European Journal of Human Genetics (2004) 12, 139–149

www.nature.com/ejhg

ARTICLE

In *NF1*, *CFTR*, *PER3*, *CARS* and *SYT7*, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons

Dieter Kaufmann*[,1], Oliver Kenner[1], Peter Nurnberg[2], Walther Vogel[1] and Britta Bartelt[1]

[1]*Department of Human Genetics, University of Ulm, Germany;* [2]*Gene Mapping Center, Max Delbrueck Center for Molecular Medicine, Berlin, Germany*

It is still not fully understood to what extent intronic sequences contribute to the regulation of the different forms of alternative splicing. We are interested in the regulation of alternative cassette exon events, such as exon inclusion and exon skipping. We investigated these events by comparative genomic analysis of human and mouse in five experimentally well-characterized genes, neurofibromatosis 1 (*NF1*), cystic fibrosis transmembrane conductance regulator (*CFTR*), period 3 (*PER3*), cysteinyl-tRNA synthetase (*CARS*) and synaptotagmin 7 (*SYT7*). In *NF1*, high intron identity around the 52 constitutive and four alternatively skipped *NF1* exons is restricted to the close vicinity of the exons. In contrast, we found on average high conservation of intron sequences over 300 base pairs up- and downstream of the five alternatively included *NF1* exons. The investigation of alternatively included exons in *CFTR*, *PER3*, *CARS* and *SYT7* supported this finding. In contrast, the mean intron identities around the alternatively skipped exons in *CTFR* and *NF1* do not differ considerably from those around the constitutive exons. In these genes, the difference in intron conservation could point to a difference between the regulation of alternative exon inclusion and alternative exon skipping or constitutive exon splicing. Additional genome-wide investigations are necessary to elucidate to what extent our finding can be generalized.
European Journal of Human Genetics (2004) **12**, 139–149. doi:10.1038/sj.ejhg.5201098
Published online 29 October 2003

Keywords: intron conservation; alternative splicing; alternatively included

Introduction

To generate correct mRNAs in eucaryotes, exons must be identified and joined together precisely and efficiently. In human, this process requires the coordinated action of small nuclear (sn)RNPs interacting with specific DNA motifs. Well-characterized DNA motifs are the 5′ and the 3′ splice sites, the branch site, and exonic splice enhancers and silencers.[1–3] Interestingly, also the promoter seems to play a crucial role in alternative splicing.[4] Several isoforms can be generated by alternative splicing through cassette exon events (ie exon skipping or exon inclusion), exon and intron isoform events (ie use of alternative 3′ or 5′ splice sites) and intron retention.[2,3,5] We are interested in the regulation of alternative cassette exon events, a cassette exon being one included or excluded from the transcript. The terms exon inclusion and skipping are defined by the absence/presence of the respective alternative exon in the predominant transcript, exons always present in the

*Correspondence: Dr D Kaufmann, Department of Human Genetics, University of Ulm, Albert Einstein Allee 11, Ulm D 89070, Germany. Tel: +49 731 500 23419; Fax: +49 731 500 23438;
E-mail: dieter.kaufmann@medizin.uni-ulm.de
Online Mendelian Inheritance in Man (OMIM) (http://www3.ncbi.nlm.nih.gov/Omim/); Promoter 2.0 Homepage (http://www.cbs.dtu.dk/services/Promoter/)
Received 5 March 2003; revised 25 June 2003; accepted 29 August 2003

transcript are termed constitutive.[5] It is still not fully understood to what extent intronic sequences around such exons contribute to the regulation of these forms of alternative splicing. A powerful method to investigate this is the comparative genomic analysis of human and mouse intron sequences.[6] The availability of both mouse and human genomes allows such investigations on a genome-wide level *in silico*.[7] However, there are often inconsistent data concerning the classification of alternative cassette exons. Therefore, we chose very well-characterized genes for our investigation. We started with the neurofibromatosis 1 (*NF1*) gene. *NF1* shows numerous constitutive, alternatively included and alternatively skipped exons. *NF1* is a tumor suppressor gene and responsible for the human disease NF1 [MIM 162200], which is one of the most common familial tumor syndromes.[8,9] In human, 61 *NF1* exons are known.[8–11] There are four alternatively included cassette *NF1* exons, 9br, 10a-2, 23a and 48a. An additional alternatively included exon, 23b, is expressed in mouse.[10] The four *NF1* exons 4b, 29, 30 and 43 are alternatively skipped in human.[11] The pattern of expression of the alternative exons is characterized.[8–11] In addition, there are diverse but rare aberrantly spliced *NF1* transcripts.[12–15] These were not considered alternative in our investigations as they most likely represent splicing noise.

Our investigations demonstrated a high identity of intron sequences over more than 300 base pairs (bp) up- and downstream of the alternatively included exons between human and mouse. In contrast, high intron identity around the constitutive and alternatively skipped *NF1* exons was restricted to the close vicinity of the exons. Investigation of one of the alternatively included *NF1* exons in several species supported the high conservation. We performed similar comparisons with four other experimentally well-characterized genes, the cystic fibrosis transmembrane conductance regulator (*CFTR*) [MIM 602421], the period 3 (*PER3*) [MIM 603427], the cysteinyl-tRNA synthetase (*CARS*) [MIM 123859] and the synaptotagmin 7 (*SYT7*) [MIM 604146]. These investigations supported the high intron identity around alternatively included exons found in *NF1*. This finding points to an unknown function of these sequences in the regulation of splicing of alternatively included exons in these genes.

Materials and methods

Cell culture

Primary human fibroblasts, a cell line derived from a rat schwannoma (*Rattus norvegicus*), fibroblast-like cells derived from mouse (*Mus musculus*), cat (*Felis catus*), dog (*Canis familiaris*), cow (*Bos taurus*), horse (*Equus caballus*), sheep (*Ovis aries*) and griffon vulture (*Gyps fulvus*) were cultured in DMEM containing 10% fetal calf serum. Fibroblasts derived from several primates as the new world monkey Geoffroy's marmoset (*Callithrix geoffroyi*), the old world monkeys gorilla (*Gorilla gorilla*) and orang-utan (*Pongo pygmaeus*), cultured as described above, were kind gifts of W Just, W Krone and W Schemp.

Isolation of cellular RNA and cDNA synthesis

Total cellular RNA was isolated using the RNeasy spin columns (Qiagen, Hilden, Germany). RNA was quantified by measuring the optical density at 260 nm and stored at −70°C. Total RNA (1 μg) was reverse transcribed using the SuperScript™ First-Strand Synthesis system and random hexamers (Invitrogen Life Technologies, Carlsbad, CA, USA). cDNA derived from *Fugu rubripes* was a kind gift of H Kehrer-Sawatzki.

