13**, 307–313.
52. Heino, M., Peterson, P., Sillanpaa, N., Guerin, S., Wu, L., Anderson, G., Scott, H.S., Antonarakis, S.E., Kudoh, J., Shimizu, N. *et al.* (2000) RNA and protein expression of the murine autoimmune regulator gene (Aire) in normal, RelB-deficient and in NOD mouse. *Eur. J. Immunol.*, **30**, 1884–1893.

53. Hau, P., Wise, P., Bosserhoff, A.K., Blesch, A., Jachimczak, P., Tschertner, I., Bogdahn, U. and Apfel, R. (2002) Cloning and characterization of the expression pattern of a novel splice product MIA (splice) of malignant melanoma-derived growth-inhibiting activity (MIA/CD-RAP). *J. Invest. Dermatol.*, **119**, 562–569.
54. Kawabata, M. and Saeki, K. (1999) Multiple alternative transcripts of the human homologue of the mouse TRAD/R51H3/RAD51D gene, a member of the rec A/RAD51 gene family. *Biochem. Biophys. Res. Commun.*, **257**, 156–162.
55. Kherrouche, Z., Begue, A., Stehelin, D. and Monte, D. (2001) Molecular cloning and characterization of the mouse E2F6 gene. *Biochem. Biophys. Res. Commun.*, **288**, 22–33.
56. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
57. Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060–1067.
58. Iida, K. and Akashi, H. (2000) A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene*, **261**, 93–105.
59. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
60. Pantel, J., Machinis, K., Sobrier, M.L., Duquesnoy, P., Goossens, M. and Amselem, S. (2000) Species-specific alternative splice mimicry at the growth hormone receptor locus revealed by the lineage of retroelements during primate evolution. *J. Biol. Chem.*, **275**, 18664–18669.
61. Baranova, A.V., Lobashev, A.V., Ivanov, D.V., Krukovskaya, L.L., Yankovsky, N.K. and Kozlov, A.P. (2001) In silico screening for tumour-specific expressed sequences in human genome. *FEBS Lett.*, **508**, 143–148.
62. Han, E., Phan, D., Lo, P., Poy, M.N., Behringer, R., Najjar, S.M. and Lin, S.H. (2001) Differences in tissue-specific and embryonic expression of mouse *Ceacam1* and *Ceacam2* genes. *Biochem J.*, **355**, 417–423.
63. Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.*, **6**, 119–129.
64. Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.*, **29**, 11–16.
65. Mironov, A.A., Novichkov, P.S. and Gelfand, M.S. (2001) Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors. *Bioinformatics*, **17**, 13–15.
66. Conboy, J.G., Cox, T.C., Bottomley, S.S., Bawden, M.J. and May, B.K. (1992) Human erythroid 5-aminolevulinate synthase. Gene structure and species-specific differences in alternative RNA splicing. *J. Biol. Chem.*, **267**, 18753–18758.
67. Mollat, P., Fournier, A., Yang, C.Z., Alsat, E., Zhang, Y., Evain-Brion, D., Grassi, J. and Thang, M.N. (1994) Species specificity and organ, cellular and subcellular localization of the 100 kDa Ras GTPase activating protein. *J. Cell Sci.*, **107**, 427–435.
68. Will, K., Warnecke, G., Bergmann, S. and Deppert, W. (1995) Species- and tissue-specific expression of the C-terminal alternatively spliced form of the tumor suppressor *p53. Nucl. Acids Res.*, **23**, 4023–4028.
69. Reddy, R.S. and Swarup, G. (1995) Alternative splicing generates four different forms of a non-transmembrane protein tyrosine phosphatase mRNA. *DNA Cell Biol.*, **14**, 1007–1015.
70. Soury, E., Olivier, E., Daveau, M., Hiron, M., Claeyssens, S., Risler, J.L. and Salier, J.P. (1998) The H4P heavy chain of inter-alpha-inhibitor family largely differs in the structure and synthesis of its prolin-rich region from rat to human. *Biochem. Biophys. Res. Commun.*, **243**, 522–530.
71. Lu, B., Dotzlaw, H., Leygue, E., Murphy, L.J., Watson, P.H. and Murphy, L.C. (1999) Estrogen receptor-alpha mRNA variants in murine and human tissues. *Mol. Cell. Endocrinol.*, **158**, 153–161.
72. Takano, J., Kawamura, T., Murase, M., Hitomi, K. and Maki, M. (1999) Structure of mouse calpastatin isoforms: implications of species-common and species-specific alternative splicing. *Biochem. Biophys. Res. Commun.*, **260**, 339–345.
73. Kimura, R., Yoshii, H., Nomura, M., Kotomura, N., Mukai, T., Ishihara, S., Ohba, K., Yanase, T., Gotoh, O., Nawata, H. and Morohashi, K. (2000) Identification of novel first exons in *Ad4BP/SF-1* (*NR5A1*) gene and their tissue- and species-specific usage. *Biochem. Biophys. Res. Commun.*, **278**, 63–71.
74. Martin, V., Bredoux, R., Corvazier, E., Papp, B. and Enouf, J. (2000) Platelet Ca(2+)ATPases: a plural, species-specific, and multiple hypertension-regulated expression system. *Hypertension*, **35**, 91–102.
75. Falardeau, J.L., Kennedy, J.C., Acierno, J.S. Jr, Sun, M., Stahl, S., Goldin, E. and Slaugenhaupt, S.A. (2002) Cloning and characterization of the mouse *Mcoln1* gene reveals an alternatively spliced transcript not seen in humans. *BMC Genomics*, **3**, 3.
76. Wang, Y. and Sheibani, N. (2002) Expression pattern of alternatively spliced *PECAM-1* isoforms in hematopoietic cells and platelets. *J. Cell. Biochem.*, **87**, 424–438.