Nucleic Acids Research, 2005, Vol. 33, Database issue ***D75–D79***
doi:10.1093/nar/gki118

ECgene: genome annotation for alternative splicing

Pora Kim, Namshin Kim[1], Younghee Lee, Bumjin Kim, Youngah Shin and Sanghyuk Lee*

Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea and [1]School of Chemistry, Seoul National University, Seoul 151-747, Korea

Received August 15, 2004; Revised October 9, 2004; Accepted October 20, 2004

ABSTRACT

ECgene provides annotation for gene structure, function and expression, taking alternative splicing events into consideration. The gene-modeling algorithm combines the genome-based expressed sequence tag (EST) clustering and graph-theoretic transcript assembly procedures. The website provides several viewers and applications that have many unique features useful for the analysis of the transcript structure and gene expression. The summary viewer shows the gene summary and the essence of other annotation programs. The genome browser and the transcript viewer are available for comparing the gene structure of splice variants. Changes in the functional domains by alternative splicing can be seen at a glance in the transcript viewer. We also provide two unique ways of analyzing gene expression. The SAGE tags deduced from the assembled transcripts are used to delineate quantitative expression patterns from SAGE libraries available publically. Furthermore, the cDNA libraries of EST sequences in each cluster are used to infer qualitative expression patterns. It should be noted that the ECgene website provides annotation for the whole transcriptome, not just the alternatively spliced genes. Currently, ECgene supports the human, mouse and rat genomes. The ECgene suite of tools and programs is available at http://genome.ewha.ac.kr/ECgene/.

INTRODUCTION

Alternative splicing (AS) is a major mechanism of increasing transcriptome diversity in eucaryotic genomes (1–3). Recent studies on AS estimated that 40–70% of human genes show evidence of encoding transcripts that are alternatively spliced (4–6). Numerous databases on AS, mostly concentrating on specific aspects of gene structure (7–14), have been published so far.

Databases on AS are generated either by data mining the experimental databases such as GenBank, Swiss-Prot/TrEMBL and Medline, or by comparing sequence alignments. The former includes ASDB (14), Xpro (12) and AEdb in ASD (7). Computational approaches are even more diverse. Many attempts compared the expressed sequence tag (EST) alignments with mRNA or known gene sequences (10,12, 13,15,16). However, examining genomic alignment of EST and mRNA sequences has many advantages once the genome map is available. This includes AltExtron (17), TAP (5), ASAP (8) and altGraphX (18). Notably, the European Bioinformatics Institute (EBI) launched the Alternative Splicing Database (ASD) consortium to annotate AS events (7). Currently, three major databases are available in ASD—AltSplice and AltExtron from computational pipeline and AEdb which is manually generated from the literature. AceView at the NCBI is another website that provides ample information on alternatively spliced genes (D. Thierry-Mieg, J. Thierry-Mieg, M. Potdevin and M. Sienkiewicz, unpublished data; http://www.ncbi.nih.gov/IEB/Research/Acembly/).

Recently, we developed a gene prediction program, ECgene (Gene prediction by EST Clustering), taking alternative splicing into consideration. The algorithm combines the genome-based EST clustering and graph-theoretic transcript assembly procedures in a coherent fashion. A brief summary of the algorithm was described with the ASmodeler work, the web server version (19). Further details of the algorithm will be published elsewhere (N. Kim, S. Shin and S. Lee, manuscript submitted).

In this paper, we describe the ECgene database and website with several application programs. Besides the gene prediction algorithm, our viewers and applications have many unique features useful for analyzing the transcript structure, gene function and expression pattern.

ECgene DATABASE

The ECgene algorithm has been applied to the human, mouse and rat genomes to produce transcriptomes that include alternative splicing events. ECgene transcriptome must be one of the most complete versions even though it contains many unreal transcripts. We classify the transcript models into

*To whom correspondence should be addressed. Tel: +82 2 3277 2888; Fax: +82 2 3277 2384; Email: sanghyuk@ewha.ac.kr

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors

three groups according to reliability. ECgene Part A represents transcripts of almost RefSeq quality with a clone evidence covering all exons of the transcript. Part B includes transcripts of slightly lower quality (but still highly probable). Their transcript structure requires concatenation of minimum two clones to recover all exons in the transcript. No single clone is available to be full-length for transcript model. All other transcripts belong to Part C and are of low reliability. Their gene structure may be questionable since it requires concatenation of more than two clones. However, individual event of alternative splicing implied in the transcripts should be real unless the genomic alignment of mRNA/EST sequences is wrong. They have a fair chance to be real transcripts with more sequence data. Statistics on the number of genes and transcripts is available at the website.