Expression analysis of *NF1* 10a-2

Human cDNA was amplified using the primers Nf9H and Nf10bR (primer sequences see Table 1) resulting in a wild-type amplification product of 306 bp representing sequences of *NF1* exons 9, 10a and 10b. The amplification conditions in the thermal cycler GeneAmp 9600 (Perkin-Elmer, Boston, MA, USA) were denaturation at 94°C for 2 min followed by 30 cycles of 93°C for 30 s, 49°C for 30 s and 72°C for 30 s and a final extension step at 72°C for 10 min. The reamplification was performed as described.[9] In addition, cDNA was amplified using the *NF1* 10a-2-specific primer Nf10a2H and primer Nf10bR, annealing at 46°C, 40 cycles, resulting in a PCR product of 157 bp. cDNA from different species was amplified using the same primer pairs as for human cDNA. *Fugu rubripes* cDNA was tested for the expression of *NF1* 10a-2 with primer pair Fugu-Nf7-H/Fugu-Nf10b-R, annealing at 61°C, yielding a wild-type product of 560 bp. The region where the *NF1* 10a-2 product was expected was gel extracted and reamplified with the same primer pair. Alternatively, a reamplification with a specific forward primer Fugu-Nf10a-2-H and Fugu-Nf10b-R, annealing at 65°C, was performed, resulting in a product of 198 bp.

DNA isolation and amplification from several species

To analyze the *NF1* intron 10a sequence of various species gDNA was isolated from cells either with the DNeasy Kit (Qiagen, Hilden, Germany) or with the salting-out method. PCR products were generated with the primer pairs in Table 1 using system 1 of the Expand Long Template PCR System (Roche Diagnostics, Mannheim, Germany). PCR conditions in a Perkin-Elmer GeneAmp 9700 were: 2 min at 92°C, then 10 cycles of 10 s, 92°C, 30 s annealing (60–65°C), 8 min, 68°C, then 20 cycles of 10 s, 92°C, 30 s annealing, 10 min, 68°C and a final elongation of 10 min at 68°C. Primer pairs for cow and sheep DNA In10aH/In10aR, Cow-In10a-H/In10aR and Cow2-In10a-H/In10aR. Annealing was 65°C for all three pairs. Primer pairs for horse DNA:

Table 1 Primers used for the amplification of sequences containing *NF1* exon 10a-2 and of *NF1* intron 10a in several species

Species	*Name*	*Sequence*
Cat	Cat10aH	5′-CTA GCT AAT GGT GTT TGT TCT TCA-3′
	Cat10aR	5′-TCA ATT GGT TCC ATA CGA GTT TTA-3′
	Cat3-In10a-H	5′-GGC GCA CAC CCA GCA ATA C-3′
	Cat3-In10a-H	5′-GGC GCA CAC CCA GCA ATA C-3′
	Cat3-In10a-H	5′-GCG CAC ACC CAG CAA TAC-3′
	Cat-In10a-H	5′-CAT TGA TTG GTG GTG CTT TGT CTT CT-3′
Cow/sheep	Cow-In10a-H	5′-CAC TTT TCT CAC ACA TAA ACA TTG GGA-3′
	Cow2-In10a-H	5′-AGC TTT CTT TGT TCT ATC AGT GTT CTT CT-3′
Dog	Dog2-In10a-R	5′-CAA CAG CAG CCA ATA AGA ACA GAA-3′
Fugu	Fugu-Nf7-H	5′-CGC CGC CAT CGC CTG TGT C-3′
	Fugu-Nf10b-R	5′-TGG GGT CTG CGT GGA TGA GTT TGA-3′
Horse	Horse3-In10a-H	5′-TAG GCT CTT TGA TAT TGA AGT TTG TGT TT-3′
	Horse3-In10a-R	5′-CCC TTT GTA AGA ATC AGA CAT CAG AAC T-3′
Mouse	Mouse-In10a-R	5′-CAT TTG TAG CTC CTT GTC TCT AGG TCT-3′
	Mouse-In10a-H	5′-AGA TTG TGC CCA TGG TTT CCT TAC TC-3′
Vulture	Vulture-In10a-H	5′-GGC CCA AGA TCG ATG CTG TTT ACT G-3′
	Vulture-In10a-R	5′-CAG CAT GGA TCA GTT TTA CCA AGG ATA A-3′
	Vulture2-In10a-H	5′-TCA TTT GTC TTG CTT GCT TTC TGT TGA G-3′
Human	Nf10aH	5′-CAT TGG ATT GGT GGC CTA AGA-3′
	Nf10a-nested-H	5′-TTG ATG CTG TGT ATT GTC ACT C-3′
	Nf9H	5′-CTG GCT CAG AAT TCA CCT TCT-3′
	Nf10bR	5′-TTA GTT TCA CCA TGG ACA AGA G-3′
	In10aH	5′-TTG GAT TGG TGG CCT AAG ATT GAT GC-3′
	In10aR	5′-TAC TTA TAG CTT CTT TCT CCA GGT CT-3′

In10aH/In10aR, annealing at 65°C and Horse3-In10a-H/Horse3-In10a-R, annealing at 63°C. Primer pairs for mouse DNA: In10aH/Mouse-In10a-R, annealing at 65°C and Mouse-In10a-H/Mouse-In10a-R, annealing at 62°C. Primer pairs for cat DNA: Cat-In10a-H/In10aR, annealing at 63°C. Primer pairs for dog DNA: Cat-In10a-H/In10aR, annealing at 60°C. Primer pairs for vulture DNA: GW-In10a-H/GW-In10a-R and GW2-In10a-H/GW-In10a-R, both annealing at 62°C. PCR products of cat and dog DNA were also generated with Taq DNA polymerase (Amersham Pharmacia Biotech, Uppsala, Sweden) and 30 cycles of 30 s at 92°C, 30 s annealing and 30 s at 72°C. Primer pairs for cat were: (1) Nf10aH/Nf10a2R, annealing at 49°C; (2) Nf10a-nested-H/Nf10a2R, annealing at 52°C and (3) Cat10aH/Cat10aR, annealing at 55°C. Primer pair for dog was: Cat3-In10a-H/Dog2-In10a-R, annealing at 54°C. Intron 9 of the griffon vulture *NF1* gene was amplified with Taq DNA polymerase and primer pair GW-In9-H/GW-In9-R and GW-In9-H3/GW-In9-R, annealing at 65°C. All PCR products were purified by gel extraction or PCR product purification with the GFX kit (Amersham Pharmacia Biotech, Uppsala, Sweden) and sequenced on an ABI 373A automated sequencer (PE Applied Biosystems, Foster City, CA, USA). PCR products from gDNA were sequenced using only the respective forward primers.

