The Contribution of Exon-Skipping Events on Chromosome 22 to Protein Coding Diversity

Winston A. Hide,[1,3] Vladimir N. Babenko,[1,2] Peter A. van Heusden, Cathal Seoighe, and Janet F. Kelso

South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

Completion of the human genome sequence provides evidence for a gene count with lower bound 30,000–40,000. Significant protein complexity may derive in part from multiple transcript isoforms. Recent EST based studies have revealed that alternate transcription, including alternative splicing, polyadenylation and transcription start sites, occurs within at least 30–40% of human genes. Transcript form surveys have yet to integrate the genomic context, expression, frequency, and contribution to protein diversity of isoform variation. We determine here the degree to which protein coding diversity may be influenced by alternate expression of transcripts by exhaustive manual confirmation of genome sequence annotation, and comparison to available transcript data to accurately associate skipped exon isoforms with genomic sequence. Relative expression levels of transcripts are estimated from EST database representation. The rigorous *in silico* method accurately identifies exon skipping using verified genome sequence. 545 genes have been studied in this first hand-curated assessment of exon skipping on chromosome 22. Combining manual assessment with software screening of exon boundaries provides a highly accurate and internally consistent indication of skipping frequency. 57 of 62 exon skipping events occur in the protein coding regions of 52 genes. A single gene, (*FBXO7*) expresses an exon repetition. 59% of highly represented multi-exon genes are likely to express exon-skipped isoforms in ratios that vary from 1:1 to 1:>100. The proportion of all transcripts corresponding to multi-exon genes that exhibit an exon skip is estimated to be 5%.

Gene expression products can have variable forms, characterized by alternate start sites of transcription and polyadenylation (Gautheret et al. 1998), exon skipping, and alternate donor and acceptor sites at exon boundaries (Mironov et al. 1999a; Brett et al. 2000; Croft et al. 2000). Exon skipping in transcript isoforms is the most frequent event altering the protein coding sequence of genes (Mironov et al. 1999b; International Human Genome Sequencing Consortium 2001) (http://industry.ebi.ac.uk/~thanaraj/gene.html). Surveys of the incidence of alternative splicing, including exon skipping, have been performed (Andreadis et al. 1987; Iida 1997; Valentine 1998; Thanaraj 1999), and a growing number of anecdotal observations confirm the utilization of exon-skipped transcripts in developmental (Dufour et al. 1998; Lambert de Rouvroit et al. 1999; Lim et al. 1999; Unsworth et al. 1999; Kawahara et al. 2000), tissue-specific (Zacharias et al. 1995), and disease-specific (Jiang and Wu 1999; Mercatante and Kole 2000; Strehler and Zacharias 2001) states.

Several approaches have successfully used hybridization experiments both in silico (Wolfsberg and Landsman 1997; Gautheret et al. 1998; Mironov et al. 1999b; Thanaraj 1999; Beaudoing et al. 2000; Brett et al. 2000; Croft et al. 2000; Schweighhoffer 2000; International Human Genome Sequencing Consortium 2001) and in vitro (Schweighhoffer 2000; Strehler and Zacharias 2001) to assess alternate transcript diversity. Nevertheless there are difficulties with interpretation of the results that include (1) the existence of gene families, paralogs, gene copies, and pseudogenes that have similar DNA sequences, providing false positive hybridization; (2) the existence of orphan genes that are located in the complementary strand of intronic or flanking regions (Mironov et al. 1999a); (3) insufficient representation of expressed sequence data in public expressed sequence tag (EST) databases to identify all transcript isoforms.

We have taken an exhaustive approach to the detection of exon skipping from carefully annotated, protein-confirmed genes to maximize the accurate assessment of the degree of isoform diversity.