WEB INTERFACE AND APPLICATIONS

The ECgene website contains several independent but closely related applications—the summary viewer, genome browser, alignment viewer, ECfunction and ECexpression. The summary viewer shows the overall picture of the gene. Each application page provides more detailed information and diverse options to explore a specific aspect of the gene.

Summary viewer

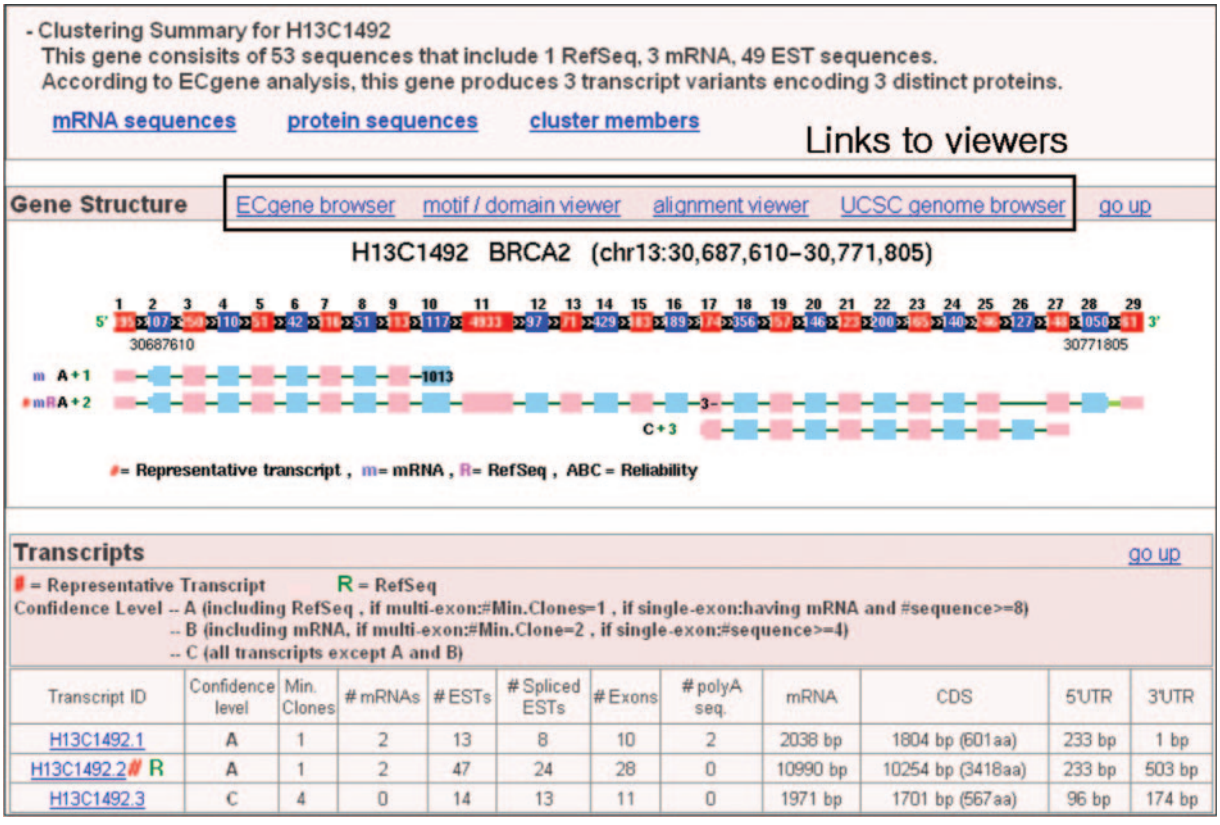
The summary viewer shows the gene summary and the essence of other annotation programs. It supports querying by gene symbol, accession number of mRNA/EST and the ECgene ID. Figure 1 shows a part of output from the ECgene summary page. The output page shows a brief summary of the gene and the cluster, a graphic illustration of the transcript structure, the summary table of splice variants, functional analysis from the InterPro (20) and Swiss-Prot databases (21), and gene expression analysis using SAGE and cDNA libraries. Links to other well-known databases, predicted mRNA and protein sequences, and sequence members in the cluster are shown at the bottom of the page. Graphic images are from ECfunction and ECexpression, which are described in later sections.