Search and criteria for the genes investigated

GenBank (http://www.ncbl.nlm.nih.gov/Genbank/GenbankSearch.html), the AltExtron database (http://www.ebi.ac.uk/asd/altextron/access.html) and PubMed (http://www.ncbi. nlm.nih.gov/entrez/query.fcgi1) were searched for genes with alternatively included cassette exons conserved between human and mouse and ideally surrounded by several hundred base pairs of intronic sequences. Of these, the majority of the genomic sequence of human and mouse or rat had to be available and they should contain at least 10 exons and sufficiently sized introns to ensure enough data for statistical analysis.

Comparative sequence analysis of intron conservation

Alignment of the *NF1* intron 10a sequences of various species to the human sequence was carried out using the BestFit agorithm (GCG, Accelrys, Burlington, MA, USA) with the most stringent gap creation and gap extension penalties yielding a complete alignment. The rat intron 10a sequence was obtained from the Rat Genome Project (http://www.hgsc.bcm.tmc.edu/projects/rat/). For the calculation of mean intron identities around all constitutive *NF1* exons except exon 1, the human and mouse *NF1* intron sequences together with the flanking exons were aligned with BestFit. Alignment was carried out in sections

of one intron plus the surrounding exons. The *NF1* sequences of human and mouse were obtained from GenBank. *Nf1* exon 1 and intron 1 have not yet been sequenced in mouse. The identities of 50 nucleotides (nt) windows 350 bp up- and downstream of the respective exons were calculated. Statistical analysis of the difference between the graphs of the mean values of alternatively included, alternatively skipped and constitutively spliced exons was performed with the paired *t*-test. The alignments of the *NF1* intron 10a sequences were used to calculate the identities for 50 nt windows 600 bp up- and downstream of exon 10a-2, which lies 764 bp downstream of exon 10a in the 8 kb sized intron 10a. To generate random sequences with the same dinucleotide frequency as the *NF1* gene, the complete human genomic *NF1* sequence was randomized with Shuffle from the GCG package under the preservation of the dinucleotide frequency. Five sequences of 1250 nt each were then taken from the resulting sequence and compared to the human sequence of the five alternatively included and five randomly chosen constitutive *NF1* exons and 600 bp of intron sequence up- and downstream as described above. The identities for 50 nt windows were calculated and averaged for all alternatively included exons ($n_{\text{windows}} = 683$) and the five constitutive exons ($n_{\text{windows}} = 725$). Likewise, five random sequences of 1250 nt each, representing each base equally, were compared to the same *NF1* sequences described above. Again, the identities for 50 nt windows were calculated and averaged for all alternatively included exons ($n_{\text{windows}} = 649$) and the five constitutive exons ($n_{\text{windows}} = 617$). Alignment of the cat and mouse *CFTR* sequences to the human sequence was carried out as described for *NF1* except for very large introns or introns extremely differing in size where only smaller parts of the sequences were aligned. Multiple alignment of the previously unknown *CFTR* intron 10 element of various species was carried out with Dialign (http://www.genomatix.de/cgi-bin/diallgn/dialign.pl). *PER3* and *WT1* sequences were also obtained from GenBank, *CARS* and *SYT7* sequences from the Celera Discovery System and the alignment carried out as described for *NF1*.

Search for intron sequence elements

In all, 600 bp up- and downstream of all alternative *NF1* and *CFTR* exons investigated were examined for several known intronic sequence elements with FindPatterns or BestFit of the GCG package. We tested for the following pattern sequences: Nova-1 binding sequence;[16] (A/U)GGG repeats;[17] PTB binding site;[18] SF1 binding site;[19] UGCAUG repeats;[20] hnRNP A1 binding site;[21] multiple copies of UGC;[22] GAR repeats;[23] CE4m repressor;[24] IAS 2 and IAS 3;[25] 5′ splice site consensus;[26] downstream control sequence;[27] conserved intronic element;[28] and muscle-specific enhancer.[29] The conserved intron sequences surrounding the alternative exons were examined for RNA polymerase II promoter sequences with CorePromoter (http://argon.cshl.org/genefinder/CPROMOTER/index.htm),[30] Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) and Promoter 2.0 (http://www.cbs.dtu.dk/services/Promoter/).[31] Search for RNA polymerase III promoter sequences was carried out with FindPatterns, including the A and B box consensus sequences of the type II promoter, the TATA box, PSE, DSE and Sp1 consensus[32] elements of the type III promoter. Sequences characteristic of snoRNAs, such as the C, D and H box consensus sequences[33] were also searched for with FindPatterns. The intron sequences containing alternative exons were searched for open reading frames in sense and antisense orientation with the algorithm Frames of the GCG package.

Accession numbers

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = Nucleotide, NCBI GenBank Entrez Nucleotide (for *NF1* intron 10a sequenced by us [AF457133] to [AF457138]; for human *NF1* [NT_010799]; for mouse *Nf1* exons 2–27b [AL591126]; for mouse *Nf1* exons 28–49 [AC008161]; for human *CFTR* [AC000111] and [AC000061]; for cat *CFTR* [AC091436] and [AC091382]; for mouse *Cftr* [AF162137]; for Pan troglodytes *CFTR* [AC087834]; for *C. familiaris CFTR* [AC091119]; for *R. norvegicus CFTR* [AC091268]; for *B. taurus CFTR* [AC089993]; for *Sus scrofa CFTR* [AC092478]; for human *PER3* [Z98884]; for mouse *Per3* [AL607143]; for human *WT1* [AL049692]; for mouse *Wt1* [AL5125841]). https://industry.ebi.ac.uk/altExtron/, AltExtron Database Project (for alternative *PER3* exon [IDB61710]). http://www.hgsc.bcm.tmc.edu/projects/rat/, Rat Genome Project (for rat *NF1* intron 10a [project_gjdy_BT]). http://www.celeradiscoverysystem.com/index.cfm, Celera Discovery System (for human *CARS* [hCG16441]; for mouse *CARS* [mCG15274]; for human *SYT7* [hCG40863]; for mouse *SYT7* [mCG1938]).