RESULTS

To develop an unambiguous assessment of the degree to which exon skipping contributes to expressed transcript isoform diversity, and to assess the impact on protein coding of exon-skipping events within coding regions of transcripts from known genomic loci, we have compared ESTs to 545 annotated genes on chromosome 22. Although no standard measure of relative spliceform frequencies for human genes exists, coverage of exon boundaries by ESTs provides a measure of the diversity of isoforms for a particular gene. The incidence of captured ESTs spanning exon junctions may also provide a reasonable, though noncomprehensive, view of transcript diversity and expression. Detection of transcripts displaying exon skipping was performed using novel software, `j_explorer`, which reduced the complexity of the gene sequences to a set of possible splice junctions that were used to search public EST databases to identify ESTs spanning the annotated exon–exon junctions. The software employs

[1]These authors contributed equally to this work.
[2]Present address: Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA, USA.
[3]Corresponding author.
E-MAIL winhide@sanbi.ac.za; FAX .

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.188001.

11:1848–1853 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

Table 1. Selection and Exon Structure of Genes for Study

Number of multiple-exon genes selected for study	347
Number of exons	3240
Number of exon junctions	2893
Mean exon length	254 bp
Minimum exon length observed	8 bp
Maximum exon length observed	7660 bp
Maximum number of exons observed in one gene	54

We selected 347 multiple-exon genes of a total 545 genes present on chromosome 22 for study. Those removed included 134 single-exon genes and 64 double-exon genes that could not be assessed for exon skipping.

standard data format (EMBL sequence format) and visualization tools (ARTEMIS, Rutherford et al. 2000) in the analysis (http://www.sanbi.ac.za/exon_skipping/). Removal of single and double exon genes reduced the set to 347 multi-exon genes (Table 1), of which 10 were annotated previously in literature or public databases as having experimentally confirmed exon skips (Table 2). Exon-skipping events were recorded when all original junctions involved in the skipping event, including flanking exons, were confirmed by EST sequences. All ESTs supporting exon-skipping events were subsequently confirmed to be unambiguous transcripts of the corresponding gene and not products of paralogous genes, pseudogenes, or related members of an extended gene family by BLAST searches against the nonredundant (nr) database at NCBI. Highly specific identification of exon-skipping and exon-repetition events has resulted.

Sensitivity was assessed using the 10 genes with experimentally confirmed exon skipping. j_explorer accurately identified the previously reported skipped exons in four of the genes (*NF2, ADSL, CLTCL,* and *GGT1*). Novel isoforms were detected in *EWSR1, PLA2G6,* and *GGT1* (Table 2), whereas previously described exon-skipping events in four genes (*CACNA1I, BZRP, MTMR3, SEP3*) were not detected because ESTs mapping to these exon junctions were not available in the public EST databases. The approach yields zero false positives, as confirmed by available mRNA and genomic data, and provides a solid basis for the development of models of transcript diversity that can be generated from a single gene.

We have discovered 62 exon-skipping events in 52 genes (Table 2); 57 of the 62 (92%) exon-skipping events occur within the protein-coding region. The remainder occur in either the 3′ (1/62) or 5′ (4/62) untranslated region (UTR). In 31/62 (50%) of cases the reading frame is maintained but regions are deleted. In 18/62 cases (29%) the introduction of a skip destroys the reading frame resulting in a frame shift. Proteins for the remaining 8/62 (13%) could not be reconstructed. In four cases an alternative stop codon is used, whereas in five cases there is an alternative start codon introduced.

Gene transcripts were scanned for exon repetition using similarity searching of repeated exon constructs against public EST data. A single tandem repetition of exon 2 of the F-box protein (NM_012179) was detected with high identity to EST AA569698. Exon repetition has only previously been reported in rodents (Frantz et al. 1999).

Ratios of transcript isoforms are difficult to resolve using only EST data, however using the relative capture frequency of skipped exons as a measure provides an indication of the incidence of more commonly occurring isoforms (four or more ESTs confirm the isoform with exon skipping in: *CLTCL1, ADSL, GGT1, GSTT1, HMG2L1, MFNG,* dJ222E13.1) as compared to rarer isoforms. In 47/62 (76%) cases, the reference isoform, constructed from the genomic EMBL entry, is represented more frequently than a skipped exon isoform (Table 2).