Links to various viewers are given next to the gene structure in Figure 1. We provide our own genome browser and the transcript/motif/domain viewer whose details are described in the following sections. The alignment viewer (picture not shown) can be used to see the alignment and tissue source of ESTs (organ and cancer). Viewers via these links use the default options, and users should access the application page for various options.

Clicking on the variant in the picture or in the table opens the mRNA page. This page provides more detailed information on the selected transcript in a similar interface. Domain/motif information is added in the picture.

ECgene genome browser

The transcript structure and cluster members can be seen via the ECgene genome browser that adds ECgene models as custom tracks in the UCSC genome browser. Figure 2 shows the transcript structure of the *BRCA2* gene using the ECgene genome browser available at http://genome.ewha.ac.kr/ECgene/gbr/.



Transcript ID	Confidence level	Min. Clones	# mRNAs	# ESTs	# Spliced ESTs	# Exons	# polyA seq.	mRNA	CDS	5'UTR	3'UTR
H13C1492.1	A	1	2	13	8	10	2	2038 bp	1804 bp (601aa)	233 bp	1 bp
H13C1492.2# R	A	1	2	47	24	28	0	10990 bp	10254 bp (3418aa)	233 bp	503 bp
H13C1492.3	C	4	0	14	13	11	0	1971 bp	1701 bp (567aa)	96 bp	174 bp

Figure 1. Part of output from the ECgene summary viewer. Clustering summary, gene structure and the table for transcript details are shown. Three splice variants are predicted for *BRCA2* gene.