Results

Intron sequence conservation between human and mouse is higher around alternatively included than constitutive *NF1* exons

We compared *NF1* sequences from human and mouse using the BestFit algorithm. First, we determined the basic identity generated by this algorithm. Random sequences representing the same dinucleotide frequency as the *NF1* gene were generated. We aligned these random sequences to the human intron sequences around the five randomly chosen constitutively spliced *NF1* exons 2, 12b, 31, 39 and 41, and around the five alternatively included *NF1* exons 9br, 10a-2, 23a, 23b and 48a. The investigation yielded an average identity of $46.5 \pm 0.28\%$ (SEM) per 50 nt window for the constitutively spliced exons and 46.2 ± 0.31 (SEM) for the alternatively included exons. We considered this

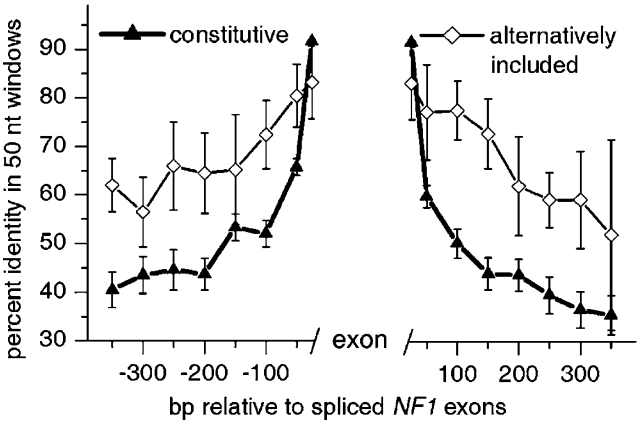

Figure 1 Comparison of the intron sequences around alternatively included and constitutive *NF1* exons between human and mouse. Mean identities of 50 nt windows 350 bp up- and downstream of 52 constitutively spliced *NF1* exons (black triangles, error bars: SEM) are lower than those of all five alternatively included *NF1* exons (white rhombuses, error bars: SEM).

identity the basic identity for our investigations. Intron alignment of 52 constitutive *NF1* exons of human and the corresponding exons of mouse yielded on average high identity (>60%) only in close vicinity to the exons. The identity declined to values nearing the basic identity at a distance of 100 bp from the exon–intron boundary (Figure 1). In contrast to the constitutively spliced exons, the mean identity of the intron sequences surrounding the alternatively included *NF1* exons is higher for every 50 nt window of the 350 bp (Figure 1). The mean values of the two graphs differ significantly (paired *t*-test, $P<0.001$). The individual alternatively included *NF1* exons show differences in intron conservation (data not shown). We found the most extended region of high identity around *NF1* exon 23a, followed by exons 10a-2, 23b and 9br. The intron sequences around the poorly conserved *NF1* exon 48a are not remarkably identical except for a sequence stretch 100 bp upstream of exon 48a, which may represent an unknown essential element (see below and Table 2). Even if the exceptionally conserved sequences surrounding exon 23a are not considered, the mean values of the intron sequences surrounding the remaining four exons differ significantly from the mean values of the introns of the constitutively spliced exons (paired *t*-test, $P<0.001$). The high intron homology around exon 23b is remarkable because expression of this exon is found in mouse but not yet in human.[8,15] We suggest that it is expressed in specific situations not yet investigated in human. The investigation of the human sequence of exon 23b and its surrounding intronic sequences showed that the splice sites are almost 100% identical. Only the reading frame is altered due to a 1 bp deletion in the human sequence, which would lead to an earlier premature stop codon as in mouse.

High conservation of intron sequences surrounding the alternatively included *NF1* exon 10a-2 in several species

The observation of the high intron homology around the *NF1* exon 10a-2 between human and mouse is also remarkable because we found the expression of this exon to be lost in mouse. The invariant T of the 5′ splice site consensus is altered to G, the putative branch point shows a single base exchange and the mouse exon 10a-2 homologous sequence comprising 40 bp would lead to a premature stop codon formed by the last base of exon 10a-2 and the first two bases of exon 10b (Figure 2). The high intron homology without the expression of the exon could on the one hand indicate that the expression was lost only recently in mouse leaving too little time for the accumulation of mutations in the intron or on the other hand that an active constraint independent of exon 10a-2 prevents mutation of the intronic sequences. To elucidate this, we investigated several species with different times of divergence (eg vulture–human 310 million years ago, mouse, rat, dog, cat, horse–human 90 million years ago[34]) for exon 10a-2 presence, expression and intron identity (Figures 2 and 3). We detected the first reliable expression of exon 10a-2 in bird (griffon vulture). The homology of the bird to the human 10a-2 sequence is high as shown by sequencing of cDNA and gDNA (Figure 2). In intron 10a of the *NF1* gene of the pufferfish *Fugu rubripes*, we identified a nonexpressed 62 bp sequence with 50% identity to the

Table 2 Previously unknown conserved sequence elements observed in the intron sequences surrounding alternatively included exons

Exon	*Location relative to exon*	*Sequence*	*Identity to human*
NF1 10a-2	+400 bp	GTT TTG GGG ATG AGT AAG GGA AGC TGA CTC CTG GGT TAG AGT GAA TGT T	Cat, horse: 93.3%; dog: 81.6%; mouse: 46.9%; vulture: 54%
NF1 48a	+99 bp	ATA ATT AAA ACC AGA TTC CTT CTG AAA ACC A	Mouse: 87.1%
CFTR 10b	−11 bp	GCA CAA CAT ATT TCA CAT AGT TTT CTG ATT TCA GT	Old world monkeys: 100%; Mouse, rat, cat: 100%; cow: 94.3% dog: 91.4%; pig: 88%

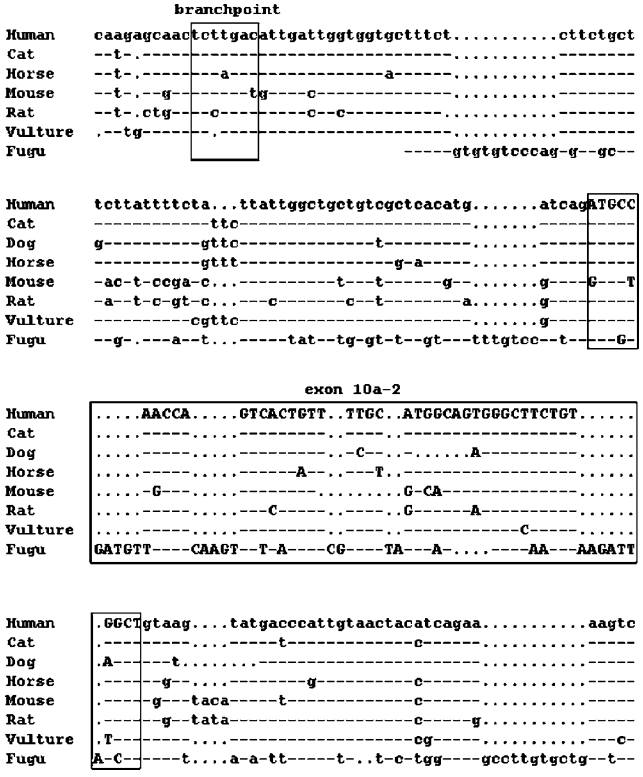

Figure 2 Sequence homology between *NF1* exon 10a-2 of human and cat, dog, horse, mouse, rat, vulture and Fugu. Nucleic acid sequence alignment of *NF1* 10a-2 and surrounding intron sequences between several species. The human sequence is shown in capital letters. A match to the human sequence is represented by a dash, a gap by dots. The exon 10a-2 homologous sequence and the putative branch point are highlighted boxes.