The degree to which the level of gene expression, and hence database representation, affects the probability of finding a skipped transcript was assessed using the number of EST exon–exon junction captures per gene as a relative measure of transcript representation. Three categories comprising equal gene numbers were selected: low capture, which corresponds to <14 EST matches per gene; medium capture, those from 14 to 50 EST matches per gene; and high capture, those with 50 or more EST matches per gene (Table 3). Forty-four genes had no matches to ESTs. We found that 33 have >50 EST matches per gene and that >60% of genes that demonstrate exon skipping have large numbers of ESTs matching to them. Although no relationship between degree of gene expression and extent of skipping can be determined from this study, the degree to which exon junctions are represented in transcripts reveals that highly represented genes demonstrate skipping more frequently. Ten of the 17 (59%) most highly represented multi-exon genes show exon skipping and of these, three (18%) express more than one isoform (Table 2, http://www.sanbi.ac.za/exon_skipping).

To calculate the proportion of mRNAs that may contain a skipped exon we treat each EST spanning an exon junction as an independent sample of the exon junctions. The number of times an EST spans an exon junction is 23,922. The number of times an EST spans a nonconsecutive junction is 149. The number of times an EST spans a consecutive junction is 23,773. From this we estimate that the probability that a given exon junction in a given mRNA is nonconsecutive is ~$f = 149/23{,}773$. There are 2893 exon junctions in the 347 multi-exon genes, therefore the average number of exon junctions per multi-exon genes is $m = 2893/347$. The probability that a given multi-exon mRNA has at least one nonconsecutive exon junction is therefore $1 - (1 - f)^m = 0.051$. As this estimate is derived from a large sample of exon junctions, it may be applicable as a genome-wide estimate.

DISCUSSION

Our approach precisely identifies exon skipping when EST transcript data that spans exon boundaries is available. The number of ESTs that cover an exon–exon boundary determines the likelihood of discovering an exon skip, but capture of exon-skipping events are dependent on the ratio of low-abundance to high-abundance isoforms of transcripts from the gene. The depth of transcript representation in EST databases, level of expression, and number and length of exons all contribute to the complexity of estimation of the number of genes that may have exon-skipped expressed transcripts. Estimation of the genome-wide extent of exon skipping is supported here by 52 of 347 multi-exon genes (~15%). This conservative estimate reflects the fact that only 68% of exon–exon junctions have EST coverage, and that this coverage is skewed towards over-representation of the 3′ UTRs. In contrast, 59% of multi-exon, highly EST-represented genes present exon skipping. If exon skipping is independent of the level of expression of a gene, then 59% of all multi-exon genes could exhibit skipping. The fact that increasing EST coverage results in the detection of increasing numbers of exon-

Table 2. Identification of Chromosome 22 Genes with Unambiguous Transcripts of Exon-Skipped Isoforms