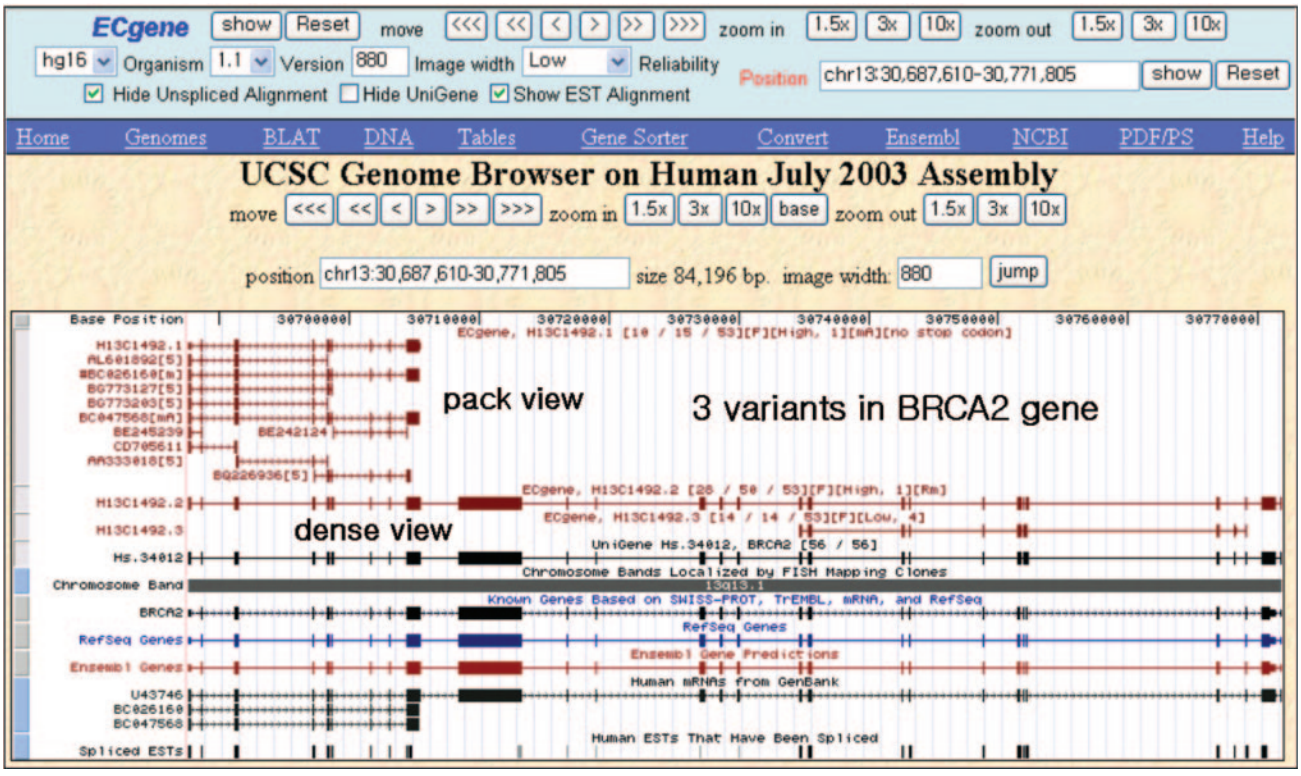


Figure 2. ECgene genome browser. Query field, options and navigating bars are provided in the upper window. If the option of showing EST alignment is unchecked, it will show all transcript models within a single track. Options for displaying unspliced alignments and the UniGene are available. '[10/15/53][F][High, 1][mA][no stop codon]' in the title line means that this cluster has 53 sequences, 15 of which belong to this specific variant and 10 sequences are multi-exon clones. The transcript is on the sense (+) strand. It has mRNA and poly(A) evidence. [High, 1] means that the transcript belongs to the ECgene part A.

The GUI design is almost identical to the UCSC genome browser as shown in Figure 2. The most useful feature would be the option of showing the EST alignment that adds each transcript model and member sequences as a separate custom track. The title line includes a brief summary of the transcript model and clones. Clicking on the transcript or the title line expands the picture to show the alignment in the pack-view mode. We also provide the option of hiding unspliced alignments since many of them are incomplete or are artifacts. Furthermore, one can see the result of UniGene clustering for comparison. It is just a collection of alignments without transcript models. We find that all 56 sequences in the UniGene cluster for the *BRCA2* gene appear in this region.

ECfunction—transcript/domain/motif viewer

ECfunction available at http://genome.ewha.ac.kr/ECgene/ECfunction/ collects information related to gene function. It includes the InterPro domains, transmembrane helices, signal peptides, BLAST hits and GO annotations.

ECfunction has a unique transcript viewer as shown in Figure 3. It shows the transcript structure in the mRNA coordinates where introns are depicted as short lines of fixed length. Overlapping exons are grouped together to show all splice variants on a single ruler scale given above the picture. The nucleotide ranges on the exons indicate exons from splice site variation so that one can see the difference of splice variants at a glance. It should be noted that merging overlapping exons to generate the master coordinate might create unrealistically long exons in case the introns are retained in one of the transcripts. To remedy this problem, we provide the option of removing intron retention events. Alternatively, the user may choose the specific isoforms to draw in the transcript viewer.

Being on the mRNA coordinate, the transcript viewer can be used to show the functional domains and motifs. This enables the user to recognize immediately any loss of functional domains due to alternative splicing.

ECexpression

ECexpression available at http://genome.ewha.ac.kr/ECgene/ECexpression/ analyzes the gene expression pattern using ~260 SAGE and ~8600 cDNA libraries available publically. Our SAGE analysis is much different from the NCBI's SAGEmap (22) in that SAGE tags are extracted from the assembled transcript models, not from EST sequences in the cluster. The clustering is also different from the UniGene even though both are genome-based at this point. We explain here the output pictures briefly.