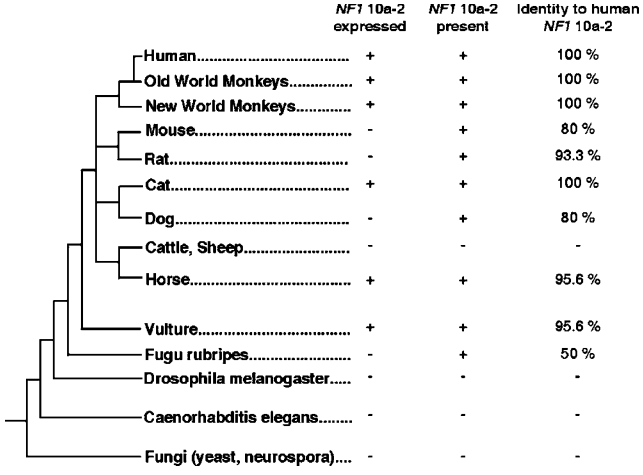

Figure 3 Phylogenetic tree showing the relationship between the species investigated for alternatively included *NF1* exon 10a-2. For each species the information whether exon 10a-2 is expressed and/or present in the genomic DNA sequence is shown, and, if present, its identity to the human sequence.

human *NF1* 10a-2 (Figure 2). We also found highly conserved exon 10a-2 homologous sequences in horse, dog, cat, rat and all primates investigated. Of these species, 10a-2 is expressed in horse, cat and primates. We did not find 10a-2 expression in rat, although the splice sites are both intact, the homology to the human exon 10a-2 is very high (93.3 %) and it does not lead to an alteration of the reading frame (Figure 2). Only the putative branch point shows a 1 bp alteration different from that seen in mouse. It seems therefore likely that the rat exon 10a-2 is expressed but in cell types not investigated by us. As even rat does not harbor the same alterations as mouse and the divergence of rat and mouse is estimated at 40 million years ago,[34] the loss of exon 10a-2 expression is clearly a very recent event in mouse. We therefore assume that the time span since this loss was insufficient to greatly alter the surrounding intronic sequences. The alignment of the introns surrounding exon 10a-2 of those species with this exon to the human introns revealed a very high identity over several hundred base pairs for cat and horse (Figures 2 and 4). Also rat and dog show remarkably high intron identity, although not expressing exon 10a-2. In vulture, high identity is restricted to about 100 bp up- and downstream

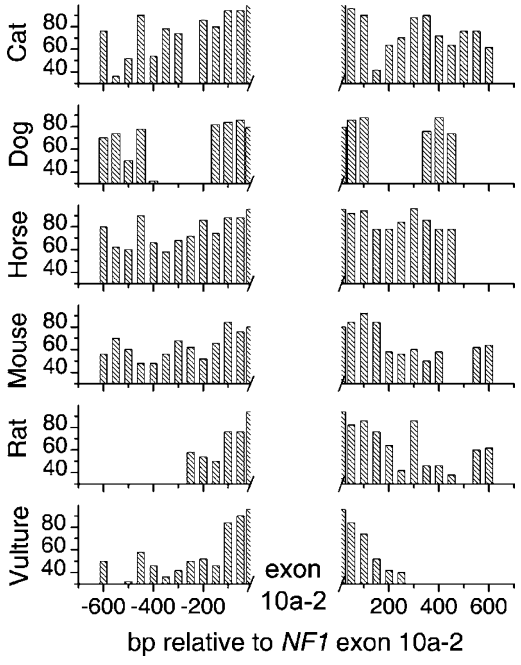

Figure 4 Comparison of *NF1* exon 10a-2 and surrounding intron sequences between human and several species. The identities of 50 nt windows of alignments between the human intron 10a sequence and that of several other species (cat, dog, horse, mouse, rat and vulture) were calculated for 600 bp up- and downstream of exon 10a-2. The identities are displayed as columns. Gaps in the alignment due to insertions/deletions are shown as an identity of zero.

of the exon. Taken together, we could confirm the high conservation of the intron sequences surrounding exon 10a-2 observed in mouse in several other species and we could show that the loss of exon 10a-2 expression is a very recent event in mouse.

Higher conservation of intron sequences surrounding alternatively included *CFTR, PER3, CARS* and *SYT7* exons

To investigate if high intron identity around alternatively included exons is a more general phenomenon, we examined four other very well-characterized genes fitting our criteria (see Materials and methods), the *CFTR*, the *PER3*, the *CARS* with one alternatively included exon each and the *SYT7* with five alternatively included exons. The intron identities around the conserved alternatively included *CFTR* exon 10b[35] are higher than that of the average intron identities around the 16 constitutively spliced *CFTR* exons between human and cat (Figure 5a) and for the most part also between human and mouse (data not shown). *PER3* comprises 21 exons in human homologous to exons 2–22 of the mouse *mPer3*[36] and contains one alternatively included exon, which we have designated 2a. The comparison revealed intron identities higher than the average identities around the constitutive exons downstream of exon 2a, while upstream only the identity of the first 50 nt window is higher than the average (Figure 5b). *CARS* is composed of 23 exons in human and mouse, 20 of which are constitutively spliced.[37] The alternatively included exon 2 shows as the *PER3* exon mainly intron identities higher than the average identities around the constitutive exons downstream of the exon (Figure 5c). *SYT7* comprises 14 exons, five of which, exons 4–8, are alternatively included.[38] The average intron identities around all alternatively included exons are clearly higher than the average identities around the constitutive exons