Locus name	Skipped exon(s)	Effect of exon skip	No. ESTs confirming exon skip	Average no. ESTs confirming reference isoform	Database annotation
J_explorer identified an experimentally confirmed isoform					
CLTCL1‡	29	C+	6	4.0	Clathrin heavy polypeptide-like 1
ADSL‡*	12	3′	11	63.0	Adenylosuccinate lyase
NF2‡	2–3	C f/s	1	5.7	Neurofibromatosis 2 (bilateral acoustic neuroma)
J_explorer identified an experimentally confirmed isoform and a novel isoform					
GGT1‡	7	C f/s	2	—	Gamma-glutamyltransferase 1
	3	5′	4	—	
J_explorer identified a novel isoform and not the experimentally confirmed isoform					
PLA2G6‡	3 + 5	C f/s	2	1.5	Phospholipase A2 groups VI
EWS‡*	6	C+	1	48.0	Ewing sarcoma breakpoint region 1
Novel exon skipping events identified by J_explorer					
*ATP6E**	2	C+	1	73.5	ATPase H+ transporting lysosomal (vacuolar proton pump)
	5–7	C+ 3′*t*	1	38.0	31kD
MIL1	3	C+	2	13.0	*Homo sapiens* MIL1 protein
*UFD1L**	2–3	C+ 3′*t*	1	21.7	Ubiquitin fusion-degradation 1-like
TR	13	C+	1	6.0	Thioredoxin reductase beta
ARVCF	19	C+	1	1.5	Armadillo repeat gene deletes in velocardiofacial syndrome
AC005500.4	2–3	C+	1	5.4	Zinc finger protein
PIK4CA	36–42	C f/s	1	5.6	Phosphatidylinositol 4-kinase catalytic alpha polypeptide
BCR	20	C f/s	1	5.5	Active BCR-related gene
AP000350.2	5	C+	2	1.0	Similar to glucose transporters SW:P22732
GSTT1	2	C f/s	4	7.5	Glutathione S-transferase theta 1
	2–3	C f/s	1	8.0	
	3–4	C+	1	9.3	
AC004997.1†	5	C N/A	1	4.0	GATS protein
	5–6	C+	1	4.0	
SEC14L2	10	C f/s	2	4.5	SEC14 (S. cerevisiae)-like 2
SMTN†	14–15	C N/A	1	5.7	Smoothelin
dJ858B16.1	27	C+ 3′*t*	1	2.3	*Homo sapiens* mRNA for KIAA0542 protein complete cds.
AC005004.1	22–23	C+	2	1.7	*Homo sapiens* mRNA for KIAA0645 protein complete cds
HMG2L1	2	5′	4	1.5	High-mobility group protein 2-like 1
	5	5′	1	4.0	
	2 + 5	5′	1	4.6	
CE132D12.1	6	C f/s	1	21.5	Similar to RAS-related protein RAB-5A (HS)
	7	C f/s	5	46.0	
MFNG†	2	C+	1	16.0	Manic fringe (*Drosophila*) homolog
*LGALS1**	3	C f/s	5	>100.0	Lectin galactoside-binding soluble 1 (galectin 1)
GCAT	2–3 + 5	C N/A	1	2.8	Glycine C-acetyltransferase (2-amino-3-ketobutyrate-CoA ligase)
dJ1014D13.1*	12	C+	1	>100.0	Weakly similar to casein kinase 1 homolog HRR25
GTPBP1	2	C+ 5′*t*	8	17.0	GTP binding protein 1
dJ508I15.1	2	C+ 5′*t*	1	8.0	Novel human gene mapping to chromosome 22
dJ508I15.4	3	C N/A	1	1.5	*Homo sapiens* mRNA for KIAA0668 protein
RPL3†*	8	C+	7	>100.0	Ribosomal protein L3
dJ1042K10.2	2	C+ 5′*t*	1	10.5	Similar to *C. elegans* predicted protein with probable rabGAP domains and src homology
SLC25A17	2–4	C f/s	1	9.0	Solute carrier family 25 (mitochondrial carrier-peroxisomal membrane protein 34kD) member 17
	3–4	C+	2	9.0	
*ST13**	8	C f/s	1	17.5	Suppression of tumorigenicity 13 (Hsp 70-interacting protein)

Table 2. (*Continued*)

Locus name	Skipped exon(s)	Effect of exon skip	No. ESTs confirming exon skip	Average no. ESTs confirming reference isoform	Database annotation
RBX1	2	C f/s	3	50.5	Ring-box 1
	3–4	C+ 3′*t*	1	92.0	
PMM1	4	C f/s	1	24.5	Phosphomannomutase 1
TCF20 (*AR1*)†	3	C N/A	1	5.0	Transcription factor 20 (AR1)
dJ222E13.3†	7	C f/s	2	12.0	Weak match to *Arabidopsis* RNA and export factor binding protein
dJ222E13.1	8–9	C f/s	5	2.0	Novel protein with some similarity to *Drosophila* KRAKEN
bK1191B2.3†	3	C+	4	2.0	Weakly similar to dJ1118 COA-ACYL carrier protein transacylase
dJ796I17.2*	3	C+	1	27.0	CGI-51
NPAP60L	4	C N/A	1	11.5	Nuclear pore-associated protein 60L
dJ355C18.1	9	C+	1	1.5	Matches KIAA0027 gene with weak similarity to GTPase activating protein
ECGF1	5	C N/A	1	3.0	Endothelial cell growth factor 1 (platelet-derived)
GTSE1 (*B99*)	8	C f/s	1	13.5	*Homo sapiens* G-2 and S-phase expressed 1 (GTSE1),
dJ1163J1.4†	3	C+	1	1.0	Novel protein similar to *C. elegans* B0035.16 and bacterial tRNA (5-Methylaminomethyl-2-thiouridylate)-Methyltransferases
DGCR2	2–3	C+	1	2.3	DiGeorge syndrome critical region gene 2
AC007050.6	2	C N/A	1	11.0	*Homo sapiens* mRNA- from clone DKFZp434G1017
UBE2L3	2	C+ 5′*t*	1	71.0	Ubiquitin-conjugating enzyme E2L3
DJ756G23.3	5	C+	1	2.0	Similar to Tr:Q24191 *Drosophila* TRANSCRIPTIONAL REPRESSOR PROTEIN
bK212A2.1	2	C+ 3′*t*	1	—	TNF-inducible protein CG12-1 mRNA
*G22P1**	3	C+	1	>100.0	Thyroid autoantigen 70kD (Ku antigen)