Figure 4 shows the normalized tag frequencies (i.e. tags per million count) for 28 organs with public SAGE data. In cases the splice variants have different SAGE tags, SAGE tags can be used to find variant-specific gene expression. We divide the normal and cancer libraries whose tag frequencies are depicted in the upper and lower half of the picture, respectively. This layout makes it particularly easy to find any tissue-specific or

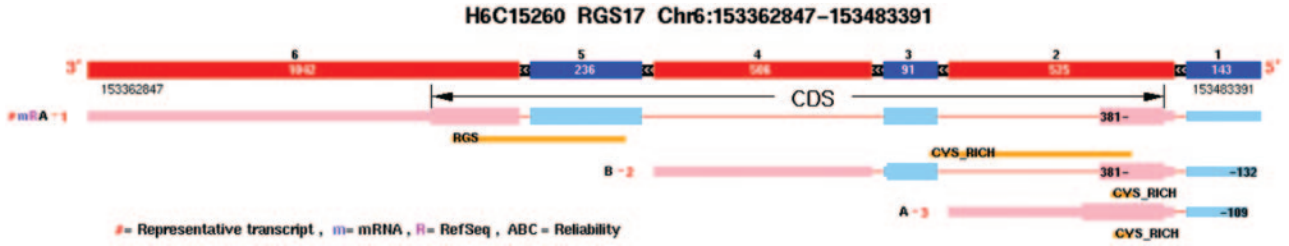


Figure 3. Output picture from the ECfunction. RGS17 (regulator of G-protein signaling 17) has three splice variants encoding three distinctive proteins. The CDS (coding sequence) region is indicated in tall boxes. Note that variants #2 and #3 lack the RGS domain.