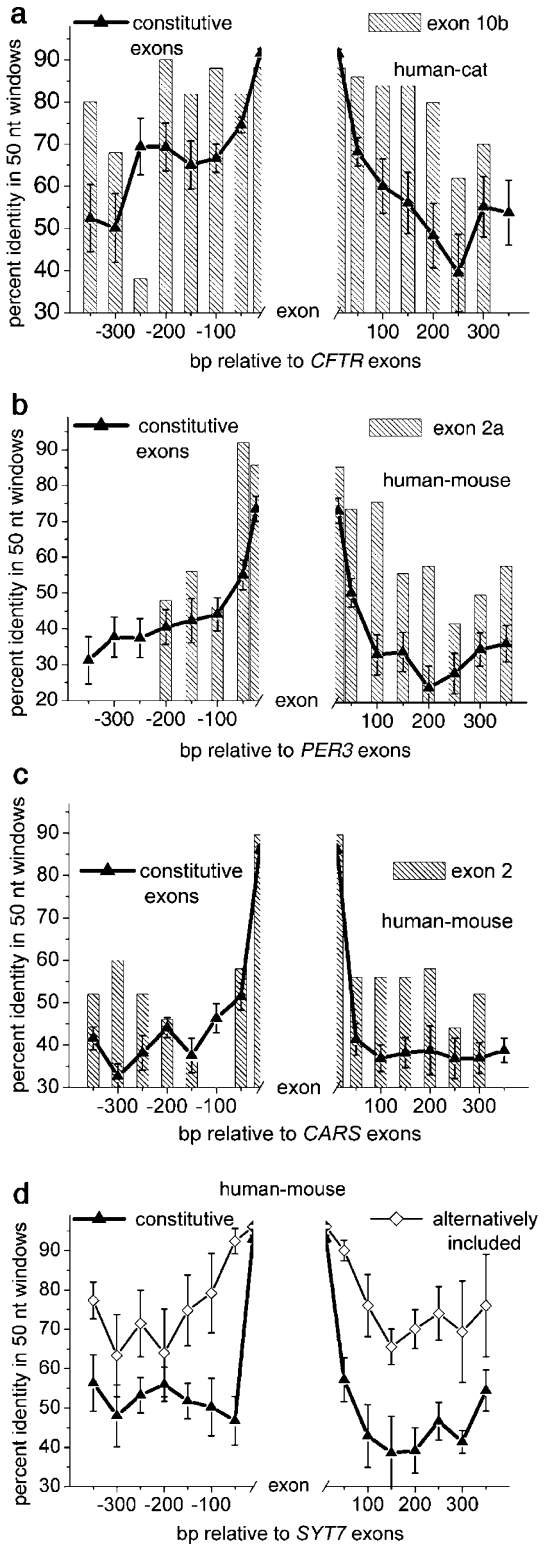

Figure 5 Intron conservation around alternatively included *CFTR*, *PER3*, *CARS* and *SYT7* exons in comparison to the intron conservation of the respective constitutively spliced exons. (a) Average intron identities of all constitutive *CFTR* exons (black triangles) in comparison to the intron identities around *CFTR* exon 10b (columns) between human and cat. The generally higher similarity of not transcribed sequences is surprising, because the evolutionary distance of these species is in the same order of magnitude as that of man and mouse. It somehow parallels the similarity of the karyotypes, which has been found to be close to the ancestral mammalian chromosome arrangement in both cases.[49] (b) Average intron identities of all constitutive *PER3* exons (black triangles) in comparison to the intron identities around *PER3* exon 2a (columns) between human and mouse. (c) Average intron identities of all constitutive *CARS* exons (black triangles) in comparison to the intron identities around *CARS* exon 2 (columns) between human and mouse. (d) The average intron identities of the nine constitutive *SYT7* exons (black triangles) in comparison to the average intron identities around the five alternatively included *SYT7* exons (white rhombuses) between human and mouse.

(Figure 5d). Taken together, this points to high intron identity around alternatively included exons being a more general phenomenon.

Restricted intron conservation around alternatively skipped *NF1* and *CFTR* exons

Then, we investigated the introns surrounding alternatively skipped cassette exons. The *NF1* exons 4b, 29, 30 and 43[11,12] and the *CFTR* exons 2, 3, 4, 5, 9, 11, 12 and 14a[39] are alternatively skipped. Interestingly, the mean intron identities around these skipped exons do not differ considerably from those around the constitutive exons (Figures 6a and b). This shows a difference in the conservation of surrounding intron sequences between alternatively included and alternatively skipped exons. Our finding is supported by the investigation of the Wilms' tumor 1 gene (*WT1*) consisting of eight constitutive exons and one alternatively skipped exon.[40] The mean intron identities around the skipped exon 5 do not differ from those around the constitutive exons (data not shown).

Other differences between alternatively included and skipped exons

Recently, other differences between alternatively included and skipped cassette exons have been reported as smaller size of the included exons.[5] The five alternatively included *NF1* exons are indeed much smaller (mean 46.4 bp) than the four skipped *NF1* exons (mean 193 bp). The one alternatively included *CFTR* exon (119 bp), however, is not that much shorter than the average of the eight skipped exons (127.5 bp). Small size does not seem to correlate with high intron identity, as in *NF1* the identities around the three smallest constitutive small exons 4c (68 bp), 35 (62 bp) and 47 (47 bp) do on average not differ from those found for the constitutive exons (data not shown). It was also shown that the donor splice sites of included cassette exons show a stronger consensus to A at position +3 (A3) than alternatively skipped and constitutive exons.[1,5] However, we see no differences between the six included (all A3) and 12 skipped (11 A3, 1 G3) alternative exons investigated in *NF1* and *CFTR*.

Known elements and unknown conserved sequences in the introns surrounding alternatively included *NF1* and *CFTR* exons

The high intron identity found around alternatively included exons raised the question about the function of these sequences. The investigation of the introns around the alternatively included *NF1* and *CFTR* exons for several known intronic splice elements mostly present in several copies revealed several matches conserved between human and mouse but no correlation with the conserved sequences. Also, the investigation of the introns showed

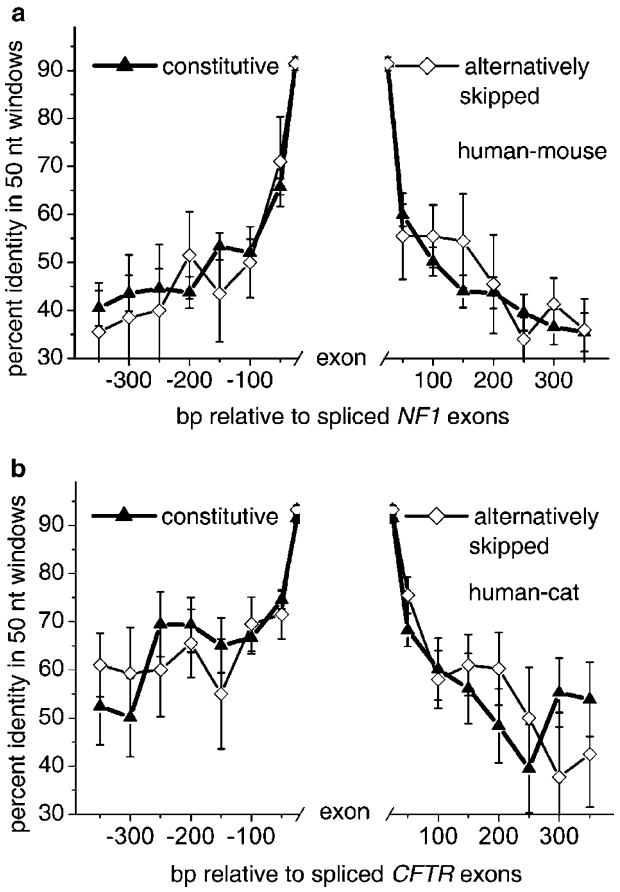

Figure 6 The intron conservation around alternatively skipped *NF1* and *CFTR* exons is not higher than that around the constitutive exons. (a) The mean identities of 50 nt windows 350 bp up- and downstream of all four alternatively skipped *NF1* exons (white rhombuses; error bars: SEM) between human and mouse do not differ significantly (paired *t*-test: $P > 0.9999$) from those of 52 constitutively spliced *NF1* exons (black triangles; error bars: SEM). (b) The mean identities of 50 nt windows 350 bp up- and downstream of all eight alternatively skipped *CFTR* exons (white rhombuses; error bars: SEM) between human and cat and of 16 constitutively spliced *CFTR* exons (black triangles; error bars: SEM) also do not differ significantly (paired *t*-test: $P > 0.939$).

several previously unknown small elements extremely high conserved between mouse and human. Such elements were found around three of the six alternatively included exons (Table 2). The comparison of the conserved intronic sequences around *NF1* exons 9br, 10a-2, 23a, 23b and 48a with each other revealed no common motifs.