We tested 347 multiple-exon genes on Chromosome 22 for exon-skipping events using `J_explorer` and EST sequences from GenBank 119. Genes in which novel exon skipping events have been identified are ordered according to their relative physical organization along chromosome 22. Genes are identified using the HUGO name if one exists. In the absence of a HUGO identifier, the accession number of the sequence or the Sanger Centre clone name is used. Exon numbering is based on the exon structure of the original EMBL entries obtained from the Sanger Centre. ESTs confirming a skip were required to span both the 3′ and 5′ flanks of the skipped exon. To calculate the average number of ESTs confirming the reference isoform, the exon flanking ESTs in the reference isoform were totalled and the sum divided by corresponding averaged number of junctions. In cases where the reference isoform was not represented in the public EST databases, the sequence was confirmed using a corresponding experimentally-determined mRNA.

Skip location and context is denoted as follows: (C) skip occurs in protein coding region; (+) ORF remains unchanged; (3′) skip occurs in 3′ UTR; (5′) skip occurs in 5′UTR; (f/s) frameshift is introduced by skip; (5′*t*) alternative start codon is used; (3′*t*) alternative stop codon is used; (N/A) not possible to reconstruct a protein; (†) genes (eight entries total) with an already-annotated exon skip in EMBL entries; (‡) (six entries total) with experimentally-confirmed notation: 2–4 indicates that exons 2, 3 and 4 skipped exons; 2–3 +5 indicates that exons 2, 3 and 5 are skipped. Experimentally confirmed skipping events in the genes *CACNA1I, BZRP, MTMR3,* and *SEP3* had no EST matches and are not included.

skipping events indicates that exon-skipped transcripts are relatively rare. We have estimated the probability of detecting an exon-skipped transcript from a pool of multi-exon transcripts to be ~5%. More sensitive transcript capture techniques may discover exon skipping to be far more widespread than the previous estimates of ~10%–~20% (Mironov et al. 1999a; Croft et al. 2000) (http://industry.ebi.ac.uk/~thanaraj/gene.html), which have been based on EST frequency-independent measures. Expression studies will clarify the relationship between level of expression and degree of exon skipping in transcripts. The diversity of skipped-exon transcript forms is likely to contribute significantly to the diversity of protein products encoded by the genome, especially because the ratio of skipped isoforms of transcripts appears to vary widely, which is likely to have significant functional impact on the proteins for which they code. At least 50% of exon skips that we have detected result in in-frame deletions in the predicted protein products. In 29% of cases, exon skipping results in a disruption of the reading frame which may change or disrupt the function of the protein product. Functional

Table 3. Capture of Exon Skipping Relative to Expression Representation

Number (n) of ESTs matching exon junctions per gene (interval)	Number of genes	Number of genes with skips detected by J_explorer
0 < n ≤ 14	101 (33%)	4 (7%)
14 ≤ n ≤ 50	101 (33%)	17 (33%)
n ≥ 50	101 (33%)	33 (60%)
Total	303	52

roles for these protein isoforms remain to be explored experimentally.