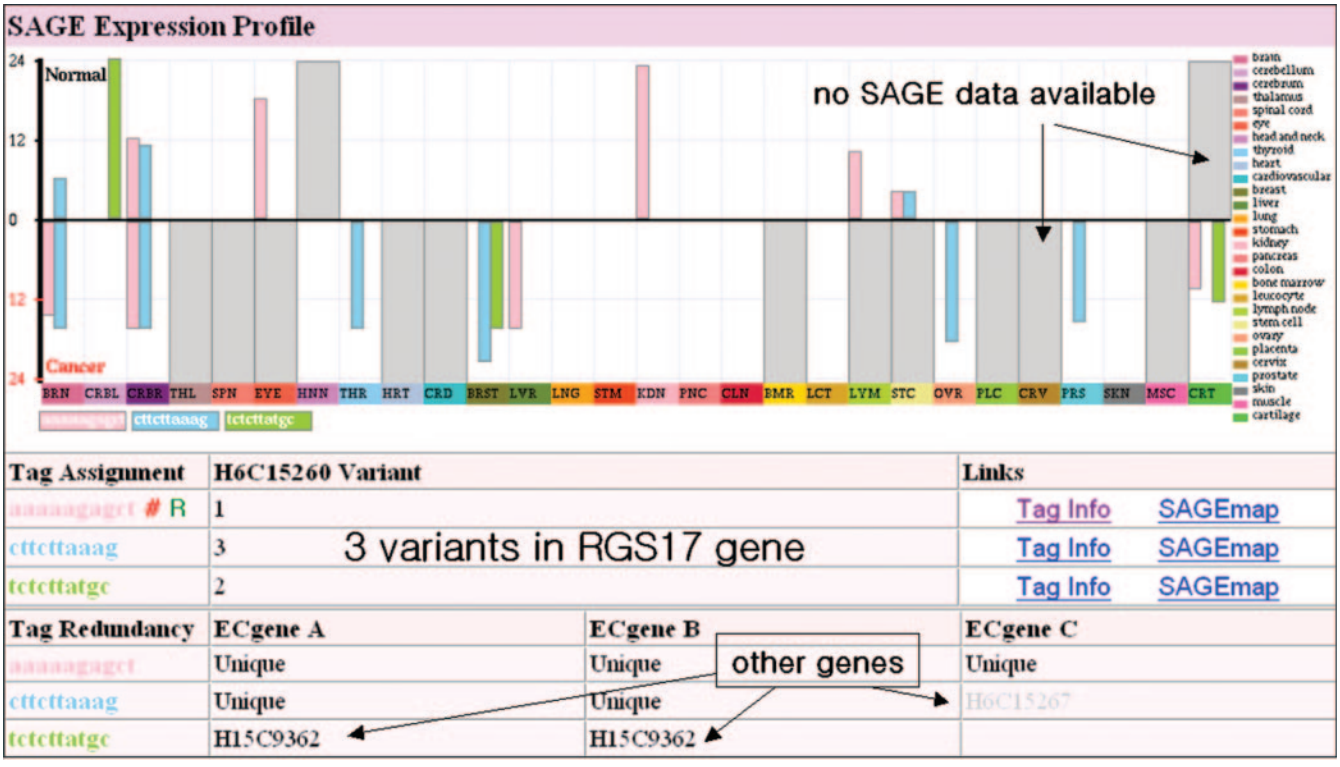


Figure 4. SAGE expression profile for RGS17. Three variants have different SAGE tags. Two tags are unique among the ECgene transcriptome. Tag for variant #2 (green tag) is the same with other gene H15C9362 (Hs.150824 in the UniGene), thereby prediction being less reliable. Variant #1 with the pink tag seems to be expressed in brain, kidney, eye and liver. One should examine critically the expression whose actual tag frequency is one (kidney, eye and liver). The ECexpression page provides the option of suppressing tags with low frequency.

cancer-specific isoforms that have tremendous value for clinical application.

Gene expression from cDNA libraries is quite similar except that the graph represents the number of EST sequences from the specified tissue in the cluster. We manually classified ~8600 human cDNA libraries in terms of tissue (organ), pathology and developmental stage. Gene expression predicted by using EST members is mostly qualitative since many cDNA libraries are normalized or subtracted to find genes with low expression level. However, the coverage of cDNA libraries is much more extensive as can be seen in the number of libraries available publically.

FUTURE DIRECTIONS

ECgene is an ongoing project. It aims to be one of the reference sites for genome annotation. A number of new features and applications are currently under development. Gene profiler, ortholog finder and ChimerDB are among the applications to be completed in the near future. We plan to expand and optimize the system resources to shorten the response time. Other model organisms will be added too.

ACKNOWLEDGEMENTS

We are grateful to the UCSC genome center for making such a wonderful resource available to the public. This work was supported by the Ministry of Science and Technology of Korea through the bioinformatics research program of MOST NRDP (M1-0217-00-0027) and the Korea Science and Engineering Foundation through the Center for Cell Signaling Research at Ewha Womans University.

REFERENCES

1. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

2. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
3. Black,D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, **72**, 291–336.
4. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
5. Kan,Z., States,D. and Gish,W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res.*, **12**, 1837–1845.
6. Johnson,J.M., Castle,J., Garrett-Engele,P., Kan,Z., Loerch,P.M., Armour,C.D., Santos,R., Schadt,E.E., Stoughton,R. and Shoemaker,D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141–2144.
7. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res.*, **32**, D64–D69.
8. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res.*, **31**, 101–105.
9. 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.
10. Pospisil,H., Herrmann,A., Bortfeldt,R.H. and Reich,J.G. (2004) EASED: Extended Alternatively Spliced EST Database. *Nucleic Acids Res.*, **32**, D70–D74.
11. Huang,H.D., Horng,J.T., Lee,C.C. and Liu,B.J. (2003) ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. *Genome Biol.*, **4**, R29.
12. Gopalan,V., Tan,T.W., Lee,B.T. and Ranganathan,S. (2004) Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Res.*, **32**, D59–D63.
13. Krause,A., Haas,S.A., Coward,E. and Vingron,M. (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.*, **30**, 299–300.
14. Dralyuk,I., Brudno,M., Gelfand,M.S., Zorn,M. and Dubchak,I. (2000) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.*, **28**, 296–297.
15. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
16. Zavolan,M., van Nimwegen,E. and Gaasterland,T. (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.*, **12**, 1377–1385.
17. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.
18. Sugnet,C.W., Kent,W.J., Ares,M.,Jr and Haussler,D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, 66–77.
19. Kim,N., Shin,S. and Lee,S. (2004) ASmodeler: gene modeling of alternative splicing from genomic alignment of mRNA, EST and protein sequences. *Nucleic Acids Res.*, **32**, W181–W186.
20. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
21. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
22. Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.