Taken together, we assume that the few identified conserved elements cannot account for the extent of the high intron conservation. Therefore, we also performed *in silico* search for evidence indicating the presence of other genes in the conserved intron regions. However, *in silico*

search for RNA polymerase II promoter sequences correlating with the conserved regions was inconclusive as different algorithms yielded different putative promoter locations or none at all. Also, no RNA polymerase III promoter sequences and open reading frames in sense and antisense orientation were identified and no matches to elements characteristic for snoRNAs were found.

Discussion

We describe here a high identity between human and mouse of intron sequences surrounding most alternatively included coding exons of *NF1* and four other genes investigated, but not the average of the constitutive exons. Our finding raises the question if this conservation is a general phenomenon. High intron conservation is described around the alternatively included exons Ve of *CDK2* and 5a of *Pax6*,[41,42] but the average conservation of the introns surrounding the constitutive exons of the respective genes was not determined. The complete sequencing of human and mouse genomes greatly facilitated comparisons of whole genes. However, the number of homologous genes of human and mouse with conserved alternatively included exons is restricted, because genome-wide investigation showed that the alternative exon inclusion comprises only approximately 15% of all alternative splicing events in contrast to exon skipping, which amounts to 48% approximately.[4] In addition, there are often inconsistent data concerning the classification of alternative cassette exons as skipped or included, for example, present or absent in the predominant transcript. Several genome-wide comparisons of expressed sequence tags or full-length cDNA clones with the genomic draft of human and mouse have been carried out to investigate the extent of alternative splicing in the respective genome.[43,44] It suggests itself to use these databases for a genomic comparison of alternatively spliced genes of human and mouse.[7,45] Such detailed comparisons will shed more light on the difference between the alternative splice forms and elucidate to what extent our finding can be generalized or if it is specific for the *NF1* and the other investigated genes. Our observation of highly conserved intron sequences around alternatively included exons in *NF1* and other genes raises the question of the reason for this conservation. The idea of a regulatory function in the expression of the respective exon suggests itself. Alternatively included exons have been reported to be smaller than the average.[5] This could mean that they need more help from accessory splicing factors and therefore more respective binding elements, in exon or intron, to ensure accurate splicing. So, first, there could be an accumulation of intronic splice elements. We found several highly conserved small elements, some of which also possess sequences of known intronic splice enhancer/silencer elements. But, we did not find a general correlation between known splice elements and the conserved intron sequences. Second, regulation of splicing may be influenced by the secondary structure of the pre-mRNA as found in the rat *CGRP* exon 4.[46] This possibility will have to be clarified experimentally. Third, the expression of genes transcribed in antisense orientation may regulate alternative exon inclusion. Such a regulation has been described *in vivo* for the *FGF-2*[47] and *HFE* genes.[48] We could identify some putative promoter elements correlating with the conserved regions *in silico*. But, since the reliability of the available prediction algorithms is not very high, this possibility remains to be tested experimentally. In addition, it is conceivable that yet unknown genes are lying in sense orientation in the regions of high identity. But in spite of a thorough investigation, we did not find an indication for this. Our finding may also be relevant for mutation screening in disease-causing genes with alternatively included exons as *NF1*. The present screening methods often do not find all mutations in *NF1*. A reason for this may be that some mutations lie in such conserved intronic regions, which are often not investigated in detail. In summary, we found differences in the conservation between introns around alternatively included and introns around alternatively skipped and constitutive exons in the *NF1* and four other genes. This could point to a difference in the regulation of exon inclusion and the regulation of constitutive splicing and alternative exon skipping.

Note

During the submission of this article, Modrek *et al* (*Nat Genet* 2003; **34**: 177–180) found by a genome-wide scan a difference in conservation of the sequences of alternatively included exons and constitutive or skipped exons between human and mosue or rat.

Acknowledgements

We thank D Viskochil for critical discussion concerning our observations. The technical assistance of A Siegel, B Dieske, H Goetz and E Winkler is also gratefully acknowledged. Furthermore, we thank H Lattke, who died recently and R Mueller for technical advice. This work was supported by the Deutsche Krebshilfe (BB, BD) and Graduate College 460 of University Ulm (OK).

References

1 Zhang MQ: Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet* 2002; **3**: 698–709.
2 Cartegni L, Chew SL, Krainer AR: Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 2002; **3**: 285–298.
3 Ladd AN, Cooper TA: Finding signals that regulates alternative splicing in the postgenomic era. *Genome Biol* 2002; **3**: (reviews0008.1-0008.16.).