METHODS

`j_explorer` (available for download from http://www.sanbi.ac.za/exon_skipping) was used to assemble exon constructs from mRNA-annotated genomic sequences produced by the Human Chromosome 22 Sequencing Group at the Sanger Centre (Chr22.genes.dna file at http://www.sanger.ac.uk/HGP/Chr22/cwa_archive/Nature_02–12–1999/Chr22Genes.tar.gz). Using a 50-bp tag from the 3′ terminus of the preceding exon and a 50-bp tag from the 5′ terminus of all downstream exons, a set of all consecutive and nonconsecutive exon–exon junctions for each gene was created. Each junction was submitted for similarity searching against dbEST (human) using `BLAST` 2.0 (Altschul et al. 1990). By combining junctions in a consecutive (i.e., exon 1–exon 2 junction) and nonconsecutive (i.e., exon 1–exon 3 junction) manner the incidence of exon skipping was assessed. A skipping event is reported when an EST is detected that does not contain the exon(s) in question, but does contain an uninterrupted tag made up of 50 bp from each of the flanking exons. Exon repetition was investigated by creating splice junctions composed of the concatenation of the 3′ and 5′ 50-bp splice junctions of the same exon. ESTs showing significant ($P<1 \times 10^{-40}$) homology to an exon junction were extracted and aligned to the corresponding genomic sequence using `sim4` (Florea et al. 1998). To exclude the possibility that ESTs confirming exon-skipping events were the products of paralogous genes or members of gene families, all ESTs identifying exon skipping were confirmed to be unique to a single target gene from Chromosome 22. Both interchromosomal and intrachromosomal specificity of the transcripts was confirmed using `BLAST` with a cut-off score of 1×10^{-30}. `sim4` was employed where ambiguous matches were encountered. The resulting 'unambiguous transcripts' can therefore be assigned unambiguously to the correct gene of origin. The effect of these transcripts on the reading frame of the protein for which they code was assessed for frameshifts and in-frame deletions. We have confirmed that the results contain no false positives by manual analysis of all cases of possible misalignments. All exon skips reported by `j_explorer` therefore represent valid skipping events. The identity and genomic location of each the ESTs was converted into EMBL format and added as annotation to the relevant EMBL sequence file. Sequences were then analyzed using `ARTEMIS` (Rutherford et al. 2000) and are presented together with supplemental information, annotated EMBL entries, and links to ENSEMBL genes and transcripts at http://www.sanbi.ac.za/exon_skipping. All exon structure annotations for the genes used (both confirmed and predicted) were confirmed by manual inspection to be correct. To prevent the detection of skips as a result of incorrectly annotated exon boundaries we required that an EST spanning consecutive (or linear) exon boundaries was present in addition to the ESTs confirming the skip. All linear junctions that could not be confirmed by ESTs resulted in that junction being excluded from further analysis. To address data consistency, we confirmed that in EMBL release 64 (GenBank 119) and 65 (GenBank 121) ~68% of splice junctions are covered with an EST. This figure does not vary significantly between the two releases.

To determine the effect of exon skipping on protein production, the genomic sequences of the 52 genes with exon-skipping events affecting the coding sequence were compared to the cognate mRNA using `sim4`. Exon-skipping events in the first or last protein-coding exon were recorded as altering the start/stop codon. A UTR-skipping event was recorded when the exon skipped was located in either the 3′ or 5′ UTR. When the skipped exon was located between two coding exons, `framefinder` (Slater 2000) was used to predict open reading frames (ORFs) for the ESTs confirming the skip. If the ORF showed homology to the protein isoform then it was considered to be a valid representation of the protein isoform. `BLASTX` alignment of the EST and the predicted protein was used to determine whether or not the exon-skipping event introduced a frameshift.

ACKNOWLEDGMENTS

We thank L. Sullivan and M. Ramsay for comments on the manuscript, and S. Bardien-Kruger, T. Broveak, and T-M. Chern for support, suggestions, and helpful discussions throughout the study. This work had financial support from the South African Government through the Department of Arts, Culture, Science, and Technology initiated Innovation Fund Program, grant 32146 (W.A.H.) and the South African National Research Foundation (J.F.K).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Andreadis, A., Gallego, M.E., and Nadal-Ginard, B. 1987. Generation of protein isoform diversity by alternative splicing: Mechanistic and biological implications. *Annu. Rev. Cell Biol.* **3:** 207–242.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10:** 1001–1010.

Brett, D., Lehmann, G., Hanke, J., Gross, S., Reich, J., and Bork, P. 2000. EST analysis online: WWW tools for detection of SNPs and alternative splice forms. *Trends Genet.* **16:** 416–418.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J.S. 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat. Genet.* **24:** 340–341.

Dufour, C., Weinberger, R.P., Schevzov, G., Jeffrey, P.L., and Gunning, P. 1998. Splicing of two internal and four carboxy-terminal alternative exons in nonmuscle tropomyosin 5 pre-mRNA is independently regulated during development. *J. Biol. Chem.* **273:** 18547–18555.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.

Frantz, S.A., Thiara, A.S., Lodwick, D., Ng, L.L., Eperon, I.C., and Samani, N.J. 1999. Exon repetition in mRNA. *Proc. Natl. Acad. Sci.* **96:** 5400–5405.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., and Claverie, J.M. 1998. Alternate polyadenylation in human mRNAs: A large-scale analysis by EST clustering. *Genome Res.* **8:** 524–530.

Iida, Y. 1997. A mechanism for unsplicing and exon skipping in human alpha- and beta-globin mutant pre-mRNA splicing. *Nucleic Acids Symp. Ser.* 183–184.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Jiang, Z.H. and Wu, J.Y. 1999. Alternative splicing and programmed cell death. *Proc. Soc. Exp. Biol. Med.* **220:** 64–72.

Kawahara, H., Kasahara, M., Nishiyama, A., Ohsumi, K., Goto, T., Kishimoto, T., Saeki, Y., Yokosawa, H., Shimbara, N., Murata, S., et al. 2000. Developmentally regulated, alternative splicing of the Rpn10 gene generates multiple forms of 26S proteasomes. *EMBO J.* **19:** 4144–4153.

Lambert de Rouvroit, C., Bernier, B., Royaux, I., de, B., V, and Goffinet, A.M. 1999. Evolutionarily conserved, alternative splicing of reelin during brain development. *Exp. Neurol.* **156:** 229–238.

Lim, S., Naisbitt, S., Yoon, J., Hwang, J.I., Suh, P.G., Sheng, M., and Kim, E. 1999. Characterization of the Shank family of synaptic proteins. Multiple genes, alternative splicing, and differential expression in brain and development. *J. Biol. Chem.* **274:** 29510–29518.

Mercatante, D. and Kole, R. 2000. Modification of alternative splicing pathways as a potential approach to chemotherapy. *Pharmacol. Ther.* **85:** 237–243.

Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999a. Frequent alternative splicing of human genes. *Genome Res.* **9:** 1288–1293.

Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S. 1999b. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* **27:** 2981–2989.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. 2000. Artemis: Sequence visualization and annotation. *Bioinformatics* **16:** 944–945.

Schweighhoffer, F. 2000. Qualitative gene profiling: A novel tool in genomics and pharmacogenomics that deciphers messenger RNA isoforms diversity. *Pharmacogenomics* **1:** 187–197.

Slater, G. 2000. "Algorithms for analysis of ESTs." Ph.D. thesis, University of Cambridge, UK.

Strehler, E.E. and Zacharias, D.A. 2001. Role of alternative splicing in generating isoform diversity among plasma membrane calcium pumps. *Physiol Rev.* **81:** 21–50.

Thanaraj, T.A. 1999. A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures. *Nucleic Acids Res.* **27:** 2627–2637.

Unsworth, B.R., Hayman, G.T., Carroll, A., and Lelkes, P.I. 1999. Tissue-specific alternative mRNA splicing of phenylethanolamine N-methyltransferase (PNMT) during development by intron retention. *Int. J. Dev. Neurosci.* **17:** 45–55.

Valentine, C.R. 1998. The association of nonsense codons with exon skipping. *Mutat. Res.* **411:** 87–117.

Wolfsberg, T.G. and Landsman, D. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25:** 1626–1632.

Zacharias, D.A., Dalrymple, S.J., and Strehler, E.E. 1995. Transcript distribution of plasma membrane Ca2+ pump isoforms and splice variants in the human brain. *Brain Res. Mol. Brain Res.* **28:** 263–272.

Received March 9, 2001; accepted in revised form August 24, 2001.