4 Nogués G, Kadener S, Cramer P, Bentley D, Kornblihtt AR: Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem* 2002; **277**: 43110–43114.
5 Clark F, Thanaraj TA: Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet* 2002; **11**: 451–464.
6 Loots GG, Locksley RM, Blankespoor CM *et al*: Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 2000; **288**: 136–140.
7 Mouse Genome Sequencing Consortium: Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; **420**: 520–562.
8 Shen MH, Harper PS, Upadhyaya M: Molecular genetics of neurofibromatosis type 1 (NF1). *J Med Genet* 1996; **33**: 2–17.
9 Cichowski K, Jacks T: NF1 tumor suppressor gene function: narrowing the GAP. *Cell* 2001; **104**: 593–604.
10 Kaufmann D, Muller R, Kenner O *et al*: The N-terminal splice product *NF1-10a-2* of the *NF1* gene codes for a transmembrane segment. *Biochem Biophys Res Commun* 2002; **294**: 496–503.
11 Mantani A, Wakasugi S, Yokota Y, Abe K, Ushio Y, Yamamura K: A novel isoform of the neurofibromatosis type-1 mRNA and a switch of isoforms during murine cell differentiation and proliferation. *Gene* 1994; **148**: 245–251.
12 Vandenbroucke I, Vandesompele J, De Paepe A, Messiaen L: Quantification of NF1 transcripts reveals novel highly expressed splice variants. *FEBS Lett* 2002; **522**: 71–76.
13 Wimmer K, Eckart M, Rehder H, Fonatsch C: Illegitimate splicing of the NF1 gene in healthy individuals mimics mutation-induced splicing alterations in NF1 patients. *Hum Genet* 2000; **106**: 311–313.
14 Kaufmann D, Leistner W, Kruse P *et al*: Aberrant splicing in several human tumors in the tumor suppressor genes neurofibromatosis type 1, neurofibromatosis type 2, and tuberous sclerosis 2. *Cancer Res* 2002; **62**: 1503–1509.
15 Thomson SA, Wallace MR: 2002) RT-PCR splicing analysis of the NF1 open reading frame. *Hum Genet* 2002; **110**: 495–502.
16 Jensen KB, Dredge BK, Stefani G *et al*: Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* 2000; **25**: 359–371.
17 Sirand-Pugnet P, Durosay P, Brody E, Marie J: An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. *Nucleic Acids Res* 1995; **23**: 3501–3507.
18 Chou MY, Underwood JG, Nikolic J, Luu MH, Black DL: Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Mol Cell* 2000; **5**: 949–957.
19 Carlo T, Sierra R, Berget SM: A 5′ splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol Cell Biol* 2000; **20**: 3988–3995.
20 Modafferi EF, Black DL: A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. *Mol Cell Biol* 1997; **17**: 6537–6545.
21 Chabot B, Blanchette M, Lapierre I, La Branche H: An intron element modulating 5′ splice site selection in the hnRNP A1 pre-mRNA interacts with hnRNP A1. *Mol Cell Biol* 1997; **17**: 1776–1786.
22 Jin W, Bi W, Huang ESC, Cote GJ: Glioblastoma cell-specific expression of fibroblast growth factor receptor-1β requires an intronic repressor of RNA splicing. *Cancer Res* 1999; **59**: 316–319.
23 Pret AM, Balvay L, Fiszman MY: Regulated splicing of an alternative exon of beta-tropomyosin pre-mRNAs in myogenic cells depends on the strength of pyrimidine-rich intronic enhancer elements. *DNA Cell Biol* 1999; **18**: 671–683.
24 Blanchette M, Chabot B: Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *EMBO J* 1999; **18**: 1939–1952.
25 del Gatto F, Plet A, Gesnel MC, Fort C, Breathnach R: Multiple interdependent sequence elements control splicing of a fibroblast growth factor receptor 2 alternative exon. *Mol Cell Biol* 1997; **17**: 5106–5116.
26 Lou H, Yang Y, Cote GJ, Berget SM, Gagel RF: An intron enhancer containing a 5′ splice site sequence in the human calcitonin/calcitonin gene-related paptide gene. *Mol Cell Biol* 1995; **15**: 7135–7142.
27 Markovtsov V, Nikolic JM, Goldman JA, Turck CW, Chou MY, Black DL: Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Mol Cell Biol* 2000; **20**: 7463–7479.
28 Standiford DM, Sun WT, Davis MB, Emerson Jr CP: Positive and negative intronic regulatory elements control muscle-specific alternative exon splicing of drosophila myosin heavy chain transcripts. *Genetics* 2001; **157**: 259–271.
29 Cooper TA: Muscle-specific splicing of a heterologous exon mediated by a single muscle-specific splicing enhancer from the cardiac troponin T gene. *Mol Cell Biol* 1998; **18**: 4519–4525.
30 Zhang MQ: Identification of human gene core promoters *in silico*. *Genome Res* 1998; **8**: 319–326.
31 Knudsen S: Promoter2.0: for the recognition of Pol II promoter sequences. *Bioinformatics* 1999; **15**: 356–361.
32 Boyd DC, Turner PC, Watkins NJ, Gerster T, Murphy S: Functional redundancy of promoter elements ensures efficient transcription of the human 7SK gene *in vivo*. *J Mol Biol* 1995; **253**: 677–690.
33 Weinstein LB, Steitz JA: Guided tours: from precursor snoRNA to functional snoRNA. *Curr Opin Cell Biol* 1999; **11**: 378–384.
34 Hedges SB, Kumar S: Vertebrate genomes compared. *Science* 2002; **297**: 1283–1285.
35 Melo CA, Serra C, Stoyanova V *et al*: Alternative splicing of a previously unidentified CFTR exon induces an in-frame stop codon 5′ of the R region. *FEBS Lett* 1993; **329**: 159–162.
36 Zylka MJ, Shearman LP, Weaver DR, Reppert SM: Three period homologs in mammals: differential light responses in the suprachiasmatic circadian clock and oscillating transcripts outside of brain. *Neuron* 1998; **20**: 1103–1110.
37 Kim JE, Kim KH, Lee SW, Seol W, Shiba K, Kim S: An elongation factor-associating domain is inserted into human cysteinyl-tRNA synthetase by alternative splicing. *Nucleic Acids Res* 2000; **28**: 2866–2872.
38 Fukuda M, Ogata Y, Saegusa C, Kanno E, Mikoshiba K: Alternative splicing isoforms of synaptotagmin VII in the mouse, rat and human. *Biochem J* 2002; **365**: 173–180.
39 Hull J, Shackleton S, Harris A: Analysis of mutations and alternative splicing patterns in the CFTR gene using mRNA derived from nasal epithelial cells. *Hum Mol Genet* 1994; **3**: 1141–1146.
40 Haber DA, Sohn RL, Buckler AJ, Pelletier J, Call MK, Housman DE: Alternative splicing and genomic structure of the Wilms tumor gene WT1. *Proc Natl Acad Sci USA* 1991; **88**: 9618–9622.
41 Ellenrieder C, Bartosch B, Lee GY *et al*: The long form of CDK2 arises via alternative splicing and forms an active protein kinase with cyclins A and E. *DNA and Cell Biol* 2001; **20**: 413–423.
42 Jaworski C, Sperbeck S, Graham C, Wistow G: Alternative splicing of Pax6 in bovine eye and evolutionary conservation of intron sequences. *Biochem Biophys Res Commun* 1997; **240**: 196–202.
43 Modrek B, Resch A, Grasso C, Lee C: Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* 2001; **29**: 2850–2859.
44 Zavolan M, Van Nimwegen E, Gaasterland T: Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res* 2002; **12**: 1377–1385.
45 Xuan Z, Wang J, Zhang MQ: Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol* 2002; **4**: R1.

46 Coleman TP, Roesser JR: RNA secondary structure: an important cis-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry* 1998; **37**: 15941–15950.
47 Li AW, Murphy PR: Expression of alternatively spliced FGF-2 antisense RNA transcripts in the central nervous system: regulation of FGF-2 mRNA translation. *Mol Cell Endocrinol* 2000; **170**: 233–242.
48 Thenie AC, Gicquel IM, Hardy S *et al*: Identification of an endogenous RNA transcribed from the antisense strand of the HFE gene. *Hum Mol Gen* 2001; **10**: 1859–1866.
49 Rettenberger G, Klett C, Zechner U *et al*: ZOO-FISH analysis: cat and human karyotypes closely resemble the putative ancestral mammalian karyotype. *Chromosome Res* 1995; **13**: 479–486.